

**Can Artificial Intelligence applied to a pre-treatment  $^{18}\text{F}$ -  
FDG PET/CT scan be used to predict two-year disease free  
survival for patients with Oesophageal Cancer?  
(CALIFORNIA)**

A thesis submitted to the University of Manchester for the degree of Doctor of Clinical  
Science in the Faculty of Biology, Medicine and Health.

**2023**

**Nicholas J Vennart**

# Contents

<b>1</b>	<b>Abstract .....</b>	<b>10</b>
<b>2</b>	<b>Declarations and Copyright .....</b>	<b>11</b>
2.1	Declaration .....	11
2.2	Copyright Statement.....	11
2.3	Acknowledgements .....	12
<b>3</b>	<b>Literature Review.....</b>	<b>13</b>
3.1	Oesophageal cancer .....	13
3.1.1	Introduction to oesophageal cancer.....	13
3.1.2	Diagnosis and staging of oesophageal cancer.....	15
3.1.3	Treatment and treatment response of oesophageal cancer .....	18
3.2	PET/CT Technology .....	20
3.2.1	Introduction to PET/CT .....	20
3.2.2	Standard Uptake Value (SUV) .....	22
3.2.3	Technological Advances in PET: Time of Flight.....	23
3.2.4	Technological Advances in PET: Point Spread Function .....	24
3.2.5	Technological Advances in PET: Block Sequential Regularized Expectation Maximum iterative reconstruction .....	28
3.2.6	Motion Correction.....	29
3.3	Radiomics.....	30
3.3.1	Introduction to Radiomics .....	30
3.3.2	Radiomics and Tumour Phenotype.....	32
3.3.3	Radiomics and “Texture” Analysis .....	34
3.3.4	Radiomics in PET .....	38
3.3.5	Effect of Reconstruction on Radiomics Parameters.....	43

3.4	Artificial intelligence and machine learning .....	46
3.4.1	Introduction and history of Artificial Intelligence .....	46
3.4.2	Machine Learning .....	48
3.4.3	Deep learning architecture.....	53
3.5	Systematic Review .....	57
3.6	Research Proposal .....	66
<b>4</b>	<b>Methodology.....</b>	<b>67</b>
4.1	Ethics and Health Research Authority Approval.....	67
4.2	Overview of Methodology.....	68
4.3	Patient Selection .....	70
4.3.1	Patient Cohort.....	70
4.3.2	Image Acquisition and Transfer.....	72
4.4	Software Review.....	74
4.4.1	Radiomics Software .....	74
4.4.2	AI software.....	75
4.5	Image Analysis and Tumour Threshold Method .....	76
4.6	Radiomics Analysis.....	77
4.6.1	OSEM only radiomics data .....	77
4.6.2	OSEM and BSREM radiomics data.....	77
4.7	Machine Learning .....	80
4.7.1	Overview .....	80
4.7.2	Train, Test and Validation .....	81
4.7.3	Summary of the tests evaluated.....	84
<b>5</b>	<b>Results .....</b>	<b>88</b>
5.1	Survival prediction with radiomics.....	88

5.1.1	Clinical Data.....	88
5.1.2	Radiomics.....	90
5.2	OSEM vs BSREM.....	98
5.2.1	Radiomics Comparison .....	99
5.3	Machine Learning .....	107
5.3.1	Pilot Study.....	107
5.3.2	Survival prediction with machine learning.....	109
5.3.3	OSEM vs BSREM.....	112
5.3.4	Summary of machine learning results .....	114
5.4	Summary of key results.....	117
<b>6</b>	<b>Discussion.....</b>	<b>118</b>
6.1	Survival prediction with radiomics.....	118
6.1.1	Clinical Data.....	118
6.1.2	Radiomics and clinical features.....	121
6.2	OSEM vs BSREM.....	126
6.3	Machine Learning .....	130
6.3.1	Pilot .....	130
6.3.2	Survival prediction with machine learning.....	131
6.3.3	Machine learning method comparison for BSREM and OSEM.....	134
6.4	Further Work .....	135
6.4.1	Clinical Data.....	135
6.4.2	Radiomics – limitations and further work.....	136
6.4.3	Machine Learning .....	138
<b>7</b>	<b>Conclusion .....</b>	<b>140</b>
<b>8</b>	<b>References .....</b>	<b>141</b>

<b>9</b>	<b>Appendices.....</b>	<b>157</b>
9.1	DClinSci Appendix – List of A units and Medical Physics B units together with assignments – Nicholas Vennart .....	157
9.2	Lay Abstract.....	159
9.3	Innovation Proposal (Business Case) .....	160
9.4	Image and region Checklist with SUV thresholds used .....	165
9.5	LiFEX v7.0.0 Patient data Download Code .....	168
9.6	Raw data for the features showing the largest difference with failed treatment....	170
9.7	Machine Learning Raw Code .....	172

## Word Count:

Total: 49627

Body of Thesis (not including reference / contents lists and appendices): 37948

Version 1: Submitted 28/09/2022, VIVA VOCE examination on 11/11/2022

Version 2: Resubmission, version 2, Submitted 05/05/2023, Accepted 27/7/2023.

## List of Figures

Figure 1 Stages I - IV oesophageal cancer .....	15
Figure 2 Treatment pathway for upper GI cancer .....	16
Figure 3 PET-CT oesophageal cancer image .....	17
Figure 4 Gastric cancer PET-CT images.....	18
Figure 5 Schematic diagram PET scanner .....	20
Figure 6 Conventional vs TOF PET .....	24
Figure 7 Measurement of Point Spread Function.....	25
Figure 8 FWHM of Point Spread Function.....	26
Figure 9 Process flow map for OSEM vs Q.Clear Reconstruction .....	29
Figure 10 Heterogeneous vs Homogeneous Tumour cell distribution .....	31
Figure 11 Radiomics analysis process map .....	33
Figure 12 Formulation of grey levels .....	34
Figure 13 A grey-level co-occurrence matrix .....	36
Figure 14 Grey-level run length matrix .....	37
Figure 15 Grey-level size zone matrix.....	37
Figure 16 Grey-level co-occurrence, run length and size zone matrix.....	38
Figure 17 Tumour thresholding methods .....	40
Figure 18 Radiomics Quality System .....	42
Figure 19 Robustness of features comparison to other works.....	45
Figure 20 History of Artificial Intelligence .....	46
Figure 21 Logistic Regression .....	49
Figure 22 Support Vector Machine.....	50

Figure 23 Random Forest Algorithm .....	52
Figure 24 Summary of machine learning algorithms .....	53
Figure 25 Machine vs deep learning .....	54
Figure 26 Example methodology flow chart for radiomics, ML and DL.....	55
Figure 27 Example deep learning system architecture.....	56
Figure 28 Flow chart overview of project methodology .....	69
Figure 29 Region thresholding in LiFEX .....	76
Figure 30 Survival curve for all upper GI patients. ....	88
Figure 31 Disease-free survival curve for all upper GI patients. ....	89
Figure 32 Survival curve for oesophageal and OGJ adenocarcinoma .....	90
Figure 33 Sample images and regions for successful treatment, failed treatment and excluded studies.....	91
Figure 34 Correlation Heatmap of all radiomics parameters downloaded for OSEM only images. ....	92
Figure 35 Heat map of the p-value of correlation coefficients .....	93
Figure 36 Features with highest average increase for failed treatments.....	96
Figure 37 Features with highest average increase for failed treatments (normalised) .....	96
Figure 38 GLZLM_ZLNU against volume for 1 and 2 year disease free survival .....	97
Figure 39 OSEM vs BSREM images for successful treatments for patients with the largest and smallest feature differences.....	99
Figure 40 OSEM vs BSREM images for failed treatments for patients with the largest differences in TLG and SUVmin. ....	100
Figure 41 SNR liver plotted against BMI.....	101

Figure 42 Absolute percentage difference for 58 radiomic features OSEM vs BSREM .....	105
Figure 43 Percentage difference for 58 radiomic features OSEM vs BSREM.....	106
Figure 44 Kaplan-Meier survival curve for oesophageal adenocarcinoma.....	107
Figure 45 Comparison of accuracy for different machine learning algorithms.....	109
Figure 46 Comparison of accuracy for different machine learning algorithms.....	110
Figure 47 Comparison of accuracy for different machine learning algorithms, 1 and 2 year disease free survival for TNM Score Only .....	111
Figure 48 Comparison of accuracy for different machine learning algorithms.....	113
Figure 49 Summary of recall scores for 2 year DFS .....	115
Figure 50 Summary of recall scores for 1 year DFS .....	116
Figure 51 Tumour scores for patients in dataset 3 .....	123
Figure 52 Nodal scores for patients in dataset 3.....	124
Figure 53 Percentage split of failed / successful treatments for GLZLM_GLNU .....	125

## List of Tables

Table 1 Summary of studies investigating oesophageal cancer with PET or PET/CT Imaging (Van Rossum, et al., 2016; Sah, et al., 2019; Xie, et al., 2021) .....	59
Table 2 Key fields extracted from NOGU Database: .....	70
Table 3 Summary of cancer location and type in cohort .....	71
Table 4 Summary of cancer location and cancer treatment .....	72
Table 5 Summary of the available radiomics programs.....	74
Table 6 A summary of Artificial Intelligence Software platforms .....	75
Table 7 Summary of radiomics parameters and definitions.....	78



Table 8 Summary of machine learning techniques used in literature.....	80
Table 9 Guide to prediction matrix .....	83
Table 10 Summary of number of Successes and Failures in each class.....	85
Table 11 Summary of versions and computational time.....	86
Table 12 Summary of machine learning tests .....	87
Table 13 Pearson's correlation coefficient >0.9 against SUVmax .....	94
Table 14 Student's T-Test comparing statistical significance of percentage differences .....	98
Table 15 Liver Noise spread .....	101
Table 16 Largest differences in radiomic features for OSEM and BSREM images ...	102
Table 17 Largest variances in the difference in radiomic features for OSEM and BSREM images.....	102
Table 18 Smallest variances in the difference in radiomic features for OSEM and BSREM images.....	103
Table 19 Average difference and variance in the difference of common radiomic features for OSEM and BSREM images .....	103
Table 20 Ten most statistically significantly different radiomic features between OSEM and BSREM .....	104
Table 21 Pilot machine learning code validation results.....	108
Table 22 False and True Positive and Negatives for SVM Algorithm .....	112
Table 23 Validation set false and true positive and negatives for LR algorithm.....	114

# 1 Abstract

---

Oesophageal cancer is one of the leading causes of cancer death in the. It is potentially curable with surgery but carries a significant risk of complication; the decision to treat is important for patients and clinicians. Several authors have explored the potential link between radiomic data in positron emission tomography / computed tomography (PET/CT) images and treatment response. We propose an artificial intelligence method to predict disease-free survival (DFS) from PET imaging for patients with upper gastro-intestinal (GI) adenocarcinoma using a larger patient cohort than has previously been described in the UK. Furthermore, we propose investigating the effect of the Block Sequential Regularized Expectation Maximum (BSREM or “Q.Clear”) BSREM image reconstruction algorithm on radiomic signatures and overall machine learning performance with 3 key questions:

1. Are radiomic feature(s) from pre-treatment PET imaging linked to DFS?
2. Can artificial intelligence predict DFS from pre-treatment PET imaging?
3. Does BSREM affect radiomic features and the ability to predict DFS?

We retrospectively analysed the staging PET/CT images of 144 patients with upper GI tract adenocarcinomas who underwent curative surgical treatment. We analysed 58 radiomic, 3 clinical features and 2 reconstruction methods (OSEM vs BSREM). We compared 6 machine learning (ML) algorithms for predicting DFS up to 2 years post treatment.

We found that larger, heterogeneously distributed tumours were associated with poorer DFS rates. Radiomic features related to grey-level run length matrix were robust to different image reconstructions but features evaluating local variations, such as grey-level co-occurrence matrix contrast, were susceptible to reconstruction method. Most ML algorithms tested did not produce sufficient accuracy for use clinically however, BSREM images with a logistic regression algorithm, provided the most clinically relevant results: an overall 75% accuracy predicting 70% of successful, and crucially, 83% of failed treatments.

Radiomic signatures from the PET images for upper GI cancer patients can aid clinicians and patients in identifying where closer monitoring for recurrence is required after surgical treatment. BSREM remains a useful tool for image quality enhancement but caution is advised when interpreting radiomic signatures. BSREM images with a logistic regression algorithm showed initial promise for predicting 2-year DFS from the radiomic signature of the primary tumour however further work with larger, standardised cohorts is required to validate this.

## 2 Declarations and Copyright

---

### 2.1 Declaration

I, Nicholas J Vennart, declare 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.

### 2.2 Copyright Statement

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 he has given the University of Manchester certain rights to use such Copyright, including for administrative purposes.

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.

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.

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.

## 2.3 **Acknowledgements**

I would like to thank Dr Jill Tipping (Clinical Scientist, The Christie NHS Foundation Trust) and Mr Andrew Knight (Clinical Scientist, South Tyneside and Sunderland NHS Foundation Trust) as the Academic and workplace supervisors for guiding me through the process at every stage and for their useful discussion and comments on my thesis structure throughout. I thank my line manager and mentor, Peter Bartholomew, for his guidance throughout the project and support allowing me time to complete the work. I would like to thank Dr Kathy Clawson (Senior Lecturer, The University of Sunderland) for her input and for pointing me in the right direction with getting started on the practical aspect of writing machine learning code. I would like to thank Dr George Petrides (Consultant Nuclear Medicine Radiologist, The Newcastle Upon Tyne Hospitals NHS Foundation Trust) for his help in contouring the tumours and for many useful discussions during the patient selection process. I would like to thank Mr Alex Phillips and Mr Shajahan Wahed (Consultant Upper GI Surgeons, The Newcastle Upon Tyne Hospitals NHS Foundation Trust) for their initial idea for the project, access to the database and assistance in stratifying the initial patient group. I would like to thank Mr David McCullough (Medical Physicist) and Mr Terry Watson (Clinical Technologist, The Newcastle Upon Tyne Hospitals NHS Foundation Trust) for their help accessing, anonymising the images. Finally, I would like to thank my wife, Lizzie, and our children, Daphne, Monty and Alma, for putting up with and supporting me throughout this entire course and particularly to Lizzie for keeping me going!

## 3 Literature Review

---

This chapter provides the background and explanation to a variety of key concepts for this project. The first section covers the clinical background to oesophageal cancer to illustrate the clinical difficulty with this disease. The second section covers Positron Emission Tomography (PET) imaging and PET technological advances to describe PET imaging in current clinical practice. The third section covers radiomics and illustrates this more generally as an imaging concept and more specifically for its use in PET imaging, concluding with details of how different factors can affect the radiomic signature for the same patient. The fourth section introduces artificial intelligence (AI) with a more detailed introduction to machine learning (ML) in terms of the different algorithms available and commonly used. The fifth and final section is a detailed literature review of all published works in relation to upper gastro-intestinal tract (GI) cancer, PET imaging, radiomics and machine learning to ultimately identify the deficits in the literature and summarise where this project contributes to the scientific community.

### 3.1 Oesophageal cancer

This section will cover the different stages of oesophageal and oesophago-gastric junction (OGJ) cancer, diagnosis and treatment, and discuss the typical patient prognosis following treatment.

#### 3.1.1 Introduction to oesophageal cancer

Oesophageal is one of the top ten most common cancers and is one of the leading causes of cancer death (Fitzmaurice, et al., 2013). “Over the past 25 years”, the natural incidence of oesophageal cancer has been increasing (Ries, et al., 2005; Yang, et al., 2008). Each year in the UK, there are around 9,200 new cases and 7,900 deaths from oesophageal cancer with a 10-year survival rate of just 12% (CRUK, 2017). Oesophageal cancer presents as either an adenocarcinoma or a squamous cell carcinoma (Mayo Clinic Staff, 2019) with several other rarer forms not considered in this study. Oesophageal SCC occurs in the cells lining the surface of the oesophagus, is the most prevalent global sub-type and accounts for “above 90% of oesophageal cancer in China” (Zhu, et al., 2019). Adenocarcinomas are most prevalent “in North America and Europe, especially among white men” (Zhu, et al., 2019) and first presents in the mucus secreting cells of the oesophageal glands (Mayo Clinic Staff, 2019). There have been several advances in both diagnosis and curative treatment; however, the number of deaths from oesophageal cancer remains high with over 50% of patients diagnosed with stage IV (incurable)

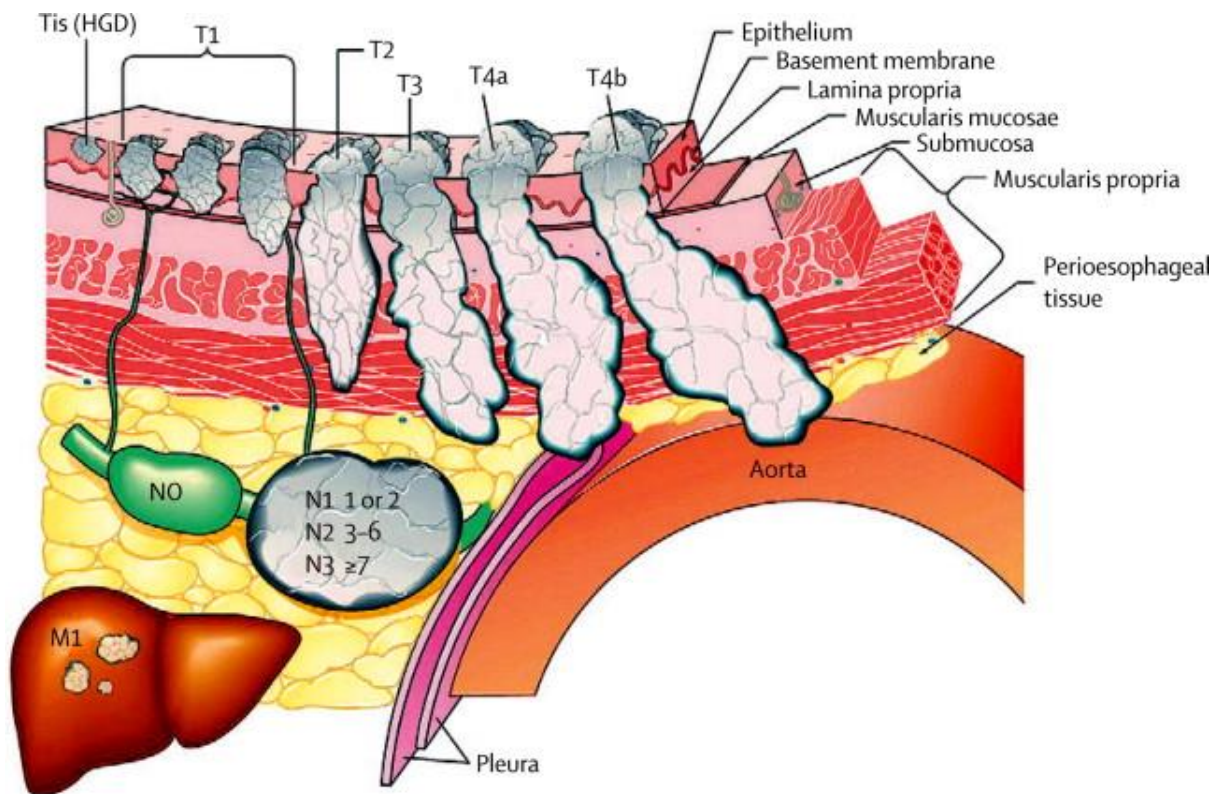
disease (Sah, et al., 2019). For patients diagnosed with disease, even when prescribed with curative intent, prognosis remains relatively poor (Kudou, et al., 2016).

An important scoring tool is the tumour-node-metastases, “TNM Score”. The tumour or T-score or “stage” in oesophageal cancer is classified into stages I – IV (Figure 1) indicating the severity of disease and currently informs the most effective treatment and management of oesophageal cancer (American Cancer Society, 2019). Accurate staging of oesophageal cancer is essential as the efficacy of each treatment option depends on the specific staging of disease (Luu, et al., 2017).

Stage 0 tumours contain abnormal cells in the epithelium but not the basement membrane. Treatment options include endoscopic photodynamic therapy, endoscopic radiofrequency ablation or surgical resection. Stage 1 cancers have affected the deeper lining of the oesophagus but have not spread to lymph nodes or other organs (loco-regional disease). Early stage T1a tumours are treated with endoscopic mucosal resection followed by endoscopic therapy (Sah, et al., 2019). A T2 tumour (which has infiltrated the muscularis propria), is treated with a combination of chemo-radiation and surgery. The combination is chosen depending on the specific location of the tumour; for example a tumour located near the stomach will likely receive chemotherapy and surgery and not radiation therapy whereas a tumour located in the neck will be treated with chemo-radiation and no surgery (Beukinga, et al., 2018; American Cancer Society, 2019). Stage 2 cancers have infiltrated the muscle layer or connective tissues on the outside of the oesophagus and have spread to no more than 2 lymph nodes (Mayo Clinic Staff, 2019). Stage 3 cancers have “grown through the wall of the oesophagus” and are now invading nearby tissues and organs (American Cancer Society, 2019). Stage 2 - 3 cancers are treated with a combination of chemo-radiation therapy and surgery depending on the location of the tumour. Stage 4 cancers are those which have spread to “distant lymph nodes or to other distant organs” (American Cancer Society, 2019); patients with stage 4 disease are usually offered palliative therapy. More detail on this in section 3.1.3.

The nodal or N-score is denoted as a score of 0-3 where: 0 means no lymph nodes contain cancer cells, 1 means there are cancerous cells in 1-2 lymph nodes local to the site of the tumour, 2 means there are 3-6 local lymph nodes with disease and N3 means there are 7 or more nearby cancerous lymph nodes (CRUK, 2019). Finally, the metastases or M score is denoted as 0 or 1 indicating the presence of metastases (1) or not (0) (CRUK, 2019).

**Figure 1 Stages I - IV oesophageal cancer**

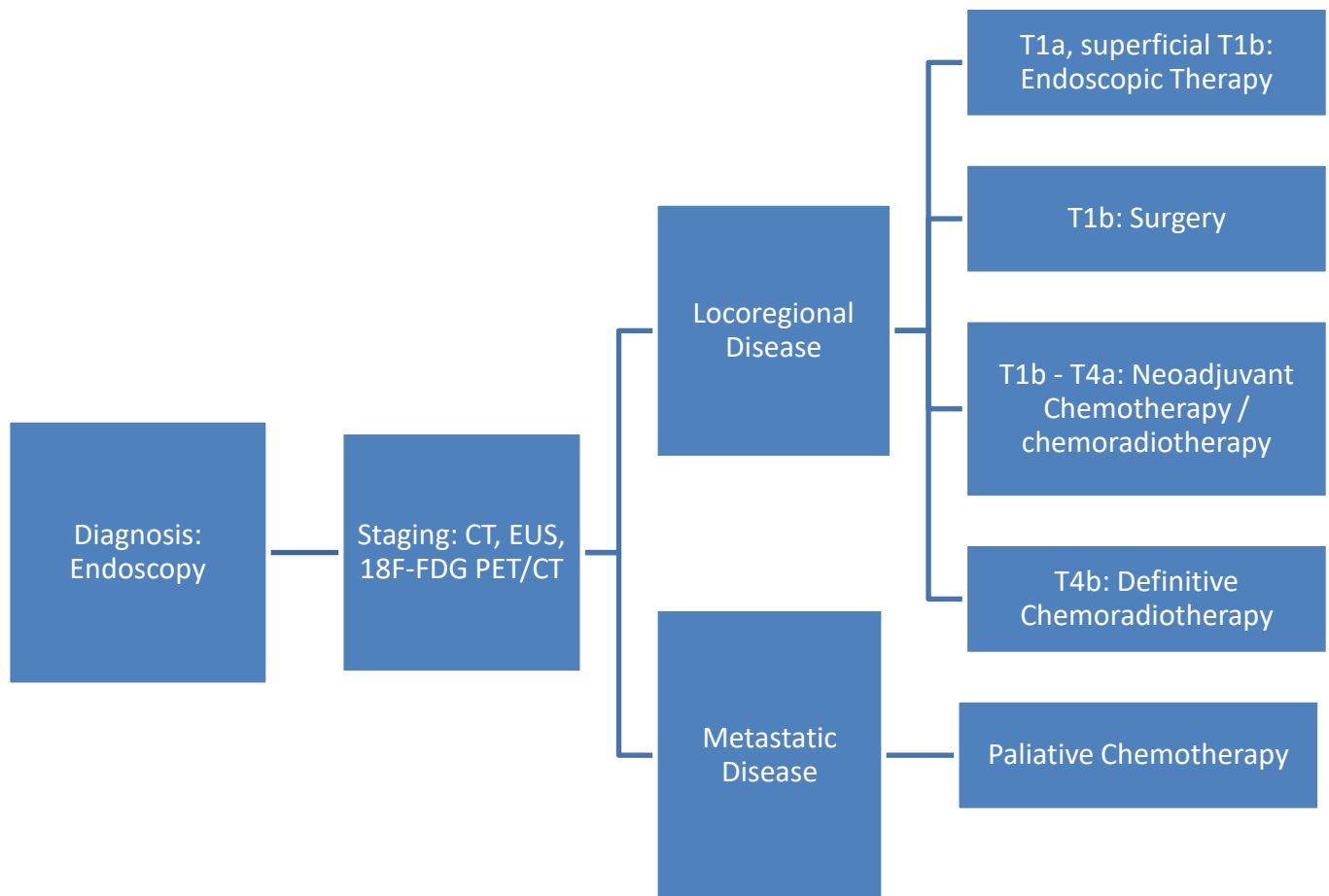


*Diagram showing how oesophageal cancer tumours are scored in relation to the depth of penetration through different soft tissue layers and graded according to the number of involved nodes and presence of metastatic spread. Source: Pennathur et al (2013).*

### **3.1.2 Diagnosis and staging of oesophageal cancer**

Patients are initially diagnosed using endoscopy and an endoscopic biopsy is taken for verification (Bruzzi, et al., 2007) (Figure 2). The biopsy determines the nature of the cells, the degree of disease and the type of cancer. The staging of oesophageal disease is determined using chest and abdomen computed tomography (CT), endoscopic ultrasound (EUS) and 18F-Fluoro-deoxy glucose positron emission tomography (<sup>18</sup>F-FDG PET/CT) (Sah, et al., 2019; Stahl, et al., 2003). In general, following initial diagnosis, patients receive a chest / abdomen CT scan to identify whether there is metastatic disease present in the liver or lungs (Beukinga, et al., 2018). EUS is then performed to determine the extent of the primary oesophageal tumour however such imaging is limited by the extent of the tumour i.e., whether the endoscope is able to circumvent the tumour (Berry, 2014). Where the CT does not show metastatic disease in the liver or lungs, a PET/CT scan is performed (Berry, 2014; Vargese, et al., 2013). Conversely, if metastases are located on CT, further studies are not necessary and patients proceed directly to palliative treatment (Vargese, et al., 2013).

**Figure 2 Treatment pathway for upper GI cancer**

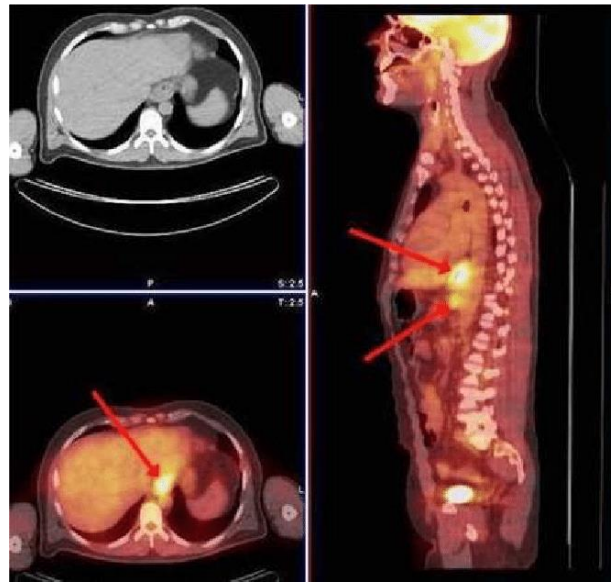


*The UK diagnosis, staging and treatment pathway for oesophago-gastric cancers, adapted from Sah et al (2019).*

One of the key advantages of PET/CT is the ability to identify distant metastases and therefore distinguish between potentially curable disease (Stage 1 – 3) and patients requiring palliation (Stage 4). Luketich et al (1997), and later confirmed by Downey et al (2003), showed that distant metastases can be detected in up to 15-20% more patients than using CT alone. For example, Rashid et al (2015) demonstrate a case with an uncertain node on CT, confirmed as FDG avid on PET imaging (Figure 3).



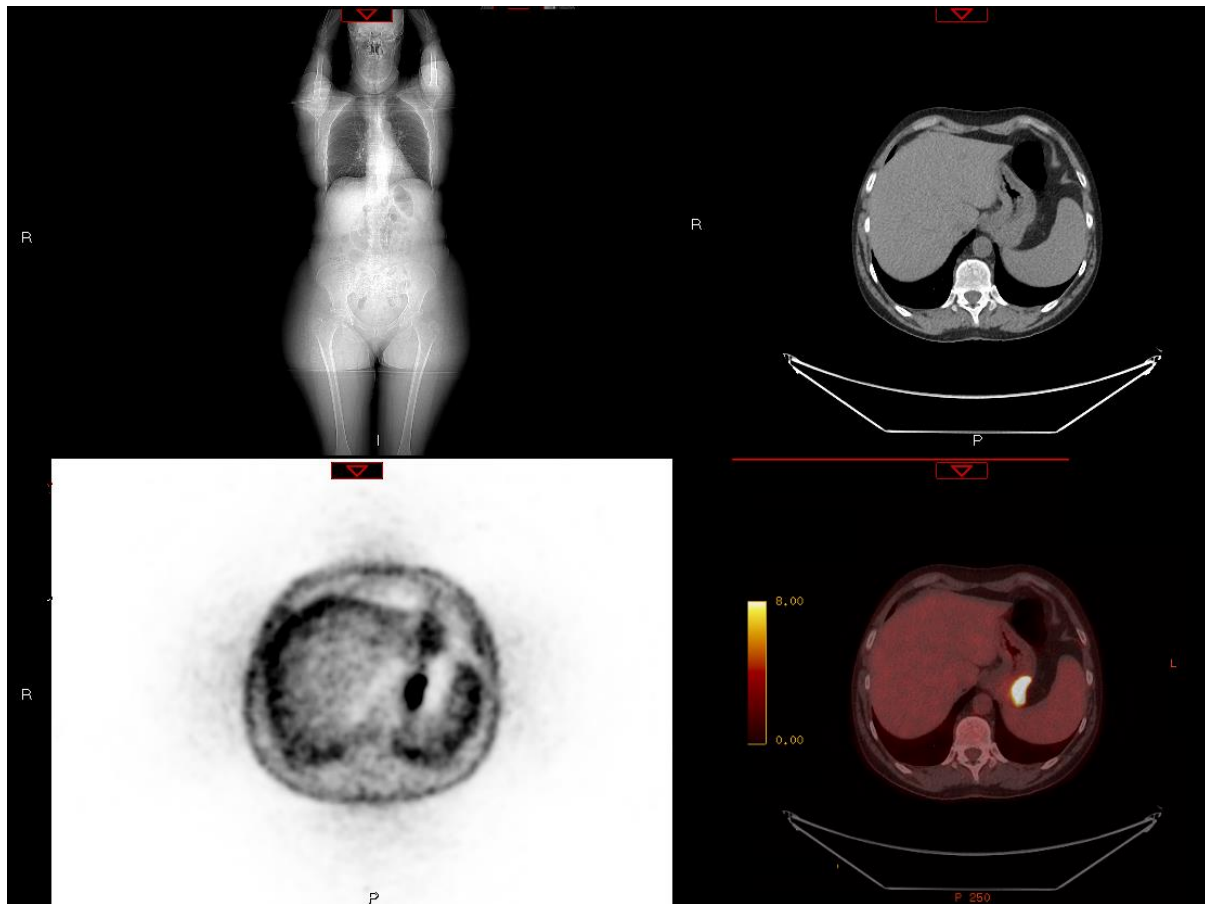
**Figure 3 PET-CT oesophageal cancer image**



*The addition of PET to CT in diagnosis of metastasis. Images show CT (top left), axial PET-CT (bottom left) and sagittal PET-CT (right). Source: Rashid et al (2015).*

Furthermore, PET/CT allows measurement of the biochemical and physiological processes of oesophageal cancer, which helps categorise the primary tumour (Yang, et al., 2008). For example, the information provided by a PET scan can give quantitative information regarding metabolic processes within the tumour such as blood flow and receptor status;  $^{18}\text{F}$ -FDG uptake relates to glycolysis of the tumour i.e. increased glucose uptake by the tumour correlates with increased cellular metabolism. An oesophagectomy is offered for patients where the PET/CT scan does not find distant metastases and patients who subsequently do not develop distant metastases will be offered radiation and neo-adjuvant chemotherapy (Bruzzi, et al., 2007). PET/CT is not without limitation, particularly in squamous cell carcinomas in the identification of false negatives (Berry, 2014). Stahl et al (2003) describe one common issue affecting adenocarcinomas of the oesophagus is low or absent FDG uptake. Yang et al (2008) further describe the link between poor FDG uptake and poor differentiation of tumour cells and believe that reduced / absent FDG uptake is related to “inert mucus which does not accumulate FDG” or “lack of expression of the glucose transporter Glut-1 on the cell membrane” (Atay-Rosenthal, et al., 2012), seen particularly in oesophago-gastric cancers, for example, Atay-Rosenthal presented a case with subtle PET uptake and the biopsy confirmed aggressive disease. In Figure 4, we present a patient from our cohort who demonstrated an area of soft tissue thickening at the greater curve of the stomach, in keeping with a gastric adenocarcinoma.

**Figure 4 Gastric cancer PET-CT images**



*PET/CT images for a gastric adenocarcinoma patient from our cohort to show the Scout image, CT Alone, PET alone and Fused PET/CT images.*

### **3.1.3 Treatment and treatment response of oesophageal cancer**

A variety of treatment options are offered depending on the extent and severity of their disease. For example, most patients opt for “definitive” chemo-radiotherapy (dCRT) as a first line. This treatment option may be chosen either because this will likely cure disease without surgery or, more likely, the disease is in such a place as surgery is either refused or not possible. The tumour may be located adjacent to other structures such as the stomach at the distal end or throat and head / neck structures in the proximal oesophagus and therefore is technically difficult to operate on. For disease suitable for surgery, patient’s may be offered “neoadjuvant” chemo-radiotherapy (nCRT) where CRT is given as a “shrinking” agent to reduce the size of the tumour to aid the success of a surgical resection (American Cancer Society, 2019). The prognosis for oesophageal cancer patients remains poor and the treatment, particularly surgery, is highly invasive. It is therefore useful to accurately identify patients who will show complete pathologic response prior to surgery, which will go on to have successful, organ-preserving surgery and therefore avoid any unnecessary surgical morbidity. In other

words, it would be very useful to identify patients, using their medical images, which will go on to have a successful surgery to avoid patients having to go through invasive CRT and surgery with no survival benefit (Van Rossum, et al., 2016). Furthermore, it is likely that patients who do not respond to CRT will experience “the toxicity of these therapies without prognostic benefit” (Van Rossum, et al., 2016). It is important to note the difference between adenocarcinomas and squamous cell carcinomas (SCC) in terms of treatment response; adenocarcinomas have shown an 8-9% pathologic complete response to chemotherapy and a 23-28% response to CRT whereas SCCs have shown a pathologic complete response rate of 49% (Van Rossum, et al., 2016).

Several trials aimed at improving treatment options for patients have yielded promising, but comparatively poor prognostic results. For example, the OEO2 (Allum, et al., 2009) and MAGIC (Cunningham, et al., 2006) trials have shown 6% and 13% improvement in 5-year overall survival for oesophageal and oesophago-gastric cancers respectively where neo-adjuvant CRT was used in addition to surgery compared to chemotherapy alone (Reynolds, et al., 2017; Sah, et al., 2019). The CROSS trial investigated neo-adjuvant CRT with surgery against surgery alone for oesophageal and oesophago-gastric patients. Van Hagen et al (2012) showed that multimodality treatment increased overall survival from 24 months (surgery alone) to 49 months (surgery plus CRT) with a complete pathological response rate of 29%.

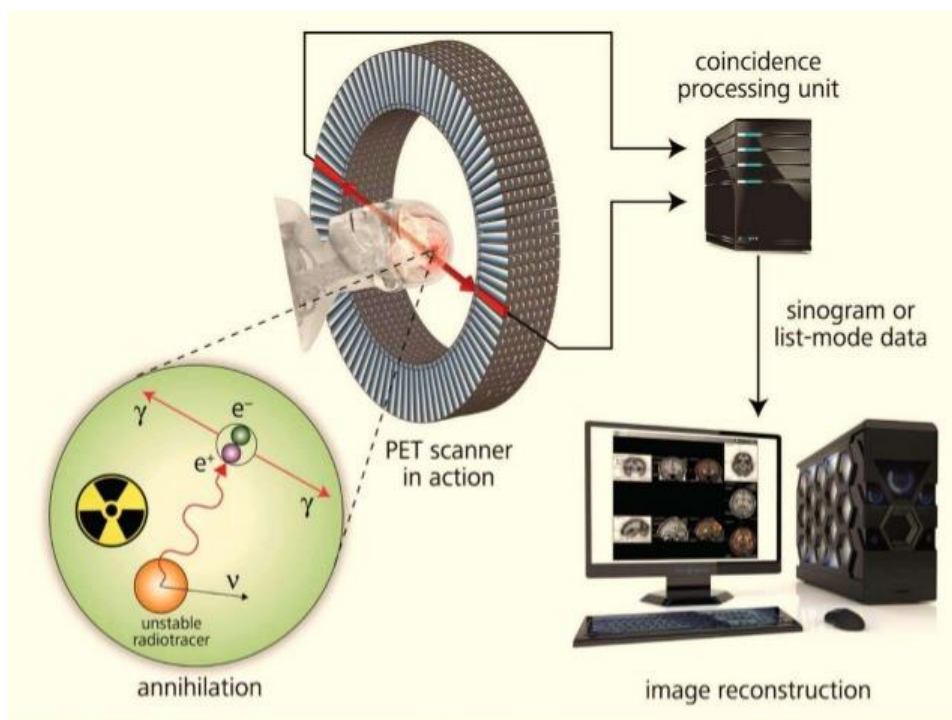
## 3.2 PET/CT Technology

This section will describe how a Positron Emission Tomography – Computed Tomography (PET/CT) scanner works and several recent technological advances to improve image quality including ‘time of flight’ (TOF), ‘point spread function’ (PSF) and Block Sequential Regularized Expectation Maximum (BSREM or “Q.Clear”), which are incorporated into an Ordered Subsets Expectation Maximisation (OSEM) iterative reconstruction loop

### 3.2.1 Introduction to PET/CT

The (PET/CT) scanner consists of a conventional CT scanner (3D x-ray system) and a ring of high-energy photon detectors. In simple terms, a positron-emitting isotope, e.g.  $^{18}\text{F}$  used in the radiopharmaceutical,  $^{18}\text{F}$ -Fluoro-Deoxyglucose ( $^{18}\text{F}$ -FDG) is administered intravenously and positrons annihilate with free electrons in the body and emit two gamma photons in opposite directions (Rich, 1997; Townsend, 2008). Co-incident photons are detected by a ring of detectors and converted, via a sinogram, into an axial image (Ruotsalainen & Viik, 2015)(Figure 5). Typically, a free positron travels approximately 0.5mm before annihilating with a free electron in the body and emitting 2 511keV photons (Schmitdz, et al., 2004).

**Figure 5 Schematic diagram PET scanner**



*An overview of the basic function of a PET scanner showing positron-electron annihilation, coincidence detection and signal processing steps. Source: Ruotsalainen and Viik (2015).*

The concept of a PET scanner was first developed much earlier in the mid-1970s when the Washington University group successfully built an annihilation coincidence detection system which was translated into the first clinically applicable scanner by 1975 (Rich, 1997). Since the widespread adoption of PET/CT in routine oncological practice, there have been several key advances in PET/CT technology (Berger, 2003; Czernin, et al., 2013; Kudou, et al., 2016; Townsend, 2008). The first scanners used the same technology as a conventional gamma camera, utilising thicker sodium-iodide (NaI) crystals for detection however, NaI crystals do not possess the required stopping power for efficient detection of 511keV photons (Melcher & Schweitzer, 1991). The development of higher density Bismuth Germanate Oxide (BGO) crystals allowed greater stopping power in smaller crystals but maintaining a similar light decay constant (Rich, 1997). The main developments in PET until the 1990's were in detector design and computing capability however in 1991, Townsend and Nutt (Townsend DW, 1993; Townsend, 2008) proposed the combination of a PET scanner with a CT scanner to acquire both functional and anatomical information simultaneously.

The next major advance was in the improvement in detector architecture; specifically, the introduction of cerium-doped Lutetium Oxyorthosilicate scintillation crystals allowed sufficient decay time to enable "Time-of-Flight" (TOF) image reconstruction (Melcher & Schweitzer, 1991; Moses & Derenzo, 1999; Mullani, et al., 1980 ). Melcher and Schweitzer (1991) described the physical properties of several possible PET crystals showing that the high density LSO crystals ( $7.4\text{g/cm}^3$ ) exhibited a better detector efficiency than NaI crystals ( $3.67\text{g/cm}^3$ ). LSO crystals also showed a similar radiation length to BGO crystals (1.14cm and 1.12cm respectively for high-energy 511keV annihilation gamma photons) (Schmitdz, et al., 2004). The main advantage of LSO over other crystal types is the combination of similar attenuation length, higher light output (20,000-30,000/MeV) but with a much shorter scintillation time (40ns) compared to BGO and NaI (300ns and 230ns respectively) (Moses & Derenzo, 1999; Melcher & Schweitzer, 1991). The new crystal enabled sufficient stopping power without the loss of detector sensitivity and efficiency and therefore a gain in signal-to-noise ratios, particularly for larger patients (Akamatsu, et al., 2012 ; Ghotbi, et al., 2014; Huang, et al., 2009).

PET images were historically acquired in 2D whereby lead or tungsten septa existed between rings of detectors and all photons arriving at an oblique angle to the detector were discounted; thereby presenting a simpler model computationally but a large loss of counts and therefore sensitivity. Modern PET scanners acquire in 3D whereby photons are detected from a block of detectors, allowing the detection of all photons within a "bed position". This is more computationally demanding but leads to higher sensitivity and ultimately higher quality images (Gundlich, et al., 2006).

The increased sensitivity of 3D acquisition also introduces several disadvantages, such as an increased sensitivity to scattered and random coincidence events. For example, with the septa removed, the detector blocks are able to accept photons from a “greater range of scattering angles” (Cherry, et al., 1991). The larger acceptance angle also increases the susceptibility to random events (Badawi, 1999). In practice, an improvement in reconstruction allows PET systems to take advantage of the improvement in sensitivity to true coincidences (Bulus, et al., 2009). An improvement in data re-binning techniques allows for improvements in correcting for scatter and random events.

### 3.2.2 Standard Uptake Value (SUV)

The key advantage of PET imaging is the ability to perform quantitative analysis where by the value of any particular pixel is related to the underlying biological properties of the tissue it represents. With this in mind, the scanner is calibrated such that the value of a pixel can be directly related to the activity concentration (Kinahan & Fletcher, 2010). This is useful for  $^{18}\text{F}$ -FDG imaging in oncology because increased accumulation of tracer is a useful marker for both identifying and grading of cancer. Furthermore, the SUV can be used to assess changes in the shape of the tumour in response to disease progression / treatment (Akamatsu, et al., 2012 ). The standardised uptake value (SUV) is the decay corrected radioactivity concentration measured by the PET scanner ( $r$ , kBq / ml) divided by the amount of activity injected ( $A$ , kBq) and then normalised to the patient’s weight ( $W$ , g). Patient weight is used as a surrogate for the volume of tracer distribution and is widely accepted by the nuclear medicine community (Kinahan & Fletcher, 2010). Using weight (rather than volume) is not without pitfalls though, for example, typically heavier patients will have more body fat which takes up less  $^{18}\text{F}$ -FDG

$$\text{SUV} = r / (A/W) \quad \text{Equation 1}$$

Use of the SUV is pertinent to further discussions regarding quantification and effects on image improvement. The ultimate goal of a ‘perfect’ SUV would be that the value as measured by an individual pixel matches the exact concentration of activity in the tissue it represents. It also assumes that the concentration of activity in the tissue is directly and robustly related to the grade of the tumour and gives an exact indication of the treatment response or progression of the tumour. However, the SUV measurement itself is not a perfect measure and is subject to several fundamental flaws, inherent to the PET imaging system. PET spatial resolution is inherently restricted by the distance a positron travels before a detectable annihilation event and the size of the detectors. The combined effect limits spatial resolution to 5-8mm, which reduces its efficacy for detecting smaller tumours and affects the ability to detect the underlying concentration in the tissue. The spatial

resolution limit gives rise to the partial volume effect whereby the measured activity concentration, particularly for objects less than a few cm<sup>3</sup> in volume, is less than the true tracer concentration (Kinahan & Fletcher, 2010). Sections 3.2.3-3.2.5 describe several technological advances in both detector hardware and image reconstruction software to address these issues and improve the accuracy of SUV quantification. In addition to the various technical and physical factors affecting the SUV measurement, FDG accumulation in tumours is not directly related to the “proliferative activity of malignant tissue and to the number of viable tumour cells” (Kinahan & Fletcher, 2010). The most useful measure is the relative uptake of FDG in tissue which, as aforementioned, is varied by the amount of injected FDG and the patient weight. Furthermore, FDG uptake is not specific enough to tumour activity; processes such as inflammation and infection can show high avidity for FDG and equally, a slowly growing malignancy can show low FDG uptake (Kinahan & Fletcher, 2010).

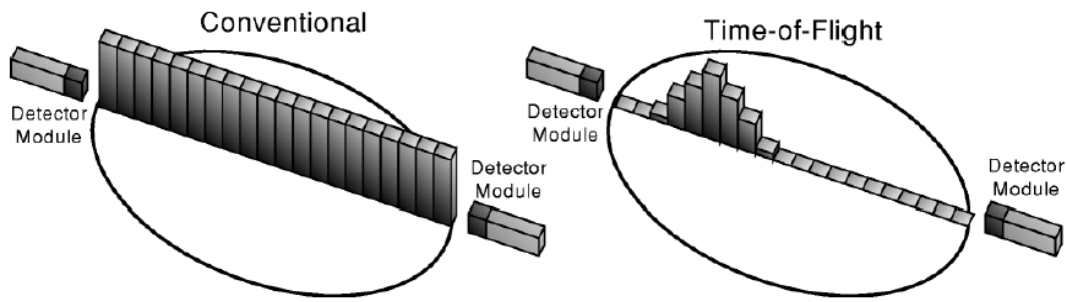
### **3.2.3 Technological Advances in PET: Time of Flight**

The much faster scintillation time of LSO crystals allows the “time-of-arrival” of each annihilation photon to be recorded, i.e., the shorter scintillation time improves the timing resolution of the scanner to the point where the scanner can be used to determine the approximate position along the line of response (LOR) at which the photon annihilation event originated. In conventional PET, for each recorded coincidence event, the counts along any particular LOR are spread evenly along the whole LOR (Moses, 2003). In TOF PET, the timing resolution (typically 300-500ps) allows the scanner to narrow down the likely position that the event took place (Moses, 2003). Budinger (1983) showed that with a timing resolution of ~500ps, it would be possible to confine a positron event to a positional line element of ~7.5cm on a chord:

$$\Delta x = \frac{c}{2} \Delta t \quad \text{Equation 2}$$

Where  $\Delta x$  = distance from the annihilation in the centre of the scanner to the detector ring,  $\Delta t$  = time difference between gamma photon arrivals and  $c$  = speed of light.

**Figure 6 Conventional vs TOF PET**



*Schematic to show the difference between count recording for conventional and time-of-flight PET. Source: Moses (2003).*

By confining the positron event to a line segment (rather than the whole line element) we can improve signal-to-noise ratios (SNR), particularly for physically larger patients with  $^{18}\text{F}$ -FDG distributed across a larger body diameter. Budinger (1983) further showed that the improvement in Signal (S) and Noise (N) was related to the diameter (D) of the imaged object and the line element error ( $\Delta x$ ):

$$\frac{S/N_{TOF}}{S/N_{NON-TOF}} = \left( \frac{\Delta x^2}{D^2} \right)^{-1/4} = \sqrt{\frac{D}{\Delta x}} \quad \text{Equation 3}$$

TOF PET has since been shown to improve SNRs and image quality, most pronounced in larger patients (Karp, et al., 2005 ; Surti, et al., 2007; Karp, et al., 2008; Kadmas, et al., 2009; Lois, et al., 2008 ).

### **3.2.4 Technological Advances in PET: Point Spread Function**

As aforementioned, the spatial resolution in PET is limited to 5-8mm. However, several advances in detector technology and image reconstruction algorithms are starting to ameliorate for the limiting spatial resolution (Reader, et al., 2003; Akamatsu, et al., 2012 ). With the advent of more sophisticated and powerful computing, further advances were made in image reconstruction algorithms such as “Point-Spread-Function” (PSF) which aims to reduce partial volume effects and correct for “spatial distortion away from the centre of the detector” (Vennart, et al., 2017) and ultimately improve image quality (Akamatsu, et al., 2013; Murray, et al., 2010; Ross & Stearns, 2010; Reader, et al., 2003).

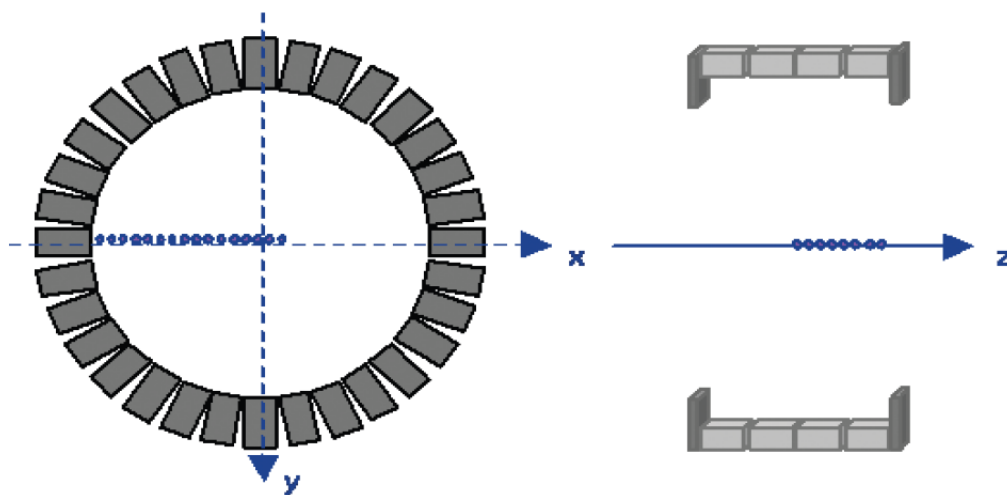
The crystal element size limits the spatial resolution in PET (Alessio & Kinahan, 2006; Cherry, 2006). One of the challenges of the circular geometry of a PET detector ring is that the ring introduces a spatial distortion away from the centre of the scanner (i.e. the depth of the interaction), which limits spatial resolution. Photons produced in the centre of the scanner are detected and localised correctly



however, as we move towards the edge of the field of view, photons arising from annihilation events are more likely to be localised incorrectly. Towards the edge of the scanner, the photon strikes the crystal at an angle and is more likely to travel to (and be detected by) the neighbouring crystal in the detector block (Akamatsu, et al., 2012 ). The General Electric (GE) scanners, from 600 series onwards, include a point-spread function correction algorithm called “Sharp IR” (Ross & Stearns, 2010). The correction was developed by measuring the response of a point source at several million points across the FOV (Alessio, et al., 2010) and then incorporating this measured response into the sinogram space and ultimately the image reconstruction algorithm (Figure 7) (Ross & Stearns, 2010; Alessio, et al., 2006).

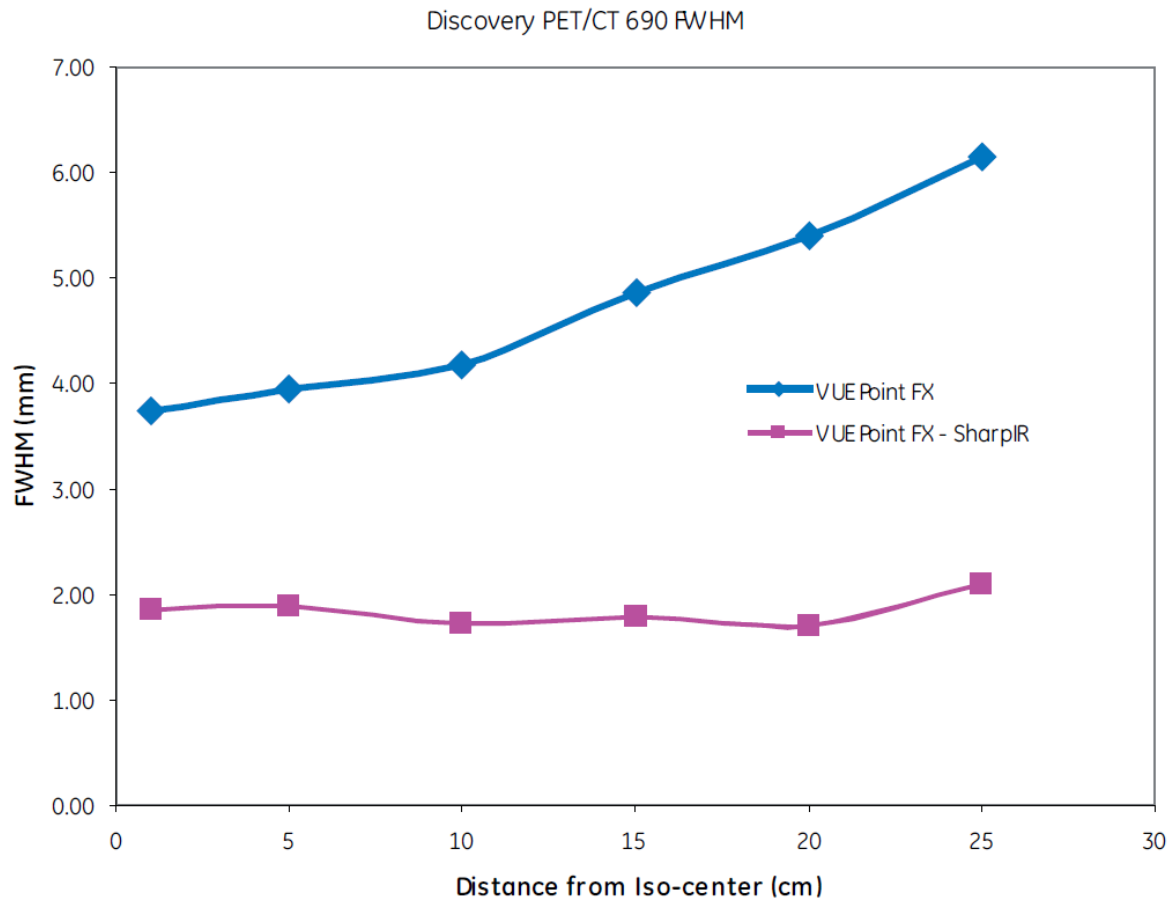
Ross et al (2010) plotted the full width half maximum (FWHM) of the point spread function at different distances across the field of view (FOV) for OSEM+TOF (“Vue Point FX”) reconstructions both with and without the PSF (‘SharpIR’) correction (Figure 8). Note, GE’s implementation of PSF correction is spatially variant however, other manufacturers, such as Siemens, have implemented a correction by incorporating the probability “that an event in an image voxel contributes to a sinogram bin” into the image estimation step of the reconstruction algorithm (Casey, 2007).

**Figure 7 Measurement of Point Spread Function**



*Schematic to show the effect and measurement of PSF on count detection. Source: Ross and Stearns (2010).*

**Figure 8 FWHM of Point Spread Function**



*Plot to show the enhancing effects of PSF on the detection FWHM away from the centre of the scanner. Source: Ross and Stearns (2010).*

Several groups have since proved the efficacy of PSF correction algorithms. Panin et al. (2006) performed measurements on a Siemens “HiRez” Scanner using a point source to develop a fully 3D PET reconstruction. Panin et al. (2006) corrected the measured data for crystal and geometrical efficiency then derived a whole system matrix from the responses in projection space.

Manjeshwar et al. (2006) performed line source measurements using the “GE Discovery STE” PET/CT scanner. Manjeshwar et al. (2006) implemented a geometry correction using a distance driven projector to include the system geometry into the reconstruction model. The distance driven projection allowed for faster reconstruction times. This work was extended by Alessio et al. (2010) and Rapisarda et al. (2010). Alessio et al. (2006) initially discussed the implementation of 3D spatially variant system response function derived from Monte Carlo simulations. Alessio et al. (2006) found that by incorporating a 3D spatially variant system response function improved image quality with more significant improvements towards the edge of the FOV. Alessio et al. (2010) continued this work

and performed measurements on the GE Discovery STE scanner using a  $^{22}\text{Na}$  source and evaluated results using the NEMA Image quality phantom, a whole-body  $^{18}\text{F}$ FDG and a brain  $^{18}\text{F}$ FDG patient scan. Alessio et al (2010) found that incorporation of PSF into the reconstruction algorithm required more iterations to converge to a final solution than algorithms without PSF. Rapisarda et al. (2010) performed a similar analysis and evaluation. Similarly, Rapisarda et al (2010) found that incorporation of PSF into the reconstruction improved image quality.

Tong et al. (2010) performed a full evaluation of noise and signal properties with PSF incorporated into the reconstruction algorithm. Tong et al. (2010) filled the NEMA image quality phantom with  $^{68}\text{Ga}$  spheres filling all six spheres in a concentration ratio of 4:1 (SBR). Tong et al. (2010) performed 50 identical scans and results compared using a variety of signal and noise metrics. Images were reconstructed using 1-10 iterations, 0-10mm (0, 4, 7 and 10mm) post-filters and results were compared for OSEM+LOR (line of response) and OSEM+LOR+PSF. Tong et al. (2010) found that at matched iterations, the incorporation of PSF reduced noise levels and improved SNRs and contrast recovery of lesions in a warm background.

### 3.2.5 Technological Advances in PET: Block Sequential Regularized Expectation Maximum iterative reconstruction

In PET imaging, the primary method of reconstruction is the Ordered Subsets Expectation Maximization (OSEM) algorithm (referred to as “VuePoint HD” on GE Scanners). OSEM provides gains in Signal-to-Noise Ratios (SNRs) over the traditional filtered back projection algorithm. The main issue with OSEM is that the algorithm cannot be run to full convergence because with each iteration, whilst the signal may increase, so too does the noise. In clinical practice, 2-4 iterations are used which often results in an under-converged (under-optimised) image (Ross, 2014). Note, ‘Ordered Subsets’ relates to the number of divisions made of the sinogram data, typically 24 - 32, and the number of iterations is then the number of complete cycles through the data that the reconstruction algorithm makes. E.g. with 24 subsets and 2 iterations, the reconstruction algorithm will make 24 attempts at the correct image, each time using a different  $1/24^{\text{th}}$  of the data and this process will happen twice (2 iterations) GE (Ross, 2014) have introduced a “Regularized Reconstruction iterative algorithm (Q.Clear)” which is a Bayesian penalized likelihood reconstruction algorithm, which uses the Block Sequential Regularized Expectation Maximum (BSREM) algorithm to solve equation 3 . This equation includes an image noise function term in the reconstruction equation where equation 4 shows an OSEM reconstruction and equation 5 shows the addition of the regularisation function. Term “ $R(x)$  is a penalty to control noise” and the  $\beta$  term “controls the relative strength of the regularizing term relative to the data statistics” (Ross, 2014). The relative strength of the noise penalty term is chosen using prior knowledge about the image quality e.g. a different  $\beta$  value depending on the part of the body being imaged. The iterative algorithm is then run to convergence.

In equations 4 and 5, term “ $y_i$  represents the measured PET coincidence data,  $x$  is the image estimate, and  $P$  is the system geometry matrix” (Ross, 2014).

$$\tilde{x} = \underset{x \geq 0}{\operatorname{argmax}} \sum_{i=1}^{nd} y_i \log[Px]_i - [Px]_i \quad \text{Equation 4}$$

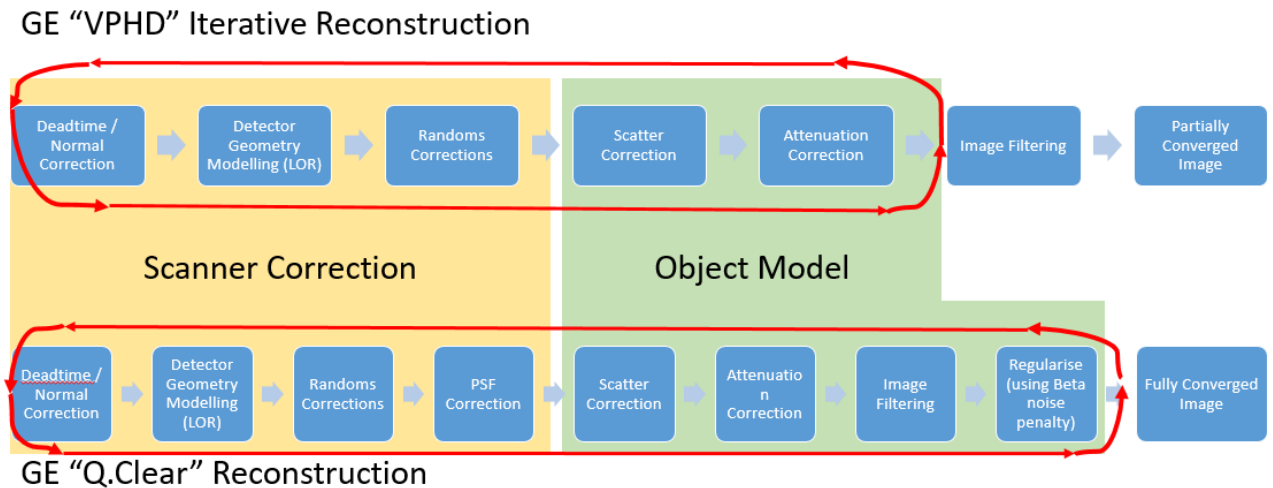
$$\tilde{x} = \underset{x \geq 0}{\operatorname{argmax}} \sum_{i=1}^{nd} y_i \log[Px]_i - [Px]_i - \beta R(x) \quad \text{Equation 5}$$

In other words, this allows the OSEM algorithm to continue performing iterations to enhance the SNR of e.g., lesions whilst maintaining acceptable image noise levels and therefore producing superior quality images compared to OSEM alone. The  $\beta R(x)$  term acts to reduce the object function, which effectively steers the algorithm away from creating noisier images.

Several authors have explored the effect of the noise penalty term on the NEMA phantom, for example, Teoh et al (2015) and Spasic (2018) who investigated different Beta terms using the NEMA

phantom. Reyes-Llompарт et al (2019) expanded on this work and investigated different values of  $\beta$  (50 – 500), concluding that a value of 350 was optimum for oncological studies however; optimisation should be performed locally depending on the scanner and type of studies being viewed.

**Figure 9 Process flow map for OSEM vs Q.Clear Reconstruction**



*Process map showing the incorporation of PSF and Regularisation to achieve a fully converged image, adapted from Ross (2014).*

Clinically, the BSREM algorithm increases the SUVmax values observed in PET/CT imaging (Wyrzykowski, et al., 2020; Reyes-Llompарт, et al., 2019). Reyes-Llompарт et al (2019) compared image quality parameters in a group of 112 PET/CT patients with both OSEM+PSF and BSREM reconstructed scans. The group concluded that BSREM provided only marginal improvements in overall image quality and interpretation but that a simple radiomics model outperformed any single image quality (IQ) parameter. Wyrzykowski et al (2020) investigated the use of "Q.Clear" in a group of 70 lymphoma patients and found that only three scans upgraded from a negative to a positive scan. Wyrzykowski et al (2020) recommended OSEM reconstruction for treatment response, but that Q.Clear may aid interpretation and lesion detection in a minority of cases. Further work has shown that Q.Clear "increases contrast recovery and decreases background variability, producing an overall increase of contrast-to-noise ratio in phantom studies" (Reyes-Llompарт, 2019; Teoh, et al., 2015).

### 3.2.6 Motion Correction

In addition to various image reconstruction techniques, there are several motion correction techniques which can be used. There are several types of motion which can impede PET imaging: patient movement, cardiac cycle, respiratory motion and organ motion (Nayyeri, 2015). Some patient movements can be limited using patient restraints and corrected using post acquisition software, depending on the type and severity of the movement. For cardiac motion, this is typically acquired

and reconstructed separately to the whole body images, with the data split into time frames and independently reconstructed (Nayyeri, 2015). Most pertinent to this work is respiratory motion; for CT the “breath-hold” technique is commonly used however this is inadequate for PET imaging. One solution to respiratory motion is to use respiratory gating to track the motion and place counts into bins depending on the respiratory cycle (Nayyeri, 2015). Motion correction was not used in this study as effective motion correction required manipulation of the raw data, which was not available for this study.

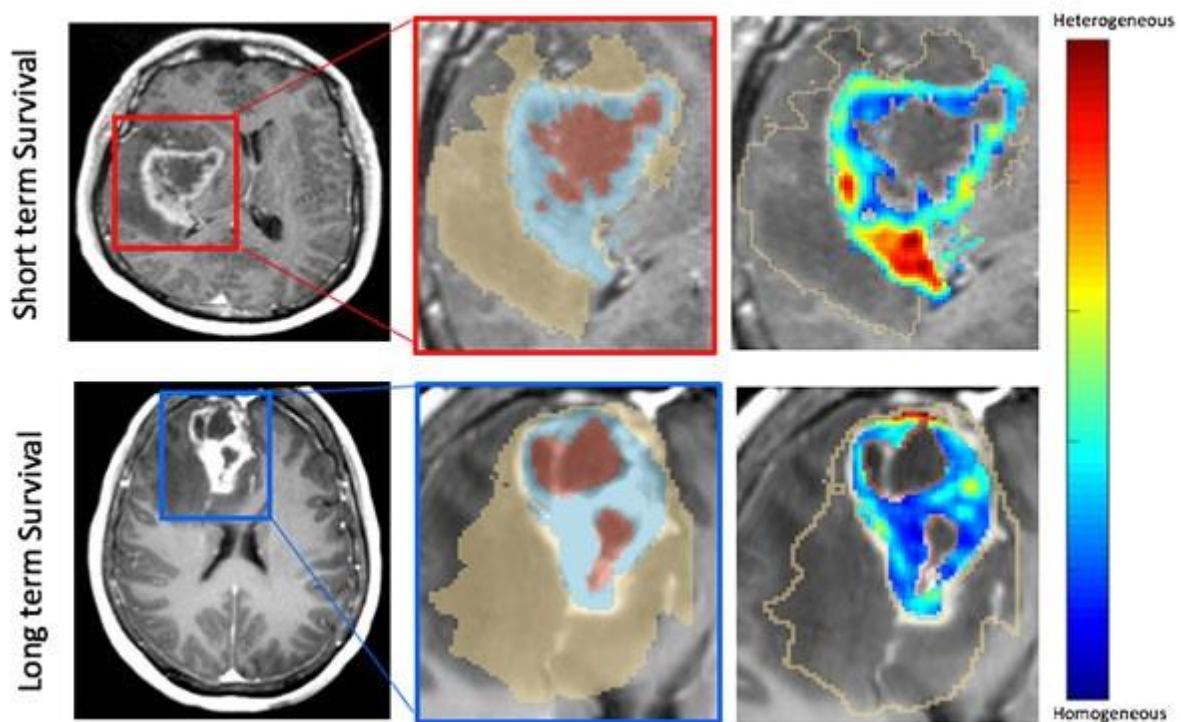
### **3.3 Radiomics**

More recent advances in PET/CT imaging have moved towards using concepts such as ‘radiomics’ to ultimately fuel and develop ‘artificial intelligence’ techniques such as ‘machine learning’ (Sah, et al., 2019). This section will introduce radiomics, describe how it may relate to underlying tumour histological features, discuss the nuances of radiomics and describe how imaging parameters can affect the absolute values.

#### **3.3.1 Introduction to Radiomics**

The concept of ‘big data’ is increasingly applicable to medical imaging, in particular since the advent of ‘radiomics’ (Gillies, et al., 2016; Aktolun, 2019). Gillies et al (2016) describe radiomics as the extraction of “innumerable quantitative features from tomographic images” by the “conversion of digital medical images into minable high-dimensional data” (Hatt, et al., 2016). The basic hypothesis behind radiomics is that medical images contain more information than can be resolved with the naked eye and that there is a relationship between these ‘hidden’ features of the image and the tumour pheno/genotype (Cook, et al., 2018). Furthermore, “it has been recognised that genetic heterogeneity exists within tumours and between metastatic tumours in the same patient” (Cook, et al., 2018). The concept of tumour “heterogeneity” and “homogeneity” is demonstrated well using magnetic resonance (MR) images; Beig et al (2020) demonstrated a group of 203 patients with MR images for Glioblastoma and showed that patients with more heterogeneously distributed tumour genetic profiles were more likely to show poorer outcomes. In other words, tumours with a more uniform distribution of tumour cells (a more homogeneous genetic profile) generally responded better to treatment and demonstrated more long-term survival (Beig, et al., 2020).

**Figure 10 Heterogeneous vs Homogeneous Tumour cell distribution**



*MR images to show heterogeneous vs homogeneous cell distribution in patients who demonstrated short and long term survival respectively. Source: Beig et al (2020).*

One of the attractions of radiomics as a tool is the ability to sample the whole tumour, using images acquired in the line of normal clinical practice. Radiomics is also a non-invasive technique, which avoids sampling errors associated with techniques such as tissue biopsies.

There are, as with any technique, disadvantages associated with radiomics, particularly in PET. The data is macroscopic in nature as opposed to microscopic and so are unlikely to closely correspond to the “underlying cellular biology on a microscopic scale” (Cook, et al., 2018). A further, current disadvantage of radiomics is the general lack of standardisation between institutions on factors such as scanner hardware, activity injected, uptake time, bed position time, CT attenuation correction algorithm used and PET image reconstruction parameters (Cook, et al., 2018). The lack of standardisation may ultimately limit the reliability in comparing with similar published works and comparisons between trials. Lovat et al (2017) present an example of such disadvantages and have shown that texture features can vary depending on the time post injection that an  $^{18}\text{F}$ -FDG scan is acquired (Cook, et al., 2018). A factor, which may be an issue in examining oesophageal cancer images, is respiratory artefacts, which may affect the initial PET tumour image and subsequently affect textural analysis (Yip, et al., 2014). Furthermore, an attempt to use motion correction techniques may

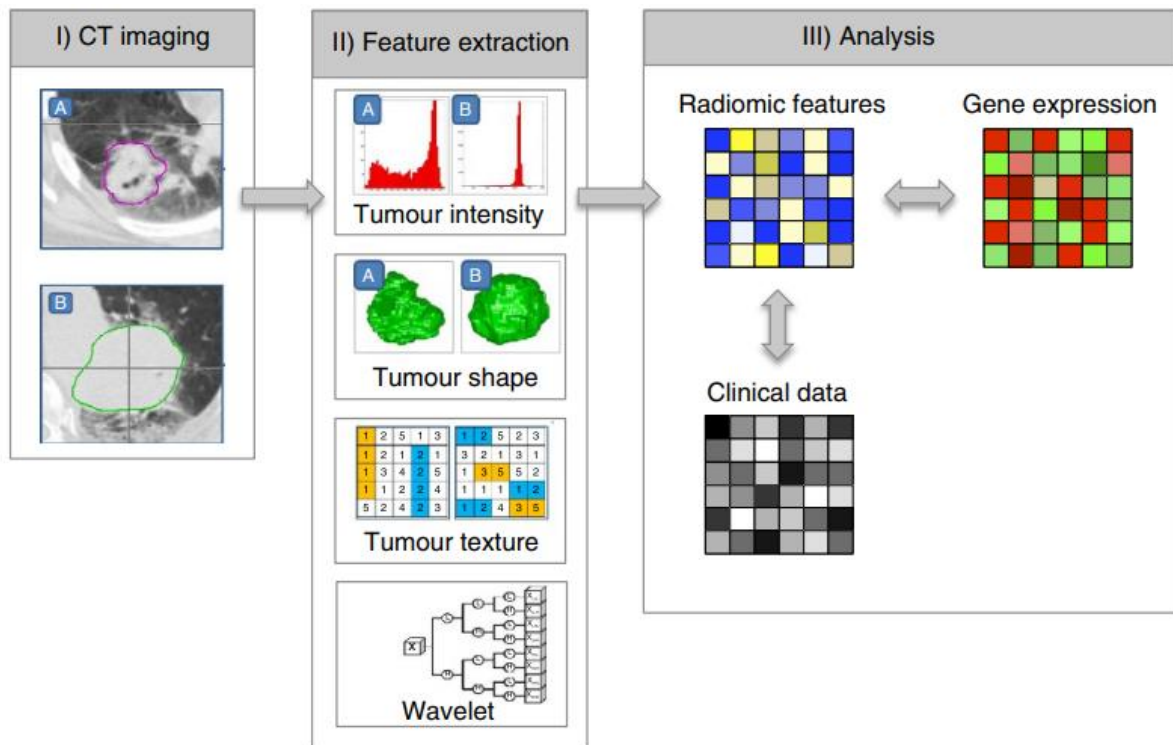
inadvertently add a uniformity to the images and therefore invalidate any radiomic signatures acquired.

### **3.3.2 Radiomics and Tumour Phenotype**

In recent years, there has been increasing support for the link between medical imaging and biological processes such as cellular density, angiogenesis and cell proliferation (Cook, et al., 2018). There is further support for the link between poor treatment outcomes and more aggressive tumour phenotypes as described by these biological processes (Junttila & Sauvage, 2013; Cook, et al., 2018). A tumours “phenotype” in simple terms, describes an observable physical characteristic of an organism in a tumour, which distinguishes it from other organisms (Honey & Shows, 1983). On a genetic level, the tumour phenotype relates to the underlying “genotype” which describes a particular genetic signature, which makes that particular physical characteristic more likely (Honey & Shows, 1983). For example, having blonde hair is a phenotype whereas ‘not having’ the gene for black hair is a genotype. Aerts et al (2014) describe successes in analysis of CT images for lung and head and neck cancers whereby radiomics signatures had great prognostic power through comparison to the genetic signature of a particular tumour phenotype. Aerts et al (2014) segmented tumour volumes from CT images, extracted 440 radiomic features and compared these to both the clinical data and the gene expression showing that heterogenic radiomic signatures link with the underlying gene expression patterns (Figure 11).



**Figure 11 Radiomics analysis process map**



*Schematic process map to show the basic process of imaging, feature extraction and analysis. Source: Aerts et al (Aerts, et al., 2014).*

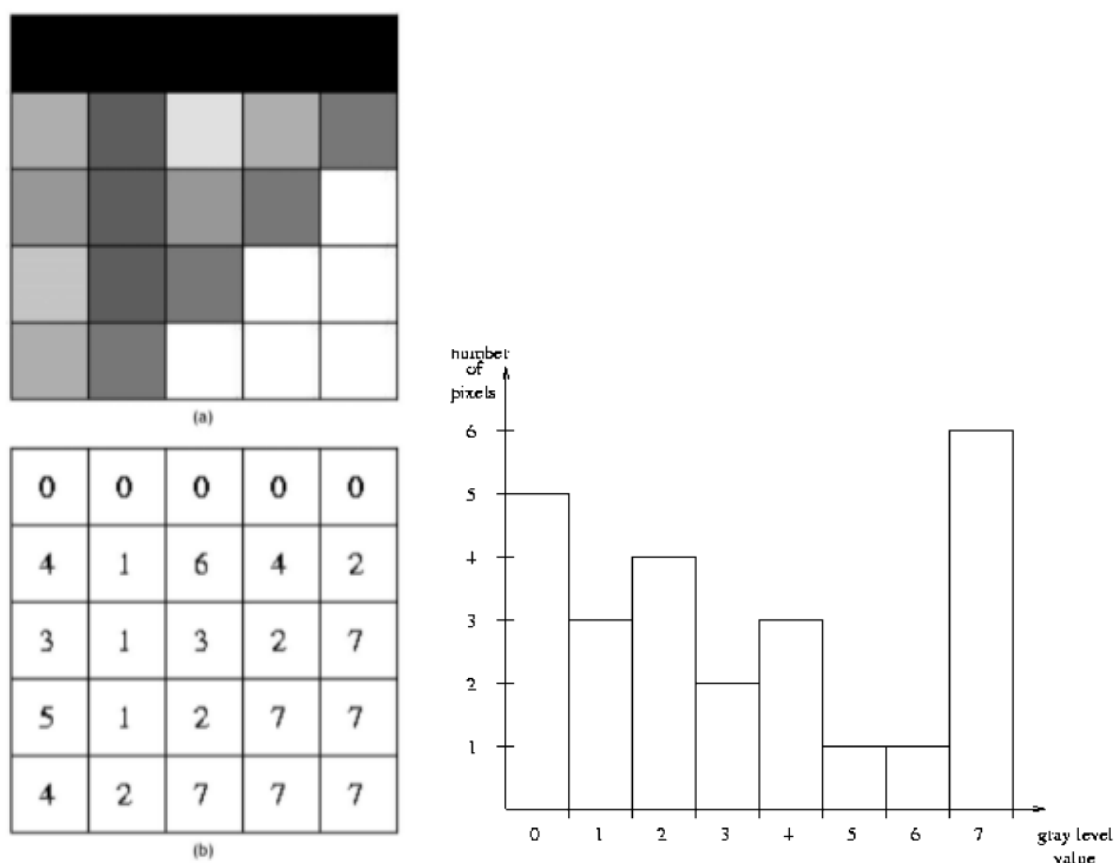
The term “radiomic phenotyping” has been used by (Velazquez, et al., 2017) who describe a study of 763 lung adenocarcinoma patients whereby they have used artificial intelligence and automatic deep-learning methods to link specific radiomic characteristics to a particular somatic mutation. In other words, Velazquez et al (2017) have shown that certain tumour phenotypes have a distinct, corresponding radiomic phenotype and hence, by re-examining the data already available via medical imaging, we are able to better classify tumours to inform management.

Recently, there has been increasing interest in the application to PET images by identifying parameters, which describe the spatial distribution of <sup>18</sup>F-FDG uptake within a tumour, and it is hypothesized that greater heterogeneity of this uptake within the tumour is linked to physiological factors such as tumour necrosis and metabolism (Ypsilantis, et al., 2015). It is further hypothesized that greater variability in heterogeneity is linked to more aggressive cancers and therefore raising the chance of futile treatment (Ypsilantis, et al., 2015).

### 3.3.3 Radiomics and “Texture” Analysis

Radiomics acquires traditional features such as tumour volume, signal density (on CT or MRI), SUV (in PET imaging) and more detailed, ‘hidden’ features related to the “voxel-intensity volume histogram” (Cook, et al., 2018); this gives information on the 3-dimensional spatial heterogeneity of voxels with differing signal intensities and gives parameters used to describe image ‘texture’. Castellano et al (2004) describe and illustrate, for example, in digital images, the allowed grey-levels for a pixel are limited by the number of bits for the image. Standard digital images (e.g. PET images) can be stored with 8 bits whereas MR images typically use 12 bits; this allows pixel values from 0 to 255 or 0 to 4095 respectively (Castellano, et al., 2004). For an image of 0 to 255, 0 equals a black pixel, 255 equals a completely white pixel and 1 to 254 are varying levels of ‘grey’ in between (Figure 12) (Castellano, et al., 2004). The numbered pixels can then be recorded in a histogram where further parameters such as mean, variance and percentiles can be derived (Castellano, et al., 2004).

**Figure 12 Formulation of grey levels**



*Schematic to show the formulation of grey levels within an image: (a) Image with grey level values (b) numbered pixels and the associated grey level histogram. Source: Castellano et al (2004).*

One of the main challenges, particularly when applying radiomics techniques to PET imaging, is spatial (voxel) resolution; in comparison to CT and MRI, PET offers poorer spatial resolution of 5-8mm and so individual voxel information does not correspond to a process occurring on a cellular level (Cook, et al., 2018; Aerts, et al., 2014). Rather, PET data corresponds to the aggregate signal from the cells within the volume captured by a single voxel; so when assessing tumour heterogeneity, the scale at which such heterogeneity may be observed is coarser than with sub-millimetre imaging with CT or MRI. However, data acquired on PET alone may well be complementary to CT or MRI data and give a more complete picture of tumour phenotype, i.e. PET data gives an image of the relative heterogeneity of function rather than of tissue density (CT) or proton density (MRI). Indeed, more recent advances in radiomics have combined the radiomic features in PET, MR and CT with strong correlation between these features (Esfahani, et al., 2022).

Statistical textural analysis splits into 3 main sub-groups: first, second and higher order statistics. First order statistics relates to measures of “central tendency” (Ypsilantis, et al., 2015) or “global measurements” (Cook, et al., 2018) i.e. measures relate to voxels in the whole tumour volume such as the mean, median, mode, percentiles, quartiles, variance, interquartile range, standard deviation and coefficient of variance (Ypsilantis, et al., 2015; Leijenaar, et al., 2015; Deantonio, et al., 2022). Second order statistics are generally described as the “texture” features (Deantonio, et al., 2022) and relate to “co-occurrence measurements” specifically between 2 adjacent pixels on any 2D slice (axial, sagittal and coronal) which are calculated using “Grey-Level Co-occurrence Matrices” (GLCM) and “Grey-Level Difference Matrices (GLDM) (Ypsilantis, et al., 2015). A GLCM is a matrix containing information on the number of times that a voxel with a given intensity co-occurs with a second adjacent voxel with a different intensity. A GLDM contains the absolute difference in intensity between a pair of voxels. Continuing the example shown in Figure 12, Castellano et al (2004) show the same image as a matrix representing the “distribution of pairs of voxels” which in turn gives rise to parameters such as the “entropy which measures the randomness or homogeneity of the pixel distribution”; the higher the entropy, the more random the distribution of pixels.

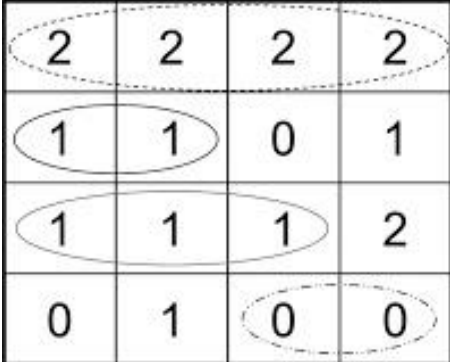
**Figure 13 A grey-level co-occurrence matrix**

	0	1	2	3	4	5	6	7
0	6	0	0	0	0	0	0	0
1	0	0	1	0	1	0	0	1
2	0	1	0	0	0	1	1	2
3	0	0	0	2	0	0	0	1
4	0	1	0	0	0	0	1	1
5	0	0	1	0	0	0	0	0
6	0	0	1	0	1	0	0	0
7	0	1	2	1	1	0	0	2

*The co-occurrence matrix created using pixels within a 2-pixel distance in the horizontal direction, based on the image shown in Figure 12. Source: Castellano et al (2004).*

Higher order statistics impose filter grids on the image (Deantonio, et al., 2022; Gillies, et al., 2016) and describe relationships between three or more voxels, which occur at “specific locations relative to each other”. Higher order metrics are extracted from “Grey-Level Run Length Matrices” (GLRLM), “Grey-Level Size Zone matrices” (GLSZM) and “Neighbourhood Grey-Tone Difference Matrices” (NGTDM) (Ypsilantis, et al., 2015; Sah, et al., 2019). The GL Run Length is the number of consecutive voxels with the same value in a given direction whereas the GL Run Zone is the number of clusters of a particular size with the same GL intensity (Figure 14). When the GLRL is converted into a matrix, each element in the matrix denotes “the total number of occurrences of runs of length  $j$  at grey level  $i$  in a specific direction  $a$ ” (Ypsilantis, et al., 2015; Buch, et al., 2015).

**Figure 14 Grey-level run length matrix**

				Horizontal Runs		Run Length				
Gray Level		0	1			2	3	4		
						1	0	0		
						1	1	0		
		2				0	0	1		

*Example illustration of how a grey-level run length matrix is formed. Source: Buch et al (2015).*

Similarly, the GLSZ matrix denotes the number of clusters of a specific size with grey-level  $i$ ; the size of the cluster is defined as the number of adjacent pixels with the same grey-level. The NGTDM matrices “describe the differences between each voxel and its neighbouring voxels in adjacent image planes” (Ypsilantis, et al., 2015); Amadasun and King (1989) describe NGTDM metrics as being similar to the human experience of imaging. Each entry in an NGTDM matrix denotes the sum of the differences between pixels with a grey-level value  $i$  and the average value of those pixels surrounding neighbours (Ypsilantis, et al., 2015; Sah, et al., 2019).

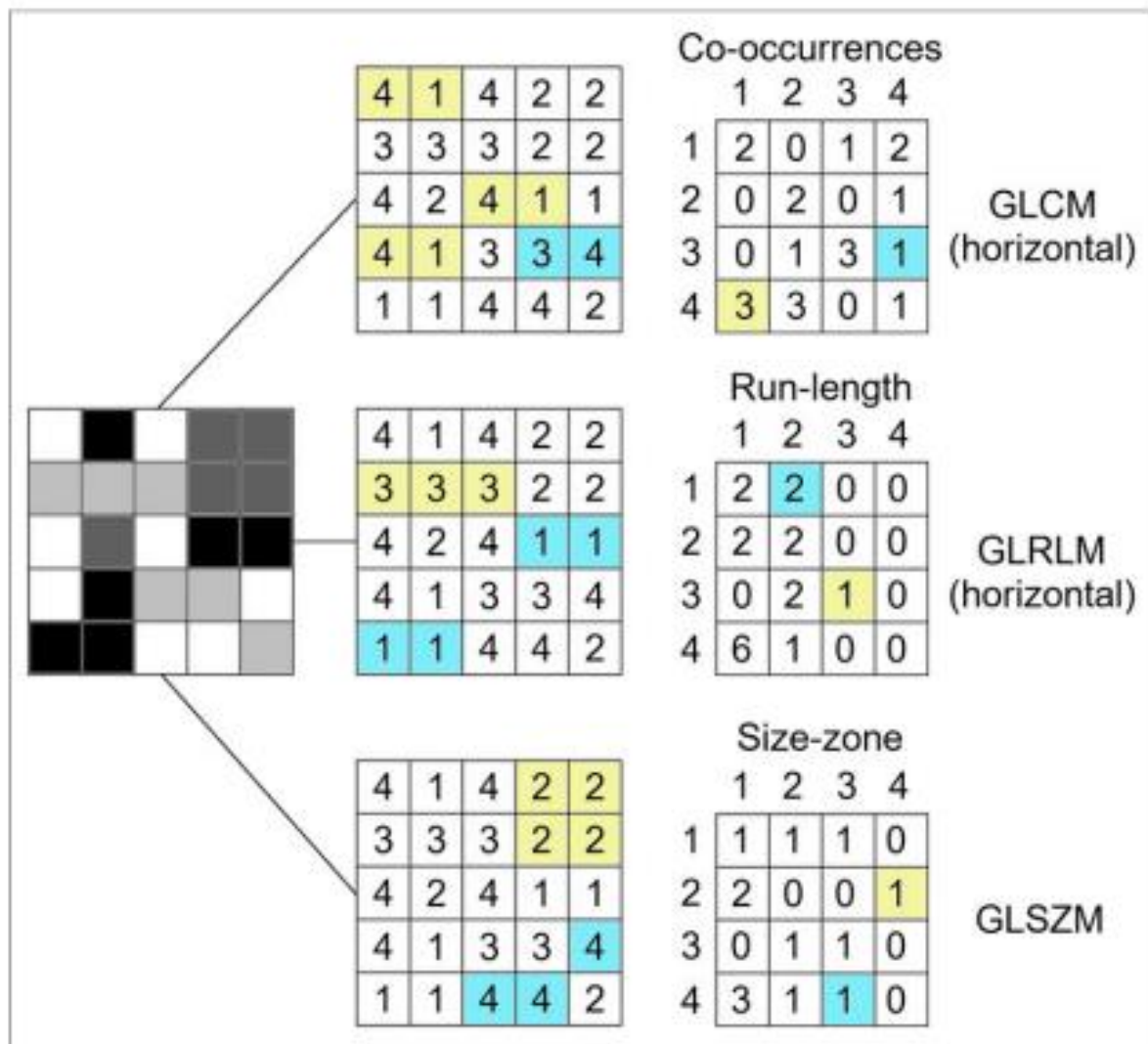
**Figure 15 Grey-level size zone matrix**

1	2	3	4
1	3	4	4
3	2	2	2
4	1	4	1

<i>Level</i>	<i>Size zone, s</i>		
<i>g</i>	1	2	3
1	2	1	0
2	1	0	1
3	0	0	1
4	2	0	1

*Example illustration of how a grey-level size zone matrix is formed. Source: Thilbault (2006).*

**Figure 16 Grey-level co-occurrence, run length and size zone matrix**



Example illustration of a grey-level co-occurrence, run length and size zone matrix is formed. Source: Mayerhoefer (2020).

A brief note on terminology, in the literature, higher order statistics are referred to both as grey level matrices relating to multiple pixels (Sah, et al., 2019) and to more complex features such as wavelets (Gillies, et al., 2016). And definitions of the individual parameters such as the grey level zone length can also be referred to as the size zone (Aerts, et al., 2014). In this project, we will refer to the definitions described in and produced by Nioche et al (2018).

### 3.3.4 Radiomics in PET

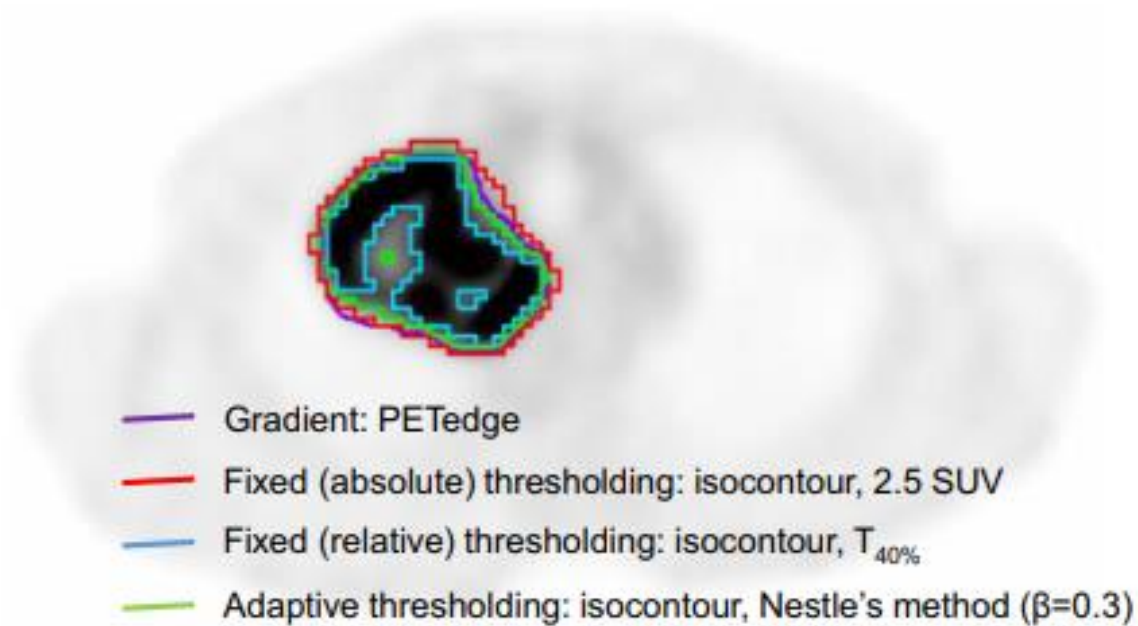
There are several fundamental differences between radiomic measurements on CT or MRI with PET. Firstly, the voxel size in PET is much larger than CT or MRI, several millimetres, and sub-millimetres respectively. Larger voxel sizes result in a much coarser tumour sampling which presents a

disadvantage in assessing tissue properties however, where PET is preferable over CT (tissue density) or MRI (proton density) is that we have a detailed map of cellular glucose metabolism, which may also hold potential in predicting disease response (Cook, et al., 2018; Sah, et al., 2019). Furthermore, this study is specifically using Fluorine-18-fluorodeoxyglucose ( $^{18}\text{F}$ -FDG) PET and alternative tracers may yield different results, for example, Ma et al (2015) compared radiomic performance with pathologic staging features for squamous cell oesophageal carcinomas. Ma et al (2015) found that  $^{18}\text{F}$ -FDG features were more significantly associated with pathologic predictors (Van Rossum, et al., 2016). Furthermore, Ma et al (2015) found that standard uptake value (SUV) max; tumour length and eccentricity were the most important features for  $^{18}\text{F}$ -FDG.

An important first step in applying and mining radiomics data from PET images is using a “relative quantization process” (Hatt, et al., 2016) whereby the image is re-sampled into SUV intensity bins, e.g. pixels are placed into bins of 0.5 SUV size. Leijenaar et al (2015) have investigated the effect of different sampling approaches to SUV discretisation on the subsequent textural features extracted via radiomics analysis. Leijenaar et al (2015) varied the SUV bin size from 0.05 to 1 and the number of bins from 8-128 bins concluding that fixing the number of bins and bin size was the key factor when comparing results between patients with small cell lung cancer. A review paper by Cook et al (2018) described 64 equally distributed bins as the consensus approach.

Ypsilantis et al (2015) describe a study using 107 oesophageal cancer patients who underwent neo-adjuvant chemotherapy; 38/107 patients responded to treatment and the overall survival time was defined as the number of days between the PET/CT scan and date of death (Responders OS = 972.5 days, Non-responders OS = 714 days). A “40% slice-wise maximum threshold” (Ypsilantis, et al., 2015) was used to delineate tumour volumes and textural analysis was performed using both statistical and model-based approaches. However, using a 40% max SUV threshold may present challenges for particularly heterogeneous tumours with ‘colder’ patches (Cook, et al., 2018) and a threshold method of either fixed SUV (e.g. 2.5) or an adaptive / gradient model, may be more preferable to ensure more of the treatable tumour is captured for analysis (Ha, et al., 2019)(Figure 17).

**Figure 17 Tumour thresholding methods**



*A comparison of different tumour thresholding methods: Gradient, fixed (absolute using a minimum SUV cut-off), fixed (relative to SUVmax) and Adaptive using Nestle's method of isocontouring. Source: Ha et al (2019).*

Another factor to be conscious of is the tumour size; for example, Brooks and Grigsby (2014) used probability theory to calculate that a tumour volume of  $45\text{cm}^3$  is required for adequate sampling without significant bias for  $^{18}\text{F}$ -FDG images of cervical cancer. Brooks and Grigsby (2014) have derived a “minimum volume” by using the probability of a count intensity in a bin being statistical sufficient; a volume of  $45\text{cm}^3$  or larger gave a probability of at least 95% that there would be sufficient samples in the least populated intensity bin to make meaningful statistical calculations.

Hatt et al (2015) reported a high correlation between texture features and tumour volume, and between second order features in much smaller volumes of  $10\text{cm}^3$ , suggesting that lower volumes can also be effective. On tumour segmentation, Ha et al (2019) conclude that there is not yet an agreed consensus on the recommended method for delineation, owing mainly to consensus on whether to “include the necrotic portion” of a tumour in PET imaging. The most popular methods for oesophageal cancer have been either manual contouring or region growing to SUV values of greater than or equal to 2.5 (Van Rossum, et al., 2016). Lambin et al (2017) describe a “radiomics quality system” (Figure 18) which should serve as a guide for ensuring that any prospective studies can be both replicated and validated against other studies; this however has limited application to this study since we focussed on using a retrospective dataset.



To complement discussions around standardisation, Tixier et al (2012) describe a study investigating the reproducibility of radiomic features between separate scans, within 4 days of each other, on the same patient with oesophageal cancer. Tixier et al (2012) conclude that whilst SUVmax and mean showed good reproducibility, some texture features including “entropy and homogeneity” showed better reproducibility between scans and that many tumour regional heterogeneity measures showed similar reproducibility to that of SUV.

One of the key challenges for radiomics in PET is to achieve appropriate standardisation of imaging protocols, smoothing levels, quantization levels, segmentation methods and feature stratification (Lambin, et al., 2017; Van Rossum, et al., 2016; Sah, et al., 2019). Shiri et al (2017) have investigated the robustness of radiomics parameters with varying scanner and reconstruction settings. Geometry and intensity based features, such as tumour volume and SUVmax respectively, were relatively robust to different reconstructions whereas some textural features such as grey-level run-length matrices were more sensitive to changes in reconstruction and varied with a coefficient of variance up to 20% (Shiri, et al., 2017). One of the main challenges to using radiomics in PET, and indeed other modalities, is the lack of standardisation, particularly when using retrospective datasets. The focus on different feature sets and use of a variety of in-house software has led to inconsistencies between published works (Lambin, et al., 2017). **Many published studies in relation to radiomics, PET imaging and oesophageal cancer have used small cohorts of patients and there is therefore a need to validate currently published works using larger cohorts of patients using all available radiomic features from the software.**

**Figure 18 Radiomics Quality System**

Criteria	Points
1 Image protocol quality - well-documented image protocols (for example, contrast, slice thickness, energy, etc.) and/or usage of public image protocols allow reproducibility/replicability	+1 (if protocols are well-documented) +1 (if public protocol is used)
2 Multiple segmentations - possible actions are: segmentation by different physicians/algorithms/software, perturbing segmentations by (random) noise, segmentation at different breathing cycles. Analyse feature robustness to segmentation variabilities	+1
3 Phantom study on all scanners - detect inter-scanner differences and vendor-dependent features. Analyse feature robustness to these sources of variability	+1
4 Imaging at multiple time points - collect images of individuals at additional time points. Analyse feature robustness to temporal variabilities (for example, organ movement, organ expansion/shrinkage)	+1
5 Feature reduction or adjustment for multiple testing - decreases the risk of overfitting. Overfitting is inevitable if the number of features exceeds the number of samples. Consider feature robustness when selecting features	-3 (if neither measure is implemented) +3 (if either measure is implemented)
6 Multivariable analysis with non radiomics features (for example, EGFR mutation) - is expected to provide a more holistic model. Permits correlating/inferencing between radiomics and non radiomics features	+1
7 Detect and discuss biological correlates - demonstration of phenotypic differences (possibly associated with underlying gene-protein expression patterns) deepens understanding of radiomics and biology	+1
8 Cut-off analyses - determine risk groups by either the median, a previously published cut-off or report a continuous risk variable. Reduces the risk of reporting overly optimistic results	+1
9 Discrimination statistics - report discrimination statistics (for example, C-statistic, ROC curve, AUC) and their statistical significance (for example, p-values, confidence intervals). One can also apply resampling method (for example, bootstrapping, cross-validation)	+1 (if a discrimination statistic and its statistical significance are reported) +1 (if a resampling method technique is also applied)
10 Calibration statistics - report calibration statistics (for example, Calibration-in-the-large/slope, calibration plots) and their statistical significance (for example, P-values, confidence intervals). One can also apply resampling method (for example, bootstrapping, cross-validation)	+1 (if a calibration statistic and its statistical significance are reported) +1 (if a resampling method technique is also applied)
11 Prospective study registered in a trial database - provides the highest level of evidence supporting the clinical validity and usefulness of the radiomics biomarker	+7 (for prospective validation of a radiomics signature in an appropriate trial)
12 Validation - the validation is performed without retraining and without adaptation of the cut-off value, provides crucial information with regard to credible clinical performance	-5 (if validation is missing) +2 (if validation is based on a dataset from the same institute) +3 (if validation is based on a dataset from another institute) +4 (if validation is based on two datasets from two distinct institutes) +4 (if the study validates a previously published signature) +5 (if validation is based on three or more datasets from distinct institutes)  *Datasets should be of comparable size and should have at least 10 events per model feature
13 Comparison to 'gold standard' - assess the extent to which the model agrees with/is superior to the current 'gold standard' method (for example, TNM-staging for survival prediction). This comparison shows the added value of radiomics	+2
14 Potential clinical utility - report on the current and potential application of the model in a clinical setting (for example, decision curve analysis).	+2
15 Cost-effectiveness analysis - report on the cost-effectiveness of the clinical application (for example, QALYs generated)	+1
16 Open science and data - make code and data publicly available. Open science facilitates knowledge transfer and reproducibility of the study	+1 (if scans are open source) +1 (if region of interest segmentations are open source) +1 (if code is open source) +1 (if radiomics features are calculated on a set of representative ROIs and the calculated features and representative ROIs are open source)
Total points (36 = 100%)	

*A comprehensive radiomics quality system. Source: Lambin et al (2017)*

### **3.3.5 Effect of Reconstruction on Radiomics Parameters**

Several groups have investigated the relative stability of various radiomics parameters to different reconstruction parameters, modes and models. Galavis et al (2010) evaluated 50 different radiomics parameters with 10 different combinations of iterations, subsets and post filters, for both 2D and 3D datasets using ordered subsets expectation maximum (OSEM) iterative reconstructions respectively. Galavis et al (2010) categorised the changes in value between different reconstructions as “Small (<5%), intermediate (10-25%) and Large >30%” concluding that features such as “entropy-first order, energy, maximal correlation coefficient, and low-grey level run emphasis” were good candidates for reproducible tumour segmentation but that the majority of other parameters used exhibited large variations between 2D and 3D scans. In our study, we present a dataset of exclusively 3D studies, specifically excluding any scans acquired in 2D.

Hatt et al (2013) describe a study investigating the “robustness of intra tumour uptake heterogeneity quantification for therapy response-prediction in oesophageal carcinoma” concluding that parameters such as entropy, homogeneity, dissimilarity, and zone percentage were most robust to delineation methods and partial volume effects. Whilst investigating for oesophageal cancer patients and 3D data, Hatt et al (2013) however only explored the effect with OSEM and a partial volume correction (PVC) and did not explore any further impact on radiomics features with other reconstruction methods such as Block Sequential Regularized Expectation Maximum (BSREM).

Yan et al (2015) explored the effect on texture parameters in OSEM, time-of-flight (TOF) and point-spread-function (PSF) reconstructions using various combinations and parameters, for a group of 20 patients with lung lesions. Yan et al (2015) concluded that entropy; low grey-level run emphasis, high grey-level run emphasis and low-grey level zone emphasis were most robust to changes in reconstruction method. Van Helden et al (2016) describe a similar study investigating the impact of reconstruction and tumour delineation on radiomic features for patients with non-small-cell lung cancer. Van Helden et al (2016) compared radiomic features on images reconstructed with OSEM+TOF and 7mm gaussian filter (EANM standard, (Boellaard, et al., 2015)) and OSEM+TOF+PSF and found a high level of repeatability of the majority of radiomic features in this patient group. Furthermore, Van Helden et al (2016) found that delineation on CT was more robust than delineation on PET imaging alone. However, Yan et al (2015) or Van Helden et al (2016) have investigated the effects of BSREM on radiomic features.

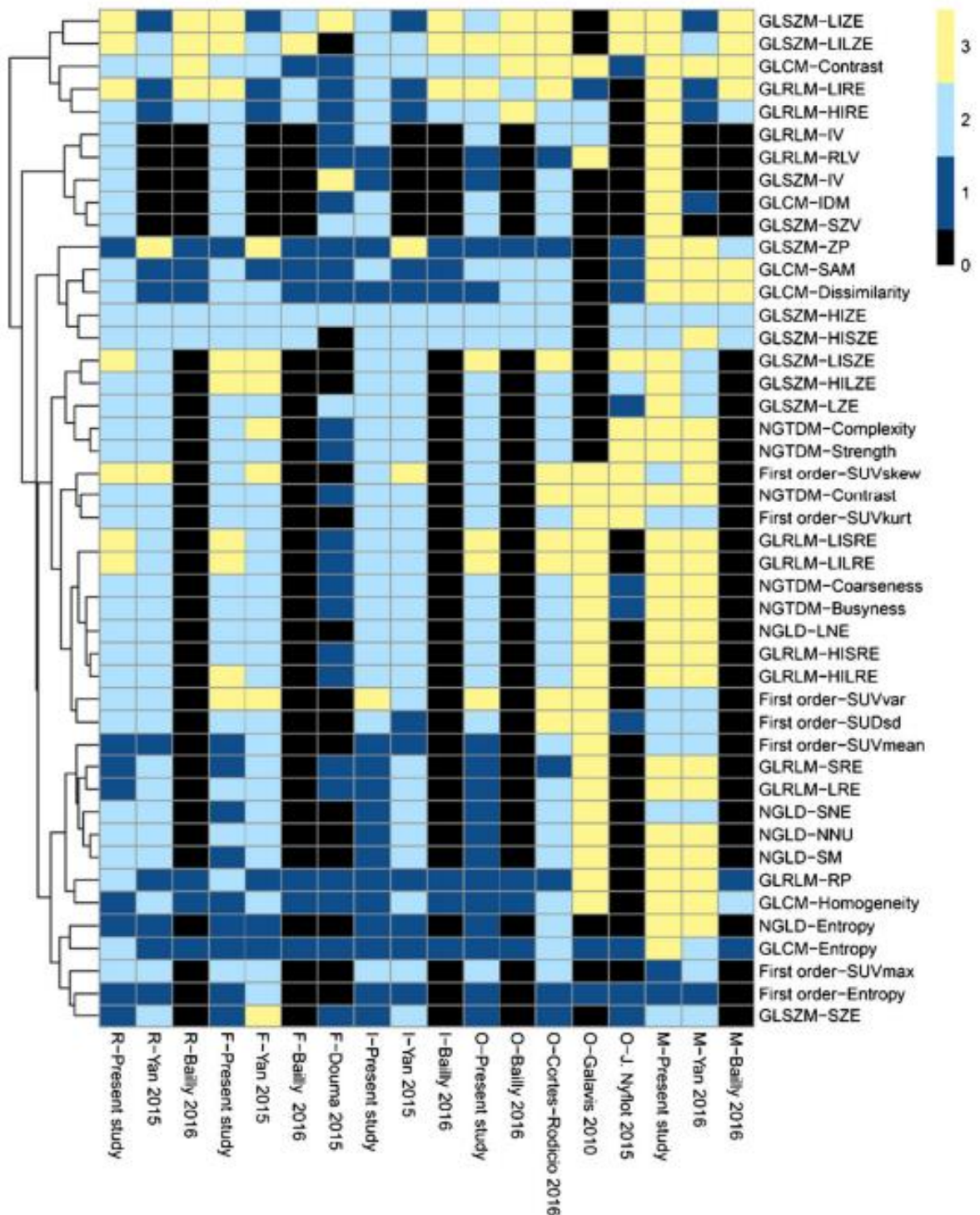
Shiri et al (2017) evaluated reconstruction algorithms in addition to different settings and showed that robustness to different reconstruction algorithms is feature dependent. Shiri et al (2017) have

performed a comparison of other studies investigating the relative robustness of different radiomics parameters (Figure 19) citing that their study is unique in comparing the stability of radiomics parameters with more advanced reconstruction techniques such as OSEM, TOF and PSF. In Figure 19, studies and parameters graded with 1 (dark blue) showed good robustness to the effects of different imaging parameters. Across all studies, the following features were found to be the most robust to different reconstruction algorithms: “GLCM (Entropy, Homogeneity, Dissimilarity, Correlation), GLRLM (SRE, LRE, RLV, RP), GLSZM (SZE, IV, ZP), NGLCM (Entropy, Homogeneity, Dissimilarity), Intensity and SUV (SUVmean, Entropy, SULpeak...)”. Shiri et al (2017) include studies describing robustness to: reconstruction (R), Gaussian Filter FWHM (F), Iteration number (I), Matrix Size (M) and Overall (O). Notably, Shiri et al (2017) have excluded BSREM from discussions, which further highlights a need for this work.

Ger et al (2019) describe a detailed study using the ‘Hoffman’ brain phantom, investigating the effects of a variety of scanners, reconstruction methods and parameters on radiomic features. In their study, they have included analysis using a GE710 series scanner (as proposed for this study) and described analysis using OSEM, OSEM+TOF and OSEM+TOF+PSF. In relation to the reproducibility of radiomic features between scanners and vendor reconstruction methods, Ger et al (2019) cite “92% of features had excellent reliability” but state alongside this that these results were produced “when Q.Clear was not included (reconstruction types QCFX-S and QCHD-S)” (NB QCFX-S = OSEM+TOF+PSF+BSREM, QCHD-S = OSEM+PSF+BSREM). The publication unfortunately does not describe or expand on how features perform once Q.Clear (BSREM) is included in analysis. Reynes-Llompart (2019) has presented in a thesis body of work a thorough investigation into the effects of BSREM reconstruction on various radiomics parameters. One of their aims was to investigate “the direct impact of BSREM reconstruction in the stability of heterogeneity features” (Reynes-Llompart, 2019). However, their study was primarily focussed on optimising the Q.Clear algorithm rather than comparing performance to OSEM alone (Reynes-Llompart, et al., 2018).

Reynes-Llompart (2018) and (2019) have studied the impact of BSREM and OSEM on image quality and Ger et al (2019) presented a review of all reconstruction methods for different manufacturers using phantom studies but both fall short of investigating the effect of BSREM, particularly in a clinical setting. **To the best of our knowledge, there is limited discussion in the literature of the effect of BSREM on radiomics parameters, and no studies on the effect of BSREM when using machine learning algorithms to predict survival of oesophageal cancer patients, based on the radiomic signature.**

Figure 19 Robustness of features comparison to other works



A comparison of the robustness of features from other works. R = Reconstruction, F = FWHM, I = Iteration, O = overall, M = matrix, 0 = not calculated in this study, 1 = Most Robustness, 2 = intermediate robustness, 3 = low robustness. Source: Shiri et al (2017).

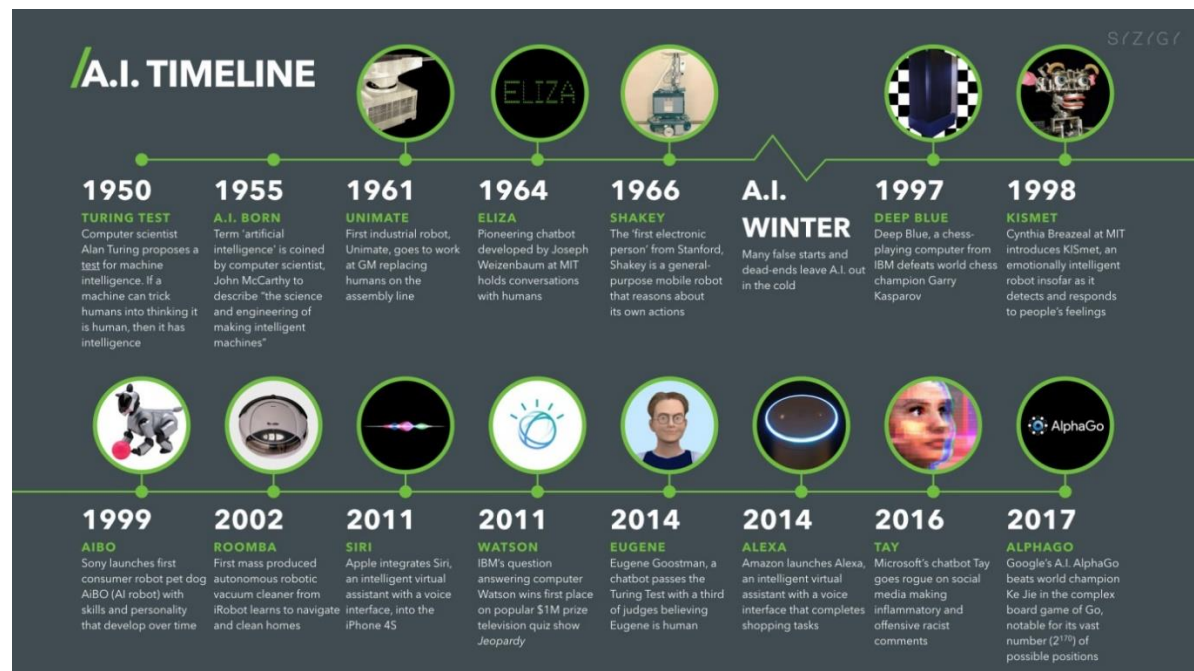
### 3.4 Artificial intelligence and machine learning

This section introduces artificial intelligence including discussion of machine learning algorithms. This section will also include a brief introduction to deep learning to and complement the context of some similar works investigating outcome prediction from PET images for oesophageal cancer.

#### 3.4.1 Introduction and history of Artificial Intelligence

The term “Artificial Intelligence” (AI) was first described in 1956 and in its simplest form describes the ability for computers to perform tasks which usually require human intelligence for example, decision making and visual perception (Uribe, et al., 2019). In recent years, “machine learning” (ML) has become an increasingly prevalent part of radiological discussion and research however, ML was first discussed in the 1950’s (Rosenblatt, 1958) and applied to medicine in the 1960s (Silink, 1961). Between 1966 and 1997 there were many failed attempts to advance AI before IBM presented “Deep Blue”, a chess playing super computer which was able to use a form of artificial intelligence to beat a world champion at chess (Marsden, 2017). Mainstream AI applications are found in everyday life such as “facial recognition, speech recognition, language translation, web searches and autonomous driving” (Lee, et al., 2019). Their increased application is being rapidly progressed by large tech companies such as Apple (“Siri”), Amazon, Microsoft and Google (Marsden, 2017) .

**Figure 20 History of Artificial Intelligence**



*A brief history of artificial intelligence. Source: Marsden (2017).*



Artificial Intelligence (AI) holds great promise for diagnostic imaging with the most attention currently being paid to fine-tuning performance to facilitate the detection of a variety of clinical conditions (Oren, et al., 2020). The majority of AI research into medical imaging has been on lesion detection but ignoring any lesion stratification and therefore creating “a skewed representation of AI’s performance” (Oren, et al., 2020). Oren et al (2020) put forward the case for more refinement in AI studies, focussing on “clinically meaningful endpoints such as survival, symptoms and need for treatment”. One of the earliest uses of AI in Healthcare was for “computer-assisted detection” (CAD) which aimed to highlight areas of concern on an image, such as for mammography x-rays (Lee, et al., 2019). CAD has been further demonstrated in reviewing chest x-rays (CXR); initial focus has been on mammography and CXRs mainly due to the volume of data giving good statistics for teaching CAD models. AI is used in a variety of other healthcare applications such as in Histopathology assisting pathologists in identifying nodal disease from slides, dermatology in identifying cancerous skin lesions from medical photographic images and ophthalmology in detecting diabetic retinopathy (Lee, et al., 2019).

One area where AI may help in Radiology is in the automation or semi-automation of reporting, by helping to bridge the gap in a workforce facing significant staff shortages (Lee, et al., 2019). Pesapane et al (2018) describe the role of Radiologists in the introduction of AI into healthcare and describe the recent dramatic increase in publications “from about 100–150 per year in 2007–2008 to 700–800 per year in 2016–2017” with “magnetic resonance imaging and computed tomography collectively account for more than 50% of current articles”. In terms of applications, Neurology accounts for approximately a third followed by an approximately even split amongst musculoskeletal, cardiovascular, breast, urogenital, lung / thorax and abdomen studies (Pesapane, et al., 2018). In Nuclear medicine, studies have been described for a variety of applications such as for SPECT and PET which, as AI continues to grow, is more capable of handling the larger datasets presented by hybrid modalities such as PET/CT, SPECT/CT and PET/MRI. However, in spite of a dramatic overall increase in AI medical imaging publications, PET and SPECT still account for around 1% of publications (Pesapane, et al., 2018). This is likely due to a combination of factors: Nuclear Medicine accounts for a much smaller share of both the imaging performed and the workforce working on such images.

Some of the key applications of AI in PET/CT imaging include image analysis (Amico, et al., 2020), tumour identification (Zhang, et al., 2020), image segmentation (Avanzo, et al., 2020) and treatment response (Wei & El Naqa, 2020). One of the main areas of interest in AI and PET/CT imaging is in response prediction by using either pre or post treatment PET/CT images. Wei and El Naqa (2020) describe a review of using a variety of machine / deep learning methods for obtaining regions of interest and predicting treatment outcomes from PET/CT images. Many authors have investigated

treatment response for oesophageal cancer in a variety of ways such as Rahman et al (2019) who pooled full clinical data from seven different centres to predict response to treatment using a random forest machine learning method. Furthermore, many authors have used AI on both CT and MRI to investigate treatment response (Pesapane, et al., 2018) and many have used PET/CT to investigate treatment response for a variety of conditions such as lung cancer. However, to date, there have been sparse publications specifically for using AI to predict treatment response in oesophageal cancer using  $^{18}\text{F}$ -FDG PET images (Xiong, et al., 2018; Yang, et al., 2019) and even fewer with a published, reproducible methodology (Ypsilantis, et al., 2015). There is therefore a need to contribute to this area of research and broaden the knowledge base for this application.

### **3.4.2 Machine Learning**

This section gives a brief introduction to machine learning, its background (in general terms) and a brief description of the machine Learning algorithms used in this project: Logistic Regression (LR), Support Vector Machine (SVM), Linear Discrimination Analysis (LDA), K-nearest Neighbours (KN), Gaussian Naive Bayes (GNB) and Decision Tree Classifier (DT) algorithms.

ML is a branch of AI and describes computer algorithms programmed to learn from observations and make decisions using statistical metrics. These metrics are enhanced and built upon with increasing amounts of data, thus 'learning' new information (Uribe, et al., 2019). The rise in the recent popularity of AI and ML is due to improved theory, hardware and availability of large amounts of training data (Kelchtermans, et al., 2014). In its simplest form, an ML algorithm requires a mathematical model, a cost function and adequate data (Uribe, et al., 2019). The mathematical model is a function, which links an input to an output. There are many models available, and the exact choice of which to use depends on a trade-off between accuracy, suitability, and implementation. The cost function is a measure of how closely the model resembles the intended result, for example the area under receiver operator curve (ROC), misclassification rate etc. In machine learning, algorithms can be supervised or unsupervised. A supervised algorithm uses known outcomes for training and determines unknown parameters in the mathematical model (ML algorithm) using the training data. An unsupervised algorithm is one which the outputs are unknown, and the task of the program may solve problems such as determining common features e.g. between two images without prior knowledge of the outcome (Uribe, et al., 2019). For regression-type models, there are several approaches such as support vector machines (Van Weegaeghe, et al., 2016), random forest (Ingrisch, et al., 2018) and Artificial Neural Networks (ANN) (Wei & El Naqa, 2020).



## Logistic Regression

The logistic regression algorithm uses a sigmoid function:

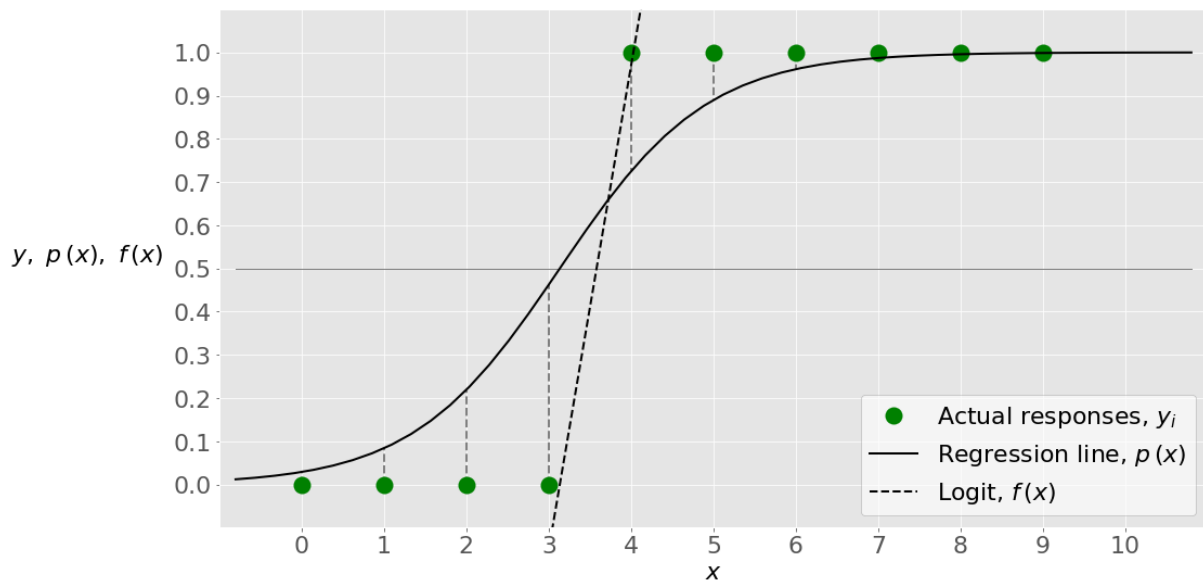
$$p(x) = 1 / (1 + \exp(-f(x))) \quad \text{Equation 6}$$

Whereby the function  $p(x)$  is interpreted as “the probability that the output for a given  $x$  is equal to 1” and function  $f(x)$  is a linear function:  $f(x) = b_0 + b_1x_1 + \dots b_r x_r$ . The coefficients  $b_0, b_1 \dots b_r$  are determined by the machine learning process to identify the values of the coefficients such that the function  $p(x)$  best matches the actual responses (Stojiljkovic, 2022). The optimum weights are determined by maximising the “log-likelihood function (LLF)” for all observations:

$$\text{LLF} = \sum_i (y_i \log(p(x_i)) + (1 - y_i) \log(1 - p(x_i))) \quad \text{Equation 7}$$

In our study, the aim is to determine the weights for the LLF which best separates a successful treatment from a failed treatment, based on multiple input features such as radiomic parameters and the TNM score. Logistic regression is ideally suited to classification problems.

**Figure 21 Logistic Regression**



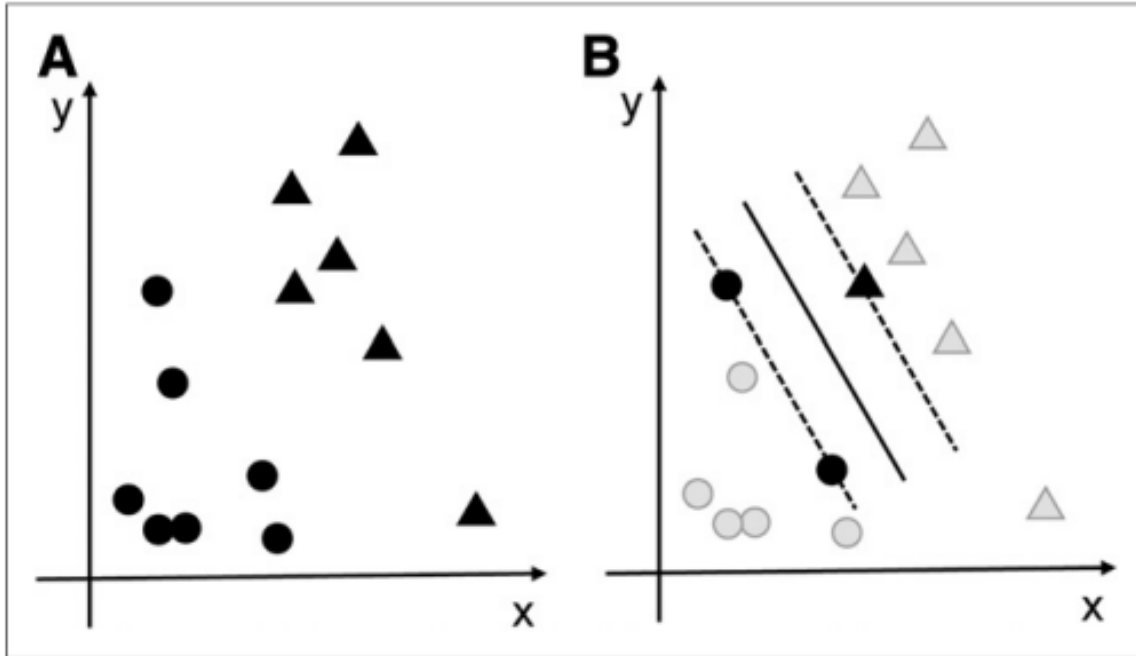
*Illustration of logistic regression separation. Source: Stojiljkovic (2022).*

## Support Vector Machine

A support vector machine identifies a curve (or hyperplane for multiple dimensions) which best separates two classes such that the resulting curve is at the maximum distance from the two closest points between the two classes (Figure 22); “the closest points to the line are called the support vectors” (Uribe, et al., 2019). For example, A indicates the raw separation of the training dataset; and

in B, the solid line indicates the maximum separation from the dashed lines (closest points or support vectors) (Uribe, et al., 2019).

**Figure 22 Support Vector Machine**



*Illustration to show how a support vector machine learning algorithm is used to separate data points. Source: Uribe et al (2019).*

### **Gaussian Naïve Bayes**

The Gaussian Naïve Bayes (GNB) approach follows a simple probability distribution, is relatively low in computational burden and is one of the simpler ML classification algorithms. Mehta et al (2017) have described a Naïve Bayes approach to classification algorithm to predict the response to <sup>90</sup>Y radio-embolisation therapy based on <sup>18</sup>F-FDG PET/CT scan features. The GNB algorithm uses ‘Bayes Theorem’, using a conditional probability formula of event A taking place given that event B has happened:

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)} \quad \text{Equation 8}$$

Where “A and B are two events; P(A|B) is the probability of event A provided event B has already happened; P(B|A) is the probability of event B provided event A has already happened; P(A) is the independent probability of A and P(B) is the independent probability of B” (Sharma, 2021).

In this instance, a Naïve Bayes algorithm assumes that all predicting features contribute equally and the Gaussian Naïve Bayes assumes that the values are continuous and follow a Gaussian distribution. The above formula can be extended to include multiple variables:

$$P(y|x_1, x_2, x_3 \dots x_N) = \frac{P(x_1|y).P(x_2|y).P(x_3|y)\dots P(x_N|y)P(y)}{P(x_1).P(x_2).P(x_3)\dots P(x_N)} \quad \text{Equation 9}$$

Where “X = x<sub>1</sub>,x<sub>2</sub>,x<sub>3</sub>,... x<sub>N</sub> are independent predictors; y is the class label and P(y|X) is the probability of label y given the predictors X” (Sharma, 2021).

### **Linear Discrimination Analysis**

Linear discrimination analysis (LDA) is a probabilistic model, developed by a specific distribution of observations for each input variable. A new feature is “is then classified by calculating the conditional probability of it belonging to each class and selecting the class with the highest probability” (Brownlee, 2020). LDA follows Bayes Theorem and is essentially a simple version of GNB and works by creating summary statistics (e.g. mean and standard deviation) for all the input features by the class label (success and failure). From the statistics associated with each feature, the LDA algorithm then calculates the probability that a given feature of a particular value belongs to which class. The advantage of LDA is that this is a relatively simple analysis tool however; LDA assumes that input variables, which follow a normal distribution, have the same variance and do not correlate with each other (Brownlee, 2020).

### **K-nearest neighbours Classifier**

The k-nearest neighbours (KN) Classifier algorithm is a supervised learning algorithm and creates “an imaginary boundary to classify the data” (Anon., 2021). KN can solve regression or classification problems, this project uses the latter. KN uses the entire dataset and is therefore susceptible to any anomalies and outliers. Predictions are made on new data by searching through the entire dataset for the K most similar instances (neighbours) and summarising the output for those K instances. KN works well for a small number of variables however, with larger sets of input dimensions, KN can struggle to classify correctly (Brownlee, 2016).

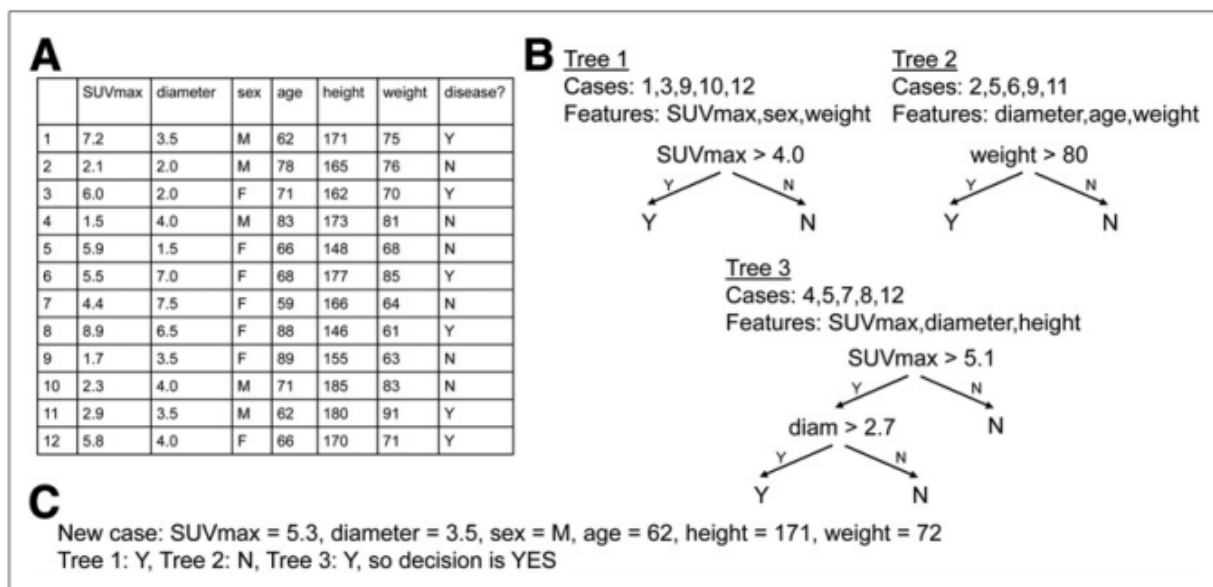
### **Decision Tree Classifier and Random Forest**

Decision tree classifiers and random forest classifiers operate in a similar way in that both algorithms take the data and separate it into two groups according to a classification question. The key difference is that a decision tree classifier plots all possible outcomes from a decision node and uses the entire dataset whereas the random forest algorithm randomly selects the features to test, according to the output (see Figure 23). The advantage of a decision tree is, as the name suggests, that it is a single tree

of decisions. Decision trees use the whole dataset so they are easier to interpret but are susceptible to over-fitting. The advantage of random forest is that this is a forest of smaller decision trees rather than one large decision tree, which is ultimately less computationally demanding. However, the user cannot control the “randomness” of a random forest algorithm making the results more difficult to interpret (Chauhan, 2022).

A random forest algorithm is a type of supervised algorithm, used mainly for classification or regression problems. The random forest algorithm is based on taking a random subset of e.g. images and a random subset of features to test on the image subset and attempts to split the group into a binary “yes / no” category. The tree ends when a branch contains all images in a single class. If the first branch does not split completely into a single class, the random forest moves to the next random feature (Uribe, et al., 2019). For example, in Figure 23, dataset (A), a training dataset with 12 cases and 6 features, creates three decision trees (B) each “generated from 5 randomly selected cases and 3 randomly selected features”. In Tree 1, all randomly selected cases fall into category “N” in that all cases have SUVmax <4.0, similarly in Tree 2, all cases fall into “N” with all randomly selected cases having a weight <80. In other words, for trees 1 and 2, a single feature is able to fully separate 5 cases. Tree 3, using SUVmax >5.1, does not sufficiently separate all randomly selected cases into 1 group and so a second classifier, diameter >2.7, is applied to achieve complete separation. When a new case is introduced (C), the data is then fed through the set of pre-trained decision trees and the new study is classified according to the majority decision of the pre-trained decision trees. In other words, a random forest is a set of decision trees, trained using a set of data but fed in using both random imaging subsets and classification subsets.

**Figure 23 Random Forest Algorithm**



*Illustration of a random forest algorithm example. A: Total data set, B: example randomly selected decision trees from randomly selected subgroups of data, C: results from a ‘newly tested’ data entry. Source: Uribe et al (2019).*

**Figure 24 Summary of machine learning algorithms**

Algorithm	Task	Supervision?	Model	Typical cost function	Computational burden	Assumptions/ comments
Naïve Bayes classification	Classification	Supervised	Several (e.g., Gaussian)	Probabilistic	Low	Relies on naïve probability distribution
Linear regression	Regression	Supervised	Hyperplane	MSE	Low	
Support vector machines	Classification or regression	Supervised	Hyperplane	Classification rate, MSE	Moderate	Handles complex problems
Random forest	Classification or regression	Supervised	Tree	Classification rate, MSE	Low–moderate	Is tolerant to overfitting
ANN	Classification or regression	Supervised (typical); unsupervised/ reinforcement learning (less common)	Neurons connected in layers	Classification rate, MSE	High	Is used for complex problems; may be convolutional or deep
k-means clustering	Clustering	Unsupervised	Cluster centroid	Distance to cluster center	Moderate (depends on problem)	Identifies centroids and assigns data to nearest centroid
Hierarchical clustering	Clustering	Unsupervised	Dendrogram	Distance between data points	Moderate (depends on problem)	Clusters data by identifying data-points that are similar
Principal-component analysis	Dimensionality reduction	Unsupervised	Principal components		Moderate (depends on problem)	

MSE = mean square error.

*A table summary of some machine learning algorithms including the supervision, cost function and computational burden. Source: Uribe et al (2019).*

In addition to Figure 24, Xiong et al (2018) describe a method “extreme machine learning” which “is defined as a single-hidden layer feedforward neural network, and it randomly chooses hidden nodes and analytically determines the output weights of the feedforward neural networks”; in other words, a simplified form of a neural network, see section 3.4.3. For more detail on the specific algorithms chosen for this project, and why, see section 0.

### 3.4.3 Deep learning architecture

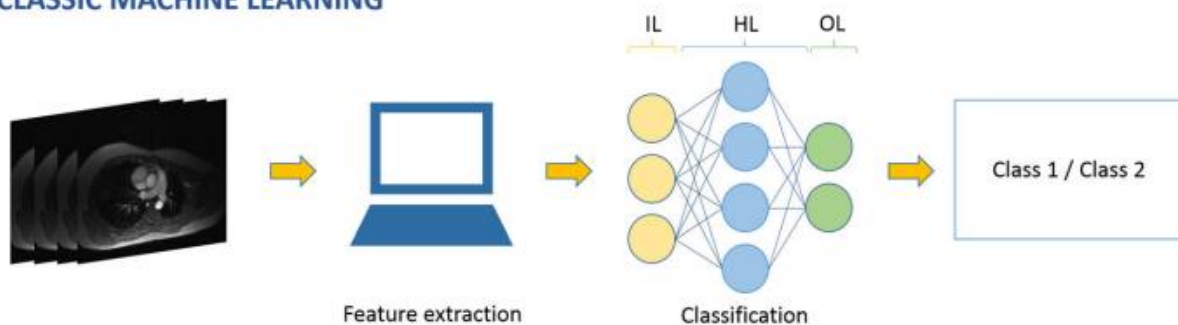
This project does not investigate convolutional neural networks (CNN) in detail. This section provides additional background and context for discussion relating to Artificial Intelligence and studies using neural network architecture to predict outcomes for oesophageal cancer PET images.

An Artificial Neural Network (ANN) forms the overarching term for algorithms commonly used for ‘deep learning’ and demonstrates the next level up in computational complexity. ANNs can solve

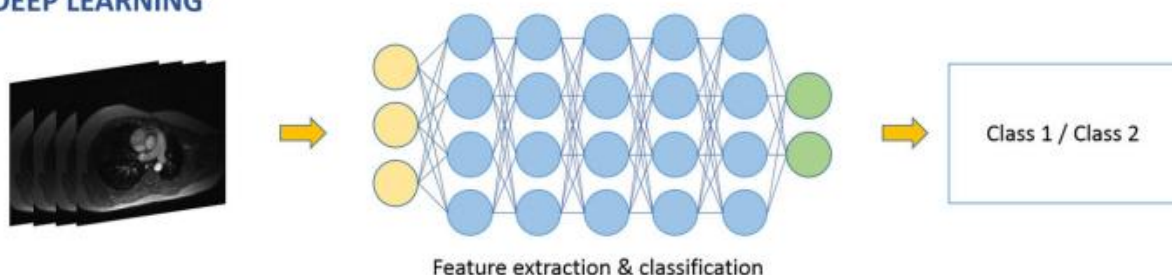
regression or classification problems and process the data in several steps or ‘layers’. Each layer involves several computational units (‘neurons’) which process the data and pass the result to multiple neurons in the next layer (Uribe, et al., 2019). Each neuron in the network will typically take a weighted sum of the input and apply a bias before applying a “non-linear transformation” before passing to the next layer (Uribe, et al., 2019). ANNs are more computationally intensive as there are often 1000s of neurons in a layer which link to another set of 1000s of neurons with 1000s of input data points. Deep learning as applied to ANNs (DNN), is essentially many more layers of neurons (10-150) and places further computational demand on solving problems (Uribe, et al., 2019). Both ML and DL contain an input layer (IL), hidden layers (HL) and output layer (OL). The key difference between machine learning (ML) and deep learning (DL), is that for classical machine learning (e.g. support vector machines), feature extraction is performed before the data is passed to the ML algorithm whereas deep learning (e.g. ANN) uses the computational layers to perform both feature extraction and classification tasks (Figure 25). In this example, class 1 and 2 may be e.g. survival analysis whereby class 1 = survived and class 2 = deceased.

**Figure 25 Machine vs deep learning**

#### CLASSIC MACHINE LEARNING



#### DEEP LEARNING



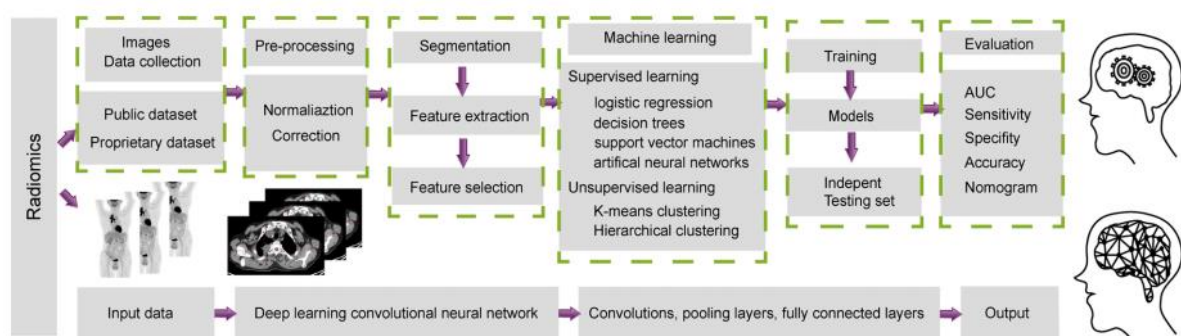
*Illustration of machine learning vs deep learning to show the key difference: in ML feature extraction is performed prior to classification whereas DL, feature extraction is included in the algorithm. Source: Pesapane et al (2018).*

Litjens et al (2017) describes an overview of deep learning techniques in medical imaging with MRI, Microscopy and CT dominating most publications. Litjens et al (2017) further cite very few studies for deep learning and PET imaging (Ypsilantis, et al., 2015). Litjens et al (2017) provide an overview of the

most common deep learning architectures and algorithms found in application to medical imaging. For deep learning algorithms, data classification or regression, and prediction tasks performed jointly in the hidden computational layers (Wei & El Naqa, 2020). The hidden layers module within a neural network may transform the data (e.g. an image) at one level, for example, one level may represent the edge of a tumour with an image in a particular orientation, the second layer may detect a particular pattern in the observed edges and a third may then recognise objects from different ensembles of patterns, linked form other layers (Wei & El Naqa, 2020). This approach removes the need for image segmentation i.e. the need to draw a region around the tumour and extract only tumour image information to feed into the algorithm. In contrast, the user can give the CNN a whole image or an image containing the tumour and the trained CNN can make its own decisions on which particular features are of importance. Allowing the CNN to make decisions on which features to prioritise removes the statistical bias associated with pre-defining an ML algorithm with a subset of features to focus on (Wei & El Naqa, 2020).

A sub-type of ANNs and a method particularly useful for image processing applications (such as this project) is using convolutional neural networks whereby each neuron in a layer is connected to only a subset of neurons in the subsequent layer. In a CNN, the first layer output is considered as an image; a convolution is then applied, followed by a mathematical operation and a “pooling of pixel data” (Uribe, et al., 2019). This method reduces the number of data points in the network and therefore reduces computational cost. Following the ‘image processing’ layers, there is another layer of neurons, which determines the output classification parameters. By this process, the CNN can directly process whole images or images with pre-defined regions / volumes of interest. The user therefore does not need to decide which radiomics parameters to extract as the CNN makes this decision (Uribe, et al., 2019) (Figure 26).

**Figure 26 Example methodology flow chart for radiomics, ML and DL**



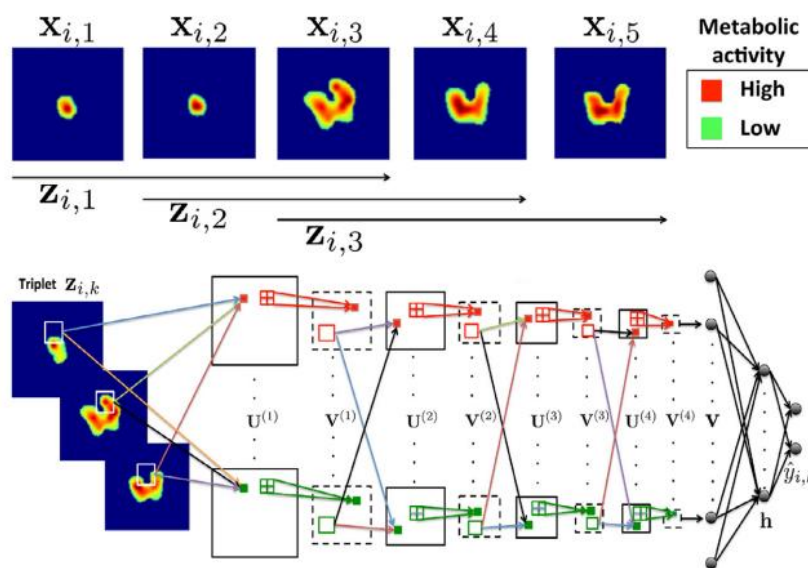
*Summary flow chart of radiomics, machine learning and deep learning methods to show the comparison of computational steps and image processing / feature extraction steps. Source: Li et al (2021).*

Litjens et al (2017) present an example of a 1-dimensional CNN (1D-CNN) and describe two key differences between an ANN and a CNN. First, the weights used in the network are linked in such a way that the network can perform convolution operations on images and the model does not need to learn from separate nodes for the same object occurring at different positions in the image. Second, the CNN contains “pooling” layers where “pixel values of neighbourhoods are aggregated using a permutation invariant function, typically the max or mean operation” (Litjens, et al., 2017); in other words, the pixels are grouped into similar values as a ‘pseudo-smoothing’ effect to reduce the variance between subsequent convolutional layers. After the convolutional section of the network, fully connected layers are added where the weighting is no longer shared across the nodes and, similar to a multi-layered perceptron; the classification process takes place by feeding forward ‘activations’ (results meeting the specified classification criteria) to a final layer.

As to the optimum architecture of a CNN, there are multiple different options, and the ideal solution is dependent on the specific application. Outside of medical imaging applications, several groups have proposed deeper networks operating on multiple sub-sets of images / data in order to reduce the memory burden on the CPU (e.g., in smart phone applications); this is however less of a concern for medical image applications and a “whole image” approach is favoured (Litjens, et al., 2017).

Ypsilantis et al (2015) have compared machine learning performance with a 3-slice CNN (3S-CNN) architecture, concluding that the 3S-CNN architecture is superior for predicting treatment response. Figure 27 shows how their 3-slice approach maps to various nodes in the neural network.

**Figure 27 Example deep learning system architecture**



*Illustrative figure of a proposed 3-slice deep learning system architecture. Source: Ypsilantis et al (2015)*



### 3.5 Systematic Review

Several groups have linked radiomics parameters to oesophageal cancer treatment response, for example, one of the earliest studies of this kind, Tixier et al (2011) showed in 41 patients that GLCM homogeneity and entropy, GLSZM variability and run length matrix (RLM) variability successfully separated responders, non-responders and partial responders with sensitivities of 72%-92% (Sah, et al., 2019). In contrast, Beukinga et al (2018) found in 97 patients that GLRL emphasis gave higher area under operator receiver curve (AUROC) when linked to therapy response prediction. Wu et al (2018) have used radiomics to predict the staging of pre-operative oesophageal squamous cell carcinoma for 154 patients, again using the PET/CT data acquired during staging; and successfully predicted tumour stage (stage I-II vs stage III-IV) more accurately using radiomics compared to conventional parameters alone. Nakajo et al (2017) showed in 52 patients that GLSZM variability and other standard parameters successfully predicted tumour response but not overall / progression-free survival. Foley et al (2018) describe a study with 403 patients linking “total lesion glycolysis, histogram energy and kurtosis” (Sah, et al., 2019) with overall survival. Here, overall survival is the total months survived post treatment. Paul et al (2017) have taken radiomics further and have used a genetic algorithm based on random forest, an artificial intelligence technique, to link GLCM homogeneity with treatment response for 65 patients (Sah, et al., 2019) concluding that the random forest approach performed better at predicting treatment response and prognosis compared to conventional methods.

Table 1 shows a summary of relevant papers specifically investigating the impact of radiomics in oesophageal cancer compiled from 3 separate review papers investigating the use of radiomics, machine learning and deep Learning as applied to oesophageal cancer. The papers collated are studies using PET or PET/CT imaging only (Van Rossum, et al., 2016; Sah, et al., 2019; Xie, et al., 2021). Most studies included a mixture of adenocarcinoma and squamous cell carcinomas and investigated treatment response. Complete pathologic response to treatment relates to overall disease-free and survival rate however, investigating radiomic signatures and their direct link to survival rates has been sparsely discussed in the literature (Van Rossum, et al., 2016). There is a mixture of studies investigating treatment response both to definitive chemo radiotherapy (CRT) treatments and neoadjuvant CRT followed by surgery. Groups investigating nCRT and Surgery have also performed analysis using both baseline and post nCRT PET/CT images (Tan, et al., 2013; Tan, et al., 2013; Zhang, et al., 2014; Van Rossum, et al., 2016; Yip, et al., 2016). Whereas other groups investigating only the pathologic response to dCRT used a baseline PET/CT image only (Tixier, et al., 2011; Beukinga, et al., 2018; Nakajo, et al., 2017; Hatt, et al., 2013).

Xiong et al (2018) extracted 440 radiomics features for a group of 30 patients with oesophageal cancer and used 4 different machine learning methods to predict local control of disease (2 years progression-

free) concluding that the random forest ML method achieved the best predictive results. Parmar et al (2015) extracted 440 radiomics parameters from CT data for 464 lung cancer patients and evaluated 12 different machine learning methods, concluding that a “classification method random forest...had the highest prognostic performance”.

Ypsilantis et al (2015) describe a study comparing two competing approaches to linking oesophageal cancer PET/CT images to treatment response to neoadjuvant chemotherapy. The group compare a “hand engineered” radiomics approach by extracting large numbers of parameters and linking them to response and they describe a biologically inspired approach using convolutional neural networks. In their radiomics approach, 85 textural parameters and 18 SUV statistical summaries gave an initial vector of 103 dimensions with the aim of minimising the misclassification rate using four different ML approaches: logistic regression, gradient boosting, random forests and support vector machines. Their deep learning architecture involved a “three-slice convolutional neural network” to produce features, which are “representative of metabolic activity in cancer” (Ypsilantis, et al., 2015). To create additional training and testing datasets, Ypsilantis et al (2015) “artificially augmented” their data by “rotating each triplet”. Ypsilantis et al (2015) describe in detail their exact network construction; of note is the pre-processing of the images to form a standard image input size by essentially normalising the size of the input image to the largest tumour in the data set; this normalisation allows images to be input of the exact same size for all patients. Furthermore, they have simplified their approach by specifying their input as multiple sets of three slices in which to perform mathematical operations on (Ypsilantis, et al., 2015). Ypsilantis et al (2015) conclude that the 3S-CNN method out performs all others in terms of superior sensitivity and specificity.

In contrast, Yang et al (2019) describe a more holistic 3D-CNN model using a “residual network” to predict overall survival at 1 year from a cohort of 548 PET scans using the whole PET image however, less detail is given regarding the specifics of their architecture, other than the number of layers in their network. Cao et al (2020) extracted 944 radiomics parameters per patient from pre-treatment PET/CT scans and used a “least absolute shrinkage and selection operator” (LASSO) machine learning algorithm to predict treatment response. Their cohort consisted of 159 oesophageal squamous cell carcinoma patients treated with CRT only.

**Table 1 Summary of studies investigating oesophageal cancer with PET or PET/CT Imaging (Van Rossum, et al., 2016; Sah, et al., 2019; Xie, et al., 2021)**

Reference, Nationality	N	Histology (AC / SCC / Other) and Stage	Approach (Rad, ML, DL)	Treatment	Image Timing	Details of Model	End point and Reference Standard	Key Findings	Software
(Tixier, et al., 2011). France	41	10/31/0 I-IV	Rad	dCRT, 60 Gy with Cis, Car or Flu.	Baseline	7 intensity and 31 texture features 4 quantization levels	Response prediction, measured using CT and Endoscopy, RECIST Criteria. AUROC	Clinical response (based on CT; RECIST: CR vs. PR vs. non-R). Tumour GLCM homogeneity, GLCM entropy, RLM intensity variability and GLSZM size zone variability can differentiate non, partial and complete responders with higher sensitivity (76%–92%) than any SUV measurement	In-House
(Beukinga, et al., 2018), Netherlands	73	65/8/0 I-IV	Rad ML	CRT: 41.4 Gy with Car / Pac And Surgery	Baseline	103 features including: First order statistics Geometry GLCM NGTDM ML using Logistic Regression	Response prediction based on Histology and pre/post SUVmax and TNM scoring	18F-FDG long run low grey level emphasis higher in responders than non-responders. Model including histologic type, clinical T stage, 18F-FDG long run low grey level emphasis and non-contrast CT run percentage has higher AUC than SUVmax: 0.74 vs. 0.54	MatLab (In-house)
(Nakajo, et al., 2017), Japan	52	0/52/0 II-IV	Rad	CRT: 41–70 Gy with Cis / Flu	Baseline	GLCM: Entropy, homogeneity, dissimilarity; GLSZM: Intensity variability, Size-zone variability, zone percentage	Response using CT / Endoscopy RECIST Criteria and Progression free / overall survival prediction, AUROC	GLSZM intensity variability and GLSZM size-zone variability predictive of response. No texture parameter independently associated with progression free or overall survival	Python PyRadiomics

Reference, Nationality	N	Histology (AC / SCC / Other) and Stage	Approach (Rad, ML, DL)	Treatment	Image Timing	Details of Model	End point and Reference Standard	Key Findings	Software
(Hatt, et al., 2013), France	50	14/36/0 I-IV	Rad	dCRT	Baseline	10 texture features 3 segmentation methods With and without PVC	Response prediction, using CT and Endoscopy, RECIST Criteria.	Clinical response (based on CT; RECIST: CR + PR vs. non-R), Entropy and homogeneity show high differentiation between CR and Non-R	MedCalc Software
(Tan, et al., 2013), USA	20	17/3/0 II-III	Rad	nCRT 50.4 Gy with Cis / Flu + Surgery	Baseline + after nCRT	34 intensity, texture, and geometry features	Response prediction determined by histology. AUROC	Pathologic response (TRG* 1-2 vs. 3-5) SUVmean decline, SUV skewness, GLCM inertia, GLCM correlation, and GLCM cluster prominence predict complete response, AUC 0.76–0.85	Insight Segmentation and Registration Toolkit
(Tan, et al., 2013), USA	20	NR NR	Rad	nCRT + Surgery	Baseline + after nCRT	SUVmax, SUVpeak, TLG, 8 texture features, and 19 histogram distances	Response Prediction determined by histology. AUROC	Pathologic response (TRG* 1-2 vs. 3-5). Histogram distances provide useful prediction information.	Insight Segmentation and Registration Toolkit
(Zhang, et al., 2014), USA	20	17/3/0 II-III	Rad, ML	nCRT + Surgery	Baseline + after nCRT	9 intensity, 8 texture, and 15 geometry features, TLG, and 16 clinical features	Response Prediction determined by histology.	Pathologic response (TRG* 1-2 vs. 3-5) “SVM model using all features including spatial-temporal PET features accurately and precisely predicted pathologic tumour response to CRT.”	In House
(Ypsilantis, et al., 2015), UK	107	86/20/1 II-IV	Rad, ML and DL	Chemo only (nChTx)	Baseline	> 100 texture features vs. (3S-CNN) trained directly from scans	Response Prediction by Mandard tumour regression and OS	Pathologic response (TRG* 1-3 vs. 4-5) NGTDM coarseness, SUVminimum, GLRL Long Run L. Grey-Level Emphasis are top 3 most important parameters.	PyRadiomics

Reference, Nationality	N	Histology (AC / SCC / Other) and Stage	Approach (Rad, ML, DL)	Treatment	Image Timing	Details of Model	End point and Reference Standard	Key Findings	Software
(Van Rossum, et al., 2016), USA	217	217/0/0 II-III	Rad, ML	nCRT (36% ChTx before nCRT) CRT: 45 or 50.4 Gy with Flu + Surgery	Baseline + after nCRT	69 texture and 12 geometry features 2 baseline scans at different institutions	Response prediction with tumour regression grade, and OS Multivariable Cox analysis	Pathologic Response (TRG* 1-3 vs. 4-5), feature selection by uni-variable logistic Regression Model including induction chemotherapy. Change in RLM run percentage, change in GLCM entropy, and post – CRT roundness is better than clinical prediction model.	MatLab, In House
(Yip, et al., 2016), USA	45	44/1/0 I-IV	Rad	nCRT 45–50.4 Gy with Cis / Flu / Car / Pac + Surgery	Baseline + after nCRT	GLCM: homogeneity, entropy RLM: high grey run emphasis, short-run high grey run emphasis GLSZM: high grey zone emphasis, short-zone high grey emphasis	Response prediction, defined by histology AUROC	Response prediction: Change in run length and size zone matrix parameters differentiates non-responders from partial/complete responders (AUC: 0.71–0.76)	MatLab-based Chang-Gung Image Texture Analysis Toolbox
(Paul, et al., 2017), France	65	8/57/0 II-III	Rad, ML	CRT: 50 Gy with Car / Flu	Baseline	84 features including, Random forest classifier	Response prediction with endoscopic biopsy, AUROC	Best subset of predictive variables: metabolic tumour volume, GLCM homogeneity	(Lambin, et al., 2012)

Reference, Nationality	N	Histology (AC / SCC / Other) and Stage	Approach (Rad, ML, DL)	Treatment	Image Timing	Details of Model	End point and Reference Standard	Key Findings	Software
(Foley, et al., 2018), UK	403	237+79/65+22/0 II-III	Rad	Mixture: Surgery, NACT, NACRT, dCRT	Baseline	First order GLCM: homogeneity, entropy, dissimilarity; coarseness;	Overall Survival, Prognostic score	TLG, histogram energy and histogram kurtosis are independently associated with overall survival	ATLAAS (In house, MatLab)
(Beukinga, et al., 2017), Netherlands	70	65/8/0 II-III	Rad	CRT: 41.4 Gy in 23 fractions with Car / Pac and Surgery	Baseline + After	113 features, 6 different prediction models, no ML	Pathologic Response prediction (Mandard Tumour scoring)	Prediction models composed of clinical T-stage and post-NCRT joint maximum adds important information to the visual PET/CT evaluation of a pathologic complete response	MatLab
(Xiong, et al., 2018), China	30	0/30/0 I-IV	Rad, ML	CRT only	Baseline and Mid treatment	440 radiomic Parameters, 4 ML methods (RF, SVM, LR and ELM)	Response Prediction using local control rate AUROC	Random Forest method was best predictor of outcome	NR
(Cao, et al., 2020), China	159	0/159/0 IIA-IV	Rad, ML	CRT	Baseline and follow up	944 radiomic Parameters, LASSO ML Algorithm	Response Prediction, Follow up PET/CT imaging RECIST criteria	LASSO logistic regression model successfully predicted high and low risk patient groups.	Slicer Radiomics (PyRadiomics), Python

Reference, Nationality	N	Histology (AC / SCC / Other) and Stage	Approach (Rad, ML, DL)	Treatment	Image Timing	Details of Model	End point and Reference Standard	Key Findings	Software
(Yang, et al., 2019), China	548	0/548/0	DL	Multiple treatment regimens	Baseline	3D CNN based on Residual Network, model expanded using lung and oes. cancer patients	1 Year Survival, 5 Year Survival,	3D-CNN model can be trained to predict more aggressive tumours	NR
(Desbordes, et al., 2017), France	65	8/57/0	Rad, ML	CRT, 14 underwent surgery also	Baseline	61 features from medical records including 45 radiomics parameters, RF algorithm,	Pathologic Response assessment	MTV, WHO Status and nutritional risk were predictive of treatment response	MaTLAB
(Chen, et al., 2019), Taiwan	44	0/44/0	Rad, ML	nCRT followed by Surgery	Baseline	Logistic Regression ML model	Overall and DFS	Risk stratification for DFS and OS	SSPS Software

Key: 18F-FDG = 18F-fluorodeoxyglucose; 3S-CNN = three-slices convolutional neural network; AC = adenocarcinoma; Car = Carboplatin chemotherapy; Cis = Cisplatin Chemotherapy; CR = complete response; CT = computed tomography; CRT = chemo-radiotherapy; dCRT = Definitive Chemo-Radiotherapy; Flu = Flurouracil; MTV = metabolic tumour volume; nChTx, neoadjuvant chemotherapy; nCRT = neoadjuvant chemo-radiotherapy; non-R = non-response; NR = not reported; Pac = Paclitaxel Chemotherapy; RECIST = Response Evaluation Criteria in Solid Tumours; PET, positron emission tomography; PR, partial response; SCC, squamous cell carcinoma; SD, standard deviation; SRHIE = short-run high-intensity emphasis; SUV = standardized uptake value; SZHIE = short-zone high-intensity emphasis; TLG = total lesion glycolysis; TRG = tumour regression grade, Rad = radiomics, ML = machine learning, DL = deep learning, OS = Overall Survival

This systematic review, at the time of writing, describes 18 studies where PET imaging, oesophageal cancer and radiomics, machine learning or deep learning are featured key words. Patient numbers in studies have varied between 20 (Tan, et al., 2013) and 548 (Yang, et al., 2019). Studies have primarily been of either squamous cell carcinoma (SCC) patients (5 studies) or a mixture of adenocarcinoma and SCC (12 studies) with only one study (Van Rossum, et al., 2016) describing an exclusively adenocarcinoma group. Patient treatments varied widely across these studies with the majority (11 studies) performing chemo-radiotherapy (CRT) in addition to surgery, 6 studies performing CRT and 1 study performing chemotherapy only (Ypsilantis, et al., 2015). In total, 15 studies used pathologic response as the end-point for describing a ‘responder’ with response defined either by histological sampling (Yip, et al., 2016; Tan, et al., 2013) or the RECIST criteria (Eisenhauer, et al., 2009) (Cao, et al., 2020; Nakajo, et al., 2017). Five studies reported using overall survival or DFS, for 1- and 5-year periods with two studies also reporting using pathologic response (Nakajo, et al., 2017; Van Rossum, et al., 2016). **Clinically, there is sparse literature using groups of exclusively adenocarcinoma patients and using disease-free or overall survival to define the ‘successful treatment’ of oesophageal cancer. Furthermore, there are no studies using 2-year disease-free survival to define a successful treatment, as is local practice in the North East oesophago-gastric unit.**

The majority (17) of studies have investigated the efficacy of a radiomics signature from the pre-treatment PET imaging of oesophageal cancer patients. Of those 17 studies, eight studies have explored the additional application of using machine learning techniques with a variety of algorithms to predict outcomes (whether overall survival or treatment response). One study has explored both radiomics, machine learning and deep learning approaches (Ypsilantis, et al., 2015) and one study has used an exclusively deep learning approach (Yang, et al., 2019). Overall, most studies have incorporated “classical” findings into analysis with texture parameters however, only two groups reported an association between texture parameters and overall survival. Furthermore, there is only one study, which used a combination of radiomics and machine learning approaches to predict overall survival however, this group have focussed more on the comparison of baseline and post-treatment scans. **There are currently no studies using radiomics, clinical parameters and machine learning to predict 2-year survival of oesophageal cancer and oesophago-gastric junction adenocarcinoma patients.**

For radiomics studies: various first, second and high-order features have successfully assessed treatment response and differentiated between responders and non-responders. In general, there is greater tumour heterogeneity in non-responders and outcome prediction has been more accurate



than conventional parameters alone, which agrees for similar studies using CT imaging (Sah, et al., 2019). Whilst there is no single parameter which emerges as a definitive 'gold standard' for separating responders with non-responders, GLCM entropy (describing local randomness and irregularity) has been the most reported feature of interest with a high tumour entropy describing heterogeneity and a low entropy describing homogeneity. Analysis of parameters such as local entropy derived from GLCMs for tumour heterogeneity characterisation, has been reported as the most robust and repeatedly showing a strong correlation with the prediction of response, tumour stage and survival (Van Rossum, et al., 2016). **Several studies have used radiomics and machine learning approaches to predict treatment responses and overall survival however there is only one study comparing the performance of each of these approaches** (Ypsilantis, et al., 2015).

### 3.6 **Research Proposal**

From the systematic review and review of the literature in relation to this project, the following gaps or areas of sparse publication identified are:

- Using radiomics to predict disease-free survival (DFS) in oesophageal or oesophago-gastric junction adenocarcinoma patients, treated primarily with surgery: 1 study to date (Van Rossum, et al., 2016)
- Using the above methodology with 2-year DFS as the end point: 1 study to date (Xiong, et al., 2018)
- Comparison of radiomics, and machine learning performance in predicting outcomes (treatment response or overall survival) using PET imaging in the context of oesophageal cancer: one study to date (Ypsilantis, et al., 2015)
- The effect of a 'Block Sequential Regularized Expectation Maximization' (BSREM) algorithm on radiomics parameters in clinical images: two similar, but not comprehensive studies to date (Reynes-Llompart, et al., 2018; Ger, et al., 2019).
- Comparison of machine learning algorithm performance on the radiomic signature from OSEM and BSREM PET images: No studies to date.

We propose a retrospective analysis of a cohort of oesophageal and oesophago-gastric junction cancer patients from the northern centre for cancer care (NCCC) with adenocarcinoma. All patients included in this study will have had a 3D PET/CT scan for staging, pre-treatment and at least 2 years of follow up to determine whether they have remained disease-free after treatment. Patients will be included who have been treated definitively (curative) with surgery and a combination of chemotherapy or chemotherapy and radiotherapy.

Our primary proposal is to investigate whether a radiomics signature, acquired from a region of interest drawn around the primary oesophageal / oesophago-gastric junction tumour, can predict 2-year DFS. We propose investigating this using a variety of 'hand-crafted' machine learning algorithms. Our secondary aim is to compare radiomics and machine learning performance in relation to PET image reconstruction with 'Ordered Subsets Expectation Maximum' (OSEM) and Block Sequential Regularized Expectation Maximization (BSREM, "Q.Clear") algorithms; we aim to determine the effect of these two reconstruction methods on raw radiomic parameters and machine learning algorithm predictive performance.

## 4 Methodology

---

This chapter outlines the methodology used for this project. This chapter includes an overview of the ethics and approvals process, a detailed overview of the methodology and details of patient selection for this project. This chapter includes a review of available radiomics and artificial intelligence (AI) software, a description of how the images were analysed and radiomic signatures downloaded (including definitions). Finally, this section includes a description of the machine learning algorithms used, key definitions from the code and a summary of the different ML experiments performed.

### 4.1 Ethics and Health Research Authority Approval

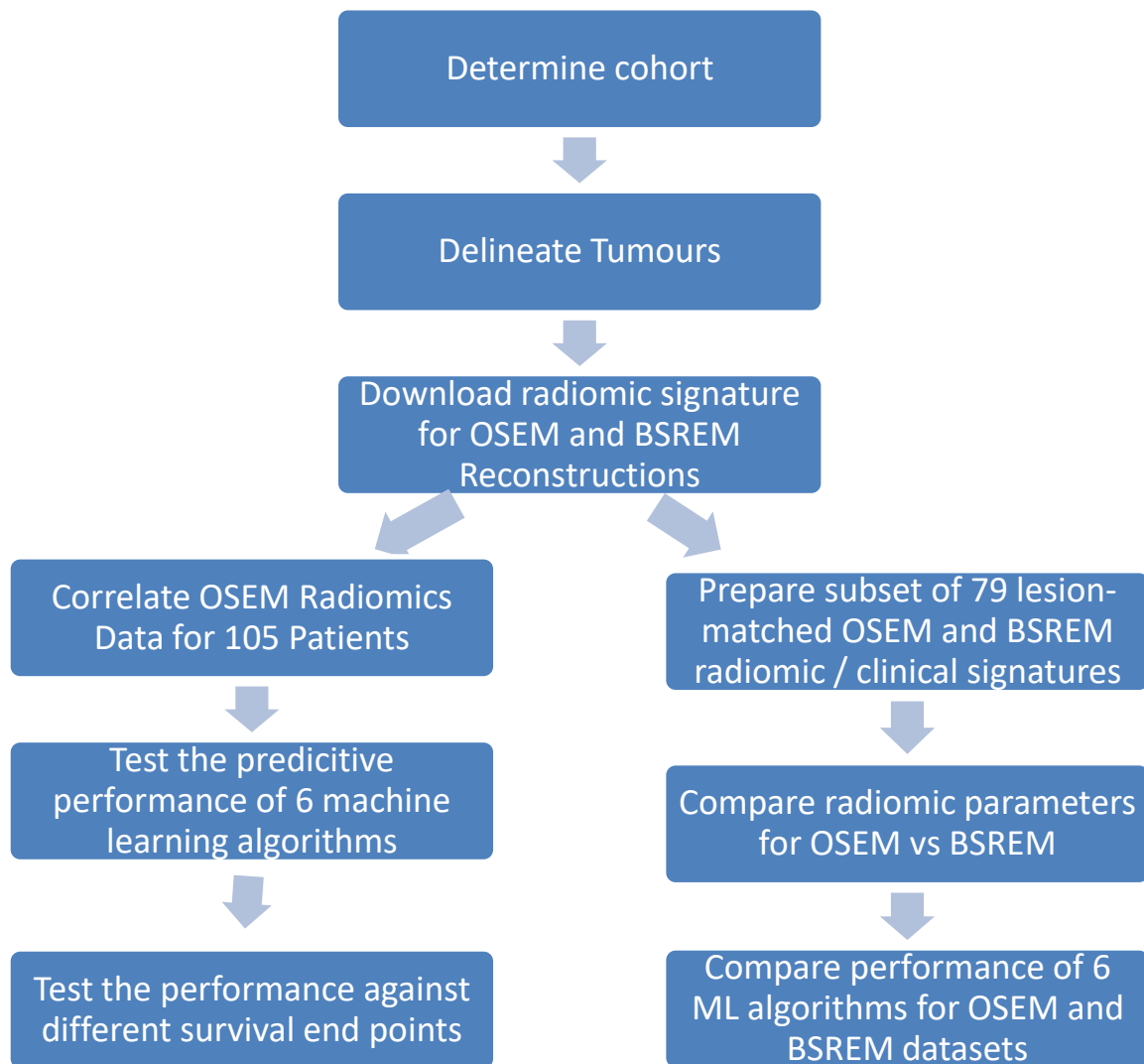
We submitted a “Project Initiation Form” to the Newcastle University and Newcastle Upon Tyne Hospitals NHS Foundation Trust (NuTH) research office on 29<sup>th</sup> November 2019. An honorary contract with NuTH was received on 10<sup>th</sup> January 2020. Sponsorship with NuTH was applied for on 24<sup>th</sup> June 2020 via the Integrated Research Application System (IRAS) under ID: 277971.

This study received favourable ethical approval from the East of Scotland Research Ethics Service (REC) on 9<sup>th</sup> November 2020 (Ref: 20/ES/0115). This study then received approval from the Health Research Authority (HRA), and the Health and Care Research Wales (HCRW) on 9<sup>th</sup> November 2020 (Ref: 20/ES/0115). The Newcastle Upon Tyne Hospitals are sponsoring this study with confirmation of capacity received on 17<sup>th</sup> February 2021.

## 4.2 **Overview of Methodology**

This section provides an overview of the methodology used to complete this project.

1. Determine the patient cohort and organise the patient data into a normalised data set to build a coherent database of anonymised patient information. Designate all patients as survived > 2-year post treatment or deceased at < 2 year post treatment.
2. Load each patient into the LiFEX v7.0.0 program (Nioche, et al., 2018). Delineate the tumours using an SUV threshold method to best match the edge of the tumour in conjunction with review of clinical and CT data and support from an experienced Nuclear Medicine Radiologist.
3. Use LiFEX v7.0.0 to output a set of radiomics parameters into a .csv file. Parameters based on:
  - a. The Raw pixel values
  - b. An image discretized to 64 bins of 0.3125 SUV value in size (Leijenaar, et al., 2015).
4. Use MS Excel to correlate the radiomics signature with the TNM score and 2-year disease-free survival (DFS) status, designating DFS at 2 years as “Success” and recurrence or death within 2 years as “Failure”.
5. Correlate the radiomic signature and clinical data with DFS time, TNM scoring, treatment and cancer site.
6. Compare the radiomic signature for both OSEM reconstructed and BSREM (“Q.Clear”) reconstructed datasets, using identical tumour regions.
7. Split the data for each machine learning algorithm into 80% of patients for training and evaluation, 20% held back as a validation dataset.
8. Using the Spyder v5.1.5 program (within Anaconda3) in conjunction with the “Keras” library, investigate the performance of 6 different machine learning algorithms: Logistic Regression, Linear Discrimination Analysis, K Neighbours Classifier, Decision Tree Classifier (“Random Forest”), Gaussian Naïve Bayes, Support Vector Machines.
9. Estimate the model accuracy using a “10-fold cross validation estimation model” (Brownlee, 2019).
10. Compare the accuracy of the 6 models for both OSEM reconstructed and BSREM (“Q.Clear”) reconstructed datasets, using identical tumour regions.

**Figure 28 Flow chart overview of project methodology**

*Flowchart to show an overview of the proposed experimental methods.*

### 4.3 Patient Selection

In this section, I will summarise the patient selection process from the Northern Upper Gastric Unit database, image acquisition and determination of the final patient cohort for this study.

#### 4.3.1 Patient Cohort

This study recruited patients, retrospectively, from the Northern Upper Gastric Unit (NOGU). The NOGU database included patients from a large geographical area (Cumbria, Northumberland, Tyne and Wear and Middlesbrough) presenting with cancer of the upper GI tract: oesophageal, oesophago-gastric junction (OGJ) and gastric cancers (see Appendix 9.4). The key data fields extracted from the database for this study were:

**Table 2 Key fields extracted from NOGU Database:**

Diagnostic Site Type	Diagnostic Histology	PET Scan (Y/N)
Overall T Stage - staging	Overall N Stage - staging	Overall M Stage - staging
Date of 1st Treatment	Treatment Received	Operation
Date of Recurrence	Date of Death	Cause of Death

*Table to show an overview of all of the individual data fields extracted from the NOGU database.*

Diagnostic site type was used to define the origin of the cancer (oesophageal, OGJ or gastric) and diagnostic histology to determine the disease type (adenocarcinoma, squamous cell carcinoma, other). The overall TNM score was included however, the M (metastases) score was not included as patients were only offered definitive, curative surgical treatment if their disease had not metastasised, therefore all patients in our final group had an “M Stage” of 0. We recorded the N-score numerically as 0-3 and the T score as 1-4. T4 graded tumours fall into subcategories T4a and T4b and were recorded as T4 only since the majority of such tumours were excluded from our study because the majority of these tumours were palliative patients, accompanied with an M score of 1.

The DFS time, used as the end point in this study, was defined as the number of days between the date of the 1<sup>st</sup> treatment (either the date of the first neo-adjuvant chemotherapy or the surgical date, whichever was first) and either the date of recurrence or date of death (if the cause of death was related to their cancer). The majority of patients in our cohort received Neo-adjuvant chemotherapy 3 months prior to their surgery date clinicians consider the “start” of treatment. We used disease recurrence (rather than death) as this often occurred prior to death and clinicians consider disease

recurrence within two years of treatment as a 'failure in management'. The average time from diagnosis to first treatment in our cohort was 61 days and the average time from first treatment to surgery was 86 days. We excluded patients if the cause of death was unrelated to their treatment. We assumed patients to have survived for at least 2 years when there was no entry for date of recurrence or date of death. DFS time was normalised to a maximum of 2 years for all patients to account for differing follow up periods at the time of recruitment.

The initial search included all patients seen from 1/1/2016 – 31/12/2018 such that, at the time of recruitment into the study, 17/2/2021 (Final Approval, see section 4.1), each patient had a minimum of 2 years of follow up. The initial database search included 778 patients within the specified period who had also had a PET/CT scan at the time of diagnosis.

**Table 3 Summary of cancer location and type in cohort**

	<b>Number of Patients (adenocarcinoma / SCC / other)</b>
<b>Gastric</b>	160 (135 / 1 / 24)
<b>OGJ</b>	149 (142 / 5 / 2)
<b>Oesophageal</b>	438 (249 / 154 / 35)
<b>Other / Unspecified</b>	31 (6 / 2 / 23)
<b>Total</b>	778 (532 / 162 / 84)

*Table to show an overview of the numbers of adenocarcinoma, squamous cell carcinoma (SCC) and other cancer subtypes for each group of upper gastrointestinal tract cancer patients: Gastric, Oesophago-gastric junction, Oesophageal and other sub-types.*

This study focussed on oesophageal and OGJ patients with adenocarcinoma, owing to the relatively similar aetiology of oesophageal and OGJ cancers and the same underlying disease process of adenocarcinoma. Moreover, this represented a larger cohort of patients for the study and gave 391 patients.

For an initial pilot study to test and develop the machine learning algorithm, we used a separate sub group of 249 oesophageal adenocarcinoma patients who received either curative or palliative treatment with 11 patients excluded due to incomplete data records, leaving a pilot sub group of 238 patients, 160 Curative and 78 Palliative.

For the final patient cohort, the initial group was filtered to include only oesophageal / OGJ adenocarcinoma patients who had received surgery with curative intent; 144 patients with the following cancers and treatment pathways:

**Table 4 Summary of cancer location and cancer treatment**

Location	No. Patients	Surgery Alone	Neo-Adjuvant Chemotherapy	Adjuvant Chemotherapy	Radiotherapy
Oesophageal	102	21	76	32	12
OGJ	42	5	34	14	1

*Table to show an overview of the numbers of patients in our final cohort of surgically treated, adenocarcinoma patients with the split of different treatments received.*

The final cohort was of patients from the North East of England, with adenocarcinoma of the oesophagus or OGJ, treated curatively with a combination of surgery alone or with chemo and / or radiotherapy. The average age in our cohort was 65.9(47-85) years, 124/19 male/female, with an average weight of 83.7 (54-142) Kg and average height of 1.73 (1.50 – 1.93) m.

#### **4.3.2 Image Acquisition and Transfer**

The NOGU Multi-Disciplinary Team (MDT) meeting in Newcastle reviewed the PET/CT images before transfer for storage on the relevant local Picture Archiving and Communication System (PACS). We contacted nine hospital PACS systems for image transfer requests to the Newcastle Hospitals PACS (Phillips, 2020); 134/144 patient images were located and successfully transferred. Images were transferred from the Newcastle PACS system to the NuTH HERMES Medical Solutions GOLD Browser for anonymization. Upon loading into the HERMES system, all relevant DICOM headers were reviewed and a further 29 patients were excluded because only a 2D dataset was available; these scans were unfortunately acquired on the previous GE Discovery ST scanner which was replaced with a GE Discovery 710 PET/CT scanner in mid-2016. We excluded patients with 2D data because the proposed radiomic methods were only compatible in 3D. 105 patient PET studies were completely anonymised using the HERMES anonymization tool and assigned a unique numerical tag, which linked back to the original patient data via a spreadsheet stored on the secure trust servers. For each patient in the final dataset, the DICOM headers were reviewed to ensure and double check full anonymization of the data has been achieved. We transferred pseudo-anonymised patient PET scans from HERMES to “LiFEX v7.0.0” (Nioche, et al., 2018) for tumour segmentation and radiomic signature analysis.



PET images were acquired by Alliance Medical using images acquired in Newcastle on the GE Discovery 710 PET/CT scanner (GE, 2022). Patients were administered a nominal 400MBq  $^{18}\text{F}$ -Fluoro-Deoxy-Glucose ( $^{18}\text{F}$ -FDG) in accordance with local practice. Ideally, the injected dose would be scaled to the patient weight as variations in the injected dose can affect the radiomic values (Cook, et al., 2018), however, since this was a retrospective analysis, injected dose was performed as per local practice. To estimate the possible implication of using a standard dose, an assessment of the liver noise was made using a 3cm diameter volume of interest according to RECIST criteria (Eisenhauer, et al., 2009). The liver signal-to-noise ratio was then calculated using:

$$\text{SNR}_{\text{liver}} = \text{SUV}_{\text{mean}} / \text{SD}_{\text{liver}} \quad \text{Equation 10}$$

Where  $\text{SNR}_{\text{liver}}$  is the signal-to-noise ratio with  $\text{SUV}_{\text{mean}}$  as the mean SUV in a 3cm region drawn in the lateral lobe of the liver and  $\text{SD}_{\text{liver}}$  as the standard deviation of SUV in that region (Yan, et al., 2016).

Patients received an ‘eyes to thighs’ PET/CT scan with a PET slice thickness of 3.27mm and a matrix size of 256 x 256. PET Images were reconstructed using 3D iterative Ordered Subsets Expectation Maximum (OSEM) algorithm with 2 iterations, 24 subsets, and a 6.4 mm width Gaussian filter. For 79 of those patients, data was also available reconstructed using GE’s “Q.Clear”, Block Sequential Regularized Expectation Maximization (BSREM) reconstruction algorithm including PSF and TOF corrections. We used a  $\beta$  value of 400, as recommended for whole body oncology studies (Teoh, et al., 2015) and optimised locally (GE, 2022; Ross, 2014).

## 4.4 Software Review

### 4.4.1 Radiomics Software

There are several programs currently available and used by researchers and clinicians. The below table by Fornacon-wood et al (2020), has been adapted to indicate suitability for this project.

**Table 5 Summary of the available radiomics programs**

Software	Cited	IBSI	Open Source?	Notes	Ref
MaZda	366	N	N	No IBSI	(Szczypiński, et al., 2009)
Chang-Gung Image Texture Analysis (CGITA)	65	N	Y	No IBSI	(Fang, et al., 2014)
IBEX	134	N	Y	No IBSI	(Zhang, et al., 2015)
Moddicom	13	N	Y	No IBSI	(Dinapoli, et al., 2015)
PyRadiomics	324	Y	Y	IBSI and High Publications	(Van Griethuysen, et al., 2017)
LiFEX v7.0.0	84	Y	Y	IBSI and used in several PET studies	(Nioche, et al., 2018)
Quantitative Image Feature Engine (QIFE)	13	N	Y	No IBSI	(Echegaray, et al., 2018)
CERR	25	Y	Y	Low Publication, MATLAB platform	(Apte, et al., 2018)
MITK Phenotyping	6	Y	Y	Low Publication	(Götz, et al., 2019)
RaCat	4	Y	Y	Low Publication	(Pfaehler, et al., 2019)
PORTS v1.1 MATLAB software	Not Published	N	Y	No IBSI	Not published
MATLAB Package	Not Published	Y	Y	No Publications	Not published

*Summary of the different programs available / discussed in publications for the extraction of radiomics data.*

*Adapted from: Fornacon-Wood et al al (2020).*

Radiomics programs can, for reliability and harmonisation across platforms, be validated using the “Image Biomarker Standardisation Initiative” (IBSI) compliance hence, all the described programs have been assessed for IBSI approval (Zwanenburg, et al., 2019). The most popular platform among oesophageal PET studies was MATLAB using a variety of ‘in-house’ programmed approaches.

However, as per Fornacon-Wood et al (2020), MATLAB approaches lack approval from the IBSI. Whilst one of the key issues for radiomics and Artificial Intelligence is the lack of standardisation, our intention is to develop an approach in line with the standards, which are already available such as IBSI. For PET specific studies, MATLAB and LiFEX appear to emerge as the preferred platforms however; Python and PyRadiomics have emerged favourite among more general published works in imaging (Fornacon-Wood, et al., 2020). Furthermore, several other groups using a variety of different tumours have validated the radiomic results produced from PET image analysis in LiFEX 7.0.0 (Nioche, et al., 2018). We used LiFEX v7.0.0 for this project, given the relative user accessibility, the ease of extracting data, compatibility with PET specific data sets and external validation of the results produced.

#### 4.4.2 AI software

There are limited options for machine and deep learning programs however, those available offer powerful, user friendly solutions to writing machine learning algorithms and deep learning neural networks with Python computing language being the most widely used amongst the scientific community and for the analysis of medical imaging. At the time of writing, the below packages are the most widely used in the application to medical imaging (Litjens, et al., 2017; Uribe, et al., 2019):

**Table 6 A summary of Artificial Intelligence Software platforms**

Program	Description	Reference
Caffe	C++ and Python Interfaces, developed by UC Berkeley	(Jia, et al., 2014)
Tensorflow	C++ and Python Interfaces, developed by Google	(Abadi, et al., 2016)
Theano	Python Interface, developed by MILA Lab	(Bastien, et al., 2012)
Torch	Lua interface, used by Facebook AI Research	(Collobert, et al., 2011)

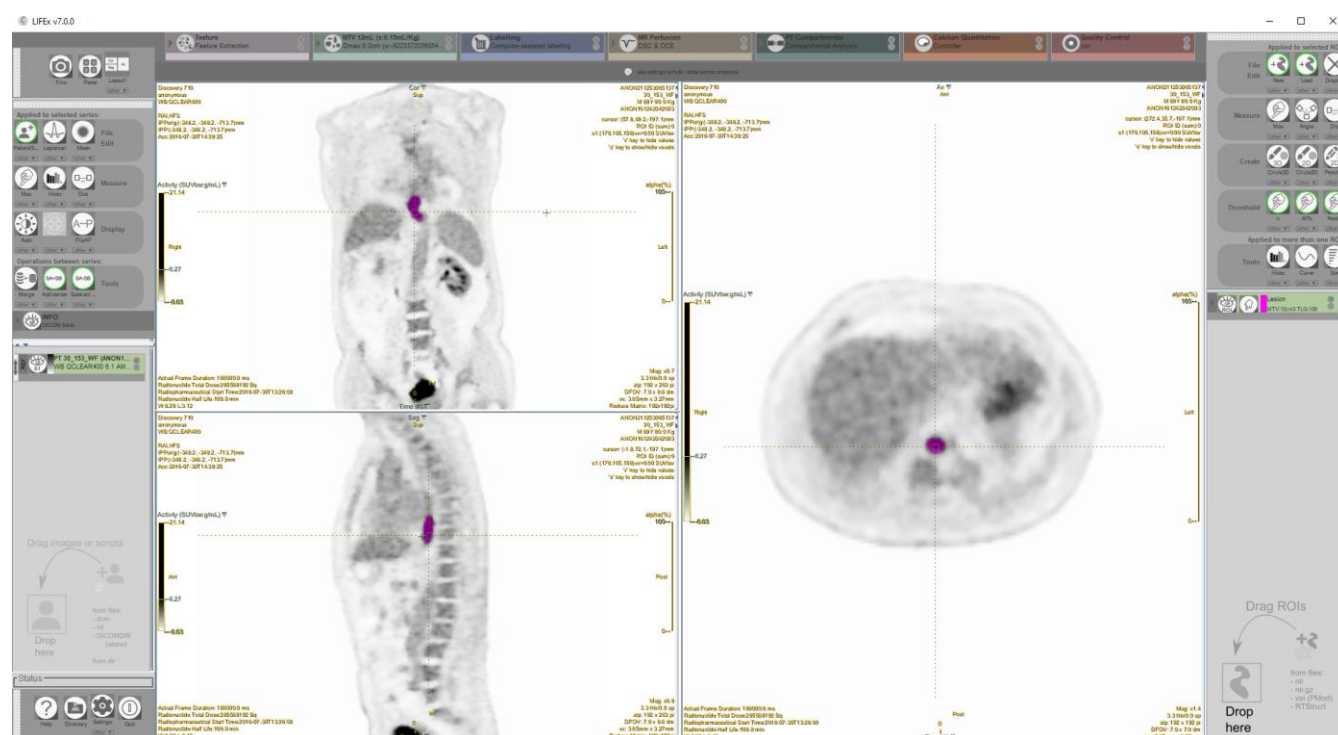
*Summary of the different programs available / discussed in publications for the formulation of artificial intelligence programs for medical imaging. Adapted from: Litjens et al (2017).*

The “Anaconda” Software was chosen because it is a free to download and includes several user-friendly platforms for loading machine / deep learning libraries and writing code using Python computer coding language (Brownlee, 2019).

## 4.5 Image Analysis and Tumour Threshold Method

PET Images were loaded into the LiFEX Program, version 7.0.0 (Nioche, et al., 2018) and a region of interest drawn around the primary tumour. Due to a lack of standardised approach to tumour segmentation, the methodology chosen for this study has replicated several other similar published works (Ypsilantis, et al., 2015; Van Rossum, et al., 2016). Regions of interest around the primary tumour were grown to an SUV threshold of 2.5-5 with the specific threshold chosen on a patient-by-patient basis to cover the maximum amount of tumour without including any normal physiological uptake (Appendix 9.4). All regions were reviewed by an experienced Nuclear Medicine Radiologist alongside the CT imaging and MDT information with changes made to the threshold as required, e.g. if using an SUV of 2.5 included any physiological uptake, the threshold was increased incrementally until the grown region included tumour only. All 105 tumour regions were approved by the experienced Nuclear Medicine Radiologist as providing adequate coverage of the primary tumour without including normal physiological uptake.

**Figure 29 Region thresholding in LiFEX**



*LiFEX v7.0.0 program with example region drawn around primary oesophageal cancer tumour (Nioche, et al., 2018)*

## 4.6 Radiomics Analysis

Radiomics data was downloaded for all 105 patients using data reconstructed using OSEM only and a further 79 patients with BSREM Data. Table 7 shows a summary, with definitions, of the radiomics parameters used in this project. To extract the radiomics texture parameters, a code written in Microsoft Notepad was loaded into the LiFEX v7.0.0 program (Nioche, et al., 2018) to download and export all features available into a spreadsheet (9.5). Texture results were discretized into 64 bins of 0.3125 SUV size according to the consensus approach described by Cook et al (2018) and as recommended in other similar published works (Leijenaar, et al., 2015). In each case, the parameters are in reference to the Volume of interest (VOI) drawn as described in section 4.5. The radiomic signature consisted of 9 Conventional, 8 Conventional (Discretized Image), 4 Discretised Histogram, 5 Shape and Volume, 7 Grey-level co-occurrence matrix, 11 Grey-level run-length matrix, 3 Neighbourhood Grey-level difference matrix, and 11 Grey-level size zone matrix features; 58 image features in total.

### 4.6.1 OSEM only radiomics data

Several comparisons were produced from the downloaded radiomics signature. A correlation “heat map” was produced to show how each feature correlated with each feature. Further correlation maps were created for each parameter against the 90 day, 1 and 2 year DFS to give an indication of which radiomic feature(s) best correlated with survival. For each survival endpoint, the average, minimum, maximum, standard deviation and variance was determined for the whole dataset and then the average and percentage difference determined for successful and failed treatments at each time point respectively.

### 4.6.2 OSEM and BSREM radiomics data

Radiomics data for the OSEM and BSREM cohort were compared directly using the raw values and values normalised to between 0 and 1. In this context, OSEM was considered to be the reference value. The percentage difference between the radiomic features determined using OSEM and BSREM image data was calculated using:

$$(\text{BSREM Value} - \text{OSEM Value}) / \text{OSEM Value} \quad \text{Equation 11}$$

The difference was determined for the raw and normalised datasets. The average, minimum, maximum, standard deviation and variance was determined for the difference percentages to give

an indication of which radiomic features were most affected by a change in reconstruction and which remained robust to reconstruction algorithm and remained relatively unchanged.

**Table 7 Summary of radiomics parameters and definitions**

	Group	Feature	Definition
Conventional and First Order Statistics	Conventional Histogram (9 Features)	SUVmin	Minimum value
		SUVmean	Average value
		SUV Stdev	Standard Deviation of values
		SUVmax	Maximum value
		SUV Skewness	Asymmetry of the grey-level distribution around the mean value
		SUV Kurtosis	Shape of the grey-level distribution (peaked or flat) relative to a normal distribution
		SUVpeak (0.5ml)	SUVmean in a sphere of 0.5ml volume, located so that the average value in the VOI is maximum
		SUVpeak (1ml)	As above for a 1ml sphere
		TLG	Product of SUVmean and Volume in ml
	Discretized Image statistics (8 Features)	Disc SUVmean	Definitions as above but applied to a discretized image. SUV values placed into 64 bins, each of 0.3125 SUV in size.
		Disc SUV Stdev	
		Disc SUVmax	
		Disc Skewness	
		Disc Kurtosis	
		Disc SUVpeak (0.5ml)	
		Disc SUVpeak (1ml)	
		Disc TLG	
	Discretized Histogram (4 Features)	Disc Histo Entropy Log10	Randomness of a discretized histogram distribution using a log base 10 formula
		Disc Histo Entropy Log2	Randomness of a discretized histogram distribution using a log base 2 formula
		Disc Histo Energy	Uniformity of the histogram distribution
		Disc Histo AUC	Area under the curve of discretized histogram
	Shape and Volume (5 Features)	Volume (ml)	Volume of the VOI in ml
		Volume (Voxels)	Volume of the VOI in Voxels
		Sphericity	How spherical the VOI is; 1 = perfect sphere
		Surface Area (mm2)	The surface area of the VOI in mm2
		Compacity	How compact the VOI is, $\text{Area}^{3/2} / \text{Volume}$
Second Order Statistics	Grey-level co-occurrence matrix (GLCM) (7 Features)	GLCM	Calculated from 13 different directions in 3D space and takes account the arrangement of pairs of voxels
		GLCM Homogeneity	Homogeneity of grey-level voxel pairs
		GLCM Energy	Uniformity of grey-level voxel pairs
		GLCM Contrast	Variance or Inertia, the local variation in GLCM
		GLCM Correlation	Linear dependency of grey-levels in GLCM
		GLCM Entropy log10	The randomness of grey-level pairs (log10)

Higher Order Statistics		<b>GLCM Entropy Log2</b>	The randomness of grey-level pairs (log2)
		<b>GLCM Dissimilarity</b>	Variation of grey-level voxel pairs averaged over 13 directions
	<b>Grey-level run-length matrix (GLRLM) (11 Features)</b>	<b>GLRLM</b>	The size of homogeneous runs for each grey level
		<b>GLRLM_SRE</b>	Short / Long Run Emphasis, distribution of the short / long homogeneous runs in an image
		<b>GLRLM_LRE</b>	
		<b>GLRLM_LGRE</b>	Low / High Grey level Run Emphasis, the distribution of low / high grey-level runs
		<b>GLRLM_HGRE</b>	
		<b>GLRLM_SRLGE</b>	Short-Run Low / High Grey-level Emphasis, distribution of short homogeneous runs with low / high grey levels
		<b>GLRLM_SRHGE</b>	
		<b>GLRLM_LRLGE</b>	Long-Run Low / High Grey-level Emphasis, distribution of short homogeneous runs with low / high grey levels
		<b>GLRLM_LRHGE</b>	
		<b>GLRLM_GLNU</b>	Grey-level / Run Length Non-Uniformity, the non uniformity of the grey-levels or the length of homogeneous runs
		<b>GLRLM_RLNU</b>	
		<b>GLRLM_RP</b>	Run percentage, homogeneity of the homogeneous runs
	<b>Neighbourhood Grey-level difference matrix (NGLDM) (3 Features)</b>	<b>NGLDM</b>	Difference of grey-levels between 1 voxel and its 26 neighbours in 3D
		<b>NGLDM_Coarseness</b>	Level of spatial rate of change in intensity
		<b>NGLDM_Contrast</b>	Intensity difference between neighbouring regions
		<b>NGLDM_Busyness</b>	Spatial frequency of changes in intensity
	<b>Grey-level size zone matrix (GLZLM) (11 Features)</b>	<b>GLZLM</b>	Grey-level zone length matrix, size of homogeneous zones for each grey-level
		<b>GLZLM_SZE</b>	Short / Long zone Emphasis, the distribution of the short / long homogeneous zones in an image
		<b>GLZLM_LZE</b>	
		<b>GLZLM_LGZE</b>	Low / High Grey level zone emphasis, distribution of low / high grey level zones
		<b>GLZLM_HGZE</b>	
		<b>GLZLM_SZLGE</b>	Short zone Low / High Grey level zone emphasis, distribution of short homogeneous zones with low / high grey-levels
		<b>GLZLM_SZHGE</b>	
		<b>GLZLM_LZLGE</b>	Long zone Low / High Grey level zone emphasis, distribution of long homogeneous zones with low / high grey-levels
		<b>GLZLM_LZHGE</b>	
		<b>GLZLM_GLNU</b>	Grey-level Non-Uniformity for zone and Zone Length Non-Uniformity, non-uniformity of the grey-levels or the length of the homogeneous zones
		<b>GLZLM_ZLNU</b>	
		<b>GLZLM_ZP</b>	Zone percentage measures homogeneity of homogeneous zones

Definitions of the radiomics features used in this study adapted from published works (Sah, et al., 2019; Nioche, et al., 2018).

## 4.7 Machine Learning

In this section, I will describe in detail the key elements of the machine learning code, the algorithms used, and the parameters used to describe and compare the efficacy of the algorithms in different applications of the data.

### 4.7.1 Overview

This project used the Anaconda platform to load the “Keras” environment for access to machine learning tools and the “Theano” environment for access to deep learning. All code was written in python language using the “Spyder” program to enable saving and easy transfer of code to external parties for discussion. Of the 18 studies reviewed in relation to oesophageal cancer PET; 8 studies have explored a machine learning application, utilising a variety of algorithm architectures summarised below:

**Table 8 Summary of machine learning techniques used in literature**

Reference	LR	SVM	GB	RF	ELM	CNN	LDA	KN	GNB	DT
(Beukinga, et al., 2018)	x									
(Zhang, et al., 2014)		x								
(Ypsilantis, et al., 2015)	x	x	x	x		x				
(Van Rossum, et al., 2016)	x									
(Paul, et al., 2017)				x						
(Xiong, et al., 2018)	x	x		x	x					
(Cao, et al., 2020)	x									
(Desbordes, et al., 2017)		x		x						
(Chen, et al., 2019)	x									
(Yang, et al., 2019)						x				
This Project	x	x					x	x	x	x

*Comparison of the different machine learning algorithms used in various similar publications. Key: Logistic Regression (LR); Support Vector Machines (SVM); Gradient Boosting (GB); Random Forest (RF); Extreme Machine Learning (ELM); Convolutional Neural Network (CNN); Linear Discrimination Analysis (LDA); K Neighbours Classifier (KN); Gaussian Naïve Bayes (GNB); Decision Tree Classifier (DT).*

Example machine learning code available from Brownlee (2019) was modified to fit the requirements of this project. This project compared the performance of six machine learning algorithms: Logistic Regression, Linear Discrimination Analysis, K Neighbours Classifier, Decision Tree Classifier, Gaussian



Naïve Bayes, Support Vector Machines. Choices are based on a balance between replicating results from the 2 most popular algorithms tested to date, comparison with 4 currently untested algorithms (in this context) and exploring a variety of fundamentally different ML algorithm architectures. Further individual justification below:

#### **Logistic Regression (LR)**

LR was chosen because it was mostly widely used in similar literature and, by design, is ideally suited to classification problems, such as classifying data as “success” or failure”.

#### **Support Vector Machine (SVM)**

SVM was similarly chosen for prevalence in the existing literature in a number of key related studies and again because SVM, is suited to classification problems.

#### **Gaussian Naïve Bayes (GNB) and Linear Discrimination Analysis (LDA)**

GNB and LDA were chosen because their inherent probabilistic basis offers an alternative architecture. GNB architecture has not been discussed in the literature related to upper GI treatment prediction and has been included as a contribution to the literature.

#### **K-nearest neighbours Classifier (KN)**

KN has been chosen, similarly because of its lack of reporting in the literature and because again, cluster-style analysis has not yet been explored in relation to upper GI treatment prediction. Note, the optimisation of the K-value was beyond the scope of this project and therefore, initial analysis was performed using the default value of five.

#### **Decision Tree Classifier (DT)**

Random Forest (RF) has previously shown success in other studies (Desbordes, et al., 2017). However, we chose a DT algorithm to maintain a level of control over the inputs and function of the algorithm; as aforementioned, the disadvantage of RF algorithms is the ‘black box’ element of the ‘randomness’ of the individual decision trees within the forest. Furthermore, including at least 1 decision-tree style algorithm gave a good spread of different ML architecture types.

### **4.7.2 Train, Test and Validation**

In machine learning, any algorithm requires a “Train”, “Test” and “Validation” set of data. For the proposed algorithm architecture, we split our data into 80% for “test and train” and kept 20% of the data as unseen by the algorithm for validation at the end. The “test and train” dataset was further split using a “stratified 10-fold cross validation estimate model” (Brownlee, 2019); this method took

the 80% portion of the data and shuffled it randomly. The k-fold process split the data into 10 parts and then the model was trained on 9, tested on 1. In this work, we have chosen  $k=10$  because this “has been found through experimentation to generally result in a model skill estimate with low bias a modest variance” (Brownlee, 2020). The stratification means that each train-test set is selected to have the same distribution of ‘successes’ and ‘failures’ from the dataset. In the models used in this work, each machine learning model was fitted based on a single set using 9/10ths of the 80% “test and train” data. The “test-train” accuracy generated was the accuracy of the trained model on the remaining 1/10<sup>th</sup> of the 80% “test and train” data.

A “random\_state” variable, set equal to 1, was used in conjunction with a “shuffle” function to essentially train and test the data 10 times using a random, but equally proportioned (between successes and failures), set of the data (Brownlee, 2019). We tested each algorithm using the same 10 sets of cross validation data. An accuracy metric assessed the accuracy of each model, defined as the ratio of correctly predicted successes divided by the total number of successes in the dataset, expressed as a percentage. The 10-fold cross validation method meant that 10 “test-train” accuracy measures were produced for each algorithm using the training dataset, giving an average accuracy from  $k-1$  (9) different attempts. For each attempt, the model was fitted on the “train” data, tested on the “test” set, the accuracy recorded and then the model discarded. The training accuracies were recorded and averaged in the python code, see Appendix 9.6. The purpose of this step was to be able to initially compare the accuracy of the different models as the program was built and developed.

Later in the program, each algorithm was then trained (built) for a final time using the full 80% portion of the “test-train” dataset and validated using the remaining 20% of previously unseen, held back, data (Brownlee, 2020). The purpose of the final validation was to give a final accuracy measure using the full dataset. The predictive power of each algorithm was evaluated using a confusion matrix, which showed the number of correct and incorrect predictions (Brownlee, 2019). In our project, this matrix predicts as follows:

**Table 9 Guide to prediction matrix**

	<b>Failure Predicted</b>	<b>Success Predicted</b>
<b>Failure Expected</b>	True Positive (Failure Predicted as Failure)	False Positive (Success Predicted as Failure)
<b>Success Expected</b>	False Negative (Failure Predicted as Success)	True Negative (Success Predicted as Success)

*Schematic illustration of the different fields of the confusion matrix, adapted from Brownlee et al (2019).*

In this context, the ability to predict a failure of treatment was the “positive” result. A perfect model, with X ‘failures’ and Y ‘successes’ predicted correctly with no successes predicted as failures and vice versa, would give a classification accuracy of 100% and be denoted:

$$\begin{vmatrix} X & 0 \\ 0 & Y \end{vmatrix}$$

Models were evaluated by providing a breakdown for each class (Success and Failure) by giving the accuracy, precision, recall, f1-score and support with the macro and weighted average:

- Accuracy: the number of correct predictions against the total number predictions, (true positives + true negatives) / total predictions.
- Precision: the ratio of true positives / (true positives + false positives), i.e. the ability of the model not to label a failure as a success (e.g. the Positive Predictive Value, PPV).
- Recall: the ratio of true positives / (true positives + false negatives), i.e. the ability of the model to find all of the true successes (e.g. the Sensitivity).
- F1-score: The equally weighted average of the recall and the precision
- Support: The number of successes / failures in each group.
- Macro Average: averaging the unweighted mean per class

- Weighted average: averaging the support weighted mean per class, i.e. the average but weighted by the number of successes and failures.

Complementary to the above terms (and not determined by the machine learning code), the specificity and negative predictive values were defined as follows (Trevethan, 2017):

- Specificity:  $\text{true negatives} / (\text{false positives} + \text{true negatives})$
- Negative Predictive value:  $\text{true negatives} / (\text{false negatives} + \text{true negatives})$

#### **4.7.3 Summary of the tests evaluated**

An initial pilot study was run to test and understand the ML algorithm; the aim of the pilot was to predict whether a patient received curative or palliative treatment, based on their TNM score. This was of limited clinical benefit because patients with an M score  $> 0$  (i.e. with metastatic spread) usually received palliative treatment; therefore, patients who received curative and palliative treatment were already well defined. The purpose of this pilot was to test the initial code as a ‘proof of concept’ which could be adapted to larger, less well-defined datasets, such as one containing the radiomic signature from a PET image.

We present several sub-datasets from the initial group of 778 patients:

- Dataset 1: Stratified to include all patients with upper GI cancer (oesophagus, OGJ and stomach, squamous cell or adenocarcinoma), treated with both curative and palliative treatments. The aim was to predict TNM 2-year, 1 year and 90 days survival from the TNM score. Total of 660 patients.
- Dataset 2: Stratified to include all oesophageal and OGJ adenocarcinoma patients, treated with definitive (curative) surgery alone / in addition to neoadjuvant / adjuvant chemotherapy. The aim was to predict 2-year, 1-year and 90 days survival from the TNM score. Total of 144 patients.
- Dataset 3: A subset of 92 patients with 3D PET scans (from dataset 2 of 144 surgically treated, oesophageal / OGJ adenocarcinoma patients), reconstructed with OSEM only. The aim was to predict 2-year, 1-year and 90 day survival. Each survival end point was tested with:
  - 58 radiomic features (from the primary tumour) and the TNM Score.
  - 58 radiomic features and TNM score with each result normalised onto the same scale between 0 and 1.

- 58 radiomic features only.
- TNM Score only
- Dataset 4: A further subset of 79 patients (from dataset 3) had both OSEM and BSREM reconstructions available and a further comparison of the radiomic signature and machine learning performance was made by using two identical datasets with identical regions of interest differing only in the image reconstruction method .

Table 10 indicates the number of patients classified as “Success” and “Failure” depending on the DFS time end point and the number of “Success” and “Failure” used in the “Test-Train” and “Validation” sets. The attempted machine learning experiments are summarised below in (Table 12). For each case, the number of patients in both the “Test-Train” (80% of the data) and the “Validation” (20% of the data) was indicated.

**Table 10 Summary of number of Successes and Failures in each class**

		2 Years Disease - Free Survival				1 Years Disease - Free Survival				90 Days Disease - Free Survival			
	Dataset	1	2	3	4	1	2	3	4	1	2	3	4
	Total Patients	660	144	92	79	660	144	92	79	660	144	92	79
A	Success	284	69	39	33	435	107	67	59	614	140	89	76
	Failure	376	75	53	46	225	37	25	20	46	4	3	3
B	Test-Train Success	221	56	33	27	342	87	53	46	489	111	71	60
	Test-Train Failure	307	59	40	36	186	28	20	17	39	4	2	3
C	Validation Success	63	13	6	6	93	20	14	13	125	29	18	16
	Validation Failure	69	16	13	10	39	9	5	3	7	0	1	0

*Overview of the number of patients assigned to each dataset and a summary of the split of patients between success and failure in A) the complete dataset, B) the test-train group and C) the validation group.*

The machine learning code itself was validated and written by replicating an example code and ensuring the published results were exactly reproduced (Brownlee, 2019). Brownlee’s (2019) initial code was written to predict, from a well-prescribed dataset, the type of flower from various parameters relating to the petal size; five features for 150 entries. Below is a brief summary of the various versions of the code with the key differences and approximate computational time:

**Table 11 Summary of versions and computational time**

Version	Details	Approx. Computational time / no. executions
1	Exact replication of Brownlee et al (2019) code.	2 min / 15
2	Repeated using the same code and dimensions but with 5 features from our dataset (T score, N Score, M Score, SUVmax and DFS time)	2 min / 10
3	Added additional feature to the dataset (6 features)	2 min / 5
4	Using all 61 features from my dataset but including the DFS time as a feature	25 min / 10
5	Patient entries from the dataset removed due to poor computation of the radiomic features resulting in “not a number” entries (see 5.1.2).	25 min / 10
6	Survival time data removed to start testing the link between radiomic / clinical features and success / failed treatment.	25 min / 5
7-9	Removing / including various features of the code to increase speed, for example removing the display of the correlation matrix (a display of plotting all features against all features, 61 x 61 individual graphs)	3-25 min / 25
10	Added prediction matrix analysis for every ML algorithm tested.	4 min / 6
11-13	Testing with different success end points: 90-day, 1-year and 2-year DFS.	4 min / 10
14	Pilot study using TNM scores only to predict DFS at 2 years	2 min / 5
15	First experimental dataset – OSEM only images, 2 year DFS with all features.	2 min / 3
16-30	Analysis of all experiments described in Table 12.	2 min / 30

*A summary of the approximate computational time, number of executions of the code and a summary of the changes between different versions to show the progression and development of the program.*

Table 12 Summary of machine learning tests

Name	No. Patients	Test – Train / Validation	Data Used	ML Prediction Goal
Pilot	238: 160C / 69P		TNM Score, Survival Status	C or P
Dataset 1: All upper GI cancer / treatment	660	528 / 132	TNM Score Only	DFS at 90 days, 1 Year and 2 Years
Dataset 2: Oeso / OGJ adeno Carcin. curative treat	144	115 / 29	TNM Score Only	
Dataset 3: DFS Prediction	92	73 / 19	58 RF, TNM	
			58 RF Only	
			TNM Only	
			58 RF and TNM Normalised	
Dataset 4: Effect of Image Reconstruction	79	63/16	58 RF, TNM for lesion matched reconstructions	

Key: RF = radiomic features; DFS = DFS; C = curative treatment, P = palliative treatment

*Summary of the different machine learning experiments performed including the dataset, number of patients in the dataset, split of test/train and validation patients, type of data used and the prediction goal: disease-free survival at 90 days, 1 and 2 years.*

## 5 Results

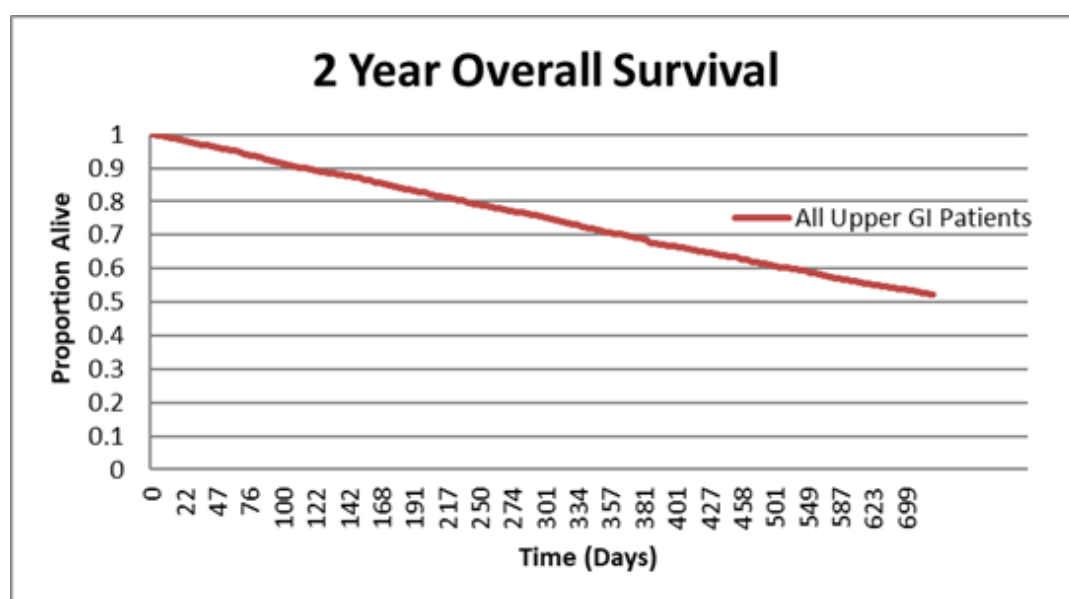
This chapter covers the key results from this project and covers analysis of the radiomic signature from PET images of upper gastro-intestinal (GI) adenocarcinomas. The first section on survival prediction covers analysis of whether any radiomic features are more closely associated with disease-free survival (DFS) at a given time-point. The second section covers a comparison of the radiomic signature for different image reconstruction methods (OSEM and BSREM). The final section describes a machine learning approach to predicting DFS based on the radiomic signature; this includes a pilot study to build a proof-of concept machine learning (ML) algorithm, the evaluation of a set of different ML algorithms in predicting survival and a comparison of ML performance where the radiomic signature was downloaded from OSEM and BSREM images.

### 5.1 Survival prediction with radiomics

#### 5.1.1 Clinical Data

A Kaplan-Meier 2 year overall survival (Figure 30) and DFS (Figure 31) curve was plotted for the total initial cohort of 778 patients with upper GI cancer (see Table 3). We found that the overall survival of patients attending the Northern Oesophago-Gastric Unit (NOGU) MDT was 52.3%. Across all pathologies, we found the 2-year DFS rate was 43.3% of patients.

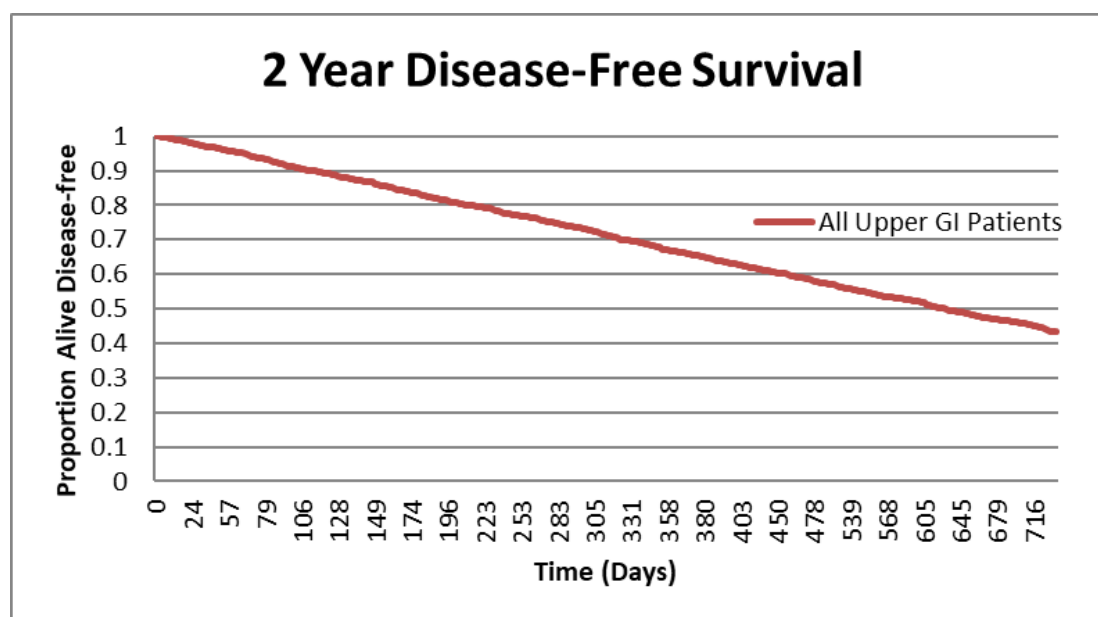
**Figure 30 Survival curve for all upper GI patients.**



*Kaplan-Meier overall survival curve for all upper GI patients covering 2 years*



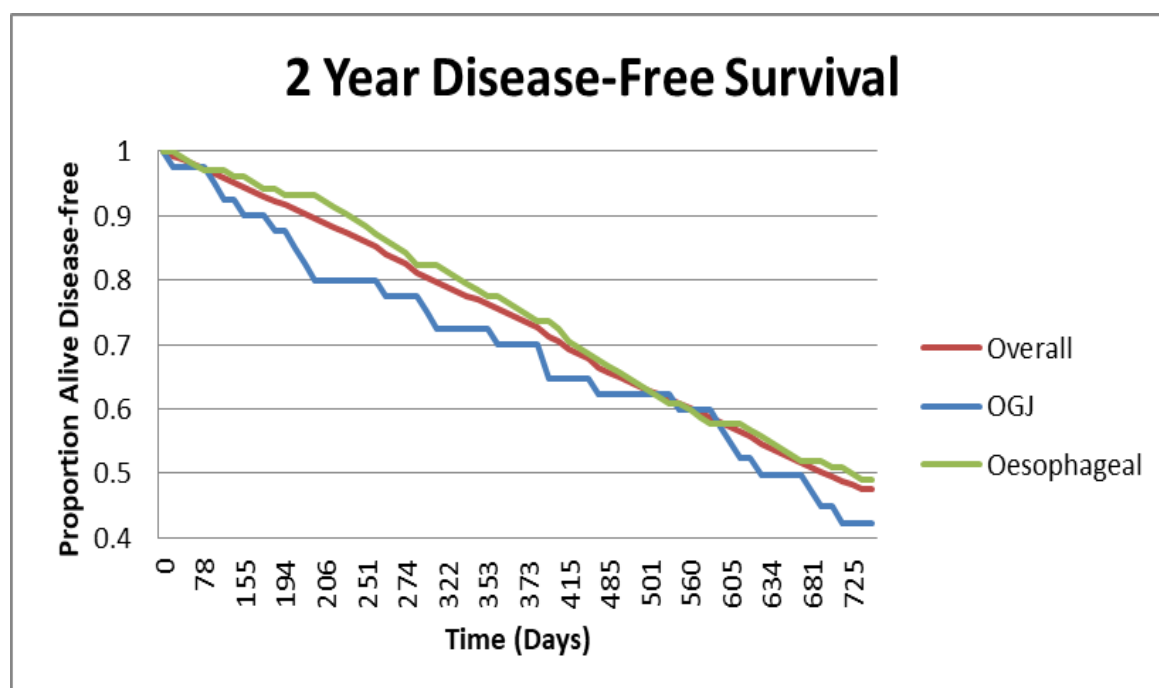
**Figure 31 Disease-free survival curve for all upper GI patients.**



*Kaplan-Meier 2 Year disease-free survival (DFS) curve for all upper GI patients.*

Our stratified patient cohort consisted of 144 patients, distributed as shown in Table 4. Survival curves shown for the subset of 144 patients (102 oesophageal, 42 OGJ) meeting the study inclusion criteria: adenocarcinoma and definitive surgical treatment. Survival curves were plotted for 2-year and 1-year and 90-day DFS curves. 1-year DFS was displayed for comparison with previous studies (Chen, et al., 2019; Foley, et al., 2018; Van Rossum, et al., 2016; Yang, et al., 2019). 90-day survival was chosen to investigate whether there were any radiomic feature(s) indicating a fast deterioration following treatment. 2-year survival was chosen a) to match local practice in defining a “successful” treatment and b) to address a gap in the published literature. Curves displayed at 2-years, 1-year, and 90-days for clearer visualisation of the relative split between oesophageal and OGJ at different survival endpoints.

Survival curves show that the 90-day survival of oesophageal and OGJ patients is greater than 95% in both groups. At 1 year, 70% of oesophageal and 75.4% of OGJ had survived disease-free. At 2 years, 42.3% of OGJ and 48.9% of oesophageal patients had survived disease-free. The largest deviation between the survival rates of OGJ and oesophageal patients was at ~205 days where the survival rate of OGJ and oesophageal patients was 80% and 93.1% respectively. Overall, 47.5% of upper GI patients survived disease-free at 2 years, with 74.1% at 1 year.

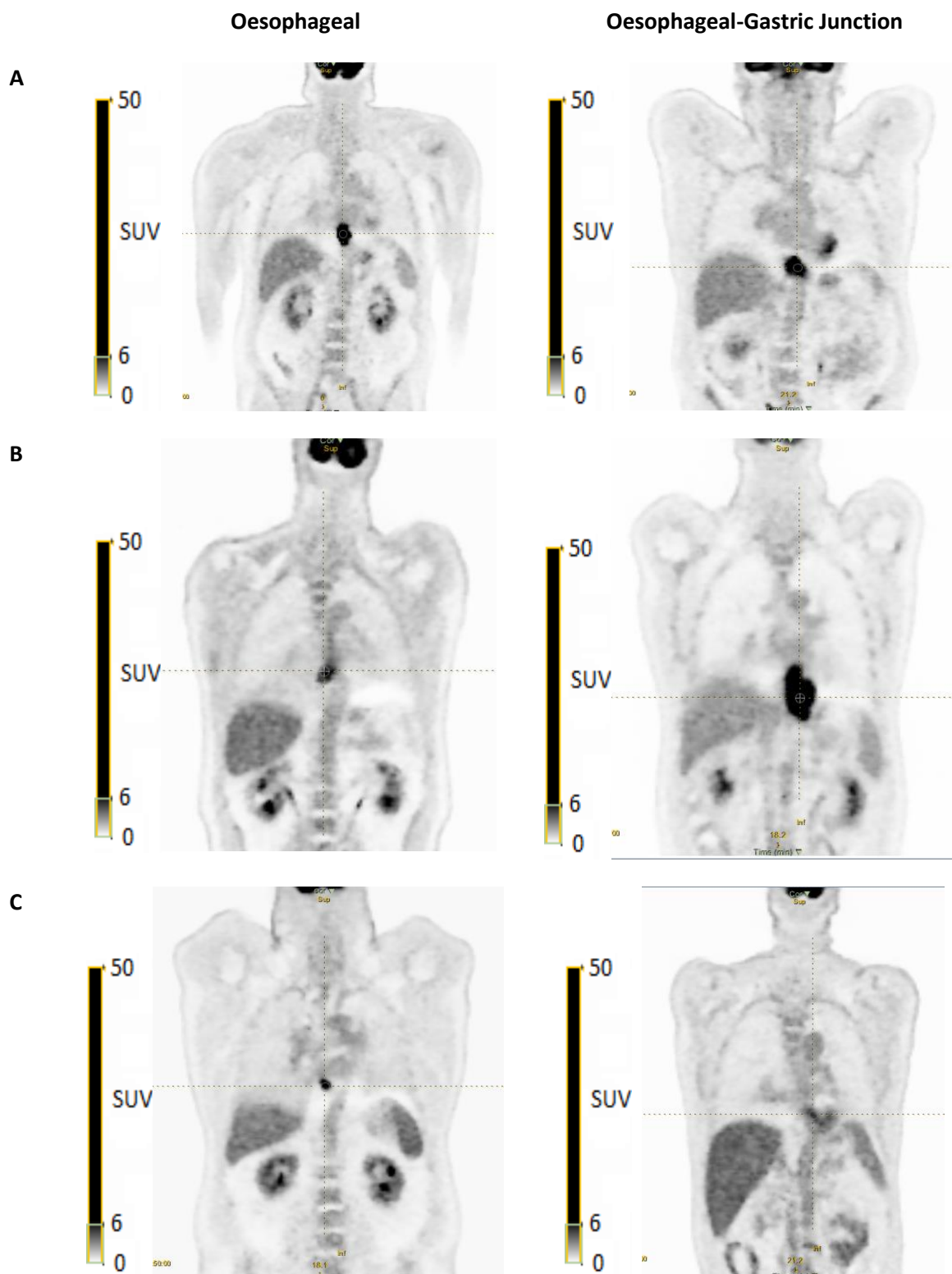
**Figure 32 Survival curve for oesophageal and OGJ adenocarcinoma**

Kaplan-Meier disease-free survival curve for all oesophageal and OGJ patients with adenocarcinoma, shown for 2 years.

### 5.1.2 Radiomics

From the initial dataset of 144 patients, PET Images were either not found (10 patients) or were using 2D data only (29 patients). We excluded patients with 2D only data because the radiomics parameters used would not compute correctly or comparably to the 3D datasets (majority of patients). The radiomics signature was downloaded for 105 patients with 3D-OSEM PET images and a primary oesophageal or OGJ adenocarcinoma. The radiomics signature consisted of 9 Conventional, 8 Conventional (Discretized Image), 4 Discretised Histogram, 5 Shape and Volume, 7 Grey-level co-occurrence matrix, 11 Grey-level run-length matrix, 3 Neighbourhood Grey-level difference matrix, and 11 Grey-level size zone matrix features: 58 image features in total. 13 further patients were excluded from radiomics analysis because the full radiomic signature did not compute correctly for a variety of reasons: lesion <70 voxels and too small to compute co-occurrence statistics (10 patients), lesion shape would not fit 10ml sphere (1 patient), lesion too small to compute SUVpeak 10ml value (2 patients). Radiomics analysis was performed on the remaining 92 patients with a complete radiomic signature (Dataset 3). The excluded group of 13 patients contained 6 patients with a successful (DFS > 2 years) and 7 patients with a failed (DFS < 2 years) treatment.

**Figure 33 Sample images and regions for successful treatment, failed treatment and excluded studies**



Key: Coronal slices to show A: Successful Treatment (DFS > 2 years). B: Failure of Treatment (Disease recurrence in < 2 years). C: Excluded on grounds of incomputable radiomic signature.





Furthermore, the standard significance threshold for a 2 tailed normal distribution,  $p = 0.05$ , was adjusted using the “Bonferroni correction” (Dunn, 1961), which adjusts the  $p$  value threshold for significance by the number of comparisons being made on the data. Here

$$\alpha_{\text{new}} = \alpha_{\text{original}} / n = 0.05 / 62 = 0.000806 \quad \text{Equation 13}$$

Therefore, we should only reject the null hypothesis of each individual test if the  $p$  value is less than 0.000806. For our data, this corresponds to a correlation coefficient of  $r < -0.34$  and  $r > 0.34$ .

Table 13 shows the radiomic features which correlated strongly ( $>0.9$ ) with SUVmax. Correlations with SUVmean, min, peak etc have been excluded for clarity.

**Table 13 Pearson’s correlation coefficient  $>0.9$  against SUVmax**

Radiomic Feature	Correlation Coefficient
GLZLM_SZHGE	0.97
GLRLM_HGRE	0.97
GLZLM_HGZE	0.96
GLRLM_SRHGE	0.93

*Table to show the parameters which showed a Pearson’s correlation coefficient of greater than 0.9 with SUVmax.*

We found that SUVmax correlated strongly with grey-level size-zone matrix short zone high grey level emphasis (GLZLM\_SZHGE), high grey-level zone emphasis (GLZLM\_HGZE) and short-zone emphasis (GLZLM\_SZE); the grey-level run length matrix high grey level run emphasis (GLRLM\_HGRE) and short run high grey-level emphasis (GLRLM\_SRHGE). In other words, SUVmax correlated strongly with parameters which describe areas of the image with a high grey-level, an expected result however one which highlights the nature of such radiomic features.

Of most pertinent interest to this project was whether the DFS time correlated with any of the radiomic features however, we found no significant correlation between any single radiomic feature and the DFS time at 1 and 2 years. Correlation for a DFS time of 90 days was unreliable as, at 90 days, only 2/92 patients had recurred / expired (Table 10).

For dataset 3, each of the 92 patients were in either the “failure” or the “success” class for DFS at 1 and 2 years. For each of the two classes (failure and success), the average value of each radiomic feature was calculated. From the average value for each class (failure and success, at 1 and 2 years), the percentage difference between classes was calculated to determine whether any features were significantly higher or lower for a particular DFS time. Analysis with 90-day survival was excluded due to inconsistent results with low numbers of patients in this category.

Figure 36 shows the features which exhibited, on average, a greater than 10% increase in patients who recurred or expired within 1 and 2 years. Patients who recurred within 1 or 2 years of treatment generally had larger (Volume (ml), Volume (Voxels), Surface Area(mm2)) tumours with a more non-uniformly distributed grey-level (GLZLM\_ZLNU, GLZLM\_GLNU, GLRLM\_RLNU). Note, not included here is the percentage increase for TLG which was increased by 37% and 31% for 1 and 2 year DFS respectively. Whilst this showed the highest average increase, this value is inherently biased because TLG (and Discretized TLG) is a product of SUV mean and Volume. In this work, the apparent increase in TLG was mainly driven by the volume term, with SUVmean accounting for less than 5% of the increase.

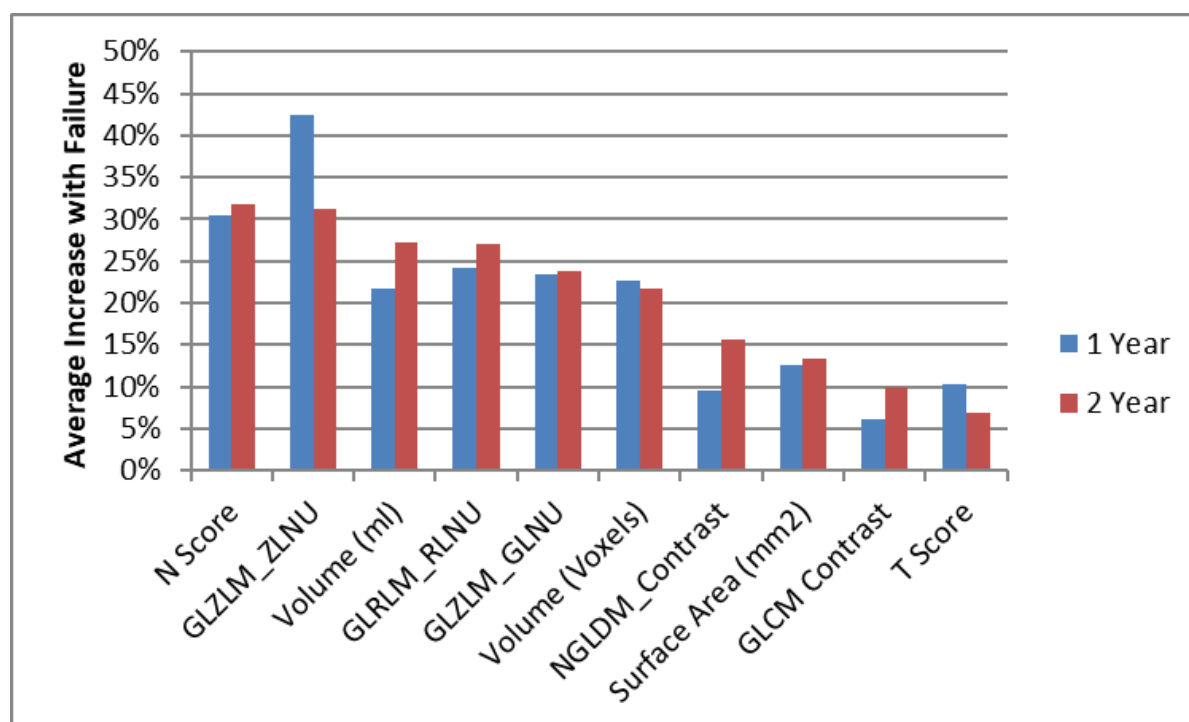
Furthermore, the average nodal score (N Score) was found to be higher in the group which recurred or expired within 1 / 2 years of treatment. Grey-level Size Zone Matrix Zone Length Non-Uniformity (GLZLM\_ZLNU) showed the highest increase for groups split at 1 year; a 42.5% increase on average.

Figure 37 shows the average value in each class (success / failure) with each feature normalised to between 0 and 1. Normalisation was calculated by subtracting the minimum value from each value and dividing by the range of values for that feature:

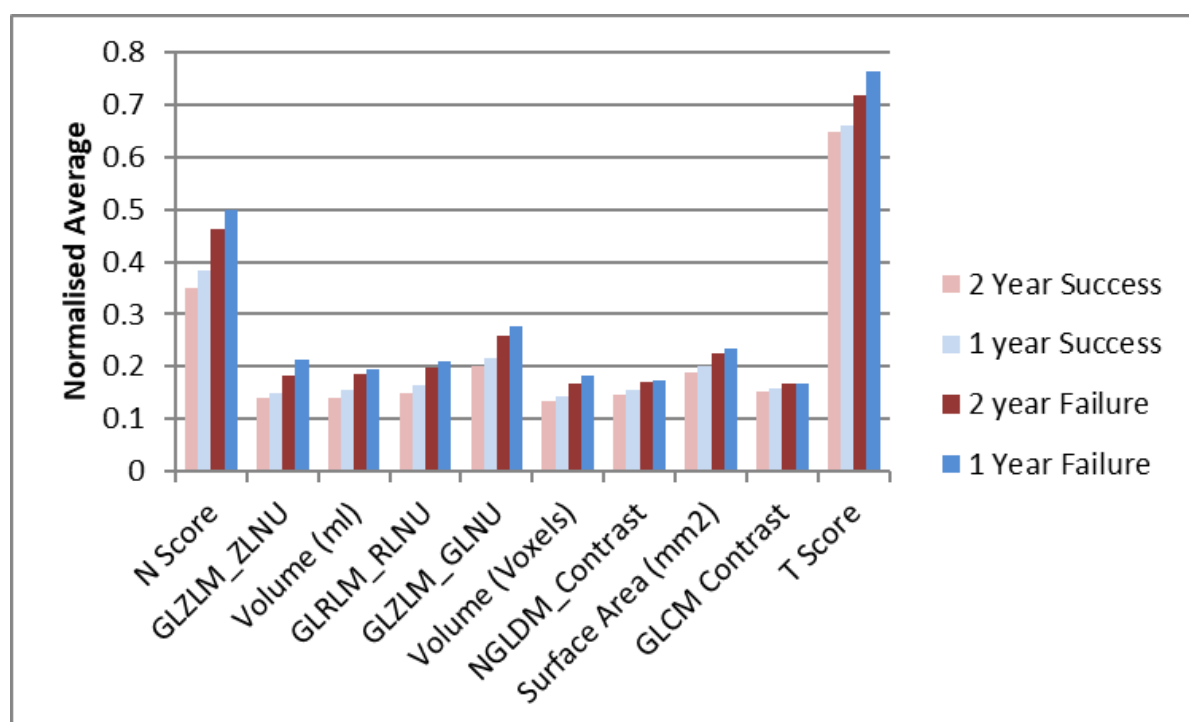
$$\text{(Value - Value}_{\text{Min}} \text{) / Range} \quad \text{Equation 14}$$

For the normalised features with a difference of greater than 10%, we found that, on average, larger and more heterogeneously distributed tumours were associated with poorer outcomes. The average, normalised value was generally higher again for 1-year DFS suggesting that larger, more heterogeneously distributed tumours were also associated with shorter DFS times.



**Figure 36 Features with highest average increase for failed treatments**

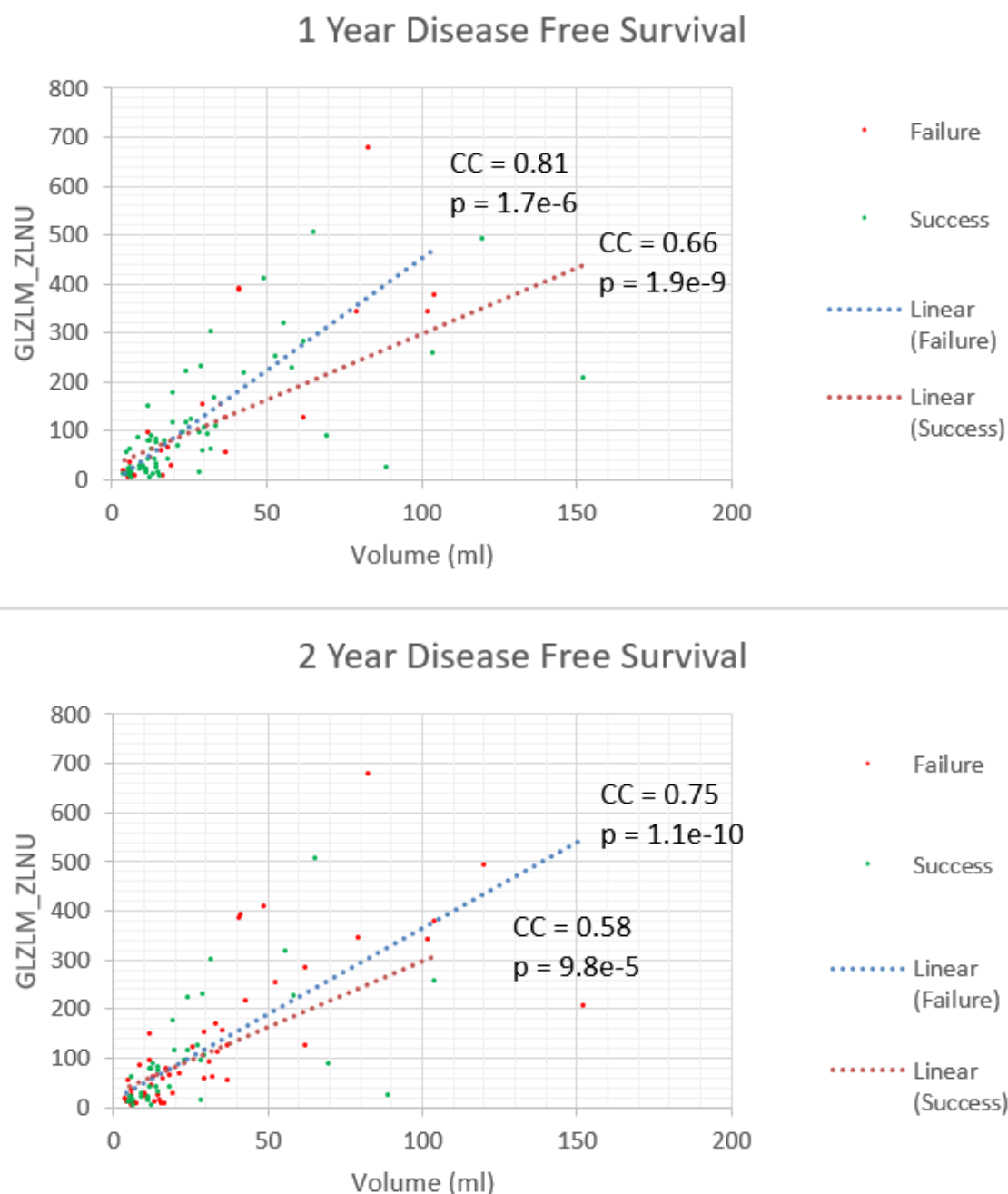
A bar chart to show a comparison of the average (92 patients) percentage increase observed in failed treatments for 1 and 2 year DFS, for all features with greater than 10% difference.

**Figure 37 Features with highest average increase for failed treatments (normalised)**

A bar chart to show a comparison of the average (92 patients) percentage increase observed in failed treatments for 1 and 2 year DFS, using normalised results and for all features showing greater than 10% difference.



Figure 38 GLZLM\_ZLNU against volume for 1 and 2 year disease free survival



The grey-level size zone matrix zone length non-uniformity (GLZLM\_ZLNU) against volume. Results shown for successful (green dot) and failed (red dot) treatments and for 1 and 2 year disease free survival.

When plotted against tumour volume (ml), the feature with the highest average difference (GLZLM\_ZLNU) showed a strong correlation coefficient (CC) between tumour volume and GLZLM\_ZLNU with stronger correlation for failed treatments and for 1 year DFS (Figure 38); all correlations were found to be statistically significant when the Bonferroni correction was applied to

the p value ( $p = 0.000806$ ). Note, the raw data supporting Figure 36 - Figure 38 can be found in Appendix 9.6.

Statistical significance between the groups was determined using a 2 tailed, unpaired unequal variance, student's t-test to compare the "failure" and "success" groups for 1- and 2-year DFS groups. As aforementioned, 90-day survival was excluded from this analysis due to a low distribution of patients in each class, leading to poor counting statistics (Table 10).

**Table 14 Student's T-Test comparing statistical significance of percentage differences**

Feature	1 Year Disease Free Survival	2 Year Disease Free Survival
T Score	0.03	0.11
N Score	0.18	0.15
GLZLM_GLNU	0.31	0.17
Disc TLG	0.30	0.22
Volume (ml)	0.42	0.25
GLRLM_RLNU	0.42	0.26
GLZLM_ZLNU	0.27	0.27
TLG	0.32	0.33
Volume (Voxels)	0.43	0.37

*The p values from a 2 tailed, unequal variance student's t-test for comparing the feature values for "failed" and "successful" treatments at 1 and 2 years.*

The highest confidence was found for the T-score (section 3.1.1, 4.3.1) when comparing failure and success of treatment at 1 year which was on average 10.3% higher for failed treatments. However, once the Bonferroni correction was applied for  $n = 62$  tests (Dunn, 1961), we should only reject the null hypothesis of each individual test if the p value is less than 0.000806. When using 2-year DFS to stratify patients, we found GLZLM\_GLNU, Disc TLG, Volume (ml), GLRLM\_RLNU and GLZLM\_ZLNU may indicate a weak trend however, more data will be required to confirm this and improve counting statistics. Ultimately, whilst we demonstrated weak trends, we were unable to demonstrate statistical significance when separating successful and failed treatment groups.

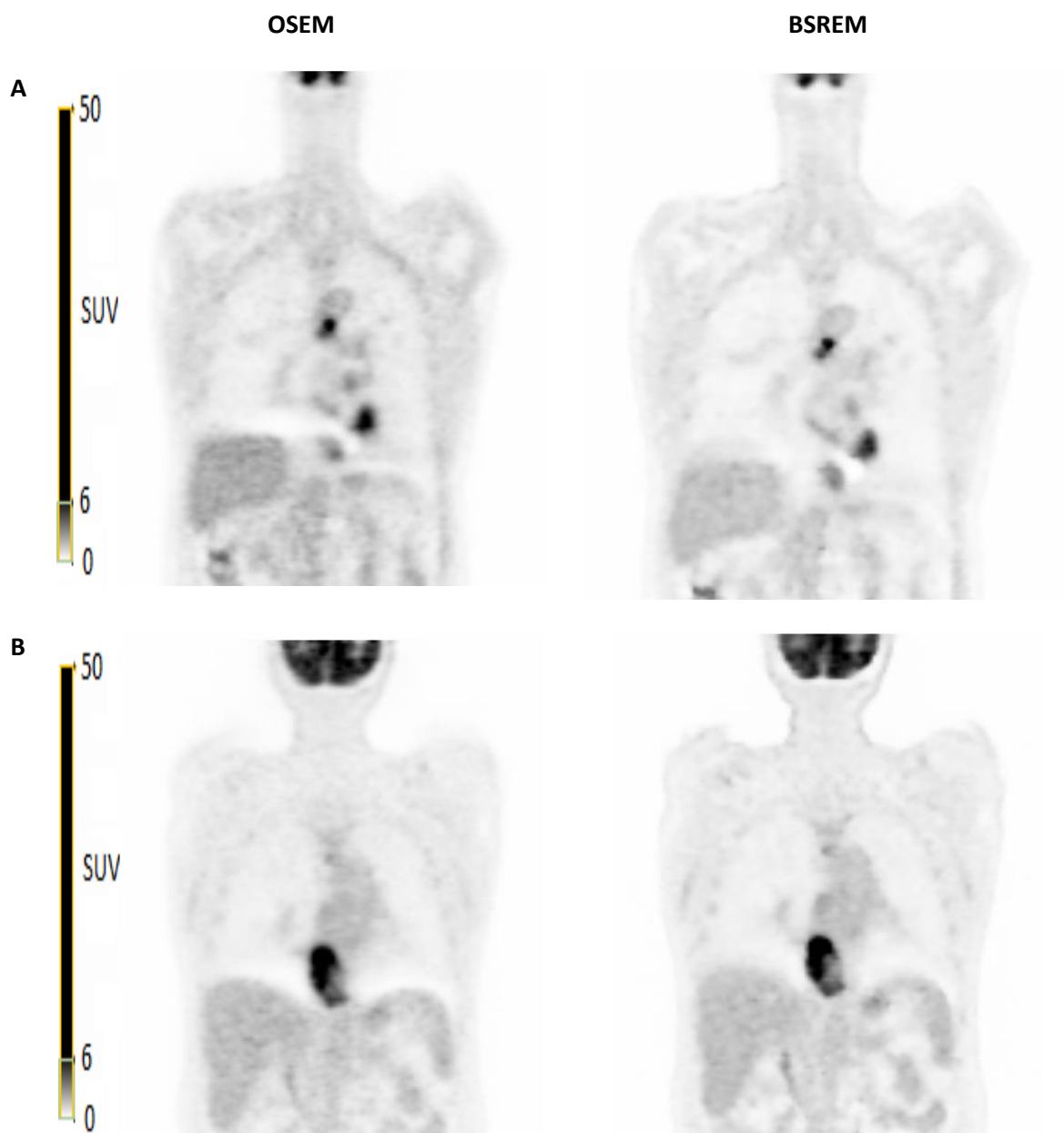
## 5.2 OSEM vs BSREM

We evaluated the radiomic signature for a further subset of 79 patients who had both OSEM and BSREM reconstructed images.

### 5.2.1 Radiomics Comparison

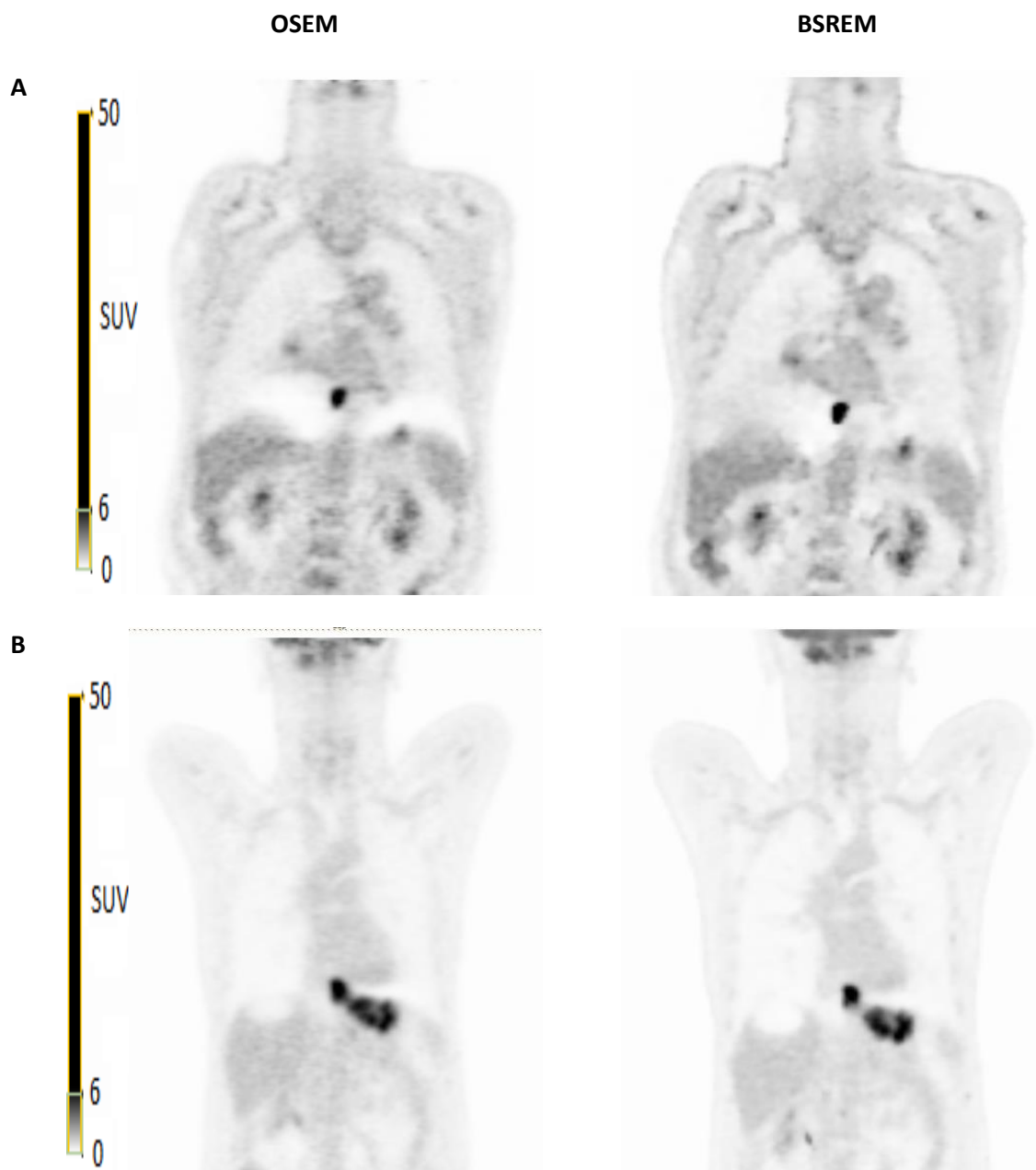
We compared the radiomic signature extracted from 79 patients for both OSEM and BSREM reconstructed images. We downloaded radiomics data from the primary tumour using PET data from both an OSEM and a BSREM reconstructed set of images.

**Figure 39 OSEM vs BSREM images for successful treatments for patients with the largest and smallest feature differences.**



*Key: A: oesophageal primary, successful treatment (DFS > 2 years), largest difference in GLCM contrast and SUVmax. B: oesophageal primary, successful treatment (DFS > 2 years), lowest difference in GLRLM\_SRE*

**Figure 40 OSEM vs BSREM images for failed treatments for patients with the largest differences in TLG and SUVmin.**



*Key: A: oesophageal primary, failed treatment (DFS < 2 years), largest difference in TLG. B: OGJ primary, failed treatment (DFS < 2 years), largest difference in SUVmin.*

For each patient, we used the same region of interest for both the OSEM and BSREM reconstructed images. For each radiomic feature, we determined an average, standard deviation, maximum, minimum and variance across all 79 patients and for each reconstruction. We calculated the percentage difference (between OSEM and BSREM images) for each radiomic feature and each

patient, and a further average, standard deviation, maximum, minimum and variance for the difference percentage results.

As aforementioned, we acquired the liver noise data using a 3cm diameter ROI as per the RECIST criteria (Eisenhauer, et al., 2009) for assurances of the spread of patient activity since we were unable to access this information. Liver noise was assessed for our group of matched OSEM and BSREM images:

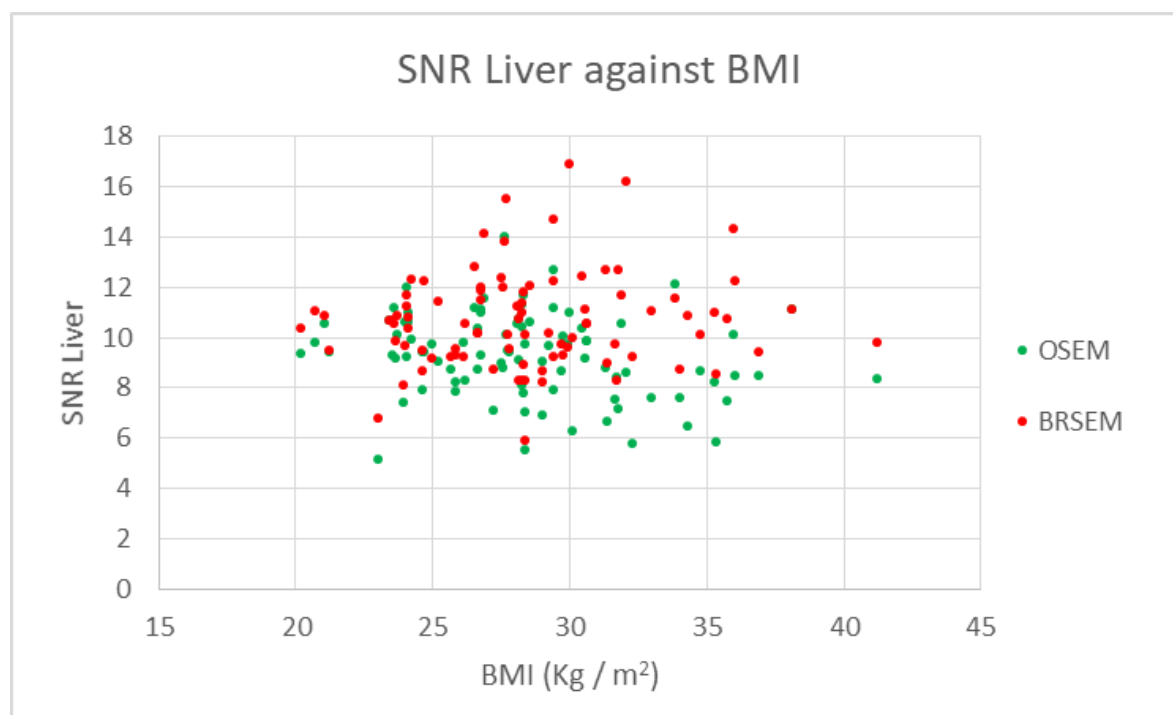
**Table 15 Liver Noise spread**

	OSEM	BSREM
Average	9.3	10.7
Range Min	5.1	5.9
Range Max	14.0	16.9
SD	1.7	1.9

*Table to show liver signal-to-noise for OSEM and BSREM images.*

Liver noise was higher for BRSEM reconstructed images however, when plotted against patient BMI, no discernible difference was noted, providing reassurance that whilst BSREM increases uptake values in the image, the image noise remained relatively consistent between patients.

**Figure 41 SNR liver plotted against BMI**



*SNR Liver plotted against BMI showed no significant pattern or correlation.*

Figure 42 and shows the average absolute difference for each radiomic feature between OSEM and BSREM reconstructed images. Grey-level co-occurrence matrix (GLCM) Contrast, neighbourhood grey-level difference matrix (NGLDM) Contrast, grey-level size zone matrix (GLZLM) short-zone high grey-level zone emphasis and zone-length non-uniformity, and grey-level co-occurrence matrix dissimilarity showed the largest differences (Table 16). Parameters related to the dynamic range of grey-levels in an image appeared to have a much larger range of values (e.g. GLCM contrast, NGLDM contrast) for BSREM images (Scapicchio, et al., 2021).

**Table 16 Largest differences in radiomic features for OSEM and BSREM images**

Radiomic Feature	Average	Variance
GLCM Contrast	104.0%	0.58
NGLDM_Contrast	91.8%	0.29
GLZLM_SZHGE	55.2%	0.31
GLZLM_ZLNU	51.3%	0.51
GLCM Dissimilarity	37.1%	0.05

*Table to show the largest average difference between OSEM and BSREM images with the associated variance.*

Furthermore, grey-level size zone matrix and grey-level co-occurrence matrix features demonstrated the largest variance in difference percentages; BSREM appears to create larger dynamic ranges of grey-level values when compared to OSEM images. I.e. a larger difference between the maximum and minimum pixel values.

**Table 17 Largest variances in the difference in radiomic features for OSEM and BSREM images**

Radiomic Feature	Variance
GLZLM_LZHGE	5.15
GLZLM_LZE	1.32
GLCM Energy	1.14
GLCM Contrast	0.58
GLZLM_ZLNU	0.51
SUV Skewness	0.41
Disc Skewness	0.37
GLZLM_SZHGE	0.31
NGLDM_Contrast	0.29
GLZLM_LZLGE	0.28

*Table to show the largest variances in the differences between radiomic features for lesion matched OSEM and BSREM images.*

Comparatively, features associated with volume and region shape remained at zero because the region remained constant however several radiomic features remained relatively stable with low average differences between OSEM and BSREM for grey-level run length matrix short and long run emphasis, run percentage, run length non uniformity, and the grey-level co-occurrence matrix entropy (log 10). In other words, features describing grey level run lengths in the image, and therefore describing small local variations (finer textures), remain relatively robust to different reconstructions of the same patient and relatively consistent between patients (low variance of the percentage differences between patients).

**Table 18 Smallest variances in the difference in radiomic features for OSEM and BSREM images**

Radiomic Feature	Average	Variance
GLRLM_SRE	0.6%	0.000
GLRLM_RP	0.5%	0.001
GLCM Entropy log10	3.0%	0.003
GLRLM_RLNU	2.2%	0.005
GLRLM_LRE	-0.3%	0.012

*Table to show the smallest variances in the differences between radiomic features for lesion matched OSEM and BSREM images, to show the most consistent parameters between the two reconstructions.*

For comparison, Table 19 shows the effect on commonly used parameters: SUVmax, SUVpeak (1ml sphere) and total lesion glycolysis (TLG); all three parameters described here show that BSREM increases the SUV relative to OSEM only images.

**Table 19 Average difference and variance in the difference of common radiomic features for OSEM and BSREM images**

Radiomic Feature	Average	Variance
SUVmax	25.9%	0.023
SUVpeak (1ml)	17.9%	0.007
TLG	14.3%	0.005

*Table to show the average difference and variance of commonly used parameters in clinical practice.*

Figure 43 shows the relative differences between parameters and indicates the relative direction for the differences; for example, GLCM contrast was considerably higher in BSREM images, whereas parameters such as Neighbourhood Grey-level difference matrix (NGLDM) coarseness and Grey-level

size zone matrix long zone low grey level emphasis (GLZLM\_LZLGE) were lower in BSREM images compared to OSEM only.

A 2-tailed, students t-test (samples of unequal variance) was performed to determine the statistical significance of any difference between OSEM and BSREM reconstructed radiomic features.

**Table 20 Ten most statistically significantly different radiomic features between OSEM and BSREM**

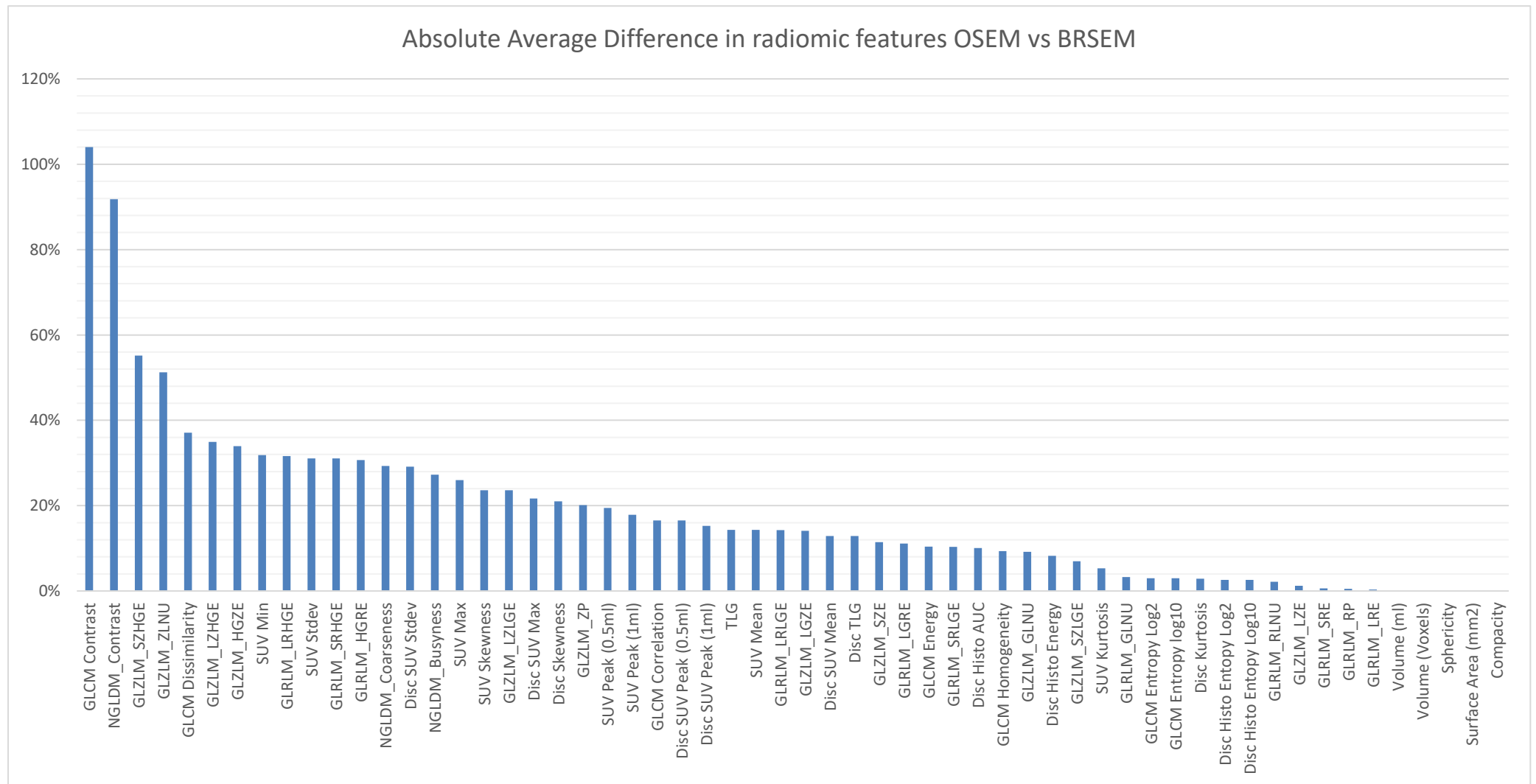
Radiomic Feature	Average Difference	Variance	Students T Test
SUVmin	31.9%	0.03	7.2558E-10
GLCM Correlation	-16.5%	0.02	0.0007
NGLDM Contrast	91.8%	0.29	0.0011
GLCM Dissimilarity	37.1%	0.05	0.0017
GLCM Contrast	104.0%	0.58	0.0019
GLZLM_SZE	11.4%	0.03	0.0043
NGLDM Coarseness	-29.3%	0.02	0.0114
Disc SUVmax	21.7%	0.03	0.0119
SUV Skewness	23.6%	0.41	0.0193
GLZLM_ZP	20.1%	0.05	0.0203
Disc SUV Stdev	29.2%	0.04	0.0292

*Table to show the most statistically significant differences in features downloaded for OSEM and BSREM images.*

SUVmin was, on average, 31.9% higher in BSREM images compared to OSEM and GLCM Correlation was, on average, -16.5% lower for BSREM images. Furthermore, differences in SUVmax (discretized image) were on average of 21.7% higher in BSREM images, indeed, conventional SUVmax had a p value of 0.055. By conventional metrics, all results in Table 20 were statistically significant to less than p value < 0.05, however, again, the Bonferroni correction for n = 62 tests (Dunn, 1961), gave a p value < 0.000806 meaning it is only the results for SUVmin and GLCM Correlation which are statistically significant here. Our results do still give weak support to there being significant differences in the other parameters shown in Table 20, however greater patient numbers will be required to verify this.

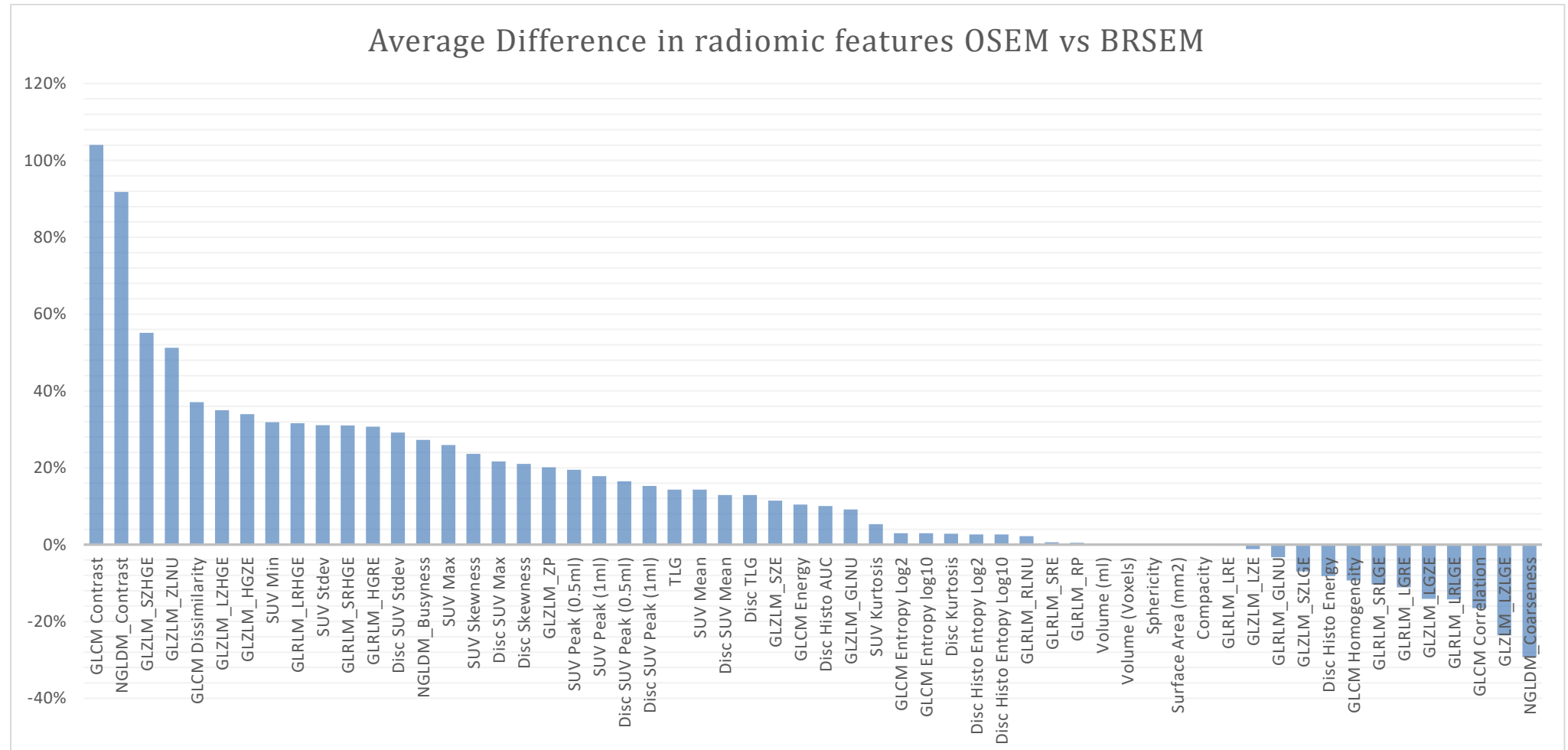


**Figure 42 Absolute percentage difference for 58 radiomic features OSEM vs BSREM**



*The absolute difference between radiomic features for OSEM and BSREM plotted in order of greatest to least difference.*

**Figure 43 Percentage difference for 58 radiomic features OSEM vs BSREM**



*The raw average percentage difference between radiomic features for OSEM and BSREM.*

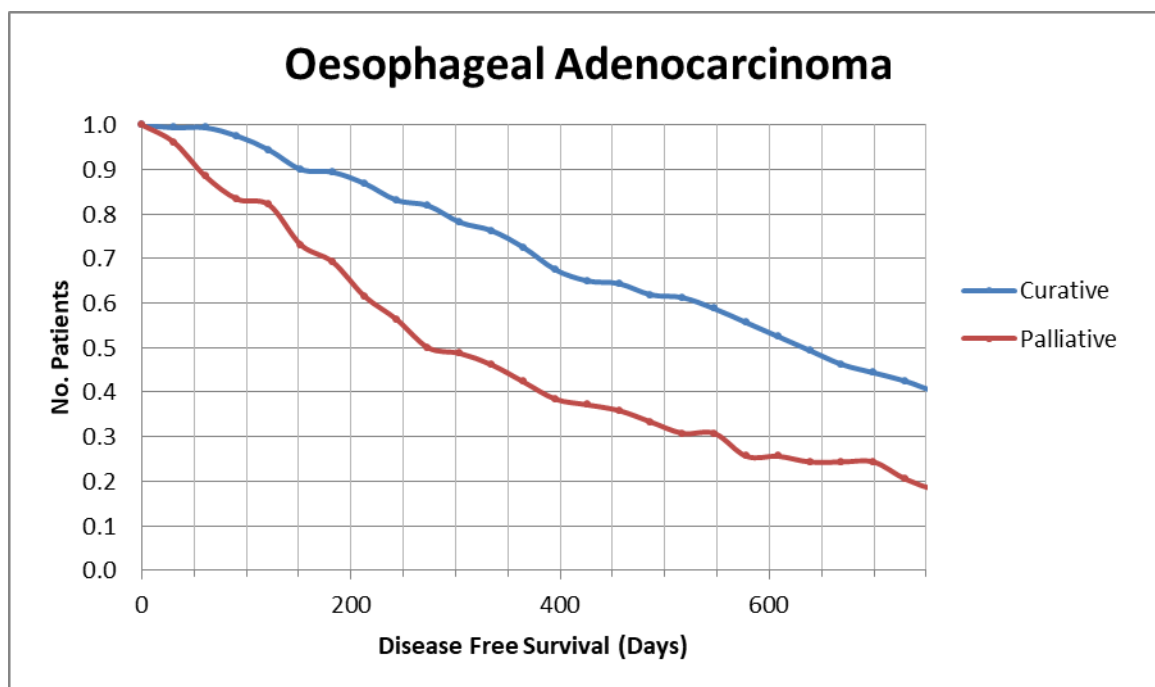
In summary, BSREM appears to create images with larger minimum and maximum SUV values and a larger range of grey-level values in the image and larger variations in local intensity, i.e. a more exaggerated difference between the maximum and minimum values. Features such as grey-level size zone matrices showed a high variation in results between OSEM and BSREM images and between individual patients with the same image reconstruction. Features describing pixel-pixel level variations, such as the distribution of short and long runs of pixels of a particular value appeared relatively robust to reconstruction method. In other words, features of this nature were, on average, very similar for both reconstruction methods and showed a low average difference between the values calculated for the two reconstruction methods (OSEM and BRSEM).

## 5.3 Machine Learning

### 5.3.1 Pilot Study

We performed an initial pilot study to test and develop the machine learning code using a simplified, well-defined dataset. A Kaplan-Meier style survival curve (for DFS) for all oesophageal adenocarcinoma patients (238 adenocarcinoma patients, 160 Curative, 78 Palliative) showed the split between curative and palliative patients.

**Figure 44 Kaplan-Meier survival curve for oesophageal adenocarcinoma**



*Kaplan-Meier style survival curve plotted to show disease-free survival for the curative and palliative groups.*

As anticipated, the DFS time of patients receiving palliative treatment was significantly lower than the curative group.

A machine learning (ML) prediction code was tested on the oesophageal adenocarcinoma group to evaluate the predictive power of 6 ML algorithms of determining whether a patient received curative or palliative treatment, based on their TNM score and whether the patient survived disease-free after 2 years (using 0 or 1 for failure or success). For each algorithm, a classification accuracy, confusion matrix, precision and recall were determined. We trained and tested the pilot algorithm with 190 patients and validated the pilot algorithm using a 20% subset of 48 patients (31 curative, 17 palliative).

**Table 21 Pilot machine learning code validation results**

Algorithm	Accuracy	True +ve	False -ve	False +ve	True -ve	Curative		Palliative	
						Precision	Recall	Precision	Recall
<b>Logistic Regression</b>	0.8125	31	0	9	8	0.78	1	1	0.47
<b>Linear Discrimination</b>	0.8125	31	0	9	8	0.78	1	1	0.47
<b>K Neighbours Classifier</b>	0.8333	30	1	7	10	0.81	0.97	0.91	0.59
<b>Decision Tree Classifier</b>	0.8333	30	1	7	10	0.81	0.97	0.91	0.59
<b>Gaussian Naive Bayes</b>	0.8125	31	0	9	8	0.78	1	1	0.47
<b>Support Vector Machine</b>	0.8125	31	0	9	8	0.78	1	1	0.47

*The accuracy, confusion matrix results, precision and recall for the pilot study for 6 machine learning algorithms.*

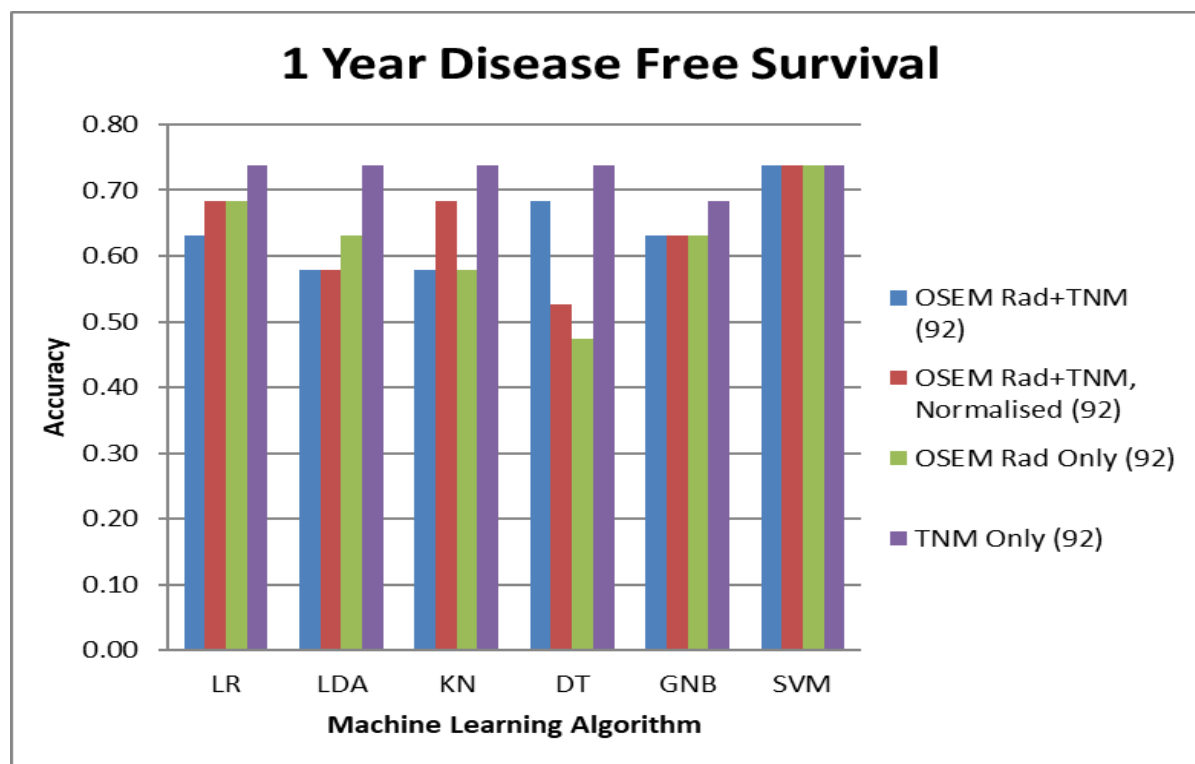
In this pilot example, Logistic Regression (LR), Linear Discrimination (LDA), Gaussian Naive Bayes (GNB) and Support Vector Machine (SVM) algorithms, all produced the same results, with an accuracy of 81.25% of correctly predicting whether a patient received curative or palliative treatment, based on the TNM score and 2 year DFS status. LR, LDA, GNB and SVM algorithms all correctly categorised all 31 Curative patients, based on their TNM score and 2 year DFS Status. LR, LDA, GNB and SVM also correctly classified 8 patients as palliative but incorrectly classified a further 9 palliative patients as curative, leading to a poor recall score of 0.47 in all cases.

The K Neighbours Classifier (KN) and Decision Tree Classifier (DT) both gave an accuracy of 83.3% and correctly classified 30/31 patients as curative and 10/17 patients as palliative. KN and DT incorrectly classified 7 palliative patients as curative and 1 curative patient as palliative.

### 5.3.2 Survival prediction with machine learning

We evaluated 6 machine learning algorithms and several datasets (see section 4.7.3) with each algorithm trained and tested using a larger dataset, then validated using a smaller set of previously unseen data. For each dataset and algorithm attempted, we measured the Accuracy, Precision, Recall and the number of True / False Positive / Negative results for the validation data. We explored DFS times of 90 days, 1 and 2 years however, too few patients recurred or expired within 90 days to give a statistically acceptable distribution of failures and successes and accuracy results for this experiment excluded. Figure 45 and Figure 46 show a comparison of the validation accuracy of each ML algorithm tested using different combinations of the data. A comparison was made using radiomic features and TNM scores, radiomic features only, TNM score only and then a dataset whereby each feature (radiomic and TNM score) was normalised to a value between 0 and 1.

**Figure 45 Comparison of accuracy for different machine learning algorithms**

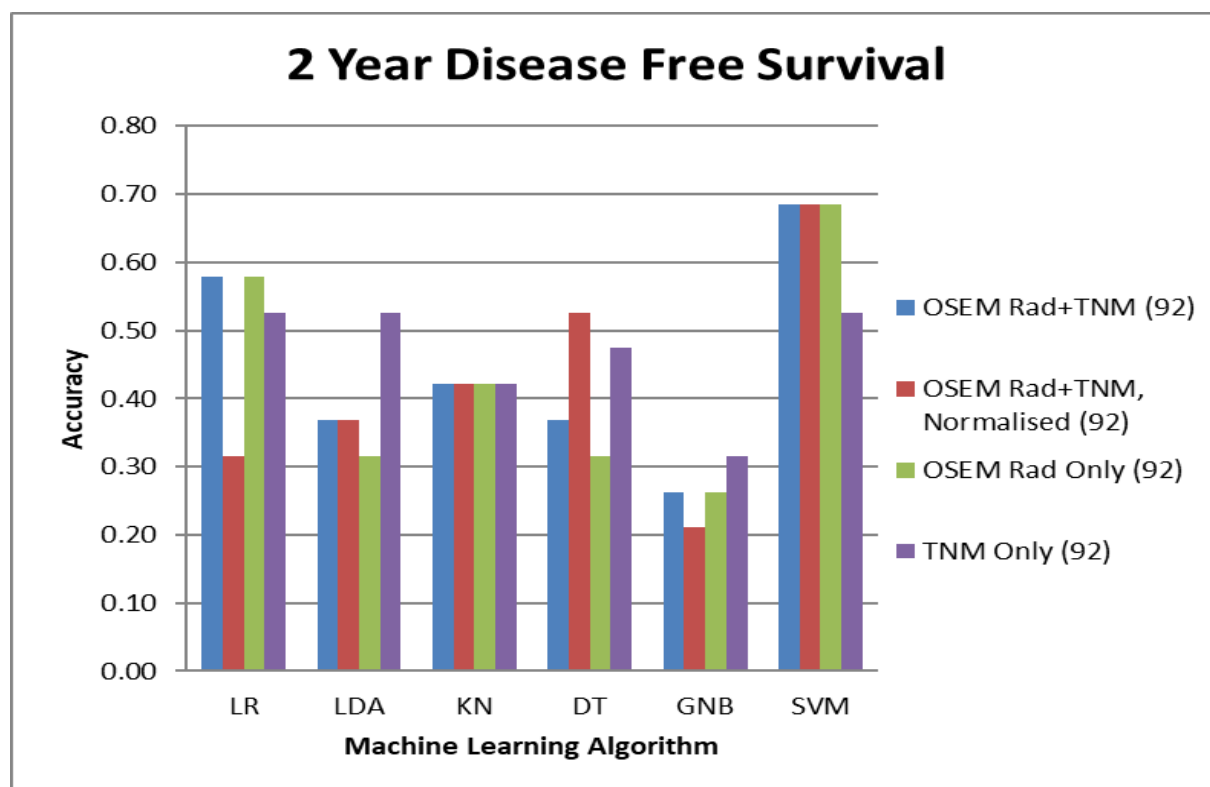


*The accuracy in predicting 1 year disease-free survival for Logistic Regression (LR), Linear Discrimination (LDA), Gaussian Naïve Bayes (GNB) and Support Vector Machine (SVM) algorithms for feature sets using all features (OSEM Rad + TNM), all features normalised (OSEM Rad + TNM, Normalised), Radiomic features only (OSEM Rad Only) and TNM Score (TNM Only).*

TNM score only for predicting 1-year DFS gave the highest overall accuracy, replicated for all algorithms except Gaussian Naïve Bayes (GNB) algorithm. For 1-year DFS, different algorithms benefited from different datasets, for example, the random forest (RF, decision tree classifier)

algorithm benefited from having radiomic features and TNM scores whereas Logistic Regression (LR) benefited from having less parameters. The Support Vector Machine (SVM) algorithm outperformed other algorithms for all scenarios, including using 1- and 2-year DFS time. Overall, algorithms performed better for predicting 1-year survival compared to 2-year survival. Conversely, the GNB algorithm predicting 2-year DFS on a normalised full dataset showed the poorest.

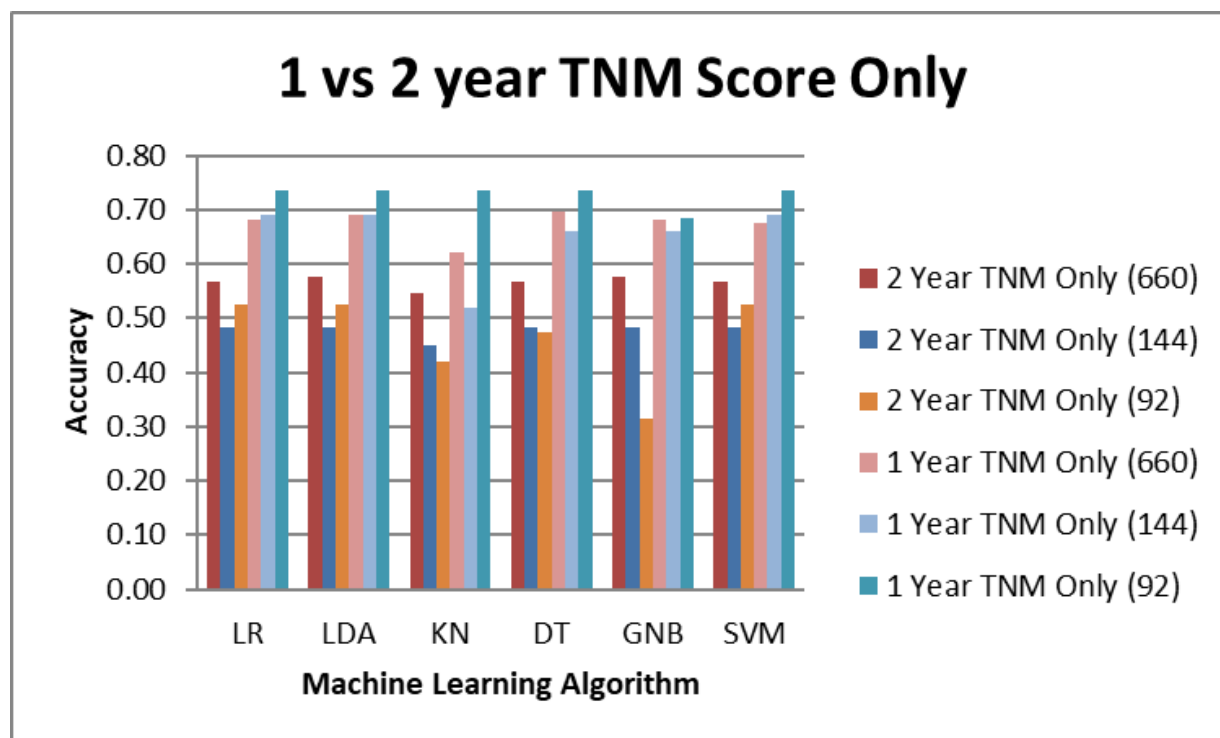
**Figure 46 Comparison of accuracy for different machine learning algorithms**



*The accuracy in predicting 2 year disease-free survival for Logistic Regression (LR), Linear Discrimination (LDA), Gaussian Naive Bayes (GNB) and Support Vector Machine (SVM) algorithms for feature sets using all features (OSEM Rad + TNM), all features normalised (OSEM Rad + TNM, Normalised), Radiomic features only (OSEM Rad Only) and TNM Score (TNM Only). Analysis for 92 oesophageal / OGJ adenocarcinoma patients with OSEM images.*

For 2-year DFS, the SVM algorithm outperformed others when radiomic feature data was available. Figure 47 shows improvements in accuracy with using only the TNM score. Accuracy was determined for 1- and 2-year DFS using the TNM scores from three datasets (see Table 10): the final dataset (dataset 3, 92 patients), the originally stratified group of oesophageal / oesophago-gastric junction adenocarcinoma patients (dataset 2, 144 patients), and the original total group of all treated upper GI cancer patients (dataset 1, 660 patients).

**Figure 47 Comparison of accuracy for different machine learning algorithms, 1 and 2 year disease free survival for TNM Score Only**



*The accuracy in predicting 1 and 2 year disease-free survival for Logistic Regression (LR), Linear Discrimination (LDA), Gaussian Naive Bayes (GNB) and Support Vector Machine (SVM) algorithms using only the TNM score for the complete upper GI cancer patient group (660 patients), all oesophageal / OGJ adenocarcinoma patients (144 patients) and the sub-group of this with 3D OSEM images, used for final radiomic and machine learning analysis.*

Again, machine learning algorithms performed better in stratifying 1-year DFS compared to 2 years. Accuracy was generally higher for datasets using less patients, e.g. the highest accuracy was found for 1 year survival using the group of 92 patients. The overall best performing algorithm was the support vector machine algorithm; Table 22 shows the number of true and false positive and negative results. In this context, a positive result was the correct prediction of a failed treatment (disease-recurrence / death within 1 and 2 years respectively).

**Table 22 False and True Positive and Negatives for SVM Algorithm**

	<b>1 Year Disease Free Survival</b>				<b>2 Year Disease Free Survival</b>			
Test	Rad + TNM	Rad Only	TNM Only	Rad + TNM Norm.	Rad + TNM	Rad Only	TNM Only	Rad + TNM Norm.
Total Success	14	14	14	14	6	6	6	6
Total Failure	5	5	5	5	13	13	13	13
Accuracy	0.74	0.74	0.74	0.74	0.68	0.68	0.53	0.68
True Positive	0	0	0	0	13	13	7	13
False Positive	5	5	5	5	0	0	6	0
False Negative	0	0	0	0	6	6	3	6
True Negative	14	14	14	14	0	0	3	0

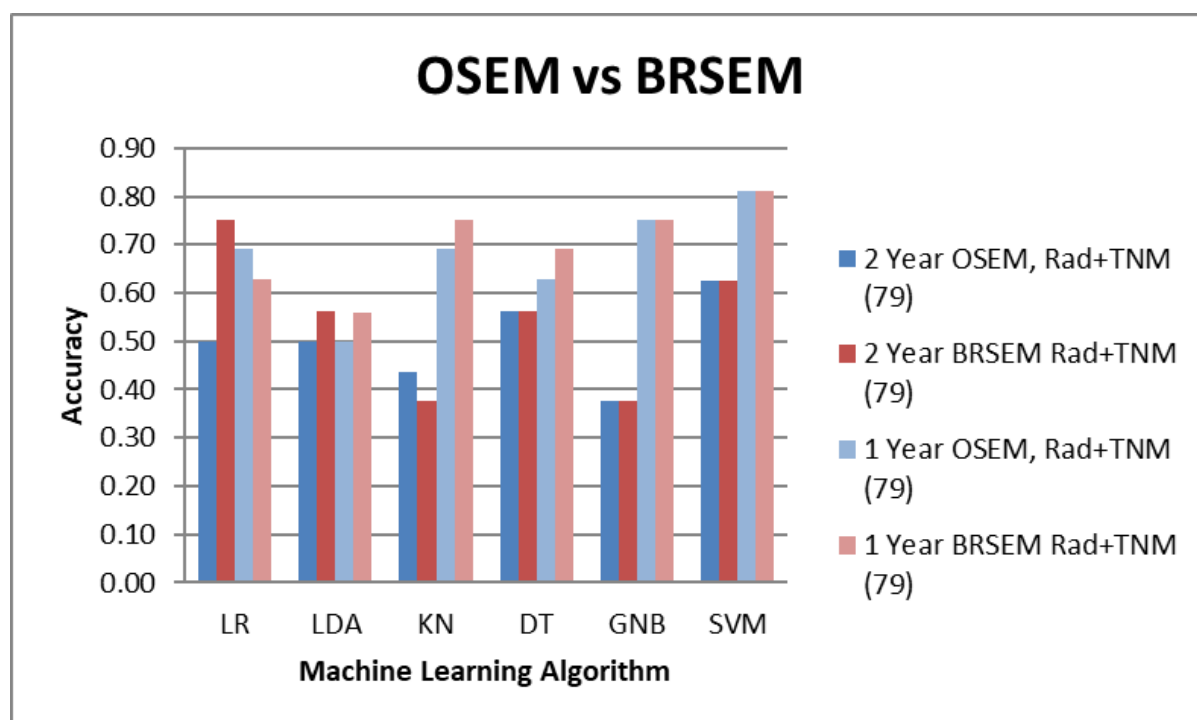
*Table to show the split of cases in the validation group for the SVM algorithm with the accuracy and confusion matrix values.*

In other words, when predicting 1-year DFS, a time point at which most patients in the validation group remained disease-free (successful treatment), the SVM algorithm correctly predicted all successes (true negative) but failed to correctly classify the failure treatments (false positive). Conversely, when predicting 2-year DFS, where most patients in the validation group had recurred or expired within 2 years, the SVM algorithm successfully classified all the failed treatments (true positive) but failed to classify patients who had a successful treatment.

### **5.3.3 OSEM vs BSREM**

We compared the performance of several machine learning algorithms when using OSEM and BSREM reconstructed images (Figure 48). The SVM algorithm again performed the best with an 81% accuracy for predicting 1-year survival (OSEM and BSREM images). The Logistic Regression (LR) algorithm performed with a 75% accuracy when predicting 2-year DFS (BSREM images). The Gaussian Naïve Bayes (GNB) algorithm performed the poorest with a 37.5% accuracy at predicting 2-year DFS (OSEM and BSREM images) but improved to 75% when predicting 1-year survival.



**Figure 48 Comparison of accuracy for different machine learning algorithms**

The accuracy in predicting 1 and 2 year disease-free survival for Logistic Regression (LR), Linear Discrimination (LDA), Gaussian Naive Bayes (GNB) and Support Vector Machine (SVM) algorithms. Shown for the subset of 79 oesophageal / OGJ adenocarcinoma patients with both OSEM and BRSEM 3D images and computable radiomic signatures.

For the best performing algorithm, Support Vector Machine (SVM), using OSEM or BRSEM images made no difference to the overall accuracy. The Logistic regression algorithm showed the largest improvement (with BRSEM): from 50% to 75% accuracy when predicting 2-year DFS. When using BRSEM images, the validation dataset (see section 4.7.2) LR algorithm identified 7/10 successful treatments and, more importantly, 5/6 failed treatments (Table 23). In summary, most machine learning algorithms remained relatively insensitive to reconstruction method except for the LR algorithm. In the context of 2-year-DFS prediction, the combination of BRSEM images and LR algorithm showed the most clinically relevant findings of all experiments, patient groups and algorithms tested.

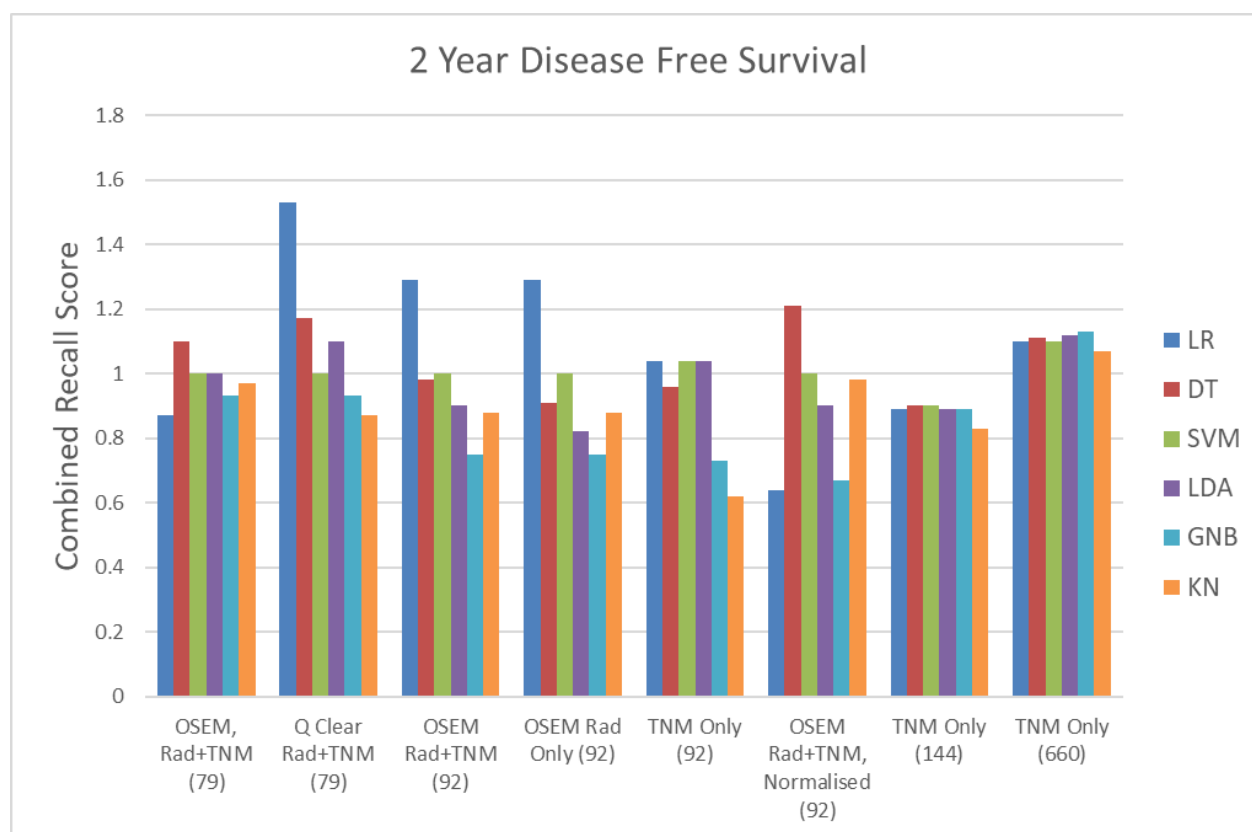
**Table 23 Validation set false and true positive and negatives for LR algorithm**

Test	1 Years Disease - Free Survival		2 Years Disease - Free Survival	
	OSEM, Rad+TNM	BSREM Rad+TNM	OSEM, Rad+TNM	BSREM Rad+TNM
Total Successes	13	13	6	6
Total Failures	3	3	10	10
Accuracy	0.69	0.63	0.50	0.75
True Positive	0	0	7	7
False Positive	3	3	3	3
False Negative	2	3	5	1
True Negative	11	10	1	5

Table to show the split of cases in the validation group for the Logistic regression algorithm with the accuracy and confusion matrix values. Shown for OSEM vs BSREM for 1 and 2 year survival.

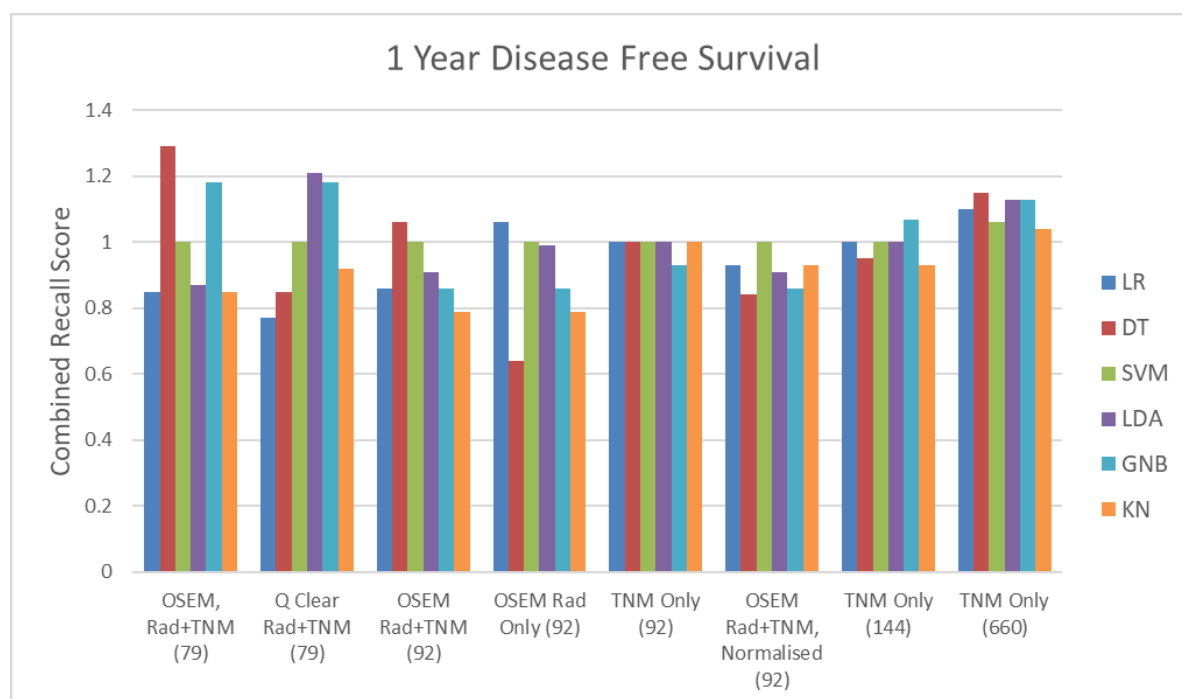
### 5.3.4 Summary of machine learning results

In summary of all machine learning experiments performed, we evaluated eight different datasets using three different DFS times, for six different machine learning algorithms. Accuracy as a measure, provides a useful summative measure for comparing the performance of individual algorithms in a variety of experiments, however, of most clinical relevance is the recall score. As aforementioned, the “recall” for each class gives the percentage of correct predictions in each class (successful and failed treatment). Clinically, the ‘best’ performing algorithm must be able to predict both success and failure sufficiently; Figure 49 and Figure 50 show the optimum algorithm (and dataset for prediction) as the sum of the recall scores for success and failure for 1- and 2-year DFS.

**Figure 49 Summary of recall scores for 2 year DFS**

*The combined recall score for predicting successful and failed treatments for 6 machine learning algorithms: Logistic Regression (LR), Linear Discrimination (LDA), Gaussian Naive Bayes (GNB) and Support Vector Machine (SVM). Plotted for 2-year disease-free survival only. Plotted for ML experiments using all datasets (Table 12)*

Clinically, when predicting 2 year survival, the logistic regression algorithm making predictions from features acquired from BSREM data gave a combined recall score of 0.7 (success) + 0.83 (failure) = 1.53; the highest score across any experiment. Comparatively, the poorest performance was for the K-nearest neighbours (KN) algorithm using the TNM scores only. Combining the scores further across all experiments, Gaussian Naive Bayes (GNB) and KN performed overall worst with Logistic regression performing overall the best across all scenarios.

**Figure 50 Summary of recall scores for 1 year DFS**

*The combined recall score for predicting successful and failed treatments for 6 machine learning algorithms: Logistic Regression (LR), Linear Discrimination (LDA), Gaussian Naive Bayes (GNB) and Support Vector Machine (SVM). Plotted for 1-year disease-free survival only. Plotted for ML experiments using all datasets (Table 12).*

The highest combined score for 1 year DFS was found for the Decision Tree Classifier analysis making predictions from features acquired from OSEM images giving a combined recall score of 0.67 (success) + 0.62 (failure) = 1.29. Comparatively, the worst performing algorithm was the decision tree classifier for OSEM images and radiomic data only; highlighting the importance of including the TNM scores when using DT algorithm.

## 5.4 **Summary of key results**

- We found that no radiomic feature correlated individually with the DFS time (in days). However, we did find that larger, more heterogeneously distributed tumours were associated with recurrence or expiration within 2 years of treatment; locally considered as an unsuccessful treatment.
- The most statistically significant difference between the treatment failure and treatment success groups was the T-score for 1-year DFS. However, once the p value had been adjusted using the bonferroni correction, we essentially found no statistically significant result but observed weak trends for features describing larger and more heterogeneously distributed tumours.
- BSREM images range of grey levels than OSEM images with radiomic features related to global differences in grey level exhibiting large differences between OSEM and BSREM images. Comparatively, features describing smaller local changes in the images appeared to remain relatively consistent between OSEM and BSREM images.
- For developing a machine learning algorithm to perform predictive analysis, we found that the SVM algorithm was able to correctly predict the most true negatives at 1 year and true positives at 2 years however, not with sufficient accuracy to be used clinically and not with a large enough validation dataset to be considered reliable.
- Machine learning performance was relatively unaffected by image reconstruction however, a Logistic Regression algorithm showed good predictive power particularly in identifying patients who survived, disease-free at 2-years when BSREM images were used. Of all experiments attempted, this scenario showed the most clinically relevant and promising machine learning result.

## 6 Discussion

---

This chapter covers a discussion and likely explanation of our results. The first section discusses the relative merits and pitfalls of our clinical data in comparison to other such published cohorts using similar patient groups and disease aetiologies. This section goes on to discuss the link between radiomic features and the prediction of overall survival following treatment for oesophageal / OGJ adenocarcinoma. The second section describes a comparison between OSEM and BSREM reconstructed images and discusses the effect of BSREM reconstruction on textural features. The final section describes a machine learning approach using the radiomic signature to predict overall survival at 1 and 2 years including a comparison of different algorithm performance.

### 6.1 Survival prediction with radiomics

#### 6.1.1 Clinical Data

Globally, squamous cell oesophageal carcinoma (SCC) is the dominant sub-type, accounting for over 85% of cases (Naghavi, et al., 2020; Watanabe, et al., 2022). The majority of studies reviewed in relation to this study (Table 1) described a patient cohort of either SCC only or a mixture of predominantly SCC with a smaller group of adenocarcinoma patients. Van Rossum et al (2016) cite the marked difference in pathologic response between adenocarcinoma (23-28% response rate) and SCC (49% response rate) which gives further support to investigating oesophageal and OGJ adenocarcinoma patients. Most patients in our cohort received Surgery in addition to adjuvant / neo-adjuvant chemotherapy; shown to improve 5-year survival from “23-34% with surgery alone, to 36-47% with the addition of neo-adjuvant chemo-radiotherapy” (Van Rossum, et al., 2016).

For the complete patient cohort, of 778 patients attending the Northern Oesophago-Gastric Unit multi-disciplinary team (NOGU MDT), our 2-year overall survival and disease-free survival (DFS) rates were 52.3% and 43.3% respectively. For all patients seen by the MDT (including palliative and curative, squamous cell, adenocarcinomas, and others), approximately half of patients survived 2 years from treatment (in any form) and approximately a fifth of those patients who survived, were alive but with disease recurrence.

Survival rate at 2 years for patients attending the NOGU MDT is broadly superior to the national average; CRUK (2017) data which quotes “almost 1 in 2 (46.5%) of people diagnosed with

oesophageal cancer in England survive their disease for one year or more” with a 10-year survival rate of 12%. However, it is important to note that patients referred to the NOGU MDT are diagnosed with upper GI cancer via another route into the healthcare system and the figures therefore do not capture any upper GI patients treated outside of this group. Furthermore, the CRUK figures relate specifically to oesophageal cancer whereas we have described all upper GI patients, to include gastric, OGJ and oesophageal cancer, which differ slightly in overall survival time (Sah, et al., 2019).

DFS curve data for our stratified group of curative adenocarcinoma patients showed more than 95% of patients remained alive and disease-free at 90 days; 70% of oesophageal and 75.4% of OGJ patients at 1 year; and 48.9% of oesophageal and 42.5% of OGJ patients at 2 years. These results highlight and agree with the basis for investigating upper GI cancer patients. We note that less than half of patients remain alive without disease-recurrence at 2 years, even with appropriate treatment stratification performed by a regional tertiary referral centre with survival rates above the national average. This clearly demonstrates that the successful treatment of oesophageal cancer remains a significant issue, particularly given the long recovery period following treatment of approximately 3-6 months. Improved stratification of patients whom are likely to benefit from treatment, may prevent patients from spending a large amount of their remaining time in recovery from a treatment, which may ultimately prove to be futile.

The DFS rate at 1-year for our stratified groups of adenocarcinoma patients receiving treatment (Figure 31, 70-75%) appears higher than the CRUK (2017) data. However, this may be because this includes only patients who, due to a variety of factors such as disease staging, lifestyle etc., are expected to do well following treatment and were chosen to receive curative surgery. Our main experimental cohort did not include any palliative or untreated patients. A further caveat to directly comparing our survival data with CRUK (2017) is that this is quoted for the survival of oesophageal cancer of all stages and subtypes whereas our data is for oesophageal and OGJ adenocarcinoma patients, treated with curative intent, and in relation to DFS, rather than survival alone.

In this study, our data was limited to a 2-year period because of two conflicting restraints on the data: the availability of 3D PET data (mid-2016 onwards) and performing the data download at a sufficient time into the future to allow a minimum of 2 years follow up. We have defined 2-year DFS as patients who have received treatment and then had at least 2 years without death or disease-recurrence. For patients from the NOGU, recurrence was determined using follow up imaging with either CT or PET/CT, with recurrence identified as either metastatic spread, nodal involvement, secondary cancer, or recurrence of the primary tumour. Of the 18 studies identified during systematic review, most studies determined the efficacy of radiomic signatures and / or artificial

intelligence predictive models in relation to pathologic response, which is a subtly different end point to DFS time sampling (Yip, et al., 2016; Tan, et al., 2013; Cao, et al., 2020; Nakajo, et al., 2017). Whilst both approaches nominally used imaging to identify the presence of recurrent disease, pathologic response is determined at a different time from patient to patient whereas DFS at 2 years is fixed. However, this is also dependent on the timing of the imaging and review in the clinic, therefore the 'date of recurrence' may not be exact. For example, the date recorded as being disease-free is dependent on either when the patient was imaged or last seen in clinic and there may have been up to 6 months where the patient has 'recurred' but not been seen in clinic and therefore this detail is not captured. If the patient was to be imaged or reviewed more regularly (e.g. daily / weekly), it may be possible to determine a more definite 'moment' of recurrence; however, this would be a significant and non-essential use of clinical resources and therefore, routine follow up dependent on disease aetiology is deemed sufficient locally for effective clinical practice.

In our study, we chose DFS, rather than overall survival (Van Rossum, et al., 2016) or pathologic response (Tan, et al., 2013) because locally, 2 years of survival without recurrence following curative treatment is considered a successful treatment. Progression-free (disease-free) survival analysis has been performed for a variety of other cancers such as non-small cell lung cancer (Cook, et al., 2013) but to date, has not been explored in this context for oesophageal cancer. Furthermore, the main goal of this project was to determine whether there were any features in pre-treatment PET imaging to suggest early recurrence or expiration following treatment. The ultimate goal was to provide additional information for patients and clinicians to aid their decision to embark on significantly invasive treatment with a long recovery time and guide a more personalised, patient-specific follow up schedule. For example, if features were found on the initial scan, which raised the possibility of an unsuccessful treatment, the clinical partnership may revise their treatment plan or provide more regular follow up appointments and imaging for that patient, moving towards a more personalised medicine approach (Vincente, et al., 2020).

Our cohort was primarily of patients seen through the tertiary referrals centre serving the north east of England population, covering patients from Teesside, County Durham, Tyne and Wear, Northumberland and Cumbria. To our knowledge, this is the first study of its kind for this population group and one of only a limited number using UK based populations (Ypsilantis, et al., 2015; Foley, et al., 2018). The nationality of the population is important to note, as the prevalence of different disease aetiologies is different depending on the geographical location, for example, SCC is the most prevalent in China whereas adenocarcinoma is most prevalent in North America and Europe (Zhu, et al., 2019).



As described in section 4.3.2 and 5.1.2, 29 studies were excluded due to absent 3D PET data, 10 studies were not found on the regional PACS systems and 13 patients yielded un-calculable radiomic signatures; a loss of 52 studies, leaving 92 patients for complete investigation. From the systematic review, we identified that more, larger cohort studies were required. Of the papers reviewed, a median of 58 patients (Nakajo, et al., 2017; Paul, et al., 2017) were used with an average of 112 (20 – 548) patients (Tan, et al., 2013; Yang, et al., 2019). Our study is therefore the 6<sup>th</sup> largest study investigating PET imaging, radiomics and oesophageal cancer and the 2<sup>nd</sup> largest to investigate adenocarcinoma only (Van Rossum, et al., 2016).

In summary of this section, we have shown comparable survival rates both to published works and to the national average for the UK. Our study was among the larger patient cohort studies to be published to date and took the more novel approach of reviewing an exclusively adenocarcinoma group using 2-year DFS to define clinical success. Further work in this area should include a larger cohort which, within the confines of extending this work, could include either the same group of patients with, at the time of writing, now a minimum of 4 year's follow up time and / or a larger cohort of patients continuing with a minimum 2 year follow up period.

### **6.1.2 Radiomics and clinical features**

Several authors have shown links between radiomic features acquired from PET images of oesophageal cancers and the likelihood of a successful response to treatment (Sah, et al., 2019). This section discusses our key findings in relating radiomic features of pre-treatment PET images for oesophageal and OGJ adenocarcinomas to DFS. This section will describe where the results of this study fit with the work from previous authors and detail several areas in relation to radiomics and PET imaging where there remains a lack of consensus, including differences in scanner, imaging methods, tumour threshold and definition of clinical success.

The wider literature largely agrees that more heterogeneously distributed tumours are associated with poorer outcomes (Ypsilantis, et al., 2015) however, there remains a lack of consensus on specifically which radiomic feature or combination of features is most closely associated with successful / failed treatment. We found that larger, more heterogeneously distributed tumours were more closely associated with a failure of treatment (recurrence or expiration within 1 / 2 years of treatment). Specifically, we found that Total Lesion Glycolysis (TLG, discretised and non-discretized), N-score, Volume (ml and voxels) and several radiomic features were greater than 20% higher (on average), for patients who failed treatment at 1 and 2 years (Figure 36). We found that the most

statistically significant feature was the T-score only, which was on average 10.3% higher for failed treatments.

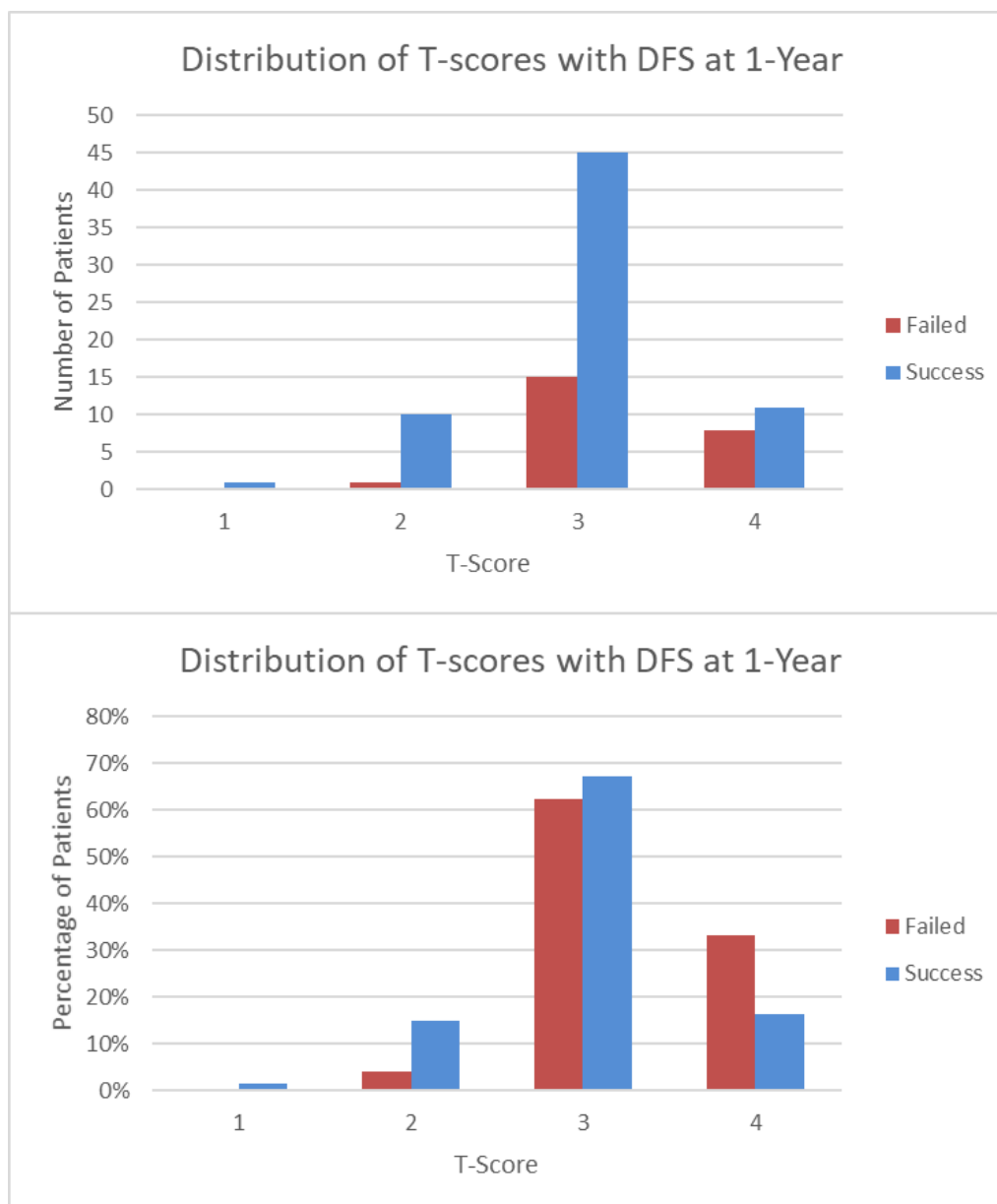
As an aside, we found a strong correlation between some higher order radiomic parameters and commonly used parameters, e.g. between SUVmax and GLZLM\_SZHGE, GLRLM\_HGRE and GLZLM\_HGZE. However, this is likely because tumours with a high SUVmax are also tumours with large areas of high uptake; as described by grey-level matrices, which relate to areas of high grey level (Galloway, 1975).

The clinical standard since PET was first introduced has been for clinicians to review the SUVmax (Eisenhauer, et al., 2009) and many studies have since reported on the utility of SUVmax for a variety of tumours and applications (Cherry, 2006; Czernin, et al., 2013). However, more recently, there has been increased interest in the use of tumour volume and total lesion glycolysis (TLG). Van Rossum et al (2016) “suggest that spatial image information, such as metabolic tumour volume (MTV), total lesion glycolysis (TLG), tumour shape, and texture features, provide more useful information than SUVmax”. Hatt et al (2011) have also shown for squamous cell oesophageal carcinoma patients, treated with chemotherapy, which TLG and tumour volume were able to confidently separate responders from non-responders. Hatt et al (2015) later linked larger volumes with poor prognosis for a mixed group of lung and oesophageal cancer FDG PET images however, citing that the most complementary prognostic gains were made on much larger volumes; limited application for oesophageal cancer because they are inherently a smaller volume tumour. Foley et al (2018) also showed that TLG from pre-treatment imaging was associated with overall survival in a group of oesophageal SCC and adenocarcinoma patients. We found that both TLG and tumour volume were associated with failed treatment. Recall that TLG is the tumour mean SUV multiplied by volume, therefore, since both TLG and volume are separately strongly associated with failed treatment, it is likely that the volume term is the main contribution driving the link between TLG and failed treatment, suggesting that larger tumours generally perform poorer. We also found little difference in the average value of SUVmean between successful and failed treatment, giving further evidence to the link between large tumour volumes and failed treatments.

The TNM scoring system remains the most important prognosticator prior to treatment in clinical practice. Indeed, although to a lesser extent, so does the SUVmax of the primary tumour (Van Rossum, et al., 2016). In this study, we were not aiming to replace the TNM scoring system, rather provide additional information to support this scoring system. Note that we have not included the M score in this analysis because all patients with metastatic disease, i.e. an M score >0, received palliative treatment and were therefore excluded from the study. Reassuringly, we found that the T score was

~10% higher in patients who failed treatment at 1 year however, this should be interpreted cautiously because the T-score in our patient cohort was comprised of scores 2-4 and the majority of patients with a T score of 4 received palliative treatment (excluded from our study).

**Figure 51 Tumour scores for patients in dataset 3**



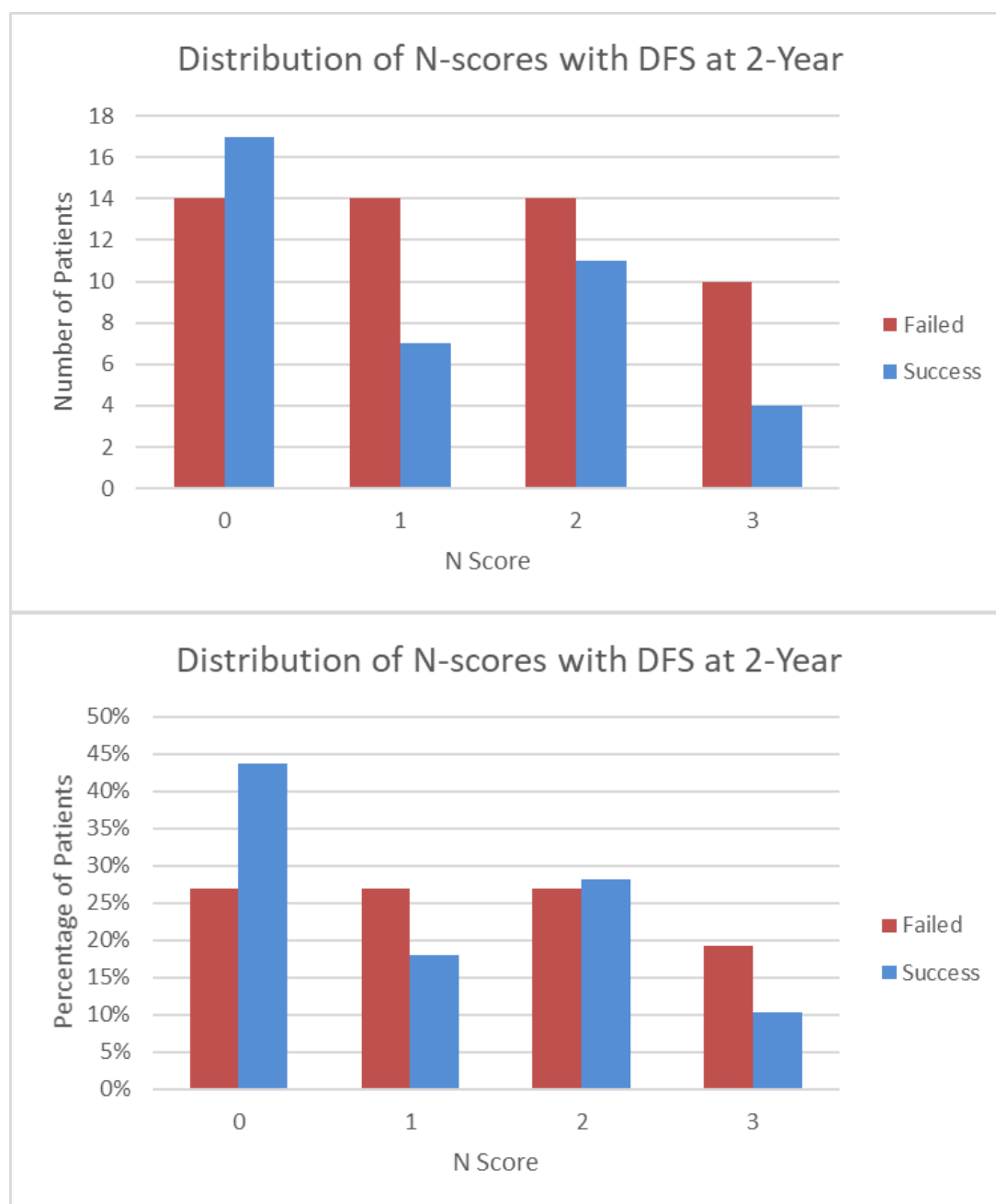
*The number and percentage number of patients with tumour grade (T) scores in dataset 3, split for successful and failed treatments. Plotted, for context, for patients with disease-free survival (DFS) at 1 year.*

In our cohort, we included 19 patients with a T4 tumour. Whilst there was no presence of metastases and were considered operable, patients with a T4 tumour still had more advanced disease. Figure 51 shows that a greater proportion of patients with T4 tumours failed treatment and a greater proportion

of patients with a T2 tumour had a successful treatment which is the likely driver of the observed difference however both categories unfortunately contained low numbers of patients.

Nodal involvement across all cancers, including oesophageal, is an important prognostic factor (Van Rossum, et al., 2016; Chirieac, et al., 2005) however, despite the availability of this information, oesophageal cancer patients continue to receive inaccurate individual survival prediction and disease control.

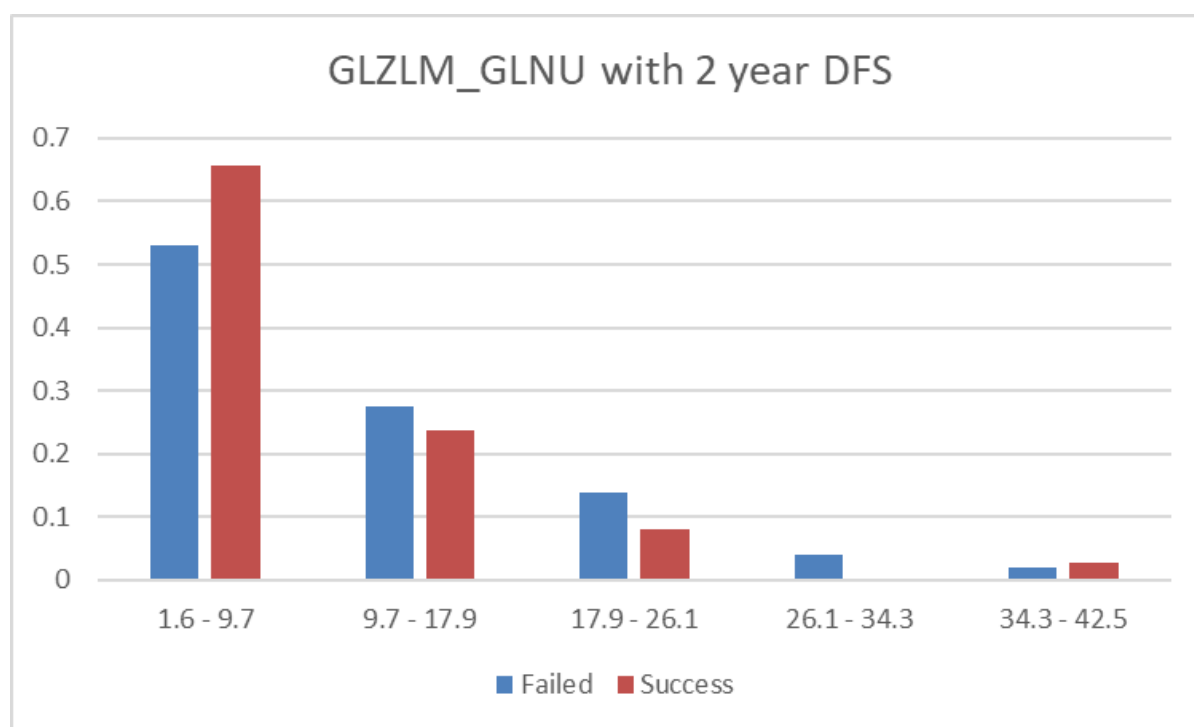
**Figure 52 Nodal scores for patients in dataset 3**



*The number and percentage of patients with nodal (N) scores in dataset 3, split for successful and failed treatments. Plotted, for context, for patients with disease-free survival (DFS) at 2 years.*

Again, reassuringly, we demonstrated 31% higher N-scores for patients who failed treatment at 2 years. However, patients were given N-scores of 0 – 3 and we would expect that patients with more nodal involvement would generally be more likely to recur because they have more widespread disease. Recall, and N-score of 0 = no lymph node involvement; 1 = 1-2 cancerous lymph nodes, 2 = 3-6 local cancerous lymph nodes and 3 = 7 or more cancerous nodes (see section 3.1.1). Figure 52 shows the distribution of nodal scores for treatment outcome at 2 years which shows a greater proportion of patients with more nodal involvement (higher N-score) fail treatment at 2 years. Inherently, with more nodal involvement, there is a greater likelihood of disease spread and therefore a greater likelihood of recurrence. In opposition to this, there appeared to be a higher proportion of patients with N1 who failed treatment however, this is likely an effect seen secondary to low numbers.

**Figure 53 Percentage split of failed / successful treatments for GLZLM\_GLNU**



*The percentage of patients with Grey-level zone length matrix grey level non-uniformity (GLZLM\_GLNU) in dataset 3, split for successful and failed treatments. Values have been plotted according to equally split bins covering the full range of results. Plotted, for context, for patients with disease-free survival (DFS) at 2 years.*

Shown in Figure 53 is the percentage split between failed and successful treatments at 2 years; recall a higher value of GLNU is related to a more heterogeneously distributed tumour. Figure 53 illustrates the potential clinical benefit, i.e. that tumours with higher values of non-uniformity (more heterogeneously distributed tumours) are generally in greater proportion for failed treatments.

Studies investigating the value of radiomic features for the prediction of disease response and survival for oesophageal cancer are summarised in Table 1. Tixier et al (2011) found that, for a mixed group of

oesophageal SCC and adenocarcinoma patients, that variability in the Grey Level Zone Length (Size zone) Matrix (GLZLM) could differentiate responders and non-responders. Nakajo et al (2017) found similar results when predicting response but found that no individual feature was predictive of overall survival. Hatt et al (2015), in a larger study of lung and oesophageal cancers, also found links between GLZLM variability but to a lesser extent for oesophageal cancer, largely due to generally smaller volumes with which to calculate such features from. Other authors, investigating response prediction, rather than overall survival, linked grey level co-occurrence matrix variabilities with poorer outcomes (Tan, et al., 2013; Van Rossum, et al., 2016). We found that grey level zone length (Size zone) matrix zone length non-uniformity (GLZLM\_ZLNU), GLZLM grey-level non-uniformity (GLZLM\_GLNU), and grey level run length matrix run length non-uniformity (GLRLM\_RLNU) were all, on average, greater than 20% higher for patients who recurred or expired within 2 years.

In summary, whilst there are several considerations when applying these results to other datasets and to PET images for patients from other centres (see section 6.4.2), we have found, consistent with other groups, that larger and more heterogeneously distributed tumours were associated with expiration or disease recurrence within 2 years of receiving treatment intended to achieve a definitive cure. Our results suggest that radiomic information would be of benefit in clinical practice when reviewing PET images for oesophageal / OGJ adenocarcinoma patients. We recommend that such patients with larger tumours, larger TLG values or greater GLZLM variabilities may benefit from either more regular follow up or more frequent imaging following treatment. From here, the options for such patients may include a more informed discussion on the likely outcome of their treatment or a revised follow up treatment plan.

## 6.2 OSEM vs BSREM

Several authors have investigated the effect of image reconstruction on radiomic features (Galavis, et al., 2010; Hatt, et al., 2013; Ger, et al., 2019) and Shiri et al (2017) provided a review detailing the robustness of radiomic features to imaging parameters. However, to the best of our knowledge, none have explicitly reported on the effect of GE's "Q.Clear" (BSREM) algorithm (Ross, 2014) on radiomic signatures, in any clinical setting. In this comparison, we evaluated the radiomic signature using identical (both in size and location) regions of interest placed separately over the primary oesophageal / oesophago-gastric junction (OGJ) tumour of OSEM and BSREM images. The aim was to investigate which features (if any) were statistically significantly different between the reconstructions (either positive or negatively skewed) and which features remained essentially unchanged. The first of our

results to cite is that reassuringly, all shape-related features remained identical for both the OSEM and the BSREM reconstructed images, assuring ourselves that the regions of interest had copied effectively between image datasets with no scaling or localisation issues.

We found that the largest, statistically significant differences were for SUVmin, GLCM contrast (the variance in grey-level co-occurrence matrix) and NGLDM contrast (variance in variations between regions, the neighbourhood gray-level difference) (Table 16). This suggests that the BSREM reconstruction appears to create a larger dynamic range of grey values across the whole image, likely as a result of two factors: performing many more iterations to achieve “convergence” and the inclusion of PSF and TOF in the reconstruction algorithm compared to OSEM alone (which uses 2-3 iterations) (Reynes-Llompart, 2019; Ross, 2014).

The linear dependence of pixels in the co-occurrence matrix (GLCM Correlation) and the level of spatial rate of change in intensity (NGLDM Coarseness) was found to be statistically significantly lower (-16.5% and -29.3% respectively) in BSREM images than for OSEM images. Recall ‘coarseness’ describes the differences between neighbouring voxels in adjacent axial images (Ypsilantis, et al., 2015). This implies that whilst BSREM creates a larger global dynamic range, it also creates images where the differences between individual neighbouring pixels is larger, again, likely a result of driving the iterations to effective convergence but with the noise controlling parameter not fully compensating for this on a pixel-to-pixel level. In other words, whilst BSREM appears to qualitatively improve visual image quality, on a pixel-to-pixel level, the image itself is more non-uniform. Furthermore, the inclusion of both TOF and PSF in the reconstruction (compared to OSEM alone) is likely also contributing to an enhancement in the range of individual pixel values. Several previous authors have linked heterogeneously distributed primary tumours with overall poorer outcomes therefore, caution should be taken in making inferences about the patient’s likely survival with a heterogeneously distributed tumour when viewing BSREM images alone (Sah, et al., 2019; Deantonio, et al., 2022).

In contrast, several features emerged as being, on average, <5% different between OSEM and BSREM, with a low variance of difference percentages (Table 18). Features associated with the grey-level run length matrix such as the distribution of short runs (GLRLM\_SRE), percentage of homogeneous runs (GLRLM\_RP) and the length of homogeneous runs (GLRLM\_RLNU), all appeared relatively unaffected by reconstruction, which is consistent with the findings of Reynes-Llompart et al (2018) who found that the same parameters remained consistent when varying the  $\beta$  value for BSREM but made no comment on the consistency between reconstruction algorithms.

Not considered here is the effect and comparison with altering the noise controlling penalty term, with a  $\beta$  value fixed at 400, however, this has already been explored: Reynes-Llompарт et al (2018) describe how “an increase in the  $\beta$  value tends to homogenize the lesion”. Further to this work, Reynes-Llompарт et al (2018) found that BSREM appears to increase contrast ratios and decrease background noise. For comparison we found an average 21.7% and 25.9% increase in Discretized SUVmax and conventional SUVmax respectively with BSREM which is consistent with other works and consistent with performing iterations to effective convergence (Reynes-Llompарт, 2019). Similarly, other parameters linked with SUVmax such as SUVpeak were also higher in BSREM images. However, it is also important to note that, unlike OSEM alone, the BSREM algorithm inherently includes a Point-spread function (PSF) and Time-of-flight (TOF) correction. TOF and PSF in their own right, also increase SUV values, so it is likely that these correction factors are a contributory factor in enhancing pixel values.

One of the challenges of comparing BSREM with an OSEM reconstruction is that similar quantitative metrics (such as contrast to noise) can be achieved by varying the  $\beta$  value for BSREM and varying the iterations, subsets, and post filter for OSEM, hence this project focussed on directly comparing patient images acquired using clinically used, pre-validated parameters. Indeed, by using a higher  $\beta$  value, we create ‘smoother’ images and therefore limit how different adjacent pixels can be. The fact that we found higher average GLCM contrast and GLCM Dissimilarity values, suggests that the  $\beta$  value has been set to increase pixel values rather than reduce image noise (recall GLCM contrast and dissimilarity give the local variation in grey-level values and a higher value of this indicates greater variation).

In comparing with other published works, the main challenge facing PET radiomics research is the lack of standardisation at present; several authors have sought to address this such as Lambin et al (2017) however, since radiomic application to PET imaging is a relatively recent area of interest, many previous studies have been performed out of adherence to a strict and reproducible image acquisition and image analysis protocol. For this study, whilst analysis methods were performed using established techniques and programs, the clinical image data was from a retrospective cohort and was therefore not performed according to Lambin et al’s (2017) prescriptive method.

In summary, when comparing images reconstructed with OSEM and BSREM using clinically validated parameters, BSREM appears to create images with qualitatively improved visual image quality and a larger dynamic range, with higher lesion contrast and lower background noise. This was likely due to combining iterating the image to effective convergence with TOF and PSF to achieve larger individual pixel values but with the  $\beta$  value noise penalty term reducing the overall image noise. Radiomic features describing pixel run length uniformities appeared relatively stable to different



reconstructions however BSREM appeared to exaggerate the heterogeneity of the primary lesion, therefore caution is advised if using BSREM images to draw conclusions about an individual patient's relative outcome following treatment. In other words, BSREM may create images with a level of heterogeneity, which at worst could falsely raise the possibility of a failed treatment or at best, could enhance the level of heterogeneity to reveal a patient's true vulnerability to failed treatment. This question has been further explored in section 6.3.3.

## 6.3 Machine Learning

### 6.3.1 Pilot

The aim of the pilot study was to test and develop a “proof of concept” machine learning code using a well-defined dataset, specifically, a machine learning code to predict whether patients were treated with curative or palliative treatment, based on their TNM score. The pilot study was successful in producing a working machine learning code with the main results being that a code was developed to, with reasonable accuracy, separate palliative from curative patients.

Figure 44 showed a clear split between curative and palliative patients in terms of their DFS time with patients in the curative group exhibiting much a higher survival rate. For machine learning analysis, asking an algorithm to predict whether patients would be curative or palliative treatment based on the TNM score was inherently biased because all patients with an M score of 1 are offered palliative treatment. However, not all patients offered palliative treatment have an M score of 1, therefore, the algorithm’s predictive power was not expected to be 100% accurate. Most patients with a T score of 4 are offered palliative treatment however, some patients with a T score of 4 may be treated curatively. Whilst metastatic spread of any kind ( $M = 1$ ) dictates palliative treatment only, one key factor not captured is the exact location of the tumour. The tumour location, provided there is no metastatic spread, can dictate whether a stage 3 tumour should be considered palliative and a stage 4 tumour could be considered curative. It is expected therefore that if, for example, all palliative patients had an M score of 1 (and all curative patients had an M score of 0), that the algorithm would perform more accurately because the M score indicated in every case whether a patient was curative or palliative. The algorithm was challenged (and appears to incorrectly classify) patients where the M score was 0 but were still given palliative treatment, perhaps because of the relative operability of the tumour.

The fact that a machine learning algorithm was able to predict with up to 83% accuracy (K-Neighbours Classifier and Decision Tree Classifier) whether a patient should be offered curative or palliative treatment based on the TNM score does not help clinicians. In clinical practice, it is the TNM score, in conjunction with several other factors which informs the decision on how to treat the patient. This pilot however, did allow for the initial building and understanding of a machine learning program to test 6 different algorithms in the context of using large datasets to predict outcomes.

Whilst we were able to produce ‘acceptable’ results from the pilot machine learning code, the pilot study highlighted a note of caution around the terms used to describe the efficacy of each algorithm; specifically, the use of the terms ‘accuracy’. Accuracy is commonly used in machine learning

classification problems, particularly in binary classifications; for example, predicting whether a patient was offered curative or palliative treatment and predicting whether a patient will survive disease-free at a given time point.

### 6.3.2 **Survival prediction with machine learning**

Our main experiment was to investigate whether machine learning algorithms could be used to predict overall survival, based on the radiomic signature from the primary tumour and the TNM score.

We evaluated ML performance for a variety of combinations of input data to identify whether any particular algorithm performed better given a particular input dataset: radiomics + TNM Score, radiomics + TNM Score (normalised), radiomics only and TNM only. Moreover, we wanted to explore what value was added in statistical predictive power by using a radiomic signature over a conventional TNM score for predicting patient outcomes. We found that ML algorithms generally performed better when predicting 1-year over 2-year survival and that the support vector machine algorithm was generally superior in most applications. Furthermore, predicting DFS at 1-year with the TNM score alone was better than or equal to the accuracy when predicting using radiomics features also. In exploring the effect of predicting with TNM only, we found that the validation accuracy of trained and tested algorithms improved when compared to a 1-year DFS end point and by using less patients. NB we tested this with 3 cohorts: all upper GI cancer, 660 patients; all adenocarcinoma with surgery, 144 patients; and all adenocarcinoma with PET radiomic signatures successfully downloaded, 92 patients.

One important factor affecting our accuracy measurement was the split of the data, i.e. using 80% of the data to test and train the algorithm (see section 4.7.2 for details) keeping 20% to validate the algorithm accuracy. This meant that in a set of 92 patients, the algorithm was trained and tested using 73 patients and then validated on 19 patients. Ideally, this analysis would be performed on a larger dataset. As discussed in section 4.7.2, the model determined for each machine learning algorithm was done based on a single iteration of the training data and was not updated with subsequent datasets. More complex machine learning algorithms, with much larger datasets, would be able to perform the ‘training – testing phase’ multiple times and update the parameters each time. Due to the constraints of both the dataset size and the project, this was not explored however further work, with a larger dataset should include more sophisticated machine learning analysis.

The number of patients in each class was unequal with a splits of 14 / 5 and 6 / 13 (success / failure) for 1- and 2-year DFS respectively; a ratio of 3:1 and 1:2 respectively. Whilst not considered “severe”, this split of the validation dataset was considered to be a “binary imbalanced classification problem” (Brownlee, 2021; Brownlee, 2019). However, when compared to the test-train data split of 53 / 20 and 33 / 40 (success / failure) for 1- and 2-year DFS respectively; the imbalance becomes less

pronounced for 2-year DFS, a “slight imbalance” (Brownlee, 2019). A “slight imbalance” is generally not considered to be too problematic and is widely accepted as a data split of up to 4:6 (Brownlee, 2019). However, as we found with the majority of experiments, particularly for SVM, the algorithm was biased towards the “majority” class and therefore, the accuracy measure here is misleadingly optimistic. In the case of 1-year, and indeed the 90-day DFS, the algorithms were heavily weighted towards successful treatment (particularly for 90-day). The issue with this in machine learning is that the algorithm is weighted and trained to recognise the characteristics associated mainly with a successful treatment with less weight given to learning the features associated with the minority class (failed treatment). The net result of this, was that the algorithm was better at selecting the successes but failed to classify any of the failures. Our algorithm trained on the 1-year DFS data (with a classification imbalance) was less effective at recognising the likelihood of treatment failure. This effect was most significant for the 90-day dataset, which contained a severe classification imbalance of up to 1:35 in dataset 2 (Table 10).

When predicting 2-year survival, the more clinically useful metric, our algorithm maintained a 68% accuracy for the SVM algorithm and was superior when radiomic features were included in addition to a TNM score. However, the classification imbalance was likely still an issue because, e.g. for predicting 2-year DFS with radiomics and TNM scores, the algorithm successfully predicted the majority class (failed treatment in this case, 13/13 True positive) but failed to classify any of the minority class (successful treatment, 6/6 False negative). Clinically, this is therefore not useful to clinicians and caution should be used when using vector-style metrics to separate similar data.

The overall performance of the Gaussian Naïve Bayes (GNB), K-Nearest Neighbours (KN) and Linear Discrimination Analysis (LDA) was poorest across all experiments. GNB architecture assumes that all values in the dataset are continuous and both GNB and LDA assume a normal distribution; since our data was not normally distributed, this has likely contributed towards the poor performance. KN also performed poorly however KN generally works best for a small number of variables and requires significant optimisation to the problem at hand. It is possible that a KN algorithm, using a smaller subset of features and an optimised K-value may yield superior results however such exploration was beyond the scope of this project.

In terms of statistical methodology, Ypsilantis et al (2015) have also compared the accuracy of several ML methods to predict treatment response for oesophageal cancer patients, mixture of squamous cell and adenocarcinomas. For an SVM algorithm, Ypsilantis et al (2015) reported an accuracy of 55.9% for the prediction of treatment response from the radiomic features of a primary tumour. The reported accuracy increased to 60.5% with principal component analysis, i.e. when the algorithm was trained using only the 10 most important features. Of the machine learning methodologies used, Ypsilantis et

al (2015) found that a Gradient-boosting (GB) algorithm performed best however we have not tested this algorithm on our data as we aimed to explore the accuracy of more, to date, untested algorithms. Recall the SVM algorithm involves the separation of any particular feature by mapping training samples to a higher dimensional space and finding hyperplanes to linearly separate the two outcomes: success and failure (see section 3.4.2 and Figure 22). Whilst our SVM algorithm performed better than results found by Ypsilantis et al (2015), a direct comparison was not possible; our clinical datasets and endpoints were different: Ypsilantis et al (2015) have used a mixture of oesophageal cancer aetiologies and used pathologic response to chemotherapy whereas we've used specifically oesophageal / OGJ adenocarcinomas treated with surgery and have investigated DFS. This highlights the need for more standardised, prospective studies in this area and that the results can depend on the initial patient database and outcome (Desbordes, et al., 2017).

Ypsilantis et al (2015) noted an improvement in algorithm accuracy when principal components analysis was used. Desbordes et al (2017) found that even when using a feature pre-selection method, there remained a high number of uncorrelated features and, in general, there was a non-linear relationship between individual features and the patient outcome. Based on our correlation matrix, we similarly found no correlation between any feature tested and DFS. Desbordes et al (2017) compared the performance of a Random Forest (RF) and SVM algorithms, concluding that RF was superior and has yielded the most promising results in the literature so far. However, the performance of an RF model depends on the number of decision trees included (which has an impact on computational power), future work should include repeating our analysis with a tuned RF model.

Xiong et al (2018) extracted 440 radiomic features (including 384 wavelet decomposition image features) predicting pathologic response for 30 SCC oesophageal cancer patients. Xiong et al (2018) reported an accuracy with the SVM algorithm of 82% however, their data included 7 M1 patients, i.e. 7 patients who are predisposed to failed treatment, therefore introducing a bias into their results. Xiong et al (2018) have similarly used a k-fold "leave-one-out" cross validation to boost the objectivity of a small patient cohort however, their accuracy values, similar to our study, are based on small validation groups. Xiong et al (2018) give further evidence to support the need for validation of such studies with larger, standardised patient datasets.

Zhang et al (2014) performed pathologic response prediction analysis on a group of 20 oesophageal squamous cell and adenocarcinoma patients treated with surgery and chemo-radiotherapy. The Zhang et al (2014) were able to produce 100% accuracy for their small patient cohort when predicting pathologic response using an SVM algorithm however, acknowledging that validation in a larger cohort is required, even though they too used a 10 fold cross validation. Zhang et al (2014)

also focussed on analysis of the primary tumour, raising the point that pathologic response based on the primary site alone may not exclude the possibility of lymph node invasion and that further work repeating their study with lymph node analysis.

In summary, we found that the SVM algorithm produced a higher accuracy measure than other algorithms tested when all features were used to predict DFS at 1 year, with performance dropping slightly when predicting 2-year DFS. The performance accuracy of our model is comparable to the limited number of similar studies however a direct comparison was not possible because of differences in the patient cohort and disease type. We produced an accuracy of 0.74 and 0.68 for 1-year and 2-year DFS respectively however, this figure was produced based on the model's sole ability to identify the majority class (success for 1-year and failure for 2-year) and therefore, in this experimental set, we were unable to produce a clinically reliable result using any machine learning algorithm. Our model requires further refinement and more patients to reduce the false classification rate. Whilst we have found similar accuracies to published works, a larger, more standardised patient cohort is required to further tune this model and yield a clinically trustworthy result.

### **6.3.3 Machine learning method comparison for BSREM and OSEM**

Using the same datasets, we performed a further comparison of machine learning algorithms to investigate the effect of image reconstruction, specifically, comparing OSEM with BSREM images. In almost all instances, the use of OSEM or BSREM images made little to no difference in the algorithm accuracy. The SVM algorithm used to predict 1-year DFS produced an accuracy of 0.81 however, this was driven by the model's ability to predict the majority class (successful treatment) and failing to classify any of the failed treatments. A further flaw in this figure was that the low number of patients in the validation dataset further skewed the distribution of successes and fails (ratio 13:3). Of note was the improvement in performance observed for the Logistic Regression (LR) algorithm with BSREM improving accuracy for 2-year DFS prediction from 0.5 to 0.75. OSEM and BSREM produced the same number of true positives however BSREM also produced more true negatives and was able to correctly identify 7/10 successful treatments and, more importantly, 5/6 failed treatments. The key finding here is the ability of the LR algorithm to correctly predict both positive and negative results and, predict a failed treatment from a group with more successes. This is an improvement compared to e.g. the performance of the SVM algorithm which correctly classified all patients in the majority class (whether success or fail depending on the end point) but failed to correctly classify any patients from the minority class. As discussed, (section 6.2) BSREM images produced grey-levels with

a larger dynamic range of values; since LR uses a logarithmic function to separate data into two classes (success and fail), it is possible that the larger range of values facilitates this separation.

The Decision Tree Classifier (DT) algorithm performed 2<sup>nd</sup> best for the BSREM dataset (combined recall score of 1.17) and performed similarly well (combined recall score of 1.21) for a normalised dataset of radiomic and clinical features for OSEM images. DT architecture is designed for classification problems however, it is possible that the performance of this algorithm was hampered by the size of the dataset and therefore the number of 'branches' in the tree; a further, useful comparison would be to the random forest algorithm to attempt to reduce the risk of over fitting the data by including too many decisions.

To the best of our knowledge, there have been no such works which have investigated the effect of PET image reconstruction on ML performance in any clinical application, including for oesophageal cancer. Furthermore, there are no such works investigating the effect of BSREM on ML predictive performance.

## **6.4 Further Work**

This section summarises several areas, highlighted by this study, for further work and exploration.

### **6.4.1 Clinical Data**

Our original clinical dataset used a highly specified and 'cleaned' dataset of only upper GI adenocarcinoma patients who underwent curative treatment. Our time for DFS was calculated from the date of 1<sup>st</sup> treatment whether by neo-adjuvant chemotherapy or surgery however, within this group, there were patients who underwent surgery alone; neo-adjuvant chemotherapy and surgery; and neo-adjuvant chemotherapy, surgery and radiotherapy. One future investigation could include whether our results are replicated in further stratified groups by treatment type. To expand our cohort, an expansion into both squamous cell and adenocarcinoma patients may also be beneficial.

Some authors have shown that the proximity of the primary oesophageal tumour (Ypsilantis, et al., 2015) to the oesophago-gastric junction, is related to poorer treatment outcomes; our patient cohort contained a mixture of oesophageal and OGJ patients and therefore, further separation of these two patient groups may yield different results. Similarly, including clinical data such as the physical distance from the oesophageal tumour to the OGJ may support classification.

Several authors have described the link between nodal involvement and treatment outcomes for oesophageal cancer (Berry, 2014) but, more specifically, the distribution of nodal involvement and the distance to the primary tumour has been linked to treatment response (Sah, et al., 2019). Further work following our study should include evaluation of a further split of the clinical data, e.g. separating the oesophageal and oesophago-gastric junction patients. It may also be possible to include further metrics such as the physical distance of an involved node to the primary tumour to investigate any association with survival outcomes.

Our treatment end point was chosen as 2 years DFS using patients from a narrow time window, balancing access to 3D data and at least 2 years of follow up. It would be useful to explore these results using overall survival as a metric but also for several different time end points such as 6 months, 3 and 5 years for further comparison to the results of other similar works.

It is important to note that this was an  $^{18}\text{F}$ -FDG based study only and that further work investigating a similar group using alternative, novel tracers may yield different results. However, in a study investigating radiomics for response prediction from  $^{18}\text{F}$ -Fluoro-Thymidine PET scans for head and neck cancers, Ulrich et al (2019) similarly found that smaller, homogeneously distributed tumours were associated with better prognosis. Ulrich et al (2019) did not make a comparison with the results of  $^{18}\text{F}$ -FDG scans from the same patient cohort.

The most pertinent further work for this study is to repeat the analysis of BSREM images on a completely new set of upper GI adenocarcinoma patients, for example, patients seen by the NOGU MDT in 2018-2020 with also now at least 2 years of follow up data.

#### **6.4.2 Radiomics – limitations and further work**

Multiple studies found that more heterogeneously distributed patterns of FDG uptake were associated with poorer outcomes (Sah, et al., 2019) however there remains a lack of agreement in terms of which feature is most closely associated with pathologic response or overall outcome. To date, studies investigating prognostic texture features in PET images have been more varied in which specific features correspond directly to treatment response and only two studies have described features associated with overall survival (Sah, et al., 2019). In comparison, authors consistently linked high GLCM entropy with poor outcomes in CT studies for oesophageal cancer where high GLCM entropy indicated a more heterogeneous tumour. Furthermore, of the related published works, there are differences in the disease aetiology, treatment choice, patient demographics, features extracted, PET imaging protocols and image reconstruction; it is therefore difficult to definitively compare our results with published works.



One of the key challenges in comparing and interpreting PET radiomic studies is the lack of standardisation and use of retrospective datasets. Lambin et al (2017) have proposed a “radiomics quality system” (Figure 18) to attempt to address this issue for prospective studies. Our study scored 11/36 (~30%) using the quality score with our main deficiency being the use of a non-trial retrospective dataset. The data used for this study was from a clinical database, held by the MDT, and therefore, all patients enrolled were treated and scanned according to clinical need not dictated a uniform clinical trial protocol. Subsequently, the exact treatment, whilst all were treated with at least surgery, some were given adjuvant or neo-adjuvant chemotherapy and some with additional radiotherapy. Furthermore, the timing of the PET/CT scan during the course of their diagnosis and treatment was not prescribed but performed according to clinical need; in a rapidly progressing cancer, the timing of the imaging can affect the nature and size of the tumour being imaged, therefore bringing in an uncertainty between patients (Lambin, et al., 2017). To mitigate this as far as possible, patients were only enrolled from a single tertiary referral centre MDT and the initial data download included only PET scans referred under a 2-week-wait protocol, therefore limiting the diagnosis to imaging time as far as possible.

Our study also lacked cross-validation of the regions drawn; in our study, the regions of interest used for the primary tumour were grown to a fixed SUV which was adjusted as appropriate to best match the tumour observed on CT and was performed in conjunction with input from an experienced nuclear medicine radiologist. This method balanced consistency with ensuring that any ‘colder’ patches were captured for more heterogeneous tumours (Cook, et al., 2018). Ha et al (2019) suggest that an adaptive or gradient model may be preferable and for consistency but that ultimately, there is no ‘perfect’ method. Because of the nature of the location of oesophageal and particularly OGJ tumours, care taken to avoid including normal physiological uptake in the regions.

One area we were able to ensure consistency between patients was the scanner and scanning protocol in that all PET scans were performed on the same GE 710 Discovery scanner (GE, 2022) with a consistent scanning protocol. Injection time (Lovat, et al., 2017; Cook, et al., 2018), bed position time, and different scanners have been shown to vary radiomic features “about equal to interpatient variability” therefore our data was specifically only included from a single scanner and single institution imaging protocol (Ger, et al., 2019). Furthermore, Van Rossum et al (2016) confirmed the work of Tixier et al (2012) describing only a limited number of textural features which were reproducible between two separately acquired baseline scans “with respect to physiological variability” associated with  $^{18}\text{F}$ -FDG scans. Caution is therefore required when comparing results between centres, studies and even scans acquired for the same patient at the same centre. In

contrast, Ger et al (2019) showed good reliability between radiomic features when imaged using typical clinical parameters however their paper did not report the effect of the most recent PET image reconstruction advance; the BSREM algorithm (section 3.2.5). To the best of our knowledge, the effect of BSREM on clinical oesophageal cancer PET images has not been investigated for any applications, including for oesophageal cancer hence the basis for investigating this further.

The radiomics software we used, LiFEX v7.0.0 (Nioche, et al., 2018), was limited to first, second and higher order radiomic features however some groups have explored using “Fractal Dimension” analysis (Cao, et al., 2020; Xiong, et al., 2018) which uses higher level image processing techniques to incorporate wavelet analysis into the features available. Analysis of higher level radiomic features may also provide further insight into the structure of the images.

Our statistical analysis took account of our non-normally distributed datasets by using a 2 tailed, samples of unequal variance students t-test however, several groups have used a Kruskal-Wallis non-parametric test for statistical analysis of non-normally distributed and such analysis may be better suited to this work. Kruskal-Wallis approach was not chosen for this work because this is designed to test more than two groups (Ypsilantis, et al., 2015; Unistat, 2022); in our work, we tested two groups only; success and failed treatment, however further extension of this work may include more direct comparisons of success and fail groups for different DFS end points.

### **6.4.3 Machine Learning**

Our key finding was that by using a logistic regression algorithm in combination with a set of BSREM images, we were able to correctly predict 2 year disease-free survival for 70% of successful and 83% of failed treatments.

Our analysis using the machine learning facility in python used a variety of different architectures however, there are several key areas for further work. Our machine learning algorithm requires further refinement and optimisation, for example, the use of “principal components analysis” (PCA) to identify the most influential features on any particular algorithm may be helpful in cutting down the number of features required to perform predictive tasks and may subsequently improve the predictive power (Ypsilantis, et al., 2015). For our project, the use of PCA would help to identify which features were the most important in developing the machine learning algorithm parameters which would help to reduce the number of inputs and speed up the running of the program. A further extension of using PCA is to then run multiple iterations of the machine learning process (test-train-validation) to improve the accuracy of the algorithms.

We used Logistic Regression, Linear Discrimination Analysis, K Neighbours Classifier, Decision Tree Classifier, Gaussian Naïve Bayes and Support Vector Machine learning algorithms. There are several other algorithms which have been used successfully in similar studies such as Gradient Boosting (Ypsilantis, et al., 2015), Random Forest (Desbordes, et al., 2017; Paul, et al., 2017), Extreme Machine Learning (Xiong, et al., 2018) and more complex Convolutional Neural Networks (Xiong, et al., 2018; Ypsilantis, et al., 2015). A further expansion of this project would be to explore the comparison of performance of our most promising result (BSREM images and the Logistic Regression algorithm) with a ‘whole image’ CNN approach. The CNN approach presents a number of challenges such as whether to take a data-driven (similar to our approach), ‘whole image’ (Xiong, et al., 2018), 3-slice tumour only image (Ypsilantis, et al., 2015) or a novel approach using e.g. a whole image of the tumour only. Furthermore, there is no optimum approach when designing a CNN to perform a particular task and in setting the number of layers and number of nodes in each layer. To date, results have been optimised using a trial and error approach in a variety of settings not limited to medical imaging (Brownlee, 2022).

We explored the machine learning performance using up to 62 features, including the TNM score. As discussed, we found that larger, more heterogeneously distributed tumours were generally more associated with poorer outcomes. What may be of interest to the wider clinical community is the utility of machine learning performance using a smaller subset of ‘easily accessible’ parameters only, such as SUVmax, volume and TLG, such to give a broader application to centres without radiomic analysis expertise or software.

With any machine learning application, having more studies for training, testing and validation is helpful in improving the algorithm. Our most important result was based on a group of 79 patients however further work with more patients would be useful to clarify, replicate and hopefully improve this result. Furthermore, with a larger dataset, improved ‘training – testing’ could be performed with multiple iterations and updates of the model parameters. Our final validation group used a small number of patients (16 -19 patients) with an uneven split between successes and failures; further work using a larger dataset and therefore a larger validation set would help to resolve any class imbalance classification accuracy issues (Brownlee, 2021). In absence of a larger dataset, repeating the analysis with a smaller but completely balanced dataset would also be useful in clarifying and confirming the accuracy of our algorithms.

## 7 Conclusion

---

In conclusion, we have analysed the PET/CT images for 144 patients with upper GI adenocarcinoma who underwent curative treatment with the intention of predicting disease-free survival (DFS) at 2 years from the radiomic signature of the primary tumour. We investigated the utility of a machine learning approach to semi-automate this prediction. Finally, we evaluated the effect of a novel image reconstruction algorithm, Block Sequential Regularized Expectation Maximum (BSREM), on radiomic values and machine learning algorithm performance.

We found that whilst no individual radiomic feature correlated with DFS, radiomic features describing larger, more heterogeneously distributed tumours, such as total lesion glycolysis (TLG) and grey-level size zone matrix zone length non-uniformity (GLSZM\_ZLNU), were associated with poorer disease-free survival rates. Radiomic features related to the run length matrix remained robust to image reconstruction when comparing Ordered-Subsets Expectation Maximum (OSEM) with BSREM images however, features evaluating local variations, such as grey level co-occurrence matrix (GLCM) Contrast, were highly sensitive to reconstruction method. We found that the majority of machine learning algorithms used in our study did not produce sufficient predictive power for both successful and failed treatments and tended to favour the majority class. However, we found that for BSREM PET/CT images, a logistic regression algorithm was able to predict 70% of successful and 83% of failed treatments correctly, making this the most clinically relevant result. The combination of different pixel value enhancing reconstruction algorithms may help to exaggerate the difference between pixels and therefore the difference in the combination of features which separate patients likely to receive successful treatment. Further work is required, in the first instance, by repeating our analysis on a larger group of similar patients to confirm the validity of this finding.

Ultimately, we found that the radiomic signature from pre-treatment PET/CT images was helpful in identifying features more strongly associated with a failed treatment. Machine learning approaches, in combination with BSREM images may help clinicians to better identify patients who would benefit from closer follow-up after curative treatment for upper GI adenocarcinoma.

## 8 References

---

Abadi, M., Agarwal, A., Barham, P. & et, 2016. Tensorflow: large-scale machine learning on heterogeneous distributed systems. *arxiv*, Volume 1603.04467.

Aerts, H. et al., 2014. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications*, pp. 5: 1-9.

Akamatsu, G. et al., 2012 . Improvement in PET/CT Image Quality with a Combination of Point-Spread Function and Time-of-Flight in Relation to Reconstruction Parameters. *J Nuc. Med* , pp. 53(11), 1716-17 .

Akamatsu, et al., 2013. Benefits of point-spread function and time of flight for PET/CT image quality in relation to the body mass index and injected dose.. *Clin Nucl Med*, 38(6), pp. 407-412.

Aktolun, C., 2019. Artificial Intelligence and radiomics in nuclear medicine: potentials and challenges. *Eur J Nucl Med*, pp. 46:2731-2736.

Alessio, A. & Kinahan, P., 2006. Image Reconstruction. In: *Nuclear Medicine*. Philadelphia: Elsevier.

Alessio, A., Kinahan, P. & Lewellen, T., 2006. Modeling and incorporation of system response functions in 3-D whole body PET. *IEEE Trans on Med Im*, pp. 25 (7): 828-837.

Alessio, A. et al., 2010. Application and Evaluation of a Measured Spatially Variant System Model for PET Image Reconstruction. *IEEE* , Volume 29(3), pp. 938-949.

Allum, W. et al., 2009. Long-term results of a randomized trial of surgery with or without preoperative chemotherapy in esophageal cancer.. *J Clin Oncol*, 27(30), pp. 5062-5067.

Amadasun, M. & King, R., 1989. Textural Features Corresponding to Textural Properties.. *IEEE Transactions on systems, man, and cybernetics*, 19(5), pp. 1264-1274.

American Cancer Society, 2019. *Treating Esophageal Cancer by Stage*. [Online] Available at: <https://www.cancer.org/cancer/esophagus-cancer/treating/by-stage.html> [Accessed 20 February 2020].

Amico, A., Borys, D. & Gorczewska, I., 2020. Radiomics and Artificial Intelligence for ET imaging analysis. *Nucl Med Rev*, Volume 23, pp. 36-39.

Anon., 2021. *k-nearest neighbor algorithm in Python*. [Online] Available at: <https://www.geeksforgeeks.org/k-nearest-neighbor-algorithm-in->

[python/#:~:text=This%20algorithm%20is%20used%20to,nearest%20of%20the%20boundary%20line.](#)  
[Accessed 16 June 2022].

Apte, A., Iyer, A., Crispin-Ortuzar, M. & al, e., 2018. Technical note: extension of CERR for computational radiomics: a comprehensive MATLAB platform for reproducible radiomics research. *Med Phys*, Volume 45, pp. 3713-3720.

Atay-Rosenthal, S., Wahl, R. & Fishman, E., 2012. PET/CT findings in gastric cancer: potential advantages and current limitations. *Imaging Med.*, 4(2), p. 241–250.

Avanzo, M., Stancanello, J. & Pirrone, G. S. G., 2020. Radiomics and deep learning in lung cancer. *Strahlenther Onkol*, Volume 196, pp. 879-887.

Badawi, R., 1999. *Introduction to PET Physics*. [Online] Available at: [depts.washington.edu/imreslab/](https://depts.washington.edu/imreslab/)  
[Accessed 10 Feb 2023].

Bastien, F., Lamblin, P., Pascanu, R. & al, e., 2012. Theano: new features and speed improvements. *Deep Learning and Unsupervised Feature Learning NIPS Workshop*.

Beig, N. et al., 2020. Radiogenomic-Based Survival Risk Stratification of Tumor Habitat on Gd-T1w MRI Is Associated with Biological Processes in Glioblastoma. *Clinical Cancer Research*, 26(8), pp. 1866-1876.

Berger, A., 2003. Positron Emission Tomography. *BMJ*, p. 326: 1449.

Berry, M., 2014. Esophageal cancer: staging system and guidelines for staging and treatment. *J Thoracic Dis*, pp. 6(S3): S289-S297.

Beukinga, R. et al., 2018. Prediction response to neoadjuvant chemotherapy and radiation therapy with baseline and restaging 18F-FDG PET imaging biomarkers in patients with esophageal cancer. *Radiology*, pp. 3: 983 - 992.

Beukinga, R., Hulshoff, J., van Dijk, L. & al., e., 2017. Predicting response to neoadjuvant chemoradiotherapy in esophageal cancer with textural features derived from pretreatment (18)F-FDG PET/ CT imaging.. *J Nucl Med*, Volume 58(5), p. 723–729.

Boellaard, R., Delgado-Bolton, R., Oyen, W. & al, e., 2015. FDG PET/CT EANM procedure guidelines for tumour imaging version 2.0. *Eur J Nucl Med Mol Imaging*, Volume 42, pp. 328-354.

Bolus, N., George, R., Washington, J. & Newcomer, B., 2009. PET/MRI: The Blended Choice for the future?. *J Nucl med Technol*, Volume 37, pp. 63-71.

Brooks, F. & Grigsby, P., 2014. The Effect of Small Tumor Volumes on Studies of intratumoral Heterogeneity of Tracer Uptake. *J Nucl Med*, Volume 55, pp. 37-42.

Brownlee, 2016. *K-nearest Neighbours for Machine Learning*. [Online]  
Available at: <https://machinelearningmastery.com/k-nearest-neighbors-for-machine-learning/>  
[Accessed 16 June 2022].

Brownlee, J., 2019. *A Gentle Introduction to Imbalanced Classification*. [Online]  
Available at: <https://machinelearningmastery.com/what-is-imbalanced-classification/>  
[Accessed 19 05 2022].

Brownlee, J., 2019. *Machine Learning Mastery: Your first machine learning prject in Python step-by-step*. [Online]  
Available at: <https://machinelearningmastery.com/machine-learning-in-python-step-by-step/>  
[Accessed 20 March 2021].

Brownlee, J., 2020. *A Gentle Introduction to k-fold Cross-Validation*. [Online]  
Available at: <https://machinelearningmastery.com/k-fold-cross-validation/>  
[Accessed 25 Aug 2022].

Brownlee, J., 2020. *Linear Discriminant Analysis*. [Online]  
Available at: <https://machinelearningmastery.com/linear-discriminant-analysis-with-python/#:~:text=Linear%20Discriminant%20Analysis%20is%20a,observations%20for%20each%20input%20variable.>  
[Accessed 16 June 2022].

Brownlee, J., 2021. *Failure of Classification Accuracy for Imbalanced Class Distributions*. [Online]  
Available at: <https://machinelearningmastery.com/failure-of-accuracy-for-imbalanced-class-distributions/>  
[Accessed 19 05 2022].

Brownlee, J., 2022. *Your First Deep Learning Project in Python with Keras Step-By-Step*. [Online]  
Available at: <https://machinelearningmastery.com/tutorial-first-neural-network-python-keras/>  
[Accessed 23 June 2022].

Bruzzi, J. et al., 2007. PET/CT of Esophageal cancer: its role in clinical management. *RadioGraphics*, pp. 27:1635-1652.

Buch, K. et al., 2015. Using Texture Analysis to Determine Human Papillomavirus Status of Oropharyngeal Squamous Cell Carcinomas on CT. *Am J Neuroradiol*, Volume 36, p. 1343– 48 .

Budinger, T., 1983. Time-of-Flight Positron Emission Tomography: Status Relative to COnventional PET. *J Nucl Med*, 24(1), pp. 73-78.

Cancer Research UK, 2019. *PET-CT Scan*. [Online] Available at: <https://www.cancerresearchuk.org/about-cancer/cancer-in-general/tests/pet-ct-scan> [Accessed 18 Dec 2019].

Cao, Q. et al., 2020. Development and validation of a radiomics signature on differentially expressed features of 18F-FDG PET to predict treatment response of concurrent chemoradiotherapy in thoracic esophagus squamous cell carcinoma.. *Radiother. Oncol.*, Volume 146, pp. 9-15.

Casey, M., 2007. *Point Spread Function Reconstruction in PET*, Malvern, USA: Siemens Medical.

Castellano, G., Bonilha, L., Li, M. & Cendes, F., 2004. Texture analysis of medical images. *Clinical Radiology (2004)*, Volume 59, p. 1061–1069.

Chauhan, N., 2022. *Decision Tree ALgorithm, Explained*. [Online] Available at: <https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html> [Accessed 16 June 2022].

Chen, Y. et al., 2019. Combining the radiomic features and traditional parameters of 18F-FDG PET with clinical profiles to improve prognostic stratification in patients with esophageal squamous cell carcinoma treated with neoadjuvant chemoradiotherapy and surgery.. *Ann Nucl. Med*, Volume 33, pp. 657-670.

Cherry, S., 2006. The 2006 Henry N. Wagner Lecture: Of Mice and Men (and Positrons) – Advances in PET imaging Technology. *J Nuc. Med*, Volume 47, pp. 1735-1745.

Cherry, S., Dahlborn, M. & Hoffman, E., 1991. 3D PET using a conventional multislice tomograph without septa. *J Comput Assist Tomograph*, Volume 15, pp. 655-68.

Chirieac, L., Swisher, S., Ajani, J. & al, e., 2005. Posttherapy pathologic stage predicts survival in patients with esophageal carcinoma receiving preoperative chemoradiation. *Cancer*, Volume 103(7), pp. 1347-1355.

Collobert, R., Kavukcuoglu, K. & Farabet, C., 2011. Torch7: a matlab-like environment for machine learning. *Advances in Neural Information Processing Systems*.

Cook, G. et al., 2018. Challenges and Promises of PET Radiomics. *Int J Radiation Oncol Biol Phys*, pp. 102(4) 1083-1089.

Cook, G. et al., 2013. Are Pretreatment 18F-FDG PET Tumor Textural Features in Non–Small Cell Lung Cancer Associated with Response and Survival After Chemoradiotherapy?. *J Nucl Med*, Volume 54 (1), pp. 19-26.



CRUK, 2017. *Oesophageal Cancer Statistics*. [Online] Available at: [https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/oesophageal-cancer#:~:text=There%20are%20around%209%2C200%20new,new%20cancer%20cases%20\(2017\).](https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/oesophageal-cancer#:~:text=There%20are%20around%209%2C200%20new,new%20cancer%20cases%20(2017).) [Accessed 12 Oct 2020].

CRUK, 2019. *TNM Staging for oesophageal cancer*. [Online] Available at: <https://www.cancerresearchuk.org/about-cancer/oesophageal-cancer/stages-types-and-grades/tnm-staging#:~:text=for%20oesophageal%20cancer-,TNM%20staging%20for%20oesophageal%20cancer,way%20to%20stage%20oesophageal%20cancer.> [Accessed 7 July 2022].

Cunningham, D., Allum, W., Stenning, S. & al., e., 2006. Perioperative chemotherapy versus surgery alone for resectable gastroesophageal cancer. *N Engl J Med*, 355(1), pp. 11-20.

Czernin, J., Allen-Auerbach, M., Nathanson, D. & Herrmann, a. K., 2013. PET/CT in Oncology: Current Status and Perspectives. *Curr Radiol Rep.*, 1(3), p. 177–190.

Deantonio, L. et al., 2022. 18F-FDG PET Radiomics as Predictor of Treatment Response in Oesophageal Cancer: A Systematic Review and Meta-Analysis. *Front Oncol.*, Volume 15.

Desbordes, P. et al., 2017. Predictive value of initial FDG-PET features for treatment response and survival in esophageal cancer patients treated with chemo-radiation therapy using a random forest classifier. *PLoS ONE*, 12(e0173208).

Dinapoli, N., Alitto, A., Vallati, M. & al, e., 2015. Moddicom: a complete and easily accessible library for prognostic evaluations relying on image features. *Conf Proc IEEE Eng Med Biol Soc*, pp. 771-774.

Dong, X., Xing, L., Wu, P. & al., e., 2013. Three-dimensional positron emission tomography image texture analysis of esophageal squamous cell carcinoma: relationship between tumor 18F-fluorodeoxyglucose uptake heterogeneity, maximum standardized uptake value, and tumour stage. *Nucl Med Commun*, Volume 34, pp. 40-46.

Downey, R., Akhurst, T., D, I. & al, e., 2003. Whole body 18FDG-PET and the response of esophageal cancer to induction therapy: results of a prospective trial. *J Clin Oncol*, pp. 21:428-432.

Dunn, O., 1961. Multiple Comparisons Among Means. *Journal of the American Statistical Association*, 56(293), pp. 52-64.

Echegaray, S., Bakr, S., Rubin, D. & Napel, S., 2018. Quantitative image feature engine (QIFE): an open-source, modular engine for 3D quantitative feature extraction from volumetric medical images. *J Digit Imaging*, Volume 31, p. 403–414.

- Eisenhauer, E. et al., 2009. New Response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *Eur J Cancer*, Volume 45, pp. 228-247.
- Esfahani, S. et al., 2022. PET/MRI and PET/CT Radiomics in Primary Cervical Cancer: A Pilot Study on the Correlation of Pelvic PET, MRI, and CT Derived Image Features. *Mol Imaging Biol*, Volume 7, pp. 60-69.
- Fang, Y., Lin, C., Shih, M. & al, e., 2014. Development and evaluation of an open-source software package “cGITA” for quantifying tumor heterogeneity with molecular images.. *Biomed Res Int*, Volume 248505.
- Fitzmaurice, C. et al., 2013. The Global Burden of Cancer 2013. *JAMA Oncol*, pp. 505-527.
- Foley, K. et al., 2018. Development and validation of a prognostic model incorporating texture analysis derived from standardised segmentation of PET in patients with oesophageal cancer. *Eur Radiol*, Volume 28, pp. 428-436.
- Fornacon-Wood, I. et al., 2020. Reliability and prognostic value of radiomic features are highly dependent on choice of feature extraction platform. *European Radiology*, Volume 30, pp. 6241-6250.
- Galavis, P. et al., 2010. Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters. *Acta Oncol*, 49(7), pp. 1012-1016.
- Galloway, M., 1975. Texture Analysis Using Gray Level Run Lengths. *Computer Graphics and Image Processing*, Volume 4, pp. 172-179.
- GE, 2022. *GE Healthcare*. Bangalore, India: GE Healthcare.
- Ger, R. et al., 2019. Effects of alterations in positron emission tomography imaging parameters on radiomics features. *PLoS One*, Volume e0221877, pp. 1-12.
- Ghotbi, A. A., Kjær, A. & Hasbak, P., 2014. Review: comparison of PET rubidium-82 with conventional SPECT myocardial perfusion imaging. *Clin Physiol Funct Imaging*, 34(3), pp. 163-170.
- Gillies, R., Kinahan, P., Hricak, H. & al, e., 2016. Radiomics: Images are more than pictures, they are data. *Radiology*, pp. 278(2):563-577.
- Götz, M., Nolden, M. & Maier-Hein, K., 2019. MITK Phenotyping: an open-source toolchain for image-based personalized medicine with radiomics.. *Radiother Oncol*, Volume 131, p. 108–111.
- Gundlich, B. et al., 2006. From 2D PET to 3D PET: issues of data representation and image reconstruction. *Z Med Phys*, 16(1), pp. 31-46.

- Ha, S., Choi, H., Paeng, J. & Cheon, G., 2019. Radiomics in Oncological PET/CT: a Methodological Overview. *Nuclear Medicine and Molecular Imaging*, Volume 53, pp. 14-29.
- Hatt, M. et al., 2009. A fuzzy locally adaptive Bayesian segmentation approach for volume determination in PET. *IEEE Trans Med Imaging*, 28(6), pp. 881-893.
- Hatt, M. et al., 2015. 18F-FDG PET Uptake Characterization Through Texture Analysis: Investigating the Complementary Nature of Heterogeneity and Functional Tumor Volume in a Multi-Cancer Site Patient Cohort. *J Nucl Med*, Volume 56, pp. 38-44.
- Hatt, M. et al., 2013. Robustness of intratumour 18-F-FDG PET uptake heterogeneity quantification for therapy response prediction in oesophageal carcinoma. *Eur J Nucl Med Mol Imaging*, Volume 40, pp. 1662-1671.
- Hatt, M., Tixier, F., Visvikis, D. & Cheze Le Rest, C., 2016. Radiomics in PET/CT: More than Meets the Eye?. *J Nucl Med*, pp. 365-6.
- Hatt, M., Visvikis, D., Pradier, O. & Cheze-Le-Rest, C., 2011. Baseline 18F-FDG PET image-derived parameters for therapy response prediction in oesophageal cancer. *Eur J Nucl Med Mol Imaging*, Volume 38 (9).
- Honey, N. & Shows, T., 1983. The tumor phenotype and the human gene map. *Cancer Genet Cytogenet*, Volume 10(3), pp. 287-310.
- Honey, N. & Shows, T., 1983. The tumor phenotype and the human gene map. *Cancer Genet Cytogenet*, Volume 10(3), pp. 287-310.
- Huang, B., Law, M. W.-M. & Khong, P.-L., 2009. Whole-Body PET/CT Scanning: Estimation of Radiation Dose and Cancer Risk. *Radiology*, 251(1), pp. 166-174.
- Ingrisch, M., Schoppe, F., Paprottka, K. & al, e., 2018. Prediction of 90Y radioembolization outcome from pretherapeutic factors with random survival forests. *J Nucl Med*, Volume 59, pp. 769-773.
- Jia, Y. et al., 2014. Caffe: convolutional architecture for fast feature embedding.. *22nd ACM International Conference on Multi-media*, pp. 675-678.
- Junttila, M. & Sauvage, F., 2013. Influence of tumour micro-environment heterogeneity on therapeutic response. *Nature*, pp. 501: 346-354.
- Kadmas, D. et al., 2009. Impact of Time-of-flight on PET Tumour Detection. *J. Nuc. Med.* , Volume 50, pp. 1315-1323.
- Karp, J. et al., 2005 . Characterization of a Time-of-Flight PET Scanner based on Lanthanum Bromide. *IEEE Nuc. Sci. Symposium Conference Record* , pp. 1919-1923.

- Karp, J., S, S., M, D.-W. & G, a. M., 2008. Benefit of Time-of-Flight in PET: Experimental and Clinical Results. *J. Nuc. Med*, Volume 49, pp. 462-470.
- Kelchtermans, P., Bittremieux, W., De Grave, K. & al, e., 2014. Machine Learning applications in proteomics research: how the past can boost the future. *Proteomics*, pp. 14:353-366.
- Kinahan, P. & Fletcher, J., 2010. PET/CT Standardised Uptake Values (SUVs) in CLinical Practice and Assessing Repsonse to Therapy. *Semin Ultrasound CT MR*, 31(6), pp. 496-505.
- Kudou, M. et al., 2016. Efficacy of PET-CT in the Diagnosis and Treatment of Recurrence After Esophageal Cancer Surgery. *AntiCancer Research*, pp. 36: 5473-5480.
- Lambin, P., Leijenaar, R., Deist, T. & al., e., 2017. Radiomics: the bridge between medical imaging and personalized medicine.. *Nat Rev Clin Oncol*, Volume 14(12), p. 749–762.
- Lambin, P. et al., 2012. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer*, 48(4), pp. 441-446.
- Langer, A., 2010. A systematic review of PET and PET/CT in oncology: A way to personalize cancer treatment in a cost-effective manner?. *BMC Health Serv Res.* , 10(283).
- Larue, R. et al., 2018. Pre-treatment CT radiomics to predict 3-year overall survival following chemoradiotherapy of esophageal cancer. *Acta Oncologica*, 57(11), pp. 1475-1481.
- Lee, L., Kanthasamy, S., Ayyalaraju, R. & Ganatra, R., 2019. The Current State of Artificial Intelligence in Medical Imaging and Nuclear Medicine.. *BJR*, Volume 1.
- Leijenaar, R., Nalbantov, G., Carvalho, S. & al., e., 2015. The effect of SUV discretization in quantitative FDG-PET radiomics: the need for standardized methodology in tumor texture analysis.. *Sci Rep*, 5(11075), pp. 1-10.
- Litjens, G., Kooi, T., Bejnordi, B. & al, e., 2017. A survey on deep learning in medical image analysis. *Medical Image Analysis*, Volume 42, pp. 60-88.
- Li, W. et al., 2021. Artificial intelligence applications for oncological positron emission tomography imaging. *Eur J Radiol*, Volume 134.
- Lois, C., Jakoby, B. W. & Long M J, H. K. F. B. D. W. C. M. E. C. M. P. V. Y. K. D. J. a. T. D. W., 2008 . An Assessment of the Impact of Incorporating Time-of-Flight Information into Clinical PET/CT. *J. Nuc. Med.* , Volume 51 , pp. 237-245.
- Lovat, E. et al., 2017. The effect of post-injection 18F-FDG PET scanning time on texture analysis of peripheral nerve sheath tumours in neurofibromatosis-1. *EJNMMI Research*, 7(35).

Luketich, J., Schauer, P., Meltzer, C. & al, e., 1997. Role of positron emission tomography in staging esophageal cancer. *Ann Thorac Surg*, pp. 64:765-769.

Luu, C. et al., 2017. Endoscopic ultrasound staging for early esophageal cancer: Are we denying patients neoadjuvant chemo-radiation?. *World J Gastroenterol*, pp. 23(46) 8193-8199.

Ma, C., Li, D., Yin, Y. & al., e., 2015. Comparison of characteristics of 18F-fluorodeoxyglucose and 18F-fluorothymidine PET during staging of esophageal squamous cell carcinoma.. *Nucl Med Commun* , Volume 36, pp. 1181-1186.

Ma, C., Li, D. & Yin, Y. e. a., 2015. Comparison of characteristics of 18F-fluorodeoxyglucose and 18F-fluorothymidine PET during staging of esophageal squamous cell carcinoma. *Nucl Med Commun*, Volume 36, pp. 1181-6.

Manjeshwar, R. et al., 2006. Fully 3D PET Iterative Reconstruction Using Distance Driven Projectors and Native Scanner Geometry. *IEEE Nuc. Sci. Symp. Conf. Rec* , pp. 2804-2807.

Marsden, P., 2017. *Artificial Intelligence Timeline Infographic – From Eliza to Tay and beyond*. [Online] Available at: <https://digitalwellbeing.org/artificial-intelligence-timeline-infographic-from-eliza-to-tay-and-beyond/> [Accessed 21 Jan 2021].

Mayerhoefer, M. et al., 2020. Introduction to Radiomics. *J Nucl Med*, Volume 61, pp. 488-495.

Mayo Clinic Staff, 2019. *Esophageal Cancer*. [Online] Available at: <https://www.mayoclinic.org/diseases-conditions/esophageal-cancer/symptoms-causes/syc-20356084> [Accessed 18 Dec 2019].

Mehta, R. et al., 2017. A Lesion-Based Response Prediction Model Using Pretherapy PET/CT Image Features for Y90 Radioembolization to Hepatic Malignancies. *Technology in Cancer Research & Treatment*, 16(5), pp. 620-629.

Melcher, C. & Schweitzer, J., 1991. Cerium-doped Oxyorthosilicate A Fast, Efficient New Scintillator. *IEEE Nuc. Sci. Symp and Med. Im. Conference*, pp. 228-231.

Moses, W., 2003. Time of Flight in PET Revisited. *IEEE Trans. on Nuc. Sci.* , Volume 50 , pp. 1325-1330.

Moses, W. & Derenzo, S., 1999. Prospects for Time-of-Flight PET using LSO Scintillator. *IEEE Trans. on Nuc. Sci.*, pp. 46: 474-478.

Mullani, N., Markham, J. & Ter-Pogossian, M., 1980 . Feasibility of Time-of-Flight Reconstruction in Positron Emission Tomography. *J Nuc Med*, pp. 21: 1095-1097.

Murray, et al., 2010. Time-of-flight PET/CT using low-activity protocols: potential implications for cancer therapy monitoring.. *Eur J Nucl Med Mol Imaging*, 37(5), pp. 1643-53.

Naghavi, M., Malekzadeh, R. & Collaborators, O. C., 2020. The global, regional, and national burden of oesophageal cancer and its attributable risk factors in 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet Gastroenterol Hepatol*, Volume 5, pp. 582-97.

Nakajo, M. et al., 2017. Texture analysis of 18F-FDG PET/CT to predict tumour response and prognosis of patients with esophageal cancer treated by chemoradiotherapy. *Eur J Nucl Med Mol Imaging*, Volume 44, pp. 206-214.

Nayyeri, F., 2015. A review on motion correction methods in PET/CT images for detection of cancer cells. *Acta Medica Bulgarica*, 62(2), pp. 68-79.

NHS, 2019. *Oesophageal Cancer*. [Online] Available at: <https://www.nhs.uk/conditions/oesophageal-cancer/> [Accessed 18 Dec 2019].

Nioche, C., Orlhac, F., Boughdad, S. & al, e., 2018. Lifex: a freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity.. *Cancer Res*, Volume 78, pp. 4786-4789.

Oren, O., Gersh, B. & Bhatt, D., 2020. Artificial intelligence in medical imaging: switching from radiographic pathological data to clinically meaningful endpoints. *Lancet Digital Health*, Volume 2, pp. e486-488.

Panin, V., Kehren, F., Michel, C. & Casey, M., 2006. Fully 3-D PET Reconstruction With System Matrix Derived From Point Source Measurements. *IEEE Trans on Med. Im*, Volume 25(7), pp. 907-921.

Parmar, C. et al., 2015. Machine Learning methods for Quantitative Radiomic Biomarkers. *Nature: Scientific Reports*, 5(13087), pp. 1-11.

Paul, D. et al., 2017. Feature selection for outcome prediction in oesophageal cancer using genetic algorithm and random forest classifier. *Computerized Medical Imaging and Graphics*, Volume 60, pp. 42-49.

Pedregosa, F. et al., 2011. SciKit-Learn Machine Learning in Python. *Journal of Machine Learning Research*, Volume 12, pp. 2825-2830.

Pennathur, A., Gibson, M., Jobe, B. & Luketich, J., 2013. Oesophageal Carcinoma. *Lancet*, Volume 381, pp. 400-412.

Pesapane, F., Codari, M. & Sardanelli, F., 2018. Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine. *European Radiology Experimental*, 2(35).

Pfaehler, E., Zwanenburg, A., de Jong, J. & Boellaard, R., 2019. RACAT: an open source and easy to use radiomics calculator tool.. *PLoS One*, Volume 14, p. 1–26.

Phillips, 2020. *CareStream PACS*. [Online] Available at: <https://collaboration.carestream.com/solutions/diagnostics/radiology> [Accessed 13 Jan 2022].

Rahman, S. et al., 2019. Machine learning to predict early recurrence after oesophageal cancer surgery. *Br J Surg*, Volume 106 Suppl 12:S7.

Rapisarda, E., V, B., K, T. & M, G., 2010. Image-based point spread function implementation in a fully 3D OSEM reconstruction algorithm for PET. *Phys. Med. Biol.*, Volume 55, pp. 4131-4151.

Rashid, N., Elshaer, M., Kosmin, M. & Riaz, A., 2015. Current management of oesophageal cancer. *BMJ*, Volume 8.

RCR, 2016. *Evidence-based indications for the use of PET/CT in the United Kingdom in 2016*, London: RCR.

Reader, A. et al., 2003. EM algorithm system modeling by image-space techniques for PET reconstruction. *IEEE Trans Nucl Sci*, 50(5), pp. 1392-1397.

Reynes-Llompart, G., 2019. *Impact of Tomographic reconstruction with Bayesian penalty in the quantification of PET/CT studies*, Pamplona, Spain: Faculty of Medicine, Universidad de Navarra.

Reynes-Llompart, G., Gamez-Cenzano, C. & Marti-Climent, J., 2018. Phantom, clinical, and texture indices evaluation and optimization of a penalized-likelihood image reconstruction method (Q.Clear) on a BGO PET/CT scanner. *Med Phys*, Volume 7.

Reynes-Llompart, G. et al., 2019. Image quality evaluation in a modern PET system: impact of new reconstructions methods and a radiomics approach. *Nature Sci Reports*, 9 (10640).

Reynolds, J., Preston, S., O'Neill, B. & al, e., 2017. ICORG 10-14: NEOadjuvant trial in adenocarcinoma of the oEsophagus and oesophagoGastric junction International Study (Neo-AEGIS).. *BMC Cancer*, 17(1), p. 401.

Rich, D., 1997. A Brief History of Positron Emission Tomography. *J Nucl Med Technol*, Volume 25, pp. 4-11.

Ries, L., Eisner, M., Kosary, C. & al, e., 2005. *SEER Cancer Statistics Review, 1975 - 2002*. [Online] Available at: [http://seer.cancer.gov/csr/1975\\_2002/](http://seer.cancer.gov/csr/1975_2002/) [Accessed 18 Dec 2019].

Rosenblatt, F., 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol Rev*, pp. 65: 386-408.

Ross, S., 2014. *Q.Clear: GE White Paper*, Milwaukee, WA: GE.

Ross, S. & Stearns, C., 2010. *SharpIR White Paper*, Milwaukee: GE.

Ruotsalainen, U. & Viik, J., 2015. *SALLA YLIPÄÄ MONTE CARLO SIMULATIONS OF AN AXIAL POSITRON EMISSION TOMOGRAPHY DEMONSTRATOR*, YLIPÄÄ, SALLA: TAMPERE UNIVERSITY OF TECHNOLOGY .

Sah, B. et al., 2019. Radiomics in esophageal and gastric cancer. *Abdominal Radiology*, pp. 44: 2048-2058.

Scapicchio, C. et al., 2021. A deep look into radiomics. *La Radiologia medica*, Volume 126, pp. 1296-1311.

Schmitdz, R., Alessio, A. & Kinahan, P., 2004. *The Physics of PET/CT scanners*, University of Washington: Imaging Research Laboratory.

Sharma, P., 2021. *Implementation of Gaussian Naive Bayes in Python Sklearn*. [Online] Available at: <https://www.analyticsvidhya.com/blog/2021/11/implementation-of-gaussian-naive-bayes-in-python-sklearn/> [Accessed 16 June 2022].

Shiri, I. et al., 2017. The impact of image reconstruction settings on 18F-FDG PET radiomic features: multi-scanner phantom and patient studies. *Eur Radiol*, Volume 27, pp. 4498-4509.

Silink, K., 1961. The possibility of designing machines which learn diagnostic: the zero systems of types and pathtypes in endocrinology. *Act Nerv Super (Praha)*, pp. 3: 148-153.

Spasic, E. et al., 2018. Phantom and Clinical Evaluation for New PET/CT Reconstruction Algorithm: Bayesian Penalized Likelihood Reconstruction Algorithm Q.Clear. *J Nucl Med Rad Ther*.

Stahl, A., Ott, K., Weber, W. & al, e., 2003. FDG-PET Imaging of locally advanced gastric carcinomas: correlation with endoscopic and histopathological findings. *Eur J Nucl Med Mol Imaging*, pp. 30:288-295.

Stojiljkovic, M., 2022. *Logistic Regression in Python*. [Online] Available at: [https://realpython.com/logistic-regression-python/#:~:text=The%20logistic%20function%20F0%9D%91%9D\(%F0%9D%90%B1](https://realpython.com/logistic-regression-python/#:~:text=The%20logistic%20function%20F0%9D%91%9D(%F0%9D%90%B1)



%20is%20the%20sigmoid%20function,that%20the%20output%20is%200.

[Accessed 16 June 2022].

Surti, S. et al., 2007. Performance of Phillips Gemini TF PET/CT Scanner with Special Consideration for Its Time-of-Flight Imaging Capabilities. *J Nuc Med*, Volume 48, pp. 471-480.

Szczypiński, P., Strzelecki, M., Materka, A. & Klepaczko, A., 2009. MaZda-a software package for image texture analysis.. *Comput Methods Program Biomed*, Volume 94, pp. 66-76.

Tan, S., Kligerman, S., Chen, W. & al., e., 2013. Spatial-temporal [18F] FDG-PET features for predicting pathologic response of esophageal cancer to neoadjuvant chemoradiation therapy.. *Int J Radiat Oncol Biol Phys*, Volume 85, pp. 1375-82.

Tan, S., Zhang, H., Zhang, Y. & al., e., 2013. Predicting pathologic tumor response to chemoradiotherapy with histogram distances characterizing longitudinal changes in 18F-FDG uptake patterns.. *Med Phys.*, Volume 40:101707.

Teoh, E. et al., 2015. Phantom and Clinical Evaluation of the Bayesian Penalized Likelihood Reconstruction Algorithm Q.Clear on an LYSO PET/CT System. *J Nucl Med*, Volume 56, pp. 1447-1452.

Teoh, E. et al., 2015. Phantom and Clinical Evaluation of the Bayesian Penalized Likelihood Reconstruction Algorithm Q.Clear on an LYSO PET/CT System. *J Nucl Med*, Volume 56, pp. 1447-1452.

Thibault, 2006. *Size Zone Matrix*. [Online] Available at: <https://www.thibault.biz/Research/ThibaultMatrices/GLSZM/GLSZM.html> [Accessed 1 Dec 2020].

Tixier, F. et al., 2011. Intratumor Heterogeneity Characterized by Textural Features on Baseline 18F-FDG PET Images Predicts Response to Concomitant Radiochemotherapy in Esophageal Cancer. *J Nucl Med*, Volume 52, p. 369–378.

Tixier, F. et al., 2012. Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in 18F-FDG PET. *J Nucl Med*, Volume 53 (5), pp. 693-700.

Tong, S., Alessio, A. & Kinahan, P., 2010. Noise and signal properties in PSF-based fully 3D PET image reconstruction: an experimental evaluation. *Phys. Med. Biol.*, Volume 55, pp. 1453-1473.

Townsend DW, W. M. B. L. e. a., 1993. A rotating PET scanner using BGO block detectors: Design, performance and applications.. *J Nucl Med*, Volume 34, p. 1367–1376.

Townsend, D., 2008. Combined PET/CT: the historical perspective. *Semin Ultrasound CT MR*, p. 29 (4): 232–235.

Tramontano, A. et al., 2019. Esophageal cancer treatment costs by phase of care and treatment modality, 2000-2013. *Cancer Medicine*, Volume 8, pp. 2158-5172.

Trevethan, R., 2017. Sensitivity, Specificity, and Predictive Values: Foundations, Plabilities, and Pitfalls in Research and Practice. *Epidemiology*, 5(307), pp. 1-7.

Ulrich, E., Menda, Y., Ponto, L. & al, e., 2019. FLT PET Radiomics for Response Prediction to Chemoradiation Therapy in Head and Neck Squamous Cell Cancer. *Tomography*, 5(1), pp. 161-170.

Unistat, 2022. *Kruskal-Wallis Test*. [Online] Available at: <https://www.unistat.com/guide/nonparametric-tests-kruskal-wallis-one-way-anova/> [Accessed 25 Aug 2022].

Uribe, C. et al., 2019. Machine Learning in Nuclear Medicine: Part 1 - Introduction. *J Nucl Med*, pp. 60:451-458.

Van Griethuysen, J., Fedorov, A., Parmar, C. & al, e., 2017. Computational radiomics system to decode the radiographic phenotype. *Cancer Res*, Volume 77, p. e104–e107.

Van Hagen, P., Hulshof, J., Van Lanshot, E. & al, e., 2012. Preoperative Chemoradiotherapy for Esophageal or Junctional Cancer. *N Engl J Med*, 17(1), pp. 2074-2084.

Van Helden, F. et al., 2016. Repeatability of Radiomic Features in Non-Small-Cell Lung Cancer [18F]FDG-PET/ CT Studies: Impact of Reconstruction and Delineation. *Mol Imaging Biol*, Volume 18, pp. 788-795.

Van Rossum, P. et al., 2016. The Incremental Value of Subjective and Quantitative Assessment of 18F-FDG PET for the Prediction of Pathologic Complete Response to Preoperative Chemoradiotherapy in Esophageal Cancer.. *J Nucl Med*, Volume 57, pp. 691-700.

Van Rossum, P. et al., 2016. The emerging field of radiomics in esophageal cancer: current evidence and future potential. *Transl Cancer Res*, pp. 5(4): 410-423.

Van Weegaeghe, D. et al., 2016. Prospective validation of 18F-FDG brain PET discriminant analysis methods in the diagnosis of amyotrophic lateral sclerosis.. *J Nucl Med*, Volume 57, pp. 1238-1243.

Vargese, T. et al., 2013. The Society of Thoracic Surgeons Guidelines on the Diagnosis and Staging of Patients With Esophageal Cancer. *Ann Thoracic Surg*, pp. 96: 346-356.

Velazquez, R. et al., 2017. Somatic Mutations Drive Distinct Imaging Phenotypes in Lung Cancer. *Cancer Research*, 77(14), pp. 3922-3932.

Vennart, N. et al., 2017. Optimization of PET/CT image quality using the GE ‘Sharp IR’ point-spread function reconstruction algorithm. *Nuclear Medicine Communications*, Volume 38, p. 471–479.

- Vincente, A., Ballensiefen, W. & Jonsson, J., 2020. How Personalised medicine will transform healthcare by 2030: the ICPeMed version. *J Transl Med*, 18(180).
- Watanabe, M. et al., 2022. Comprehensive registry of esophageal cancer in Japan, 2014. *Esophagus*, Volume 19, pp. 1-26.
- Wei, L. & El Naqa, I., 2020. AI for Response Evaluation With PET/CT. *Sem Nucl Med*, Volume 0, pp. 1-13.
- Wu, L. et al., 2018. Radiomics approach for preoperative identification of stages I–II and III–IV of esophageal cancer. *Chin J Cancer Res*, Volume 30(4), pp. 396-405.
- Wyrzykowski, M. et al., 2020. Impact of the Q.Clear reconstruction algorithm on the interpretation of PET/CT images in patients with lymphoma. *EJNMMI Research*, 10(99).
- Xie, C. et al., 2021. Machine Learning and Radiomics Applications in Esophageal Cancers Using Non-Invasive Imaging Methods - A Critical Review of Literature. *Cancers*, 13(2469).
- Xiong, J. et al., 2018. The Role of PET-Based Radiomic Features in Predicting Local Control of Esophageal Cancer Treated with Concurrent Chemoradiotherapy. *Nature*, 8(9902), pp. 1-11.
- Yang, C. et al., 2019. Deep Convolutional Neural Network-Based Positron Emission Tomography Analysis Predicts Esophageal Cancer Outcome. *J Clin Med*, 8(844), pp. 1-9.
- Yang, G., Wagner, T., Jobe, B. & Thomas, C., 2008. The Role of Positron Emission Tomography in Esophageal Cancer. *Gastrointest Cancer Res*, pp. 2: 3-9.
- Yan, J. et al., 2015. Impact of Image Reconstruction Settings on Texture Features in 18F-FDG PET. *J Nucl Med*, Volume 56, pp. 1667-1673.
- Yan, J., Schaefferkoette, J., Conti, M. & Townsend, D., 2016. A method to assess image quality for Low-dose PET: analysis of SNR, CNR, bias and image noise. *Cancer Imaging*, 16(1)(26).
- Yasaka, K. et al., 2018. Deep learning with convolutional neural network in radiology. *Japanese Journal of Radiology*, Volume 36, p. 257–272.
- Yip, S., Coroller, T., Sanford, N. & al., e., 2016. Use of registration-based contour propagation in texture analysis for esophageal cancer pathologic response prediction. *Phys Med Biol*, Volume 61, pp. 906-22.
- Yip, S. et al., 2014. Comparison of Texture Features Derived from Static and Respiratory-Gated PET Images in Non-Small Cell Lung Cancer. *PLOS One*, 9(12).
- Ypsilantis, P. et al., 2015. Predicting Response to Neoadjuvant Chemotherapy with PET Imaging Using Convolutional Neural Networks. *PLOS One*, 10(9), pp. 1-18.

- Zhang, H., Tan, S., Chen, W. & al., e., 2014. Modeling pathologic response of esophageal cancer to chemoradiation therapy using spatial-temporal 18F-FDG PET features, clinical parameters, and demographics.. *Int J Radiat Oncol Biol Phys*, Volume 88, pp. 195-203.
- Zhang, L. et al., 2015. IBEX: An open infrastructure software platform to facilitate collaborative work in radiomics. *Med Phys*, Volume 42, p. 1341–1353.
- Zhang, Y., Guo, L., Yuan, X. & Hu, B., 2020. Artificial intelligence assisted esophageal cancer management now and future. *World J Gastroenter*, 26(35), pp. 5256-5271.
- Zhuang, H. & Codreanu, I., 2015. Growing applications of FDG PET-CT imaging in non-oncologic conditions. *The Journal of Biomedical Research*, 29(3), pp. 189-202.
- Zhu, Y. et al., 2019. The value of magnetic resonance imaging in esophageal carcinoma: Tool or Toy?. *Asia-Pac J Clin Oncol*, pp. 15: 101-107.
- Zwanenburg, A., Leger, S., Vallières, M. & Löck, S., 2019. *Image biomarker standardisation initiative*, s.l.: arXiv Prepr arXiv161207003.

## 9 Appendices

### 9.1 DClSci Appendix – List of A units and Medical Physics B units together with assignments – Nicholas Vennart

<b>AMBS – A Units</b>		
<b>Unit title</b>	<b>Credits</b>	<b>Assignment wordcount</b>
A1: Professionalism and professional development in the healthcare environment	30	A1 – assignment 1 – 2500 words  Group work/presentation – 10 minutes (10%)  A1 – assignment 2 – 3000 words
A2: Theoretical foundations of leadership	20	A2 – assignment 1 – 3000 words  A2 – assignment 2 – 3000 words
A3: Personal and professional development to enhance performance	30	A3 – assignment 1 – 1500 words  A3 – assignment 2 – 4000 words
A4: Leadership and quality improvement in the clinical and scientific environment	20	A4 – assignment 1 – 3000 words  A4 – assignment 2 – 3000 words
A5: Research and innovation in health and social care	20	A5 – Group work/presentation – 15 minutes (25%)  A5 – assignment – 4000 words
<b>Medical Physics – B Units</b>		
B1: Medical Equipment Management	10	Group presentation  1500 word assignment
B2: Clinical and Scientific Computing	10	Group presentation  1500 word assignment

B3: Dosimetry	10	Group presentation 1500 word assignment
B4: Optimisation in Radiotherapy and Imaging	10	Group presentation 1500 word assignment
B6: Medical statistics in medical physics	10	3000 word assignment
B8: Health technology assessment	10	3000 word assignment
B9: Clinical applications of medical imaging technologies in radiotherapy physics	20	Group presentation 2000 word assignment
B10b: Assessment of Image Quality	10	Group presentation 1500 word assignment
B10f: Radiation Protection Advice	10	1500 word portfolio item
B10k: Radiopharmaceuticals and Radiopharmacy	10	Examination
<b>Generic B Units</b>		
B5: Contemporary issues in healthcare science	20	1500 word assignment + creative project
B7: Teaching Learning Assessment	20	20 minute group presentation

\*AMBS = Alliance Manchester Business School

## 9.2 **Lay Abstract**

We present a study investigating patients who have undergone treatment for cancer of the upper digestive system, e.g. oesophageal cancer. As part of the treatment pathway, patients receive a variety of medical imaging tests. Medical imaging is used to identify the extent of the patient's disease. In addition to the visual information from medical images, sophisticated computer analysis can be used to look at the "non-visual" information in the image such as the relationships between individual pixels in the image. We set out to investigate 3 key questions:

1. Are any of these 'non-visual' image features can be linked with patient survival, without disease returning, for 2 years (locally considered a successful treatment)?
2. How are the 'non-visual' features affected by the way the images are acquired?
3. Can we use a computer program (artificial intelligence) to predict whether the patient will survive for 2 years after treatment, based on the non-visual features of the image acquired before treatment?

We analysed the images of 144 patients with oesophageal cancer and investigated 58 non-visual image features. We found that tumours which were more non-uniform on imaging generally did not survive as long as patients with more uniform tumours. The way the images are acquired can affect the non-visual features and so caution should be used when comparing results between patients and between other studies. Using artificial intelligence for this patient group showed some early promising results whereby we were able to predict, with 83% certainty, which patients would fail treatment within 2 years, based on the initial images however further work with much larger patient groups is required to confirm this.

### 9.3 Innovation Proposal (Business Case)

#### Executive Summary

In the UK, oesophageal cancer is the 14<sup>th</sup> most common cancer but the 5<sup>th</sup> most common cause of cancer death; highlighting the relatively poor prognosis for this disease and a need for further development to improve outcomes for these patients (NHS, 2019). As part of the work up for treatment, patients undergo a variety of medical imaging procedures to help to guide clinicians in determining what type of disease the patient has and therefore, how best to treat it. The main treatment for oesophageal cancer is surgery to remove the tumour which is a highly invasive procedure with long, often futile recovery periods.

One of the main imaging modalities used to determine the best course of treatment is a “PET Scan” – a map of glucose uptake. Recently, researchers have used the data in the medical images acquired as part of routine care to perform detailed mathematical calculations on how the individual pixels in an image relate to each other. In other words, producing hundreds of parameters for a single medical image which describe underlying features of the tumour, not perceived with the naked eye such as how uniformly the glucose uptake is distributed in the tumour (“texture”). This technique is called ‘radiomics’ and has been shown to be linked to underlying genetics of a particular tumour. Several groups have shown that tumours with non-uniformly distributed uptake are more likely to respond poorly to treatment and represent poorer prognoses for these patients.

This innovation proposal forms part of a larger research project which aims to use artificial intelligence “machine learning” methods to determine how well radiomics parameters can predict 2-year DFS for patients with oesophageal cancer by examining the PET scan acquired as part of the routine patient pathway. The research project will develop an algorithm which can be applied to a PET scan and determine the likelihood of a successful treatment, based on the PET image acquired before starting treatment. This prediction value can be used to facilitate better information sharing and discussion between patients and clinicians and to help guide treatment decisions for oesophageal cancer patients where treatment is invasive and often ineffective.

This innovation proposal is to include routine medical physics expert (MPE) involvement in the radiology reporting of PET scans acquired as part of the work up for oesophageal cancer patients. This innovation proposal is to use MPE time to run the PET scan through a machine learning algorithm, developed as part of this larger research project and ultimately to add a prediction value, based on the radiomic data of the PET tumour image, estimating the chances of a 2-year DFS following treatment. It is anticipated that this additional prediction value will either give patients and clinicians greater confidence in the decision to proceed with treatment or persuade some patients against undergoing invasive and costly treatment when the prognosis is particularly poor.



## **Description and Background**

### **Oesophageal Cancer**

Despite advances in staging and treatment for oesophageal cancer, around 50% of patients are categorised as having stage IV (incurable) disease (Sah, et al., 2019). For patients diagnosed with Stage I-III disease (curative intent), prognosis remains relatively poor (Kudou, et al., 2016). Indeed the 5-year survival rate is around 17%, compared to 85% for breast cancer (NHS, 2019). For curable disease, the tumour remains confined to localised tissue around the oesophagus and any lymph nodes (if involved) are local rather than 'distant' i.e. disease has spread to other organs (American Cancer Society, 2019). Localised disease is then treated with a combination of chemo-radiation therapy (CRT) and surgery (Mayo Clinic Staff, 2019).

The initial diagnosis is usually performed using endoscopy and staging determined by a chest and abdomen computed tomography (CT), endoscopic ultrasound (EUS) and  $^{18}\text{F}$ -Fluoro-deoxy glucose positron emission tomography ( $^{18}\text{F}$ -FDG PET/CT) (Sah, et al., 2019; Stahl, et al., 2003). Where the CT scan does not show metastases, the patient proceeds to receive an  $^{18}\text{F}$ -FDG PET/CT scan which is used to further stratify and plan the most effective course of treatment (Berry, 2014; Vargese, et al., 2013). One of the key advantages of the PET information is that this allows the measurement of the underlying biochemical and physiological processes associated with the primary oesophageal cancer and gives a 'map' of the tumour glycolysis (Yang, et al., 2008). Patients identified to have disease without distant metastases are offered CRT to "shrink" the tumour before being offered surgery as a definitive treatment (American Cancer Society, 2019).

The surgery offered involves an open chest procedure which is highly invasive and carries a long recovery time and patients who do not respond to CRT are harmed by "the toxicity of these therapies without prognostic benefit" (Van Rossum, et al., 2016). It is therefore useful to accurately identify patients prior to treatment that will have a complete pathological response, widely regarded as 2 years of DFS following treatment (Ypsilantis, et al., 2015). Currently, imaging provides much of the prognostication for oesophageal cancer, traditionally using the TNM cancer staging system to describe the tumour type, nature of the nodal involvement and stratification of metastatic status.

### **Radiomics and Artificial Intelligence**

More recently, authors have used medical images to extract "innumerable quantitative features from tomographic images" by the "conversion of digital medical images into minable high-dimensional data" (Hatt, et al., 2016), more specifically, by using the concept "radiomics" to extract features describing the tumour which are related to the tumour phenotype / genotype (Aktolun, 2019; Cook, et al., 2018; Gillies, et al., 2016). In CT and MR, several authors have demonstrated a link between

radiomics parameters and the tumour phenotype (Aerts, et al., 2014; Beig, et al., 2020; Gillies, et al., 2016), in other words, the ‘hidden’ data already available in medical imaging has been shown to describe observable physical characteristics of an organism in a tumour which can distinguish it from other organisms (Honey & Shows, 1983). Radiomics allows the extraction of currently used parameters such as tumour volume and tumour signal / uptake but the key advance is in exploiting the “voxel-intensity volume histogram” (Cook, et al., 2018) which describes various relationships between individual pixels and ultimately provides parameters for describing the tumour ‘texture’ i.e. how homogeneously individual pixels are distributed within the primary tumour.

Whilst radiomics in CT and MR have already been well described in several tumour types, the literature in relation to PET imaging sparse and furthermore, the use of PET radiomics for oesophageal cancer has been further limited. PET data can provide a detailed map of glucose metabolism which may hold potential in predicting disease response with several groups showing that heterogeneously distributed tumours are linked to poor prognosis (Cook, et al., 2018; Sah, et al., 2019; Ypsilantis, et al., 2015).

More recently, the task of linking hundreds of radiomics parameters with the ultimate clinical outcome has been performed using Artificial Intelligence (AI) methodology to determine which single (or subset) of radiomics parameters best links to disease prognosis (Cook, et al., 2018; Sah, et al., 2019; Ypsilantis, et al., 2015). Artificial intelligence methods such as machine learning which describes algorithms which have been trained to learn from images and data it has been given. For example, an ML algorithm may be given 50 images and told to extract radiomics parameters, then be told which of those patients survived without disease recurrence for 2 years and determine the radiomic parameter which best describes this. The algorithm can then be given a ‘new’ image and the DFS probability determined based on the information learned in the previous 50 images.

This innovation project is a sub-strand of a larger research project “Can artificial intelligence be used to predict overall survival from a pre-treatment <sup>18</sup>F-FDG PET/CT scan for patients with oesophageal cancer?”. The main research project aims to characterise and produce the baseline testing and training dataset for using ML algorithms to perform this prediction. This innovation proposes utilising machine learning methods in the PET/CT radiology reporting clinic in conjunction with standard radiologist reporting to provide a numerical probability of 2 year DFS based on the radiomic signature of the tumour.

### **Stakeholder Engagement**

As part of this project, various stakeholders have been consulted including the surgeons managing the patient’s care, the Nuclear Medicine Radiologist reporting the PET/CT scans and the Ethics approval board. At the time of writing, the project is in progress and the results will be presented to the

Newcastle cancer treatment patient focus group for comments and feedback, specifically regarding the proposed innovation.

The clinical data underpinning this project has been collated in-conjunction with the surgeons and a database manager; patients have been appropriately stratified by the surgical team. Patients have been included in this study who presented with an adenocarcinoma of the oesophagus / gastric oesophageal junction and proceeded to surgery; NB surgery is considered as a “definitive” treatment with the patient being disease free for two years post-surgery being considered a “successful treatment” by the clinical team. The contributions from engagement with the clinical team were:

1. to provide an appropriately stratified patient cohort so as to yield a meaningful end result
2. provide guidance on what is considered “treatment” and what is considered “success”
3. to illustrate the feedback they receive from patients being that the treatment is invasive and recovery is difficult and having more information about the chances of a success will be helpful in promoting a more informed discussion between clinicians and patients.

Engagement with Radiology in accessing and processing the PET scans used has formed an essential part of this project. The key contribution from the nuclear medicine radiologist has been in ensuring the regions capturing the primary tumour are accurate and therefore the results of the ML algorithm more reliable.

The proposal is to include a machine learning generated 2-year DFS prediction value based on the radiomic signature of the tumour on PET imaging as part of the PET/CT radiology report. Further engagement with the clinicians and radiologists will be paramount to ensure the prediction value facilitates and supports the patient-clinician conversation; there is a danger that the prediction value will dissuade patients from receiving treatment who would still receive a net benefit. Engagement with the patient groups on exactly how this prediction value should be used and communicated will be essential and plans to present to patient groups are in place. This project proposes that the prediction value be used in a radiology report alongside other prognostic indicators such as nodal spread, tumour location and overall TNM scoring.

This study received favourable ethical approval from the East of Scotland Research Ethics Service REC 1 on 9th November 2020 (Ref: 20/ES/0115) and received approval from the Health Research Authority (HRA) and Health and Care Research Wales (HCRW) on 9<sup>th</sup> November 2020 (Ref: 20/ES/0115).

### **Business Case**

This innovation project proposes to include a machine learning generated 2-year DFS prediction value based on the radiomic signature of the tumour on PET imaging as part of the PET/CT radiology report.

As discussed, the prognosis for oesophageal cancer patients remains relatively poor and the treatment is both invasive, costly and carries long recovery times. Unfortunately for many patients, they will embark upon treatment only to spend their remaining time recovering and then either die from the disease, from post-surgical complications or will experience disease recurrence within 2 years necessitating palliative treatment.

This business case proposes the use of a small amount of Medical Physics Expert time at band 8A, for 30 minutes per patient to perform the machine learning analysis and participate in dual reporting sessions with the nuclear medicine radiologist. Including 15% for employer costs, the charge for band 8A MPE time is £26.87 / hour. Set up costs have been covered as part of agreed funding for this HSST research project.

**Phase 1:** All adenocarcinoma oesophagus / gastric oesophageal junction patients proceeding to definitive surgery (included within this study), estimated at 48 patients / year, requesting £1290 / year.

**Phase 2:** All upper GI patients with a PET scan to receive joint MPE / radiologist reporting, requesting £6986 per year.

The main argument for the implementation of this innovation proposal is to promote a more informed conversation between clinicians and patients however, there will likely be situations where the predictive score adds weight to the clinicians overall opinion that patients will not benefit from receiving treatment and the predictive score may help to guide that conversation. Data on exact costs for the treatment of oesophageal cancer are unavailable for UK treatments however Tranmontano et al (2019) suggest that the total cost during the surgery phase of treatment in the USA is an average of \$62,760 per patient (£45,572 on 8/9/21). A total of 144 patients from 3 years of data were included in this study; of those, 37 patients either died or had disease recurrence within 1 year, regarded by the surgical team as a failure in treatment plan and / or staging. It could be reasonably argued that these 37 patients have therefore not benefited from going through invasive treatment and therefore should have received an alternative pathway. Based on the estimates from Tranmontano et al (2019), this equates to a potential saving of £1.69M or approximately £562,000 per year. Of course cost saving should never be a driver for denying patients receiving treatment however it is anticipated that the machine learning data will be able to contribute to improved stratification and staging of patients prior to treatment and therefore treat patients more appropriately according to their exact disease, providing a cost saving as a by-product.

#### 9.4 Image and region Checklist with SUV thresholds used

HERMES Link No	PACS Req?	Anon?	PACS Transfer to Hermes	Downloaded?	Lesion Region	13/12/21 notes (after review with George)
1	y	y	y	y	SUV 4.0 Min	New region drawn
9	y	y	y	y	SUV 3.0 Min	
14	y	y	y	y	SUV 4.0 Min	
16	y	y	y	y	SUV 4.0 Min	
17	y	y	y	y	SUV 5.0 Min	
19	y	y	y	y	SUV 3.0 Min	
22	y	y	y	y	SUV 3.0 Min	New region drawn
23	y	y	y	y	SUV 2.5 Min	
24	y	y	y	y	SUV 4.0 Min	
26	y	y	y	y	SUV 2.0 Min	
27	y	y	y	y	SUV 3.5 Min	New region drawn
28	y	y	y	y	SUV 6.0 Min	
30	y	y	y	y	SUV 4.0 Min	
31	y	y	y	y	SUV 4.0 Min	
32	y	y	y	y	SUV 2.5 Min	New region drawn
33	y	y	y	y	SUV 4.0 Min	
35	y	y	y	y	SUV 3.0 Min	
38	y	y	y	y	SUV 5.0 Min	
39	y	y	y	y	SUV 4.0 Min	
40	y	y	y	y	SUV 4.0 Min	
41	y	y	y	y	SUV 4.0 Min	
42	y	y	y	y	SUV 4.0 Min	New region drawn
43	y	y	y	y	SUV 3.5 Min	
44	y	y	y	y	SUV 3.5 Min	
45	y	y	y	y	SUV 4.0 Min	
46	y	y	y	y	SUV 3.0 Min	
47	y	y	y	y	SUV 3.5 Min	
48	y	y	y	y	SUV 4.0 Min	New region drawn
49	y	y	y	y	SUV 4.0 Min	
50	y	y	y	y	SUV 3.5 Min	
52	y	y	y	y	SUV 3.0 Min	
53	y	y	y	y	SUV 4.0 Min	
58	y	y	y	y	SUV 4.5 Min	
59	y	y	y	y	SUV 3.5 Min	
60	y	y	y	y	SUV 3.0 Min	
61	y	y	y	y	SUV 3.5 Min	
62	y	y	y	y	SUV 2.5 Min	
63	y	y	y	y	SUV 4.0 Min	

64	y	y	y	y	SUV 3.0 Min	
65	y	y	y	y	SUV 3.0 Min	
66	y	y	y	y	SUV 3.0 Min	
68	y	y	y	y	SUV 4.0 Min	
70	y	y	y	y	SUV 4.0 Min	
71	y	y	y	y	SUV 2.5 Min	
72	y	y	y	y	SUV 2.5 Min	New region drawn
73	y	y	y	y	SUV 3.5 Min	
74	y	y	y	y	SUV 2.5 Min	
75	y	y	y	y	SUV 3.5 Min	
77	y	y	y	y	SUV 2.5 Min	
78	y	y	y	y	SUV 3.5 Min	
79	y	y	y	y	SUV 3.5 Min	
80	y	y	y	y	SUV 4.5 Min	
81	y	y	y	y	SUV 2.5 Min	
82	y	y	y	y	SUV 3.5 Min	
83	y	y	y	y	SUV 2.0 Min	
84	y	y	y	y	SUV 2.5 Min	
85	y	y	y	y	SUV 5.5 Min	
86	y	y	y	y	SUV 4.0 Min	
87	y	y	y	y	SUV 4.0 Min	
89	y	y	y	y	SUV 4.0 Min	
92	y	y	y	y	SUV 3.0 Min	New region drawn
93	y	y	y	y	SUV 2.5 Min	New region drawn
94	y	y	y	y	SUV 3.5 Min	
95	y	y	y	y	SUV 3.5 Min	
96	y	y	y	y	SUV 4.0 Min	
97	y	y	y	y	SUV 3.0 Min	New region drawn
98	y	y	y	y	SUV 4.5 Min	
99	y	y	y	y	SUV 3.0 Min	
101	y	y	y	y	SUV 4.0 Min	
104	y	y	y	y	SUV 2.5 Min	
105	y	y	y	y	SUV 5.0 Min	
108	y	y	y	y	SUV 3.0 Min	
109	y	y	y	y	SUV 3.5 Min	
110	y	y	y	y	SUV 4.0 Min	
111	y	y	y	y	SUV 2.5 Min	New region drawn
112	y	y	y	y	SUV 3.5 Min	
113	y	y	y	y	SUV 3.5 Min	New region drawn
114	y	y	y	y	SUV 4.5 Min	
115	y	y	y	y	SUV 5.0 Min	
116	y	y	y	y	SUV 4.0 Min	
117	y	y	y	y	SUV 5.0 Min	
118	y	y	y	y	SUV 3.0 Min	
119	y	y	y	y	SUV 3.0 Min	

120	y	y	y	y	SUV 3.0 Min	
122	y	y	y	y	SUV 4.0 Min	
123	y	y	y	y	SUV 4.0 Min	
124	y	y	y	y	SUV 4.5 Min	
125	y	y	y	y	SUV 3.0 Min	
126	y	y	y	y	SUV 5.0 Min	
127	y	y	y	y	SUV 3.0 Min	
128	y	y	y	y	SUV 3.5 Min	
129	y	y	y	y	SUV 3.5 Min	
131	y	y	y	y	SUV 2.5 Min	New region drawn
132	y	y	y	y	SUV 4.0 Min	
133	y	y	y	y	SUV 3.0 Min	
134	y	y	y	y	SUV 4.0 Min	
135	y	y	y	y	SUV 4.0 Min	
137	y	y	y	y	SUV 3.0 Min	
138	y	y	y	y	SUV 4.0 Min	
139	y	y	y	y	SUV 4.5 Min	
143	y	y	y	y	SUV 4.0 Min	
144	y	y	y	y	SUV 4.0 Min	

## 9.5 LiFEX v7.0.0 Patient data Download Code

The below code was written using MS Notepad and the .txt file loaded into the LiFEX v7.0.0 Program (Nioche, et al., 2018).

```
#texture: Common (LiFEX>=5.1.0)

LiFEX.texture.BinSize=0.3125

LiFEX.texture.NbGrey=64

LiFEX.texture.SessionCsv=C:/Users/nicholas.vennart/HSST_PROJ_PET/TextureResults.csv

#texture: Absolute (LiFEX>=5.1.0) [ true || false ]

LiFEX.texture.ButtonAbsolute=true

LiFEX.texture.MinBound=0

LiFEX.texture.MaxBound=20

#texture: RelativeMeanSd (LiFEX>=5.1.0) [ true || false ]

LiFEX.texture.ButtonRelativeMeanSd=false

#texture: RelativeMinMax (LiFEX>=5.1.0) [ true || false ]

LiFEX.texture.ButtonRelativeMinMax=false

#texture: DistanceWithNeighbours (LiFEX>=5.1.0)

LiFEX.texture.GLCM.DistanceWithNeighbours=1

#texture: dimension calculation (3D is default) (LiFEX>=5.1.0) [ 3D || 2D ]

LiFEX.texture.DimensionProcessing=3D


# Patient1

LiFEX.texture.Session0.Img0=C:/Users/nicholas.vennart/HSST_PROJ_PET/1_1_PD/

# SpatialResampling of Img0 of session0 (0 = no spatial resampling = native spacing voxels)

LiFEX.texture.Session0.Img0.ZSpatialResampling=0

LiFEX.texture.Session0.Img0.YSpatialResampling=0

LiFEX.texture.Session0.Img0.XSpatialResampling=0
```



LiFEX.texture.Session0.Roi0=C:/Users/nicholas.vennart/HSST\_PROJ\_PET/1\_1\_PD/ROI/Lesion.uint16.  
nii.gz

LiFEX.texture.Session0.Roi1=C:/Users/nicholas.vennart/HSST\_PROJ\_PET/1\_1\_PD/ROI/Liver.uint16.ni  
i.gz

.....

**Above section from “#Patient 1” repeated for all 105 patients**

.....

# Patient105

LiFEX.texture.Session104.Img0=C:/Users/nicholas.vennart/HSST\_PROJ\_PET/144\_778\_WD/

# SpatialResampling of Img0 of session104 (0 = no spatial resampling = native spacing voxels)

LiFEX.texture.Session104.Img0.ZSpatialResampling=0

LiFEX.texture.Session104.Img0.YSpatialResampling=0

LiFEX.texture.Session104.Img0.XSpatialResampling=0

LiFEX.texture.Session104.Roi0=C:/Users/nicholas.vennart/HSST\_PROJ\_PET/144\_778\_WD/ROI/Lesion  
.uint16.nii.gz

LiFEX.texture.Session104.Roi1=C:/Users/nicholas.vennart/HSST\_PROJ\_PET/144\_778\_WD/ROI/Liver.  
uint16.nii.gz

## 9.6 Raw data for the features showing the largest difference with failed treatment

Patient	SUV Max	TLG	Volume (ml)	Surface Area (mm2)	GLRLM_RLNU	NGLDM_Contrast	GLZLM_GLNU	GLZLM_ZLNU	T Score	N Score	2 year survival, normalised
1	4.33	38.54	16.92	4325.86	117.26	0.10	2.91	4.00	4	2	354
2	37.61	683.08	41.38	7655.91	545.14	1.65	12.05	390.02	3	1	78
3	17.66	125.49	12.21	2962.24	256.75	0.58	6.47	94.67	3	0	206
4	35.92	675.87	41.29	7168.75	544.16	1.53	11.51	384.20	4	2	328
5	6.28	26.43	6.09	2255.92	96.39	0.09	4.09	8.83	4	1	204
6	6.77	21.15	5.43	2068.89	96.96	0.21	3.63	11.20	3	1	364
7	6.12	23.61	6.78	2675.25	120.03	0.13	4.12	8.40	3	1	274
8	10.38	350.93	62.55	9998.92	1099.08	0.13	22.68	122.73	3	2	269
9	8.92	55.37	12.43	3366.09	243.16	0.24	8.02	41.52	3	1	22
10	37.46	1116.71	82.98	11196.29	1339.09	1.04	23.29	675.40	4	1	261
11	6.27	63.59	19.82	6159.31	328.89	0.11	11.91	25.59	3	3	275
12	6.35	31.33	7.69	2475.78	129.29	0.11	4.50	7.17	3	0	345
13	13.97	254.54	35.42	7019.65	693.85	0.26	15.77	152.62	3	2	275
14	27.90	299.16	29.96	6341.48	370.33	0.93	9.79	149.86	3	3	310
15	7.02	19.88	4.26	1534.57	80.00	0.18	3.84	14.53	3	0	353
16	19.41	576.27	79.45	13829.93	1562.24	0.20	28.64	341.33	3	2	205
17	4.81	19.34	5.95	2293.82	99.71	0.11	3.17	3.26	3	0	261
18	9.92	191.65	37.29	8862.56	427.69	0.23	15.74	51.67	4	3	253
19	16.30	787.92	102.23	14118.87	1962.29	0.19	37.04	339.89	4	3	155
20	12.54	95.35	16.21	3894.72	308.93	0.22	8.96	57.74	3	2	251
21	14.77	95.95	18.58	4318.15	170.53	0.94	6.39	62.58	3	1	97
22	11.67	250.46	37.08	6731.01	711.17	0.20	18.49	125.05	4	3	197
23	9.18	30.12	6.30	2062.61	128.56	0.30	4.70	32.44	2	0	322
24	13.77	758.04	104.49	18518.39	2018.47	0.20	42.46	375.89	4	2	286
25	11.62	134.61	21.56	5714.31	406.74	0.17	12.79	65.76	4	3	452
26	15.39	291.25	43.16	8449.31	857.73	0.25	19.01	215.92	4	3	560
27	6.77	19.42	4.35	1626.25	81.31	0.20	4.06	9.89	3	0	394
28	13.66	234.21	33.38	7033.61	665.81	0.36	16.26	165.77	3	2	386
29	6.44	23.90	5.48	2114.48	101.97	0.14	4.64	8.28	3	0	415
30	7.78	67.45	14.95	4275.07	271.24	0.15	8.24	21.00	3	1	527
31	11.38	130.71	29.63	6926.89	263.36	0.47	9.68	57.77	3	1	490
32	6.52	46.25	11.00	3738.83	200.56	0.17	8.38	20.00	3	1	405
33	14.18	672.43	152.52	21967.96	2409.64	0.07	27.36	204.73	4	3	556

34	11.18	110.23	17.34	3741.98	336.77	0.22	10.16	76.04	3	1	452
35	29.50	1445.97	120.09	15783.95	1085.38	1.39	19.92	489.91	3	2	373
36	5.02	23.92	6.48	2095.17	105.40	0.10	5.41	6.24	2	0	729
37	12.42	240.37	33.82	6661.24	648.32	0.18	17.34	108.48	3	2	718
38	15.76	359.97	52.98	10215.91	1058.72	0.33	20.36	250.85	3	3	681
39	10.74	80.69	13.30	3915.27	265.19	0.30	8.45	59.36	3	2	666
40	12.21	185.26	26.17	5208.96	520.89	0.36	13.81	119.50	3	2	638
41	16.91	565.75	62.24	10177.81	1240.37	0.33	23.91	280.61	4	3	620
42	12.71	30.53	5.13	1707.83	109.78	0.65	3.66	51.89	3	0	641
43	10.12	168.67	31.34	6500.67	603.17	0.24	16.81	89.25	2	0	634
44	5.13	61.54	16.08	4380.24	213.63	0.06	4.31	5.34	3	1	610
45	8.49	54.22	10.61	3172.37	203.34	0.34	6.54	24.57	3	2	619
46	21.14	74.84	8.78	2528.49	187.18	0.84	4.24	82.93	2	0	571
47	6.28	51.31	13.65	3787.09	224.80	0.09	6.31	9.37	4	1	577
48	34.06	150.06	11.87	2922.88	211.09	2.25	4.61	148.10	3	0	417
49	28.63	689.54	49.25	8189.28	789.70	1.34	14.56	407.60	3	1	550
50	6.88	60.31	15.47	3963.18	242.74	0.05	5.39	12.61	2	0	394
51	7.71	160.66	32.34	7160.94	576.71	0.16	17.79	58.23	4	2	485
52	8.37	30.81	6.39	1925.12	126.46	0.20	5.56	22.68	2	0	429
53	18.82	161.46	19.86	4498.61	421.02	0.55	8.99	174.02	2	0	730
54	12.45	152.70	23.30	5294.72	456.02	0.29	11.62	93.47	3	2	730
55	22.76	338.91	32.16	6899.34	681.08	0.74	12.41	299.08	4	2	730
56	7.61	24.68	6.39	2124.42	122.73	0.23	4.77	17.73	2	0	730
57	4.04	68.79	28.86	9312.45	392.77	0.06	8.88	10.68	3	0	730
58	14.71	88.20	12.13	3564.07	250.25	0.51	6.30	77.51	3	1	730
59	5.77	44.98	11.65	3699.23	190.84	0.07	5.84	12.73	4	0	730
60	9.40	84.36	18.73	4887.35	338.79	0.17	8.59	38.97	3	1	730
61	10.25	74.17	14.60	3541.52	273.71	0.15	7.37	29.68	4	2	730
62	26.57	807.45	65.76	10561.04	1245.86	0.83	20.55	503.97	3	0	730
63	13.15	91.75	15.04	3932.61	302.54	0.25	9.29	78.76	3	3	730
64	9.75	78.40	14.04	3654.13	267.40	0.18	8.31	38.97	3	1	730
65	5.91	45.78	11.65	3781.94	184.66	0.08	6.83	12.96	2	0	730
66	8.01	32.62	6.61	2167.98	121.26	0.16	5.44	13.04	3	0	730
67	10.31	30.00	5.69	1753.03	114.46	0.32	4.05	20.69	3	2	730
68	27.82	294.62	29.17	5696.88	596.18	0.73	10.48	227.70	3	2	730
69	3.05	27.91	12.65	5090.72	143.83	0.03	5.60	3.00	3	3	730
70	16.52	102.83	13.39	3128.31	275.20	0.42	7.38	87.99	4	0	730
71	8.97	308.62	70.02	12329.35	1184.76	0.11	21.46	87.81	3	0	730
72	10.81	59.30	11.95	3360.55	230.09	0.30	6.86	38.46	2	0	730
73	7.84	40.23	9.20	2828.89	109.71	0.31	5.56	24.25	3	1	730
74	5.58	20.11	5.39	2049.30	99.45	0.19	5.00	10.48	3	0	730
75	57.71	1468.86	55.94	9299.65	359.62	0.74	10.32	317.12	3	0	730
76	17.44	152.00	20.33	4663.16	263.08	0.74	7.75	112.75	3	2	730
77	15.68	407.63	58.59	9257.49	1122.25	0.24	21.44	226.18	3	2	730

78	14.39	232.40	30.03	6753.15	367.56	0.39	12.00	104.83	3	2	730
79	11.80	671.58	104.06	15435.54	1925.32	0.18	41.16	255.46	4	3	730
80	8.17	46.21	11.83	3270.37	102.48	0.39	3.72	18.32	3	2	730
81	3.17	14.42	6.61	2167.98	67.64	0.04	1.60	2.20	3	3	730
82	4.76	233.09	89.06	15942.98	1070.73	0.03	17.81	21.40	3	0	730
83	5.82	20.92	5.52	1764.79	96.80	0.12	3.00	7.52	3	0	730
84	8.82	40.16	9.29	2866.92	82.64	0.36	4.22	19.90	3	2	730
85	14.91	91.73	12.39	3156.85	258.62	0.44	6.27	75.89	3	0	730
86	15.58	48.82	6.17	1970.59	131.08	1.02	4.62	59.95	1	0	730
87	12.48	144.21	24.51	5340.12	492.80	0.37	11.81	112.70	3	0	730
88	11.50	139.18	28.73	9326.54	548.69	0.21	14.25	94.03	2	1	730
89	27.86	265.98	24.38	5349.16	488.17	1.28	7.41	220.06	3	1	730
90	14.06	108.39	14.91	3520.70	297.27	0.34	7.64	72.05	3	2	730
91	12.24	186.65	27.47	6134.47	545.03	0.26	15.11	123.66	3	1	730

## 9.7 Machine Learning Raw Code

```
# -*- coding: utf-8 -*-
```

```
"""
```

```
Created on Thu Dec 23 14:07:42 2021
```

```
@author: nicholas.vennart
```

```
"""
```

```
# Python version
```

```
import sys
```

```
print('Python: {}'.format(sys.version))
```

```
# scipy
```

```
import scipy
```

```
print('scipy: {}'.format(scipy.__version__))
```

```
# numpy
```

```
import numpy
```

```
print('numpy: {}'.format(numpy.__version__))
```

```
# matplotlib
```

```
import matplotlib
```

```
print('matplotlib: {}'.format(matplotlib.__version__))
```

```
# pandas
```

```
import pandas
```

```
print('pandas: {}'.format(pandas.__version__))
```

```

# scikit-learn
import sklearn
print('sklearn: {}'.format(sklearn.__version__))

# Load libraries
from pandas import read_csv
from pandas.plotting import scatter_matrix
from matplotlib import pyplot
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import StratifiedKFold
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC

# summarize the data
#Load data
from pandas import read_csv

names = ['SUVmin','SUVmean','SUV Stdev','SUVmax','SUV Skewness','SUV Kurtosis','SUVpeak (0.5ml)','SUVpeak (1ml)','TLG','Disc SUVmean','Disc SUV Stdev','Disc SUVmax','Disc Skewness','Disc Kurtosis','Disc SUVpeak (0.5ml)','Disc SUVpeak (1ml)','Disc TLG','Disc Histo Entopy Log10','Disc Histo Entopy Log2','Disc Histo Energy','Disc Histo AUC','Volume (ml)','Volume (Voxels)','Sphericity','Surface Area (mm2)','Compacity','GLCM Homogeneity','GLCM Energy','GLCM Contrast','GLCM Correlation','GLCM Entropy log10','GLCM Entropy Log2','GLCM Dissimilarity','GLRLM_SRE','GLRLM_LRE','GLRLM_LGRE','GLRLM_HGRE','GLRLM_SRLGE','GLRLM_SRHGE','GLRLM_LRLGE','GLRLM_LRHGE','GLRLM_GLNU','GLRLM_RLNU','GLRLM_RP','NGLDM_Coarseness','NGLDM_Contrast','NGLDM_Busyness','GLZLM_SZE','GLZLM_LZE','GLZLM_LGZE','GLZLM_HGZE','GLZLM_SZLGE','GLZLM_SZHGE','GLZLM_LZLGE','GLZLM_LZHGE','GLZLM_GLNU','GLZLM_ZLNU','GLZLM_ZP','T','N','TNM Score','SuccessFailure']

dataset = read_csv(r'C:\Users\nicholas.vennart\MLOSEM_v2.csv', names=names)
print (dataset)

# shape

```

```

print(dataset.shape)
# head
print(dataset.head(20))
# descriptions
print(dataset.describe())
# class distribution
print(dataset.groupby('SuccessFailure').size())

# box and whisker plots
dataset.plot(kind='box', subplots=True, layout=(7,10), sharex=False, sharey=False)
pyplot.show()

# histograms
dataset.hist()
pyplot.show()

# scatter plot matrix
#scatter_matrix(dataset, figsize=(30,30))
scatter_matrix(dataset)
pyplot.show()

#Set up correlation matrix
import pandas as pd
import seaborn as sn
dataset.corr()
df = pd.DataFrame(dataset,columns=names)
corrMatrix = df.corr()
print(corrMatrix)
#heatmap
sn.heatmap(corrMatrix, annot=True)
pyplot.show()
corrMatrix.to_csv(r'C:\Users\nicholas.vennart\MLOSEM_v2_correlationresults.csv', index = False)

# Split-out test/train and validation dataset, NB we have 62 columns of data (0-61) and column 62
contains success/fail data; the Y array
array = dataset.values

```

```

X = array[:,0:61]
y = array[:,61]
X_train, X_validation, Y_train, Y_validation = train_test_split(X, y, test_size=0.20, random_state=1)

# Spot Check Algorithms
models = []
models.append(('LR', LogisticRegression(solver='liblinear', multi_class='ovr')))
models.append(('LDA', LinearDiscriminantAnalysis()))
models.append(('KNN', KNeighborsClassifier()))
models.append(('CART', DecisionTreeClassifier()))
models.append(('NB', GaussianNB()))
models.append(('SVM', SVC(gamma='auto')))
# evaluate each model in turn
results = []
names = []
for name, model in models:
    kfold = StratifiedKFold(n_splits=10, random_state=1, shuffle=True)
    cv_results = cross_val_score(model, X_train, Y_train, cv=kfold, scoring='accuracy')
    results.append(cv_results)
    names.append(name)
    print('%s: %f (%f)' % (name, cv_results.mean(), cv_results.std()))

# Compare Algorithms
pyplot.boxplot(results, labels=names)
pyplot.title('Algorithm Comparison')
pyplot.show()

## Evaluation 1
# Make predictions on validation dataset
model = LogisticRegression(solver='liblinear', multi_class='ovr')
model.fit(X_train, Y_train)
predictions = model.predict(X_validation)

# Evaluate predictions
print("Linear Regression")
print(accuracy_score(Y_validation, predictions))

```

```
print(confusion_matrix(Y_validation, predictions))  
print(classification_report(Y_validation, predictions))
```

### ## Evaluation 2

```
# Make predictions on validation dataset  
model = LinearDiscriminantAnalysis()  
model.fit(X_train, Y_train)  
predictions = model.predict(X_validation)
```

#### # Evaluate predictions

```
print("Linear Discrimination")  
print(accuracy_score(Y_validation, predictions))  
print(confusion_matrix(Y_validation, predictions))  
print(classification_report(Y_validation, predictions))
```

### ## Evaluation 3

```
# Make predictions on validation dataset  
model = KNeighborsClassifier()  
model.fit(X_train, Y_train)  
predictions = model.predict(X_validation)
```

#### # Evaluate predictions

```
print("K Neighbors Classifier")  
print(accuracy_score(Y_validation, predictions))  
print(confusion_matrix(Y_validation, predictions))  
print(classification_report(Y_validation, predictions))
```

### ## Evaluation 4

```
# Make predictions on validation dataset  
model = DecisionTreeClassifier()  
model.fit(X_train, Y_train)  
predictions = model.predict(X_validation)
```

#### # Evaluate predictions

```
print("Decision Tree Classifier")  
print(accuracy_score(Y_validation, predictions))
```



```
print(confusion_matrix(Y_validation, predictions))  
print(classification_report(Y_validation, predictions))
```

### ## Evaluation 5

```
# Make predictions on validation dataset  
model = GaussianNB()  
model.fit(X_train, Y_train)  
predictions = model.predict(X_validation)
```

#### # Evaluate predictions

```
print("Gaussian")  
print(accuracy_score(Y_validation, predictions))  
print(confusion_matrix(Y_validation, predictions))  
print(classification_report(Y_validation, predictions))
```

### ## Evaluation 6

```
# Make predictions on validation dataset  
model = SVC(gamma='auto')  
model.fit(X_train, Y_train)  
predictions = model.predict(X_validation)
```

#### # Evaluate predictions

```
print("Support Vector")  
print(accuracy_score(Y_validation, predictions))  
print(confusion_matrix(Y_validation, predictions))  
print(classification_report(Y_validation, predictions))
```