The use of machine learning/deep learning in PET/CT interpretation to aid in outcome prediction in lymphoma

Russell Thomas Frood

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds
Leeds Institute of Medical Research

December 2022

The candidate confirms that the work submitted is his own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

The following joint publications have been written as a result of the work in this thesis:

i) Frood R, Burton C, Tsoumpas C, Frangi AF, Gleeson F, Patel C, Scarsbrook A. Baseline PET/CT imaging parameters for prediction of treatment outcome in Hodgkin and diffuse large B cell lymphoma: a systematic review. Eur J Nucl Med Mol Imaging. 48(10):3198-3220

   Contribution: R Frood was responsible for the creation of the review question, design, literature search, data extraction and analysis and writing of the manuscript

   Contribution of other authors: A Scarsbrook was lead supervisor and was the second author reviewing the literature. A Scarsbrook, C Burton, C Tsoumpas, A Frangi, F Gleeson and C Patel all reviewed the manuscript and provided comments

ii) Frood R, Clark M, Burton C, Tsoumpas C, Frangi AF, Gleeson F, Patel C, Scarsbrook A. Discovery of Pre-Treatment FDG PET/CT-Derived Radiomics-Based Models for Predicting Outcome in Diffuse Large B-Cell Lymphoma. Cancers (Basel). 14(7):1711.

   Contribution: R Frood was responsible for the study design, image segmentation, prediction model creation, data analysis, literature search, manuscript preparation and submission.

   Contribution of other authors: A Scarsbrook was lead supervisor. M Clark contributed to image segmentation. A Scarsbrook, M Clark, C Burton, C Tsoumpas, A Frangi, F Gleeson and C Patel all reviewed the manuscript and provided comments

iii) Frood R, Clark M, Burton C, Tsoumpas C, Frangi AF, Gleeson F, Patel C, Scarsbrook A. Utility of pre-treatment FDG PET/CT derived machine learning models for outcome prediction in classical Hodgkin lymphoma. Eur Radiol. 32(10):7237-7247.

Contribution: R Frood was responsible for the study design, image segmentation, prediction model creation, data analysis, literature search, manuscript preparation and submission.

Contribution of other authors: A Scarsbrook was lead supervisor. M Clark contributed to image segmentation. A Scarsbrook, M Clark, C Burton, C Tsoumpas, A Frangi, F Gleeson and C Patel all reviewed the manuscript and provided comments

Acknowledgements

I would like to express utmost gratitude to my supervisors Professor Andrew Scarsbrook, Professor Alejandro Frangi, Professor Charalampos Tsoumpas and Professor Fergus Gleeson for all the support and guidance throughout the fellowship.

I would like to thank Rachel Hyland, Head of School for Radiology Yorkshire, and Hannah Lambie, Training Programme Director at the West Yorkshire Radiology Academy, for supporting my out of programme research during my Radiology training.

I would like to thank Atif Rabani, Matthew Clark and Chirag Patel with their help in data acquisition and reviewing segmentations.

I would also like to thank Cathy Burton and the University of York's Haematological Malignancy Research Network for providing follow up data for the patients included in the study.

I would like to thank my wife Lucy Elliott. I could not have done this without her support and proof reading skills.

I would like to dedicate this thesis to my late parents, Yvonne and Colin Merrick, who sadly passed away before I could complete my PhD, and to my daughter Erin Yvonne Elliott-Frood who was born during my PhD.

Abstract

Lymphoma is a haematopoietic malignancy consisting of two broad categories: Hodgkin lymphoma (HL) and non-Hodgkin lymphoma (NHL). These categories can be further split into subtypes with classical HL (cHL) and diffuse large B cell lymphoma (DLBCL) being the commonest subtypes. The gold standard imaging modality for staging and response assessment for cHL and DLBCL is 2-deoxy-2-[fluorine-18]fluoro-D-glucose (FDG) positron emission tomography/computed tomography (PET/CT), with patients having a worse prognosis if they do not demonstrate complete metabolic response (CMR). However, approximately 15% of patients will relapse even after CMR. Therefore, being able to identify patients who are likely to relapse it may be possible to stratify treatment early to improve patient outcomes. The aim of this project is to develop and test image derived predictive models based on the baseline PET/CT to risk stratify patients pre-treatment.

Table of Contents

List of Tables

List of Figures

Abbreviations

3D	3-dimensional
aa-IPI	age adjusted - International prognostic index
ABC	Activated-B-cell
ABVD	Doxorubicin (adriamycin), bleomycin, vinblastine and dacarbazine
AE	Application entity
AGM	Aorta–gonad–mesonephros
ALCL	Anaplastic large cell lymphoma
ANN	Artificial neural network
BEACOPP	Bleomycin, etoposide, doxorubicin (adriamycin), cyclophosphamide, vincristine (oncovin), procarbazine and prednisone
BCL	B-cell lymphoma
BCL	B-cell lymphoma extra-large
BCR	B-cell receptors
BMB	Bone marrow biopsy
BMI	Body mass index
BQML	Becquerel/millilitre
CD	Cluster of differentiation
cHL	Classical Hodgkin lymphoma
CHOP	Cyclophosphamide, doxorubicin, hydrochloride vincristine (oncovin) and prednisolone
CLL	Chronic lymphocytic leukaemia
CLP	Common lymphoid progenitor
CMP	Common myeloid progenitor

CNN	Convolutional neural network
CNS	Central nervous system
COO	Cell of origin
CT	Computed tomography
DC	Dendritic cells.
DICOM	Digital imaging and communications in medicine
DEL	Double expression lymphoma
DHL	Double hit lymphoma
DLBCL	Diffuse large B-cell lymphoma
DNA	Deoxyribonucleic acid
DS	Deauville score
eBEACOPP	escalated-Bleomycin, etoposide, doxorubicin (adriamycin), cyclophosphamide, vincristine (oncovin), procarbazine and prednisone
EBF	Early B-cell factor
EBV	Epstein-barr virus
ECOG	Eastern Cooperative Oncology Group
EFS	Event free survival
EORTC	European Organisation of Research and Treatment of Cancer
ESR	Erythrocyte sedimentation rate
FDG	2-deoxy-2-[fluorine-18]fluoro-D-glucose
FOXO	Forkhead box o
GCB	Germinal-centre-cell
GHSG	German Hodgkin Study Group
GELA	Groupe d'Etudes des Lymphomes de l'Adulte
GLCM	Grey-level co-occurrence matrix

GLDZM	Grey-level distance zone matrix
GLRLM	Grey-level run-length matrix
GLUT	Glucose transporters
GLSZM	Grey-level size zone matrix
HIV	Human immunodeficiency virus
HL	Hodgkin lymphoma
HLA	Human leukocyte antigen
HRS	Hodgkin and Reed-Sternberg
HSC	Haematopoietic stem cell
HU	Hounsfield units
IBSI	Image biomarker standardisation initiative
ID	Identifier
IgD	Immunoglobulin D
IgH	Immunoglobulin H
IgGM	Immunoglobulin M
IL-3R	interleukin-3 receptor
IL-3Rα	interleukin-3 receptor subunit alpha
IL-6	Interleukin-6
IP	Internet protocol
IPI	International prognostic index
IPS	International prognostic score
KNN	K-nearest neighbour
LDH	Lactate dehydrogenase
LMPP	Lymphoid primed multipotent progenitor
MPP	Multipotent progenitors
MYC	Myelocytomatosis-cellular

NCCN	National comprehensive cancer network
NCI	National cancer institute
NGLDM	Neighbouring grey-level-dependence matrix
NGTDM	Neighbourhood grey-tone difference matrix
NIfTI	Neuroimaging informatics technology initiative
NHL	Non Hodgkin lymphoma
NOS	Not otherwised specified
OS	Overall survival
PACS	Picture archiving and communication system
PAX	Paired box
PMBL	Primary mediastinal large B-cell lymphoma
PET	Positron emission tomography
PFS	Progressive free survival
RCHOP	Rituximab, cyclophosphamide, doxorubicin, hydrochloride vincristine (oncovin) and prednisolone
R-IPI	Revise international prognostic index
ROI	Region of interest
RQS	Radiomic quality score
Sca	Stem cells antigen
SCU	Service class user
SCP	Service class provider
SOP	Service-object pair
SQL	Structured query language
SUV	Standardised uptake value
SVM	Support Machine Vector
TARC	Thymus and activation regulated chemokine

THRLBCL	T-cell/histocyte rich large B-cell lymphoma
TRIPOD	Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis
TUNEL	Terminal deoxynucleotidyl transferase-mediated mdeoxyuridine triphosphate-biotin nick-labelling
UID	Unique identifier
Vcam	Vascular cell adhesion molecule
WBC	White blood cell

Chapter 1
Introduction

Lymphoma is a haematopoietic malignancy affecting lymphocytes and their progenitors. There are two main categories: Hodgkin lymphoma (HL) and non-Hodgkin lymphoma (NHL), with approximately 90% of cases within the adult population being NHL [1]. HL is further divided into classical (cHL) and nodular lymphocyte predominant groups, with 95% of HL cases being cHL [1]. cHL comprises four distinct histological subtypes: nodular sclerosis, mixed cellularity, lymphocyte predominant and lymphocyte depleted. NHL consists of approximately 50 different subtypes, the most common of which is diffuse large B-cell lymphoma (DLBCL) representing around 30-40% of adult NHL [2]. DLBCL is sub-classified into a further 13 different variants based on morphological, clinical and immunophenotypic findings, the most common, representing 80% of cases being not otherwise specified (NOS) [3]. Given the large variation in histological subtypes this thesis will focus on cHL and DLBCL NOS which represent the commonest types of HL and NHL affecting patients.

1.1 Normal B-cell Lymphopoiesis

Both cHL and DLBCL malignant cells are derived from the B-cell linage. Understanding normal B-cell lymphopoiesis is key to understanding the genetic phenotyping of the different subtypes of cHL and DLBCL [4].

During development haematopoiesis first occurs within the yolk sac, where primitive erythrocytes are produced. It is unclear if lymphopoiesis is possible during this phase [5]. After approximately day 19 of embryogenesis, a second wave of haematopoiesis takes place in the aorta–gonad–mesonephros (AGM). The AGM is the site of origin of haematopoietic stem cells (HSC), which are derived from haemopoietic progenitors, and are the source of all types of mature blood cells [6]. At approximately week 5 of gestation the HSC and progenitors migrate to the liver which takes over as the main centre of haematopoiesis [6]. Differentiation and expansion of the stem cells takes place during this period. The placenta and spleen also contribute to haematopoiesis at this stage, but to a lesser extent. At approximately 4 months gestation haematopoiesis starts in the red bone marrow, which by birth becomes the sole site of haematopoiesis. During normal ageing, the volume and location of red marrow decreases, being replaced by yellow marrow, which is made up of adipocytes. In adulthood, the remaining centres of haematopoiesis are located in the axial skeleton and the metaphysis of long bones.

B-cells can be divided into two main sub-linages, B1 and B2 B-cells, with B2 cells forming most of all B-cells in the human body. B2 cells represent the transitional, follicular, germinal centre, plasma and memory B-cells which form part of the adaptive immune response. B1 cells are mainly found in the peritoneal and pleural cavities and produce natural antibodies as part of the innate immune system. There are mixed reports within the literature regarding the origin of B cells, with a lineage model, selection model and combined layer model all being proposed. The linage model posits that B1 and B2 cells are defined distinct separate linages. This is supported by murine studies demonstrating that whilst transplanted foetal liver is able to reconstitute both B1 and B2 cells, adult bone marrow only forms B2 cells [7, 8]. The selection model suggests that B1 cells are formed, like B2 cells, from the HSC and that these cells can differentiate into either B1 or B2 cells. The selection model is supported by the fact that swapping B1 and B2 B-cell receptors (BCR) it is possible to change a B2 into a B1 cell which will promptly migrate to the pleural and peritoneal cavities [9]. The combined layer model suggests that foetal liver and bone marrow have the capacity to produce both B1 and B2 cells, but foetal liver has a higher predilection for producing B1 cells and bone marrow has a higher predilection for producing B2 cells [10].

The concept of the nature of a lymphoid or myeloid cell being predetermined from the first branch of the HSC has changed [11]. HSC were also once considered to be made up of long-term (cluster of differentiation (CD)34-) and short-term (CD34+) variants, with long-term variants more likely to self-renew and short-term more likely to differentiate further down a cell lineage [12]. However, it is now believed that HSCs are more complex in nature with an intermediate-term HSC demonstrated [13], and the HSC pool made up of HSCs and multipotent progenitors (MPP) which are more biased to certain linages [14] (Figure 1.1).

There are many factors which have an influence on actions of HSCs, with signals arising from cells in arterioles and sinusoids as well as other HSCs themselves. These are all transmitted to the HSC whilst it is within a dedicated bone marrow environment known as the bone marrow niche. As cells differentiate, they present different surface antigen markers as cluster of differentiation (CD) (Table 1.1). The CD can be used in conjunction with histology to determine specific subtype of lymphoma. For example, where there is histological uncertainty between NS HL and anaplastic large cell lymphoma (ALCL) the presence of CD15 and lack of the T-cell receptor gene would indicate cHL rather than ALCL [15].

During B-cell differentiation the BCR is formed from two heavy and two light chains. The

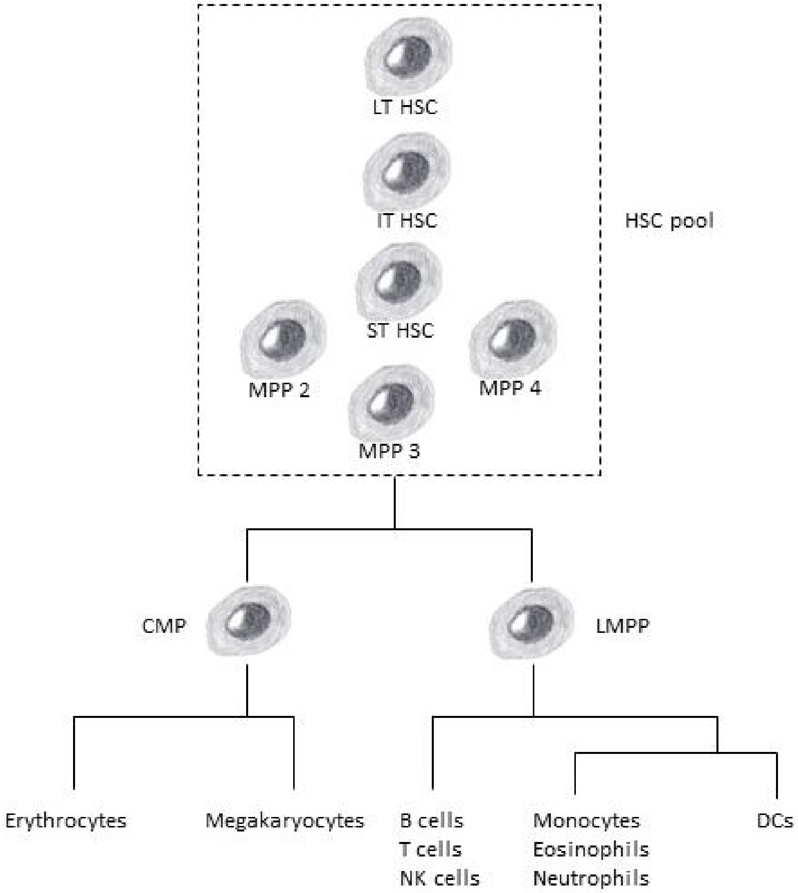


Figure 1.1 Flow diagram of haematopoiesis. HSC = haematopoietic stem cell, LT HSC = long-term haematopoietic stem cell, IT HSC = intermediate-term haematopoietic stem cell, ST HSC = short-term haematopoietic stem cell, MPP = multipotent progenitor, LMPP = lymphoid primed multipotent progenitor, common myeloid progenitor (CMP), DCs = dendritic cells.

heavy chains are made up of three gene segments: variable (V), diversity (D) and joining (J), whereas the light chains are formed from only two segments: V and J [16]. During the early pro B-cell stage there is random rearrangement of the D and J aspects of the heavy chain to try and join them together, and if this is successful, the cell progresses to the late pro B-cell stage [17]. In the late pro B-cell stage the V attempts to join to the V-J complex. At the large B cell stage the heavy chain is paired with a surrogate light chain creating a pre-BCR within the cytoplasm of the cell. The pre-BCR when signalled causes the cell to proliferate which in turn will allow different combinations of light chains to the heavy chain which was created and "tested" during the initial stages. The small pre B-cell begins the process of light chain rearrangement [16]. There are two types of light chain, kappa and lambda, and the cell will try to pair a kappa light chain with the heavy chain

Cell	Surface Marker
Haematopoietic stem cell	CD117 hi, Sca-1hi, CD135−, CD34lo, CD150+, CD48−
Common Lymphoid Progenitor	CD10+, CD34+, Pax5+
lymphoid primed multipotent progenitor	CD117hi, Sca-1hi, CD135hi, CD27+, Vcam-1−
Early progenitor B-cell	CD117lo, CD10+, CD34+, CD38+, Pax5+
Late progenitor B -cell	CD117lo, CD10+, CD19+, CD20+, CD24+, CD34+, CD38+, CD93+, IL-3R+, IL-7R +, Pax5+
Large precursor B-cell	CD19+, CD43−, internal IgH+, surface IgM−
Small precursor B-cell	CD19+, CD43−, internal IgH+, Surface IgM−
T1	CD19+, CD24hi, CD38hi, CD27IgMhi, IgDlo, CD10hi, CD21lo, CD32hi
T2	CD19+, CD24hi, CD38hi, CD27IgMin, IgDin, CD10in, CD21lo, CD32in
T3	CD19+, CD24hi, CD38hi, CD27IgMlo, IgDlo, CD10lo, CD21lo, CD32lo, CD27+
Activated B-cell	CD19+, CD80+, CD86+, CD44+, CD69+, PD-L1+
Plasma Cell	CD20-, CD24-, CD27hi, CD38hi

Table 1.1 Surface and transcription markers associated with different stages of B-cell development. CD = cluster of differentiation, Sca-1 = stem cells antigen-1, Pax5 = Paired Box-5, Vcam-1 = vascular cell adhesion molecule-1, IL-3R = interleukin-3 receptor, IL-7R = interleukin-3 receptor subunit alpha, IgH = Immunoglobulin heavy, IgM = Immunoglobulin M, IgD = Immunoglobulin D.

in the first instance, if this fails the cell will try to form a pair with the lambda light chain. If both fail the cell undergoes cell death by apoptosis. If successful it forms an immature B-cell expressing an IgM antigen surface receptor. Immature B-cells undergo several transitional phases before eventually progressing to naïve B-cells within the secondary lymphoid tissues. The naïve B-cell expresses both IgM and IgD antigen surface receptors and does not become a mature B-cell until exposed to its antigen. If a transitional B-cell demonstrates high affinity for self-antigens, the cell undergoes apoptosis. Mature B-cells become activated when an antigen encounters their surface IgM or IgD surfacer receptor.

1.2 Classical Hodgkin Lymphoma

1.2.1 Epidemiology

There is a general trend of HL to be more prevalent in Western societies. A study by Singh *et al.* using GLOBOCAN 2020 data reported the highest incidence for men was in Italy, with an age-standardised rate (ASR) of 3.5-4 per 100,000, the lowest incidence was in the Caribbean island of Martinique with an ASR of 0.25 per 100,000 [18]. The highest incidence for women was in Cyprus (ASR 3.2 per 100,00), and lowest in Niger (ASR <0.1 per 100,000). However, the highest mortality rates were demonstrated in Iraq for both sexes highlighting a potential disparity between treatment availability in different regions.

Singh *et al.* predict a global 30% increase in incidence of HL by 2040 and a 50% increase in mortality rate based on predicted demographic changes.

As well as global variation in incidence, there are historically four different age-related incidence patterns which vary depending on a country's ethnic and socio-economic make up. These patterns were first reported in studies performed between 1950 and 1970 [19, 20]. The first is demonstrated in developing countries where there is a peak incidence of HL cases among young boys, and then a low incidence until the 4th decade of life when there is an increasing incidence with age. The next is demonstrated in affluent western countries where there is a bimodal distribution with a relatively low incidence in childhood with a first peak incidence of nodular sclerosing (NS) HL in adolescent/early adults and another peak in older adults. This pattern of disease is typically demonstrated within the UK, with NS HL accounting for 70% of cases. A third pattern, which lies between that of developing and affluent western countries, typically demonstrated within rural Western areas. Finally, a fourth pattern exists in Asian countries, where there is low incidence through the first four decades of life, and after this incidence increases with age.

1.2.2 Clinical Presentation

cHL has a highly variable clinical presentation [21], the most common being a painless lump or swelling representing an enlarged lymph node or confluent mass of lymph nodes. This often occurs in the neck but can occur anywhere in the body [22]. Patients can also present with symptoms related to mass effect (e.g. superior vena cava obstruction, cauda equina, etc.) or with systemic symptoms such as fatigue, pyrexia, weight loss, night sweats, pruritus or increased frequency of infections. Since the Ann Arbor staging system was introduced in 1971, night sweats, pyrexia and weight loss are referred to as B-symptoms [23]. Prior to this, pruritus rather than weight loss was included as a B-symptom. The cause of the B-symptoms is believed to be due to raised inflammatory proteins such as cytokines, with serum levels of interleukin-6 (IL-6) correlating with their presence [24]. The probability of a patient presenting with B-symptoms varies depending on histological subtype.

1.2.3 Histology

HL, and its subtypes, are characterised by presence of Hodgkin and Reed-Sternberg (HRS) cells [15]. Reed-Sternberg cells are large binucleated cells, with the nucleoli often being eosinophilic (Figure 1.2), formed from the incomplete cytokinesis of Hodgkin cells.

Hodgkin cells are formed from clonal transformation of mature B-cells with evidence of somatic mutation in the immunoglobulin genes which indicates they are post-germinal centre B-cells [15].

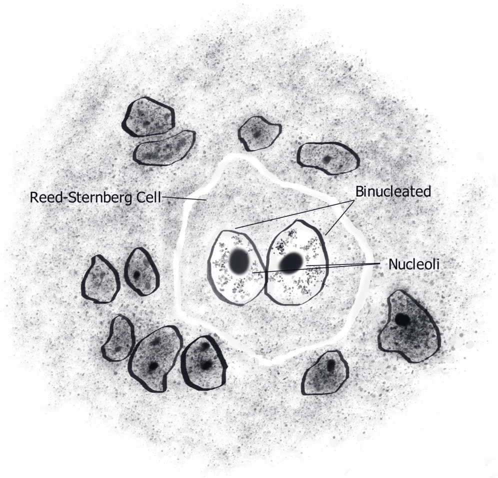


Figure 1.2 Graphical representation of a histological slide depicting the binucleated Reed-Sternberg cell surrounded by background eosinophils.

Nodular sclerosis HL is identified histologically by the presence of both HRS cells and fibrous tissue with sclerosing bands within the lymph node. Mixed cellularity cHL has HRS cells with a mixed background of inflammatory cells; fibrosis may be present but there are no sclerotic bands. Histologically lymphocyte rich cHL is characterised by HRS cells which are distributed at the edge of reactive follicles [25]. The appearances of lymphocyte rich cHL can often resemble those of nodular lymphocyte predominant lymphoma, however, the two can be distinguished by immunophenotypic features. It is believed that approximately 30% of historically diagnosed nodular lymphocyte predominant lymphoma cases were actually lymphocyte rich cHL.

Lymphocyte depleted cHL has two histological variations: a reticular type and a diffuse fibrotic type [26]. The diffuse fibrotic type consists of HRS cells with few lymphocyte infiltrations and increased histocytes and fibroblasts. The reticular type again demonstrates HRS cells with only a few infiltrating lymphocytes; however, the HRS cells can demonstrate sheet like proliferation. Histologically this can make the reticular type difficult to distinguish from DLBCL [26]. Although histologically these subtypes of cHL are distinct they are all imaged and treated in a similar way.

1.2.4 Risk factors

A UK based population cohort study by Rafiq *et al.* demonstrated a link between socioeconomic deprivation and incidence of HL (both cHL and nodular lymphocyte predominant), and that there was 60% higher incidence in more affluent areas [27]. However, patients with HL from deprived areas were more likely to have Epstein Barr virus (EBV)-positive disease. As previously discussed, there is a relationship between age and development of cHL and in the UK this follows a bimodal distribution with peaks in early (age 20-24 years) and late adulthood (age 75-79 years) (Cancer Research UK). This pattern of incidence may also be linked to race; an American study by Shenov *et al.* found the bimodal pattern held true for Caucasian and Asian populations with cHL but was not observed in black populations [28]. They proposed that these variations in distribution, and the higher incidence in white populations, could be related to differences in early microbiome exposure. Chang *et al.* further explored the link between childhood exposure to pathogens and development of HL in a population-based case-control study of 565 HL cases and 679 controls [29]. They found that attendance at nursery or day school was associated with lower incidence of HL in early adulthood.

cHL can be associated with EBV, with 80-100% of patients with human immunodeficiency virus (HIV) also having EBV, the association between EBV with cHL was first implied when raised EBV antigens were detected prior to the development of HL [30]. This was later confirmed when EBV RNA and DNA was detected in HRS cells. EBV-positive childhood cases are thought to be a direct reaction to exposure to EBV whereas HIV related cHL and EBV positive cases in the older population are thought to be due to decreasing immunity [31].

1.2.5 Clinical prognostic markers

Different pathways are influenced by the classification of cHL disease as either early or advanced. Early-stage disease consists of stage I and II patients, and can be further divided into favourable or unfavourable categories. The definition of early favourable and unfavourable varies depending on which of the many scoring systems is used (Table 1.2). This was highlighted by a study by Advani *et al.* who compared three different scoring systems: German Hodgkin Study Group (GHSG), European Organisation of Research and Treatment of Cancer (EORTC) and Groupe d'Etudes des Lymphomes de l'Adulte (GELA) when applied to a cohort of patients who had been classified as early favourable using the National Cancer Institute (NCI) score. Patients are regarded as being favourable using the NCI score if they do not have bulky disease (>10cm in maximum diameter) or B-symptoms [32]. They found that there was no significant difference in the prediction of OS and that there was only a significant in PFS when using the GHSG scoring system. Further work needs to be performed to determine a universal approach to the classification of favourable vs unfavourable.

	NCRI	**EORTC**	**GHSG**	**NCCN 2010**	**GELA**
Risk Factors	Bulky disease	Large mediastinal mass	Large mediastinal mass	Large mediastinal mass	Male
	B symptoms	Age $\geq$ 50	Extra nodal disease	>1 extra nodal area	Age $\geq$ 45
		ESR $\geq$ 50 without B-symptoms or $\geq$ 30 with	ESR $\geq$ 50 without B-symptoms or $\geq$ 30 with	ESR $\geq$ 50 or any B symptoms	Any increase in ESR
		$\geq$ 4 nodal areas	$\geq$ 3 nodal areas	$\geq$3 nodal areas	haemoglobin 105 g/L
					lymphocyte count $0 \cdot 6 \times 109$/L
Favourable	Stage I-II without risk factors	Stage I-II (supradiaphragmatic) without risk factors	Stage I-II without risk factors	Stage I-II without risk factors	Stage I-II without risk factors

Table 1.2 Different scoring systems for determining early favourable disease. NRCI = National Cancer Research Institute, EORTC= European Organization for Research and Treatment of Cancer, GHSG= German Hodgkin Study Group, NCCN = National Comprehensive Cancer Network, GELA = Groupe d'Etude des Lymphomes de l'Adulte, ESR = erythrocyte sedimentation rate.

The International Prognostic Score (IPS) was developed by Diehl and Hasenclever in 1998 as a way to prognosticate advanced HL [33]. The score was based on data from 5151 patients using seven predictors, with a patient scoring a point for each of the following: age greater than 45 years, male gender, albumin of less than 40g/L, haemoglobin of over 105 g/L, stage IV disease, white blood count (WBC) more than 15,0000mm3 and lymphopenia. The original study split the cohort into six distinctive prognostic groups with scores 0 to 4 having distinct groups and a score of 5 or greater having its own group. However, the OS curves for scores 0 and 1, and 2 and 3 were similar [33]. The IPS was re-evaluated by Deinfenbach *et al.* in a more recent patient cohort, as it was suggested that 20% of the original study cohort were treated with outdated treatment regimes [34]. They reported that only stage and age were predictive in multivariate analysis for PFS and age, stage and haemoglobin level were predictive for OS and proposed a new scoring system based on these factors. On their data the new clinical prediction model outperformed the original IPS. However, with subsequent widespread implementation of image-guided treatment adaption it has been suggested in more recent studies by Gallamini *et al.* and Bari *et al.* that only the Deauville score (DS) from interim 2-deoxy-2-[Fluorine-18]fluoro-D-glucose (FDG) positron emission tomography/computed tomography (PET/CT) is prognostic of 2-year PFS and that IPS has lost its predictive value [35, 36]. This suggests that there is value in exploring further prognostic features and models, for which imaging could play a role.

1.2.6 Imaging of classical Hodgkin lymphoma

FDG PET/CT is now widely accepted as the gold standard imaging technique for staging and response assessment in cHL. FDG is a glucose analogue actively taken up intracellularly by glucose transporters (GLUTs). Once inside the cell it is phosphorylated with hexokinase to glucose-6-phosphate [37]. The expression of GLUT and hexokinase is upregulated in cancer cells. In normal cells glucose-6-phosphate can be dephosphorylated by glucose-6-phosphatase which allows FDG to exit the cell, however, in malignant cells glucose-6-phosphatase is down regulated and therefore FDG accumulates [37]. Fluorine-18 decays by positron emission, the positron travels a short distance before annihilation with an electron causing the emission of two photons, which occur at near 180 degrees from each other [38]. Detection and timing of these pairs of photons allows localisation of the origin and therefore the location of cells where FDG is aggregating [38]. CT is used for attenuation correction and to provide accurate anatomical localisation which can be combined with the physiological aspect from the PET. An example of a cHL PET/CT study is demonstrated in Figure 1.3.

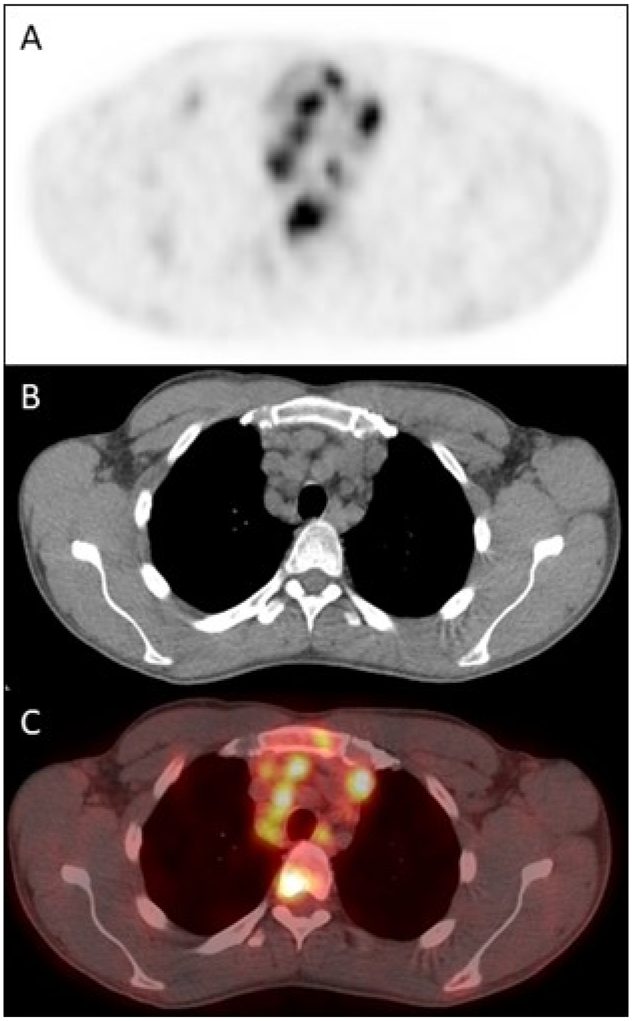


Figure 1.3 Select axial PET (A), CT (B) and fused images (C) from a pre-treatment FDG PET/CT study in a patient with stage 4 cHL demonstrating tracer-avid mediastinal lymphadenopathy and sternal and vertebral body lymphomatous disease.

cHL is clinically staged using the modified Ann Arbor staging classification. Ann Arbor classification was originally adopted for use in 1971 as a replacement to a pathological staging system which consisted of laparotomy and multiple nodal, organ and bone marrow biopsies (BMB) following a transition to multiagent chemotherapy where it was deemed unnecessary to have pathological staging [23]. CT staging was adopted as the modality of choice, however, due to low sensitivity for detecting bone marrow involvement BMB was continued. In 1989 the Ann Arbor staging classification was modified at the Cotswold meeting [39]. The new modification added X, S, E, A and B designations which represent bulky disease, splenic involvement, extra nodal disease and the presence and absence of B symptoms, respectively. The introduction of metabolic imaging occurred with the use of gallium-67 scintigraphy, but this was later replaced by FDG PET/CT which was able to demonstrate better ability to detect disease when compared to gallium-67, CT and non-targeted bone biopsy [40]. The Lugano classification is the staging method most commonly adopted and varies when compared to the Cotswold modification system by grouping stage I and II as limited disease and stage III and IV disease as advanced, the grouping of stage III disease and the X descriptor is not applied for bulky disease, but the longest diameter mass is recorded [41].

In terms of response assessment, a 5-point scale (Deauville Score, DS) is used when assessing interim (following 2 cycles of chemotherapy) or end of treatment FDG PET/CT [42]. The DS groups patients into complete response, partial response, stable disease, and progressive disease depending on the metabolic activity within sites of disease in comparison to background FDG uptake in the mediastinal blood pool and liver or the presence of new disease (Table 1.3). Response assessment is based solely on FDG uptake and not on residual soft tissue masses on the CT component of the study.

Deauville Score	**PET/CT Finding**
1	No uptake
2	Uptake equal or below mediastinal blood pool
3	Uptake equal to below liver uptake but above mediastinal blood pool
4	Uptake above liver uptake
5	Markedly increased uptake or any new lesions

Table 1.3 The Deauville criteria for determining response assessment on PET/CT.

1.2.7 Baseline Imaging Prognostic Markers

Standardised uptake value (SUV) is the most common metric extracted from PET imaging data. This parameter was designed to try and compensate for tracer tissue distribution variation, most significantly influenced by injected radiopharmaceutical dose and patient body weight. SUV is defined as the ratio of measured radioactivity within an image at a given timepoint when compared to the whole body concentration of injected radioactivity [43]. There are several limitations to the standard calculation of SUV. One of these is that body fat contributes to body weight but does not have significant FDG uptake when in a fasting state as it is less metabolically active than muscle, which means that obese patients can have a higher measured SUV [44]. To compensate for this, lean body weight can be estimated, often based on predictive equations according to height, sex and age. SUV calculated from surface area can also be calculated [44]. High blood glucose levels can reduce uptake of FDG into metabolically active cells due to competition for the same cellular uptake mechanism leading to reduced intracellular FDG accumulation [44]. There are also technical factors which influence SUV, including scanner spatial resolution, image acquisition and PET reconstruction parameters [44, 45].

Different iterations of SUV within a defined region can be used in the assessment of disease [46]. SUVmax and SUVmean are the maximum and mean values within a defined region of interest (ROI) respectively. SUVpeak is the average SUV of a ROI centred on the highest uptake region within a contoured area. There a number of definitions of SUVpeak as the value can be affected by the size and shape of the ROI. SUV forms the basis of other metabolic parameters such as metabolic tumour volume (MTV) and total lesion glycolysis (TLG). MTV is the volume of contoured disease at a specified SUV threshold, whereas TLG is MTV multiplied by SUVmean. Textural parameters can also be extracted from PET/CT data. A more in-depth exploration of baseline prognostic markers is provided in Chapter 2.

1.2.8 Treatment of classical Hodgkin lymphoma

Chemotherapy is the mainstay of first-line treatment in HL. The most common regimes used are doxorubicin (Adriamycin), bleomycin, vinblastine and dacarbazine (ABVD), or bleomycin, etoposide, doxorubicin (Adriamycin), cyclophosphamide, vincristine (Oncovin), procarbazine, and prednisone (BEACOPP). However, the number of cycles varies depending on prognostic score, patient factors, and initial treatment response. Recent guidelines produced by the British Society of Haematology recommend that patients with early favourable disease are treated with 2-3 cycles followed by radiotherapy if they have a negative interim PET/CT scan [15]. A negative interim PET/CT is regarded as DS of 3 or below.

Treatment for early disease without a negative interim PET CT is two cycles of escalated BEACOPP (eBEACOPP) followed by radiotherapy. Patients with advanced disease can be treated with ABVD or BEACOPP depending on patient factors and the balance of toxicity versus efficacy. eBEACOPP should not be given to patients older than 60 years of age. Treatment pathways are based on the RATHL and HD18 trials [47, 48]. Patients treated with ABVD who have a negative interim PET/CT should have their treatment de-escalated for the remaining four cycles. If the interim PET/CT is positive, and there is no evidence of progression, four cycles of eBEACOPP can be given and radiotherapy should be considered. For patients receiving eBEACOPP, if they have a negative interim PET/CT they should only have two further cycles of chemotherapy or can be de-escalated to four cycles of ABVD. If the interim PET/CT is positive, the patient should have a further four cycles of eBEACOPP.

1.2.9 Genetic markers

Variation associated with the human leukocyte antigen (HLA) region of the human genome have been identified which are associated with an individual's susceptibility to develop cHL [49]. Depending on the specific allele, these may be protective or lead to vulnerability for the development of a specific subtype or for all cHL variations.

When it comes to predicting outcomes, Montalbán *et al.* investigated the predictive value of forty genetic markers in a cohort of 259 patients [50]. They found that three molecular markers p53, B-cell lymphoma extra-large (Bcl-XL) and terminal deoxynucleotidyl transferase-mediated deoxyuridine triphosphate-biotin nick-labelling (TUNEL) were independent predictors of complete remission for 12 months or greater. Plattel *et al.* looked at use of serum levels of soluble Galectin-1, soluble CD163 and soluble CD30 when compared to thymus and activation regulated chemokine (TARC). They measured levels of each of these potential biomarkers pre and post treatment and found that only TARC varied with treatment response with 6/7 non-responders having a level which remained high and 95/96 who responded having a significant decrease in levels [51].

1.3 Diffuse Large B-Cell Lymphoma

1.3.1 Epidemiology

The annual incidence of DLBCL within the UK is 8.2 per 100,000, with a slightly higher predominance in males [52]. The incidence of DLBCL increases with age with the median age of diagnosis being 69.7 years [52]. There is a higher proportion of patients with DLBCL who are Caucasian when compared to other ethnicities [53]. There is a paucity of data regarding global patterns of incidence of DLBCL, however, there are higher rates of NHL within western countries with the lowest rates in Middle Africa and Central America [54].

1.3.2 Clinical presentation

Clinical presentation of DLBCL is similar to that of cHL with around 20% of patients presenting with B-symptoms. DLBCL is typically more aggressive with approximately 40% presenting with extra-nodal disease and 23% presenting with two or more extra-nodal sites of involvement [55]. DLBCL can also present following transformation of another lower-grade B-cell neoplasm. This is termed Richter's transformation when B-cell chronic lymphocytic leukaemia (CLL) transforms into DLBCL, and occurs in approximately 2-10% of CLL patients [56].

1.3.3 Histology

In general, DLBCL NOS is characterised histologically by a diffuse growth pattern with the presence of large cells which typically are five time larger than normal lymphocytes and resemble immunoblasts or centroplasts [57]. However, it is ultimately a diagnosis of exclusion where the characteristics do not fit with a specific primary site such as CNS, cutaneous or intravascular and there are no features of T-cell/histocyte rich large B-cell lymphoma (THRLBCL) or mediastinal large B-cell lymphoma (PMBL).

1.3.4 Risk factors

Although the exact aetiology of DLBCL is unknown there are several risk factors which have been identified which increase an individual's likelihood of developing DLBCL. As previously mentioned there is an associated risk of developing DLBCL with increasing age. There is a higher incidence in males and Caucasian populations. In a pooled analysis of 4667 cases and 22639 controls from 19 studies Cerhan *et al.* demonstrated that positive hepatitis B virus serology, family history of NHL and high young adult body mass index (BMI) increase the risk of a patient developing DLBCL [58]. Certain occupations may place individuals at a higher risk of developing DLBCL. In the female subcategory these included farm workers and hairdressers. In the male subcategory factory workers were deemed to be at higher risk of developing DLBCL. There was a negative association, with the odds ratios being <1 for increased levels of recreational sun exposure, presence of an atopic disorder or being in a higher socio-economic group.

1.3.5 Clinical prognostic markers

There are a number of different prognostic scoring systems proposed to stratify DLBCL patients. The first of which, the international prognostic index (IPI) was developed by Ship *et al.* in 1993. The score was based on retrospective analysis of 2031 patients treated with cyclophosphamide, hydroxydaunorubicin hydrochloride (doxorubicin hydrochloride), vincristine (Oncovin) and prednisone (CHOP) chemotherapy [59]. Patients received a point for each of the following if present: age over 60 years, stage II or III disease, raised serum lactate dehydrogenase LDH, Eastern Cooperative Oncology Group (ECOG) performance status of greater than one and involvement of two or more extra-nodal sites. Patients were split into four prognostic groups depending on their accumulated scores. The IPI was updated in 2007 by Sehn *et al.* to account for the use of rituximab with

CHOP chemotherapy (RCHOP) [60]. This retrospective analysis of 365 patients, reported that the original IPI was no longer able to stratify patients into four distinct outcome groups, but instead grouped them into two outcome groups. This revised IPI (R-IPI) used the same features of the original IPI but grouped them into 3 different outcome groups. A larger study by Ziepart *et al.*, using data from 1062 patients in three different trials, found that the IPI was still valid but the use of rituximab did improve event free survival and reduced the discrepancy between survival curves [61]. Zhou *et al.* developed a further scoring system based on data from the national comprehensive cancer network database (NCCN) [62]. The NCCN-IPI is again based on age, LDH levels, stage, extra-nodal disease and performance status but splits age and LDH levels into different scoring categories when compared to the previous systems. The NCCN-IPI splits patients into four prognostic groups. The NCCN-IPI outperformed IPI for stratification of patients into high and low risk groups when predicting 5-year OS. A more recent study by Rupert *et al.* compared IPI, R-IPI and NCCN-IPI in a cohort of 2124 DLBCL patients treated with RCHOP and reported that NCCN-IPI was superior for predicting OS compared to IPI and R-IPI [63]. However, the concordance (C)-index for NCCN-IPI was only 0.63 suggesting that prognostic scoring systems could be improved to more accurately stratify patients.

1.3.6 Imaging of diffuse large B-cell lymphoma

Like cHL, the gold standard imaging technique for staging and response assessment in DLBCL is FDG PET/CT. However, unlike in cHL, treatment is not stratified around an interim PET/CT in routine clinical practice. Studies have assessed the utility of interim PET/CT, with promising results which suggest an interim PET/CT following 2 cycles being suggested for de-escalation trials and an interim PET/CT following 4 cycles being suggested for randomised control trials assessing new treatments [64]. Although, unlike cHL there is a well-recognised pitfall in DLBCL with a high proportion of false positive results at interim FDG PET/CT which could lead to inappropriate treatment stratification [65]. The Lugano classification is used for the staging of disease and DS for the disease monitoring/response assessment.

1.3.7 Imaging markers

Similar to cHL SUV derived metrics and radiomic predictive markers have been explored. A detailed exploration of the literature is provided in Chapter 2.

1.3.8 Treatment of diffuse large B-cell lymphoma

The mainstay of treatment in DLBCL is with immunochemotherapy using RCHOP [66]. Radiotherapy can be added if there is residual bulky disease [67]. Prophylactic intrathecal methotrexate or intravenous treatment with chemotherapy which can cross the blood brain barrier (methotrexate (high dose), cytarabine or ifosphamide) is given to patients who are at high risk of central nervous system (CNS) involvement [68]. Patients are deemed high risk if they have involvement of the kidneys, adrenal glands, breasts of testes, or if they have two or more risk factors listed in the CNS-International Prognostic Index (IPI). The risk factors include: Eastern Cooperative Oncology Group (ECOG) performance score of two or more, stage III/IV disease, two or more extra-nodal sites, raised lactate dehydrogenase (LDH) and age over 60 years [68].

1.3.9 Genetic Markers

The cell of origin (COO) is gaining traction as a prognostic marker in DLBCL. Three distinct subtypes have been identified: activated-B-cell (ABC), germinal-centre-cell (GCB) and type 3, which cannot be classified as either ABC or GCB. Patients with ABC have worse 5-year PFS outcome compared to patients with tGCB subtype; 31-48% compared to 76%-78% respectively when treated with RCHOP [69, 70]. The Myelocytomatosis-cellular (MYC) proto-oncogene and B-cell lymphoma 2 (BCL2) gene expression are also prognostic markers. There are mixed reports of whether these gene expressions are prognostic markers in their own right or are related to the COO. Liu *et al.* reported that ABC subtype was associated with expression of both these genes, double expression lymphoma (DEL) [71]. Conversely, GCB is associated with translocation of the MYC and BCL2 genes, double hit lymphoma (DHL). However, the prognostic ability of both DEL and COO may not be appropriate for assessment of early stage (I//II) disease as Barraclough *et al.* were unable to demonstrate significance in OS or PFS in this cohort of patients.

1.4 Data acquisition, optimisation and analysis

Radiological imaging data within health institutions is stored using a picture archiving and communication system (PACS). PACS architecture consists of modality acquisition equipment e.g. a CT scanner, a PACS gateway which acts as quality assurance ensuring all necessary data is associated with the image, an archiving storage network and reporting workstations [72]. Each of these components within the network are known as an application entity (AE) and have their own AE title, internet protocol (IP) address and port number which act as the identifier and location for imaging data to be pushed or pulled between different AEs [73]. Not all AEs are directly connected e.g. reporting workstations are not usually directly connected to the modality acquisition equipment. An AE sending a request, is regarded as the service class user (SCU) and the AE dealing with the request is termed the service class provider (SCP) [74]. An AE can be both an SCP and SCU depending on the direction of the traffic requests.

Medical imaging data is stored in a digital imaging and communications in medicine (DICOM) format with the DICOM dataset being made up of information in the form of a key-value associative array [75]. The key is also known as a DICOM tag. DICOM tags are standardised and take the form of a 16-bit hexadecimal number. A hexadecimal number is a number which is to the base 16 rather than base 10 (decimal) with the 16 digits being represented as 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, and F . The decimal number 15 is therefore represented as F in hexadecimal. The number 16 in decimal would be 10 in hexadecimal and the number 105 in decimal would be 69 in hexadecimal. An example of a DICOM tag would be (0x0008, 0x0020) which is the tag associated with the study date. Each tag has a common nickname or description e.g. for tag (0x0008, 0x0020) it is "StudyDate" which can also be used to reference a tag [76]. The value associated with the data tag can take the form of a string, a date, integer, floating point or bytes, with tags storing information about patient details, referral information, scan acquisition details and pixel data all stored under DICOM tags. DICOM tags permit correct display, storage, and transfer of imaging data.

Pixel data is associated with the (0x7fe0, 0x0010) data tag, and is commonly stored in an uncompressed format [76]. The transfer syntax can be used to identify if data is compressed or not, and the nature of compression. Examples of transfer syntaxes for uncompressed pixel data include implicit VR little-endian, explicit VR little-endian and explicit VR big-endian [75]. To allow images to be transferred between AEs, transfer syntaxes need to be supported by both applications. If an AE does not recognise the transfer syntax, the request is rejected, and the dataset is not transferred.

To allow development of machine learning models there needs to be a streamlined process to retrieve images from PACS and convert DICOM images into image files which are suitable for the application of radiomic analysis. The next sections will describe methods undertaken to address these steps as part of the projects within thesis, the flow diagram for which is depicted in Figure 1.4.

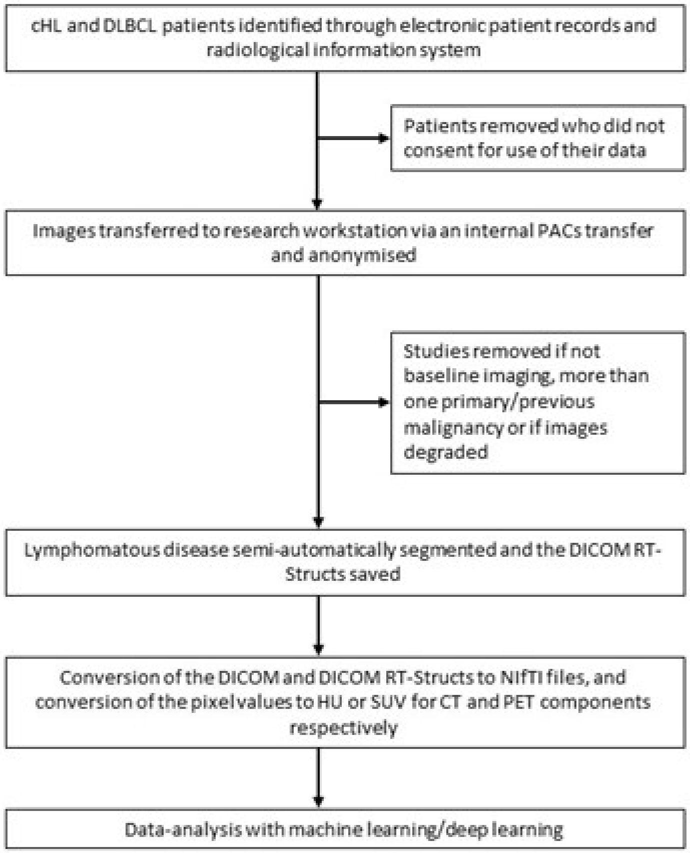

cHL and DLBCL patients identified through electronic patient records and radiological information system
Patients removed who did not consent for use of their data
Images transferred to research workstation via an internal PACs transfer and anonymised
Studies removed if not baseline imaging, more than one primary/previous malignancy or if images degraded
Lymphomatous disease semi-automatically segmented and the DICOM RT-Structs saved
Conversion of the DICOM and DICOM RT-Structs to NIfTI files, and conversion of the pixel values to HU or SUV for CT and PET components respectively
Data-analysis with machine learning/deep learning

Figure 1.4 Data pre-processing and analysis steps undertaken within the project.

1.5 Image extraction

PET/CT studies for DLBCL and cHL patients were identified by the data analyst AR using a structured query language (SQL) server search of institutional electronic medical records and radiology information system servers using International Classification of Diseases Version 10 (ICD-10) classifications and PET/CT imaging codes respectively. All patients included had provided informed written consent prospectively at the time of imaging for use of their FDG-PET/CT imaging data in research and service development projects. Following discussion with the Research and Innovation Department at LTHT it was agreed that this represented a service improvement project and formal ethics committee approval was not required. The study was approved by the University of Leeds ethics committee (Appendix A).

Details of potentially eligible patients were initially used to manually push PET/CT images from an institutional PACS (Agfa Enterprise Imaging (EI)) workstation to a dedicated research workstation with a DICOM server (DBx, Mirada Medical). The transfer process was augmented by an automated batch query/retrieve system written in Python using the libraries pynetdicom and pydicom. The AE title, IP address and port number of the Python script were recorded in the Agfa configuration file and reciprocal details of the PACS archive were inputted into the Python script to allow negotiation between the two. It should be noted when implementing a c-move request to pull images, that different PACS systems allow requests in different formats and a script written for one may not be universally applicable. C-move is the name of the operation which allows the transfer of a DICOM image from one PACs system to another and is used in combination with the operation to store the image (c-store). For example, if a c-move request is initiated from Python to Orthanc (a free PACS server), the request can be made using "Patient Root Query Retrieve Information Model Move" Service-Object Pair (SOP) class with the query level of "STUDY" using only the accession number. However, if the same method is used to try and pull images from Agfa PACS the code will fail with an error indicating the SOP class is wrong. Agfa PACS does not allow a "STUDY" level query using the "Patient Root Query Retrieve Information Model Move" SOP class and does not allow the accession number as the study identifier in a move query. To overcome this, the SOP class needs to be changed to "Study Root Query Retrieve Information Model Move" and "c_find" request needs to be performed to get the instance SOP unique identifier (UID) from the accession number which can then be used to initiate the move request.

All images were downloaded to a dedicated password-protected workstation within the LTHT firewall where the images were anonymised. No patient details left the institutional, and no anonymised imaging data was used without the approval of Information Governance at LTHT.

1.6 Anonymisation

DICOM files for each patient were initially pseudo-anonymised using the anonymisation tool within the Mirada Medical software at the point of segmentation. The anonymisation tool removes all private tags and changes identifiable patient details to those of an agreed new tag for example "LYM-E-001". However, to facilitate more efficient large-scale batch anonymisation a bespoke Python script was created using the Python library pydicom (https://gitfront.io/r/user-8522243/JQMsK3nk56iq/PhD/), a graphical user interface version is also available to download. The script iterates through all DICOM files within a folder, accesses the DICOM tags and deletes or changes tags which are unknown, private or known to have patient sensitive information. Patient age, weight, height and sex were retained to allow for the calculation of SUVs.

The anonymisation script uses a hash function with a salt-key to update sensitive tags which were retained. Although it is a one-way process, the salt-key is not shared to eliminate any risk of an individual being able to reverse engineer the process to decrypt the values. Private tags were removed as they do not usually have a description, and our dataset was found to contain patient NHS numbers in some cases. Any date related to the study, including injection times in the case of PET, were pseudo-anonymised by changing the first study of a patient's set to the date 08/11/2015 with all subsequent exams for that patient given a date which preserved the temporal relationship of the original studies. Some commercial PACS have a cut-off date for which it will allow studies to be uploaded. Also, some will not load images unless they have a value for the patient's date of birth and have a patient ID assigned.

UIDs apart from the transfer syntax UID and the SOP Class UID were also converted using a hash function as they are, as the name suggests, universally unique [77]. Whereas study accession numbers and other patient IDs may be unique within an institution they could be present somewhere else in the world, UIDs are generated in such a way that they are unique to the scan, series, study, scanner, and hospital and therefore also needed to be anonymised [77]. The syntax UID and SOP Class UID are required to define how to encode the dataset for transfer between AEs, and although they are UIDs they are not universally unique to a single study or patient. The modality tag and SOP class UID are used to identify and delete patient reports or reformats which often contain patient details burned into the pixel values. It should be noted that some institutions have a modality referred to as "PR", which stands for presentation state, which at our institution have the radiographer's details stored as the series description.

The overlay data tag (0x60xx, 0x3000) was removed from the DICOM dataset, if present, as the information within this tag can contain patient information, ward number and the date and time of the study [76]. If the ward number is displayed within this data, it can often be accompanied by a "J" or "L" prefix which when the ward and prefix are typed into an internet search engine will often allow accurate identification of the hospital where the study was performed. This is more of a concern when dealing with computed radiography. Some scanners or modalities save patient details into pixel data, the images can be discarded, or the software can overwrite the pixel data.

All studies were reviewed manually during segmentation to check that all patient identifiable information had been removed.

1.7 Segmentation

The three main semi-automated segmentation techniques reported in the literature are either based on a fixed SUV threshold, an adaptive thresholding method based on the lesion(s) being contoured or based on background physiological uptake. There is ongoing work to establish a consensus on the optimal segmentation method for deriving MTV [78], and how the segmentation affects the performance of the derived model [79].The most commonly reported segmentation technique is based on the 41% or 40% SUVmax of a lesion, this was demonstrated to have the best correlation with a defined source and then validated on a cohort of 10 lung carcinoma patients [80] and has been validated in HL and DLBCL studies [81][82]. However, lymphomatous masses can be heterogeneous and this can potentially lead to volumes being underestimated due to a small area of high SUV disproportionately influencing the threshold. Also, given the need to identify the SUVmax of each lesion to calculate the threshold needed, there is a potential for this method to be more time consuming when compared to other methods. A fixed threshold technique has the benefit of being easily applied without needed to take into consideration the lesion or background SUV uptake. A fixed threshold of 4.0SUV was demonstrated by Burggraaf *et al.* to have a higher interobserver reliability than the other thresholding methods studied (41%SUVmax, 50%SUVpeak and SUV 2.5) [83]. A threshold based on background physiological uptake is not as commonly reported in the literature in terms of lymphoma segmentation, however, its use has been reported in other cancer types [84, 85]. The premise being that variations in SUV in study acquisition, dose or body composition are normalised, and that the SUVmean of the liver has been demonstrated to be one of the most reproducible metrics for liver uptake [86]. For the studies as part of this thesis, segmentation was performed using specialised multimodality imaging software (RTx v1.8.2, Mirada Medical, Oxford, UK). Two different semi-automated segmentation methods using two different threshold methods were used: 1.5 times mean liver SUV and a fixed threshold of 4.0SUV (Figure 1.5). The mean liver SUV was calculated by defining a 110cm^3 ROI within the right lobe of the liver and recording the SUVmean. Physiological or non-lymphomatous disease was manually excluded from contoured volumes.

1.8 Neuroimaging Informatics Technology Initiative conversion

Neuroimaging Informatics Technology Initiative (NIfTI) is a file format often used in radiomic analysis [87]. The main reason for this is that NIfTI pixel data is stored as a 3D data set, whereas DICOM images are often stored as 2D slices [88]. This makes it simpler to input and navigate when extracting radiomic features from imaging data stored in this

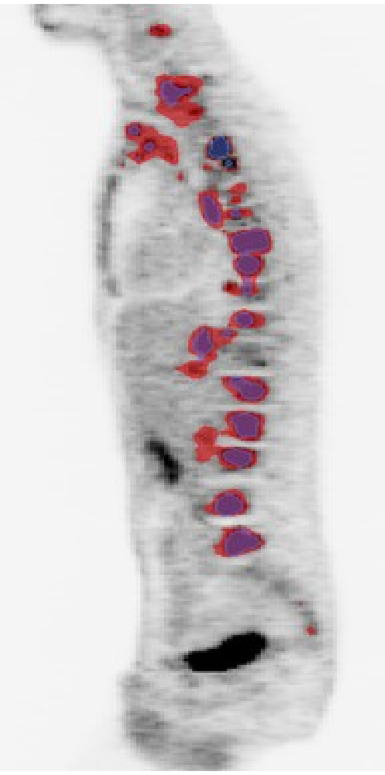


Figure 1.5 Sagittal slice of a PET study of a patient with DLBCL demonstrating the 1.5 times mean liver SUV segmentation (red) and the 4.0SUV fixed threshold segmentation (blue/purple).

file format. NIfTI files contain metadata which allows the image to load in the correct spatial plane based on the header details and affine matrix [88]. NiFTI files do not store identifiable information within the metadata and therefore can be considered a method for anonymisation. However, if the data was to be loaded into a PACS or radiotherapy planning system such as the one used in this study, the files would need to be converted back to DICOM images [89]. Also, if pixel data is burnt on this would be transferred to the NIfTI pixel data.

The steps used in the conversion of DICOM images to NIfTI files will be discussed in the next sections.

1.8.1 Pixel conversion

Firstly, pixel data within DICOM files was converted into the desired unit for the modality, i.e. Hounsfield units (HU) and SUV for CT and PET respectively. This is necessary to allow extraction of HU and SUV and any metrics based on these to be used as features. This is arguably more important in PET, where SUV is time corrected and pixels are adjusted accordingly when displayed [45]. If no correction is applied different bed positions may have inherently different background uptake which would affect any model derived from this data. Therefore, false patterns or associations may be derived. The method for conversion of SUVs and HUs are slightly different.

1.8.2 Standardised uptake value conversion

Values from DICOM tags relating to patient weight, acquisition time, injection time of radiopharmaceutical, dose and half-life of radiopharmaceutical, rescale slope and intercept were used to convert pixel values to body weight derived SUV (SUVbw) [45]. To allow for conversion of pixel values to body surface area (SUVbsa) patient height needs to be recorded, and for conversion into lean body mass (SUVlbm) the height and gender of the patient are needed. The DICOM tags for the information are detailed in Table 1.4. From a practical point of view, it is important to know which manufacturer the imaging study is performed on as different manufacturers require the use of different times as their start point. Also, the radiopharmaceutical tags are subtags to the tag (0054,0016) "RadiopharmaceuticalInformationSequence" and therefore the information needs to be extracted by first accessing the parent tag and then looping through the other tags.

DICOM Tag	DICOM Description
0x7FE0, 0x0010	Pixel Data
0x0028, 0x1053	Rescale Slope
0x0028, 0x1052	Rescale Intercept
0x0008, 0x0032	Acquisition Time
0x0018, 0x1071	Radiopharmaceutical Volume
0x0018, 0x1074	Radionuclide Total Dose
0x0018, 0x1075	Radionuclide Half Life
0x0018, 0x1072	Patient Weight
0x0010, 0x0040	Patient Sex
0x0010, 0x1020	Patient Size

Table 1.4 List of DICOM values needed for conversion of DICOM pixel values into different SUV measurements.

Pixel values were first converted into Bq/ml by using the following equation:

$$ActivityConcentration\left(\frac{Bq}{ml}\right) = PixelValue * Slope + Intercept$$

The equation will only work if the PET/CT Units tag (0054,1001) is Bq/ml (BQML). Also, it should be noted that when applying this equation, the intercept in PET/CT is generally set to a value of 0. Lastly, this equation may already be applied to the pixel data when the DICOM dataset is opened by the library SimpleITK (a simplified toolkit which supports analysis of 15 different types of images included DICOM) whereas the equation needs to be applied manually when using the library pydicom.

Next the dose decay correction factor was calculated. This factor corrects the SUV for decay which occurs from tracer injection to the time of the study and was calculated using the following equation:

$$CorrectionFactor = 2\left(-\left(\frac{ScanTime\,(s) - MeasuredTime\,(s)}{HalfLife\,(s)}\right)\right)$$

Again, there are important points to note when performing this calculation. The first thing to consider is what is documented in the tag (0054, 1101) DecayCorrection as this determines which time is used to correct the study to. If it reads "START" it will refer to the acquisition time or series time depending on the manufacturer. If is reads "ADMIN" it is the time of the pharmaceutical administration. As well as there being variation in the suggested method of calculation between scanner manufacturers, DICOM viewers can also have variation in their calculation. In the dataset used as part of this study Mirada RTx was used to contour cases and the SUV thresholds used as part of the semi-automated process were derived from this software. Therefore, the SUV calculation was performed in a similar manner to Mirada RTx where the start time was taken as the earliest acquisition time. The values given in the time data tags are written in a 5- or 6-digit number e.g. 151619.000000. This number does not represent seconds but represents the time in the form hh:mm:ss and therefore needs to transformed into seconds to allow for correct conversion.

The final calculations were applied to the pixel values to determine the SUV. The equations for SUVbw, is detailed below:

$$SUVbw = \frac{AcitivityConcentration\left(\frac{Bq}{ml}\right) * BodyWeight\,(g)}{TotalDose * CorrectionFactor}$$

To calculate SUVbsa and SUVlbm, body weight is replaced with lean body mass or body surface area which are calculated with the equations below [90]:

$$BodySurfaceArea = BodyWeight^{0.425} * Height^{0.725} * 0.007184$$

Male

$$LeanBodyMass = 1.10 * BodyWeight - 120\left(\frac{BodyWeight}{Height}\right)^{2}$$

Female

$$LeanBodyMass = 1.07 * BodyWeight - 148 \left(\frac{BodyWeight}{Height}\right)^2$$

1.8.3 Hounsfield unit conversion

The conversion of CT pixel values into Hounsfield units (HU) is less complex in its implementation and only requires the pixel value, rescale slope and intercept. The equation is detailed below:

$$HounsfieldUnits = PixelValue * Slope + Intercept$$

1.8.4 Space and orientation

1.8.4.1 Alignment of CT and PET images

Voxel size varies between CT and PET imaging due to respective scanner resolution, and the slice thickness and number of pixels in columns and rows between slices is inherently different. When extracting the data array from NIfTI images an important consideration is that the NumPy array (grid of values in Python library) does not take into consideration pixel size or orientation with a resultant mismatch between CT and PET images when trying to display them. This leads to a mismatch in the size of the images when displayed from the NumPy array which is corrected when the affine matrix is applied to the array. The affine matrix provides the information to convert the NumPy array into one that represents the relationship of the pixels in physical space as the pixel values do not have a defined pixel size and they are not necessarily orientated in direction they needed to be displayed when being reviewed [88].

Affine matrices allow for linear transformations, and for a 3D image any translation, rotation or scaling can be represented by a 4x4 matrix [88]. The first row is concerned with the x axis, the second row is concerned with the y axis and the third row represents the z axis. The final array allows for the combination of different transformations within the same matrix. The value in first column in the x axis, the second column in the y axis and the third column in the z axis are used for scaling along the given axis.

1.8.4.2 Mask creation

The creation of a single NIfTI file from a series of DICOM images was performed using simple insight toolkit (ITK). Metadata was extracted from DICOM images for the PET or CT datasets using the “ImageSeriesReader” and this was combined with the converted pixel data from the previous section. The NIfTI file was written using the “ImageFileWriter” and setting the “SetImageIO” to NiftiImageIO.

Simple ITK was also used to convert the segmentations from the DICOM radiotherapy structure (RT Struct) to NIfTI file. Unlike the DICOM files for modalities, all segmentations are stored in a single DICOM RT Struct file [91]. The data arrays, however, are not stored as pixel values but stored as co-ordinates which relate to the associated DICOM image on which the segmentation was drawn. To create a NIfTI segmentation the co-ordinates were converted into a mask using the NumPy.polygon function using the pixel spacing from the original DICOM images as the basis for the mask. The pixel values were then passed into the “ImageFileWriter” module of simple ITK in the same manner as the conversion of the DICOM images.

In this study only PET data was contoured, and the mask then transferred to the CT data. Due to the volume of PET uptake being generally larger than the underlying soft tissue. The masks were automatically adjusted by removing any pixel values not within the -10 to 100 HU range, this cut-off was chosen to allow for some areas of necrosis and higher density material to be included whilst minimising fat and bone inclusion. The mask was then reformed and any holes filled in using the binary_closing function in scipy.ndimage.

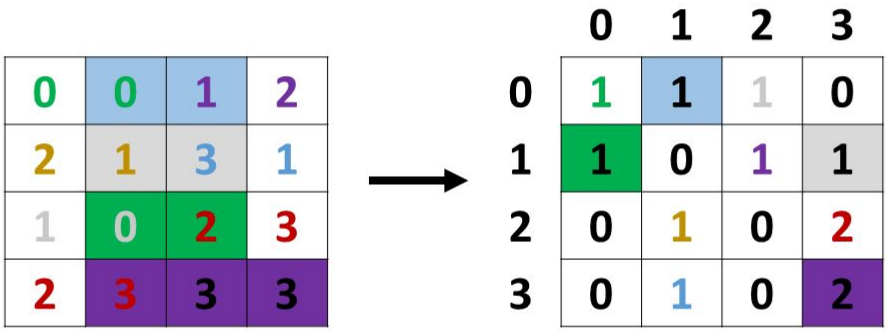
1.9 Radiomic Feature Extraction

Radiomics is the process of transforming images into mineable data which can be utilised in the creation of predictive modelling. It offers a numerical value for features perceived visually, however; it also has the ability to uncover features which are not perceivable to the human eye. Radiomic features can be studied in isolation or combined with clinical, genomic and other features as part of the broader field of integrated diagnostics [92]. PyRadiomics was used to extract the radiomic features for this study. The pipeline involved in feature extraction of radiomic features from the images and masks within the study is detailed below.

1.9.1 Feature Calculation

There is potential for different definitions of radiomic features, which is why the image biomarker standardisation initiative (IBSI) has set definitions to try and standardise practice. PyRadiomics has some variation in its methodology when compared to IBSI [93]. The kurtosis values extracted are always +3 of those defined by IBSI. When binning using fixed bin width with re-segmentation PyRadiomics always uses bin edges equally spaced from 0 whereas IBSI uses bin edges equally spaced from the minimum re-segmentation range. PyRadiomics aligns the resampling grid to the corner of the origin voxel compared to the centre of the image. PyRadiomics does not round grey values, whereas IBSI does.

First order parameters are histogram statistical features which do not examine spatial relationship and therefore cannot be termed textural analysis. Second order parameters allow for the measurement of spatial arrangement. The parameters are derived from matrices such as grey-level co-occurrence matrix (GLCM), the grey-level run-length matrix (GLRLM), grey-level size zone matrix (GLSZM), grey-level distance zone matrix (GLDZM), neighbourhood grey-tone difference matrix (NGTDM) or neighbouring grey-level-dependence matrix (NGLDM) [94]. The GLCM matrix is created by plotting the number of different combinations of neighbouring voxel values in a defined direction (Figure 1.6) [94]. The default distance weighting and distance was used in the analysis.

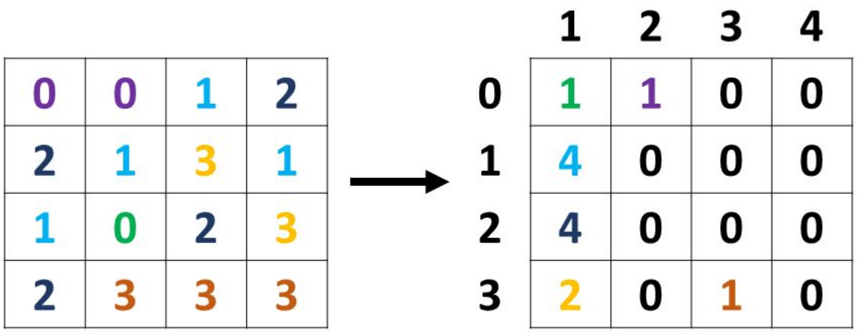


0	0	1	2
2	1	3	1
1	0	2	3
2	3	3	3

	0	1	2	3
0	1	1	1	0
1	1	0	1	1
2	0	1	0	2
3	0	1	0	2

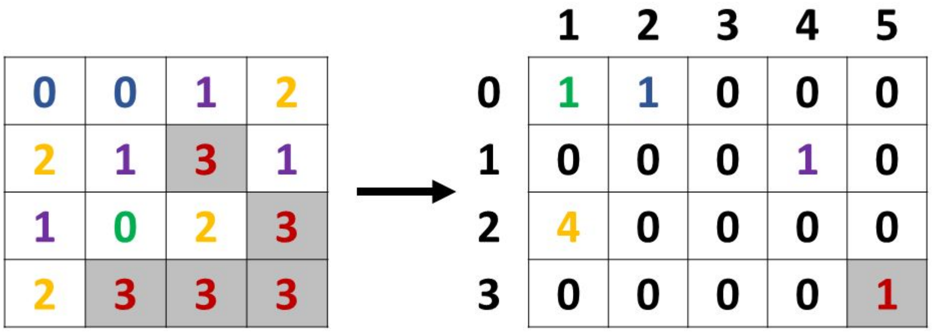
Figure 1.6 A grey-level co-occurrence matrix in a 90 degree direction is formed by plotting the number of times neighbouring pixel values occur. For example pixel values of 2 and 3 occur twice. The colours of the the pixel values and the shading of the cells help identify how these make up the matrix.

The GLRLM is formed by plotting the number of continuous voxels with the same value in a defined direction. The matrix can extend from a single run length up to a maximum number of voxels in the within the ROI in the defined direction (Figure 1.7) [94]. The GLSZM is created by plotting the number of voxels that have a same value lying adjacent to each other (Figure 1.8) [94]. The GLDZM matrix is formed by plotting the distance to the closet border of the ROI for a region of voxels of the same values (Figure 1.9). The NGTDM is the average intensity of the neighbouring voxels from a central voxel (Figure 1.10) [94].



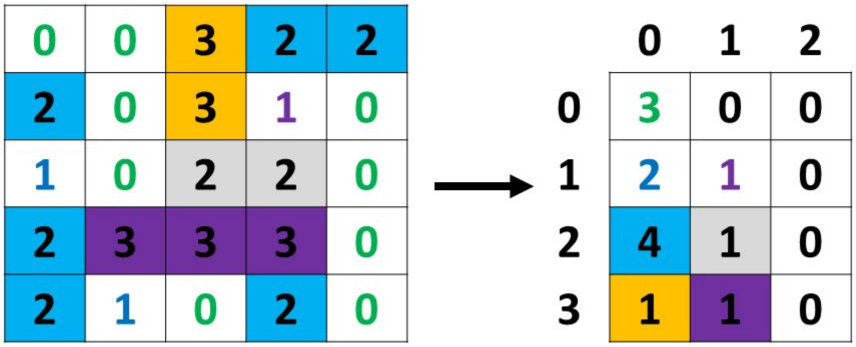
$$\begin{array}{|c|c|c|c|}\hline 0 & 0 & 1 & 2 \\ \hline 2 & 1 & 3 & 1 \\ \hline 1 & 0 & 2 & 3 \\ \hline 2 & 3 & 3 & 3 \\ \hline \end{array} \longrightarrow \begin{array}{c|c|c|c|c|} & 1 & 2 & 3 & 4 \\ \hline 0 & 1 & 1 & 0 & 0 \\ \hline 1 & 4 & 0 & 0 & 0 \\ \hline 2 & 4 & 0 & 0 & 0 \\ \hline 3 & 2 & 0 & 1 & 0 \\ \hline \end{array}$$

Figure 1.7 A grey-level run-length matrix created in the 90-degree direction. The matrix on the left represents the grouped pixel values, the matrix on the right is the GLRLM. The GLRLM represents the number of time a particular value occurs in a row. For example there is 1 time that the pixel value 3 occurs three times in a row.



$$\begin{array}{|c|c|c|c|}\hline 0 & 0 & 1 & 2 \\ \hline 2 & 1 & 3 & 1 \\ \hline 1 & 0 & 2 & 3 \\ \hline 2 & 3 & 3 & 3 \\ \hline \end{array} \longrightarrow \begin{array}{c|c|c|c|c|c|} & 1 & 2 & 3 & 4 & 5 \\ \hline 0 & 1 & 1 & 0 & 0 & 0 \\ \hline 1 & 0 & 0 & 0 & 1 & 0 \\ \hline 2 & 4 & 0 & 0 & 0 & 0 \\ \hline 3 & 0 & 0 & 0 & 0 & 1 \\ \hline \end{array}$$

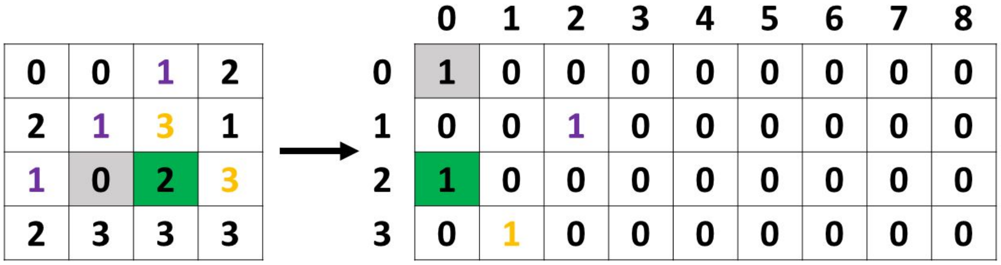
Figure 1.8 The matrix on the left represents the grouped pixel values, the matrix on the right is the grey-level size zone matrix (GLSZM). The GLSZM represents the number of time a particular value is neighboured in any direction by a pixel with the same value. For example there is 1 time that the pixel value 3 occurs in a chain of five pixels with the same value.



0	0	3	2	2
2	0	3	1	0
1	0	2	2	0
2	3	3	3	0
2	1	0	2	0

	0	1	2
0	3	0	0
1	2	1	0
2	4	1	0
3	1	1	0

Figure 1.9 The matrix on the left represents the grouped pixel values, the matrix on the right grey-level distance zone matrix (GLDZM). The GLDZM plots the distance of groups of the same pixel value to the closet border. For example the value 2 occurs next to the border in 4 distinct places and 1 pixel away from the border in 1 instance.



0	0	1	2
2	1	3	1
1	0	2	3
2	3	3	3

	0	1	2	3	4	5	6	7	8
0	1	0	0	0	0	0	0	0	0
1	0	0	1	0	0	0	0	0	0
2	1	0	0	0	0	0	0	0	0
3	0	1	0	0	0	0	0	0	0

Figure 1.10 The left matrix represents the pixel intensities of a region of interest. The matrix on the right represents the NGLDM. The NGLDM measures the number of neighbouring eight pixels with the same intensity of a central pixel.

Further higher order features are created by applying different filters to the ROI to extract repetitive or non-repetitive patterns [94]. These include Minkowski functionals which assess patterns of voxels above a threshold intensity, three-dimensional discrete wavelet transform (Figure 1.11) which applies high or low pass features highlighting the details or approximations respectively. in each dimension, and Laplacian of Gaussian filters (Figure 1.12). The Laplacian filter detects edges in images due to rapid changes of the pixel values, however, it is sensitive to noise and therefore the image is smoothed using a Gaussian filter before its application. Further filters include transforming the image via logarithm, exponentially or via the square root of the values (Figure 1.13). As part of the thesis the default settings for the filters within PyRadiomics were kept.

Once the features have been extracted from the ROI, a harmonisation step is required if multiple scanners or scanning protocols are used. Combating batch effects when combining batches (ComBat) harmonisation was originally designed to account for variations in batch effects in gene microarray expression [95]. Batch effect is analogous with imaging protocol effect in radiomics. ComBat harmonisation has since been successfully applied to radiomics to adjust for scanner variation in PET/CT [96]. It uses empirical Bayes to estimate a normalised value for a feature for a specific ROI and scanner protocol. ComBat then determines and applies a transformation for each feature based on the effect of the scanner protocol on the features [97]. The equation is detailed below where γ_{ij} = value of each feature for the ROI j and scanner i, a = average value for the feature, X_{ij} = design matrix for the covariates of interest, β= the vector of regression coefficients for each covariate, γ_{i} = additive effect of scanner, δ_{i} = the multiplicative scanner effect, ϵ_{ij} = error term.

$$y_{ij} = a + X_{ij}\beta + \gamma_i + \delta_i\epsilon_{ij}$$

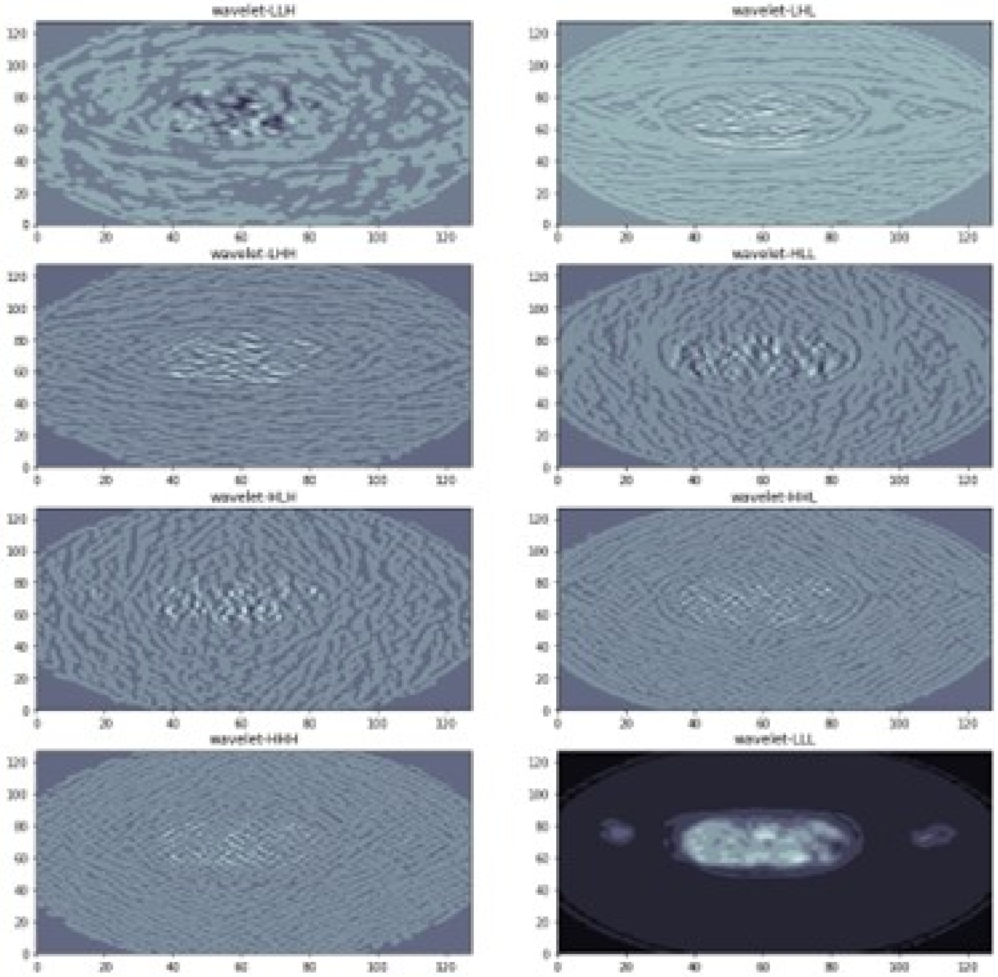


Figure 1.11 Wavelet decompositions for an axial slice of a PET image.

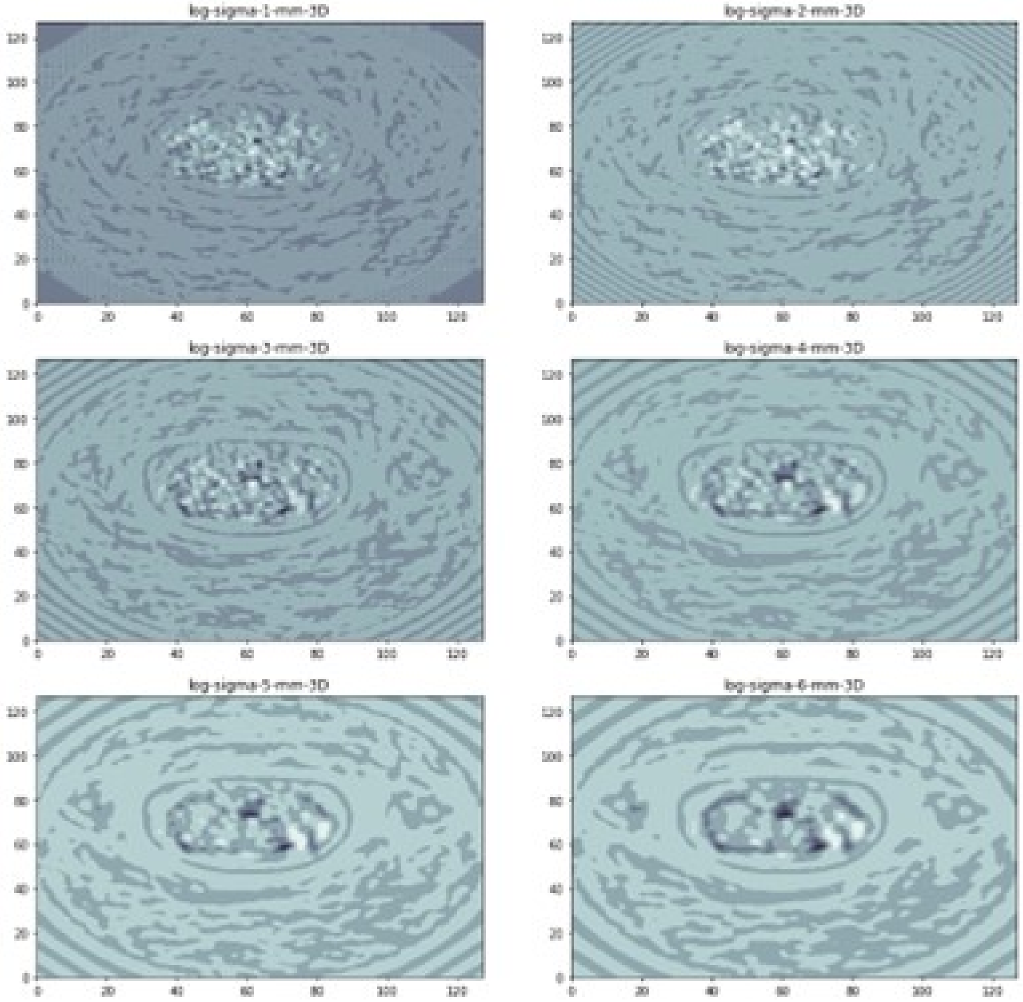


Figure 1.12 Select axial slice of a PET image with different Laplacian of Gaussian sized filters applied.

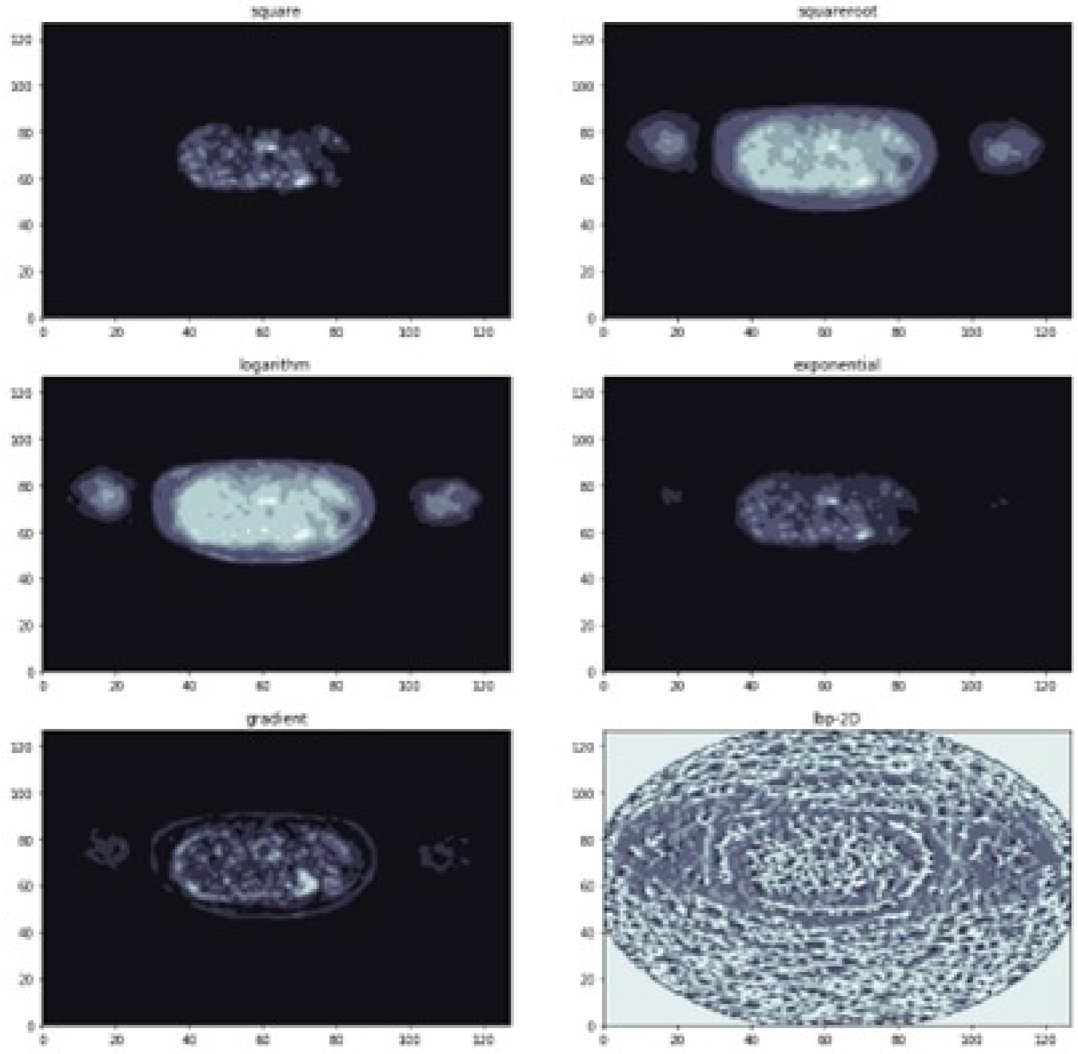


Figure 1.13 Select axial sliced of a PET image with different filters applied.

1.10 Machine learning

There are several machine learning (ML) techniques which can be utilised to create classification models [98]. Given the need to censor patients if they have come to the end of the follow up period in predictive models and implanting this into the ML algorithms, the project will focus on a binary outcome and therefore techniques for classification can be implemented. These include logistic regression, random forest (RF), K-nearest neighbour and support vector machines (SVM)[98]. Each technique has its own hyperparameters which can be tuned to try and improve performance. The algorithms will need to be adjusted to account for the fact that the dataset will inherently be unbalanced, as there are far more patients who do not relapse/progress when compared to those that do. When using unbalanced datasets, the choice of scoring method of model performance needs to be considered as some metrics may overestimate the model performance e.g. accuracy [99]. The use of the receiver operator characteristics (ROC) curve area under the curve (AUC) can offer a more effective assessment of performance in unbalanced cases. The ROC curve is created by plotting the sensitivity (true positive rate) against the 1-specificity (false positive rate) from confusion matrices created at different probability thresholds [100].

Models are developed using a training and validation dataset, the validation dataset allows for the performance of the machine learning model to be scored whilst it is being tuned [99]. The testing data set is an unseen data set which is used to evaluate the final model and is not involved in the tuning of the data. In cases where there is small dataset cross validation can be utilised to act as the training and validation data sets. Cross validation is the process of defining and splitting the data into groups and using the data from all the groups bar one (the validation set) to train a model and then the model is test model on the validation dataset [99]. One of the groups from the training data then becomes the validation dataset and the model is trained is on the remaining groups and again tested (Figure 1.14). This process is performed until all groups have acted as the validation dataset. This can then be repeated by splitting the whole data again into different groups, this is known as repeated cross validation. The groups can be defined randomly and can be stratified around important features or the outcome to maintain the ratio of the outcome or feature between the groups. Following the process of cross validation, a range of predictive score are produced which can be used to determine the best performing model, feature selection and hyperparameters to be used.

The work performed as part of this thesis utilised stratified cross validation around the 2-year EFS (2-EFS), age, sex, ethnicity, disease stage, having radiotherapy, having ABVD-based chemotherapy and being treated as advanced disease for the cHL study. For the DLBCL patients the cross validation was split around 2-EFS, disease stage, age and sex, as the treatment was standard across all patients. The ratio and split of the subsets were based around the number of patients and event rates to allow for the stratified cross validation.

1.10.1 Feature Selection

Having a large proportion of features compared to the number of events in a model can lead to the model becoming dependent on the data it is trained on causing overfitting, "the curse of dimensionality", and therefore poor generalizability to any dataset outside of the training set. There are a number of methods which can be utilized when reducing features.

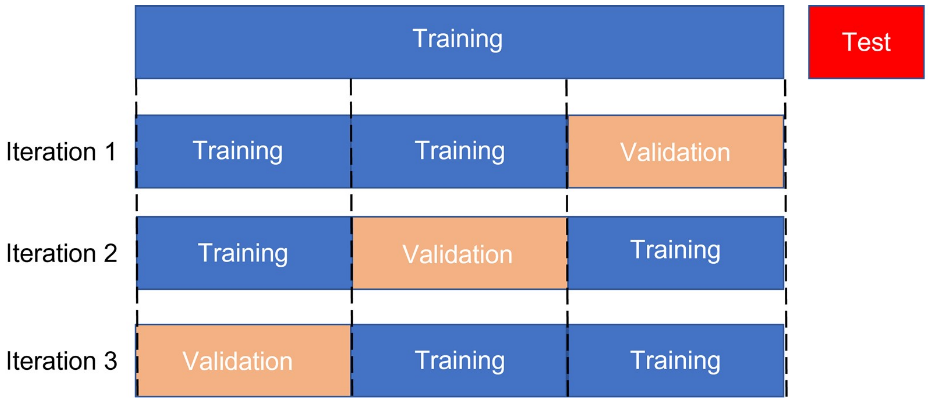


Figure 1.14 Depiction of cross validation. The training dataset is split into 3 groups each taking turns to be the validation dataset with the other groups being used for training. The test dataset it not touched during this process.

Firstly, features which are not reproducible can be removed as they are not going to limit the generalizability of the model to future datasets. The interclass correlation coefficient can be used as a marker of reproducibility, and a threshold for removing features, on repeat segmentation and extraction. The Pearson (linear) or Spearman (monotopic) coefficient can be used to remove correlated features to avoid the issue of multicollinearity. Multicollinearity within the model can reduce performance due to unreliable estimates of the derived coefficients.

A forward wrapper method tries every feature in a model to determine the best feature, then tries that feature combined with each of the different remaining features to see which provides the best score. It continues the process for the set number of parameters selected. A backward wrapper performs the process in reverse, starting with all features and removing each feature in turn to find out which feature when removed provides the best score. The recursive feature elimination method is broadly based on backward wrapper feature selection, however, features are ranked by their coefficients or by the feature importance score. The univariate feature selection method uses the F-test to determine the best performing features.

1.10.2 Naive Bayes classifier

The naive Bayes classifier is an algorithm based on Bayes Theorem [101]. The points of the training data are plotted according to their features. The test observation is then plotted, and the algorithm determines the probability of this observation belonging to either of the target groups and then compares those probabilities. These probabilities are calculated from the prior probability multiplied by the likelihood and divided by the marginal likelihood. The prior probability is the number of previous observations being classified as the target classification divided by the total number of observations. The marginal likelihood is based on the data-points which are within a radius drawn around the new observation, which are assumed to have similar features. The number of similar observations is divided by the total number of observations, to give the marginal likelihood. The likelihood is again based on the similar observations within the input radius and is the number of similar observations which are the target classification divided by the total observations with the target classification. The input radius can be adjusted, and it is one of the hyperparameters which can be tuned to improve the model. The probability of the second classifier can be calculated by subtracting the first classifier probability from one. The new observation is classified as the classifier with the greatest probability, and is calculated by the equation below:

$$PosteriorProbability = \frac{PriorProbability * Likelihood}{MarginalLikelihood}$$

1.10.3 Logistic regression

Logistic regression is a statistical method for binary classification, which uses a logistic function to bound the regression range between 0 and 1 [101]. The logistic curve used to for predictions is created by selected the curve with the maximum value for the likelihood of the observed classification for the features plotted. Penalised regression approaches can be applied reduce the coefficients in the regression and reduce the chance of over fitting when using multiple features in a predictive model. Ridge regression aims to shrink the features which do not contribute to the predictive model to close to zero, least absolute shrinkage and selection operator (LASSO) forces features which do not contribute significantly to the predictive model to zero and elastic net is a combination of both ridge and LASSO penalisation.

1.10.4 Random forest classifier

A random forest classifier consists of multiple decisions trees, where each decision tree classifies the input and then a consensus is taken from all trees to classify the input (Figure 1.15) [101, 102]. Nodal points, branches and leaves are defined by the ML algorithm from a labelled test dataset. The algorithm will randomly pick features and subsets of the training data (bootstrapped) for each of the trees when creating the forest, and this is repeated until multiple decision trees are created. The number of trees, number of maximum splits, maximum number of samples at the terminal branch and minimum number of samples that can split are all able to be adjusted to improve the predictive ability of the model.

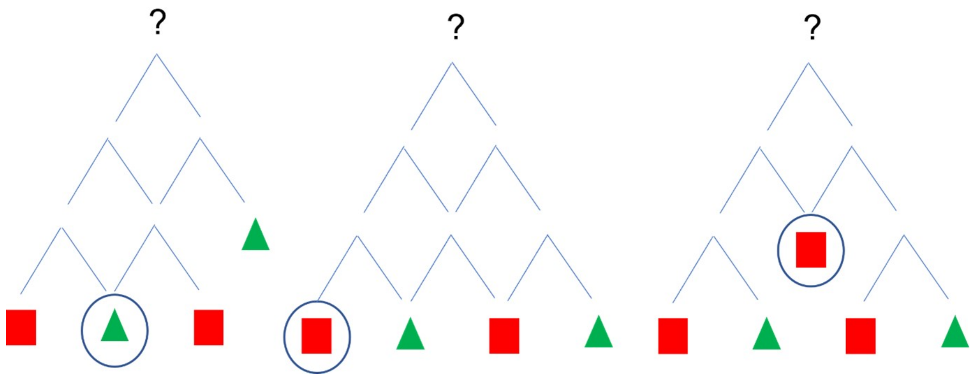


Figure 1.15 Random forest classifier made from 3 decision trees. The branches represent randomly selected features. The prediction of each of the trees on whether the "?" is a triangle or a square is given by the blue circle.

1.10.5 K-nearest neighbour (KNN) classifier

A K-nearest neighbour classifier involves plotting the features of the training data on a multi-dimensional axis [101]. The input data is then plotted on the same grid and a vote from the closest points (closest neighbours) from the training data is carried out to determine how to classify the input data (Figure 1.16). The number of neighbours which are consulted can be adjusted (but needs to be an odd value) as well as the type of distance used to determine which neighbours are the closest. The Euclidean distance is the straight-line distance between two points; the Manhattan (cityblock) distance is the summed distance measured at right angles; the Chebyshev distance, also known as the chessboard distance, is the minimum distance that the King would need to get from one point to the next by moving in the centre of a square on a Chess board; the cosine distance is the similarity of the angle of a point; and the Minkowski distance is a generalisation of both the Euclidean distance and the Manhattan distance.

1.10.6 Support vector machines (SVM)

The SVM classifier creates a model for classification by plotting the features from the training data and then trying to define a vector which splits the groups, with multiple features this a multidimensional vector (Figure 1.17) [101]. The points from each of the groups which are at the boundaries between the different classifications are known as the support vectors. The boundary may not always be linear and therefore kernel functions can be used. The most commonly used kernel functions are Gaussian, sigmoid and polynomial kernels.

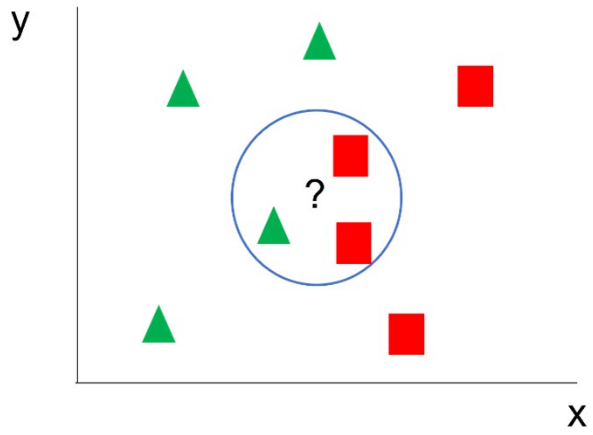


Figure 1.16 K-nearest neighbour classifier using two features (X and Y). The neighbours selected to help classifier the "?" are shown within the blue circle.

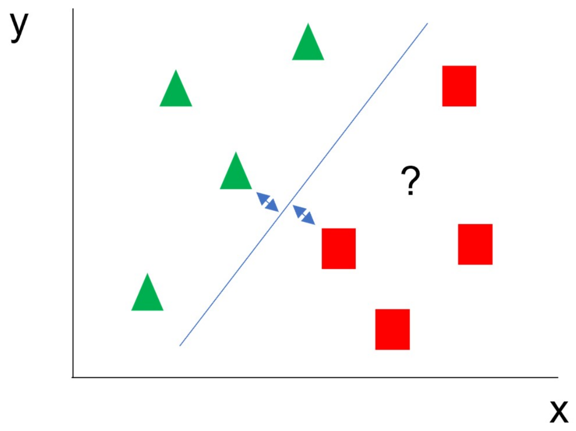


Figure 1.17 Support vector machine using two features (X and Y). The vector is defined by the blue line with the support vectors being the triangle and square closest to the vector highlighted by the blue arrows. In this case the new data point (?) would be classified as a square.

1.10.7 Artificial neural networks

Neural networks are inspired by the neural architecture of the brain and are composed of interconnecting nodes (or neurons) making up an input layer, a number of hidden layers and an output layer (Figure 1.18) [101]. The input layer output layers are defined by the shape of the input and the output whereas the number of hidden layers and quantity of nodes within the model can be adjusted as part of hyperparameter tuning. Connections between nodes of the different layers are assigned a weight and the input data is multiplied by the weight and added to the bias value of the node of the deeper layer. The resultant value is passed through an activation function which determines if the value is above the threshold to allow the node to be activated and permitted to transmit data to the next layer. This is known as forward-propagation. During training the weights are assigned initially at random and then adjusted by the network to reduce the error rate when comparing the predictions to the expected outcomes. This is termed back-propagation. The hyperparameters which can be tuned include the dropout, activation function, learning rate, momentum, number of epochs and batch size.

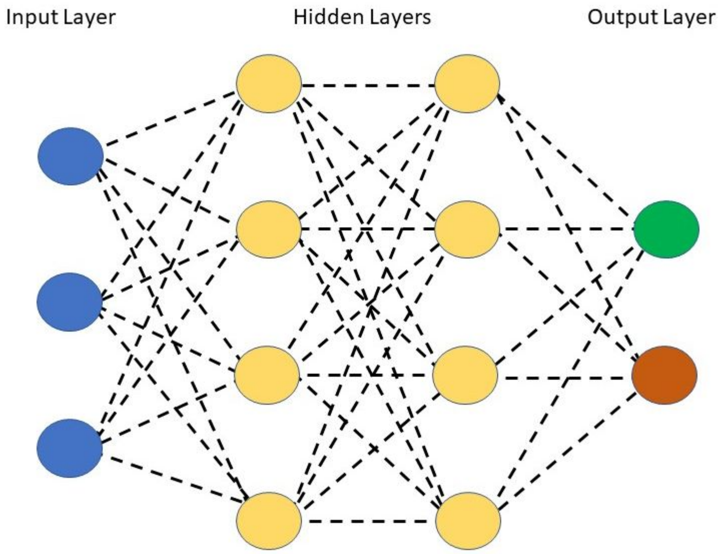


Figure 1.18 Diagrammatic representation of an artificial neural network (ANN) with three nodes forming an input layer (blue), two hidden layers (yellow) both containing four nodes and a binary classification output with two nodes within the output layer (green and orange). The dashed lines represent the connections between the neurons, each of these connections is assigned a weight which can be adjusted during training.

1.10.8 Convolutional neural networks

A convolutional neural network (CNN) is a type of deep learning which is commonly applied to visual classification tasks [103]. A filter/feature detector comprising of a defined pixel array (e.g. 3x3) which has been highlighted from the networks training, is applied to an image sequentially along the row of pixels throughout the image. The distance the feature detector moves is defined as the 'stride' which can be manually determined [103]. The values of pixels of the filter map in each position are multiplied by the input image at that position and summed together. This value is then added to a new array, known as the feature map/activation map, which reduces the size of the array but keeps the spatial information. This is then repeated for other feature detectors to create a layer of feature maps. These are then rectified to reduced linearity [103]. A pooling/down sampling step is then applied to the feature maps, which often takes the form of max pooling where the feature map is again sequentially passed through region by region (e.g. 2x2 with a stride of 1) and the maximum pixel value within that region is recorded in a new array. Other methods of pooling include sum pooling, where the sum of the pixels within the region are added together, or mean pooling, which takes the mean of the pixels recorded. The new array is defined as the pooled feature map which reduces the size of the array but retains the spatial information. All the pooled feature maps are flattened, which produces a linear array [103]. This linear array is the input for the hidden/fully connected layers of an ANN which create the prediction for the image classifier.

The hyperparameters which can be adjusted include the number of hidden layers, dropout, network weight initialisation, learning rate, momentum, number of epochs and batch size.

1.11 Practical implementation of imaging biomarkers

The next chapters will build on the information presented in the introduction to explore the current literature surrounding the use of imaging biomarkers in cHL and DLBCL, as well as train and internally test machine learning based predictive models.

1.12 References

1. Shanbhag S, Ambinder RF. Hodgkin lymphoma: A review and update on recent progress. CA Cancer J Clin. 2018;68:116–32.

2. Armitage JO, Gascoyne RD, Lunning MA, Cavalli F. Non-Hodgkin lymphoma. Lancet. 2017;390:298–310.

3. Sukswai N, Lyapichev K, Khoury JD, Medeiros LJ. Diffuse large B-cell lymphoma variants: an update. Pathology. 2020;52:53–67.

4. Harris NL. Shades of gray between large B-cell lymphomas and hodgkin lymphomas: Differential diagnosis and biological implications. Mod Pathol. 2013;26:57–70.

5. Tavian M, Robin C, Coulombel L, Péault B. The human embryo, but not its yolk sac, generates lympho-myeloid stem cells: Mapping multipotent hematopoietic cell fate in intraembryonic mesoderm. Immunity. 2001;15:487–95.

6. Ng AP, Alexander WS. Haematopoietic stem cells: Past, present and future. Cell Death Discov. 2017;3:2–5.

7. Ghosn EEB, Yamamoto R, Hamanaka S, Yang Y, Herzenberg LA, Nakauchi H, *et al.* Distinct B-cell lineage commitment distinguishes adult bone marrow hematopoietic stem cells. Proc Natl Acad Sci U S A. 2012;109:5394–8.

8. Ghosn EEB, Waters J, Phillips M, Yamamoto R, Long BR, Yang Y, *et al.* Fetal Hematopoietic Stem Cell Transplantation Fails to Fully Regenerate the B-Lymphocyte Compartment. Stem Cell Reports. 2016;6:137–49.

9. Graf R, Seagal J, Otipoby KL, Lam KP, Ayoub S, Zhang B, *et al.* BCR-dependent lineage plasticity in mature B cells. Science. 2019;363:748–53.

10. Yoshimoto M. The ontogeny of murine B-1a cells. Int J Hematol. 2020;111:622–7.

11. Laurenti E, Göttgens B. From haematopoietic stem cells to complex differentiation landscapes. Nature. 2018;553:418–26.

12. Engelhardt M, Lübbert M, Guo Y. CD34+ or CD34-: Which is the more primitive? Leukemia. 2002;16:1603–8.

13. Yamamoto R, Morita Y, Ooehara J, Hamanaka S, Onodera M, Rudolph KL, *et al.* XClonal analysis unveils self-renewing lineage-restricted progenitors generated directly from hematopoietic stem cells. Cell. 2013;154:1112–26.

14. Cheng H, Zheng Z, Cheng T. New paradigms on hematopoietic stem cell differentiation. Protein Cell. 2020;11:34–44.

15. Follows GA, Barrington SF, Bhuller KS, Culligan DJ, Cutter DJ, Gallop-Evans E, *et al.* Guideline for the first-line management of Classical Hodgkin Lymphoma — A British Society for Haematology guideline. Br J Haematol. 2022;197(5):558-72.

16. Lebien TW, Tedder TF. B lymphocytes: How they develop and function. Blood. 2008;112:1570–80.

17. Pieper K, Grimbacher B, Eibel H. B-cell biology and development. J Allergy Clin Immunol. 2013;131:959–71.

18. Singh D, Vaccarella S, Gini A, De Paula Silva N, Steliarova-Foucher E, Bray F. Global patterns of Hodgkin lymphoma incidence and mortality in 2020 and a prediction of the future burden in 2040. Int J Cancer. 2022;150:1941–7.

19. Engert A, Anas Y. Hodgkin Lymphoma. 3rd Ed. Engert A, Younes A, editors. Cham: Springer International Publishing; 2020.

20. MacMahon B. Epidemiological evidence of the nature of Hodgkin's disease. Cancer. 1957;10:1045–54.

21. Yung L, Linch D. Hodgkin ' s Lymphoma. Lancet. 2003;361:943–51.

22. Mauch PM, Kalish LA, Kadin M, Coleman CN, Osteen R, Hellman S. Patterns of presentation of Hodgkin disease. Implications for etiology and pathogenesis. Cancer. 1993;71:2062–71.

23. Carbone PP, Kaplan HS, Musshoff K, Smithers DW, Tubiana M. Report of the Committee on Hodgkin's Disease Staging Classification. Cancer Res. 1971;31:1860–1.

24. Kurzrock R, Redman J, Cabanillas F, Jones D, Rothberg J, Talpaz M. Serum Interleukin 6 Levels Are Elevated in Lymphoma Patients and Correlate with Survival in Advanced Hodgkin's Disease and with B Symptoms. Cancer Res. 1993;53:2118–22.

25. Nam-Cha SH, Montes-Moreno S, Salcedo MT, Sanjuan J, Garcia JF, Piris MA. Lymphocyte-rich classical hodgkin's lymphoma: Distinctive tumor and microenvironment markers. Mod Pathol. 2009;22:1006–15.

26. Karube K, Niino D, Kimura Y, Ohshima K. Classical Hodgkin lymphoma, lymphocyte depleted type: Clinicopathological analysis and prognostic comparison with other types of classical Hodgkin lymphoma. Pathol Res Pract. 2013;209:201–7.

27. Rafiq M, Hayward A, Warren-Gash C, Denaxas S, Gonzalez-Izquierdo A, Lyratzopoulos G, *et al.* Socioeconomic deprivation and regional variation in

Hodgkin's lymphoma incidence in the UK: A population-based cohort study of 10 million individuals. BMJ Open. 2019;9:1–8.

28. Shenoy P, Maggioncalda A, Malik N, Flowers CR. Incidence Patterns and Outcomes for Hodgkin Lymphoma Patients in the United States. Adv Hematol. 2011;2011:1–11.

29. Chang ET, Zheng T, Weir EG, Borowitz M, Mann RB, Spiegelman D, *et al.* Childhood social environment and Hodgkin's lymphoma: New findings from a population-based case-control study. Cancer Epidemiol Biomarkers Prev. 2004;13:1361–70.

30. Levine PH, Ablashi D V., Berard CW, Carbone PP, Waggoner DE, Malan L. Elevated antibody titers to epstein-barr virus in Hodgkin's disease. Cancer. 1971;27:416–21.

31. Jarrett RF, Gallagher A, Jones DB, Alexander FE, Krajewski AS, Kelsey A, *et al.* Detection of Epstein-Barr virus genomes in Hodgkin's disease: Relation to age. J Clin Pathol. 1991;44:844–8.

32. Advani RH, Hoppe RT, Maeda LS, Baer DM, Mason J, Rosenberg SA, *et al.* Stage I-IIA Non-Bulky Hodgkin's Lymphoma. Is Further Distinction Based on Prognostic Factors Useful? The Stanford Experience. Int J Radiat Oncol. 2011;81:1374–9.

33. Hasenclever D, Diehl V, Armitage JO, Assouline D, Björkholm M, Brusamolino E, *et al.* A Prognostic Score for Advanced Hodgkin's Disease. N Engl J Med. 1998;339:1506–14.

34. Diefenbach CS, Li H, Hong F, Gordon LI, Fisher RI, Bartlett NL, *et al.* Evaluation of the International Prognostic Score (IPS-7) and a Simpler Prognostic Score (IPS-3) for advanced Hodgkin lymphoma in the modern era. Br J Haematol. 2015;171:530–8.

35. Gallamini A, Barrington SF, Biggi A, Chauvie S, Kostakoglu L, Gregianin M, *et al.* The predictive role of interim positron emission tomography for Hodgkin lymphoma treatment outcome is confirmed using the interpretation criteria of the Deauville five-point scale. Haematologica. 2014;99:1107–13.

36. Bari A, Marcheselli L, Marcheselli R, Pozzi S, Cox MC, Baldessari C, *et al.* Absolute monocyte count at diagnosis could improve the prognostic role of early FDG-PET in classical Hodgkin lymphoma patients. Br J Haematol. 2018;180:600–2.

37. Pauwels EKJ, Ribeiro MJ, Stoot JHMB, McCready VR, Bourguignon M, Mazière B. FDG accumulation and tumor biology. Nucl Med Biol. 1998;25:317–22.

38. Allisy-Roberts PJ, Williams J. Farr's Physics for Medical Imaging. 2nd ed. Philadelphia: Saunders Ltd.; 2007.

39. Lister TA, Crowther D, Sutcliffe SB, Glatstein E, Canellos GP, Young RC, *et al.* Report of a committee convened to discuss the evaluation and staging of patients with Hodgkin's disease: Cotswolds meeting. J Clin Oncol. 1989;7:1630–6.

40. Kanoun S, Rossi C, Casasnovas O. [18F]FDG-PET/CT in hodgkin lymphoma: Current usefulness and perspectives. Cancers (Basel). 2018;10:1–11.

41. Cheson BD, Fisher RI, Barrington SF, Cavalli F, Schwartz LH, Zucca E, *et al.* Recommendations for initial evaluation, staging, and response assessment of hodgkin and non-hodgkin lymphoma: The lugano classification. J Clin Oncol. 2014;32:3059–67.

42. . Barrington SF, Mikhaeel NG, Kostakoglu L, Meignan M, Hutchings M, Müeller SP, *et al.* Role of imaging in the staging and response assessment of lymphoma: Consensus of the international conference on malignant lymphomas imaging working group. J Clin Oncol. 2014;32:3048–58.

43. Fletcher JW, Kinahan PE. PET/CT SUVs in clinical practice. Semin Ultrasound CT MR. 2010;31:496–505.

44. Adams MC, Turkington TG, Wilson JM, Wong TZ. A systematic review of the factors affecting accuracy of SUV measurements. Am J Roentgenol. 2010;195:310–20.

45. Fletcher JW, Kinahan PE. PET/CT Standardized Uptake Values (SUVs) in Clinical Practice and Assessing Response to Therapy. Natl Inst Heal. 2010;31:496–505.

46. Decazes P, Becker S, Toledano MN, Vera P, Desbordes P, Jardin F, *et al.* Tumor fragmentation estimated by volume surface ratio of tumors measured on 18F-FDG PET/CT is an independent prognostic factor of diffuse large B-cell lymphoma. Eur J Nucl Med Mol Imaging. 2018;45:1672–9.

47. Johnson P, Federico M, Kirkwood A, Fosså A, Berkahn L, Carella A, *et al.* Adapted Treatment Guided by Interim PET-CT Scan in Advanced Hodgkin's Lymphoma. N Engl J Med. 2016;374:2419–29.

48. Kreissl S, Goergen H, Buehnen I, Kobe C, Moccia A, Greil R, *et al.* PET-guided eBEACOPP treatment of advanced-stage Hodgkin lymphoma (HD18): follow-up analysis of an international, open-label, randomised, phase 3 trial. Lancet Haematol. 2021;8:e398–409.

49. Kushekhar K, Van Den Berg A, Nolte I, Hepkema B, Visser L, Diepstra A. Genetic associations in classical Hodgkin lymphoma: A systematic review and insights into susceptibility mechanisms. Cancer Epidemiol Biomarkers Prev. 2014;23:2737–47.

50. Montalbán C, García JF, Abraira V, González-Camacho L, Morente MM, Bello JL, *et al.* Influence of biologic markers on the outcome of Hodgkin's lymphoma: A study by the Spanish Hodgkin's Lymphoma Study Group. J Clin Oncol. 2004;22:1664–73.

51. Plattel WJ, Alsada ZND, van Imhoff GW, Diepstra A, van den Berg A, Visser L. Biomarkers for evaluation of treatment response in classical Hodgkin lymphoma: comparison of sGalectin-1, sCD163 and sCD30 with TARC. Br J Haematol. 2016;175:868–75.

52. HMRN. Haematological Malignancy Research Network (HMRN) - Incidence statistics: Disease-specific incidence. https://hmrn.org/statistics/incidence. Accessed 25 May 2021.

53. SEER. SEER Cancer Statistics Review. 1975-2016. 2018. https://seer.cancer.gov/statfacts. Accessed 25 May 2021.

54. Miranda-Filho A, Piñeros M, Znaor A, Marcos-Gragera R, Steliarova-Foucher E, Bray F. Global patterns and trends in the incidence of non-Hodgkin lymphoma. Cancer Causes Control.2019;30:489–99.

55. Shi Y, Han Y, Yang J, Liu P, He X, Zhang C, *et al.* Clinical features and outcomes of diffuse large B-cell lymphoma based on nodal or extranodal primary sites of origin: Analysis of 1,085 WHO classified cases in a single institution in China. Chinese J Cancer Res. 2019;31:152–61.

56. Ben-Dali Y, Hleuhel MH, da Cunha-Bang C, Brieghel C, Poulsen CB, Clasen-Linde E, *et al.* Richter's transformation in patients with chronic lymphocytic leukaemia: a Nationwide Epidemiological Study. Leuk Lymphoma. 2020;61:1435–44.

57. Menon MP, Pittaluga S, Jaffe ES. The Histological and Biological Spectrum of Diffuse Large B-Cell Lymphoma in the World Health Organization Classification. Cancer J. 2012;18:411–20.

58. Cerhan JR, Kricker A, Paltiel O, Flowers CR, Wang SS, Monnereau A, *et al.* Medical history, lifestyle, family history, and occupational risk factors for Diffuse Large B-cell Lymphoma: The interLymph non-Hodgkin lymphoma subtypes project. J Natl Cancer Inst - Monogr. 2014;15–25.

59. Shipp MA, Anderson J. A predictive model for aggressive non-Hodgkin's lymphoma. New English J Med. 1993;329:987–94.

60. Sehn LH, Berry B, Chhanabhai M, Fitzgerald C, Gill K, Hoskins P, *et al.* R-IPI is a better predictor of outcome than the standard IPI for patients with DLBCL treated with R-CHOP. Blood. 2007;109:1857–62.

61. Ziepert M, Hasenclever D, Kuhnt E, Glass B, Schmitz N, Pfreundschuh M, *et al.* Standard international prognostic index remains a valid predictor of outcome for patients with aggressive CD20+ B-cell lymphoma in the rituximab era. J Clin Oncol. 2010;28:2373–80.

62. Zhou Z, Sehn LH, Rademaker AW, Gordon LI, LaCasce AS, Crosby-Thompson A, *et al.* An enhanced International Prognostic Index (NCCN-IPI) for patients with diffuse large B-cell lymphoma treated in the rituximab era. Blood. 2014;123:837–42.

63. Ruppert AS, Dixon JG, Salles GA, Wall A, Cunningham D, Poeschel V, *et al.* International prognostic indices in diffuse large B-cell lymphoma (DLBCL): a comparison of IPI, R-IPI and NCCN-IPI. Blood. 2020;135:2041-8

64. Eertink JJ, Burggraaff CN, Heymans MW, Dührsen U, Hüttmann A, Schmitz C, *et al.* Optimal timing and criteria of interim PET in DLBCL: A comparative study of 1692 patients. Blood Adv. 2021;5:2375–84.

65. Adams HJA, Kwee TC. Prognostic value of interim FDG-PET in R-CHOP-treated diffuse large B-cell lymphoma: Systematic review and meta-analysis. Crit Rev Oncol Hematol. 2016;106:55–63.

66. Coiffier B, Thieblemont C, Van Den Neste E, Lepeu G, Plantier I, Castaigne S, *et al.* Long-term outcome of patients in the LNH-98.5 trial, the first randomized study comparing rituximab-CHOP to standard CHOP chemotherapy in DLBCL patients: A study by the Groupe d'Etudes des Lymphomes de l'Adulte. Blood. 2010;116:2040–5.

67. NICE. Non-Hodgkin's lymphoma: diagnosis and management. 2016. www.nice.org.uk/guidance/ng52. Accessed 25 May 2021.

68. Kansara R. Central Nervous System Prophylaxis Strategies in Diffuse Large B Cell Lymphoma. Curr Treat Options Oncol. Current Treatment Options in Oncology; 2018;19:52

69. Gutiérrez-García G, Cardesa-Salzmann T, Climent F, González-Barca E, Mercadal S, Mate JL, *et al.* Gene-expression profiling and not immunophenotypic algorithms predicts prognosis in patients with diffuse large B-cell lymphoma treated with immunochemotherapy. Blood. 2011;117:4836–43.

70. Scott DW, Mottok A, Ennishi D, Wright GW, Farinha P, Ben-Neriah S, *et al.* Prognostic significance of diffuse large B-cell lymphoma cell of origin determined by digital gene expression in formalin-fixed paraffin-embedded tissue biopsies. J Clin Oncol. 2015;33:2848–56.

71. Liu Y, Barta SK. Diffuse large B-cell lymphoma: 2019 update on diagnosis, risk stratification, and treatment. Am J Hematol. 2019;94:604–16.

72. Langer S. Issues surrounding PACS archiving to external, third-party DICOM archives. J Digit Imaging. 2009;22:48–52.

73. El Hajal G, Abi Zeid Daou R, Ducq Y, Börcsök J. Designing and validating a cost effective safe network: Application to a PACS system. Int Conf Adv Biomed Eng ICABME. 2019;23–6.

74. Valente F, Silva LAB, Godinho TM, Costa C. Anatomy of an Extensible Open Source PACS. J Digit Imaging. Journal of Digital Imaging; 2016;29:284–96.

75. Riddle WR, Pickens DR. Extracting data from a DICOM file. Med Phys. 2005;32:1537–41.

76. DICOM Library. Dicom-tags. 2022. https://www.dicomlibrary.com/dicom/dicom-tags. Accessed 16 Jul 2022.

77. Aryanto KYE, Oudkerk M, van Ooijen PMA. Free DICOM de-identification tools in clinical research: functioning and safety of patient privacy. Eur Radiol. 2015;25:3685–95.

78. Barrington SF, Meignan M. Time to prepare for risk adaptation in lymphoma by standardizing measurement of metabolic tumor burden. J Nucl Med. 2019;60:1096–102.

79. Botta F, Ferrari M, Raimondi S, Corso F, Lo Presti G, Mazzara S, *et al.* The Impact of Segmentation Method and Target Lesion Selection on Radiomic Analysis of 18F-FDG PET Images in Diffuse Large B-Cell Lymphoma. Appl Sci. 2022;12:9678.

80. Erdi YE, Mawlawi O, Larson SM, Imbriaco M, Yeung H, Finn R, *et al.* Segmentation of lung lesion volume by adaptive positron emission tomography image thresholding. Cancer. 1997;80:2505–9.

81. Triumbari EKA, Morland D, Cuccaro A, Maiolo E, Hohaus S, Annunziata S. Classical Hodgkin Lymphoma: A Joint Clinical and PET Model to Predict Poor Responders at Interim Assessment. Diagnostics. 2022;12:2325.

82. Thieblemont C, Chartier L, Duhrsen U, Vitolo U, Barrington S, Zaucha JM, *et al.* A tumor volume and performance status model to predict outcome prior to treatment in diffuse large B-cell lymphoma. Blood Adv. 2022; doi: 10.1182/bloodadvances.2021006923.

83. Burggraaff CN, de Jong A, Hoekstra OS, Hoetjes NJ, Nievelstein RAJ, Jansma EP, *et al.* Predictive value of interim positron emission tomography in diffuse large B-cell lymphoma: a systematic review and meta-analysis. Eur J Nucl Med Mol Imaging. European Journal of Nuclear Medicine and Molecular Imaging; 2019;46:65–79.

84. Zhong J, Frood R, Brown P, Nelstrop H, Prestwich R, McDermott G, *et al.* Machine learning-based FDG PET-CT radiomics for outcome prediction in larynx and hypopharynx squamous cell carcinoma. Clin Radiol. 2021;76:78.e9-78.e17.

85. Brown PJ, Zhong J, Frood R, Currie S, Gilbert A, Appelt AL, *et al.* Prediction of outcome in anal squamous cell carcinoma using radiomic feature analysis of pre-treatment FDG PET-CT. Eur J Nucl Med Mol Imaging. 2019;46:2790–9.

86. Zwezerijnen GJC, Eertink JJ, Ferrández MC, Wiegers SE, Burggraaff CN, Lugtenburg PJ, *et al.* Reproducibility of [18F]FDG PET/CT liver SUV as reference or normalisation factor. Eur J Nucl Med Mol Imaging. 2022; doi: 10.1007/s00259-022-05977-5.

87. Larobina M, Murino L. Medical image file formats. J Digit Imaging. 2014;27:200–6.

88. Li X, Morgan PS, Ashburner J, Smith J, Rorden C. The first step for neuroimaging data analysis: DICOM to NIfTI conversion. J Neurosci Methods. 2016;264:47–56.

89. Shakeshaft J. Picture Archiving and Communications System in radiotherapy. Clin Oncol. 2010;22:681–7.

90. Ahmed A, Ali M, Salah H, Eisa RE, Mohieldin H, Omer H, *et al.* Evaluation of uptake values of FDG: Body surface area Vs. body weight correction. Radiat Phys Chem. 2022;201:110482.

91. Law MYY, Liu B. Informatics in radiology: DICOM-RT and its utilization in radiation therapy. Radiographics. 2009;29:655–67.

92. Lambin P, Leijenaar RTH, Deist TM, Peerlings J, De Jong EEC, Van Timmeren J, *et al.* Radiomics: The bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol. 2017;14:749–62.

93. Zwanenburg A, Leger S, Vallières M, Löck S. Image biomarker standardisation initiative. Radiology. 2020;295:328-38.

94. PyRadiomics. Radiomic Features. 2016. https://pyradiomics.readthedocs.io/en/latest/features.html. Accessed 10 Jul 2022.

95. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. England; 2007;8:118–27.

96. Orlhac F, Boughdad S, Philippe C, Stalla-Bourdillon H, Nioche C, Champion L, *et al.* A postreconstruction harmonization method for multicenter radiomic studies in PET. J Nucl Med. 2018;59:1321–8.

97. Orlhac F, Frouin F, Nioche C, Ayache N, Buvat I. Validation of A Method to Compensate Multicenter Effects Affecting CT Radiomics. Radiology. 2019;291:53–9.

98. Choy G, Khalilzadeh O, Michalski M, Do S, Samir AE, Pianykh OS, *et al.* Current applications and future impact of machine learning in radiology. Radiology. 2018;288:318–28.

99. Whalen S, Schreiber J, Noble WS, Pollard KS. Navigating the pitfalls of applying machine learning in genomics. Nat Rev Genet. 2022;23:169–81.

100. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit. 1997;30:1145–59.

101. Biship CM. Pattern recognition and machine learning (information science and statistics). 1st ed. New York: Spinger; 2007.

102. Breiman L. Random Forests. Mach Learn. 2001;45:5–32.

103. Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. Insights Imaging. Insights into Imaging; 2018;9:611–29.

Chapter 2
Baseline PET/CT imaging parameters for prediction of treatment outcome in Hodgkin and diffuse large B cell lymphoma: a systematic review

2.1 Abstract

2.1.1 Purpose

To systematically review the literature evaluating clinical utility of imaging metrics derived from baseline fluorine-18 fluorodeoxyglucose positron emission tomography/computed tomography (PET/CT) for prediction of progression-free (PFS) and overall survival (OS) in patients with classical Hodgkin lymphoma (HL) and diffuse large B cell lymphoma (DLBCL).

2.1.2 Methods

A search of MEDLINE/PubMed, Web of Science, Cochrane, Scopus and clinicaltrials.gov databases was undertaken for articles evaluating PET/CT imaging metrics as outcome predictors in HL and DLBCL. PRISMA guidelines were followed. Risk of bias was assessed using the Quality in Prognosis Studies (QUIPS) tool.

2.1.3 Results

Forty-one articles were included (31 DLBCL, 10 HL). Significant predictive ability was reported in 5/20 DLBCL studies assessing SUVmax (PFS: HR 0.13–7.35, OS: HR 0.83–11.23), 17/19 assessing metabolic tumour volume (MTV) (PFS: HR 2.09–11.20, OS: HR 2.40–10.32) and 10/13 assessing total lesion glycolysis (TLG) (PFS: HR 1.078–11.21, OS: HR 2.40–4.82). Significant predictive ability was reported in 1/4 HL studies assessing SUVmax (HR not reported), 6/8 assessing MTV (PFS: HR 1.2–10.71, OS: HR 1.00–13.20) and 2/3 assessing TLG (HR not reported). There are 7/41 studies assessing the use of radiomics (4 DLBCL, 2 HL); 5/41 studies had internal validation and 2/41 included external validation. All studies had overall moderate or high risk of bias.

2.1.4 Conclusion

Most studies are retrospective, underpowered, heterogenous in their methodology and lack external validation of described models. Further work in protocol harmonisation,

automated segmentation techniques and optimum performance cut-off is required to develop robust methodologies amenable for clinical utility.

2.2 Background

Lymphoma is a haematopoietic malignancy, which can be broadly categorised into Hodgkin and non-Hodgkin disease. Hodgkin lymphoma (HL) accounts for approximately 10% of all newly diagnosed cases, and its hallmark is the presence of Hodgkin and Reed–Sternberg (HRS) cells [1]. HL can be further sub-divided based on morphology and immunohistochemistry into classical Hodgkin lymphoma (cHL), which has four further sub-categories, or nodular lymphocyte-predominant Hodgkin lymphoma (NLPHL) [1]. The majority (90%) of disease is due to cHL. HL is associated with a good prognosis having an overall 5-year survival of 86.6% [2]. Non-Hodgkin lymphoma (NHL) is the most prevalent form of lymphoma with over 50 sub-types, the most common being diffuse large B cell lymphoma (DLBCL) [3]. The overall 5-year survival rate is 72% for NHL but this varies by stage and subtype [2]. DLBCL has a 5-year survival of approximately 60–80%, which has improved since the use of anthracycline-containing chemotherapy and rituximab (R-CHOP) [2, 4].

There are several pretreatment clinical prognostic tools developed to stratify both DLBCL and HL. In 1993, Shipp *et al.* introduced the international prognostic index (IPI) for predicting overall survival in DLBCL patients based on a retrospective study of 2031 patients treated with CHOP. The IPI has been further refined with an age-adjusted version (aa-IPI), a revised version developed following the use of R-CHOP (R-IPI), and a version based on the National Comprehensive Cancer Network database (NCCN-IPI). HL disease can be split into early (stage I and II) or advanced (stage III or stage IV) with early being split into favourable or unfavourable depending on one of the many scoring systems including, but not limited to, the German Hodgkin Study Group (GHSG), European Organisation of Research and Treatment of Cancer (EORTC), Groupe d'Etudes des Lymphomes de l'Adulte (GELA), National Cancer Institute (NCI) or National Comprehensive Cancer Network 2010 (NCCN 2010) scores. However, given the variation in the prognostic groups derived from the different scoring systems, further information obtained from imaging may improve prognostication.

2-deoxy-2-[Fluorine-18]fluoro-D-glucose (FDG) positron emission tomography/computed tomography (PET/CT) is widely used for staging and response assessment in HL and NHL [5]. Response assessment PET/CT studies are performed at various time points, including during and after treatment [5]. The parameter most commonly used in assessment is the standardised uptake value (SUV) at sites of disease, which is compared to physiological activity in reference areas such as the mediastinal blood pool and liver and is reported using an ordinal (qualitative) scale (Deauville Score (DS)).

A variety of imaging-derived quantitative parameters have been reported in the literature with potential utility for predicting prognosis or treatment outcome. These metrics range from those based on tumour volume to metabolic features, including shape and texture. At present, none have been translated into routine clinical practice. The purpose of this study was to perform a systematic review of the literature reporting the use of quantitative imaging parameters derived from pretreatment FDG PET/CT for prediction of treatment outcome for HL and DLBCL. Due to the varied nature of NHL, DLBCL was chosen as it is the most common subtype of NHL.

2.3 Methods and Materials

2.3.1 Search strategy and selection criteria

A search of MEDLINE/PubMed, Web of Science, Cochrane, Scopus and clinicaltrials.gov databases was performed for articles on PET/CT imaging parameters in lymphoma treatment assessment. The search strategy included three primary operator criteria linked with the "AND" function. The first criteria consisted of "lymphoma", the second of "PET" or "positron emission tomography", and the third of "outcome", "prognosis", "prediction", "parameter", "radiomics", "machine learning", "deep learning" or "artificial intelligence". Case studies, articles not published in English, phantom studies, studies not assessing treatment outcomes using baseline imaging in HL or DLCBL, studies assessing primary anatomical presentations of lymphoma or HIV-related lymphoma, mixed pathology studies and studies assessing novel treatments were excluded. After duplications were excluded, studies were screened for eligibility based on the title, abstract and subsequently on full text. The references of the articles included in the systematic review were manually reviewed to identify further publications which met the inclusion criteria. The results were stored in bibliographic management software. Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) criteria were adhered to [6].

2.3.2 Quality assessment

The Quality in Prognosis Studies (QUIPS) tool was used to evaluate validity and bias which considers six areas: inclusion, attrition, prognostic factor measurement, confounders, outcome measurement, and analysis and reporting [7]. Prompting questions and modifications applied to the QUIPS tool are detailed in Supplemental Table 2.1. Two authors (RF and AS) independently reviewed all studies which met inclusion criteria and scored each of the six domains as high, moderate or low risk of bias. Any discrepancies were agreed in consensus. Overall risk of bias for each paper was further categorised based on the following criteria: if all domains were classified as low risk, or there was up to one moderate risk, the paper was classified as low risk of bias. If one or more domains were classified as high risk, the paper was classified as high risk of bias. All papers in between were classified as having moderate risk of bias [8].

2.4 Results

Results are current to July 2020. The database search strings yielded 2717 results after duplicates were excluded. Following screening and assessment of eligibility, 41 articles meeting the study inclusion criteria were included. Figure 2.1 details the study selection.

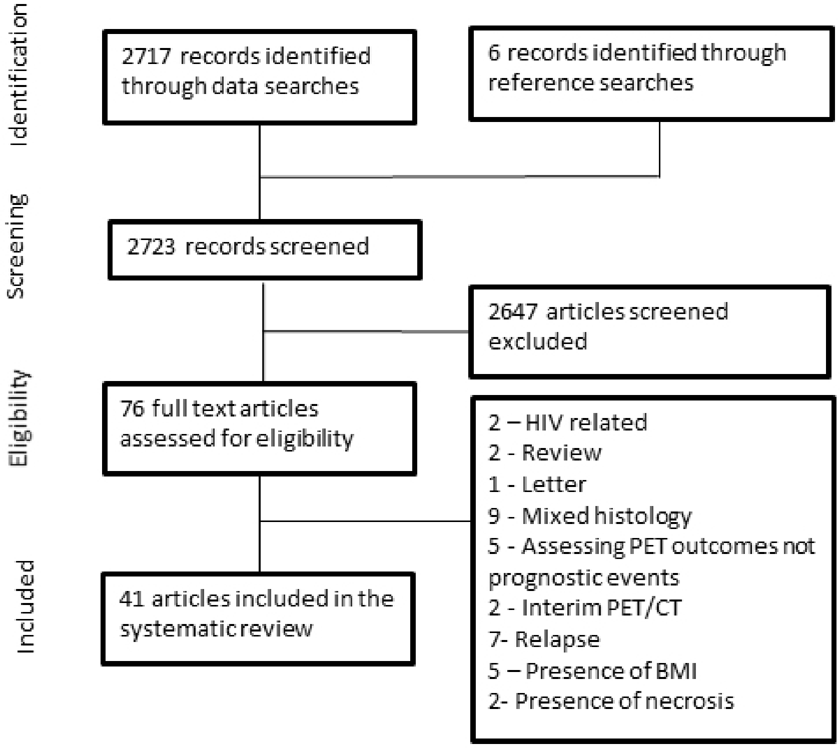

Identification
Screening
Eligibility
Included
2717 records identified through data searches
6 records identified through reference searches
2723 records screened
2647 articles screened excluded
76 full text articles assessed for eligibility
2 – HIV related
2 - Review
1 - Letter
9 - Mixed histology
5 - Assessing PET outcomes not prognostic events
2 - Interim PET/CT
7- Relapse
5 – Presence of BMI
2- Presence of necrosis
41 articles included in the systematic review

Figure 2.1 PRISMA flow diagram illustrating the methodology for study selection for the systematic review of lymphoma imaging parameters. BMI bone marrow involvement, Relapse indicates studies investigating previously treated cases

2.4.1 Quality assessment

No studies showed low risk of bias in all six domains (Supplemental Table 2.2). Only two studies demonstrated a low risk for participation; no studies had a low risk in attrition, prognostic measurement, outcome measurement or confounding factors; 33 studies had low risk for analysis and reporting. All studies were assessed as having either moderate (24/41, 59%) or high (17/41, 41%) overall risk of bias. Of the high risk studies, 6 had high risk scores of bias in participation, 5 in attrition, 8 in prognostic measurement, 8 in outcome measurement, 10 in confounding factors and 7 in analysis and reporting categories.

All studies were retrospective, with 28/41 single centre. Six reports were based on retrospective analysis of trial data from prospective studies. Four studies stated that they were compliant with the European Association of Nuclear Medicine (EANM) guidelines with their scanning protocol; 10/41 did not take into consideration important co-founders such as different treatment regimes, stage, prognostic scores or histology. Only six studies defined the method for calculation of SUV, and 7 studies used a validation cohort to test the predictive models (Table 2.1). Of the radiomic studies, one study referenced the image biomarker standardisation initiative (IBSI) within the discussion but none of the papers explicitly stated that they had complied with IBSI guidelines.

Study	Prospective	Multi-Centre	PET Scanners used	EANM guidelines stated	SUV Defined	Definition of prognostic factor provide	Follow up period	Separate Validation Cohort	Overall risk of bias
Adams [9]	N	N	Siemens Biograph 40 TruePoint	N	N	PFS - relapse / progression / death attributable to PFS OS - Death from any cause	Median: 994 days	N	Moderate
Aide [10]	N	N	Siemens Biograph TrueV	N	N	EFS - relapse / progression / unplanned treatment / death attributable to EFS	2-year EFS	Y	High
Aide [11]	N	N	Siemens Biograph TrueV	Y	N	PFS - relapse / progression OS - death from lymphoma or treatment	Median: 25.7 months	N	Moderate
Akhtari [12]	N	N	GE Discovery ST GE Discovery RX GE Discovery STE	N	Y(bw)	FFP - relapse or refractory disease OS - death from any cause	Median: 4.96 years	N	Moderate
Albano [13]	N	Y	GE Discovery ST GE Discovery 690	Y	N	PFS - progression/relapse/death OS - death from any cause	Median: 40 months	N	Moderate
Angelopulou [14]	N	N	Multiple not defined	N	N	FFP - relapse or refractory disease OS - death from any cause	Median: 56 months	N	High
Capobianco [15]	N	Y	Multiple	N	N	Not defined	Median: 5 years	Y	High
Ceriani [16]	N	Y	Multiple not defined	N	N	Not defined	Median: 64 months, 34 months	Y	High
Chang [17]	N	N	GE Discovery ST	N	N	PFS - progression / relapse / death OS - death from any cause	Median: 28.7 months	N	Moderate
Chang [18]	N	N	GE Discovery ST	N	N	PFS - progression / relapse / death OS - death from any cause	median 36 months	N	Moderate

Study	Prospective	Multi-Centre	PET Scanners used	EANM guidelines stated	SUV Defined	Definition of prognostic factor provide	Follow up period	Separate Validation Cohort	Overall risk of bias
Chihara [19]	N	N	GE Discovery LS	N	Y(bw)	PFS - progression / relapse / death from any cause OS - death from any cause	Median: 34.4 months	N	Moderate
Cottereau [20]	N	Y	Multiple not defined	N	N	Not defined	Median: 44 months	N	High
Cottereau [21]	N	Y	Multiple not defined	N	N	PFS - progression / death from any cause OS - death from any cause	Median: 55 months	Y	Moderate
Cottereau [22]	N	N	Siemens Biograph 16	N	N	OS and PFS were defined according to the revised NCI criteria	Median: 64 months	N	Moderate
Decazes [23]	N	N	Siemens Biograph Sensation 16 HiRes	N	N	Both OS and PFS were defined according to the revised NCI criteria	Median: 44 months	N	Moderate
Esfahani [24]	N	N	Siemens Biograph	N	N	PFS - recurrence	Mean: 51 months	N	High
Gallicchio [25]	N	N	GE Discovery VCT GE Discovery LS VCT	N	N	Progression / disease-related death	Median: 18 months	N	High
Huang [26]	N	N	GE Discovery LS	N	Y(bw)	PFS - progression / relapse / death OS - death from any cause	Median: 30 months	N	Moderate
Ilyas [27]	N	N	GE Discovery ST GE Discovery VCT	N	N	PFS - progression/death from any cause OS - death from any cause	Median: 3.8 years	N	High
Jegadesh [28]	N	N	Not defined	N	N	Not defined	Median: 43.9 months	N	Moderate
Kanoun [29]	N	N	Philips Gemini GXL Philips Gemini TOF	N	N	PFS - progression / relapse / death from any cause	Median: 50 months	N	High
Kim [30]	N	N	Siemens Biograph 6	N	N	EFS - relapse / progression / stopping of treatment / death from any cause OS - death from any cause	Median: 27.8 months	N	Moderate

Study	Prospective	Multi-Centre	PET Scanners used	EANM guidelines stated	SUV Defined	Definition of prognostic factor provide	Follow up period	Separate Validation Cohort	Overall risk of bias
Kim [31]	N	N	Philips Gemini Siemens Biograph 40	N	N	PFS - progression / relapse / death OS - death (? any cause)	Median: 25.8 months	N	Moderate
Kwon [32]	N	Y	GE Discovery ST	N	Y(bw)	PFS - progression / relapse / death from any cause OS - death from any cause	Median: 30.8 months	N	High
Lanic [33]	N	Y	Siemens Biograph LSO Sensation 16	N	N	PFS - progression / relapse / death from any cause OS - death from any cause	Median: 28 months	N	High
Lue [34]	N	N	GE Discovery ST	N	N	PFS - progression / relapse / death from any cause OS - death from any cause	Median: 48 months	N	Moderate
Mettler [35]	N	Y	Multiple not defined	N	N	PFS - progression / relapse / death from any cause OS - death from any cause	Not defined	N	High
Mikhaeel [36]	N	N	GE Discovery ST GE Discovery VCT	N	N	PFS - progression / death from any cause OS - death	Median: 3.8 years	N	Moderate
Milgrom [37]	N	N	GE Discovery ST GE Discovery RX GE Discovery STE	N	N	Relapse or progression or death	Not defined	Y	High
Miyazaki [38]	N	N	GE Discovery STE	N	N	PFS - relapse / death from any cause OS - death	Median: 32.7 months	N	Moderate
Park [39]	N	N	GE Discovery LS, GE Discovery STE	N	N	PFS - progression / relapse / death from any cause OS - death from any cause	Median: 21 months	N	High
Sasanelli [40]	N	Y	Philips Gemini GXL Siemens Biograph 2 GE Discovery ST	N	Y(bw)	PFS - relapse OS - death from any cause	Median: 39 months	N	Moderate

Study	Prospective	Multi-Centre	PET Scanners used	EANM guidelines stated	SUV Defined	Definition of prognostic factor provide	Follow up period	Separate Validation Cohort	Overall risk of bias
Senjo [41]	N	Y	Philips Gemini GXL GE Discovery ST	Y	Y(bw)	PFS - progression / relapse / death OS - death	Median: 33.1 months, 32.8 months	Y	High
Song [42]	N	Y	Siemens Biograph	N	N	PFS progression OS - death from any cause	Median: 40.8 months	Y	Moderate
Song [43]	N	Y	Siemens Biograph	N	N	Not defined	Median: 45.8 months	N	Moderate
Song [44]	N	Y	Siemens Biograph	N	N	PFS progression, death related to lymphoma OS - death from any cause	Median: 36 months	N	Moderate
Toledano [45]	N	N	Siemens Biograph Sensation 16 HiRes	N	N	OS and PFS were defined according to the revised NCI criteria	Median: 40 months	N	Moderate
Tseng [46]	N	N	GE Discovery LS	N	N	Not defined	Median: 50 months	N	High
Xie [47]	N	N	Siemens Biograph 64	Y	N	PFS - progression / relapse / death from any cause	Median: 17 months	N	High
Zhang [48]	N	N	Siemens Biograph 64	N	N	PFS - progression, death related to lymphoma	Median: 34 months	N	Moderate
Zhou [49]	N	N	GE Discovery ST Siemens Biograph 64	N	N	N - not defined	Median: 30 months	N	Moderate

Table 2.1 Overview of study design and risk of bias for each of the studies included in the systematic review. PFS progressive free survival; EFS event free survival; OS overall survival; FFP free from progression; bw body weight; 1 Discovery 690, STE, ST, RX, 600, 710, LS, Biograph HiRez, Truepoint, mCT, LSO, BGO and Gemini TF and GXL; EANM European Association of Nuclear Medicine

As there were no studies deemed to be of low risk for overall bias, a decision was made to include the high risk studies in the systematic review, as removal of these would introduce its own inherent bias.

2.4.2 Metabolic Parameters

SUV is the commonest metric extracted from PET studies. This represents a ratio of radioactivity at a given image location compared to injected whole-body radioactivity [50]. There are several iterations of SUV, including the maximum or mean SUV within a contoured area (SUVmax and SUVmean), or SUVpeak which is the average SUV of a region of interest centred on the highest uptake region within the contoured area. SUV supports other metabolic parameters such as metabolic tumour volume (MTV), which is the volume of disease contoured at a specified SUV threshold, and total lesion glycolysis (TLG), which is the MTV multiplied by SUVmean. Published evidence regarding metabolic parameters used in the pretreatment assessment of lymphoma is summarised below.

2.4.2.1 SUV metrics for prediction of outcome

2.4.2.2 a) DLBCL

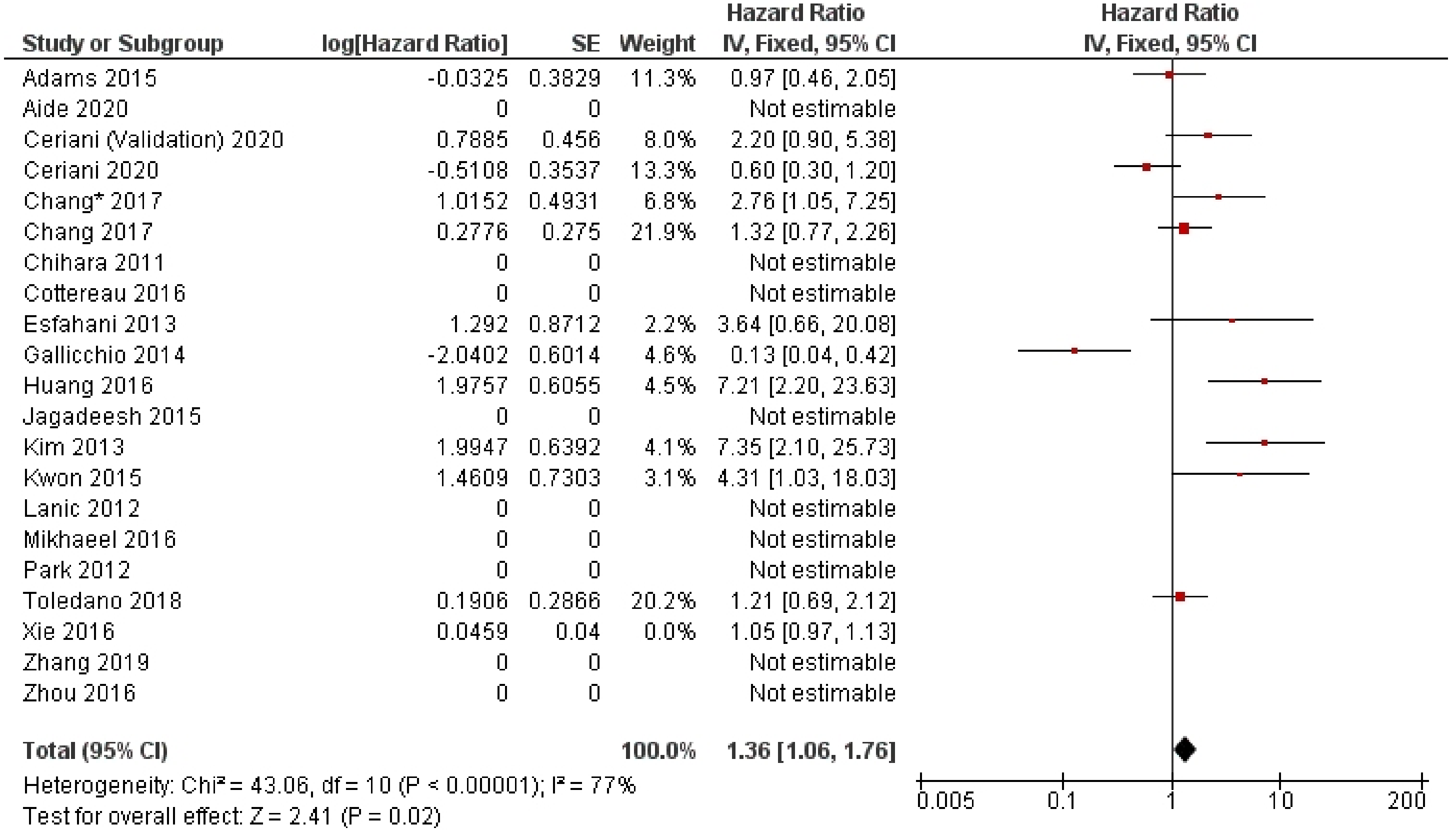
The majority of studies assessing the use of baseline SUVmax in DLBCL report no significant ability to predict progression-free survival (PFS) or overall survival (OS) (Table 2.2). Forest plots illustrating hazard ratios (HR) for PFS and OS are demonstrated in Figures 2.2 and 2.3. From the results included in the forest, the overall HR was 1.35 (CI 95% 1.06–1.76) for PFS and 1.52 (CI 95% 1.15–2.02). However, there is considerable heterogeneity specifically in the PFS analysis (I2 = 77%) and reporting bias is present because a number of studies which did not report any significance did not provide the results required to calculate a HR.

First Author	Year	Type	Study Type	Patient No.	Stage		Treatment	Events (Follow up cut off)	SUV Type	Cut-off Value	Predictive Univariate Analysis HR (95% CI)		Predictive Multivariate Analysis HR (95% CI)	
					I/II	III/IV					PFS	OS	PFS	OS
Aide [10]	2020	DLBCL	R	132 (80:20 Training / Validation)	NR	NR	R-CHOP, R-ACVBP	Relapse/death: 102 (2-year)	SUVmax	32.21	NS	NR	NR	NR
Albano [13]*	2020	HL (Aged 65-92)	R	123	36	87	ABVD, BEACOPP, R-CHOP, +/- RT, RT	Relapse: 51 Died: 37 (No defined cut-off)	L-L SUV R	9.3	0.447 (0.237–0.748)	0.526 (0.261–0.992)	0.228 (0.049–0.765)	0.200 (0.033–0.353)
									L-BP SUV R	6.4	0.469 (0.229–0.774)	0.523 (0.241–0.983)	0.354 (0.069–0.989)	0.555 (0.201–1.002)
Ceriani [16]	2020	DLBCL	R	141 – Testing	61	80	R-CHOP +/- RT	NR	Max	20	NS	NS	NR	NR
				113 -Validation	49	64	R-CHOP +/- RT	NR	Max	31	NS	NS	NR	NR
Zhang [48]	2019	DLBCL	R	85	32	53	R-CHOP/R-CHOP like	Relapse/Died:23 (3-year)	Max	NR – AUC 0.573	NR	NR	NR	NR
Akhtari [12]	2018	HL	R	267	205	62	ABVD +/- RT/ other*	Relapsed / refractory: 27 (5-year)	Max	NR	NS	NR	NR	NR
Cottereau [21]	2018	HL		258	258	0	ABVD +/- RT	PFS: 27 events OS: 12 (5-year)	Max	NR	NS	NS	NR	NR
Toledano [45]	2018	DLBCL	R	114	26	88	R-CHOP/R-CHOP like	Relapse: 52 Died: 43 (5-year)	Max	NR	NS	NS	NR	NR
Angelopoulou [14]	2017	HL	R	162	76	86	ABVD +/- BEACOPP, +/- RT	PFS: 81OS: 93(5-year)	Max	<9, 9-18, >18 93%, 81%, 58%	NR	NR	NR	
Chang [17]	2017	DLBCL	R	118	48	70	R-CHOP	Relapse: 55 Died: 49 (5-year)	Max	18.8	NS	NS	NR	NR
Chang [18]	2017	DLBCL	R	70	35	35	R-CHOP	NR	Tumour Max	19	2.76 (1.05–7.61)	NS	3.27 (1.11–9.60)	NS
									Sternal Max	1.6	NS	2.34 (1.01–5.44)	NS	2.62 (1.10–6.28)

First Author	Year	Type	Study Type	Patient No.	Stage		Treatment	Events (Follow up cut off)	SUV Type	Cut-off Value	Predictive Univariate Analysis HR (95% CI)		Predictive Multivariate Analysis HR (95% CI)	
					I/II	III/IV					PFS	OS	PFS	OS
Cottereau [22]	2016	DLBCL	R	81	16	65	R-CHOP, R-ACVBP	Relapse: 34 (5-year)	Max	NR	NS	NS	NR	NR
Huang [26]	2016	DLBCL	R	140	62	78	R-CHOP/CHOP	PFS: 73.8OS: 86.1(30 month)	Max	9	7.2 (2.201–23.631)	11.4 (1.514–86.350 0.018)	4.7 (1.429–16.022 0.011)	NS
Mikhaeel [36]	2016	DLBCL	R	147	46	101	R-CHOP	PFS: 65.4OS: 73.7(5-year)	Max	Split into tertiles	NS	NS	NR	NR
Xie [47]	2016	DLBCL	R	60	12	48	R-CHOP	Relapse: 17 Died: 3 (40 month)	Max	NR	NS	NR	NS	NR
Zhou [49]	2016	DLBCL	R	91	34	57	R-CHOP	Relapse: 37 Died: 11 (5-years)	Max	PFS – 19 OS – 15.8	NS	NS	NR	NR
Adams [9]	2015	DLBCL	R	73	11	62	R-CHOP	Relapse: 27 Death: 24 (No defined cut-off)	Max	NR	NS	NS	NR	NR
Jagadeesh [28]	2015	DLBCL	R	89	0	89	R-CHOP/R + other	LR: 50(5-year)	Max	15	NS for LR	NR	NS for LR	NR
Kwon [32]	2015	DLBCL	R	92	54	38	R-CHOP	Relapse: 33 Died: 3 (No defined cut-off)	Max	10.5	4.31 (1.03-18.1)	NR	NS	NR
Gallicchio [25]	2014	DLBCL	52	26	26	R-CHOP, R-COMP	Relapse: 15 Died: 2 (18 month)	Max	13.5	0.13 (0.04–0.46)	NR	NR	NR	
Esfahani [24]	2013	DLBCL	R	20	8	12	R-CHOP	Relapse: 6 (No defined cut-off)	Max	13.84	NS	NR	NR	NR
									Mean	6.44	NS	NR	NR	NR
Kim [31]	2013	DLBCL	R	140	77	63	R-CHOP	Relapse: 21 Died: 16 (2-year)	Max	16.4	NS	NS	NR	NR
Lanic [33]	2012	DLBCL	R	57	NR	NR	R-CHOP, intensified R-CHOP	NR (2-year)	Max	NR	NS	NS	NR	NR

First Author	Year	Type	Study Type	Patient No.	Stage		Treatment	Events (Follow up cut off)	SUV Type	Cut-off Value	Predictive Univariate Analysis HR (95% CI)		Predictive Multivariate Analysis HR (95% CI)	
					I/II	III/IV					PFS	OS	PFS	OS
Park [39]	2012	DLBCL	R	100	55	45	R-CHOP	NR	Max	NR	NS	NS	NR	NR
									Sum	NR	1.011 (1.002–1.020)	1.016 (1.006–1.026)	NR	NR
Tseng [46]	2012	HL	R	30	11	19	Standford V, ABVD, VAMP, BEACOPP	Relapse =6 (4-year)	Max	NR	NS	NS	NR	NR
									Mean	NR	NS	NS	NR	NR
Chihara [19]	2011	DLBCL	R	110	65	45	R-CHOP +/- RT	PFS: 75% OS: 84% (3-year)	Max	30	Sig.	Sig.	HR6.74	NS

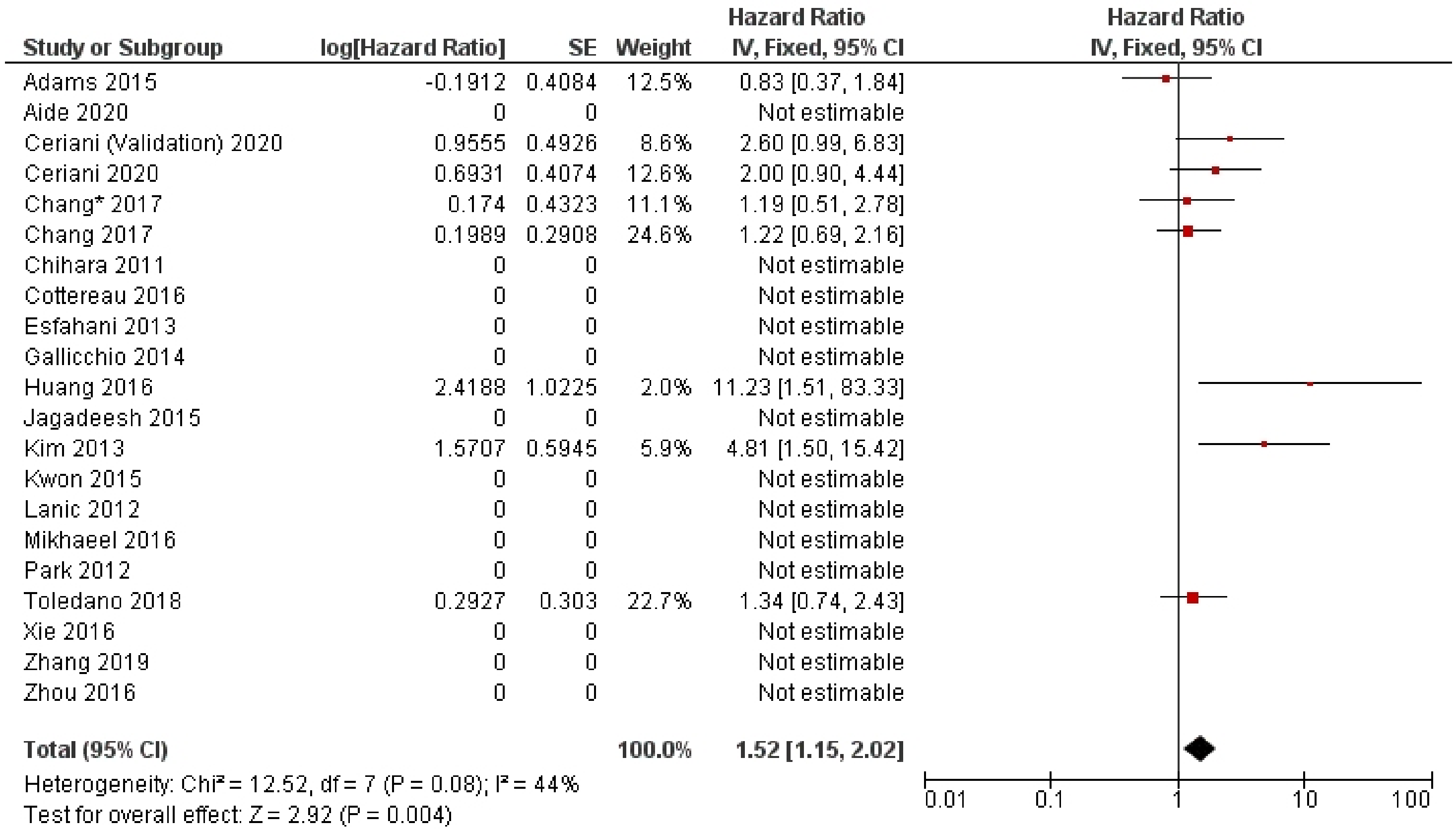
Table 2.2 Studies assessing the use of standardised uptake value (SUV) in predicting outcomes in diffuse large B-cell lymphoma (DLBCL) and Hodgkin lymphoma (HL). R = retrospective, NR = not reported, NS = not significant, Sig. = significant, HR= hazard ratio, CI = confidence interval, PFS = progressive free survival, OS = overall survival, R-CHOP = rituximab cyclophosphamide, doxorubicin hydrochloride, vincristine (Oncovin) and prednisolone, R-ACVBP - Rituximab, Doxorubicin, Cyclophosphamide, Vindesine, Bleomycin, prednisolone, R-COMP = prednisolone, Cyclophosphamide, Vincristine, Myocet and Rituximab, RT = radiotherapy, ABVD = doxorubicin (Adriamycin), bleomycin, vinblastine and dacarbazine, eBEACOPP = escalated dose bleomycin, etoposide, doxorubicin (Adriamycin), cyclophosphamide, vincristine (Oncovin), procarbazine, and prednisone, VAMP =.vincristine, doxorubicin hydrochloride, methotrexate, prednisolone. *The HRs presented as presented in the study but are inverse to the other HRs within the table.



Study or Subgroup	log[Hazard Ratio]	SE	Weight	Hazard Ratio IV, Fixed, 95% CI
Adams 2015	-0.0325	0.3829	11.3%	0.97 [0.46, 2.05]
Aide 2020	0	0		Not estimable
Ceriani (Validation) 2020	0.7885	0.456	8.0%	2.20 [0.90, 5.38]
Ceriani 2020	-0.5108	0.3537	13.3%	0.60 [0.30, 1.20]
Chang* 2017	1.0152	0.4931	6.8%	2.76 [1.05, 7.25]
Chang 2017	0.2776	0.275	21.9%	1.32 [0.77, 2.26]
Chihara 2011	0	0		Not estimable
Cottereau 2016	0	0		Not estimable
Esfahani 2013	1.292	0.8712	2.2%	3.64 [0.66, 20.08]
Gallicchio 2014	-2.0402	0.6014	4.6%	0.13 [0.04, 0.42]
Huang 2016	1.9757	0.6055	4.5%	7.21 [2.20, 23.63]
Jagadeesh 2015	0	0		Not estimable
Kim 2013	1.9947	0.6392	4.1%	7.35 [2.10, 25.73]
Kwon 2015	1.4609	0.7303	3.1%	4.31 [1.03, 18.03]
Lanic 2012	0	0		Not estimable
Mikhaeel 2016	0	0		Not estimable
Park 2012	0	0		Not estimable
Toledano 2018	0.1906	0.2866	20.2%	1.21 [0.69, 2.12]
Xie 2016	0.0459	0.04	0.0%	1.05 [0.97, 1.13]
Zhang 2019	0	0		Not estimable
Zhou 2016	0	0		Not estimable
Total (95% CI)			**100.0%**	**1.36 [1.06, 1.76]**

Heterogeneity: Chi² = 43.06, df = 10 (P < 0.00001); I² = 77%
Test for overall effect: Z = 2.41 (P = 0.02)

Figure 2.2 Forest plot demonstrating hazard ratios for progression free/event free survival for patients with DLBCL using a dichotomous cut-off value derived from SUVmax. Studies which do not provide hazard ratios are included but no estimate is given.



Study or Subgroup	log[Hazard Ratio]	SE	Weight	Hazard Ratio IV, Fixed, 95% CI
Adams 2015	-0.1912	0.4084	12.5%	0.83 [0.37, 1.84]
Aide 2020	0	0		Not estimable
Ceriani (Validation) 2020	0.9555	0.4926	8.6%	2.60 [0.99, 6.83]
Ceriani 2020	0.6931	0.4074	12.6%	2.00 [0.90, 4.44]
Chang* 2017	0.174	0.4323	11.1%	1.19 [0.51, 2.78]
Chang 2017	0.1989	0.2908	24.6%	1.22 [0.69, 2.16]
Chihara 2011	0	0		Not estimable
Cottereau 2016	0	0		Not estimable
Esfahani 2013	0	0		Not estimable
Gallicchio 2014	0	0		Not estimable
Huang 2016	2.4188	1.0225	2.0%	11.23 [1.51, 83.33]
Jagadeesh 2015	0	0		Not estimable
Kim 2013	1.5707	0.5945	5.9%	4.81 [1.50, 15.42]
Kwon 2015	0	0		Not estimable
Lanic 2012	0	0		Not estimable
Mikhaeel 2016	0	0		Not estimable
Park 2012	0	0		Not estimable
Toledano 2018	0.2927	0.303	22.7%	1.34 [0.74, 2.43]
Xie 2016	0	0		Not estimable
Zhang 2019	0	0		Not estimable
Zhou 2016	0	0		Not estimable
Total (95% CI)			**100.0%**	**1.52 [1.15, 2.02]**

Heterogeneity: Chi² = 12.52, df = 7 (P = 0.08); I² = 44%
Test for overall effect: Z = 2.92 (P = 0.004)

Figure 2.3 Forest plot demonstrating hazard ratios for overall survival for patients with DLBCL using a dichotomous cut-off value derived from the SUVmax. Studies which do not provide hazard ratios are included but no estimate is given.

Of the studies which showed a prognostic ability for SUVmax, Gallicchio *et al.* reported this was the only imaging parameter able to predict PFS when compared to TLG and MTV in a small study of 52 DLBCL patients (26 early and 26 advanced stage) with a higher SUVmax associated with a longer PFS, the hazard ratio (HR) was 0.13 (0.04–0.46) [25]. A study by Kwon *et al.* assessing 92 DLBCL (54 stage I/II, 38 stage II/IV) patients reported that a SUVmax of 10.5 was significant in predicting PFS, but this was not an independent prognostic predictor at multivariate analysis with clinical factors such as age, Lactate Dehydrogenase (LDH) level, stage, IPI score or Eastern Cooperative Oncology Group (ECOG) status [32]. Conversely, Miyazaki *et al.* demonstrated that SUVmax was an independent predictor of 3-year PFS and R-IPI [38]. Chang *et al.* found that tumour SUVmax $>$19 was a significant predictor of 3-year PFS, whereas the SUVmax of sternal uptake was an independent predictor of 3-year OS in a study of 70 DLBCL patients [18]. The most extensive study evaluating SUVmax as a predictor of PFS and OS was performed by Ceriani *et al.* with a test cohort of 141 patients and a validation cohort of 113 patients, both containing a similar mix of stage and prognostic scores. SUVmax was not significant in predicting PFS or OS in either cohort [16].

2.4.2.3 b) HL

Five studies have assessed the use of SUVmax as a predictive parameter in HL patients with only one reporting significance (Table 2.2). The largest by Akharti *et al.* showed no significant ability of SUVmax to predict PFS and OS in 267 stage I and II HL patients (74 early favourable) [12]. These findings were concordant with a study by Cottereau *et al.*, who also found no significant ability of SUVmax to predict PFS or OS in 258 stage I and II patients. Angelopoulou *et al.* reported that SUVmax was a significant predictor of 5-year PFS in a study of 162 patients with a split of stages (stage I/II $=$ 76, stage III/IV $=$ 86) [14]. The cohort was stratified into three risk groups, SUVmax $<$9, 9–18 and $>$ 18 with five-year PFS rate being 93%, 81% and 58% respectively, multivariate analysis was not performed. Albano *et al.* studied the prognostic ability of liver to lesion SUV ratio and blood pool to lesion ratio in 123 older (age $>$ 65 years) HL patients [13]. They found that both parameters were significant (at univariate analysis) for PFS and OS. They also demonstrated these metrics to be independent prognostic markers when analysed with tumour stage, German Hodgkin Study Group (GHSG) risk group, MTV and TLG for PFS, and tumour stage, GHSG risk group and Deauville score for OS.

Factors affecting SUV such as scanner spatial resolution, image acquisition and PET reconstruction parameters combined with a relatively small number of events, variation in the number of early and advanced patients, differences in treatment and definition of PFS all influence the results [51, 52]. This is reflected by the variation in cut-off/threshold values used to risk-stratify patients within each of the studies.

2.4.3 Metabolic Tumour Volume and Total Lesion Glycolysis for prediction of outcome

2.4.3.1 a) DLBCL

The potential utility of baseline MTV and TLG for predicting PFS and OS in patients with DLCBL, has been reported in multiple studies (Table 2.3, Figures 2.4 and 2.5). However, similar to SUVmax, there is heterogeneity in the cut-off values used which has led to variability in the reported survival rates between groups. Overall, the HR for MTV in PFS was 3.47 (CI 95% 2.80 – 4.30) and 4.20 (CI 95% 2.80 – 4.30) for OS. Again reporting bias is present because a number of studies which did not report any significance did not provide the results required to calculate a HR.

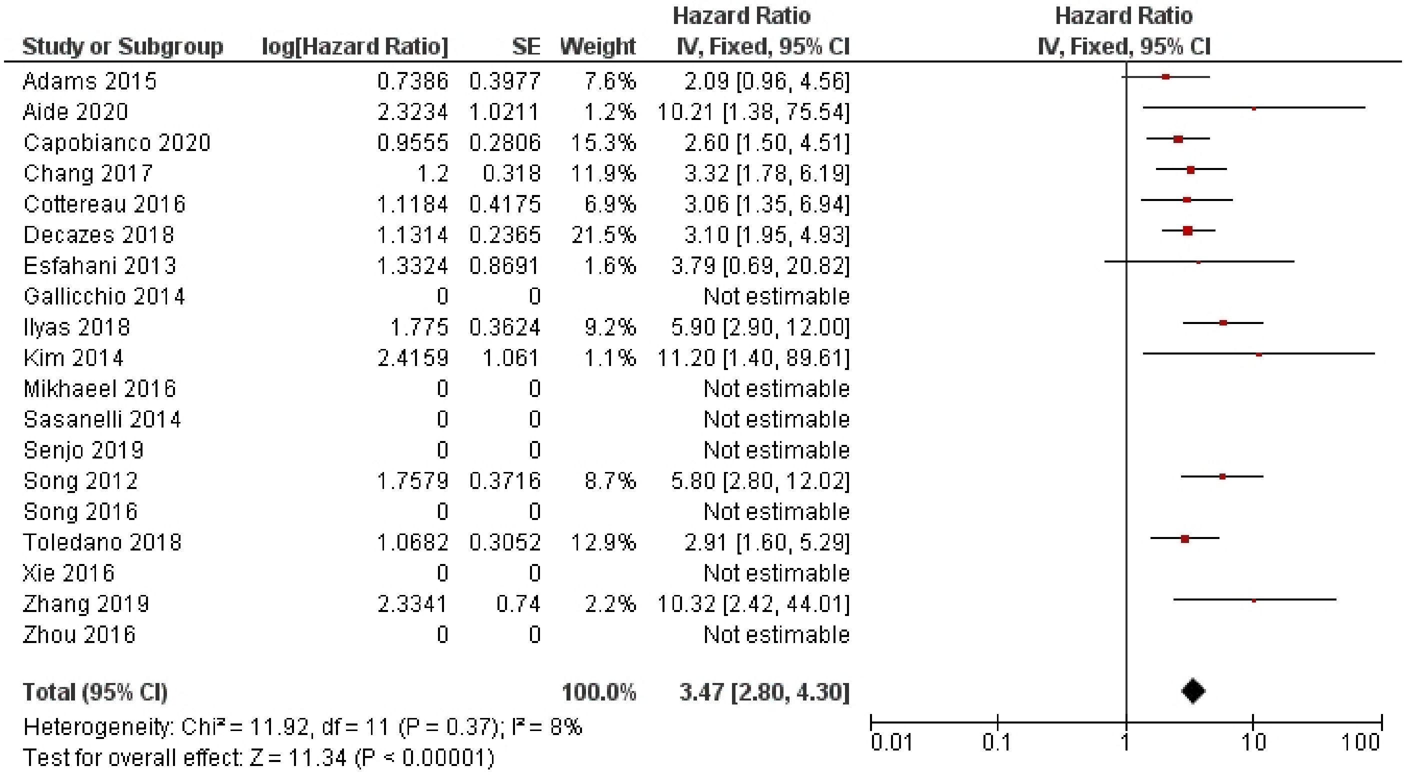
First Author	Year	Patient No.	Stage		Treatment	Events (Follow up cut off)	Segmentation Threshold	MTV / TLG	Suggested Cut-Off	Predictive Univariate Analysis HR (95% CI)		Predictive Multivariate Analysis HR (95% CI)	
			I/II	III/IV						PFS	OS	PFS	OS
Aide [10]	2020	132	NR	NR	R-CHOP, R-ACVBP	Relapse/death: 102 (2-year)	SUVmax of liver	MTV	111ml	10.2 (1.4-75.5) (training data set) NS (Validation dataset)	NR	NS aa-IPI LZHGE	NR
Capobianco [15]	2020	280	26	264	R-CHOP	Relapse: 86 Died: 51 (4-year)	41% SUVmax	MTV	242ml	NR	3.7 (1.9 – 7.2)	NR	NR
							CNN segmentation	MTV	110ml	NR	2.8 (1.6 – 5.1)	NR	NR
Decazes [23]	2018	215	51	164	R-CHOP, R-CHOP like, R-ACVBP	Relapse: 92 Died: 74 (5-year)	41% SUVmax	MTV	487ml	3.10(1.95-4.95)	4.09(2.32-7.21)	2.20 (1.26-3.83) IPI, chemo-therapy, TVSR	2.78(1.41-5.48) IPI, chemo-therapy, TVSR
Ilyas [27]	2018	147	46	101	R-CHOP	PFS: 65.4% OS: 73.7% (5-year)	PETTRA 2.5	MTV	PFS:396.1ml OS: 457.8ml	5.9 (2.9–12.2)	5.5 (2.4–12.5)	NR	NR
							HERMES 2.5	MTV	PFS:401.4ml OS: 401.4ml	5.9 (2.9–12.2 CI)	5.5 (2.4– 12.5)	NR	NR
							HERMES PERCIST	MTV	PFS:327.4ml OS: 669.8ml	4.8 (2.4–9.5 CI)	3.7 (1.8–7.8)	NR	NR
							HERMES 41%	MTV	PFS:165.7ml OS: 189.3ml	4.2 (2.2–7.9 CI)	3.5 (1.8–7.0)	NR	NR
Senjo [41]	2019	150 (combined training and validation)	66	84	R-CHOP, R-THP-COP, R-CVP	Relapse 21 Died 48 (5-year)	>4.0 SUV	MTV	150ml	NR	NR	2.49 (1.57-3.94)	2.75 (1.72-4.38)
Zhang [48]	2019	85	32	53	R-CHOP/ R-CHOP-like	Relapse: 23 Died: 6 (3-years)		MTV	80.74ml	10.32 (2.42 – 44.08)	NR	NR Correlated with TLG	NR
								TLG	1036.61g	10.39 (2.43-44.39)	NR	10.42, (2.35-46.30)	NR
Toledano [45]	2018	114	26	88	R-CHOP/ R-CHOP like	Relapse: 52 Died: 43 (5-year)	41% SUVmax	MTV	261.4ml	2.91 (1.60-5.29)	4.32 (2.07-8.99)	2.05 (HR 1.02-4.15) GEP, IPI	2.70 (1.16-6.33) GEP, IPI

First Author	Year	Patient No.	Stage		Treatment	Events (Follow up cut off)	Segmentation Threshold	MTV / TLG	Suggested Cut-Off	Predictive Univariate Analysis HR (95% CI)		Predictive Multivariate Analysis HR (95% CI)	
			I/II	III/IV						PFS	OS	PFS	OS
								TLG	1325.8g	NS MC	4.82 (2.67-8.71)	NR	NR
Chang [17]	2017	118	48	70	R-CHOP	Relapse: 55 Died: 49 (5-year)	2.5 SUV	MTV	165.4ml	3.32 (1.78-6.20)	4.05 (2.07-7.95)	2.31 (1.16 – 4.60) IPI	2.38 (1.12-5.04) Age, IPI
								TLG	1204.9ml	2.57 (1.43-4.61)	2.96 (1.61-5.45)	NR	NR
Cottereau [22]	2016	81	16	65	R-CHOP, R-ACVBP	Relapse: 34 (5-year)	41% SUVmax	MTV	300ml	3.06 (1.43–6.54)	3.01 (1.35–6.70)	1.61 (0.70–3.69)	3.0 (1.35–6.70)
								TLG	3904g	2.92 (1.45-5.90)	2.39 (1.16–4.92)	NS	NS
Song [42]	2016	107*	107	R-CHOP	NR	2.5 SUV	MTV	601.2ml	Sig.	Sig.	5.21 (2.54–10.69) IPI, bulky disease, BMI, IM MTV, CAs	5.33 (2.60–10.90) IPI, bulky disease, BMI, IM MTV, CAs	
								IM MTV	260.5ml	Significant	Significant	NS	NS
Zhou [49]	2016	91	34	57	R-CHOP	Relapse: 37 Died: 11 (5-year)	SUVmean of liver + 3 SD	MTV	PFS: 70ml OS: 78ml	88% vs 37%	98% vs 60%	NS	NS
								TLG	PFS: 826.5g OS: 726g	83% vs 34%	92% vs 67%	5.21 (2.21-12.28) MTV, NCCN-IPI, Stage, B symptoms, LDH level Ki-67	9.1 (1.83 – 45.64) MTV, NCCN-IPI, Stage, B symptoms, LDH level Ki-67
Mikhaeel [36]	2016	147	46	101	R-CHOP	PFS5: 65.4% OS5: 73.7% (5-year)	41% SUVmax	MTV	Terties	Upper: 5.81 (2.38-14.14) Middle: 3.77 (1.49-9.51)	Sig.	Upper: 3.46 (1.10-10.86) Middle: 2.73 (0.89-8.40)	NR

First Author	Year	Patient No.	Stage		Treatment	Events (Follow up cut off)	Segmentation Threshold	MTV / TLG	Suggested Cut-Off	Predictive Univariate Analysis HR (95% CI)		Predictive Multivariate Analysis HR (95% CI)	
			I/II	III/IV						PFS	OS	PFS	OS
Xie [47]	2016	60	12	48	R-CHOP	Relapse: 17 Died: 3 (40 months)	SUVmean of liver + 2SD	TLG	Tertiles	Upper: 4.90 (2.11- 11.38) Middle: (2.96 1.24 7.10)	Sig.	NR	NR
								MTV	Continuous	1.030 (1.017–1.044)	NR	1.028 (1.014–1.043) NCCN-IPI	NR
								TLG	Continuous	1.078 (1.042–1.116)	NR	1.071 (1.032–1.112) NCCN-IPI	NR
Adams [9]	2015	73	11	62	R-CHOP	Relapse: 27 Death: 24 (No defined cut-off)	40% SUVmax	MTV	445ml	NS	2.40 (1.03-5.60)	NR	NS NCCN-IPI
								TLG	4897.5g	NS	NS	NR	NR
Kim [30]	2014	96	49	47	R-CHOP	PFS3: 69.5OS3: 72.9(No defined cut-off)	2.5 SUV	MTV	130.7ml	11.2 (1.4-88.1)	NR	10.4 (1.3-83.4) IPI >/equal to 3	NS with IPI as individual parameters NR
Gallicchio [25]	2014	52	41	11	R-CHOP like	Relapse: 15 Death: 2 (18 months)	42% SUVmax	MTV	16.1ml	NS	NR	NR	NR
								TLG	589.5g	NS	NR	NR	NR
Sasanelli [40]	2014	114	20	94	R-CHOP/ R-ACVBP	Relapse: 31 Died: 25 (3-year)	41% SUVmax	MTV	550ml	77% vs 60% 87% vs 60%/ 59% vs 78% vs 84% vs 93%	NS	4.70 (1.82-12.18) Stage, LDH, Bulky disease	4.11 (1.67-10.16) aa-IPI, bulky disease
								TLG	4576g	NS	64% vs 85%	NR	NR
Esfahani [24]	2013	20	8	12	R-CHOP	Relapse: 6 Died: 0 (No defined cut-off)	50% SUVmax	MTV	379.2ml	NS	N/A	NR	N/A
								TLG	704.8g	11.21 (1.29-97)	N/A	NR	N/A
Kim [31]	2013	140	77	63	R-CHOP	Relapse: 21 Died: 16 (2-year)	25%, 50% and 75% SUVmax	TLG25	817.8g	2.8 (1.1-7.1)	NS	NR	NR
								TLG50	415.5g	3.6 (1.3-10.0)	3.3 (1.0-10.0)	3.6 (1.3-10.0) IPI (2 splits)	3.1 (1.0-9.6) IPI(2 splits)
								TLG75	102.0g	3.5 (1.3-9.5)	NS	NR	NR

First Author	**Year**	**Patient No.**	Stage		**Treatment**	**Events (Follow up cut off)**	**Segmentation Threshold**	**MTV / TLG**	**Suggested Cut-Off**	**Predictive Univariate Analysis HR (95% CI)**		**Predictive Multivariate Analysis HR (95% CI)**	
			I/II	III/IV						PFS	OS	PFS	OS
Park [39]	2012	100	55	45	R-CHOP	NR	Blood Pool threshold	TLG	NR	NS	NR	NR	NR
Song [44]	2012	169	100	69	R-CHOP	PFS: 73.4OS: 76.3 (3-year)	2.5 SUV	MTV	220ml	5.80 (2.79–12.06)	8.10 (3.40–19.31)	5.30 (2.51–11.16) Stage 3	7.01 (2.90–16.93) Stage 3

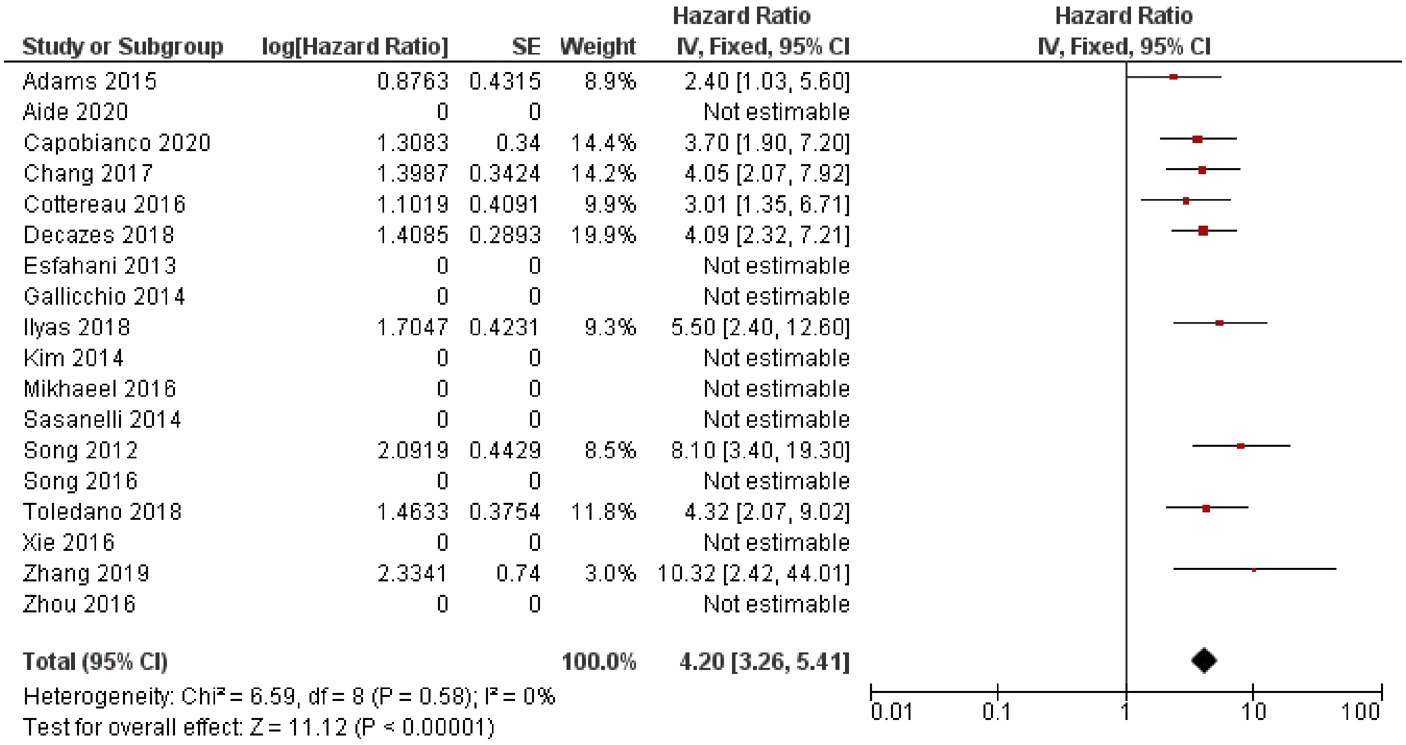
Table 2.3 Studies assessing the use of metabolic tumour volume (MTV) and total lesion glycolysis (TLG) in predicting outcomes in diffuse large B-cell lymphoma (DLBCL). NR = not reported, NS = not significant, Sig. = significant, HR= hazard ratio, CI = confidence interval, PFS = progressive free survival, OS = overall survival, R-CHOP = rituximab cyclophosphamide, doxorubicin hydrochloride, vincristine (Oncovin) and prednisolone, R-ACVBP - Rituximab, Doxorubicin, Cyclophosphamide, Vindesine, Bleomycin, prednisolone, R-THP-COP = rituximab, pirarubicin, cyclophosphamide, vincristine, and prednisolone, R- CVP = rituximab, cyclophosphamide, vincristine, prednisolone, BMI = bone marrow involvement, aa-IPI= age-adjusted International Prognostic Index, NCCN-IPI = National Comprehensive Cancer Network – International Prognostic Index, IM = intramedullary, CAs = cytogenetic abnormalities, LZHGE = Long-Zone High Grey-level Emphasis.



Study or Subgroup	log[Hazard Ratio]	SE	Weight	Hazard Ratio IV, Fixed, 95% CI
Adams 2015	0.7386	0.3977	7.6%	2.09 [0.96, 4.56]
Aide 2020	2.3234	1.0211	1.2%	10.21 [1.38, 75.54]
Capobianco 2020	0.9555	0.2806	15.3%	2.60 [1.50, 4.51]
Chang 2017	1.2	0.318	11.9%	3.32 [1.78, 6.19]
Cottereau 2016	1.1184	0.4175	6.9%	3.06 [1.35, 6.94]
Decazes 2018	1.1314	0.2365	21.5%	3.10 [1.95, 4.93]
Esfahani 2013	1.3324	0.8691	1.6%	3.79 [0.69, 20.82]
Gallicchio 2014	0	0		Not estimable
Ilyas 2018	1.775	0.3624	9.2%	5.90 [2.90, 12.00]
Kim 2014	2.4159	1.061	1.1%	11.20 [1.40, 89.61]
Mikhaeel 2016	0	0		Not estimable
Sasanelli 2014	0	0		Not estimable
Senjo 2019	0	0		Not estimable
Song 2012	1.7579	0.3716	8.7%	5.80 [2.80, 12.02]
Song 2016	0	0		Not estimable
Toledano 2018	1.0682	0.3052	12.9%	2.91 [1.60, 5.29]
Xie 2016	0	0		Not estimable
Zhang 2019	2.3341	0.74	2.2%	10.32 [2.42, 44.01]
Zhou 2016	0	0		Not estimable
Total (95% CI)			**100.0%**	**3.47 [2.80, 4.30]**

Heterogeneity: Chi² = 11.92, df = 11 (P = 0.37); I² = 8%
Test for overall effect: Z = 11.34 (P < 0.00001)

Figure 2.4 Forest plot demonstrating hazard ratios for progression free survival for patients with DLBCL using a dichotomous cut-off value derived from the metabolic tumour volume. Studies which do not provide hazard ratios are included but no estimate is given.



Study or Subgroup	log[Hazard Ratio]	SE	Weight	Hazard Ratio IV, Fixed, 95% CI
Adams 2015	0.8763	0.4315	8.9%	2.40 [1.03, 5.60]
Aide 2020	0	0		Not estimable
Capobianco 2020	1.3083	0.34	14.4%	3.70 [1.90, 7.20]
Chang 2017	1.3987	0.3424	14.2%	4.05 [2.07, 7.92]
Cottereau 2016	1.1019	0.4091	9.9%	3.01 [1.35, 6.71]
Decazes 2018	1.4085	0.2893	19.9%	4.09 [2.32, 7.21]
Esfahani 2013	0	0		Not estimable
Gallicchio 2014	0	0		Not estimable
Ilyas 2018	1.7047	0.4231	9.3%	5.50 [2.40, 12.60]
Kim 2014	0	0		Not estimable
Mikhaeel 2016	0	0		Not estimable
Sasanelli 2014	0	0		Not estimable
Song 2012	2.0919	0.4429	8.5%	8.10 [3.40, 19.30]
Song 2016	0	0		Not estimable
Toledano 2018	1.4633	0.3754	11.8%	4.32 [2.07, 9.02]
Xie 2016	0	0		Not estimable
Zhang 2019	2.3341	0.74	3.0%	10.32 [2.42, 44.01]
Zhou 2016	0	0		Not estimable
Total (95% CI)			**100.0%**	**4.20 [3.26, 5.41]**

Heterogeneity: $\text{Chi}^2 = 6.59$, df = 8 (P = 0.58); $I^2 = 0\%$

Test for overall effect: Z = 11.12 (P < 0.00001)

Figure 2.5 Forest plot demonstrating hazard ratios for overall survival for patients with DLBCL using a dichotomous cut-off value derived from the metabolic tumour volume. Studies which do not provide hazard ratios are included but no estimate is given.

One of the largest studies by Song *et al.* evaluated 169 patients with DLBCL (stage II and III without extranodal disease) treated with R-CHOP [44]. Patients with an MTV of $<$220cm^3 had significantly better PFS and OS; 89.8 versus 55.6%, and 93.2% versus 58.0%, respectively [44]. MTV was predictive of PFS and OS regardless of stage. MTV remained significant when assessed using multivariate Cox regression with stage III disease, HR = 5.30 (95% 2.51–11.16) and HR = 7.01 (2.90–16.93) for 3-year PFS and 3-year OS, respectively. In another study, Song *et al.* reported that MTV was a prognostic predictor in 107 patients with bone marrow involvement (BMI); patients with an MTV of $>$601.2cm^3 and BMI had worse PFS and OS survival compared to those with a smaller MTV and BMI [42]. Again, this was demonstrated to be an independent predictor when analysed with IPI, bulky disease, BMI, involved marrow MTV and $>$ 2 cytogenetic abnormalities with an HR = 5.21 (95% CI 2.54–10.69) and HR = 5.33 (95% CI 2.60–10.90) for PFS and OS, respectively. However, there was no significant difference in survival between the smaller MTV with BMI group and a comparison cohort of patients without BMI. MTV summarises disease burden; however, it does not account for spread. Cottereau *et al.* studied four different spatial metrics besides TLG and MTV in 95 DLBCL patients on baseline scans to determine if a predictive model could be created [20]. The spatial parameters consisted of Dmax (distance between two of the furthest lesions), Dmax bulk (distance between the largest lesion and furthest lesion away from this), SPREADbulk (sum of all distances between bulky lesions) and SPREAD (sum of all distances between lesions). They found that a model combining MTV and Dmax could significantly distinguish between three prognostic groups. The low-risk group with an MTV $<$394cm^3 and a Dmax $<$58 cm had a 4-year PFS of 94% and OS of 97%, the intermediate group with either an MTV $>$394cm^3 or a Dmax $>$58 cm had a 4-year PFS of 73% and OS of 88% and the high-risk group with a MTV $>$394cm^3 and a Dmax $>$58 cm had a 4-year PFS of 50% and OS of 53%.

Zhou *et al.* reported that although high baseline MTV and TLG were associated with poorer prognosis, only TLG was an independent predictor of PFS and OS in a study of 91 patients [49]. In this study, patients who demonstrated complete or partial remission were more likely to relapse if they had a high baseline TLG (40 versus 9%, p = 0.012). A possible explanation for the discrepancy between the prognostic ability of MTV and TLG in this study may be related to the correlation between MTV and TLG, confounded by relatively small sample sizes. Kim *et al.* evaluated TLG calculated using different MTVs derived using 25, 50 and 75% SUVmax thresholds in a mixed cohort (n = 140) of early and advanced stage DLBCL patients being treated with R-CHOP [31]. They found that all methods for calculating TLG were predictive of 2-year PFS, but only TLG50 was predictive of 2-year OS. Ilyas *et al.* also studied variation in segmentation technique and its potential to impact on predicting outcome in 147 DLBCL patients (46 stage I/II, 101 stage III/IV) all treated with R-CHOP [27]. The four segmentation techniques consisted of a threshold of SUV 2.5 on two software packages (PETTRA and Hermes), 41% SUVmax on Hermes software and an uptake higher than SUVmean of a 3-cm^3 region of interest (ROI) within the right lobe of the liver (PERCIST) using the Hermes software. They found a strong agreement between all four methods, with the lowest intraclass coefficient being between PERCIST and 41% SUVmax thresholds being 0.86. They also reported similar receiver operator curves (ROC) between the four methods with the area under the curve (AUC) ranging from 0.74 to 0.76 for PFS, and 0.71 to 0.75 for OS. All four methods were significant predictors of PFS and OS. However, as stated in the paper, no method is likely to apply to all patients generally. Large heterogeneous masses are likely to be undersized with percentage thresholds, low uptake lesions may be missed using a standard threshold method and disease involving the liver may impede its use as the background value. This may have a more significant impact when further metrics are introduced, such as those based on texture when the size of the contour can also influence the reported values. The segmentation technique of choice also needs to be easily replicated. Recently, Capobianco *et al.* assessed the use of artificial intelligence (AI) using a convolutional neural network (CNN) to segment the MTV [15]. They found that AI-derived MTV correlated with reference MTV derived by two independent readers with a classification accuracy of 85%. Automatic segmentation is a key step required to enable implementation of MTV or TLG into clinical practice.

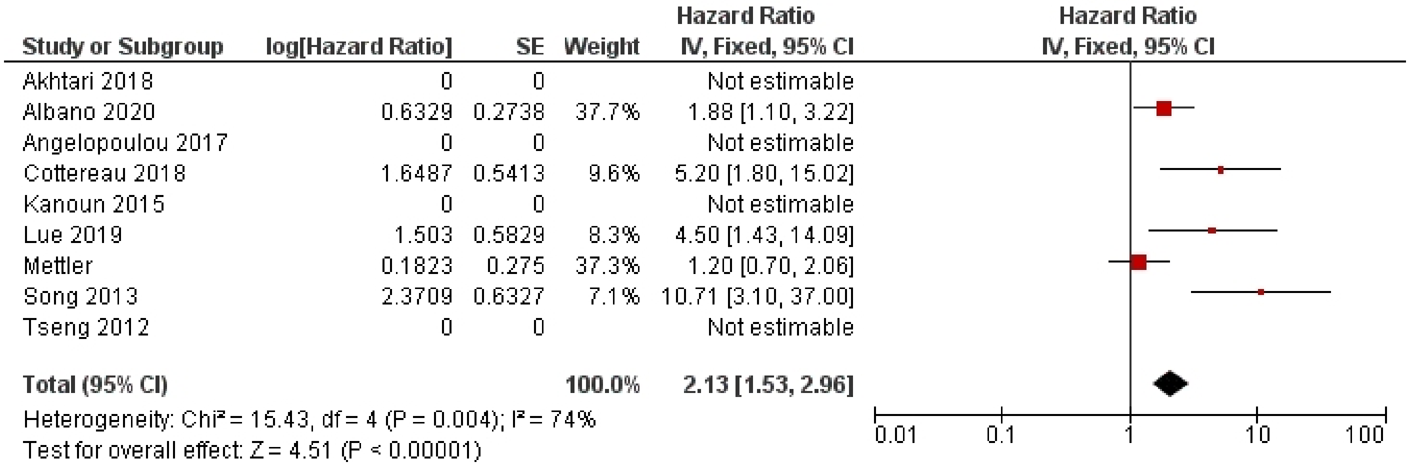
2.4.3.2 b) HL

Fewer studies have investigated the predictive ability of MTV and TLG in HL patients than in DLBCL (Table 2.4, Figures 2.6 and 2.7). This is likely due to the higher survival rate of HL limiting the number of events demonstrated in a single centre and the variation in treatments and scoring systems for a favourable and unfavourable disease, which affect multi-centre studies. The majority of studies involved patients on an adaptive ABVD treatment regime, and results may not be transferrable to patients being treated with an adaptive BEACOPP regime. This confounding issue was highlighted in a study by Mettler *et al.* who assessed the prognostic ability of MTV in 310 patients with advanced HL being treated with eBEACOPP using four different contouring methods involving summation of the volume of each disease site using different defined thresholds: 41% SUVmax of each disease site, a threshold of liver SUVmax, a threshold of liver SUVmean and a fixed threshold of 2.5 SUV [35]. They found that MTV was predictive of interim PET response regardless of segmentation methodology; however, none was able to predict OS and PFS reliably. The divergent findings compared to previous studies are likely related to low event numbers and using a different treatment regime. Albano *et al.* demonstrated the significant ability of both MTV and TLG derived from 41% SUVmax in predicting PFS in both univariate and multivariate analysis in a cohort of 123 elderly patients with a mix of different treatment regimens. However, neither TLG nor MTV were predictive of OS. Cottereau *et al.* and Akhtari *et al.* both assessed the ability of MTV in cohorts of patients consisting of stage I and II disease [12,21]. Cottereau *et al.* found that MTV derived from $>$2.5 SUV was significant in predicting five-year PFS and OS and was significant in multivariate analysis when assessed with different early disease scoring systems. Akhtari *et al.* found that MTV and TLG derived from $>$2.5 SUV thresholding and manual soft tissue contouring were significant predictors of five-year PFS. Reporting bias is present because a number of studies which did not report any significance did not provide the results required to calculate a HR. The overall HR for MTV in PFS was 2.13 (CI 95% 1.53-2.96) and 2.13 (1.43-3.16) in OS. Both were associated with high levels of heterogeneity, I2 $=$ 74% for PFS and I2 $=$ 70% for OS.

First Author	Year	Patient No.	Stage		Treatment	Events (Follow up cut off)	Segmentation Threshold	MTV / TLG	Suggested Cut-Off	Predictive Univariate Analysis HR (95% CI)		Predictive Multivariate Analysis HR (95% CI)	
			I/II	III/IV						PFS	OS	PFS	OS
Albano [13]*	2020	123 Elderly	36	87	ABVD, BEACOPP, R-CHOP, +/- RT, RT	Relapse: 51 Died: 37 (No defined cut-off)	41% SUVmax	MTV	89ml	0.531 (0.294–0.908)	NS	0.555 (0.249–0.965)	NR
								TLG	2199g	0.544 (0.240–0.963)	NS	0.602 (0.111–0.989)	NR
Lue [34]	2019	42	20	22	anthracycline-based chemotherapy +/-RT	Relapse: 12 Died: 9 (5-years)	41% SUVmax	MTV	183ml	4.495 (1.434–14.09)	4.500 (1.205–16.81)	NS	NS
Mettler [35]	2019	310		310	eBEACOPP (4 or 6 cycles)	PFS: 16 events OS: 7 events (No defined cut-off)	41% SUVmax, >2.5 SUV, Liver SUVmax, Liver SUVmean	MTV41%	NR	NS	NS	NS	
								MTV2.5	NR	NS	NS	NS	NS
								MTVlmax	NR	NS	NS	NS	NS
								MTVlmean	NR	NS	NS	NS	MTVlmax
Akhtari. [12]	2018	267	267	0	ABVD +/- RT	Relapse/ refractor = 27 (5-year)	2.5 SUV or manually contour	MTV2.5	Continuous	1.00 (1.0007-1.0025)	NR	NR	NR
								TLG2.5	1703g	1.00 (1.0001-1.0004)	NR	NR	NR
								MTVman	NR	1.00 (1.0006-1.0019)	NR	NR	NR
								TLGman	NR	1.00 (1.0001-1.0004)	NR	NR	NR
Cottereau [21]	2018	258	258		ABVD +/- RT	PFS: 27 events OS: 12 (5-year)	2.5 SUV	MTV	147ml	5.2 (1.8-14.7)	7.2 (1.6-33.4)	Sig with individual factors, EORTC, GHCS and NCCN	Sig with individual factors, EORTC, GHCS and NCCN
Angelopoulou [14]	2017	162	76	86	ABVD +/- BEACOPP, +/- RT	PFS: 81% OS: 93% (5-year)	TLG from maximal largest diameter x SUVmax	TLG	<35, 35 -100, <100	70% vs 81% vs 94%	NR	NR	NR

First Author	Year	Patient No.	Stage		Treatment	Events (Follow up cut off)	Segmentation Threshold	MTV / TLG	Suggested Cut-Off	Predictive Univariate Analysis HR (95% CI)		Predictive Multivariate Analysis HR (95% CI)	
			I/II	III/IV						PFS	OS	PFS	OS
Kanoun [29]	2015	59			Anthracycline-based 4-6-8 cycles	Relapse: 5 Died: 5 (No defined cut-off)	41% SUVmax	MTV	225ml	42% vs 85%	NR	Sig when analysed with tumour change in SUVmax	NR
Song. [43]	2013	127	127		ABVD +/- RT	PFS: 85.8OS: 88.2 (No defined cut-off)	2.5 SUV	MTV	198ml	10.707 (3.098–37.002)	13.201 (2.975–58.579)	13.008 (3.441–49.174) Age, B symptoms, mediastinal bulky disease	15.831 (3.301–75.926 Age, B symptoms, mediastinal bulky disease
Tseng [46]	2012	30	11	19	Standford V, ABVD, VAMP, BEACOPP	Relapse =6 Died: 4 (4-year)	NR	MTV		NS	NS	NR	NR

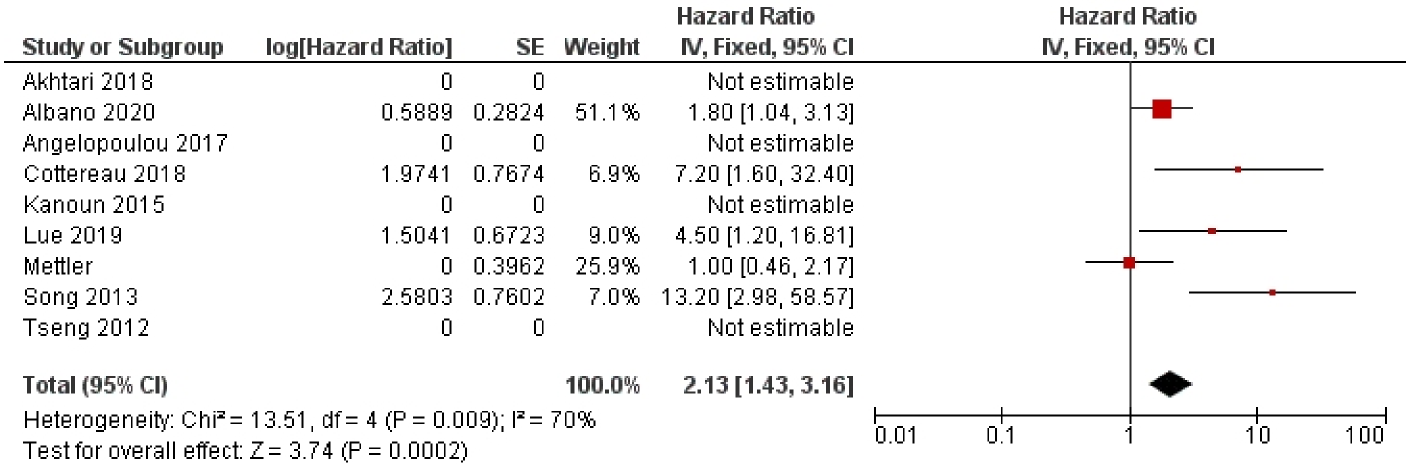
Table 2.4 Studies assessing the use of metabolic tumour volume (MTV) and total lesion glycolysis (TLG) and Hodgkin lymphoma (HL). NR = not reported, NS = not significant, Sig. = significant, HR= hazard ratio, CI = confidence interval, PFS = progressive free survival, OS = overall survival, R-CHOP = rituximab cyclophosphamide, doxorubicin hydrochloride, vincristine (Oncovin) and prednisolone, R-ACVBP - Rituximab, Doxorubicin, Cyclophosphamide, Vindesine, Bleomycin, prednisolone, RT = radiotherapy, ABVD = doxorubicin (Adriamycin), bleomycin, vinblastine and dacarbazine, eBEACOPP = escalated dose bleomycin, etoposide, doxorubicin (Adriamycin), cyclophosphamide, vincristine (Oncovin), procarbazine, and prednisone, VAMP = vincristine, doxorubicin hydrochloride, methotrexate, prednisolone, EORTC = European Organisation for Research and Treatment of Cancer, GHSC = German Hodgkin lymphoma study group, NCCN = National Comprehensive Cancer Network. *The HRs presented as presented in the study but are inverse to the other HRs within the table.



Study or Subgroup	log[Hazard Ratio]	SE	Weight	Hazard Ratio IV, Fixed, 95% CI
Akhtari 2018	0	0		Not estimable
Albano 2020	0.6329	0.2738	37.7%	1.88 [1.10, 3.22]
Angelopoulou 2017	0	0		Not estimable
Cottereau 2018	1.6487	0.5413	9.6%	5.20 [1.80, 15.02]
Kanoun 2015	0	0		Not estimable
Lue 2019	1.503	0.5829	8.3%	4.50 [1.43, 14.09]
Mettler	0.1823	0.275	37.3%	1.20 [0.70, 2.06]
Song 2013	2.3709	0.6327	7.1%	10.71 [3.10, 37.00]
Tseng 2012	0	0		Not estimable
Total (95% CI)			**100.0%**	**2.13 [1.53, 2.96]**

Heterogeneity: Chi^2 = 15.43, df = 4 (P = 0.004); I^2 = 74%
Test for overall effect: Z = 4.51 (P < 0.00001)

Figure 2.6 Forest plot demonstrating hazard ratios for progression free survival for patients with HL using a dichotomous cut-off value derived from the metabolic tumour volume. Studies which do not provide hazard ratios are included but no estimate is given



Study or Subgroup	log[Hazard Ratio]	SE	Weight	Hazard Ratio IV, Fixed, 95% CI
Akhtari 2018	0	0		Not estimable
Albano 2020	0.5889	0.2824	51.1%	1.80 [1.04, 3.13]
Angelopoulou 2017	0	0		Not estimable
Cottereau 2018	1.9741	0.7674	6.9%	7.20 [1.60, 32.40]
Kanoun 2015	0	0		Not estimable
Lue 2019	1.5041	0.6723	9.0%	4.50 [1.20, 16.81]
Mettler	0	0.3962	25.9%	1.00 [0.46, 2.17]
Song 2013	2.5803	0.7602	7.0%	13.20 [2.98, 58.57]
Tseng 2012	0	0		Not estimable
Total (95% CI)			**100.0%**	**2.13 [1.43, 3.16]**

Heterogeneity: Chi² = 13.51, df = 4 (P = 0.009); I² = 70%
Test for overall effect: Z = 3.74 (P = 0.0002)

Figure 2.7 Forest plot demonstrating hazard ratios for overall survival for patients with HL using a dichotomous cut-off value derived from the metabolic tumour volume. Studies which do not provide hazard ratios are included but no estimate is given.

Similar to DLBCL, clinical implementation of MTV and TLG in HL depends on reaching a consensus regarding segmentation methodology, each giving different variations in the volumes measured and will be facilitated by an automated process. However, variation in treatment is likely also to play an impact, and this aspect needs assessing in larger multi-centre studies.

2.4.4 Textural and Shape Analysis for outcome prediction

Textural analysis or radiomics relates to transformation of images into mineable high-dimensional data permitting invisible feature extraction, analysis and modelling for non-invasive phenotyping and outcome prediction [53]. Radiomic features can be studied in isolation or increasingly are being combined with clinical and genomic features as part of the rapidly expanding field of integrated diagnostics [54].

Aide *et al.* studied the use of PET/CT-derived textural features, clinical and imaging parameters to predict two-year PFS in DLBCL patients [10]. They split patients into training (n=105) and validation sets (n=27) and found that Long-Zone High-Grey Level Emphasis (LZHGE) was the only independent predictor when analysed with IPI and MTV. On the validation set, it was found that a high LZHGE $>$ 1,264,925.92 was associated with a two-year PFS of 60% whereas patients with a low LZGHE had a PFS of 94.1%. The study has some limitations as only the largest area of disease was analysed, a breakdown of disease stage was not presented, and 14 patients did not have standard (R-CHOP) therapy. Another study by Aide *et al.* investigated the diagnostic and prognostic value of axial skeletal textural features derived from PET/CT in patients with DLBCL in a retrospective cohort of 82 patients [11]. The CT dataset was initially contoured using a segmentation threshold of >150 Hounsfield units (HU) with the spinal column and half of the pelvis included. They reported that the first-order parameter skewness had the highest AUC for predicting BMI and that a cut-off value of 1.26 produced a sensitivity, specificity, PPV and NPV of 82%, 82%, 62% and 93%, respectively. In addition, a skewness value of <1.26 was associated with a greater two-year PFS and OS. This was true even for 60 patients without BMI. The study had a low event rate (22 patients had BMI), which limits the ability to create a robust prognostic model.

Lue *et al.* investigated the use of 11 first-order, 39 higher-order features and 400 wavelet features for predicting PFS and OS in 42 HL patients (20 stage I/II, 22 stage III/IV) with 21 events within the cohort (12 relapses, 9 deaths) [34]. They found 173 radiomic features, which were significant predictors of progression after correction for multiple testing. To avoid multicollinearity, they only selected the top two features according to the AUC from each group to be included in the univariate and multivariate analysis. MTV was selected based on previous studies. They demonstrated that SUV kurtosis, stage and intensity non-uniformity (INU) derived from Grey-Level Run Length Matrix (GLRLM) were independent predictors of PFS and only disease stage and INU derived from GLRLM were independent predictors of OS.

Decazes *et al.* retrospectively studied PET/CT scans of 215 DLBCL patients to assess the utility of total tumour surface (TTS) and tumour volume surface ratio (TVSR) as predictive biomarkers [23]. TVSR being the ratio between MTV and TTS. MTV had the highest AUC for both OS and PFS (0.71 and 0.67) when compared to TTS (0.69 and 0.66) and TVSR (0.65 and 0.61) [23]. It was reported that TVSR, MTV, IPI and type of chemotherapy were all independent prognostic parameters. Milogrom *et al.* investigated the use of a support vector machine model based on first and second-order radiomic features derived from baseline PET/CT to predict relapse or refractory disease in 167 stage I-II HL patients with mediastinal involvement [37]. Ten of the groups formed the training set, and two were designated the validation set with each group containing a single event (n=12). Five features were selected as the most predictive (SUVmax, MTV, InformationMeasureCorr1, InformationMeasureCorr2, and InverseVariance derived from GLCM 2.5). InformationMeasureCorr1 and InformationMeasureCorr2 are the first and second measures of theoretic correlation and Inverse-Variance is weighting of random variables to minimise variance. By combining these features, the AUC for predicting relapse for patients with mediastinal disease was 0.95. This outperformed TLG and MTV. This work highlights the potential for using AI-methods in lymphoma assessment. However, the study is limited to HL with mediastinal involvement with again small numbers of events.

Senjo *et al.* demonstrated that a high metabolic heterogeneity (MH) was a predictor of five-year PFS and OS in DLBCL across both training (n=86) and validation cohorts (n=64) treated at two centres [41]. They found that MH remained a significant predictor for five-year OS for both cohorts when analysed in multivariate analysis with an ECOG score of >2, and an LDH with a relative risk of 4.75 (95% CI 1.25-18.1) and relative risk of 4.92 (95% CI 1.09-17.03) in the training and validation groups respectively. A model was created which combined MH and MTV, which successfully risk stratified the combined training and validation cohorts into three risk groups: low MH and low MTV, low MH and high MTV or high MH and low MTV, and high MH and high MTV, with the five-year OS being 90.4% vs 69.5% vs 34.8%, respectively; P <0.001 and five-year PFS, 84.1% vs 43.6% vs 27.0%, P <0.001 respectively.

2.5 Current limitations and future challenges

One issue needing to be addressed when using imaging parameters derived from PET for predictive modelling is the relatively low spatial resolution, which influences how much of the avidity is included within a volume when different thresholding techniques are utilised (Figure 2.8) [55]. Meignan *et al.* used a phantom model to validate their MTV thresholding method for a patient cohort [56]. They found that a 41% SUVmax threshold gave the best concordance between contoured and actual volumes. 41% SUVmax thresholding also gave the best agreement between reviewers using the Lin concordance correlation coefficient (pc) (c=0.986, CI 0.97 – 0.99). However, for successful clinical implementation, the time it takes to implement as well as the accuracy of the thresholding method needs be considered. The use of a semi-automated method such as the one reported by Burggraaff *et al.* [57] or a deep learning derived volume as reported by Capobianco *et al.* is required [15]. Predictive models also need to be tested and adapted for new treatments or histological markers [58]. The ability to be able to predict worse outcomes could allow for future treatment stratification. There is an area of unmet need with few active studies at present. There are currently only two open/recruiting studies listed on clinicaltrials.gov assessing PET/CT parameters for outcome prediction in DLBCL, and no registered studies assessing outcomes in HL patients.

Other important limitations of the published work highlighted in this systematic review are variability in methodology and lack of external validation (Table 2.5). This presents a number of opportunities for the future (Table 2.5). Further study into the use of AI for imaging-based outcome prediction in lymphoma which may permit more accurate prediction of prognosis/treatment outcome is needed. This might also facilitate more efficient image analysis and actionable clinical decision support potentially guiding tailored treatment for individual patients. However, there is the requirement for large volumes of data necessary to train algorithms which can then be vigorously validated for reproducibility and generalizability which will require cross-institutional collaboration via imaging networks to support the establishment of multi-centre trials. Implementation studies and health economic research will also be critical for clinical adoption by demonstrating that any AI application is reliable and value-based.

All the described limitations have led to a medium and high risk of bias within the literature as evaluated with our QUIPS tool. The decision to retain papers with a high risk of bias was taken as it was felt that this itself would introduce bias into the review. However, this does mean the results need to be interpretated with caution. Further work in this area is

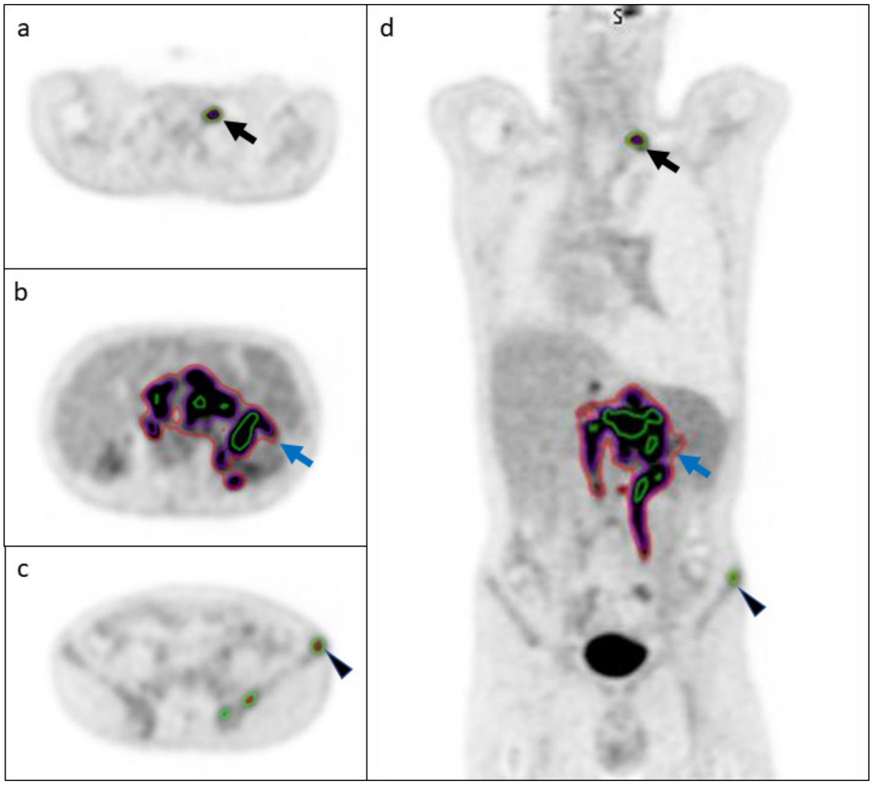


Figure 2.8 Select axial (A-C) and coronal slices (D) from an FDG PET/CT study from a patient with DLBCL demonstrating three different contouring methods (green = 41% SUVmax; red = 1.5 x SUVmean of the liver; purple = 4.0 SUV). For smaller lesions the 41% SUVmax contour is larger than the other 2 methods, black arrow and arrowhead. For larger more heterogenous lesions the 41% SUVmax is the smallest of the 3 contours (blue arrow).

clearly warranted and efforts should be made when designing future studies to carefully consider the methodology employed so as to minimise the risk of bias which is prevalent in this field of work to date.

2.6 Conclusion

Multiple reports suggest the potential utility of various PET/CT derived imaging parameters in lymphoma outcome modelling. Most studies are retrospective and lack external validation of described models. Robustness across different scanning protocols and institutions has also not been verified, and clinical implementation remains a future aspiration. AI techniques may offer a potential solution to some limitations of predictive modelling in this clinical scenario and warrant further evaluation.

Limitation	Opportunity
1. Relatively small retrospective cohorts with limited events	Establishing multi-centre networks for future larger-scale studies
2. Multiple segmentation techniques used	Consensus on segmentation technique for MTV and TLG and development of automated AI-methods which are implemented within reporting software by manufacturers
3. Single site models using a single dataset	Internal and external validation should be routinely performed and facilitated by networks
4. Varying predictive end points	Consensus on clinically relevant predictors
5. Small numbers of papers using non-linear analysis	Using different machine learning and deep learning models to aid in imaging analysis and outcome prediction

Table 2.5 Limitations of the current literature and opportunities for future work

2.7 References

1. Shanbhag S, Ambinder RF. Hodgkin lymphoma: A review and update on recent progress. CA Cancer J Clin. 2018;68:116–32.

2. SEER. SEER Cancer Statistics Review. 1975-2016. 2018.

3. Armitage JO, Gascoyne RD, Lunning MA, Cavalli F. Non-Hodgkin lymphoma. Lancet. 2017;390:298–310.

4. Horvat M, Zadnik V, Šetina TJ, Boltežar L, Goličnik JP, Novaković S, *et al.* Diffuse large B-cell lymphoma: 10 years' real-world clinical experience with rituximab plus cyclophosphamide, doxorubicin, vincristine and prednisolone. Oncol Lett. 2018;15:3602–9.

5. El-Galaly TC, Gormsen LC, Hutchings M. PET/CT for Staging; Past, Present, and Future. Semin Nucl Med. 2018;48:4–16.

6. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, *et al.* The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. J. Clin. Epidemiol. 2009;62(10)e1-e34.

7. Hayden JA, van der Windt DA, Cartwright JL, Co P. Research and Reporting Methods Annals of Internal Medicine Assessing Bias in Studies of Prognostic Factors. Ann Intern Med. 2013;158:280–6.

8. Grooten WJA, Tseli E, Äng BO, Boersma K, Stålnacke B-M, Gerdle B, *et al.* Elaborating on the assessment of the risk of bias in prognostic studies in pain rehabilitation using QUIPS—aspects of interrater agreement. Diagnostic Progn Res. 2019;3:1–11.

9. Adams HJA, de Klerk JMH, Fijnheer R, Heggelman BGF, Dubois S V., Nievelstein RAJ, *et al.* Prognostic superiority of the National Comprehensive Cancer Network International Prognostic Index over pretreatment whole-body volumetric-metabolic FDG-PET/CT metrics in diffuse large B-cell lymphoma. Eur J Haematol. 2015;94:532–9.

10. Aide N, Fruchart C, Nganoa C, Gac A, Lasnon C. Baseline 18 F-FDG PET radiomic features as predictors of 2-year event-free survival in diffuse large B cell lymphomas treated with immunochemotherapy. Eur Radiol. 2020; 30(8):4623-4632.

11. Aide N, Talbot M, Fruchart C, Damaj G, Lasnon C. Diagnostic and prognostic value of baseline FDG PET/CT skeletal textural features in diffuse large B cell lymphoma. Eur J Nucl Med Mol Imaging. 2018;45:699–711.

12. Akhtari M, Milgrom SA, Pinnix CC, Reddy JP, Dong W, Smith GL, *et al.* Reclassifying patients with early-stage Hodgkin lymphoma based on functional radiographic markers at presentation. Blood. 2018;131:84–94.

13. Albano D, Mazzoletti A, Spallino M, Muzi C, Zilioli VR, Pagani C, *et al.* Prognostic role of baseline 18F-FDG PET/CT metabolic parameters in elderly HL: a two-center experience in 123 patients. Ann Hematol. 2020;99:1321–30.

14. Angelopoulou MK, Mosa E, Pangalis GA, Rondogianni P, Chatziioannou S, Prassopoulos V, *et al.* The significance of PET/CT in the initial staging of hodgkin lymphoma: Experience outside clinical trials. Anticancer Res. 2017;37:5727–36.

15. Capobianco N, Meignan MA, Cottereau A-S, Vercellino L, Sibille L, Spottiswoode B, *et al.* Deep learning FDG uptake classification enables total metabolic tumor volume estimation in diffuse large B-cell lymphoma. J Nucl Med. 2020;jnumed.120.242412.

16. Ceriani L, Gritti G, Cascione L, Pirosa MC, Polino A, Ruberto T, *et al.* SAKK38/07 study: Integration of baseline metabolic heterogeneity and metabolic tumor volume in DLBCL prognostic model. Blood Adv. 2020;4:1082–92.

17. Chang C-C, Cho S-F, Chuang Y-W, Lin C-Y, Chang S-M, Hsu W-L, *et al.* Prognostic significance of total metabolic tumor volume on 18F-fluorodeoxyglucose positron emission tomography/ computed tomography in patients with diffuse large B-cell lymphoma receiving rituximab-containing chemotherapy. Oncotarget. 2017;8:99587–600.

18. . Chang C, Cho S, Tu H, Lin C, Chuang Y, Chang S, *et al.* Tumor and bone marrow uptakes on [18F] fluorodeoxyglucose positron emission tomography/computed tomography predict prognosis in patients with diffuse large B-cell lymphoma receiving rituximab-containing chemotherapy. Med. 2017;96(45):e8655.

19. Chihara D, Oki Y, Onoda H, Taji H, Yamamoto K, Tamaki T, *et al.* High maximum standard uptake value (SUVmax) on PET scan is associated with shorter survival in patients with diffuse large B cell lymphoma. Int J Hematol. 2011;93:502–8.

20. Cottereau AS, Nioche C, Dirand AS, Clerc J, Morschhauser F, Casasnovas O, *et al.* 18F-FDG PET dissemination features in diffuse large B-cell lymphoma are predictive of outcome. J Nucl Med. 2020;61:40–5.

21. Cottereau AS, Versari A, Loft A, Casasnovas O, Bellei M, Ricci R, *et al.* Prognostic value of baseline metabolic tumor volume in early-stage Hodgkin lymphoma in the standard arm of the H10 trial. Blood. 2018;131:1456–63.

22. Cottereau AS, Lanic H, Mareschal S, Meignan M, Vera P, Tilly H, *et al.* Molecular profile and FDG-PET/CT Total metabolic tumor volume improve risk classification at diagnosis for patients with diffuse large B-Cell lymphoma. Clin Cancer Res. 2016;22:3801–9.

23. Decazes P, Becker S, Toledano MN, Vera P, Desbordes P, Jardin F, *et al.* Tumor fragmentation estimated by volume surface ratio of tumors measured on 18F-FDG PET/CT is an independent prognostic factor of diffuse large B-cell lymphoma. Eur J Nucl Med Mol Imaging. 2018;45:1672–9.

24. Esfahani SA, Heidari P, Halpern EF, Hochberg EP, Palmer EL, Mahmood U. Baseline total lesion glycolysis measured with (18)F-FDG PET/CT as a predictor of progression-free survival in diffuse large B-cell lymphoma: a pilot study. Am J Nucl Med Mol Imaging. 2013;3:272–81.

25. Gallicchio R, Mansueto G, Simeon V, Nardelli A, Guariglia R, Capacchione D, *et al.* F-18 FDG PET/CT quantization parameters as predictors of outcome in patients with diffuse large B-cell lymphoma. Eur J Haematol. 2014;92:382–9.

26. Huang H, Xiao F, Han X, Zhong L, Zhong H, Xu L, *et al.* Correlation of pretreatm ent 18F-FDG uptake with clinicopathological factors and prognosis in patients with newly diagnosed diffuse large B-cell lymphoma. Nucl Med Commun. 2016;37:689–98.

27. . Ilyas H, Mikhaeel NG, Dunn JT, Rahman F, Möller H, Smith D, *et al.* Is there an optimal method for measuring baseline metabolic tumor volume in diffuse large B cell lymphoma? Eur J Nucl Med Mol Imaging; 2019;46:520–1.

28. Jegadeesh N, Rajpara R, Esiashvili N, Shi Z, Liu Y, Okwan-Duodu D, *et al.* Predictors of local recurrence after rituximab-based chemotherapy alone in stage III and IV diffuse large b-cell lymphoma: Guiding decisions for consolidative radiation. Int J Radiat Oncol Biol Phys. 2015;92:107–12.

29. Kanoun S, Tal I, Berriolo-Riedinger A, Rossi C, Riedinger JM, Vrigneaud JM, *et al.* Influence of software tool and methodological aspects of total metabolic tumor volume calculation on baseline [18F] FDG PET to predict survival in Hodgkin lymphoma. PLoS One. 2015;10:1–15.

30. Kim CY, Hong CM, Kim DH, Son SH, Jeong SY, Lee SW, *et al.* Prognostic value of whole-body metabolic tumour volume and total lesion glycolysis measured on 18F-FDG PET/CT in patients with extranodal NK/T-cell lymphoma. Eur J Nucl Med Mol Imaging. 2013;40:1321–9.

31. Kim TM, Paeng JC, Chun IK, Keam B, Jeon YK, Lee SH, *et al.* Total lesion glycolysis in positron emission tomography is a better predictor of outcome than the International Prognostic Index for patients with diffuse large B cell lymphoma. Cancer. 2013;119:1195–202.

32. Kwon SH, Kang DR, Kim J, Yoon JK, Lee SJ, Jeong SH, *et al.* Prognostic value of negative interim 2-[18 F]-fluoro-2-deoxy-d-glucose PET/CT in diffuse large B-cell lymphoma. Clin Radiol. 2016;71:280–6.

33. Lanic H, Mareschal S, Mechken F, Picquenot JM, Cornic M, Maingonnat C, *et al.* Interim positron emission tomography scan associated with international prognostic index and germinal center B cell-like signature as prognostic index in diffuse large B-cell lymphoma. Leuk Lymphoma. 2012;53:34–42.

34. Lue KH, Wu YF, Liu SH, Hsieh TC, Chuang KS, Lin HH, *et al.* Prognostic Value of Pretreatment Radiomic Features of 18F-FDG PET in Patients with Hodgkin Lymphoma. Clin Nucl Med. 2019;44:E559–65.

35. Mettler J, Müller H, Voltin CA, Baues C, Klaeser B, Moccia A, *et al.* Metabolic tumor volume for response prediction in advanced-stage hodgkin lymphoma. J Nucl Med. 2019;60:207–11.

36. Mikhaeel NG, Smith D, Dunn JT, Phillips M, Møller H, Fields PA, *et al.* Combination of baseline metabolic tumour volume and early response on PET/CT improves progression-free survival prediction in DLBCL. Eur J Nucl Med Mol Imaging. 2016;43:1209–19.

37. Milgrom SA, Elhalawani H, Lee J, Wang Q, Mohamed ASR, Dabaja BS, *et al.* A PET Radiomics Model to Predict Refractory Mediastinal Hodgkin Lymphoma. Sci Rep. 2019;9:1–8.

38. Miyazaki Y, Nawa Y, Miyagawa M, Kohashi S, Nakase K, Yasukawa M, *et al.* Maximum standard uptake value of 18F-fluorodeoxyglucose positron emission tomography is a prognostic factor for progression-free survival of newly diagnosed patients with diffuse large B cell lymphoma. Ann Hematol. 2013;92:239–44.

39. Park S, Moon SH, Park LC, Hwang DW, Ji JH, Maeng CH, *et al.* The impact of baseline and interim PET/CT parameters on clinical outcome in patients with diffuse large B cell lymphoma. Am J Hematol. 2012;87:937–40.

40. Sasanelli M, Meignan M, Haioun C, Berriolo-Riedinger A, Casasnovas RO, Biggi A, *et al.* Pretherapy metabolic tumour volume is an independent predictor of outcome in patients with diffuse large B-cell lymphoma. Eur J Nucl Med Mol Imaging. 2014;41:2017–22.

41. Senjo H, Hirata K, Izumiyama K, Minauchi K, Tsukamoto E, Itoh K, *et al.* High metabolic heterogeneity on baseline 18FDG-PET/CT scan as a poor prognostic factor for newly diagnosed diffuse large B-cell lymphoma. Blood Adv. 2020;4:2286–96.

42. . Song MK, Yang DH, Lee GW, Lim SN, Shin S, Pak KJ, *et al.* High total metabolic tumor volume in PET/CT predicts worse prognosis in diffuse large B cell lymphoma patients with bone marrow involvement in rituximab era. Leuk Res. 2016;42:1–6.

43. Song MK, Chung JS, Lee JJ, Jeong SY, Lee SM, Hong JS, *et al.* Metabolic tumor volume by positron emission tomography/computed tomography as a clinical parameter to determine therapeutic modality for early stage Hodgkin's lymphoma. Cancer Sci. 2013;104:1656–61.

44. Song MK, Chung JS, Shin HJ, Lee SM, Lee SE, Lee HS, *et al.* Clinical significance of metabolic tumor volume by PET/CT in stages II and III of diffuse large B cell lymphoma without extranodal site involvement. Ann Hematol. 2012;91:697–703.

45. Toledano MN, Desbordes P, Banjar A, Gardin I, Vera P, Ruminy P, *et al.* Combination of baseline FDG PET/CT total metabolic tumour volume and gene expression profile have a robust predictive value in patients with diffuse large B-cell lymphoma. Eur J Nucl Med Mol Imaging. 2018;45:680–8.

46. Tseng D, Rachakonda LP, Su Z, Advani R, Horning S, Hoppe RT, *et al.* Interim-treatment quantitative PET parameters predict progression and death among patients with hodgkin's disease. Radiat Oncol. 2012;7.

47. Xie M, Zhai W, Cheng S, Zhang H, Xie Y, He W. Predictive value of F-18 FDG PET/CT quantization parameters for progression-free survival in patients with diffuse large B-cell lymphoma. Hematology. 2016;21:99–105.

48. Zhang YY, Song L, Zhao MX, Hu K. A better prediction of progression-free survival in diffuse large B-cell lymphoma by a prognostic model consisting of baseline TLG and %ΔSUVmax. Cancer Med. 2019;8:5137–47.

49. Zhou M, Chen Y, Huang H, Zhou X, Liu J, Huang G. Prognostic value of total lesion glycolysis of baseline F-fluorodeoxyglucose positron emission tomography/ computed tomography in diffuse large B-cell lymphoma Other factors including MTV, National Comprehensive Cancer Network International Prognostic Ind. Oncotarget. 2016;7:83544–53.

50. . Fletcher JW, Kinahan PE. PET/CT Standardized Uptake Values (SUVs) in Clinical Practice and Assessing Response to Therapy. Semin Ultrasound CT MR. 2010;31:496–505.

51. Adams MC, Turkington TG, Wilson JM, Wong TZ. A systematic review of the factors affecting accuracy of SUV measurements. Am J Roentgenol. 2010;195:310–20.

52. Fletcher JW, Kinahan PE. PET/CT Standardized Uptake Values (SUVs) in Clinical Practice and Assessing Response to Therapy. Natl Inst Heal. 2010;31:496–505.

53. Lambin P, Leijenaar RTH, Deist TM, Peerlings J, De Jong EEC, Van Timmeren J, *et al.* Radiomics: The bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol. 2017;14:749–62.

54. Lundström CF, Gilmore HL, Ros PR. Integrated diagnostics: The computational revolution catalyzing cross-disciplinary practices in radiology, pathology, and genomics. Radiology. 2017;285:12–5.

55. Barrington SF, Meignan M. Time to prepare for risk adaptation in lymphoma by standardizing measurement of metabolic tumor burden. J Nucl Med. 2019;60:1096–102.

56. Meignan M, Sasanelli M, Casasnovas RO, Luminari S, Fioroni F, Coriani C, *et al.* Metabolic tumour volumes measured at staging in lymphoma: Methodological evaluation on phantom experiments and patients. Eur J Nucl Med Mol Imaging. 2014;41:1113–22.

57. Burggraaff CN, Rahman F, Kaßner I, Pieplenbosch S, Barrington SF, Jauw YWS, *et al.* Optimizing Workflows for Fast and Reliable Metabolic Tumor Volume Measurements in Diffuse Large B Cell Lymphoma. Mol Imaging Biol. 2020;22:1102–10.

58. Vercellino L, Cottereau A-S, Casasnovas O, Tilly H, Feugier P, Chartier L, *et al.* High total metabolic tumor volume at baseline predicts survival independent of response to therapy. Blood. 2020;135:1396–405.

2.8 Supplementary Material

2.8.1 Supplemental Table 2.1

Category	Components
Study Details	Author Year Title (shortened path name) DLBCL/HL
Study Participation	The source population or population of interest is adequately described (y = describes cohort and method for assessment) Methods to identify the sample sufficient to limit potential bias (y = consecutive patients without inappropriate exclusions) Description of the baseline study sample (y = describes a breakdown of the study population including age, gender, treatment and confounding factors) Period of recruitment is adequately described (y = gives a precise study period) Place of recruitment adequately described (y = describes if single centre, database or multicentre. If database or multicentre describes how many centres are used) Inclusion and exclusion criteria are adequately described (y = gives a definitive exclusion and inclusion criteria)
Study Attrition	Adequate follow up rate (y = median follow up of over 2 year or a cut-off of over 2 years) Adequate description of participants loss to follow up if any (y = discusses patients lost follow up)
Prognostic Factor Measurement	A clear definition or description of prognostic factor is provided (y = There is a definition of the factor used) Valid and reliable measurement of prognostic factors The method and setting of measurement of PF is the same for all study participants (y = the same criteria was applied to all)
Outcome Measurement	Definition of outcome (y = outcome measure clearly defined e.g. PFS, OS) Valid and reliable measurement of outcome (y = there is a description of how outcome was measured e.g. for relapse was this clinical, histology, imaging or all)
Study Confounding	All important confounders, including treatments are measured (y = consideration of clinical and treatment confounders) Appropriate methods are used if imputation is used for missing confounder data (y = Patients with missing data included but results adjusted for this, ? = patients removed, n = not mentioned)

Category	Components
	Important potential confounders are accounted for in the study design (y = multi-variate analysis or matching training/validation and testing groups, treatment taken into consideration))
Statistical Analysis and Reporting	There is sufficient presentation of data to assess the adequacy of the analysis (y = HR univariate/multivariate analysis with p-values or AUC in machine learning models)
	The strategy for model building (i.e. inclusion of variables in the statistical model) is appropriate (y = appropriate selection of features from univariate analysis or using a feature selection method)
	There is no selective reporting of results (y = all results are reported)

Table S2.1 The questions used as part of the 6 domains of the Quality in Prognosis Studies (QUIPS)

2.8.2 Supplemental Table 2.2

Study	Participation	Attrition	Prognostic Measurement	Outcome Measurement	Confounding	Analysis and Reporting
Adams [9]	Moderate	Moderate	Moderate	Moderate	Moderate	Low
Aide [10]	Low	Moderate	High	High	Moderate	Moderate
Aide [11]	Moderate	Moderate	Moderate	Moderate	Moderate	Low
Akhtari [12]	Moderate	Moderate	Moderate	Moderate	Moderate	Low
Albano [13]	Moderate	Moderate	Moderate	Moderate	Moderate	Low
Angelopulou [14]	High	Moderate	Moderate	Moderate	High	High
Capobianco [15]	Moderate	Moderate	High	High	High	High
Ceriani [16]	Moderate	Moderate	High	High	Moderate	Low
Chang [17]	Moderate	Moderate	Moderate	Moderate	Moderate	Low
Chang [18]	Moderate	Moderate	Moderate	Moderate	Moderate	Low
Chihara [19]	Moderate	Moderate	Moderate	Moderate	Moderate	Low
Cottereau [20]	Moderate	Moderate	High	High	Moderate	Low
Cottereau [21]	Moderate	Moderate	Moderate	Moderate	Moderate	Low
Cottereau [22]	Moderate	Moderate	Moderate	Moderate	Moderate	Low
Decazes [23]	Low Moderate	Moderate	Moderate	Moderate	Low	
Esfahani [24]	Moderate	Moderate	High	High	High	High
Gallicchio [25]	High	High	High	High	High	High
Huang [26]	Moderate	Moderate	Moderate	Moderate	Moderate	Low
Ilyas [27]	High	Moderate	Moderate	Moderate	High	High
Jegadesh [28]	Moderate	Moderate	Moderate	Moderate	Moderate	Low
Kanoun [29]	Moderate	Moderate	High	High	Moderate	Low

Study	Participation	Attrition	Prognostic Measurement	Outcome Measurement	Confounding	Analysis and Reporting
Kim [30]	Moderate	Moderate	Moderate	Moderate	Moderate	Low
Kim [31]	Moderate	Moderate	Moderate	Moderate	Moderate	Low
Kwon [32]	High	Moderate	Moderate	Moderate	Moderate	Low
Lanic [33]	High	Moderate	Moderate	Moderate	High	High
Lue [34]	Moderate	Moderate	Moderate	Moderate	Moderate	Low
Mettler [35]	Moderate	High	Moderate	Moderate	Moderate	Low
Mikhaeel [36]	Moderate	Moderate	Moderate	Moderate	Moderate	Low
Milgrom [37]	Moderate	High	Moderate	Moderate	High	Low
Miyazaki [38]	Moderate	Moderate	Moderate	Moderate	Moderate	Low
Park [39]	Moderate	High	Moderate	Moderate	High	High
Sasanelli [40]	Moderate	Moderate	Moderate	Moderate	Moderate	Low
Senjo [41]	Moderate	Moderate	Moderate	Moderate	High	Low
Song [42]	Moderate	Moderate	Moderate	Moderate	Moderate	Low
Song [43]	Moderate	Moderate	Moderate	Moderate	Moderate	Low
Song [44]	Moderate	Moderate	Moderate	Moderate	Moderate	Low
Toledano [45]	Moderate	Moderate	Moderate	Moderate	Moderate	Low
Tseng [46]	Moderate	Moderate	High	High	High	Low
Xie [47]	High	High	Moderate	Moderate	Moderate	Low
Zhang [48]	Moderate	Moderate	Moderate	Moderate	Moderate	Low
Zhou [49]	Moderate	Moderate	Moderate	Moderate	Moderate	Low

Table S2.2 Break down of consensus risk of bias gradings across 6 domains using the Quality in Prognosis Studies (QUIPS) tool for all studies included within the systematic review

Chapter 3
Discovery of pre-treatment FDG PET/CT-derived radiomics-based models for predicting outcome in diffuse large B-cell lymphoma

3.1 Simple Summary

Diffuse large B-cell lymphoma (DLBCL) is the most common type of lymphoma. Even with the improvements in the treatment of DLBCL around a quarter of patients will experience recurrence. The aim of this single centre retrospective study was to predict which patients would have recurrence within 2 years of their treatment using machine learning techniques based on radiomics extracted from the staging PET/CT images. Our study demonstrated that in our dataset of 229 patient (training data = 183, test data = 46) that a combined radiomic and clinical based model had a good predictive ability which was maintained when tested on an unseen test set.

3.2 Abstract

3.2.1 Background

Approximately 30% of patients with diffuse large B-cell lymphoma (DLBCL) will have recurrence. The aim of this study was to develop a radiomic based model derived from baseline PET/CT to predict 2-year event free survival (2-EFS).

3.2.2 Methods

Patients with DLBCL treated with R-CHOP chemotherapy undergoing pre-treatment PET/CT between January 2008 and January 2018 were included. The dataset was split into training and internal unseen test sets (ratio 80:20). A logistic regression model using metabolic tumour volume (MTV) and six different machine learning classifiers created from clinical and radiomic features derived from the baseline PET/CT were trained and tuned using four-fold cross validation. The model with the highest mean validation receiver operator characteristic curve area under the curve (AUC) was tested on the unseen test set.

3.2.3 Results

229 DLBCL patients met inclusion criteria with 62 (27%) having 2-EFS events. The training cohort had 183 patients with 46 patients in the unseen test cohort. The model

with the highest mean validation AUC combined clinical and radiomic features derived using ridge regression mean validation AUC of 0.75 ±0.06, with a test AUC of 0.73.

3.2.4 Conclusion

The ability of radiomics based predictive models demonstrate promise in predicting outcomes in DLBCL patients.

3.3 Introduction

Diffuse large B-cell lymphoma (DLBCL) is the commonest subtype of non-Hodgkin lymphoma (NHL), accounting for approximately 30-40% of adult cases [1]. Gold standard treatment is with immunochemotherapy: rituximab, cyclophosphamide, doxorubicin hydrochloride, vincristine (Oncovin) and prednisolone (RCHOP) [2]. Radiotherapy can be added if there is bulky or residual disease. Prophylactic intrathecal methotrexate or intravenous treatment with chemotherapy that crosses the blood-brain barrier may be included if there is high risk for central nervous system (CNS) involvement [3]. Even with current therapy regimes approximately 20-30% of patients will recur following treatment [4][5]. Staging and response assessment is performed using 2-deoxy-2-[fluorine18]-fluoro-D-glucose (FDG) positron emission tomography / computed tomography (PET/CT). Treatment stratification based on mid-treatment (interim) PET/CT is commonly used in the management of patients with Hodgkin lymphoma but is less established in DLBCL due to the reduced ability to accurately predict treatment outcome in this lymphoma subtype mid-treatment [6][7]. There is increasing interest in the use of PET/CT derived metrics for treatment stratification at baseline in lymphoma to improve patient outcome. A number of groups have explored the potential utility of baseline metabolic tumour volume (MTV) for predicting event free survival (EFS) with promising results, but this has yet to be adopted clinically [8–16] [17]. Others have explored the potential utility of radiomic features extracted from PET/CT for modelling purposes [8][18]. Initial results are promising, however, published studies are heterogenous with relatively small numbers of patients.

This aim of this study was to develop and test models combining baseline clinical information and radiomic features extracted from PET/CT imaging in DLBCL patients to predict 2-year EFS (2-EFS) using data from our tertiary centre. The secondary aim was to compare model performance to the predictive ability of baseline MTV.

3.4 Materials and Methods

The transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) guidelines were adhered to as part of this study (Supplemental Material 3.1).

3.4.1 Patient selection

Radiological and clinical databases were retrospectively reviewed to determine patients who underwent baseline PET/CT for DLBCL at our institution between January 2008 and January 2018. A cut-off of January 2018 was chosen to allow a minimum of 2 years follow up without interference or confounding factors introduced by the Covid-19 pandemic. Patients were excluded if they did not have DLBCL, were under 16 years of age, had no measurable disease on PET/CT, had hepatic involvement, had a concurrent malignancy, were not treated with R-CHOP or if the images were degraded/incomplete. A 2-EFS event was defined as recurrence or death from any cause within the 2- year follow up period.

3.4.2 PET/CT acquisition

All imaging was performed as part of routine clinical practice. Patients fasted for 6 hours prior to administration of intravenous Fluorine-18 FDG (4 MBq/kg). PET acquisition and reconstruction parameters for the four scanners used at our institution are detailed in Table 3.1. Attenuation correction was performed using a low-dose unenhanced diagnostic CT component acquired using the following settings: 3.75mm slice thickness; pitch 6; 140 kV; 80mAs; pitch 6.

Scanner	**Voxel Size in mm**	**Matrix**	**Reconstruction**	**Scatter Correction**	**Randoms Correction**
Philips Gemini TF64	4 x 4 x 4	144 or 169	BLOB-OS-TF	SS-Simul	DLYD
GE Healthcare Discovery 690	3.65 x 3.65 x 3.27	192	VPFX	Model based	Singles
GE Healthcare Discovery 710	3.65 x 3.65 x 3.27	192	VPFX	Model based	Singles
GE Healthcare STE	4.6875 x 4.6875 x 3.27	128	OSEM	Convolution subtraction	Singles

Table 3.1 Reconstruction parameters for the different scanners used. BLOB-OS-TF = an ordered subset iterative TOF reconstruction algorithm using blobs instead of voxels; DLYD = delayed event subtraction; OSEM = ordered subsets expectation maximization; SS-Simul = single-scatter simulation; VPFX = Vue Point FX (OSEM including point spread function and time of flight).

3.4.3 Image segmentation

All PET/CT images were reviewed and contoured using a specialised multimodality imaging software package (RTx v1.8.2, Mirada Medical, Oxford, UK). FDG-positive disease segmentation was performed by either a clinical radiologist with six years' experience or a research radiographer with two years' experience. Contours were then reviewed by dual-certified Radiology and Nuclear Medicine Physicians with $>$15 years' experience of oncological PET/CT interpretation. Any discrepancies were agreed in consensus.

Two different semi-automated segmentation techniques were used. The first applied a fixed standardised uptake value (SUV) threshold of 4.0, and the second used a threshold derived from 1.5 times mean liver SUV. The 4.0 SUV threshold was selected based on previous work assessing different segmentation techniques in a cohort of DLBCL patients by Burggraaff *et al.* which found it had a higher interobserver reliability [19]. The 1.5 times mean liver SUV threshold was chosen as an adaptive threshold technique which has been used in different cancer types [20, 21]. Mean liver SUV was calculated by placing a 110 cm^3 spherical region of interest (ROI) in the right lobe of the liver. The PET image contour was translated to the CT component of the study with the contours matched to soft tissue with a value of -10 to 100 Hounsfield units (HU). Contours were saved and exported as digital imaging and communications in medicine (DICOM) radiotherapy (RT) structures. Both the images and contours were converted to Neuroimaging Informatics Technology Initiative (NIfTI) files using the python library Simple ITK (v2.0.2).

3.4.4 Feature extraction

Feature extraction was performed using PyRadiomics (v2.2.0). Both the CT and PET images were resampled to a uniform voxel size of 2mm3 . Radiomic features were extracted from the entire segmented disease for each patient. A fixed bin width of 2.5 HU was used for the CT component. Two different bin-widths were used when extracting the radiomic features from the PET component. The first being derived by finding the contour with the maximum range of SUVs and dividing this by 130, the second being derived by dividing the maximum range by 64. This methodology was based on previous work by Orlhac *et al.* and from PyRadiomics documentation [22]. First and second order features were extracted from both the PET and CT components. Further higher order features were explored by extracting the first and second order features following applying wavelet, log-sigma, square, square root, logarithm, exponential, gradient and local binary pattern (lbp)-3D filters to the images. All features extracted and the filters applied are detailed in

Supplemental Material 3.2. The mathematical definition of each of the radiomic features can be found within the PyRadiomics documentation [23]. PyRadiomics deviates from the image biomarker standardisation initiative (IBSI) by applying a fixed bin width from 0 and not the minimum segmentation value, and the calculation of first order kurtosis being $+3$ [24, 25]. Otherwise, PyRadiomics adheres to IBSI guidelines. Patient age, disease stage and sex were also included as clinical features in the models. Disease stage and sex were dummy encoded using (Pandas v1.2.4). This resulted in a total of 3935 features extracted per patient. ComBat harmonisation was applied to account for the different scanners used within the study (https://github.com/Jfortin1/ComBatHarmonization) [26].

3.4.5 Machine learning

The dataset was split into a training and test set stratified around 2-EFS, disease stage, age and sex with an 80:20 split using the scikit-learn (v0.24.2). Concordance between the demographics of the training and test groups was assessed using a t-test for continuous data and a 2 test for categorical data. A p-value of <0.05 was regarded as significant. Continuous data was normalised using a standard scaler (scikit-learn v0.24.2) which was trained and fit on the training set and subsequently applied to the test set. Highly correlated features were removed from the training and test sets if they had a Pearson coefficient over 0.8. This reduced the number of features from 3935 down to 130 for each patient.

Six different machine learning (ML) classifiers were used: logistic regression with lasso, ridge and elasticnet penalties, support vector machine (SVM), random forest and k-nearest neighbour. A maximum number of 5 features were included within each model, apart from in the lasso and elasticnet models where these classifiers determined the optimum number of features. A maximum number of 5 features was chosen using the rule of thumb of 1 feature per 10 events within the training set to avoid false discovery (Type 1 error). Feature selection for the remaining models was performed using three different methods: a forward wrapper method (mlxtend 0.18.0), a univariate analysis method (scikit-learn v0.24.2), and a recursive feature extraction method (where applicable) (scikitlearn v0.24.2). Each method was used to create a list of features from two to the maximum five features which were to be explored in the training phase. The features selected were based on the highest mean receiver operating characteristic (ROC) curve area under the curve (AUC) in four-fold stratified cross validation with 25 repeats.

Training of ML models and tuning of hyperparameters was performed using a grid search with a stratified four-fold cross validation stratified around 2-EFS with 25 repeats. The list of hyperparameters explored within the grid search are detailed in Supplemental Material 3.2. Features and hyperparameters with highest mean validation AUC which was within 0.05 of the mean training AUC were selected. A 0.05 cut-off was chosen to try and minimise selection of an overfitted model. The model which had the highest mean validation AUC overall was tested once on the unseen test set. The Youden index was used to discover the optimum cut-off value from the ROC curve and the accuracy, sensitivity, specificity, negative predictive value (NPV) and positive predictive value (PPV) were calculated from this for the unseen test set. The pipeline for patient inclusion, feature selection and predictive model creation and testing is depicted in (Figure 1). Given the growing evidence surrounding MTV as a predictor of outcome two further logistic regression models were derived from the different segmentation techniques' MTV. A comparison between results from the different cross validation splits between the radiomic model with the mean highest AUC and the MTV model with the mean higher AUC was performed using a Wilcoxon signed ranked test.

3.5 Results

229 DLBCL patients met the inclusion criteria (136 male, 93 female) with 62 2-EFS events. There were 183 patients within the training cohort and 46 patients in the unseen test cohort, there was no statistically significant difference identified between the training and test sets (Table 3.2).

None of the machine learning models created using elasticnet regression, lasso regression or k-nearest neighbour algorithms had a mean validation AUC within 0.05 of the mean training AUC. The remaining model results are presented in Table 3.3 and Table 3.4.

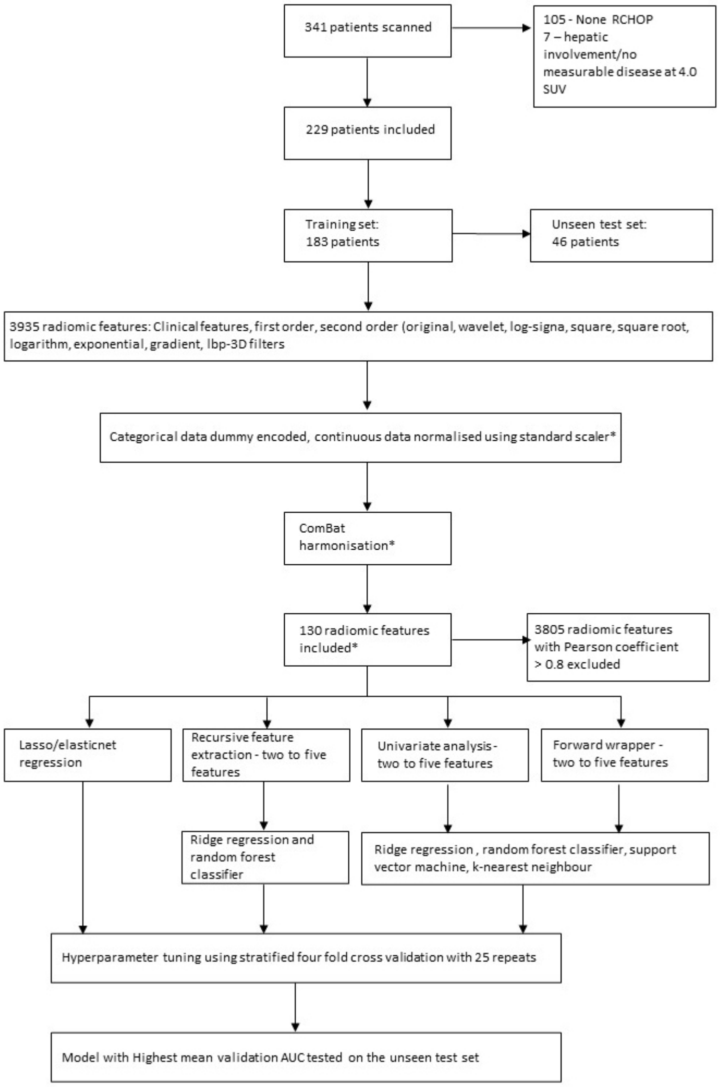

341 patients scanned
105 - None RCHOP
7 – hepatic involvement/no measurable disease at 4.0 SUV
229 patients included
Training set: 183 patients
Unseen test set: 46 patients
3935 radiomic features: Clinical features, first order, second order (original, wavelet, log-signa, square, square root, logarithm, exponential, gradient, lbp-3D filters
Categorical data dummy encoded, continuous data normalised using standard scaler*
ComBat harmonisation*
130 radiomic features included*
3805 radiomic features with Pearson coefficient > 0.8 excluded
Lasso/elasticnet regression
Recursive feature extraction - two to five features
Univariate analysis - two to five features
Forward wrapper - two to five features
Ridge regression and random forest classifier
Ridge regression , random forest classifier, support vector machine, k-nearest neighbour
Hyperparameter tuning using stratified four fold cross validation with 25 repeats
Model with Highest mean validation AUC tested on the unseen test set

Figure 3.1 Pathway for patient inclusion, feature selection and model creation. * = initially applied to the training data and then to the test data.

Demographic	Training Cohort	Test Cohort	p-value
Age	67 (IQR =17)	65 (IQR = 22.5)	0.35
Sex			
Male	107	29	0.69
Female	76	17	
Radiotherapy			
Yes	78	20	0.95
No	105	26	
Stage			
One	42	17	0.26
Two	46	6	
Three	31	6	
Four	64	17	
2-EFS Event			
Yes	50	12	0.98
No	133	34	

Table 3.2 Demographics of the training and testing groups. 2-EFS = 2-year event free survival. The p-values were calculated using a t-test for age and a 2 test for the remaining demographic features.

Machine Learning Model	Hyperparameters	Features	Mean Training	Mean Validation
SUVmax/130				
Ridge Regression	C: 1e-5, penalty: l2, solver: liblinear	Stage One, PET wavelet-LLH GLSZM Large Area Emphasis, PET wavelet-HHH GLSZM Grey Level Non-Uniformity Normalized, PET square 10th Percentile, PET square GLDM Grey Level Non Uniformity	0.75 (0.02)	0.74 (0.07)
Support Vector Machine	C: 1, gamma: 0.00891543, kernel: sigmoid	PET wavelet-HHH GLSZM Grey Level Non-Uniformity Normalized, PET square 10th Percentile, PET lbp-3D-m1 Interquartile Range, PET lbp-3D-m1 GLDM Large Dependence Low Grey Level Emphasis, PET lbp-3D-k 90th Percentile	0.74 (0.02)	0.73 (0.07)

Machine Learning Model	**Hyperparameters**	**Features**	**Mean Training**	**Mean Validation**
Random Forest	bootstrap: False, max depth: 1, max features: log2, min samples leaf: 50, min samples split: 50, n estimators: 10	PET original shape Maximum 2D Diameter Column, MTV, PET original first order Kurtosis, PET original GLSZM Large Area Emphasis, PET wavelet-LHL GLCM Correlation, PET wavelet-LHL GLCM Imc2	0.76 (0.02)	0.71 (0.08)
SUVmax/64				
Ridge Regression	C: 0.001, penalty: l2, solver: newton-cg	Stage Four, PET original GLSZM Large Area Emphasis, PET wavelet-HHL GLSZM Small Area Emphasis, PET wavelet-HHH GLSZM Grey Level Non-Uniformity Normalized, PET square 10th Percentile	0.77 (0.02)	0.75 (0.06)

Machine Learning Model	Hyperparameters	Features	Mean Training	Mean Validation
Support Vector Machine	C: 0.1, gamma: 0.07938667031015, kernel: rbf	PET original GLDM Large Dependence Low Grey Level Emphasis, PET wavelet-HHH GLSZM Grey Level Non-Uniformity Normalized, PET square 10th Percentile, PET lbp-3D-k 90 Percentile, PET lbp-3D-k GLSZM Size Zone Non Uniformity Normalized	0.75 (0.02)	0.72 (0.06)
Random Forest	bootstrap: True, max depth: 1, max features: log2, min samples leaf: 44, min samples split: 6, n estimators: 243	PET original shape Maximum 2D Diameter Column, PET original shape Surface Volume Ratio, PET original 10th Percentile	0.71 (0.02)	0.69 (0.08)

Table 3.3 Mean training and validation scores for the best performing machine learning models using the 4.0 SUV threshold segmentation technique. l2 = Ridge regression penalty, liblinear = A library for large linear classification, GLSZM = grey level size zone matrix, GLDM = grey level dependence matrix, lbp-3D-m1 = local binary pattern filtered image at level 1, lbp-3D-k = local binary pattern kurtosis image, GLCM = grey level co-occurrence matrix, rbf = radial basis function.

Machine Learning Model	Hyperparameters	Features	Mean Training	Mean Validation
SUVmax/130				
Ridge Regression	C: 1e-05, penalty: l2, solver: saga	Stage Four, Age, PET original GLDM Large Dependence Low Grey Level Emphasis, PET original GLSZM Large Area High Grey Level Emphasis	0.74 (0.03)	0.71 (0.09)
Support Vector Machine	C: 1, gamma: 0.437273674187265, kernel: rbf	PET square 10th Percentile, square first order Energy	0.78 (0.02)	0.73 (0.07)
Random Forest	bootstrap: True, max depth: 10, max features: sqrt, min samples leaf: 33, min samples split: 5, n estimators: 90	Age, PET original shape Elongation, PET original shape Least Axis Length, PET original shape Major Axis Length, PET original shape Maximum 2D Diameter Column, PET original shape Mesh Volume		
SUVmax/64				

Machine Learning Model	Hyperparameters	Features	Mean Training	Mean Validation
Ridge Regression	C: 1.0, penalty: l2, solver: liblinear	Stage Three, Age, PET wavelet-LHL GLCM Imc1, PET square GLDM Dependence Variance, PET square GLSZM Small Area Low Grey Level Emphasis	0.76 (0.02)	0.73 (0.07)
Support Vector Machine	C: 1, gamma: 0.437273674187265, kernel: rbf	PET square first order 10 Percentile, PET square first order Energy	0.78 (0.02)	0.73 (0.07)
Random Forest	bootstrap: True, max depth: 10, max features: log2, min samples leaf: 42, min samples split: 6, n estimators: 237	PET original shape Sphericity, PET original GLSZM Large Area Emphasis	0.70 (0.02)	0.69 (0.07)

Table 3.4 Mean training and validation scores for the best performing machine learning models using the 1.5 times mean liver SUV thresholding segmentation technique. l2 = Ridge regression penalty, liblinear = A library for large linear classification, GLSZM = grey level size zone matrix, GLDM = grey level dependence matrix, rbf = radial basis function.

The model within the highest mean validation ROC AUC was the ridge regression model created using radiomic features extracted from a fixed threshold of 4.0 SUV segmentation using a bin width of the maximum range of SUVs divided by 64. The mean training AUC was 0.77 ±0.02, the mean validation AUC was 0.75 ±0.06 and the AUC when tested on the unseen dataset was 0.73 Figure 3.2. The features selected with their coefficients and intercept are presented in Table 3.5. A threshold of 0.5 was chosen and led to an accuracy of 0.70, sensitivity of 0.44, specificity of 0.86, positive predictive value of 0.67, and a negative predictive value of 0.71. The confusion matrix is presented in Table 3.6.

Feature	Coefficient
Stage Four	0.01153414
PET original GLSZM Large Area Emphasis	0.0161316
PET wavelet-HHL GLSZM Small Area Emphasis	0.01482446
PET wavelet-HHH GLSZM Grey Level Non-Uniformity Normalized	-0.01923886
PET square 10 Percentile	-0.01923886
Intercept	-0.01166859

Table 3.5 Features selected and their associated coefficients and intercept in the ridge regression model tested on the unseen test dataset.

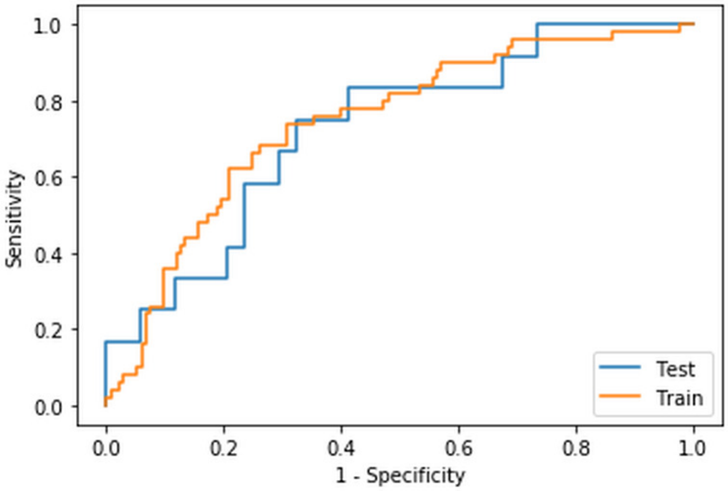


Figure 3.2 ROC Curve of the training and unseen test data AUCs for the model derived using a 4.0 SUV thresholding segmentation technique with a bin width derived from SUVmax/64.

The logistic regression model created solely from MTV using the 4.0 SUV fixed threshold segmentation technique had a mean training AUC of 0.66±0.03 and a mean validation AUC of 0.66 ±0.08. The logistic regression model derived from MTV using the 1.5 times mean liver SUV segmentation technique had a mean training AUC of 0.67±0.03 and a mean validation AUC of 0.67 ± 0.08. There was a statistically significant difference when comparing the cross validation AUCs for the 100 splits between the highest performing MTV based model and the radiomic based ridge regression model $p<0.001$(Figure 3.3).

	Negative	Postive
Predicted Negative	24	10
Predicted Positive	4	8

Table 3.6 Confusion matrix for the threshold of 0.5. Positive = recorded 2-EFS event, Negative = no recorded 2-EFS event, Predicted Positive = predicted to have had a 2-EFS event, Predicted Negative = predicted to not have had a 2-EFS event.

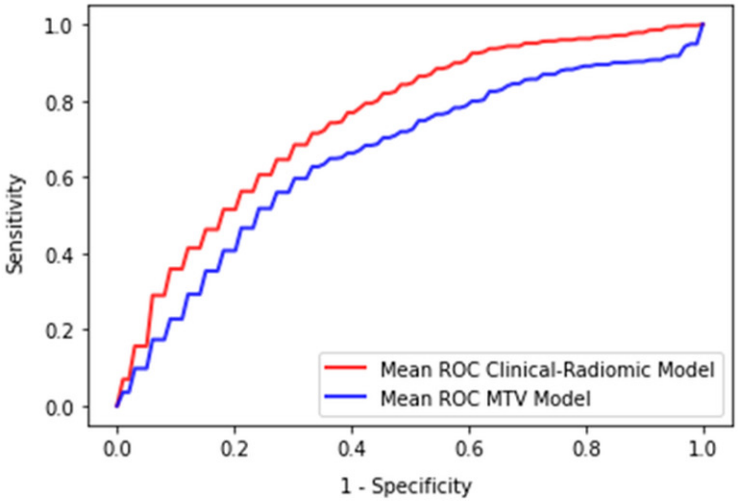


Figure 3.3 ROC curves of the mean validation AUCs for the best performing combined clinical and radiomic model and MTV model.

3.6 Discussion

Our study found that a prediction model combining clinical and radiomic features derived from pretreatment PET/CT using a ridge regression model had the highest mean validation AUC when predicting 2-EFS in DLBCL patients. This model had significantly higher validation AUCs than those achieved by a model solely derived from MTV and achieved an AUC of 0.73 on the unseen test set. The radiomic features used within the model were extracted from a segmentation derived from a fixed threshold of 4.0 SUV using a bin-width calculated from the maximum range of SUVs divided by 64 led to the highest mean validation AUC. The model was formed using five features (Stage Four, PET original GLSZM large area emphasis, PET wavelet-HHL GLSZM small area emphasis, PET wavelet-HHH GLSZM grey level non-uniformity normalized, PET square 10th percentile).

The biological correlate of radiomic features and how these relate to the lesion or disease process can often be overlooked, and can become more complex when image filtering is involved [27]. Three of the radiomic features included in the best model were derived from GLSZM which is a matrix formed by the number of connected voxels with the same grey level intensity. The first being PET GLSZM Large Area Emphasis which is a measure of distribution of large area size zones, extracted from the PET data without any a filter applied. This feature is higher in lesions which have a coarser texture based on the original image. The other two GLZMs are calculated after applying a wavelet filter. Wavelet filters highlight or suppress certain spatial frequencies within an image. In PyRadiomics a combination of high and low filters are passed in each of the different dimensions, which result in 8 different decompositions. PET wavelet-HHL GLSZM small area emphasis is a measure of the distribution of small size zones, which are higher in lesions with fine textures following the application of the wavelet filter. PET wavelet-HHH GLSZM grey level non-uniformity is a measure of the variability of the grey level intensity within the image. A lower value indicates a higher number of similar SUVs on the wavelet filtered image. The last radiomic feature included was PET square 10th percentile which is the 10th percentile value of the SUV after a square of the image SUVs has been taken and normalised to the original SUV range.

Other studies which have explored the use of radiomic features in outcome prediction in DLBCL are heterogenous [12, 28–32]. This is mainly due to differences in segmentation methodology, modelling techniques and outcome measures between groups. Aide *et al.* studied the use of radiomic features in predicting 2-EFS in 132 patients (training = 105, validation = 27) and found that MTV as well as four second order metrics and five third order metrics were selected from ROC analyses. However, longzone high-grey level emphasis was the only independent predictor when analysed with the international prognostic index (IPI) and MTV [29]. In our study long-zone high-grey level emphasis was dropped when checking for multicollinearity. This highlights a potential issue of radiomic model development when applying a methodology on different datasets. It may be that the same features would be chosen between the different datasets, but each method removes the alternate correlated feature and therefore looks to create an entirely new model. Both Zhang *et al.* and Ceriani *et al.* used lasso in their cox regression models to select the most appropriate features [31, 32]. Zhang *et al.* in a study of 152 patients (training = 100, validation = 52) treated with R-CHOP or R-EPOCH (rituximab, etoposide, prednisone, vincristine, cyclophosphamide, and doxorubicin) found that a survival model created with radiomic features and MTV had a validation time dependent ROC AUC of 0.748 (95% CI

0.596–0.886). A model created with radiomic features and metabolic bulk volume had a validation time dependent ROC AUC 0.759 (95% CI 0.595–0.888). Ceriani et al. reported that a radiomic score derived from a training set of 133 patients and tested on an external dataset of 107 patients had an AUC of 0.71 in both the test and validation datasets. The features selected within their cox regression model were GLCM sum squares, maximum 3D diameter and GLDM grey level variance, GLSZM grey level non-uniformity normalized.

In our study both lasso and elasticnet methods failed to produce a model that achieved mean training and validation scores within 0.05 of each other. Even when allowing for a more generous difference between the training and validation scores, mean validation scores remained below 0.65. This 0.05 cut-off is arbitrary and was applied to try and reduce the impact of overfitting on the dataset and allow selection of a potentially more generalisable model. Despite this, there is still a risk that both training and validation datasets are overfitted and the model would need external validation on an external dataset.

One of the largest published studies by Decazes *et al.* in 215 DLBCL patients, explored use of tumour volume surface ratio and total tumour surface as outcome predictors for 5-year progression free survival (PFS), but found that MTV outperformed both features with MTV having an AUC of 0.67 [12]. This AUC for MTV is like the findings in our study, with the mean validation AUC for MTV prediction of 2-EFS being 0.66 for the 4.0 SUV threshold and 0.67 for the 1.5 times liver threshold segmentation techniques respectively. Although, there is growing interest in the use of MTV as an imaging biomarker, Adams *et al.* reported, in a study of 73 DLBCL patients, that the prognostic ability of MTV does not add anything to the prognostic ability of the clinical scoring system National Comprehensive Cancer Network-International Prognostic Index (NCCN-IPI) [33]. Unfortunately, due to missing clinical data it was not possible to compare IPI performance in our patient cohort. However, this does highlight the potential impact of confounders on the generalisability of predictive models. Although, causality is not generally considered in predictive modelling its use in future models could allow for greater transparency of a model. The issues of generalisability may be compounded by learnt biases towards groups of patients in the training process.

The TRIPOD checklist was completed to increase transparency of model development [34, 35]. However, there are limitations to our study including the retrospective nature and uncertainty surrounding the exact timing and recording of recurrence. Use of 2-EFS partially mitigates against this by allowing a wider window for the relapse to be recorded, however, it does mean that data is lost which could have been included in a time to survival type model. 2-EFS was chosen as the majority of patients relapse within the first 2 years. Time to event ML models could be used in future studies to reduce the need to exclude data. Lack of clinical data surrounding the IPI and cell of origin (COO) information, meant that these could not be used as direct comparators to radiomic models created.

3.7 Conclusions

A combined clinical and PET/CT derived radiomics model using ridge regression demonstrated the highest mean validation AUC validation (AUC = 0.75) when predicting 2-EFS in DLBCL patients treated with R-CHOP, which outperformed a model derived solely from MTV (AUC = 0.67).

3.8 References

1. Armitage, J.O.; Gascoyne, R.D.; Lunning, M.A.; Cavalli, F. Non-Hodgkin lymphoma. Lancet 2017, 390, 298–310, doi:10.1016/S0140-6736(16)32407-2.

2. Coiffier, B.; Thieblemont, C.; Van Den Neste, E.; Lepeu, G.; Plantier, I.; Castaigne, S.; Lefort, S.; Marit, G.; Macro, M.; Sebban, C.; *et al.* Long-term outcome of patients in the LNH-98.5 trial, the first randomized study comparing rituximab-CHOP to standard CHOP chemotherapy in DLBCL patients: A study by the Groupe d'Etudes des Lymphomes de l'Adulte. Blood 2010, 116, 2040–2045, doi:10.1182/blood-2010-03-276246.

3. Kansara, R. Central Nervous System Prophylaxis Strategies in Diffuse Large B Cell Lymphoma. Curr. Treat. Options Oncol. 2018, 19, doi:10.1007/s11864-018-0569-2.

4. SEER SEER Cancer Statistics Review.

5. Cheson, B.D.; Fisher, R.I.; Barrington, S.F.; Cavalli, F.; Schwartz, L.H.; Zucca, E.; Lister, T.A. Recommendations for initial evaluation, staging, and response assessment of hodgkin and non-hodgkin lymphoma: The lugano classification. J. Clin. Oncol. 2014, 32, 3059–3067, doi:10.1200/JCO.2013.54.8800.

6. Yoo, C.; Lee, D.H.; Kim, J.E.; Jo, J.; Yoon, D.H.; Sohn, B.S.; Kim, S.W.; Lee, J.S.; Suh, C. Limited role of interim PET/CT in patients with diffuse large B-cell lymphoma treated with R-CHOP. Ann. Hematol. 2011, 90, 797–802, doi:10.1007/s00277-010-1135-6.

7. Mikhaeel, N.G.; Cunningham, D.; Counsell, N.; McMillan, A.; Radford, J.A.; Ardeshna, K.M.; Lawrie, A.; Smith, P.; Clifton-Hadley, L.; O'Doherty, M.J.; *et al.* FDG-PET/CT after two cycles of R-CHOP in DLBCL predicts complete remission but has limited value in identifying patients with poor outcome – final result of a UK National Cancer Research Institute prospective study. Br. J. Haematol. 2021, 192, 504–513, doi:10.1111/bjh.16875.

8. Frood, R.; Burton, C.; Tsoumpas, C.; Frangi, A.F.; Gleeson, F.; Patel, C.; Scarsbrook, A. Baseline PET/CT imaging parameters for prediction of treatment outcome in Hodgkin and diffuse large B cell lymphoma: a systematic review. Eur. J. Nucl. Med. Mol. Imaging 2021, doi:10.1007/s00259-021-05233-2.

9. Song, M.K.; Chung, J.S.; Shin, H.J.; Lee, S.M.; Lee, S.E.; Lee, H.S.; Lee, G.W.; Kim, S.J.; Lee, S.M.; Chung, D.S. Clinical significance of metabolic tumor volume by PET/CT in stages II and III of diffuse large B cell lymphoma without extranodal site involvement. Ann. Hematol. 2012, 91, 697–703, doi:10.1007/s00277-011-1357-2.

10. Song, M.K.; Yang, D.H.; Lee, G.W.; Lim, S.N.; Shin, S.; Pak, K.J.; Kwon, S.Y.; Shim, H.K.; Choi, B.H.; Kim, I.S.; *et al.* High total metabolic tumor volume in PET/CT predicts worse prognosis in diffuse large B cell lymphoma patients with bone marrow involvement in rituximab era. Leuk. Res. 2016, 42, 1–6, doi:10.1016/j.leukres.2016.01.010.

11. Cottereau, A.S.; Nioche, C.; Dirand, A.S.; Clerc, J.; Morschhauser, F.; Casasnovas, O.; Meignan, M.; Buvat, I. 18F-FDG PET dissemination features in diffuse large B-cell lymphoma are predictive of outcome. J. Nucl. Med. 2020, 61, 40–45, doi:10.2967/jnumed.119.229450.

12. Decazes, P.; Becker, S.; Toledano, M.N.; Vera, P.; Desbordes, P.; Jardin, F.; Tilly, H.; Gardin, I. Tumor fragmentation estimated by volume surface ratio of tumors measured on 18F-FDG PET/CT is an independent prognostic factor of diffuse large B-cell lymphoma. Eur. J. Nucl. Med. Mol. Imaging 2018, 45, 1672–1679, doi:10.1007/s00259-018-4041-0.

13. Ilyas, H.; Mikhaeel, N.G.; Dunn, J.T.; Rahman, F.; Möller, H.; Smith, D.; Barrington, S.F. Is there an optimal method for measuring baseline metabolic tumor volume in diffuse large B cell lymphoma? Eur. J. Nucl. Med. Mol. Imaging 2019, 46, 520–521, doi:10.1007/s00259-018-4200-3.

14. Toledano, M.N.; Desbordes, P.; Banjar, A.; Gardin, I.; Vera, P.; Ruminy, P.; Jardin, F.; Tilly, H.; Becker, S. Combination of baseline FDG PET/CT total metabolic tumour volume and gene expression profile have a robust predictive value in patients with diffuse large B-cell lymphoma. Eur. J. Nucl. Med. Mol. Imaging 2018, 45, 680–688, doi:10.1007/s00259-017-3907-x.

15. Chang, C.-C.; Cho, S.-F.; Chuang, Y.-W.; Lin, C.-Y.; Chang, S.-M.; Hsu, W.-L.; Huang, Y.-F. Prognostic significance of total metabolic tumor volume on 18F-fluorodeoxyglucose positron emission tomography/ computed tomography in patients with diffuse large B-cell lymphoma receiving rituximab-containing chemotherapy. Oncotarget 2017, 8, 99587–99600, doi:10.18632/oncotarget.20447.

16. Cottereau, A.S.; Lanic, H.; Mareschal, S.; Meignan, M.; Vera, P.; Tilly, H.; Jardin, F.; Becker, S. Molecular profile and FDG-PET/CT Total metabolic tumor volume improve risk classification at diagnosis for patients with diffuse large B-Cell lymphoma. Clin. Cancer Res. 2016, 22, 3801–3809, doi:10.1158/1078-0432.CCR-15-2825.

17. Mikhaeel, N.G.; Smith, D.; Dunn, J.T.; Phillips, M.; Møller, H.; Fields, P.A.; Wrench, D.; Barrington, S.F. Combination of baseline metabolic tumour volume and early response on PET/CT improves progression-free survival prediction in DLBCL. Eur. J. Nucl. Med. Mol. Imaging 2016, 43, 1209–1219, doi:10.1007/s00259-016-3315-7.

18. Lambin, P.; Leijenaar, R.T.H.; Deist, T.M.; Peerlings, J.; De Jong, E.E.C.; Van Timmeren, J.; Sanduleanu, S.; Larue, R.T.H.M.; Even, A.J.G.; Jochems, A.; *et al.* Radiomics: The bridge between medical imaging and personalized medicine. Nat. Rev. Clin. Oncol. 2017, 14, 749–762, doi:10.1038/nrclinonc.2017.141.

19. Burggraaff, C.N.; Rahman, F.; Kaßner, I.; Pieplenbosch, S.; Barrington, S.F.; Jauw, Y.W.S.; Zwezerijnen, G.J.C.; Müller, S.; Hoekstra, O.S.; Zijlstra, J.M.; *et al.* Optimizing Workflows for Fast and Reliable Metabolic Tumor Volume Measurements in Diffuse Large B Cell Lymphoma. Mol. Imaging Biol. 2020, 22, 1102–1110, doi:10.1007/s11307-020-01474-z.

20. Brown, P.J.; Zhong, J.; Frood, R.; Currie, S.; Gilbert, A.; Appelt, A.L.; Sebag-Montefiore, D.; Scarsbrook, A. Prediction of outcome in anal squamous cell carcinoma using radiomic feature analysis of pre-treatment FDG PET-CT. Eur. J. Nucl. Med. Mol. Imaging 2019, 46, 2790–2799, doi:10.1007/s00259-019-04495-1.

21. Zhong, J.; Frood, R.; Brown, P.; Nelstrop, H.; Prestwich, R.; McDermott, G.; Currie, S.; Vaidyanathan, S.; Scarsbrook, A.F. Machine learning-based FDG PET-CT radiomics for outcome prediction in larynx and hypopharynx squamous cell carcinoma. Clin. Radiol. 2021, 76, 78.e9-78.e17, doi:10.1016/j.crad.2020.08.030.

22. Orlhac, F.; Soussan, M.; Chouahnia, K.; Martinod, E.; Buvat, I. 18F-FDG PET-derived textural indices reflect tissue-specific uptake pattern in non-small cell lung cancer. PLoS One 2015, 10, 1–16, doi:10.1371/journal.pone.0145063.

23. PyRadiomics Radiomic Features Available online: https://pyradiomics.readthedocs.io/en/latest/features.html.

24. Zwanenburg, A.; Leger, S.; Vallières, M.; Löck, S. Image biomarker standardisation initiative. 2016, doi:10.1148/radiol.2020191145.

25. PyRadiomics Frequently Asked Questions Available online: https://pyradiomics.readthedocs.io/en/latest/faq.html.

26. Fortin, J.-P.; Cullen, N.; Sheline, Y.I.; Taylor, W.D.; Aselcioglu, I.; Cook, P.A.; Adams, P.; Cooper, C.; Fava, M.; McGrath, P.J.; *et al.* Harmonization of cortical thickness measurements across scanners and sites. Neuroimage 2018, 167, 104–120, doi:10.1016/j.neuroimage.2017.11.024.

27. Tomaszewski, M.R.; Gillies, R.J. The biological meaning of radiomic features. Radiology 2021, 298, 505–516, doi:10.148/radiol.2021202553.

28. Senjo, H.; Hirata, K.; Izumiyama, K.; Minauchi, K.; Tsukamoto, E.; Itoh, K.; Kanaya, M.; Mori, A.; Ota, S.; Hashimoto, D.; *et al.* High metabolic heterogeneity on baseline 18FDG-PET/CT scan as a poor prognostic factor for newly diagnosed diffuse large B-cell lymphoma. Blood Adv. 2020, 4, 2286–2296, doi:10.1182/bloodadvances.2020001816.

29. Aide, N.; Fruchart, C.; Nganoa, C.; Gac, A.; Lasnon, C. Baseline 18 F-FDG PET radiomic features as predictors of 2-year event-free survival in diffuse large B cell lymphomas treated with immunochemotherapy. 2020.

30. Bera, K.; Braman, N.; Gupta, A.; Velcheti, V.; Madabhushi, A. Predicting cancer outcomes with radiomics and artificial intelligence in radiology. Nat. Rev. Clin. Oncol. 2021, doi:10.1038/s41571-021-00560-7.

31. Ceriani, L.; Milan, L.; Cascione, L.; Gritti, G.; Dalmasso, F.; Esposito, F.; Pirosa, M.C.; Schär, S.; Bruno, A.; Dirnhofer, S.; *et al.* Generation and validation of a PET radiomics model that predicts survival in diffuse large B cell lymphoma treated with R-CHOP14: A SAKK 38/07 trial post-hoc analysis. Hematol. Oncol. 2021, doi:10.1002/hon.2935.

32. Zhang, X.; Chen, L.; Jiang, H.; He, X.; Feng, L.; Ni, M.; Ma, M.; Wang, J.; Zhang, T.; Wu, S.; *et al.* A novel analytic approach for outcome prediction in diffuse large B-cell lymphoma by [18F]FDG PET/CT. Eur. J. Nucl. Med. Mol. Imaging 2021, doi:10.1007/s00259-021-05572-0.

33. Adams, H.J.A.; de Klerk, J.M.H.; Fijnheer, R.; Heggelman, B.G.F.; Dubois, S. V.; Nievelstein, R.A.J.; Kwee, T.C. Prognostic superiority of the National

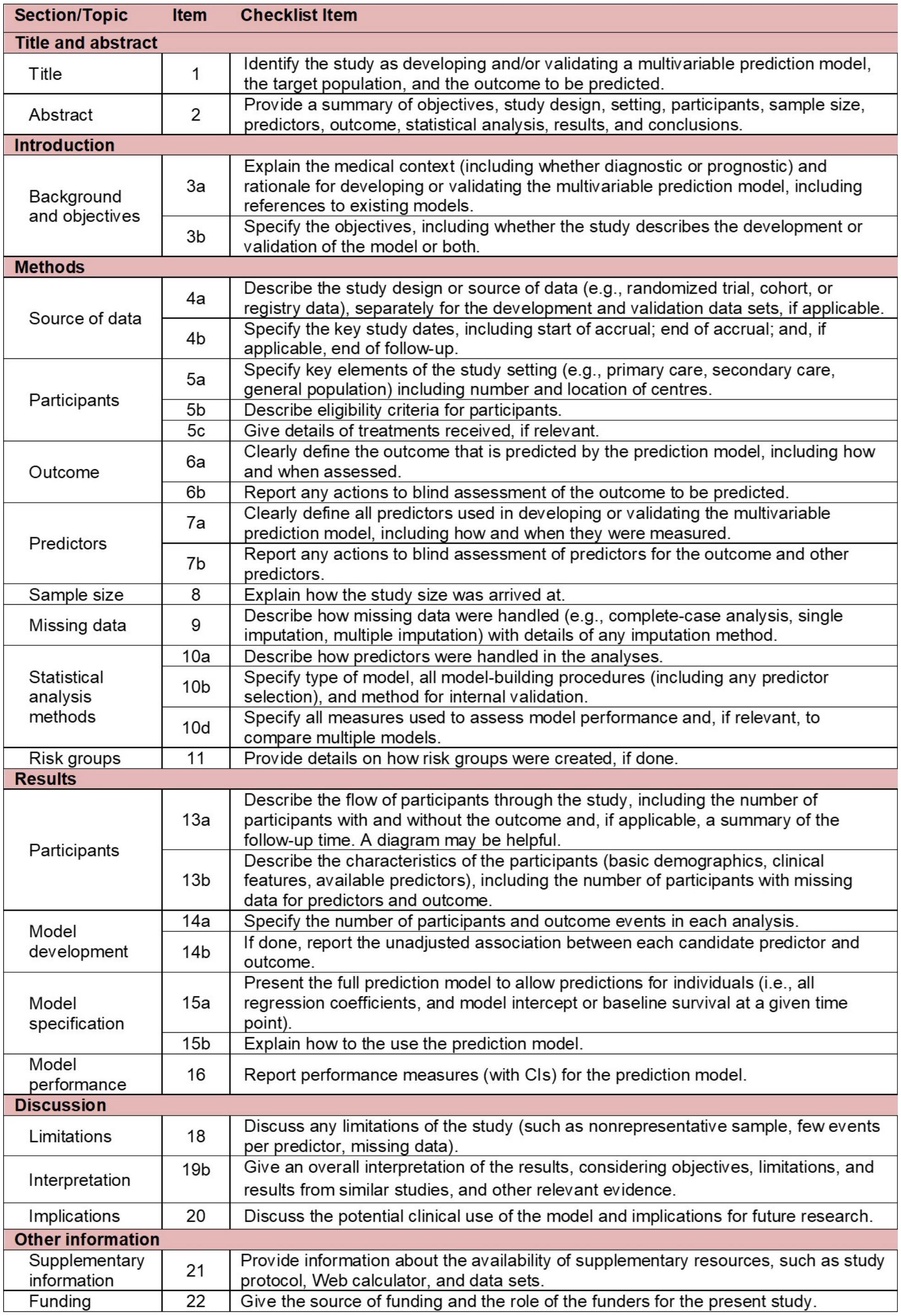
Comprehensive Cancer Network International Prognostic Index over pretreatment whole-body volumetric-metabolic FDG-PET/CT metrics in diffuse large B-cell lymphoma. Eur. J. Haematol. 2015, 94, 532–539, doi:10.1111/ejh.12467.

34. Park, J.E.; Kim, D.; Kim, H.S.; Park, S.Y.; Kim, J.Y.; Cho, S.J.; Shin, J.H.; Kim, J.H. Quality of science and reporting of radiomics in oncologic studies: room for improvement according to radiomics quality score and TRIPOD statement. Eur. Radiol. 2020, 30, 523–536, doi:10.1007/s00330-019-06360-z.

35. Pinto dos Santos, D.; Dietzel, M.; Baessler, B. A decade of radiomics research: are images really data or just patterns in the noise? Eur. Radiol. 2021, 31, 2–5, doi:10.1007/s00330-020-07108-w.

3.9 Supplementary Material

3.9.1 Supplemental Material 3.1



Section/Topic	Item	Checklist Item
Title and abstract		
Title	1	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.
Abstract	2	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.
Introduction		
Background and objectives	3a	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.
	3b	Specify the objectives, including whether the study describes the development or validation of the model or both.
Methods		
Source of data	4a	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.
	4b	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.
Participants	5a	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.
	5b	Describe eligibility criteria for participants.
	5c	Give details of treatments received, if relevant.
Outcome	6a	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.
	6b	Report any actions to blind assessment of the outcome to be predicted.
Predictors	7a	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.
	7b	Report any actions to blind assessment of predictors for the outcome and other predictors.
Sample size	8	Explain how the study size was arrived at.
Missing data	9	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.
Statistical analysis methods	10a	Describe how predictors were handled in the analyses.
	10b	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.
	10d	Specify all measures used to assess model performance and, if relevant, to compare multiple models.
Risk groups	11	Provide details on how risk groups were created, if done.
Results		
Participants	13a	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.
	13b	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.
Model development	14a	Specify the number of participants and outcome events in each analysis.
	14b	If done, report the unadjusted association between each candidate predictor and outcome.
Model specification	15a	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).
	15b	Explain how to the use the prediction model.
Model performance	16	Report performance measures (with CIs) for the prediction model.
Discussion		
Limitations	18	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).
Interpretation	19b	Give an overall interpretation of the results, considering objectives, limitations, and results from similar studies, and other relevant evidence.
Implications	20	Discuss the potential clinical use of the model and implications for future research.
Other information		
Supplementary information	21	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.
Funding	22	Give the source of funding and the role of the funders for the present study.

1. This study looks at the utility of pre-treatment FDG PET/CT derived machine learning models for outcome prediction in diffuse large B-cell lymphoma (DLBCL). (Title)

2. A summary of all the requested information is presented in the Abstract.

3. a) The introduction presents the background, aim of the model and the previous studies which have explored models containing radiomic features from the baseline PET/CT. (Introduction)b) The aim of this study was to create a predictive model using both clinical

and radiomic features derived from pre-treatment FDG PET/CT to predict 2-year EFS in DLBCL patients using a cohort from a tertiary treatment centre (Introduction)

4. a) Retrospective single centre cohort study. The study cohort was randomised on a ratio of 4:1 into training and testing cohorts stratified around 2-EFS, age, sex and disease stage. (Patient selection)
b) Consecutive patients with histologically proven DLBCL NOS who underwent baseline FDG-PET/CT at a single large tertiary referral centre between June 2008 and January 2018 were included. The follow up information recorded is set out in the patient selection section. (Patient selection)

5. a) This is a single tertiary centre study. (Patient selection) b) Patients were excluded if they did not have DLBCL NOS, were under 16 years of age, had no measurable disease on PET/CT when using a 4.0SUV or had hepatic involvement, had a concurrent malignancy, were not treated with R-CHOP or if the images were degraded on incomplete. The follow up information recorded is set out (Patient selection) c) The treatment regimen was with RCHOP for all patients. No change to departmental standard treatment was performed. (Patient selection)

6. a) A 2-EFS event was defined as recurrence or death from any cause within the 2-year follow up period. (Patient selection) b) As this was a retrospective study the primary outcomes were defined from clinical records. The investigator reviewing the records was blinded to the imaging parameters.

7. a) The description of the contouring method, resampling, harmonisation, radiomic feature extraction, feature selection and model training and testing is documented within the materials and methods section. The features selected as part of the chosen model are described the results section. (Materials and methods, Results) b) The images were contoured blinded to the outcome data.

8. All patients which met the inclusion criteria were included. The cut-off of January 2018 was chosen to allow for 2 years follow up whilst minimising confounding factors introduced by the Covid-19 pandemic. For feature selection 5 features were chosen as the maximum number of features to be include in each model. This was derived from

10 events per parameter, with 50 events within the training cohort. This was not the case for the lasso and elastinet models whose feature selection was derived by the penalty applied to the models (Materials and methods, Results)

9. Only complete data sets were used in the analysis. (Results)

10. a) Clinical factors were included in the variable selection process alongside radiomic features. The categorical data was dummy encoded and the continuous features were normalised using a standard scaler. (Machine learning analysis) b) Random forest, support vector machine, ridge regression, lasso regression, elasticnet regression and k-nearest neighbour models were trained and tuned on the training cohort using four fold cross validation with 25 repeats. The models were created using different feature selection methods, using two different segmentation techniques and two different PET bin widths. The model with the highest mean receiver operating characteristic (ROC) area under the curve (AUC) was tested once on the unseen test cohort. (Machine learning analysis) d) When comparing models, the mean validation AUC was used to determine the best performing model, the model with the highest being tested on the unseen test set (Machine learning analysis)

11. Risk groups were not created within the model.

13. a) 229 patients were included, with demographics detailed in Table 2. (Results) b) The characteristics of the participants are presented in Table 2.

14. a) The number of events per cohort are presented in Table 2. b) This has not been performed. The training and testing cohorts were stratified around key clinical features but the results are not adjusted for these.

15. a/b) The features and hyperparameters used to create the model are presented in the results section.

16. The mean validation and test ROC curves are presented, standard deviations are presented for the mean training and mean validation scores. The confusion matrix with the accuracy, sensitivity, specificity, negative predictive value and negative predictive value for the best performing model with the a threshold derived using the Youden index from the ROC curve (Results)

18. The limitations of the study are presented in the discussion. These include the retrospective nature of the study, the relative low number of events, reliance on clinical

records, the exclusion of patients with hepatic disease or not having measurable disease above 4.0 SUV, and that there was no external validation. (Discussion)

19. b)/20. The discussion section gives an overall interpretation of the results, it highlights the potential use of a pre-treatment model, but discusses the next steps needed to make this clinically viable (Discussion)

21. The python libraries used are references within the text and the radiomic features extracted using PyRadiomics are detailed in Supplementary Material Table 1.

22. Individual author's funding is declared within the Declaration. The study was not externally funded.

3.9.2 Supplemental Material 3.2

First Order	Shape	GLCM	GLRLM	GLDM	GLSZM	NGTDM
10th Percentile	Elongation	Autocorrelation	Grey Level Non-Uniformity	Dependence Entropy	Grey Level Non-Uniformity	Busyness
90th Percentile	Flatness	Cluster Prominence	Grey Level Non-Uniformity Normalized	Dependence Non-Uniformity	Grey Level Non-Uniformity Normalized	Coarseness
Energy	Least Axis Length	Cluster Shade	Grey Level Variance	Dependence Non-Uniformity Normalized	Grey Level Variance	Complexity
Entropy	Major Axis Length	Cluster Tendency	High Grey Level Run Emphasis	Dependence Variance	High Grey Level Zone Emphasis	Contrast
Inter quartile Range	Maximum 2D Diameter Column	Contrast	Long Run Emphasis	Grey Level Non-Uniformity	Large Area Emphasis	Strength
Kurtosis	Maximum 2D Diameter Row	Correlation	Long Run High Grey Level Emphasis	Grey Level Variance	Large Area High Grey Level Emphasis	
Maximum	Maximum 2D Diameter Slice	Difference Average	Long Run Low Grey Level Emphasis	High Grey Level Emphasis	Large Area Low Grey Level Emphasis	
Mean Absolute Deviation	Maximum 3D Diameter	Difference Entropy	Low Grey Level Run Emphasis	Large Dependence Emphasis	Low Grey Level Zone Emphasis	
Mean	Mesh Volume	Difference Variance	Run Entropy	Large Dependence High Grey Level Emphasis	Size Zone Non-Uniformity	
Median	Minor Axis Length	Id	Run Length Non-Uniformity	Large Dependence Low Grey Level Emphasis	Size Zone Non-Uniformity Normalized	
Minimum	Sphericity	Idm	Run Percentage	Low Grey Level Emphasis	Small Area Emphasis	
Range	Surface Area	Idmn	Run Variance	Small Dependence Emphasis	Small Area High Grey Level Emphasis	
Robust Mean Absolute Deviation	Surface Volume Ratio	Idn	Short Run Emphasis	Small Dependence High Grey Level Emphasis	Small Area Low Grey Level Emphasis	

First Order	Shape	GLCM	GLRLM	GLDM	GLSZM	NGTDM
Root Mean Squared	Voxel Volume	Imc1	Short Run High Grey Level Emphasis	Small Dependence Low Grey Level Emphasis	Zone Entropy	
Skewness		Imc2	Short Run Low Grey Level Emphasis		Zone Percentage	
Total Energy		Inverse Variance			Zone Variance	
Uniformity		Joint Average				
Variance		Joint Energy				
		Joint Entropy				
		MCC				
		Maximum Probability				
		Sum Average				
		Sum Entropy				
		Sum Squares				

Table S3.1 The radiomic features extracted for both the PET and CT components. The equations for the features can be found at https://pyradiomics.readthedocs.io/en/latest/features.html. GLCM = grey level co-occurrence matrix, GLDM = grey level dependence matrix, GLRLM = grey level run length matrix, GLSZM = grey level size zone matrix, NGTDM = neighbouring grey tone difference matrix, Id = inverse difference, Idn = inverse difference normalised, Imc = informational measure of correlation, Idm = inverse difference moment, Idmn = inverse difference moment normalised, MCC = maximal correlation coefficient. Each of the first and second order features were extracted from the original imaging and then from the images following filters applied. The filters used were: wavelet; log-signa; square; square root; logarithm; exponential; gradient; lbp-3D.

Model	List of hyperparameters explored	Static hyperparameters
Lasso Logistic Regression	C = [0.0000001, 0.000001, 0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10, 100] solver = ['liblinear', 'saga']	penalty = 'l1' random_state=0 class_weight="balanced" max_iter =10000
Elasticnet Logistic Regression	C = [0.0000001, 0.000001, 0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10, 100] l1_ratio = [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]	penalty = 'elasticnet' solver = "saga" random_state=0, class_weight="balanced" max_iter =10000

Ridge Logistic Regression	C = [0.0000001, 0.000001, 0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10, 100] solver = [’liblinear’, ’saga’, ”sag”, ’lbfgs’, ’newton-cg’]	penalty = ’l2’ random_state=0 class_weight=”balanced” max_iter =10000
Support Vector Machine	C = [0.0000001, 0.000001, 0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10] gamma = expon(scale=.1) kernel = [’linear’, ’rbf’, ”sigmoid”, ”poly”]	random_state=0 class_weight=”balanced” probability = True
Random Forest	n_estimators = [int(x) for x in np.linspace(start = 10, stop = 250, num = 40)] max_features = [’log2’, ’sqrt’] max_depth = [int(x) for x in np.linspace(start = 1, stop = 10, num = 5)] min_samples_split = [int(x) for x in np.linspace(start = 2, stop = 50, num = 49)] min_samples_leaf = [int(x) for x in np.linspace(start = 2, stop = 50, num = 49)] bootstrap = [True, False]	random_state = 0 class_weight = ”balanced” probability = True
K-Nearest Neighbour	n_neighbors = range(1, 21, 2) weights = [’uniform’, ’distance’] metric = [’euclidean’, ’manhattan’, ’minkowski’]	

Table S3.2 The hyperparameters explored within the grid search and the hyperparameters kept static when training the different machine learning models. If a hyperparameter is not documented, it was left as the default within the library.

Chapter 4
Utility of pre-treatment FDG PET/CT derived machine learning models for outcome prediction in classical Hodgkin lymphoma

4.1 Abstract

4.1.1 Objectives

Relapse occurs in 20% of patients with classical Hodgkin lymphoma (cHL) despite treatment-adaption based on 2-deoxy-2-[18F]fluoro-D-glucose positron emission tomography/computed tomography response. The objective was to evaluate pre-treatment FDG PET/CT-derived machine learning (ML) models for predicting outcome in patients with cHL.

4.1.2 Methods

All cHL patients undergoing pre-treatment PET/CT at our institution between 2008-2018 were retrospectively identified. A 1.5 x mean liver standardised uptake value (SUV) and a fixed 4.0 SUV threshold were used to segment PET/CT data. Feature extraction was performed using PyRadiomics with ComBat harmonisation. Training (80%) and test (20%) cohorts stratified around 2-year event free survival (EFS), age, sex, ethnicity and disease stage were defined. Seven ML models were trained and hyperparameters tuned using stratified 5-fold cross validation. Area under the curve (AUC) from receiver operator characteristic analysis was used to assess performance.

4.1.3 Results

289 patients (153 males), median age 36 (range 16-88 years) were included. There was no significant difference between training (n=231) and test cohorts (n=58) (p-value >0.05). A ridge regression model using a 1.5 x mean liver SUV segmentation had the highest performance, with mean training, validation and test AUCs of 0.82 ± 0.002, 0.79 ± 0.01 and 0.81±0.12. However, there was no significant difference between a logistic model derived from metabolic tumour volume and clinical features or the highest performing radiomic model.

4.1.4 Conclusions

Outcome prediction using pre-treatment FDG PET-CT-derived ML-models is feasible in cHL patients. Further work is needed to determine optimum predictive thresholds for clinical use.

4.2 Introduction

Hodgkin's lymphoma (HL) is a haematopoietic malignancy characterised by the presence of Reed-Sternberg cells [1]. There are five different sub-classes of HL: nodular lymphocyte-predominant HL (NLPHL), and four under the umbrella category of classical HL (cHL): nodular sclerosing, mixed cellularity, lymphocyte-rich and lymphocyte-depleted. Ninety percent of HL cases are cHL [2]. NLPHL is often treated differently to cHL and is associated with more indolent progression [2]. Given the higher proportion of cHL cases, difference in treatment regimens and higher relapse rate in cHL compared to NLPHL, this paper will focus on cHL only [3].

Chemotherapy is the mainstay of frontline treatment of cHL; the most common regimes being doxorubicin (adriamycin), bleomycin, vinblastine and dacarbazine (ABVD), or bleomycin, etoposide, doxorubicin (adriamycin), cyclophosphamide, vincristine (Oncovin), procarbazine, and prednisone (BEACOPP) [4]. The treatment regime and number of cycles can vary depending on patient risk factors, disease stage and initial treatment response. Radiotherapy is used in patients with stage 1 or localised stage 2 disease or in residual bulky disease [4]. The gold standard imaging modality for staging and response assessment in HL is 2-deoxy-2-[18F]fluoro-D-glucose (FDG) positron emission tomography/computed tomography (PET/CT) [5]. Patients typically undergo PET/CT pre-treatment, following two cycles of chemotherapy (interim) and post-treatment. Interim PET/CT is used to guide treatment adaption, balancing the risk of chemotherapy associated toxicity with maximising chances of event free survival (EFS) [6]. 5-year survival in HL is approximately 86% [7]. However, even following complete metabolic response (CMR), approximately 20% of cHL patients will relapse with 72% of relapses occurring within the first 2 years of diagnosis [8]. The ability to identify patients at greater risk of relapse pre-treatment would allow upfront treatment stratification and could improve outcomes.

Previous studies assessing imaging parameters derived from baseline PET/CT for outcome prediction have mainly focused on metabolic tumour volume (MTV), total lesion glycolysis (TLG) and maximum or mean standardised uptake value (SUVmax and SUVmean) [9]. SUV is defined as the ratio of injected radioactivity within an image at a given timepoint when compared to the whole-body [10]. MTV is the volume of metabolically active segmented disease ; with different segmentation techniques described [11]. The TLG is MTV multiplied by the SUVmean. Radiomics transforms images into mineable high-dimensional data permitting invisible feature extraction, analysis and modelling [12]. A limited number of studies using small sample sizes have demonstrated the potential of radiomic features in predicting progression free survival (PFS) or overall survival (OS) in HL patients [13–16]. The aim of this work was to evaluate the performance of models using radiomic features derived from pre-treatment FDG PET/CT to predict 2-year EFS in cHL patients using a larger tertiary centre cohort of patients.

4.3 Methods

This study adhered to the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) guidelines (Supplemental Material 4.1).

4.3.1 Patient selection

Retrospective review of radiology and clinical databases was performed to identify patients who had undergone FDG PET/CT for baseline staging of cHL at our institution between January 2008 and January 2018. This was chosen as the cut-off to allow a minimum of 2-year follow up without confounding factors introduced by the Covid-19 pandemic. Patients were excluded if they were under 16 years of age, did not have cHL, had treatment prior to their staging PET/CT study, did not have measurable disease on PET/CT, had a concurrent malignancy or if the images were degraded or incomplete. Patients who had hepatic disease or had no measurable disease above 4.0 SUV were removed as this would influence the segmentation techniques used.

Patient age, ethnicity, disease stage, date of PET/CT, scanner model and protocol used, type and length of treatment, date of recurrence (confirmed by imaging or clinical examination), last clinical contact and length of follow up or date of death were all recorded from electronic notes, radiological records and from a regional haematological malignancy database. An event was defined as relapse, recurrence or death within the 2-year follow up period. Due to missing clinical data, it was not possible to evaluate scoring systems such as the international prognostic score.

Informed written consent was obtained prospectively from all patients at the time of imaging for use of anonymised images in research and service development projects. As this is a retrospective study, which does not involve patient contact or the alteration of treatment, following discussion with the Research and Innovation Department at LTHT it was agreed that this represented a service improvement project and was approved by the University of Leeds School of Medicine Research Ethics Committee (SoMREC).

4.3.2 PET/CT acquisition

PET/CT studies were performed as part of routine clinical care using a standardised protocol. All patients fasted for 6h prior to administration of intravenous FDG (4MBq/kg). If serum blood glucose was > 10 mmol/L, the study was rescheduled following a clinical review of the patient's diabetic control. Patients were scanned 1 hour following FDG administration. Scans were acquired using a 16-slice Discovery STE PET/CT scanner (GE Healthcare) prior to June 2010; a 64-slice Philips Gemini TF64 scanner (Philips Healthcare) between June 2010 and October 2015; and 64-slice Discovery 690 or 710 scanners (GE Healthcare) after October 2015 (Table 4.1). Attenuation correction was performed using a CT component acquired with the settings: 140 kV; 80mAs; pitch 6; 3.75mm slice thickness.

Scanner	**Voxel Size in mm**	**Matrix**	**Reconstruction**	**Scatter Correction**	**Randoms Correction**
GE Healthcare STE	4.6875 x 4.6875 x 3.27	128	OSEM	Convolution subtraction	Singles
GE Healthcare Discovery 690	3.65 x 3.65 x 3.27	192	VPFX	Model based	Singles
GE Healthcare Discovery 710	3.65 x 3.65 x 3.27	192	VPFX	Model based	Singles
Philips Gemini TF64	4 x 4 x 4	144 or 169	BLOB-OS-TF	SS-Simul	DLYD

Table 4.1 Reconstruction parameters for the scanners used: DLYD – delayed event subtraction; OSEM – ordered subsets expectation maximization; SS-Simul – single-scatter simulation; VPFX – Vue Point FX (3D Time of Flight); BLOB-OS-TF – a 3D ordered subset iterative TOF reconstruction algorithm (spherically symmetric basis function ordered subset)

4.3.3 Image segmentation, feature extraction and machine learning analysis

Image segmentation, feature extraction and machine learning analysis A detailed methodology including detail of who performed the segmentation and interpretation of images is available in Supplemental Material 4.2. Two semi-automated segmentation techniques were used to contour the total lymphomatous disease within each study; the first using a fixed threshold of 4.0 SUV, and the second using a threshold of 1.5 x liver SUVmean. This method has been used in different cancer types [17, 18] (RTx v1.8.2, Mirada Medical). Ten percent of cases were re-segmented using the same methodology following a 3-month washout period using Slicer (v4.11). These re-segmentations were used to investigate the robustness of the extracted radiomic features using different bin widths/bin numbers. Both the CT and PET images were resampled to a uniform voxel of 2mm3. Features were extracted using PyRadiomics (v2.2.0) with 3935 features (PET/CT component x (shape features + first and second order features x number of filters) extracted per segmentation technique for each patient (Supplemental Material 4.2: Table S4.1). Harmonisation to account for the different scanners was applied using the ComBat method (https://github.com/Jfortin1/ComBatHarmonization) [19].

The data was split into training and test cohorts stratified around 2-year EFS (2-EFS), age, sex, ethnicity, disease stage, having radiotherapy, having ABVD-based chemotherapy and being treated as advanced disease using scikit-learn (v0.24.2). The cohorts were split using an 80:20 ratio. Mann-Whitney U and 2 tests (SciPy v1.6.3) were used to assess for significance in continuous and categorical clinical characteristics between the training and test cohorts respectively. A p-value less than 0.05 was regarded as significant. Correlated features were removed if the Pearson coefficient was over 0.8. Seven different machine learning methods were used to create prediction models (scikit-learn v0.24.2): random forest, logistic regression (elastic net, lasso and ridge penalties explored), k-nearest neighbour (KNN), single-layer perceptron (SLP), multi-layer perceptron (MLP), Gaussian process classifier (GCP) and support vector machine (SVM). A maximum number of five features were selected for each of these models. The features selected in each method are based on the highest mean receiver operating characteristic (ROC) area under the curve (AUC) in five-fold stratified cross validation with 20 repeats.

Each model was trained and tuned on the training cohort, using a five-fold cross validation stratified around 2-EFS, again with 20 repeats. The model, hyperparameter and feature selection combination with the highest mean validation score from both the 4.0 SUV and 1.5 x mean liver segmentation were tested once on the unseen test cohort data. Given the growing literature surrounding the use of MTV as an outcome predictor, a separate logistic regression model using total MTV was trained in addition to a model using only clinical features and a combined clinical and MTV model. AUCs were compared using the DeLong method [20]. An appropriate threshold from the ROC curve for each of the best performing models was derived using the Youden index with the Matthews correlation coefficient (MCC), sensitivity, specificity, positive predictive value (PPV) and negative predictive value (PPV) presented.

4.4 Results

4.4.1 Patient demographics

289 patients were included in the study, with the patient demographics detailed in Table 4.2. There were no significant differences in the clinical characteristics between training and test cohorts.

	Training (n=231)	Test (n=58)	p-value
Age (median)	36	41.5	0.10
Sex			0.72
Male	124	29	
Female	107	29	
Ethnicity			0.35
Caucasian	155	37	
Non-Caucasian	26	4	
Not disclosed	50	17	
Stage			0.13
1	14	5	
2	120	20	
3	46	17	
4	51	16	
Chemotherapy			0.11
ABVD/AVD	199	55	
Other	32	13	
Radiotherapy			0.87
No	179	45	
Yes	52	13	
Treated as advanced disease			0.55
No	59	12	
Yes	172	46	
2-year EFS Event			0.99
No	177	45	
Yes	54	13	

	Training (n=231)	**Test (n=58)**	**p-value**

Table 4.2 Demographics of the training and testing groups. 2-EFS = 2-year event free survival. The p-values were calculated using a t-test for age and a 2 test for the remaining demographic features.

4.4.2 Bin widths

For both the 4.0 SUV and 1.5 x mean liver SUV segmentation techniques bin widths for PET and CT data were most robust when derived from the maximum range of SUV or HU respectively divided by 128 (Supplementary Material 4.2: Figure S4.1 and S4.2). Overall, the 4.0 SUV segmentation technique resulted in more radiomic features being robust than the 1.5 x mean liver SUV segmentation method.

4.4.3 Clinical and MTV derived models of 2-EFS

Patients who had a 2-EFS event had a significantly larger MTV compared to those who did not have a 2-EFS event. This was true for both segmentation techniques. With the 4.0 SUV method, the median MTVs were 167.4cm^3 versus 87.9cm^3 (p=0.03); for the 1.5 x mean liver SUV method, 324.3cm^3 versus 148.6cm^3 (p=0.009). The median volumes were significantly greater in patients treated as advanced disease. For the 4.0 SUV method, the median MTVs were 250.6cm^3 (2-EFS event) versus 110.4cm^3 (no event) (p=0.03); for the 1.5 x mean liver SUV method, 457.8cm^3 (2-EFS event) versus 227.9cm^3 (no event) (p=0.02).

A logistic regression model using MTV derived from a 4.0 SUV method resulted in a mean training AUC of 0.61 ± 0.02 (mean ± 95% CI) and a mean validation AUC of 0.61 ± 0.10 with the odds ratio being 1.00038 (Table 4.3). The logistic regression model derived from MTV using the 1.5 x mean liver SUV method had a mean training AUC of 0.63 ± 0.02 and a mean validation AUC of 0.63 ± 0.10, with the odds ratio being 1.00038.

Model	**Features**	**Hyperparameters**	**Mean Training**	**Mean Validation**
Logistic Regression – Clinical	Cancer stage 1, Cancer stage 4, Age	C: 10, Penalty: l2, Solver: newton-cg	0.74 ± 0.004	0.74 ± 0.02
Logistic Regression – MTV (1.5 x mean liver SUV)	MTV	C: 1e-07, Penalty: l2, Solver: liblinear	0.63 ± 0.02	0.63 ± 0.10
Logistic Regression – MTV (4.0 SUV)	MTV	C: 1e-07, Penalty: l2, Solver: liblinear	0.62 ± 0.02	0.61 ± 0.10
Logistic Regression – Clinical and MTV (1.5 x mean liver SUV)	Cancer stage 1, Cancer stage 4, Age, MTV	C: 1, Penalty: l2, Solver: saga	0.75 ± 0.004	0.74 ± 0.02

Table 4.3 Mean training and validation scores for the best performing clinical and metabolic tumour volume (MTV) based logistic regression models. l2 = Ridge regression penalty, liblinear = A library for large linear classification.

Cancer stage 1, cancer stage 4 and age were selected as features for the clinical based logistic regression model. This had a mean training AUC of 74 ± 0.004 and a mean validation AUC of 0.74 ± 0.02. When combing the features from this model with 1.5 x times mean liver SUV MTV the model had a mean training AUC of 0.74 ± 0.004 and a mean validation AUC of 0.72 ± 0.01. This model was tested on the unseen test set and achieved an AUC of 0.68 ± 0.11 (Figure 4.1), MCC of 0.27, sensitivity of 0.31, specificity of 0.91, NPV of 0.47 and PPV of 0.85 at a threshold of 0.45.

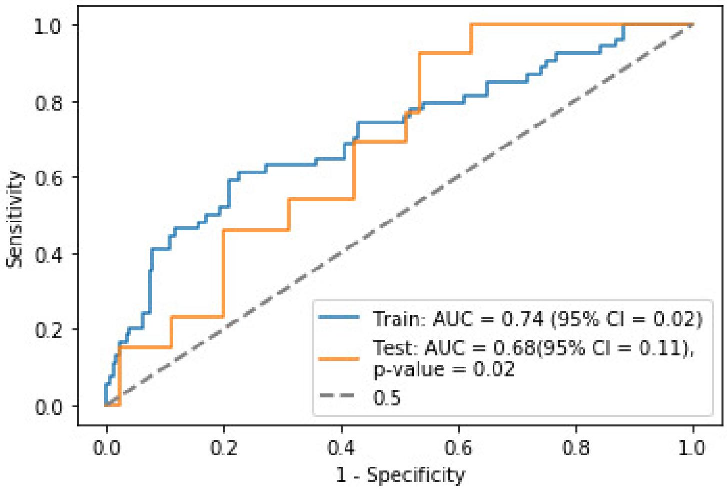


Figure 4.1 Receiver operator characteristic curve for the best performing predictive model derived from a logistic regression using MTV extracted from a 1.5 x mean liver SUV threshold segmentation technique and clinical features. The p-value represents the comparison of the ROC to that of the 0.5 curve.

4.4.4 Clinical and radiomic model for the prediction of 2-EFS

The predictive model with the highest AUC was a ridge regression model derived from clinical and radiomic features extracted from 1.5 x mean SUV threshold segmentation technique (Table 4.4). The model was constructed using features selected using a forward wrapper with five features chosen. The hyperparameters of the model were: C=1, penalty=l2 and solver=sag, class weight=balanced. The features chosen were age, PET flatness, PET major axis length, PET logarithm GLSZM size zone non-uniformity normalized, PET lbp-3D-m1 GLCM correlation and PET lbp-3D-m2 first order skewness. The mean training AUC was 0.82 ± 0.002, the mean validation AUC was 0.79 ± 0.01 and the test AUC was 0.81 ± 0.12 (Figure 4.2), with MCC=0.43, sensitivity=0.42, specificity=0.94, NPV=0.67, PPV=0.85. The demographics of the mislabelled patients are presented in Supplementary Material 2: Table S4.2.

Model	Features	Hyperparameters	Mean Training	Mean Validation
4.0 SUV				
Support Vector Machine	Age, PET GLCM Imc1, PET wavelet-LLH GLCM Imc2, PET wavelet-HLL GLSZM small area emphasis, PET log-sigma-2-0-mm-3D GLSZM small area emphasis	C: 15.78, Gamma: 0.000794, Kernel: sigmoid	0.68 ± 0.004	0.66 ± 0.02
Logistic Regression	Age, PET least axis length, PET wavelet-HLL GLCM correlation, PET wavelet-HLH GLCM Idmn, CT wavelet-HLL GLSZM large area low gray level emphasis	C: 1, Penalty: l2, Solver: lbfgs	0.80 ± 0.002	0.78 ± 0.01
Random Forest	Age	Bootstrap: True, Max depth: 1, Min samples per leaf: 11, Min samples per split: 32, Number of estimators: 213	0.67 ± 0.004	0.64 ± 0.02

Model	Features	Hyperparameters	Mean Training	Mean Validation
Multi-layer perceptron	Age, PET major axis length, PET wavelet-HHL GLCM Imc1, PET lbp-3D-k first order 10th percentile	Learning rate: invscaling, Solver: sgd	0.68 ± 0.004	0.68 ± 0.02
1.5 x mean liver SUV				
Support Vector Machine	PET first order 90th percentile, PET wavelet-LHH GLDM dependence non-uniformity normalized	C: 3.398, Gamma: 0.1005, Kernel: sigmoid	0.54 ± 0.008	0.55 ± 0.02
Logistic Regression	Age, PET flatness, PET major axis length, PET logarithm GLSZM size zone non-uniformity normalized, PET lbp-3D-m1 GLCM correlation, PET lbp-3D-m2 first order skewness	C: 1, Penalty: l2, Solver: sag	0.82 ± 0.002	0.79 ± 0.01

Model	**Features**	**Hyperparameters**	**Mean Training**	**Mean Validation**
Random Forest	Age	Bootstrap: True, Max depth: 1, Min samples per leaf: 11, Min samples per split: 48, Number of estimators: 213	0.67 ± 0.004	0.64 ± 0.02
Multi-layer perceptron	Age, PET flatness, PET major axis length	Learning rate: invscaling, Solver: adam	0.77 ± 0.004	0.75 ± 0.01

Table 4.4 Mean training and validation scores for the best performing machine learning models using a fixed threshold of 4.0SUV and 1.5 x mean liver SUV thresholding segmentation techniques. The K-nearest neighbours, single layer perceptron and Gaussian process classifier models were over-fitted with the mean training and validation AUCs with >0.10 difference between the two. l2 = Ridge regression penalty, liblinear = A library for large linear classification, GLSZM = grey level size zone matrix, GLCM = grey level co-occurrence matrix, GLDM = grey level dependence matrix, rbf = radial basis function, L = low, H = high, Imc1 = informational measure of correlation 1, Imc2 = informational measure of correlation 2, idmn = inverse difference moment normalized, lbp = local binary pattern.

The highest performing predictive model using the 4.0 SUV threshold was a regression model using a ridge regression penalty with a mean training AUC of 0.79 ± 0.002, the mean validation AUC was 0.77 ± 0.01 and the test AUC was 0.74 ± 0.13 (Figure 4.3). The MCC=0.30, sensitivity=032, specificity=0.95, NPV=0.42, PPV=0.92 at a threshold of 0.27. The model was constructed using features selected from a forward wrapper method of feature selection with five features chosen. The hyperparameters of the model were: C=100, penalty=l2 and solver=saga, class weight=balanced.

There was no significant difference between the test set AUCs of the best performing clinical and radiomic based models with each other and with the best performing clinical and MTV based model (Figure 4.4 and Table 4.5). The intercept and coefficients for each model are presented in Supplementary Material 2: Table S4.3.

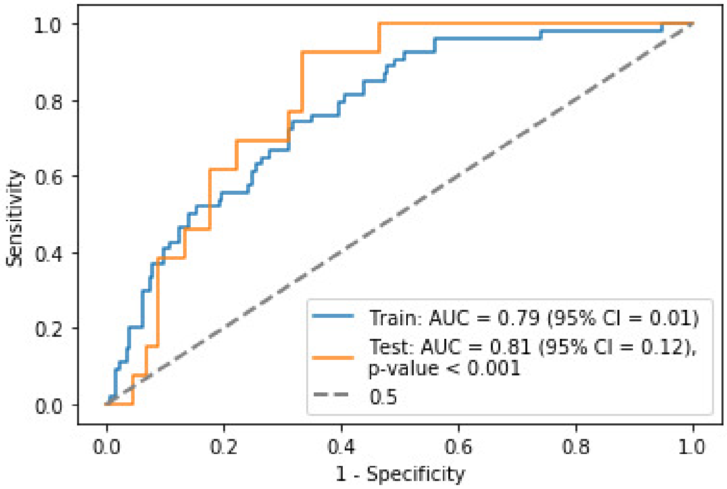


Figure 4.2 Receiver operator characteristic curve for the best performing predictive model derived from ridge regression using age and radiomic features extracted from a 1.5 x mean liver SUV threshold segmentation technique. The p-value represents the comparison of the ROC to that of the 0.5 curve.

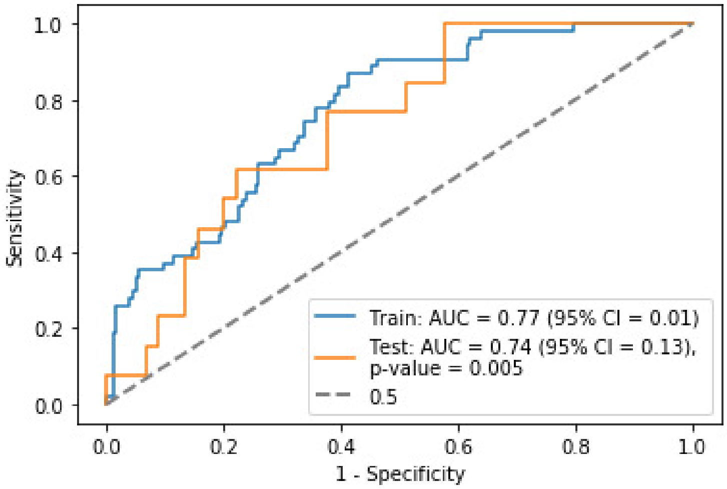


Figure 4.3 Receiver operator characteristic curve for the best performing predictive model derived from ridge regression using age and radiomic features extracted using a 4.0 SUV fixed threshold segmentation technique. The p-value represents the comparison of the ROC to that of the 0.5 curve.

4.5 Discussion

This study confirms that pre-treatment outcome prediction using FDG PET/CT derived radiomic features is feasible in patients with cHL. The best performing model was created using ridge regression combining age and four radiomic features (PET flatness, PET major

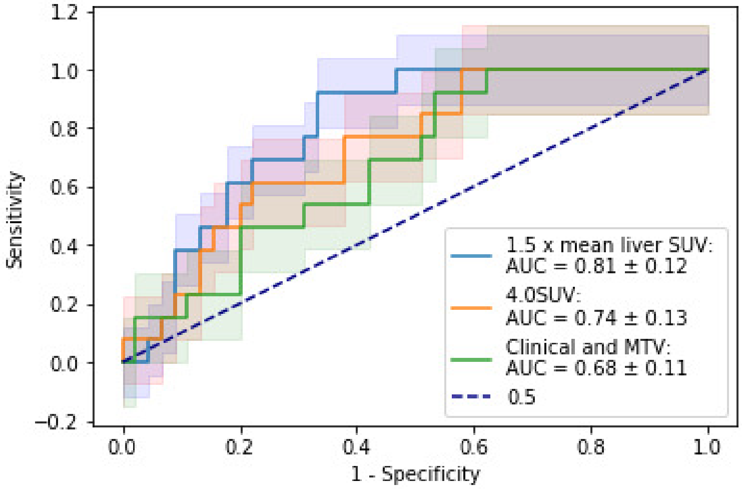


Figure 4.4 Receiver operator curves, with associated confidence intervals, for the best performing MTV and radiomic models derived from 4.0 SUV fixed threshold and 1.5 x mean liver SUV threshold segmentation techniques.

	Clinical and MTV	**1.5 x mean liver SUV**	**4.0 SUV**
Clinical and MTV	n/a	0.11	0.53
1.5 x mean liver SUV	0.11	n/a	0.22
4.0 SUV	0.53	0.22	n/a

Table 4.5 Comparison of the different test AUCs using the DeLong method, p-values presented.

axis length, PET logarithm GLSZM size zone non-uniformity normalized, PET lbp-3D-m1 GLCM correlation and PET lbp-3D-m2 first order skewness) extracted from PET images using a 1.5 x mean liver SUV method with a bin width of 0.24. It must be noted that there was no significant difference between the test AUC of this model and those of a combined clinical and MTV model and a model created using 4.0 SUV fixed threshold segmentation. This is likely due to small numbers involved given the relatively large confidence intervals. Due to missing clinical data, it was not possible to adjust for features used to stratify patients into early and advanced disease. A surrogate, treatment intent, was used instead which demonstrated that the models created remained reasonable predictors of outcome for patients treated as having advanced disease.

Further work should be performed to assess the relationship of ethnicity and socio-economic status on a model's predictive ability to avoid creation of a model which discriminates against under-represented subsets of patients due to lack of data to train and test the model on [21, 22]. Unadjusted confounders are likely one of the reasons for a minority of studies reporting the poor ability of MTV as an outcome predictor in lymphoma [23][24][25]. Most notably Adams *et al.* found that MTV was not an independent predictor of overall survival or PFS in diffuse large B-cell lymphoma once adjusting for the National Comprehensive Cancer Network International Prognostic Index [26]. To allow for transparency our study has provided the demographic information for the patients who were mislabelled using the predictive model with the highest test AUC.

Two different segmentation techniques were explored. The first was a fixed threshold of 4.0 SUV which has been demonstrated to be a reproducible, efficient method for contouring disease [27]. The second was 1.5 x mean liver SUV which has been explored in other malignancies and provides an adaptive threshold which adjusts for background SUV uptake [17, 18]. Our study echoed previous work demonstrating a fixed threshold led to more features being robust following re-segmentation. The fixed thresholding segmentation technique required less steps, and less manual adaption [28]. However, a fixed SUV thresholding technique does not scale with the physiological uptake and therefore the contours may vary on repeat studies due to external effects on the SUV rather than tumour pathophysiology [27]. The study also demonstrated the variability which can occur when repeating a segmentation methodology on different software (Figure 4.5), with radiomic features not being deemed robust following repeated segmentation even when using the same SUV thresholds. ComBat harmonisation was employed to mitigate against the effects of scanner variation. This is based on Bayes theorem and attempts to predict scanner influence whilst maintaining biological variation [29]. For this to be effective however, there must be enough samples from different scanners to apply the harmonisation method [30] and it cannot be applied prospectively to scanner acquisitions outside those used for training of the predictive model.

Previous studies have explored the use of radiomic features in the prediction of outcomes in HL [13, 14][15]. Lue *et al.* found SUV kurtosis, stage and intensity non-uniformity (INU) derived from Grey-Level Run Length Matrix (GLRLM) were independent predictors of PFS in a small cohort of 42 patients. Milogrom *et al.* demonstrated that the combination of SUVmax, MTV, InformationMeasureCorr1, InformationMeasureCorr2, and InverseVariance derived from GLCM 2.5 had an AUC of 0.95 when predicting relapse in 167 patients with stage I-II HL. However, there were very few events, with the validation cohort only having two patients who relapsed. Sollini *et al.* assessed a radiomic fingerprint using principal component analysis to classify patients who would relapse within 4 years of treatment in a cohort of 85 patients. They explored fingerprints created from a single largest nodal or extra-nodal

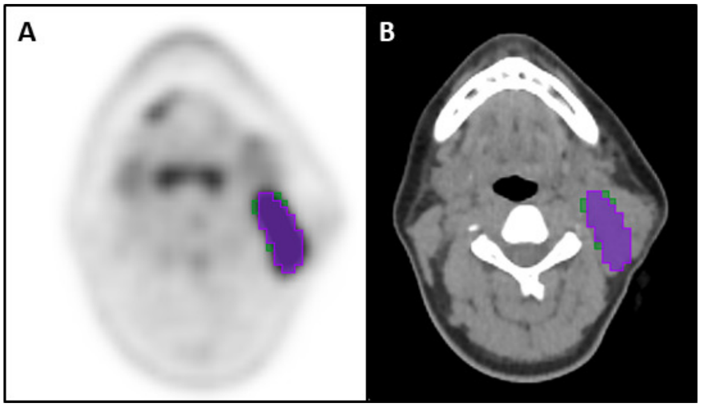


Figure 4.5 Select axial slice through PET (A) and CT (B) images of a patient with Hodgkin lymphoma demonstrating a pathological left level II lymph node. The purple segmentation represents the original 4.0 SUV fixed threshold segmentation performed using Mirada Medical RTx (v1.8.2) software and the green areas represent the additional area included when segmented with a fixed 4.0 SUV threshold using 3D Slicer (v4.11) software.

lesion versus using all lesions and found that the intra-patient similarity was low, and that the highest accuracy was achieved when using all lesions within the model [15]. This highlights the inherent heterogeneity of radiomic features within different lesions and that by restricting analysis to a single lesion, the predictive model may also be limited. The current study of 289 patients is one of the largest to assess potential utility of radiomic features derived from pre-treatment FDG PET/CT for predicting outcome in cHL patients. It demonstrates that radiomics could feasibly improve prediction of 2-EFS. However, this requires validation on an independent external dataset and although the AUC for the test set was 0.81, no clear predictive threshold could be derived. This must be a key target when creating any machine learning or AI based model. In terms of HL, it would be the ability to balance side effects of escalated treatment, with the rates of EFS and toxicity vary between treatment regimens [31]. The advent of newer therapeutic strategies limits the use of predictive models made on retrospective data; future efforts should focus on validating imaging, genetic and clinical predictive features in carefully designed prospective, multi-centre clinical trials.

A TRIPOD checklist was used to ensure transparency of the study's methodology; a concern in previous radiomic studies [32, 33]. However, no external validation was performed, and although contouring was undertaken without knowledge of clinical outcome, no measures to blind assessors was specifically undertaken. Although patients with other concurrent malignancies were excluded from analysis, other pathologies were not taken into consideration when looking at mortality. Other study limitations include its retrospective nature, the relatively small event rate, reliance on clinical records to determine date of relapse/recurrence, exclusion of patients with hepatic disease/or without disease $>$4.0 SUV and variation in different patient's treatment regimen.

4.6 Conclusion

There is potential for models derived from radiomic features extracted from pre-treatment FDG PET/CT to predict 2-EFS in cHL patients. Further work is needed to determine optimum thresholds for clinical use.

4.7 References

1. Connors JM, Cozen W, Steidl C, *et al.* (2020) Hodgkin lymphoma. Nat Rev Dis Prim 6:. https://doi.org/10.1038/s41572-020-0189-6

2. Shanbhag S, Ambinder RF (2018) Hodgkin lymphoma: A review and update on recent progress. CA Cancer J Clin 68:116–132.

3. Spinner MA, Varma G, Advani RH (2019) Modern principles in the management of nodular lymphocyte-predominant Hodgkin lymphoma. Br J Haematol 184:17–29.

4. Follows GA, Ardeshna KM, Barrington SF, *et al.* (2014) Guidelines for the first line management of classical Hodgkin lymphoma. Br J Haematol 166:34–49.

5. Kanoun S, Rossi C, Casasnovas O (2018) [18F]FDG-PET/CT in hodgkin lymphoma: Current usefulness and perspectives. Cancers (Basel) 10:1–11.

6. El-Galaly TC, Villa D, Gormsen LC, *et al.* (2018) FDG-PET/CT in the management of lymphomas: current status and future directions. J Intern Med 284:358–376.

7. SEER (2018) SEER Cancer Statistics Review. In: 1975-2016

8. Hapgood G, Zheng Y, Sehn LH, *et al.* (2016) Evaluation of the risk of relapse in classical hodgkin lymphoma at event-free survival time points and survival comparison with the general population in British Columbia. J Clin Oncol 34:2493–2500.

9. Frood R, Burton C, Tsoumpas C, *et al.* (2021) Baseline PET/CT imaging parameters for prediction of treatment outcome in Hodgkin and diffuse large B cell lymphoma: a systematic review. Eur J Nucl Med Mol Imaging. 48(10):3198-3220 10. Fletcher JW, Kinahan PE (2011) PET/CT SUVs in clinical practice. 31:496–505.

10. Im HJ, Bradshaw T, Solaiyappan M, Cho SY (2018) Current Methods to Define Metabolic Tumor Volume in Positron Emission Tomography: Which One is Better? Nucl Med Mol Imaging (2010) 52:5–15.

11. Lambin P, Leijenaar RTH, Deist TM, *et al.* (2017) Radiomics: The bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol 14:749–762. https://doi.org/10.1038/nrclinonc.2017.141

12. Milgrom SA, Elhalawani H, Lee J, *et al.* (2019) A PET Radiomics Model to Predict Refractory Mediastinal Hodgkin Lymphoma. Sci Rep 9:1–8. https://doi.org/10.1038/s41598-018-37197-z

13. Lue KH, Wu YF, Liu SH, *et al.* (2019) Prognostic Value of Pretreatment Radiomic Features of 18F-FDG PET in Patients with Hodgkin Lymphoma. Clin Nucl Med 44:E559–E565.

14. Sollini M, Kirienko M, Cavinato L, *et al.* (2020) Methodological framework for radiomics applications in Hodgkin's lymphoma. Eur J Hybrid Imaging 4:. https://doi.org/10.1186/s41824-020-00078-8

15. Zhou Y, Zhu Y, Chen Z, *et al.* (2021) Radiomic Features of 18F-FDG PET in Hodgkin Lymphoma Are Predictive of Outcomes. Contrast Media Mol Imaging 2021:. https://doi.org/10.1155/2021/6347404

16. Brown PJ, Zhong J, Frood R, *et al.* (2019) Prediction of outcome in anal squamous cell carcinoma using radiomic feature analysis of pre-treatment FDG PET-CT. Eur J Nucl Med Mol Imaging 46:2790–2799. https://doi.org/10.1007/s00259-019-04495-1

17. Zhong J, Frood R, Brown P, *et al.* (2021) Machine learning-based FDG PET-CT radiomics for outcome prediction in larynx and hypopharynx squamous cell carcinoma. Clin Radiol 76:78.e9-78.e17. https://doi.org/10.1016/j.crad.2020.08.030

18. Fortin J-P, Cullen N, Sheline YI, *et al.* (2018) Harmonization of cortical thickness measurements across scanners and sites. Neuroimage 167:104–120. https://doi.org/10.1016/j.neuroimage.2017.11.024

19. DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44:837–45

20. Berkman AM, Andersen CR, Puthenpura V, *et al.* (2021) Impact of race, ethnicity, and socioeconomic status over time on the long-term survival of adolescent and young adult hodgkin lymphoma survivors. Cancer Epidemiol Biomarkers Prev 30:1717–1725. https://doi.org/10.1158/1055-9965.EPI-21-0103

21. Pahwa P, Karunanayake CP, Spinelli JJ, *et al.* (2009) Ethnicity and incidence of Hodgkin lymphoma in Canadian population. BMC Cancer 9:1–9. https://doi.org/10.1186/1471-2407-9-141

22. Mettler J, Müller H, Voltin CA, *et al.* (2019) Metabolic tumor volume for response prediction in advanced-stage hodgkin lymphoma. J Nucl Med 60:207–211. https://doi.org/10.2967/jnumed.118.210047

23. Tseng D, Rachakonda LP, Su Z, *et al.* (2012) Interim-treatment quantitative PET parameters predict progression and death among patients with hodgkin's disease. Radiat Oncol 7:. https://doi.org/10.1186/1748-717X-7-5

24. Gallicchio R, Mansueto G, Simeon V, *et al.* (2014) F-18 FDG PET/CT quantization parameters as predictors of outcome in patients with diffuse large B-cell lymphoma. Eur J Haematol 92:382–389. https://doi.org/10.1111/ejh.12268

25. Adams HJA, de Klerk JMH, Fijnheer R, *et al.* (2015) Prognostic superiority of the National Comprehensive Cancer Network International Prognostic Index over pretreatment whole-body volumetric-metabolic FDG-PET/CT metrics in diffuse large B-cell lymphoma. Eur J Haematol 94:532–539. https://doi.org/10.1111/ejh.12467

26. Barrington SF, Zwezerijnen BGJC, de Vet HCW, *et al.* (2021) Automated Segmentation of Baseline Metabolic Total Tumor Burden in Diffuse Large B-Cell Lymphoma: Which Method Is Most Successful? A Study on Behalf of the PETRA Consortium. J Nucl Med 62:332–337. https://doi.org/10.2967/jnumed.119.238923

27. Driessen J, Zwezerijnen GJ, Schöder H, *et al.* (2022) The impact of semi-automatic segmentation methods on metabolic tumor volume, intensity and dissemination radiomics in 18 F-FDG PET scans of patients with classical Hodgkin lymphoma. J Nucl Med jnumed.121.263067. https://doi.org/10.2967/jnumed.121.263067

28. Orlhac F, Boughdad S, Philippe C, *et al.* (2018) A postreconstruction harmonization method for multicenter radiomic studies in PET. J Nucl Med 59:1321–1328. https://doi.org/10.2967/jnumed.117.199935

29. Orlhac F, Eertink JJ, Cottereau AS, *et al.* (2022) A Guide to ComBat Harmonization of Imaging Biomarkers in Multicenter Studies. J Nucl Med 63:172–179. https://doi.org/10.2967/jnumed.121.262464

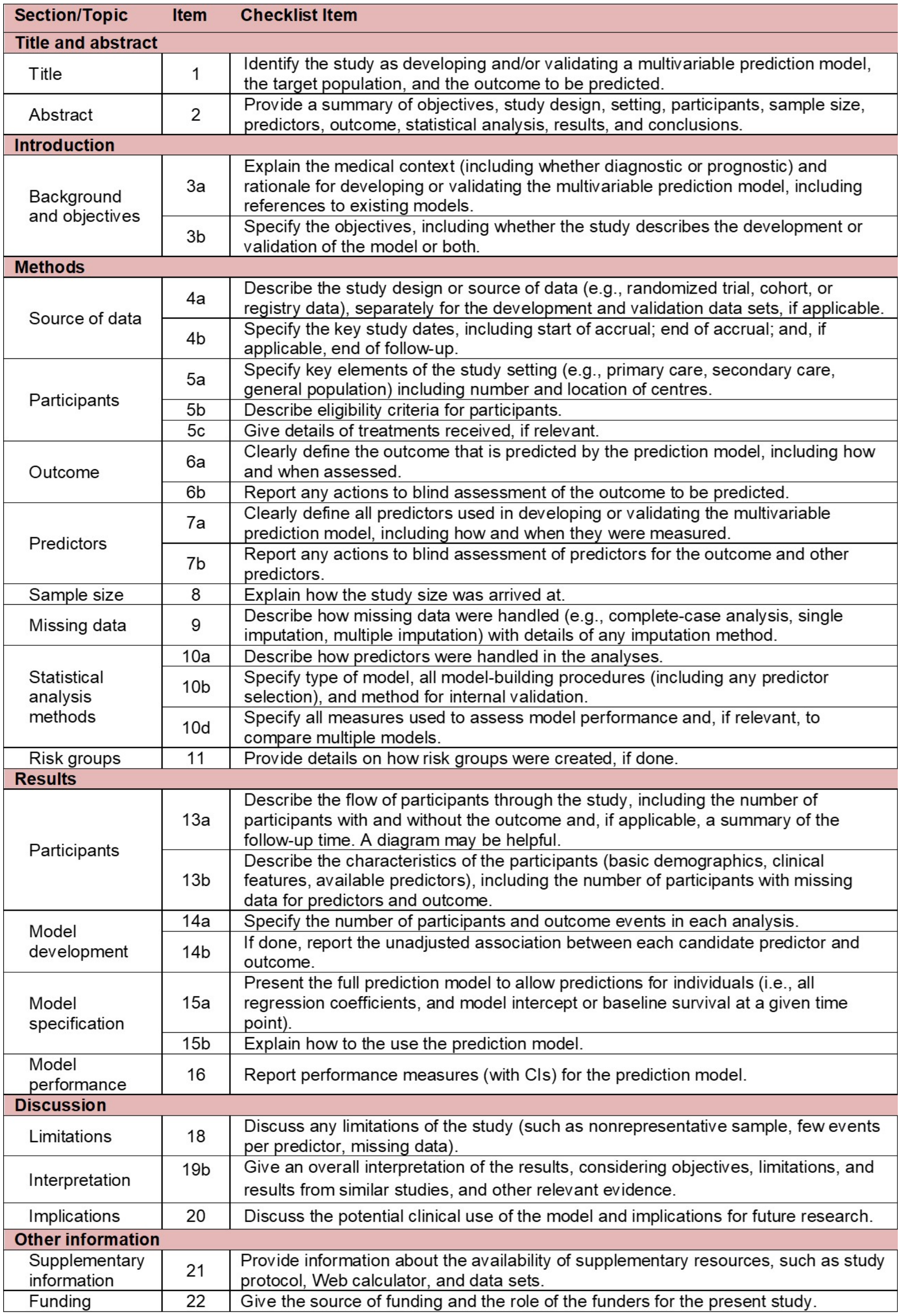
30. Lang N, Crump M (2020) PET-adapted approaches to primary therapy for advanced Hodgkin lymphoma. Ther Adv Hematol 11:204062072091449. https://doi.org/10.1177/2040620720914490

31. Park JE, Kim D, Kim HS, *et al.* (2020) Quality of science and reporting of radiomics in oncologic studies: room for improvement according to radiomics quality score and TRIPOD statement. Eur Radiol 30:523–536. https://doi.org/10.1007/s00330-019-06360-z

32. Pinto dos Santos D, Dietzel M, Baessler B (2021) A decade of radiomics research: are images really data or just patterns in the noise? Eur Radiol 31:2–5. https://doi.org/10.1007/s00330-020-07108-w

4.8 Supplementary Material

4.8.1 Supplemental Material 4.1



Section/Topic	Item	Checklist Item
Title and abstract		
Title	1	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.
Abstract	2	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.
Introduction		
Background and objectives	3a	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.
	3b	Specify the objectives, including whether the study describes the development or validation of the model or both.
Methods		
Source of data	4a	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.
	4b	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.
Participants	5a	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.
	5b	Describe eligibility criteria for participants.
	5c	Give details of treatments received, if relevant.
Outcome	6a	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.
	6b	Report any actions to blind assessment of the outcome to be predicted.
Predictors	7a	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.
	7b	Report any actions to blind assessment of predictors for the outcome and other predictors.
Sample size	8	Explain how the study size was arrived at.
Missing data	9	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.
Statistical analysis methods	10a	Describe how predictors were handled in the analyses.
	10b	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.
	10d	Specify all measures used to assess model performance and, if relevant, to compare multiple models.
Risk groups	11	Provide details on how risk groups were created, if done.
Results		
Participants	13a	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.
	13b	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.
Model development	14a	Specify the number of participants and outcome events in each analysis.
	14b	If done, report the unadjusted association between each candidate predictor and outcome.
Model specification	15a	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).
	15b	Explain how to the use the prediction model.
Model performance	16	Report performance measures (with CIs) for the prediction model.
Discussion		
Limitations	18	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).
Interpretation	19b	Give an overall interpretation of the results, considering objectives, limitations, and results from similar studies, and other relevant evidence.
Implications	20	Discuss the potential clinical use of the model and implications for future research.
Other information		
Supplementary information	21	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.
Funding	22	Give the source of funding and the role of the funders for the present study.

1. This study looks at the utility of pre-treatment FDG PET/CT derived machine learning models for outcome prediction in classical Hodgkin lymphoma. (Title)

2. The abstract covers a summary of all the requested information. (Abstract)

3. a) The introduction presents the background of HL sets out the aim of creating a predictive model using machine learning techniques using radiomic features derived from the baseline PET/CT. Two previous papers are discussed which aim to create a similar model. (Introduction) b) The aim of this study was to create a predictive model using radiomic features derived from pre-treatment FDG PET/CT to predict 2-year EFS in HL patients using a larger tertiary centre cohort of patients (Introduction)

4. a) This is a retrospective single centre cohort study. The study cohort was randomised on a ratio of 4:1 into training and testing cohorts stratified around 2-EFS, age, gender, ethnicity and disease stage. (Patient selection)

b) Consecutive patients with histologically proven cHL who underwent baseline FDG-PET/CT at a single large tertiary referral centre between June 2008 and January 2018 were included. The follow up information recorded is set out in the patient selection section. (Patient selection)

5. a) This is a single tertiary centre study. (Patient selection) b) Patients were excluded if they were under 16 years of age, did not have cHL, had treatment prior to their staging PET/CT study, did not have measurable disease on PET/CT, had a concurrent malignancy, they did not have disease over 4.0SUV, had hepatic involvement or if the images were degraded or incomplete. The follow up information recorded is set out (Patient selection) c) The treatment regimen for the cohort is set out in Table 4.2. No change to departmental standard treatment was performed. (Table 4.2)

6. a) An event was defined as relapse, recurrence or death within the 2 year follow up period. (Patient selection) b) As this was a retrospective study the primary outcomes were defined from clinical records. The investigator reviewing the records was blinded to the imaging parameters.

7. a) The description of the contouring method, resampling, harmonisation, radiomic feature extraction and the methods used for feature selection are documented within the method section and Supplementary Material 4.2. The features selected as part of the models are described in Table 4.3. (Materials and methods, Supplementary Material 4.2, Results) b) The images were contoured and analysed without reference to the outcome data.

8. All patients which met the inclusion criteria were included. The cut-off of January 2018 was chosen to allow for 2 year follow up without confounding factors introduced due

to the covid-19 pandemic. For feature selection 5 features were chosen as the maximum number of features to be include in each model. This was derived from 10 events per parameter, with 54 events within the training cohort. (Materials and methods, Results)

9. Only complete data sets were used in the analysis. (Results)

10. a) Clinical factors were included in the variable selection process alongside radiomic features. The categorical data was dummy encoded. Continuous features were normalised using a standard scaler. (Machine learning analysis, Supplementary Material 4.2) b) Random forest, support vector machine, logistic regression, k-nearest neighbour, single layer perceptron, multi-layer perceptron and Gaussian process classifier models were trained and tuned on the training cohort using cross validation. The models were created using different feature selection methods, the bin width or bin number was selected based on the method which had the greatest robust features (intraclass correlation coefficient >0.8) following regimentation. A model was generated using radiomic features from a fixed 4.0 SUV threshold segmentation technique and a 1.5 x mean liver SUV threshold segmentation technique. A model was also created using metabolic tumour volume. The models with the highest mean receiver operating characteristic (ROC) area under the curve (AUC) were tested and compared on the unseen test cohort. (Machine learning analysis, Supplementary Material 4.2) d) When comparing models, the mean validation AUC was used to determine the best performing model. A Delong test was used to compare the AUCs of the test set. (Machine learning analysis, Supplementary Material 4.2)

11. Risk groups were not created within the model.

13. a) 289 patients were included, with demographics detailed in Table 4.2. (Results) b) The characteristics of the participants are presented in Table 4.2.

14. a) The number of events per cohort are presented in Table 4.2. b) This has not been performed. The training and testing cohorts were stratified around key clinical features, but the results are not adjusted for these. Further analysis was performed looking at how the model performed on patients treated as having advanced disease.

15. a/b) The features and hyperparameters used to create the model are presented in the Clinical and radiomic model for the prediction of 2-EFS section.

16. The mean validation and test ROC curves are presented. The 95% confidence intervals are presented. (Results)

18. The limitations of the study are presented. These include the retrospective nature of the study, the relative low number of events, reliance on clinical records, the exclusion of patients with hepatic disease or disease not meeting the 4.0 SUV threshold, variation in patient treatment and that there was no external validation. (Discussion)

19. b)/20. The discussion section gives an overall interpretation of the results and highlights the potential use of a pre-treatment model to aid in early personalised treatment for patients. (Discussion)

21. The python libraries used are references within the text. The radiomic features extracted using PyRadiomics are detailed in Supplementary Material 2: Table S4.1.

22. The study was not externally funded. Individual author's funding is declared within the Declaration.

4.8.2 Supplemental Material 4.2

4.8.2.1 Image segmentation

Image data were viewed and contoured using specialised multimodality imaging software (RTx v1.8.2, Mirada Medical). Lymphomatous disease segmentation was performed by a clinical radiologist with six years' experience and a research radiographer with 2 years' experience of segmenting cross-sectional imaging and reviewed by two dual-certified Radiology and Nuclear Medicine Physicians with >15 years' experience of oncological PET/CT interpretation. Any discrepancies were agreed in consensus. Two segmentation techniques were utilised, the first using a fixed threshold of 4.0 SUV and the second using a threshold of 1.5 x liver SUVmean was used to contour disease sites on PET, this method has been used in different cancer types [16, 17]. The mean liver SUV was determined by placing a 110 cm^3 region of interest in the right lobe of the liver. The contour from the PET was translated to the co-registered unenhanced low-dose CT component of the study with the contours matched to soft tissue with a value of -10 to 100 Hounsfield units (HU). Contours were exported as digital imaging and communications in medicine (DICOM) radiotherapy (RT) structures. Ten percent of the cases were re-segmented using the same methodology described by the radiologist who performed the initial segmentation after a 3-month washout period using Slicer (v4.11). These segmentations were used to test the repeatability of the segmentation techniques and to test the robustness of the extracted features.

4.8.2.2 Feature extraction

DICOM images and DICOM-RT structures were converted to Neuroimaging Informatics Technology Initiative (NIfTI) files using the python library Simple ITK (v2.0.2). Absolute PET voxel values were converted to body weight SUV and voxel values for CT were converted to HU using the equations detailed below

$$ActivityConcentration\left(\frac{Bq}{ml}\right) = PixelValue * Slope + Intercept$$

$$CorrectionFactor = 2\left(-\left(\frac{ScanTime\,(s) - MeasuredTime\,(s)}{HalfLife\,(s)}\right)\right)$$

$$SUVbw = \frac{AcitivityConcentration\left(\frac{Bq}{ml}\right) * BodyWeight\,(g)}{TotalDose * CorrectionFactor}$$

$$HounsfieldUnits = PixelValue * Slope + Intercept$$

Both CT and PET data were resampled to a uniform voxel size of 2 mm3. The robustness of radiomic features to re-segmentation using different software was used to identify the optimum bin width for the dataset. Radiomic features were extracted using a fixed bin number of 32, 64 and 128, and bin widths derived from either dividing the maximum or median voxel range by 32, 64 and 128. Features were deemed to be robust if the intraclass correlation coefficient (ICC) calculated using the python library pingouin (v0.3.12) was $>$0.8. First and second order parameters were extracted using PyRadiomics (v2.2.0). There are some deviations between PyRadiomics and the image biomarker standardisation initiative (IBSI), with Pyradiomics starting the fixed bin width from 0 and not the minimum segmentation value, and the calculation of first order kurtosis being +3 larger in PyRadiomics [18, 19]. Patient age, histology and sex were also included as clinical features in the models. Disease stage and sex were dummy encoded using (Pandas v1.2.4). This resulted in a total of 3935 features extracted per segmentation technique for each patient (Table S4.1). Harmonisation to account for the different scanners was applied to the radiomic features using the ComBat method (https://github.com/Jfortin1/ComBatHarmonization) [20].

4.8.2.3 Machine learning analysis

The study cohort was split into training and test cohorts stratified around 2-year EFS (2-EFS), age, sex, ethnicity, stage of disease, having radiotherapy, having ABVD-based chemotherapy and being treated as advanced disease using scikit-learn (v0.24.2). Ethnicity was defined by the volunteered information from patients. Given the low numbers of some ethnic groups, it was not possible to stratify the training and tests around ethnicity without splitting the data into Caucasian and non-Caucasian ethnic groups. The cohorts were split using an 80:20 ratio. Mann-Whitney U and 2 tests (SciPy v1.6.3) were used to assess for significance in continuous and categorical clinical characteristics between the training and test cohorts respectively. A p-value less than 0.05 was regarded as significant. Categorical data was dummy encoded (Pandas v1.2.4), and continuous data was normalised using a standard scaler (scikit-learn v0.24.2). Correlated features were removed if the Pearson coefficient was over 0.8. Seven different machine learning methods were used to create prediction models (scikit-learn v0.24.2): random forest, logistic regression (elastic net, lasso and ridge penalties explored), k-nearest neighbour (KNN), single layer perceptron (SLP), multi-layer perceptron

(MLP), Gaussian process classifier (GCP) and support vector machine (SVM). A maximum number of five features was selected for the model and this was based on one feature per 10 events. Three feature selection methods were used: a forward wrapper method (mlxtend 0.18.0), a univariate analysis method (scikit-learn v0.24.2), and a recursive feature extraction method (for the models where this was applicable i.e. random forest and logistic regression) (scikit-learn v0.24.2). For each of these methods, two to five selected features were evaluated in the machine learning models. The features selected in each method are based on the highest mean receiver operating characteristic (ROC) area under the curve (AUC) in five-fold stratified cross validation with 20 repeats.

Each model was then trained and tuned on the training cohort, using a stratified five-fold cross validation stratified around 2-EFS, again with 20 repeats. Hyperparameters were initially tuned using a random search cross validation with 1000 different combinations explored (scikit-learn v0.24.2). For all models the random state hyperparameter was set to a value of 0, and, where applicable, the class weight hyperparameter was set to "balanced" to help mitigate the unbalanced nature of the data. The hyperparameters of the 10 top highest validation scores from the random search cross validation were further explored using grid search cross validation (scikit-learn v0.24.2). For the combination of hyperparameters explored in the tuning process, if the mean training and mean validation AUC were not within 0.03 the model was discarded. The remaining models were ranked by the highest mean validation score. The model, hyperparameter and feature selection combination with the highest mean validation score from both the 4.0 SUV threshold segmentation and the 1.5 x mean liver SUV threshold were tested once on the unseen test cohort data. Given the growing literature surrounding the use of MTV as an outcome predictor a separate logistic regression model using MTV was trained on the training set and tested on the unseen test cohort as was used as a comparison to the best performing model. AUCs were compared using the DeLong method. An appropriate threshold from the ROC curve for each of the best performing models was derived using the Youden index with the Matthews correlation coefficient (MCC), sensitivity, specificity, positive predictive value (PPV) and negative predictive value (PPV) presented.

Missing clinical data meant that a comparison with commonly utilised clinical scoring methods was not possible and the treatment regime used was used a surrogate indicator of whether the patient was deemed to have early or advanced disease.

First Order	Shape	GLCM	GLRLM	GLDM	GLSZM	NGTDM
10th Percentile	Elongation	Autocorrelation	Grey Level Non-Uniformity	Dependence Entropy	Grey Level Non-Uniformity	Busyness
90th Percentile	Flatness	Cluster Prominence	Grey Level Non-Uniformity Normalized	Dependence Non-Uniformity	Grey Level Non-Uniformity Normalized	Coarseness
Energy	Least Axis Length	Cluster Shade	Grey Level Variance	Dependence Non-Uniformity Normalized	Grey Level Variance	Complexity
Entropy	Major Axis Length	Cluster Tendency	High Grey Level Run Emphasis	Dependence Variance	High Grey Level Zone Emphasis	Contrast
Inter quartile Range	Maximum 2D Diameter Column	Contrast	Long Run Emphasis	Grey Level Non-Uniformity	Large Area Emphasis	Strength
Kurtosis	Maximum 2D Diameter Row	Correlation	Long Run High Grey Level Emphasis	Grey Level Variance	Large Area High Grey Level Emphasis	
Maximum	Maximum 2D Diameter Slice	Difference Average	Long Run Low Grey Level Emphasis	High Grey Level Emphasis	Large Area Low Grey Level Emphasis	
Mean Absolute Deviation	Maximum 3D Diameter	Difference Entropy	Low Grey Level Run Emphasis	Large Dependence Emphasis	Low Grey Level Zone Emphasis	
Mean	Mesh Volume	Difference Variance	Run Entropy	Large Dependence High Grey Level Emphasis	Size Zone Non-Uniformity	
Median	Minor Axis Length	Id	Run Length Non-Uniformity	Large Dependence Low Grey Level Emphasis	Size Zone Non-Uniformity Normalized	
Minimum	Sphericity	Idm	Run Percentage	Low Grey Level Emphasis	Small Area Emphasis	
Range	Surface Area	Idmn	Run Variance	Small Dependence Emphasis	Small Area High Grey Level Emphasis	
Robust Mean Absolute Deviation	Surface Volume Ratio	Idn	Short Run Emphasis	Small Dependence High Grey Level Emphasis	Small Area Low Grey Level Emphasis	
Root Mean Squared	Voxel Volume	Imc1	Short Run High Grey Level Emphasis	Small Dependence Low Grey Level Emphasis	Zone Entropy	

First Order	Shape	GLCM	GLRLM	GLDM	GLSZM	NGTDM
Skewness		Imc2	Short Run Low Grey Level Emphasis		Zone Percentage	
Total Energy		Inverse Variance			Zone Variance	
Uniformity		Joint Average				
Variance		Joint Energy				
		Joint Entropy				
		MCC				
		Maximum Probability				
		Sum Average				
		Sum Entropy				
		Sum Squares				

Table S4.1 The radiomic features extracted for both the PET and CT components. The equations for the features can be found at https://pyradiomics.readthedocs.io/en/latest/features.html. GLCM = grey level co-occurrence matrix, GLDM = grey level dependence matrix, GLRLM = grey level run length matrix, GLSZM = grey level size zone matrix, NGTDM = neighbouring grey tone difference matrix, Id = inverse difference, Idn = inverse difference normalised, Imc = informational measure of correlation, Idm = inverse difference moment, Idmn = inverse difference moment normalised, MCC = Matthews correlation coefficient. Each of the first and second order features were extracted from the original imaging and then from the images following filters applied. The filters used were: wavelet (LLL, LLH, LHL, LHH, HHH, HLH, HHL, HLL); log-signa (1.0, 2.0, 3.0, 4.0); square; square root; logarithm; exponential; gradient; lbp-3D (m1, m2, k).

2-year EFS: Prediction	2-year EFS: True	Age Group	Sex	Ethnicity	Cancer Stage	Treated as advanced disease	Radiotherapy
0	1	60-69	Female	Caucasian	3	1	0
1	0	70-79	Male	Caucasian	3	1	1
1	0	60-69	Female	Caucasian	4	1	0

1	0	40-49	Female	Caucasian	3	1	1
1	0	40-49	Female	Caucasian	2	0	1
1	0	80-89	Male	Caucasian	2	0	1
1	0	70-79	Male	Caucasian	2	1	0
1	0	30-39	Female	Caucasian	2	1	0
1	0	60-69	Male	Caucasian	3	1	0
1	0	70-79	Female	Caucasian	2	0	1
1	0	40-49	Male	Caucasian	2	1	1
1	0	50-59	Male	Caucasian	3	1	0
1	0	20-29	Male	Non-Caucasian	3	1	0
0	1	40-49	Male	Caucasian	4	1	0
1	0	40-49	Male	Caucasian	4	1	0
1	0	70-79	Male	Not disclosed	4	1	0
1	0	70-79	Female	Not disclosed	3	1	1

Table S4.2 Patient information for mislabelled test cases when using the 1.5 x mean liver SUV combined clinical and radiomic ridge regression model.

Model	Intercept	Coefficients
Clinical and MTV	-0.35815567	Cancer stage 1: 5.02009465 , Cancer stage 4: -1.27629249, Age: 0.4807701, MTV: 0.15398729
1.5 x mean liver SUV	-0.42846688	Age: 0.86012792, PET flatness: 0.75497062, PET major axis length: 1.05538773, PET logarithm GLSZM size zone non-uniformity normalized: -0.57813534, PET lbp-3D-m1 GLCM correlation: 0.61007467, PET lbp-3D-m2 first order skewness: -0.84823908
4.0 SUV	-0.41354898	Age: 0.73897899, PET least axis length: 1.10580035, PET wavelet-HLL GLCM correlation: -0.75524818, PET wavelet-HLH GLCM Idmn: -0.488136, CT wavelet-HLL GLSZM large area low gray level emphasis: -0.85812909

Table S4.3 Intercept and coefficients for the best performing clinical and MTV, and radiomic logistic regression models. GLSZM = grey level size zone matrix, GLCM = grey level co-occurrence matrix, GLDM = grey level dependence matrix, rbf = radial basis function, L = low, H = high, Imc1 = informational measure of correlation 1, Imc2 = informational measure of correlation 2, idmn = inverse difference moment normalized, lbp = local binary pattern.

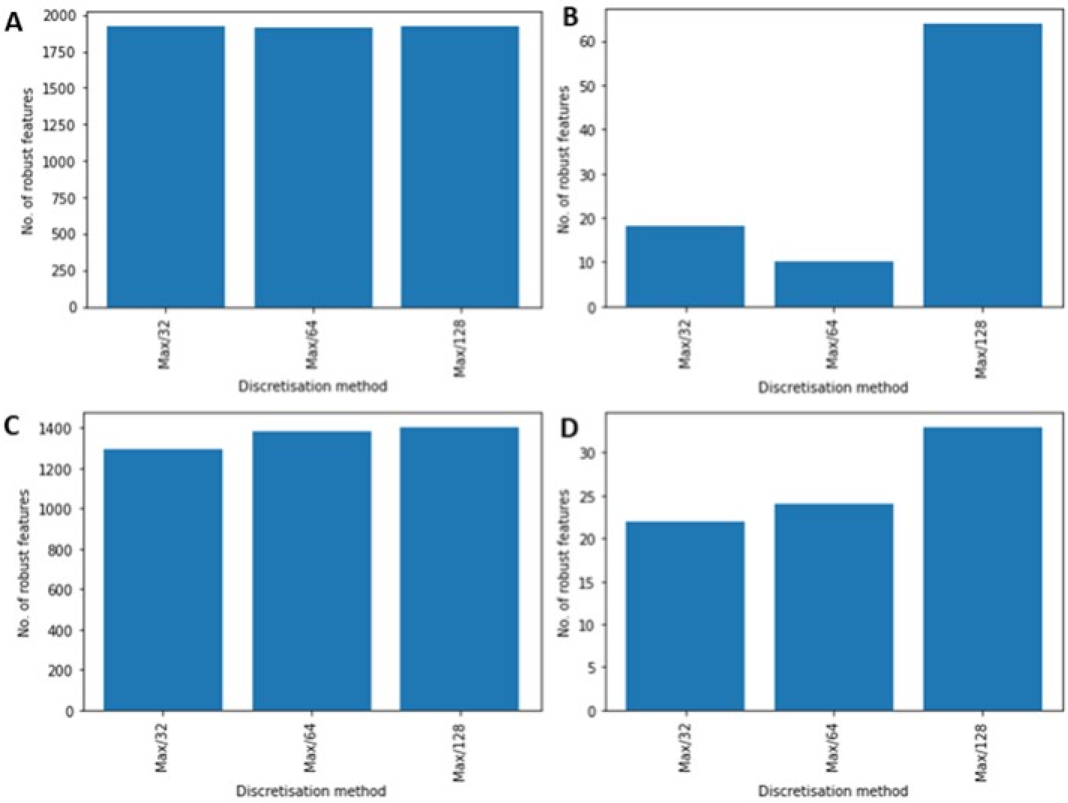


Figure S4.1 Number of robust radiomic features when using a bin width derived from the maximum SUV or HU. A and B represent PET and CT studies, respectively, using 1.5 times mean liver SUV segmentation and C and D represent PET and CT studies, respectively, using a fixed threshold of 4.0SUV segmentation.

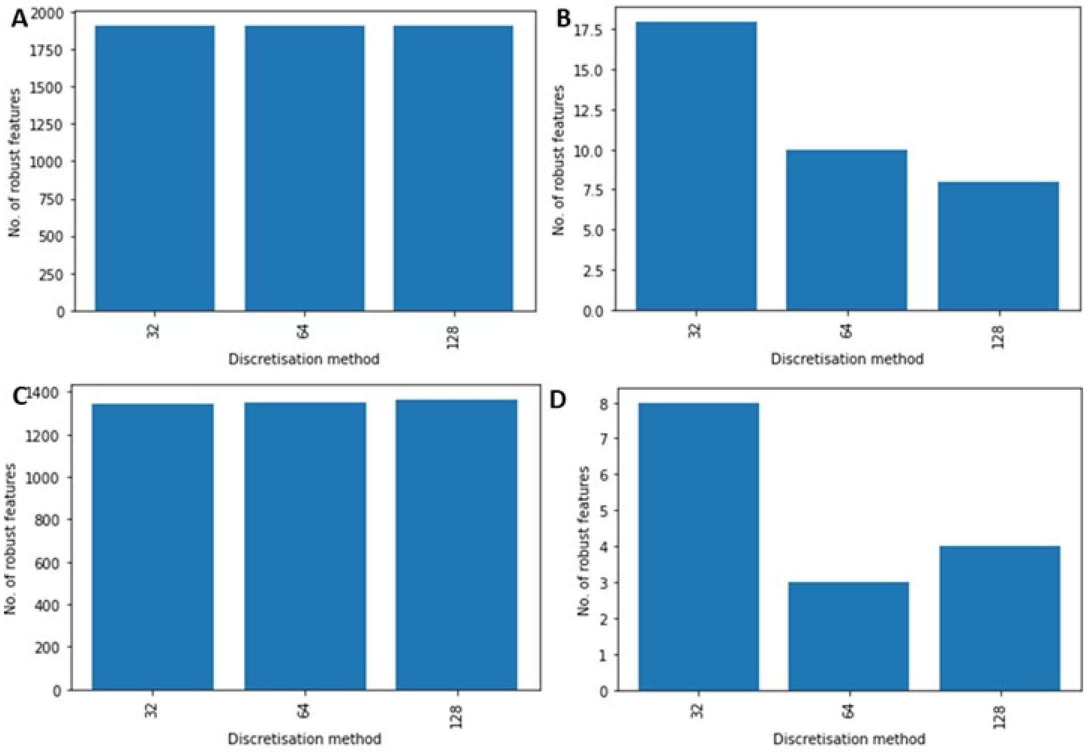


Figure S4.2 Number of robust radiomic features when using a bin width derived from the maximum SUV or HU. A and B represent PET and CT studies, respectively, using 1.5 times mean liver SUV segmentation and C and D represent PET and CT studies, respectively, using a fixed threshold of 4.0SUV segmentation.

Chapter 5 Discussion

5.1 Summary of aims

In this thesis I have reviewed the published literature surrounding use of PET derived imaging biomarkers for outcome prediction in DLBCL and cHL (Chapter 2) and explored, trained and internally validated machine learning derived radiomic models to predict 2-year EFS in DLBCL (Chapter 3) and cHL (Chapter 4).

The following sections will summarise each study and provide a discussion around the limitations and potential future work.

5.2 Baseline PET/CT imaging parameters for prediction of treatment outcome in Hodgkin and diffuse large B cell lymphoma: a systematic review (Chapter 2)

5.2.1 Summary

A systematic review of the literature evaluating the clinical utility of imaging biomarkers derived from baseline FDG PET/CT for outcome prediction in DLBCL and cHL was undertaken. Forty-one articles were included (31 DLBCL, 10 HL) with all of them demonstrating a moderate to high risk of bias. Significant predictive ability was reported in 5/20 DLBCL studies assessing SUVmax (PFS: HR 0.13-7.35, OS: HR 0.83-11.23), 17/19 assessing metabolic tumour volume (MTV) (PFS: HR 2.09-11.20, OS: HR 2.40-10.32) and 10/13 assessing total lesion glycolysis (TLG) (PFS: HR 1.08-11.21, OS: HR 2.40-4.82). Significant predictive ability was reported in 1/4 HL studies assessing SUVmax (HR not reported), 6/8 assessing MTV (PFS: HR 1.2-10.71, OS: HR 1.00-13.20). Six papers explored the use of radiomics for outcome prediction (4 DLBCL, 2 cHL).

The review identified opportunities for future research including establishing multi-centre networks which would help facilitate larger training data and external validation of models; a consensus on an appropriate segmentation technique which could be utilised universally; a consensus on clinically relevant predictors; and, in the exploration of different machine learning models which did not rely on monotonic relationships.

5.2.2 Limitations

One of the main limitations to the study was inclusion of studies which were deemed to have a high risk of bias. However, as detailed in the paper, there were no publications which had a low risk of bias and therefore it may be that the reviewers were "hawkish" and that excluding high-risk studies may have introduced bias into the systematic review. One further limitation is the differences in segmentation technique and scanners used meaning that there is likely to be variation in the results which would influence the forest plots and is the likely cause of wide confidence intervals. The literature tended to focus on a dichotomous split for continuous variables such as MTV and TLG. The rationale for this is likely due to the non-linear relationship of MTV to outcome, however, by stratifying the data in such a manner valuable information is potentially lost.

Since this systematic review was conducted there have been further articles published describing the use of imaging biomarkers for predicting outcomes in DLBCL and cHL. One of the main emerging themes is an increased number of papers exploring the use of tumour based radiomic features in combination with machine learning techniques. A brief overview of the published literature since the systematic review is provided below.

One of the most notable and largest studies recently published is a proposed international metabolic prognostic index by Mikhaeel *et al.* which developed a predictive model using 1242 baseline PET/CTs in patients with DLBCL from five separate previously published research studies [1]. A model based on MTV, age and stage was able to provide an effective predictive model for 3-year PFS. The methodology used in this work is transparent and a simple to use calculator is provided as part of the supplemental material which lends itself well to be easily utilised in future studies. There are a couple of points to consider regarding the adaptation of this model clinically. The first is the need for a standardised automated segmentation process as this aspect can be time consuming in cases with large volume disease or where disease is adjacent to physiological uptake which may be a barrier to clinical translation. The second is the reliability of SUV measurements, with variation in SUVmax reported as being up 10% between repeated studies. This has the potential to lead to changes in recorded volume when a fixed threshold is used [2]. However, the likelihood is that this is would only be a small change, not significantly influencing a patient's probability given the coefficients involved.

There has been an increase in papers including external validation for their trained models more recently. Jiang *et al.* trained and externally validated a 3D U-Net to segment DLBCL and demonstrated that a high predicted MTV ($>$201.2cm^3) was associated with PFS (HR = 3.097, P$<$0.001) and OS (HR=6.601, p$<$0.001) [3]. The U-NET was trained using segmentations which were manually adjusted from a semi-automated 41% SUVmax contour. Although the U-NET was trained and externally validated, it is unclear if survival analysis was performed on the dataset as a whole and therefore has not been validated within the paper.

Another study by Jiang *et al.* developed a radiomic based prediction model in a group of 383 DLBCL patients (273 training and 110 external validation) using a cross combination approach for feature selection and classification model creation [4]. A signature created from 12 radiomic features was a significant predictor of PFS (HR 8.037 (95% CI 3.304-19.551)) and a signature created from 31 radiomic features resulted in a significant predictor of OS (HR 4.054 (95% CI 2.337-7.031)). A combined Cox regression model created using radiomic features, age, Ann Arbor score, presence of bulky disease, SUVmax and TMTV resulted in a c-index of 0.76 (95% CI 0.618-0.795) when predicting PFS in the validation cohort and 0.79 (95% CI 0.668-0.881) when predicting OS in the validation cohort. A number of aspects of the methodology were not described, which make this work difficult to replicate in its current form: the discretisation method is not provided, hyperparameters used are not detailed and the classifier models used to generate radiomic signatures for PFS and OS, had no cut off value provided. It is therefore unclear how the time aspect was handled. Ceriani *et al.* explored the use of baseline PET radiomic in DLBCL for patients treated with either 14-day (training) or 21-day (validation) cycle chemotherapy (training = 156, external validation = 107) [5]. A fixed threshold of 4.0 SUV was used for segmentation. The study used LASSO regression for feature selection and the four radiomic features selected resulted in good AUCs in the validation dataset for PFS (0.71) and OS (0.70). Zhang *et al.* explored a radiomic signature created from LASSO repression in 152 DLBCL patients (training = 100, internal validation = 52), with features being extracted from both the metabolic bulk volume, which is the metabolic volume of the largest lesion, and total MTV [6]. Both radiomic signatures were significant predictors of PFS and OS, segmentation was performed 41% SUV.

Eertink *et al.* explored six different predictive models, a combination of clinical, basic imaging biomarkers and radiomics in assessing 2-year time to progression in 317 DLBCL patients. Segmentation was performed by using a fixed threshold of 4.0 SUV [7]. There was no separate validation group, but five-fold stratified cross validation with 2000 repeats was utilised. They found that a combination of radiomic and clinical features had the best performance. A more recent paper by Eertink *et al.* explored standard conventional PET features, features assessing dissemination of disease and radiomic features extracted from specific lesions (defined by size or SUV uptake) or from the full disease burden to assess the predictive ability in assessing 2-year PFS in DLBCL patients [8]. They found that models consisting of conventional PET and dissemination features had the highest predictive values, and that no lesion selection approach had a significantly higher predictive value than any other. Cottereau *et al.* further explored the use of disease dissemination using the largest distance between lesions standardised to body surface area and found that it was a significant predictor of PFS and OS in a cohort of 290 patients who were part of the REMARC trial [9]. When combining this with MTV it provided three distinct risk groups for 4-year OS (95%, 79% and 66%,) and 4-year PFS (90%, 63% and 41%).

There have also been further smaller studies published , which have a higher risk of bias. A small retrospective study of 35 patients by Ortiz *et al.* demonstrated that both MTV and TLG derived from a fixed threshold segmentation of 2.5 SUV were significantly associated with PFS and OS and that combining immunohistochemical and chromosomal translocations with these imaging features improved the c-index of models [10]. The best performing PFS model having a c-index of 0.923 and OS model having a c-index 0.863 utilising MTV with histological and genetic factors. The main limitation to this model is that it is a small dataset with no internal or external validation and there was no comparison with established clinically based prediction models. Mazzara *et al.* assessed the relationship between genetic metabolic signatures and radiomic defined signatures in outcome prediction in DLBCL patients [11]. They found that first order kurtosis, first order energy, sphericity, NGLDM contrast was associated with the genetic metabolic signature and with PFS and highlights the principle of radiomics to provide a non-invasive assessment of cancer metabolism. Zhou *et al.* performed a small study on 65 HL patients (training = 49 and testing = 16) using LASSO regression for feature selection and then undertaking Cox analysis [12]. They found that GLZLM LZHGE and Dmax were predictive of PFS, (HR = 9.007; p=0.044) and (HR = 3.641; p=0.048) respectively. However, the study included 14 patients with non-classical HL which often have a different treatment and outcome course when compared to cHL and may not be generalisable to populations entirely made up of cHL patients as a whole.

5.2.3 Future work

The frequency of publications in this area demonstrates that this is a very topical area of research which has yet to translate clinically. The issues previously discussed surrounding generalisability of models still remain [13]. More recent studies are becoming more robust and transparent with their model training and testing and are using multi-centre data. Studies such as the one by Mikhaeel *et al.* detailing the international metabolic prognostic index provide the means to easily validate their model on datasets from other institutions in retrospective and prospective trial situations. My intention would be to externally validate this model on the LTHT dataset. There still needs to be consensus agreement on segmentation technique as this not only influences the measured tumour volume but also potentially radiomic features extracted from the segmented volume. A 4.0 SUV threshold technique has been reported as being easily applicable and robust to user variability when compared to other segmentation techniques, but variation between protocol repeatability should also be considered [14].

There has been less research into imaging biomarkers in cHL when compared to DLBCL. cHL lends itself more easily to early risk stratification and prognostic trials based on imaging derived prediction models due to the already clinically adopted treatment stratification strategy [15]. By developing a predictive model using a multi-centre network/trial imaging data, there is the opportunity to perform prospective imaging trials looking at early treatment stratification using imaging and clinical biomarkers.

5.3 Discovery of pre-treatment FDG PET/CT-derived radiomics-based models for predicting outcome in diffuse large B-cell lymphoma (Chapter 3)

5.3.1 Summary

The study explored the use of radiomics and clinical features for prediction of 2-year EFS in 229 DLBCL patients (training = 183, test = 46) treated with R-CHOP. Six different machine learning models (LASSO, ridge and elasticnet regression, random forest, support vector machines and k-nearest neighbour) as well as a simple MTV regression model were trained and tuned using stratified four-fold cross validation with 25 repeats. The best performing model based on the mean AUC derived from the ROC curve was tested on the unseen test set. Ridge regression using the features: stage four, PET original GLSZM large area emphasis, PET wavelet-HHL GLSZM small area emphasis, PET wavelet-HHH GLSZM grey level non-uniformity normalized, PET square 10th percentile was the best performing model. This model outperformed one derived from MTV in the training dataset and had an AUC of 0.73 on the unseen test set.

5.3.2 Limitations

As described within Chapter 3 there are several limitations associated with the study. One of the main limitations is the use of relatively small numbers of patients from a single tertiary centre, and although four different scanners were utilised during the study period, there is still a potential issue with the generalisability of the models to different study populations. ComBat harmonisation was used to try and account for inter-scanner variations. The method uses empirical Bayes to estimate scanner variability which is then adjusted for. However, because of the method of application if a scanner was not part of the training data, then it is not possible to apply the same correction used in the training dataset which would limit its use in prospective trials.

The methodology applied requires semi-automated segmentation with manual adjustment around physiological uptake. This process, although not investigated here, has been shown previously to have some operator dependency which could affect reproducibility. Also, the ease of use of a tool is likely to influence how likely it is to be adopted in clinical practice. Therefore, automated segmentation needs to be prioritised to allow for the adoption of the metabolic prognostic score as well as future radiomic type models, or a CNN for outcome prediction needs to be developed . A study by Yousefirizi *et al.* and Liu *et al.* have both demonstrated the application of CNNs to segment lymphomatous disease [16][17]. The benefit of having a segmentation step is that it can be reviewed visually for any discrepancy, and if needed, the radiomic features can be explained mathematically unlike with a CNN where most of the time these operate as "black boxes" without explanation of how results have been derived.

Overfitting was present within some of the models created, and a penalty was included to mitigate against this. However, the penalty may have been too stringent which might have meant that the performance of the models created were underestimated. The cross-validation split was chosen as it gave the most consistent scores, however, different training splits may have benefited model development. Although the best model had a good performance, no useful threshold for outcome prediction could be derived. The models presented were based on treatment with R-CHOP and are not necessarily applicable to other therapy regimens. Therefore, further models would have to be derived if treatment strategies were to change. Missing clinical data meant that it was not possible to compare the presented models with clinically used prognostic models without severely limiting the number of samples used within the study. Similarly, only 70 patients had cell of origin information, and therefore this was not included in the analysis.

5.3.3 Future work

The study demonstrated the potential for a machine learning radiomics model derived from pre-treatment FDG PET-CT to predict 2-EFS in DLBCL patients. The model was saved in a trained state ("pickled") at the time of development and can be tested on external datasets from other institutions, and this is something to pursue in future work. Access has since been granted to trial data from approximately 150 DLBCL patients from the National Cancer Institute Data Archive (NCT00118209) to externally validate our model [18]. Further work to explore optimal scanner harmonisation allowing generalisability of the models is also required. There is a potential opportunity to use autoencoders to replicate the acquisition parameters of different scanners in image harmonisation. Recent work by our group in Leeds has explored the potential of this approach focusing on brain MRI image harmonisation [19]. Time to event machine learning models could be explored within the dataset to minimise the loss of data which does not meet the follow up time for a binary classifier; the literature is dominated by Cox regression models utilising LASSO regression, however, other machine learning models should be explored [20].

There is growing evidence surrounding the use of combined genetic, radiological and clinical prediction models, and this is something which should be further studied within DLBCL [11,21,22]. The lack of genetic/COO information inhibited the ability to create this model during the thesis. One option which could be explored is the use of multichannel variational autoencoders [23]. Each parameter would take the form of an input and would be convoluted down or reduced in size via dense layers into a concatenated latent space and then decoded back into the different outputs. The variational aspect resamples the latent space to generate realistic outputs, therefore this method could be used to create realistic missing data (genetic, clinical, or imaging) from the data already present.

5.4 Utility of pre-treatment FDG PET/CT derived machine learning models for outcome prediction in classical Hodgkin lymphoma (Chapter 4)

5.4.1 Summary

The study explored the use of radiomic features derived from pre-treatment FDG PET-CT and clinical features for prediction of 2-EFS in 289 cHL patients (training = 231, test = 58). Seven machine learning models were explored (random forest, logistic regression (elastic net, LASSO and ridge penalties explored), k-nearest neighbour, single-layer perceptron, multi-layer perceptron, Gaussian process classifier and support vector machine) and trained and tuned using stratified five-fold cross validation on the training set. The best performing radiomics models from two different segmentation techniques (fixed threshold of 4.0 SUV and a threshold based on 1.5 x liver mean SUV), judged by the mean AUC derived from ROC curve analysis, were tested on an unseen test dataset. The different machine learning models were compared with each other, and a logistic regression model created using MTV and clinical features. The best performing model, trained and internally tested, was a logistic regression predictive model with ridge penalisation based on the 1.5 x mean liver SUV using the features PET flatness, PET major axis length, PET logarithm GLSZM size zone non-uniformity normalized, PET lbp-3D-m1 GLCM correlation and PET lbp-3D-m2 first order skewness. Although there was no significant difference between the top performing radiomics model and a model derived from clinical features and MTV when tested on the unseen test set.

5.4.2 Limitations

There are similar limitations demonstrated in Chapter 4 to those in Chapter 3. Again, the numbers of patients studied is relatively small and from a single tertiary centre which affects the generalisability. The study aimed to be more transparent about the patient population by presenting the mislabelled cases. The patient population the model is based on will affect how generalisable it is to other populations, and with a model trained on sparse data there is a higher chance that there are groups of patients for which there was not training data. The relatively small numbers of patients likely impacts the ability to demonstrate significant differences between the models' performance, as DeLong's test has been reported to be a conservative measure of significance [24].

The discretisation method chosen was based on the method which produced the highest number of radiomic features with an ICC score of over 0.8. However, this method does not take into consideration that there is a potential for radiomic features with little variation between lesions depending on the filters applied and therefore they may be extremely robust but do not provide information to a model. Consideration surrounding the variation present in radiomic features as well as the robustness of features needs to be considered.

5.4.3 Future work

There needs to be further study focused on model generalisability and the likely impact these models would have on patients and clinicians. As mentioned previously cHL lends itself to the setup of prognostic treatment stratification-based imaging trials as there are treatment escalation regimes available [15]. The causality of features used in model creation should be explored further, as although this is not something which is considered in pure predictive modelling the ability to understand and explain possible confounders of a model allows for a greater understanding of its limitations [25]. As discussed in Chapter 4 determining how a model reacts to patients from under represented populations may help improve model performance [26].

5.5 Future perspectives and considerations

As with advances in lymphoma treatment, recent developments in imaging acquisition and improvements in PET/CT technology have the potential to make the current reported predictive models redundant in the future. However, this should not dissuade research in this field, but should be welcomed and explored as these techniques may offer a higher signal to noise ratio for features extracted, and therefore more information for any predictive models. Two acquisition techniques facilitating improved signal to noise ratio which would be of interest are total body PET/CT and 4D PET/CT. As the name implies total body PET/CT has an extended axial field of view which permits simultaneous imaging of the entire body without having to change bed positions, which is the current standard imaging practice [27]. This allows for an increased efficiency in signal collection and higher spatial resolution, and in turn shorter acquisition times, reduction in administrated dose, higher imaging quality and the ability for total-body dynamic imaging [27–29]. Dynamic imaging with kinetic modelling also allows further features to be interrogated when exploring the behaviour of different lesions, disease process and tissues [30]. PET imaging is derived from the detected radionuclide activity over multiple breathing cycles, and consequently the signal to noise ratio of thoracic and upper abdominal lesions can be negatively affected by respiratory motion. This is often confounded in smaller lesions by the low spatial resolution of PET. The use of 4D (respiratory gated) PET/CT aims to negate this motion artefact and improve the signal to allow for more accurate depiction and quantification of lesions [31]. The use of this technique is more widely reported in lung carcinoma, but given the potential of lymphoma to present at diagnosis in these tissues prone to motion artefact its use in this cohort would be provide an area of further research [32].

Lastly, it is easy for a researcher's or research team's subspecialty or research interests to be isolated from other subspecialties without the larger picture considered. The likelihood is that a model derived from genomic, biochemical, socioeconomic, ethnicity, imaging and clinical data will provide more generalisable outcome prediction. The consideration of data other than imaging biomarkers in model creation should potentially be included on any radiomic or AI predictive modelling scoring system to encourage the practice of a more holistic approach.

5.6 Conclusions

This thesis has explored the use of imaging biomarkers derived from pre-treatment FDG PET-CT for outcome prediction in common types of high-grade lymphoma (DLBCL and cHL), demonstrating the potential use of radiomics in combination with clinical features to aid in treatment stratification. The work has added to field by providing a road map for areas of further research and the considerations and limitations which are often overlooked. This work should help inform design of future prospective multicentre studies with the ultimate aim to improve patient outcomes.

5.7 References

1. Mikhaeel NG, Heymans MW, Eertink JJ, de Vet HCW, Boellaard R, Dührsen U, *et al.* Proposed New Dynamic Prognostic Index for Diffuse Large B-Cell Lymphoma: International Metabolic Prognostic Index. J Clin Oncol. 2022;40:2352–61.

2. Lodge MA. Repeatability of SUV in Oncologic 18 F-FDG PET. J Nucl Med. 2017;58:523–32.

3. Jiang C, Chen K, Teng Y, Ding C, Zhou Z, Gao Y, *et al.* Deep learning–based tumour segmentation and total metabolic tumour volume prediction in the prognosis of diffuse large B-cell lymphoma patients in 3D FDG-PET images. Eur Radiol. 2022;32:4801–12.

4. Jiang C, Li A, Teng Y, Huang X, Ding C, Chen J, *et al.* Optimal PET-based radiomic signature construction based on the cross-combination method for predicting the survival of patients with diffuse large B-cell lymphoma. Eur J Nucl Med Mol Imaging. 2022;49:2902–16.

5. Ceriani L, Milan L, Cascione L, Gritti G, Dalmasso F, Esposito F, *et al.* Generation and validation of a PET radiomics model that predicts survival in diffuse large B cell lymphoma treated with R-CHOP14: A SAKK 38/07 trial post-hoc analysis. Hematol Oncol. 2022;40:11–21.

6. Zhang X, Chen L, Jiang H, He X, Feng L, Ni M, *et al.* A novel analytic approach for outcome prediction in diffuse large B-cell lymphoma by [18F]FDG PET/CT. Eur J Nucl Med Mol Imaging. 2022;49:1298–310.

7. Eertink JJ, van de Brug T, Wiegers SE, Zwezerijnen GJC, Pfaehler EAG, Lugtenburg PJ, *et al.* 18F-FDG PET baseline radiomics features improve the prediction of treatment outcome in diffuse large B-cell lymphoma. Eur J Nucl Med Mol Imaging. 2022;49:932–42.

8. Eertink JJ, Zwezerijnen GJC, Cysouw MCF, Wiegers SE, Pfaehler EAG, Lugtenburg PJ, *et al.* Comparing lesion and feature selections to predict progression in newly

diagnosed DLBCL patients with FDG PET/CT radiomics features. Eur J Nucl Med Mol Imaging. 2022;49:4642–51.

9. Cottereau AS, Meignan M, Nioche C, Capobianco N, Clerc J, Chartier L, *et al.* Risk stratification in diffuse large B-cell lymphoma using lesion dissemination and metabolic tumor burden calculated from baseline PET/CT. Ann Oncol. 2021;32:404–11.

10. Guzmán Ortiz S, Mucientes Rasilla J, Vargas Núñez JA, Royuela A, Rodríguez Carrillo JL, Dotor de Lama A, *et al.* Evaluation of the prognostic value of the metabolic volumetric parameters calculated with 18F-FDG PET/CT and its value added to the molecular characteristics in patients with diffuse large B-cell lymphoma. Rev Esp Med Nucl Imagen Mol (Engl Ed). 2022;41:215–22.

11. Mazzara S, Travaini L, Irccs O, Botta F, Irccs O, Granata C, *et al.* Gene expression profiling and FDG-PET radiomics uncover radiometabolic signatures associated with outcome in DLBCL. Blood Advances. 2022; doi: 10.1182/bloodadvances.2022007825.

12. Zhou Y, Zhu Y, Chen Z, Li J, Sang S, Deng S. Radiomic Features of 18F-FDG PET in Hodgkin Lymphoma Are Predictive of Outcomes. Contrast Media Mol Imaging. 2021;22;2021:6347404.

13. Horng H, Singh A, Yousefi B, Cohen EA, Haghighi B, Katz S, *et al.* Generalized ComBat harmonization methods for radiomic features with multi-modal distributions and multiple batch effects. Sci Rep. 2022;12:1–12.

14. Driessen J, Zwezerijnen GJ, Schöder H, Drees EE, Kersten MJ, Moskowitz AJ, *et al.* The impact of semi-automatic segmentation methods on metabolic tumor volume, intensity and dissemination radiomics in 18 F-FDG PET scans of patients with classical Hodgkin lymphoma. J Nucl Med. 2022;63:1424-1430

15. Follows GA, Ardeshna KM, Barrington SF, Culligan DJ, Hoskin PJ, Linch D, *et al.* Guidelines for the first line management of classical Hodgkin lymphoma. Br J Haematol. 2014;166:34–49.

16. Yousefirizi F, Dubljevic N, Ahamed S, Bloise I, Gowdy C, Hyun O. J, *et al.* Convolutional neural network with a hybrid loss function for fully automated segmentation of lymphoma lesions in FDG PET images. 2022;33.

17. Liu P, Zhang M, Gao X, Li B, Zheng G. Joint Lymphoma Lesion Segmentation and Prognosis Prediction from Baseline FDG-PET Images via Multitask Convolutional Neural Networks. IEEE. 2022;10:81612–23.

18. Bartlett NL, Wilson WH, Jung SH, Hsi ED, Maurer MJ, Pederson LD, *et al.* Dose-adjusted EPOCH-R compared with R-CHOP as frontline therapy for diffuse large B-cell lymphoma: Clinical outcomes of the Phase III intergroup trial alliance/CALGB 50303. J Clin Oncol. 2019;37:1790–9.

19. Fatania K, Clark A, Frood R, Scarsbrook A, Al-Qaisieh B, Currie S, *et al.* Harmonisation of scanner-dependent contrast variations in magnetic resonance imaging for radiation oncology, using style-blind auto-encoders. Phys Imaging Radiat Oncol. 2022;22:115–22.

20. Spooner A, Chen E, Sowmya A, Sachdev P, Kochan NA, Trollor J, *et al.* A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. Sci Rep. 2020;10:1–10.

21. Gutiérrez-García G, Cardesa-Salzmann T, Climent F, González-Barca E, Mercadal S, Mate JL, *et al.* Gene-expression profiling and not immunophenotypic algorithms predicts prognosis in patients with diffuse large B-cell lymphoma treated with immunochemotherapy. Blood. 2011;117:4836–43.

22. Liu Y, Barta SK. Diffuse large B-cell lymphoma: 2019 update on diagnosis, risk stratification, and treatment. Am J Hematol. 2019;94:604–16.

23. Qiu YL, Zheng H, Gevaert O. Genomic data imputation with variational auto-encoders. Gigascience. 2020;9:1–12.

24. Vickers AJ, Cronin AM, Begg CB. One statistical test is sufficient for assessing new predictive markers. BMC Med Res Methodol. 2011;11:1–7.

25. Lin L, Sperrin M, Jenkins DA, Martin GP, Peek N. A scoping review of causal methods enabling predictions under hypothetical interventions. Diagnostic Progn Res. 2021;5:3

26. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366:447–53.

27. Cherry SR, Jones T, Karp JS, Qi J, Moses WW, Badawi RD. Total-body PET: Maximizing sensitivity to create new opportunities for clinical research and patient care. J Nucl Med. 2018;59:3–12.

28. Ng QK-T, Triumbari EKA, Omidvari N, Cherry SR, Badawi RD, Nardo L. Total-body PET/CT – First Clinical Experiences and Future Perspectives. Semin Nucl Med. 2022;52:330–9.

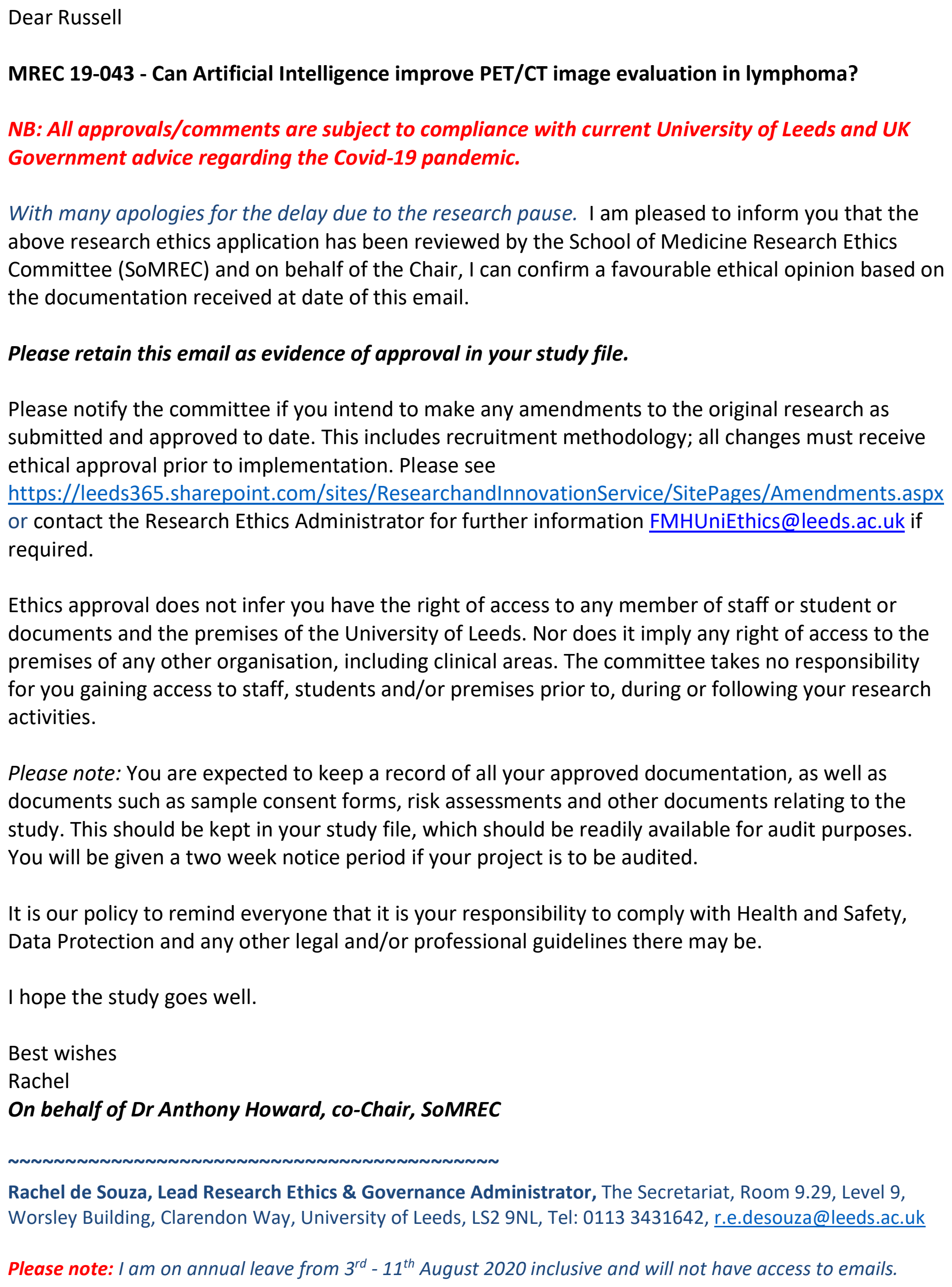
29. van Sluis J, van Snick JH, Brouwers AH, Noordzij W, Dierckx RAJO, Borra RJH, *et al.* Shortened duration whole body 18F-FDG PET Patlak imaging on the Biograph Vision Quadra PET/CT using a population-averaged input function. EJNMMI Phys. 2022;9.

30. Dimitrakopoulou-Strauss A, Pan L, Sachpekidis C. Kinetic modeling and parametric imaging with dynamic PET for oncological applications: general considerations, current clinical applications, and future perspectives. Eur J Nucl Med Mol Imaging. 2021;48:21–39.

31. Frood R, McDermott G, Scarsbrook A. Respiratory-gated PET/CT for pulmonary lesion characterisation—promises and problems. Br J Radiol. 2018;20170640.

32. Frood R, Prestwich R, Tsoumpas C, Murray P, Franks K, Scarsbrook A. Effectiveness of Respiratory-gated Positron Emission Tomography/Computed Tomography for Radiotherapy Planning in Patients with Lung Carcinoma – A Systematic Review. Clin Oncol. 2018;30:225–32

Appendix A Ethics approval



Dear Russell

MREC 19-043 - Can Artificial Intelligence improve PET/CT image evaluation in lymphoma?

NB: All approvals/comments are subject to compliance with current University of Leeds and UK Government advice regarding the Covid-19 pandemic.

With many apologies for the delay due to the research pause. I am pleased to inform you that the above research ethics application has been reviewed by the School of Medicine Research Ethics Committee (SoMREC) and on behalf of the Chair, I can confirm a favourable ethical opinion based on the documentation received at date of this email.

Please retain this email as evidence of approval in your study file.

Please notify the committee if you intend to make any amendments to the original research as submitted and approved to date. This includes recruitment methodology; all changes must receive ethical approval prior to implementation. Please see https://leeds365.sharepoint.com/sites/ResearchandInnovationService/SitePages/Amendments.aspx or contact the Research Ethics Administrator for further information FMHUniEthics@leeds.ac.uk if required.

Ethics approval does not infer you have the right of access to any member of staff or student or documents and the premises of the University of Leeds. Nor does it imply any right of access to the premises of any other organisation, including clinical areas. The committee takes no responsibility for you gaining access to staff, students and/or premises prior to, during or following your research activities.

Please note: You are expected to keep a record of all your approved documentation, as well as documents such as sample consent forms, risk assessments and other documents relating to the study. This should be kept in your study file, which should be readily available for audit purposes. You will be given a two week notice period if your project is to be audited.

It is our policy to remind everyone that it is your responsibility to comply with Health and Safety, Data Protection and any other legal and/or professional guidelines there may be.

I hope the study goes well.

Best wishes
Rachel
On behalf of Dr Anthony Howard, co-Chair, SoMREC

~~~~~~

Rachel de Souza, Lead Research Ethics & Governance Administrator, The Secretariat, Room 9.29, Level 9, Worsley Building, Clarendon Way, University of Leeds, LS2 9NL, Tel: 0113 3431642, r.e.desouza@leeds.ac.uk

Please note: *I am on annual leave from 3rd - 11th August 2020 inclusive and will not have access to emails.*
~~~~~~