Determining the utility of clinical and dosimetric factors for the prediction of radiation induced oesophagitis and pneumonitis

A THESIS SUBMITTED TO THE UNIVERSITY OF MANCHESTER FOR THE DEGREE OF DOCTOR OF CLINICAL SCIENCE IN THE FACULTY OF BIOLOGY, MEDICINE AND HEALTH

Rushil Patel
2022 SCHOOL OF MEDICAL SCIENCES

Contents

Table of Figures

Table of Tables

Abbreviations

#	Fraction
BED	Biological Equivalent Dose
CTCAE	Common Terminology Criteria for Adverse Events
CTV	Clinical Target Volume
DICOM	Digitally Imaging and Communications in Medicine
D_{max}	Maximum dose planned to be delivered to a structure
D_{mean}	Mean dose planned to be delivered to a structure
$D_X$cc	Dose (Gy) planned to be received by a certain volume (cc)
EQD2	Equivalent dose in 2 Gy fractions
FDG	Fluoro-DeoxyGlucose
FFF	Flattening filter free
GTV	Gross Tumour Volume
Gy	Gray
IMRT	Intensity modulated radiotherapy
ITV	Internal Target Volume
LR	Logistic Regression
MD	Mean Dose
ML	Machine Learning
MLD	Mean Lung Dose
MRI	Magnetic Resonance Imaging
NSCLC	Non-Small Cell Lung Cancer
OAR	Organ at risk
PET	Positron Emission Topography
PRV	Planning at Risk Volume
PTV	Planning Target Volume
QA	Quality Assurance
RT	Radiotherapy
RTOG	Radiation Therapy Oncology Group
SABR	Stereotactic Ablative Body Radiotherapy
SBRT	Stereotactic Body Radiation Therapy
TPS	Treatment planning system
V_X	Percentage volume planned to receive dose X (Gy)
VMAT	Volumetric modulated arc therapy

Abstract

Background and Purpose: Predicting toxicity from radiotherapy (RT) is a complex problem because there are usually multiple organs at risk irradiated and protecting all these structures requires compromise. Multiple methods can be used to predict toxicity such as Lyman Kutcher Burman (LKB) modelling, logistic regression (LR) and supervised machine learning (ML). Several trials have used isotoxic RT to treat non-small cell lung cancer (NSCLC) patients, a technique that escalates and individualises RT doses to the tumour to improve local control. Dose escalation is often constrained by the dose unavoidably delivered to oesophagus and lungs during treatment. Here we model toxicity using data from the IDEAL-CRT trial.

Methods: Data from 116 IDEAL-CRT patients were analysed in this study. Clinical data including sex, age, disease stage, forced expiratory volume (FEV), force vital capacity (FVC) and diffusing capacity of lung for carbon dioxide (DLCO) were collected for the trial. Dosimetric information was generated from RT datasets, including V5Gy, V20Gy, Mean dose (MD) and Equivalent uniform dose (EUD) for lung and V35Gy, V50Gy, D1cc and MD for the oesophagus. All doses were reported as equivalent dose in 2 Gy fractions corrected for overall treatment time.

Uni-variable statistical analysis was performed on all metrics using LR, with p-values used to determine which metrics would be most useful for toxicity modelling. The bootstrap method was evaluate the accuracy of LR. This information was used to inform toxicity modelling with ML using the classification learner application in MATLAB v2020a. ML Models reported overall predictive accuracy, sensitivity, specificity and Area Under the Curve (AUC) from recover operator characteristics (ROC) analysis. Resulting ML models were compared with LKB analysis and multi-variable LR.

Results: Uni-variable LR found a statistically significant ($p<0.01$) correlation between oesophagitis and the MD, V35 Gy, V50 Gy, and D1cc values. LR tests were unable to find a statistically significant relationship between any clinical or dosimetric factors and pneumonitis, here FEV and FVC produced $p<0.05$ using LR analysis.

The ML model for oesophagitis with the highest AUC had an overall accuracy of 73.3%, sensitivity of 93.6%, specificity of 31.6% and AUC of 0.79. Model inputs were V50 Gy and sex. The model for pneumonitis with the highest AUC had a predictive accuracy of 76.7%,

sensitivity of 3.8%, specificity of 97.8% and an AUC of 0.53. The model used the EUD and sex as inputs. There was statistically significant relationship ($p<0.01$) between LKB NTCP values and oesophagitis using LR analysis and the Mann-Whitney U test, but not for pneumonitis. The multi-variable LR model for oesophagitis had a predictive accuracy of 71.6%, sensitivity of 85.9%, specificity of 42.1% and an AUC of 0.77. The multi-variable LR model for pneumonitis had a predictive accuracy of 77.6%, sensitivity of 0%, specificity of 100% and an AUC of 0.58.

Conclusion: Predictive models require high specificity and selectivity in order to be clinically useful. Both LR and ML techniques can predict toxicity with similar accuracy when there is good correlation between metrics and toxicity. When there is not, machine learning's ability to utilise more diverse data and customise parameters of learning classifiers enables superior toxicity models to be generated. Further development is required for these models to be clinically useful; this includes testing a wider range of features such as genetic information or imaging biomarkers and validation on independent datasets is vital prior to adoption in the clinic. The results of this study have shown that ML approaches are well suited to radiotherapy toxicity modelling with promising results for the prediction of oesophagitis.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning

Copyright Statement

The Author

Rushil Patel completed a BSc in Natural Sciences and a MSc in Medical Physics. The author is a state registered clinical scientist and is the Lead Clinical Scientist within the National Radiotherapy Trials Quality Assurance Group (RTTQA). The author provides QA and Radiotherapy Physics support for numerous clinical trials but specialises in Lung and SABR trials and is a member of the trial management group for the ADSCAN, HALT, Radiant-BC and SARON trials

List of Published Work

1. Terparia, Samsara; Mir, Romaana; Tsang, Yatman; Clark, Catharine; **Patel, Rushil**. (2020). *Automatic evaluation of contours in radiotherapy planning utilising conformity indices and machine learning.*
2. Dimitriadis, Alexis; Tsang, Yatman; Thomas, Russell; Palmer, Antony; Eaton, David; Lee, Jonathan; **Patel, Rushil**; Silvestre Patallo, Ileana; Gouldstone, Clare; Snaith, Julia; Kirkby, Karen; Nisbet, Andrew; Clark, Catharine. (2020). *Multi-institutional dosimetric delivery assessment of intracranial stereotactic radiosurgery on different treatment platforms.*
3. Lee, Jonny; Dean, Christopher; **Patel, Rushil**; Webster, Gareth; Eaton, David. (2019). Multi-center evaluation of dose conformity in stereotactic body radiotherapy.
4. Conibear, John; Chia, Brendan; Ngai, Yenting; Bates, Andrew; Counsell, Nicholas**; Patel, Rushil**; Eaton, David; Faivre-Finn, Corinne; Fenwick, John; Forster, Martin; Hanna, Gerard; Harden, Susan; Mayles, William; Moinuddin, Syed; Landau, David. (2018). *Study protocol for the SARON trial: A multicentre, randomised controlled phase III trial comparing the addition of stereotactic ablative radiotherapy and radical radiotherapy with standard chemotherapy alone for oligometastatic non-small cell lung cancer.*
5. Eaton, David; Lee, Jonathan; **Patel, Rushil**; Millin, Tony; Paddick, Ian; Walker, Christopher. (2018). *Stereotactic radiosurgery for benign brain tumors: Results of multicenter benchmark planning studies.*
6. Eaton, David; Tyler, Justine; Backshall, Alex; Bernstein, David; Carver, Antony; Gasnier, Anne; Henderson, Julia; Lee, Jonathan; **Patel, Rushil**; Tsang, Yatman; Yang, Huiqi;

Zotova, Rada; Wells, Emma. (2017). *An external dosimetry audit programme to credential static and rotational IMRT delivery for clinical trials quality assurance.*

7. Growcott S, Dembrey T, **Patel R**, Eaton D, Cameron A. *Inter-Observer Variability in Target Volume Delineations of Benign and Metastatic Brain Tumours for Stereotactic Radiosurgery: Results of a National Quality Assurance Programme. UK Consensus on Normal Tissue Dose Constraints for Stereotactic Radiotherapy.*
8. Hanna GG, Murray L, **Patel R**, Jain S, Aitken KL, Franks KN, van As N, Tree A, Hatfield P, Harrow S, McDonald F, Ahmed M, Saran FH, Webster GJ, Khoo V, Landau D, Eaton DJ, Hawkins MA. *UK Consensus on Normal Tissue Dose Constraints for Stereotactic*
9. P. Diez, G.G. Hanna, K.L. Aitken, N. van As, A. Carver, R.J. Colaco, J. Conibear, E.M. Dunne, D.J. Eaton, K.N. Franks, J.S. Good, S. Harrow, P. Hatfield, M.A. Hawkins, S. Jain, F. McDonald, **R. Patel**, T. Rackley, P. Sanghera, A. Tree, L. Murray. *UK 2022 Consensus on Normal Tissue Dose-Volume Constraints for Oligometastatic, Primary Lung and Hepatocellular Carcinoma Stereotactic Ablative Radiotherapy*
10. Joerg Lehmann, Mohammad Hussein, Miriam A. Barry, Shankar Siva, Alisha Moore, Michael Chu, Patricia Díez, David J. Eaton, Jeffrey Harwood, Peta Lonski, Elizabeth Claridge Mackonis, Carole Meehan, **Rushil Patel**, Xenia Ray, Maddison Shaw, Justin Shepherd, Gregory Smyth, Therese S. Standen, Brindha Subramanian, Peter B. Greer, Catharine H. Clark. *SEAFARER – A new concept for validating radiotherapy patient specific QA for clinical trials and clinical practice*

1 Introduction

Lung cancer is the leading cause of cancer related death in the UK, with approximately 47,000[1] new cases a year. Radiotherapy is a common treatment used for both palliative and curable patients. The response of normal healthy tissues that are incidentally and unavoidably irradiated during radiotherapy treatments is often the main limiting factor to increasing the prescription dose to the tumour. Higher doses to the tumour improve the chance of tumour control, but high doses to healthy tissue can also increase the chance of toxicity. Optimising this trade-off is known as the therapeutic ratio, but predicting toxicity is a complex problem because there are usually multiple organs at risk (OARs) irradiated and protecting all these structures requires compromise. We must also consider that various structures will respond differently to radiotherapy due to the difference in their organ structures and the types of tissues involved.

Each radiotherapy treatment is planned by a team of experts and is optimized for the individual patient. The treatment planning process determines the way in which the prescribed dose is delivered to the patient, with the planner able to use a variety of techniques to shape the dose distribution in the patient to attempt to find the best treatment solution. For lung cancer in particular, doses to healthy tissues are often a limiting factor as treatment commonly causes toxicities such as radiation induced pneumonitis or oesophagitis, which can disrupt treatment and may lead to a reduction in tumour control. When optimizing a radiotherapy treatment plan, the planner utilises objectives regarding the required dose to be delivered to the tumour to ensure tumour control and dose constraints for Organs at Risk (OARs), which ensure acceptable levels of toxicity are not exceeded. Intensity modulated radiation therapy (IMRT) is an advanced type of radiotherapy that is commonly used today. IMRT treatment plans are inversely optimized, with the planner setting objectives which specialist radiotherapy treatment software will interpret to create the desired dose distribution. The accuracy of these dose constraints is important to be able to achieve the optimal therapeutic ratio.

OAR dose constraints are derived through toxicity modelling using large data sets. There are multiple methods used to determine the relationship between the occurrence of toxicity and various features in the data which can be used as predictors. Traditionally these outcomes are modelled using information about the dose distribution and fractionation, but it is

recognized that the response to radiation is multifactorial and can include a variety of data sources such a clinical factors, bioinformatics, and genetic information. Toxicity models can be described as analytical, which employ a biophysical understanding of irradiation effects such as the linear quadratic equation or data driven models which are phenomenological models that depend on parameters available from collected clinical and dosimetric data. Data driven models consider observed treatment outcome as a result of mathematical modelling of one or more predictive factors.

Traditionally a Normal Tissue Complication Probability[2] (NTCP) model attempts to reduce complex dosimetric and anatomical information to a single risk measure. Most models fall into three categories[3], Dose Volume Histogram (DVH) reduction models, tissue architecture models and multiple metric models. DVH-reduction models are mainly applicable to non-uniform dose distributions, although they are based on estimated complication probability under uniform irradiation, to do this they generally use the concept of generalised equivalent uniform dose. The tissue architecture model employs the concept that a portion of an organ can be damaged by dose without clinical effect. This model describes organs as either parallel or serial. In parallel organs sub-volumes of the organ function relatively independently and so some small portions of the organ can be damaged without clinical effect, meaning that toxicities are only really seen when a volume greater than a critical volume has been damaged. In contrast for serial organs, complications occur when even a small portion of the organ suffers damage. We often equate the risk of complications to a single DVH value, although this can be overly simplistic. The multi-metric model looks at a larger area of the DVH curve and so can be considered more robust. This approach will often select several uni-variable data features in combination with complex analysis techniques such as machine learning (ML) or predicative statistical techniques such as regression analysis to determine predictors for toxicity. Multi-variable analysis is not only restricted to dosimetric information so a variety of clinical or genetic information can be used in addition to improve the accuracy of these models.

It is important that the toxicity models for OARs are correct as overestimating risk can lead to conservative treatments which could lead to lower chances of tumour control, while underestimating risk can subject patients to unplanned adverse events. NTCP models are not ideal, there are issues with regards to the consistency of grading toxicities, selection of

appropriate statistical models and a number of other factors (e.g. difference between planned and accumulated dose[4]) which can limit the usefulness of the constraints generated by these models. Data from controlled clinical trials is useful for this type of modelling as they are considered to provide high quality data, trials they mandate strict grading of toxicity and high-quality data collection which can reduce the number of issues and sources of bias when compared to using lower quality evidence. Data from clinical trials is often accrued from multiple centres and undergoes rigorous quality assurance, ensuring data is of high quality. Clinical trials data is often highly curated due to strict recruitment criteria, this can mean that models derived from trials data may not always be generalisable to the real world and could contain biases. These data sets often involve small sample sizes, which can lead to problems with the robustness and accuracy of models and caution must be taken to ensure that models do not overfit their sample data. Ideally models developed using clinical trials should be tested on real world data to assess their clinical performance.

Early NTCP modelling techniques[2] assumed that OARs received a uniform dose, which was reflective of the non-conformal radiotherapy techniques used at the time (pre-1990[5]). With 3D conformal radiotherapy becoming more widely used between 1980-1990[6] and IMRT gaining widespread adoption between 2000-2010[6], OARs have become much more likely to receive inhomogeneous doses. Contemporary literature, through sources such as the QUANTEC[3] papers take into account modern radiotherapy delivery techniques such as IMRT, which deliver uniform doses to target volumes whilst delivering a highly variable dose to surrounding tissues. The Lyman Kutcher Burman[7] (LKB) model is one of the most widely used analytical NTCP models, it takes into account partial voluming effects in healthy tissue and was widely used in the QAUNTEC papers. The use of LR and ML is increasing in the radiotherapy community, in part due to improvements in technology that have allowed easier access to these tools in experimental and clinical settings[8], but also due to the digitisation of healthcare which has improved access to patient data making large scale multi-variable analysis easier. However, there are relatively few papers which use radiobiological modelling which attempts to quantify the effect of radiation on tissue rather than using absolute doses, even though the effects of fractionation of dose are well known.

For the treatment of lung cancer there are a number of different treatment fractionation regimes available for patients, each of these have their own dose constraints[9,10]. With the

advent of Isotoxic radiotherapy where treatment doses are escalated until an OAR dose constraint has been met, there is substantial variation in the doses OARs receive. In this case it can be difficult to know which dose constraints to use and it would be useful to have OAR constraints based on Biologically Effective dose (BED) or Equivalent Dose in 2 Gy Fractions (EQD2). These would allow a single set of dose constraints to be used for patients regardless of their treatment regime. Data from the IDEAL-CRT[10] trial, which followed an Isotoxic fractionation scheme with a fixed number of fractions would provide a useful data set to generate these values.

1.1 Background

1.1.1 Radiobiology and LQ model

The biological effect of radiation is not only determined by the total dose but may also be characterised by the way the dose is delivered. Important factors include the fraction dose, dose rate and overall treatment duration. There are a number of different models that have been proposed to predict radiobiological response. The linear quadratic (LQ) model is the most widely used and best validated through experimental and clinical data[11].

In the 1960s Elkind[12] showed that if a dose of radiation was divided into fractions, patient survival increased. This is because most of the radiation induced biological damage is repaired. Recovery occurs during fractionated radiotherapy, and this is more beneficial for normal tissue because there are differences in the ability to repair damage between tumour cells and normal tissue. Normal tissue is better able to repair damage than tumour cells. Through fractionating treatments, we are better able to control tumours whilst reducing the side effects.

Radiobiological experiments by Fowler in 1963[13] and 1965[14] showed that the isoeffective dose (the biologically weighted dose for therapy) was affected not only by the time over which the dose was delivered but also by the dose per fraction. Ellis argued that recovery should be a function of fraction size rather than time. At the end of the 1960s[15] the concept arose that the effect of fraction number and overall time could be separated. This is the basis of the Ellis normal standard dose model[16] (NSD). The NSD model allowed for routine adjustments of dose in the clinic and the calculation of isoeffective regiments with different numbers of fractions and overall treatment times. The NSD model was reasonably effective in predicting

acute effects but failed to predict late effects at large dose fractions (<3 Gy) as the model is empirical and so is only useful when used within the range of dose/fractionation data used to derive it[17], hence it is no longer used clinically. This led to the recognition of the importance the α/β ratio in characterising fractionation sensitivity of tissues. α and β are the respective radiosensitivity coefficients for creating lethal damage by large ionising events or by interactions between two smaller events. In practice the individual values of these coefficients are not known, only the ratio can be inferred from clinical results. α/β is a measure of fractionation sensitivity, so for a large α/β the tissue will be relatively insensitive to fractionation changes and for small α/β tissues will be highly sensitive to fractionation changes. In 1976 Douglas and Fowler[18] applied linear quadratic theory to the study of isoeffective schedules. In 1982 Barendson[19] came up with the concept of extrapolated tolerance dose (ETD), which was subsequently renamed extrapolated response dose and then biologically effective dose. The BED is a measure of the biological effect produced by a radiotherapy schedule and is directly related to the surviving fraction of cells. An increase in the BED will mean a reduction in the number of surviving cells and so greater biological effect. In 1983 Withers[20] recommended the concept of EQD2, which is the equivalent dose if given in 2 Gy per fraction. Radiotherapy treatments are most commonly delivered in 2 Gy fractions and the concept of EQD2 allows for easier comparison of non-standard treatment regimens with more common treatments schedules.

The basic LQ model describes the surviving fraction (SF) of clonogenic or stem cells as a function of radiation dose. The main parameters of this model are α and β which represent the intrinsic radio sensitivity of an irradiated cell. The LQ model uses an α/β ratio, which is a measure of the fractionation sensitivity of the cells. The LQ model has shown its clinical usefulness in predicting the sparing effect of fractionated radiotherapy and in comparing the equivalent total dose of different fractionation schedules. The estimation of the radiotherapeutic outcome and the therapeutic window strongly depends on a reliable estimation of LQ parameter α/β.

The radiosensitive parameters α and β can be measured in vitro in tumour cell lines, but these may not be representative of clinical radiobiological calculations. Although the α/β value can be inferred from clinical data, this is potentially more difficult for organs at risk due to the

inhomogeneity of the doses they receive, but doses could still be correlated to toxicity to determine these values.

The Biologically effective dose is a measure of the biological effect produced by a dose of radiation, but BED is not only dose dependent, it also takes into account the type of dose delivery. The basic formula allows for the pattern of fractionation, but more advanced formulas can also take into account other factors such as the dose rate or Relative Biological Effectiveness (RBE) of the type of radiation. The BEDs are tissue specific and normally calculated for the tumour and late responding normal tissues. This means that any given radiotherapy schedule will generally only utilise two BED values, with normal tissues routinely using an α/β of 3 Gy in clinical practice. These BED values should relate to specific biological/clinical endpoints such as tumour control or toxicity. When considering the BED for tumours, we may also consider two additional factors. These are the k factor and T_{delay}. The K factor represents the dose required to offset one days' worth of repopulation. It is inversely proportional to the tumour radiosensitivity and the tumour doubling time and is therefore largest for tumours which are radioresistant and fast growing. T_{delay} is the delay time after the initiation of treatment before fast tumour repopulation begins. These factors effectively reduce the BED to the tumour, but they do not impact late reacting normal tissues.

The BED for acute reacting normal tissues may be calculated using the formula 1.1.

$$BED = N \times d \times \left[1 + \frac{d}{\left(\frac{\alpha}{\beta}\right)}\right] - k(T - T_{delay}) \qquad (1.1)$$

Where N is the number of fractions and d is the dose per fraction

The LQ model also allows any unusual fractionation schedule to be expressed in terms of a more familiar schedule. This is usually done to express these regimes as equivalent dose in 2 Gy fractions (EQD2), this is the total dose required to produce the same biological response as a non-standard scheme if all the fractions were of 2 Gy. This can be calculated using the formula 1.2.

$$EQD2 = \frac{BED}{1 + \frac{2}{\left(\frac{\alpha}{\beta}\right)}} \qquad (1.2)$$

There are some limitations with using BED. In particular care should be taken when applying the LQ to fraction sizes significantly above 6 Gy per fraction. In this case alternative models may need to be considered. The basic BED formulation also only applies when the fractions have been appropriately spaced (e.g. 24 hrs apart), a more complex formula is required if the fractions are more closely spaced. Many existing NTCP models assume that OARs receive the same dose as the tumours, however with modern radiotherapy there is more likely to be inhomogeneity in the dose distribution and it is prudent to base late normal tissues response on maximum dose or another dose metric. Finally, BEDs are an indirect measure of cell surviving fraction. As tumour control is related to surviving fraction then the tumour BEDs are fairly reliable indicators of tumour response. For normal tissue response, toxicity is not simply related to cell survival. Physiology and hierarchical structure are also very important but are not accounted for in BED. This is less important when comparing treatments with a similar dose distribution, but perhaps more important when the dose distributions are very different. Estimates of the α/β have been derived for a wide range of tumour and normal tissues, the ratios for acute normal tissue reactions are usually 10 Gy or greater while for late effects they are typically closer to 3 Gy. Potentially a lot of these values have been generated from data that predates 3D conformal radiotherapy, whereas the data from modern radiotherapy uses 3DCRT and IMRT where the dose per fractions for and across OARs will vary significantly between different patients due to the greater emphasis on individual dose optimisation for these techniques. This means values in the literature for α/β ratios may not be applicable to modern radiotherapy data sets and therefore may need to be recalculated to take into effect changes in the way modern radiotherapy treatments are delivered.

1.1.2 Alternatives to LQ model

The fractionation effect is most commonly quantified using the linear-quadratic (LQ) model, however alternative models may be needed in some settings, for example to describe the effects of low-dose hypersensitivity[21] or hypofractionation[22,23]. The LQ model can generally be considered to characterise the effects of fractionation reasonably well up to 6 Gy per fraction.

One of the proposed alternatives to the LQ model for large fractions sizes is the lethal-potentially lethal (LPL) model proposed by Curtis[24]. Although, there is little experimental data

that shows that the LPL model describes radiation response better than the LQ model. The LPL model is also more complicated to use and is not well characterized in clinical practice. The LPL model combines the ideas of lesion interaction, irreparable damage caused by single tracks, linear lesion fixation, lesion repair via first order kinetics and binary misrepair. It essentially hypothesises two different types of damage to cells from radiation to quantify the effective dose, irreparable (lethal) and repairable (potentially lethal) damage. Guerrero[22] proposed work to extend the conventional LQ model with an additional parameter in order to more accurately describe the effects of radiation at high dose per fraction. This has essentially created a modified LQ model (MLQ) which tracks the LPL model at both high and low doses and dose rates.

The MLQ equations use the parameter δ, which is calculated from LPL parameters which are all derived from a table of LPL parameter tables per cell line. As shown in the equation 1.3.

$$\delta = \frac{3\eta_{PL}}{2\varepsilon} \tag{1.3}$$

The δ value could also be calculated from the α and β and the final slope of the survival curve D_0 as shown in equation 1.4

$$\delta = \frac{2\beta D_0}{1 - \alpha D_0} \tag{1.4}$$

In practice this would be difficult to implement as the individual α and β are rarely known by themselves and it is difficult to apply values from cell survival curves in the lab to those in clinical practice. This value is then used in the final MLQ model as shown below, where a_1/a_2 is the equivalent of α/β.

$$\frac{1}{D_{tot}} = a_1 + a_2 dG(\delta d) \tag{1.5}$$

1.2 Toxicity Modelling

1.2.1 Lyman Kutcher Burman model

The Lyman Kutcher Burman (LKB) model is probably the most well-known model used for predicating Normal Tissue Control Probability (NTCP) for a radiotherapy treatment plan. The model was developed by Lyman[25] for heavy charged particles beams where partial volumes of homogenous dose could be achieved and was adapted for conventional radiotherapy

through the histogram reduction work of Kutcher[26] and the parameter values of Emami[27] and Burman[7].

There are three parameters to the LKB model. TD50(1) represents the dose for a homogeneous dose distribution to an organ at which 50% of patients are likely to experience a defined toxicity at 5 years. m is related to the standard deviation of the TD50(1) and describes the steepness of the dose response curve and n indicates the volume effect of the organ being assessed. The publication of the QUANTEC report[7] has brought together much of the literature and experience of normal tissue toxicity. The LKB model was used by Gulliford[28] to analyse rectal toxicity, the method used in this paper is detailed below.

The LKB model from the original publication[29] is given in equation 1.6

$$NTCP = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{u} e^{-t^2/2}\,dt \tag{1.6}$$

Where

$$u = \frac{D - TD50(V)}{m \times TD50(V)} \tag{1.7}$$

$$\mathrm{TD50(V)} = \frac{\mathrm{TD50(1)}}{V^n} \tag{1.8}$$

TD50(V) is the tolerance dose for a partial volume V. The parameter m multiplied by TD50(V) approximates the standard deviation of volume V and n indicates the volume effect of the organ being assessed. N=0 indicates a completely serial organ where maximum dose dominates outcome and n=1 indicates a parallel organ where the mean dose is related to outcome. D is the maximum dose of the DVH to ensure V<1. Histogram reduction can be performed to calculate the effective volume V according to the method described in Kutcher et al[26].

$$V = \sum_i (\frac{D_i}{D})^{\frac{1}{n}} \Delta V_i \tag{1.9}$$

Where D_i is the dose defined for each bin in a differential dose volume histogram and D is the maximum dose the to the organ. A Maximum Likelihood Estimation (MLE)[30] can be used to best fit values of the parameters TD50(1), m and n of the NTCP model for known binary

outcomes y(i) of the available data by maximising the natural log of the likelihood (LLH) that the fitted model describes the data correctly.

1.2.2 Logistic Regression Analysis

Logistic regression analysis is the standard statistical tool for binary data. The technique is used to obtain an odds ratio and is a classification algorithm that is used where the response variable is categorical. The idea of logistic regression is to find a relationship between features and the probability of a particular outcome. When utilising this for toxicity prediction we are using binomial logistic regression as the response variable has a value of either 0 or 1 dependent on whether the patient has incurred a treatment related toxicity above a certain grade or not.

The logistic regression model is expressed as

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \qquad (1.10)$$

Where p is the probability of the event, β are the coefficients, x are the explanatory variables and $\left(\frac{p}{1-p}\right)$ are the odds

If x is a binary variable:

$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1$ When x = 1 and

$\log\left(\frac{p}{1-p}\right) = \beta_0$ when x = o

So the odds ratio is given in equation 1.11

$$e^{\beta} = \frac{\left(\frac{p_1}{1}-p_1\right)}{\left(\frac{p_0}{1}-p_0\right)} \qquad (1.11)$$

Binomial logistic regression estimates the probability of an event occurring and if this probability is greater than or equal to a pre-set cut off value (typically 0.5), it will be classified as the event having occurred. If it is less than the cut-off value, then it will be classified as the event having not occurred. These predictions will be compared with the observed toxicity to generate overall predictive accuracy, sensitivity, and specificity of the model. LR analysis will

also show the contribution of each independent variable to the model and its statistical significance.

1.2.3 Machine Learning

The theoretical framework for machine learning has been around since the 1950s[31] and is increasingly being used for toxicity prediction in radiotherapy. ML is typically considered as a subset of artificial intelligence (AI) and generally refers to a set of algorithms that can learn to perform a specific task without an explicit implementation of the solution. These ML techniques are particularly well suited to model the relationship between treatment induced toxicity and data features as they excel at dealing with large and complex data sets. For toxicity modelling, supervised machine learning where the algorithms are presented with a known outcome via a labelled dataset is most appropriate. ML models that are able to predict outcomes from a set of data features by tuning model parameters on a number of training cases, are referred to as a classifier. Some of the most common classifiers are Naïve Bayes, logistic regression, k-nearest neighbour, random forest, support vector machine, artificial neural networks, and ensemble.

ML classifiers will learn toxicity model parameters from the available data and so the characteristics of the dataset are very important. If the datasets are too sparse or are not representative of the overall population, models parameters will not be generalizable to real world data. Often this is a result of the modelling overfitting to the training dataset. To avoid overfitting of models, datasets are often split into training and validation cohorts but for smaller datasets k-fold cross validation is used. For k-fold validation datasets are split into k sets of roughly equal size, the first subset will be held out when training the model and the outcomes from the held-out datasets are then predicted by the model and used for validation and evaluation of the performance of the model in terms of predictive accuracy. The first subset is then returned to the training dataset and the second is held out and so forth. The k resampled estimates of performance are then summarised and used to understand the relationship between the model tuning parameters and the model utility.

It can be very difficult to rank the performance of ML models as the performance is related to the particular problem and dataset being analysed, but the acceptable metric commonly used in the literature is to approximate model performance using the Area Under the Curve (AUC) value from Receiver Operator Characteristic (ROC) curve. The AUC value of a model

ranges between 1 and 0.5, with 1 corresponding to perfect classification of the validation set and 0.5 corresponding to a random classification. Although it should be noted that the AUC value can be misleading where there are flaws in the model such as heavily imbalanced datasets or poor validation of the model. The Transparent Reporting of a multivariable model of Individual Prognosis Or Diagnosis (TRIPOD)[32] standard has been developed to improve the reporting and critical appraisal of prediction model studies, however the TRIPOD-AI standard has not yet been published. So this study conforms to the MINIMAR[33] (MINimum Information for Medical AI Reporting) standard (Appendix 4).

1.2.4 Studies using the LKB Model

The LKB model has successfully been used in a number of different studies, in particular there are numerous papers where it has been used to assess toxicity for patients that received radiotherapy to the pelvis. Gulliford et al[28] stipulated that it would be useful to predict a range of toxicities that are commonly observed including rectal toxicity that concern patients and have quality of life implications, although less data is available for these end points. Gulliford used data from the MRC-RT01[34] trial to attempt to predict additional toxicities for these patients, thereby demonstrating the versatility of the LKB model.

The ability to fit the parameters to the model hinged on the availability of detailed dosimetric information and corresponding follow up of the clinical data with accurate reporting of the toxicity. Data from the MRC-RT01 randomised clinical trial provided detailed treatment and follow up data. Gulliford[28] used rectal NTCP data for 388 patients treated with prostate radiotherapy as part of the trial. The trial randomised prescription between 64 Gy or 74 Gy. Treatment delivery was via 3D conformal radiotherapy and all patients received 64 Gy with some patients randomised to receive a 10 Gy boost. All contours were reviewed by a single observer. The LKB model was fitted to five different toxicity end points; three were clinically reported and two were patient reported. In each of these cases the fit was made separately for grade 0 vs. grades 1 and 2 and grade 0 and 1 vs. grade 2. In each case the maximum grade recorded over the entire length of the follow up was used. Patients who experienced a defined endpoint prior to treatment were excluded from the parameter fitting for that endpoint.

In this case fits were made for the five specific rectal toxicities and the two grades of complication. The parameter fits were to a specific set of data, so a bootstrap method was

employed. 1,000 different cohorts of 388 patients were generated from the patient data using sampling with replacement. The LKB model was refitted for the five end points for each of the 1000 sampled populations using MLE. The non-parametric Mann Whitney U test was also calculated to test for the significant difference in the NTCP values of the groups who reported a specific endpoint compared to patients that did not.

The MLE estimate of the TD50(1) of 59.2 Gy for rectal bleeding for both G1&2 and G2 only, were significantly lower than the QUANTEC[35] value of 76.9 Gy. The MRC-RT01 trial was conducted in an era when 3DCRT was first being implemented and the only rectal constraint in the trial was the maximum dose. This meant that there was wide variation of the doses of the rectum, which was useful for toxicity modelling. The results presented by Gulliford et al emphasise the benefit of the bootstrapping process and the leave one out analysis where the effect on a wider population can be tested. The leave one out analysis shows the effect of a single case, particularly useful when the number of events is low. This type of analysis is subject to the accuracy of incident reporting, as this can skew the data.

The Lyman[29] model has traditionally been used in combination with histogram reduction methods to take into account the heterogeneous dose distribution received by normal tissues. It was developed at a time when partial volumes of homogeneous dose were more prevalent and the ability to spare normal tissues was limited. Advances in radiotherapy planning and treatment delivery allow us to create highly conformal dose distributions using inverse optimisation. DVH histogram reduction methods are generally insufficient to fully characterise these dose distributions as they condense all the dosimetric information into a single value, which may not be representative of the response of an OAR to a particular dose distribution. Through the use of a large number of patients, it should be possible to overcome these biases. The applicability of the LKB model is dependent on the quality of the data put into the model. Gulliford concluded that the DVH response of the rectum is different for different endpoints and that quality of life related issues may not be fully predicted by the classic n value of the LKB model. The degeneracy of the model also means that it may be influenced by single cases and that this should be fully explored when deriving constraints for clinical use.

The LKB model has also been combined with the LQ equation. A paper by Tucker et al[36] tried to estimate the α/β ratio from the LQ model for grade 2 late rectal toxicity for patients that

were part of the RTOG protocol 94-06 trial in order to determine whether correcting the rectal DVH for differences in dose per fraction based on the LQ model improves the fit of this data to the LKB NTCP model. The rational for this project was that evidence has shown the severity of normal tissue toxicity depends in part on the number of dose fractions into which the total radiation dose is divided[37] and that the fractionation effect is quantified by the α/β parameter from the LQ model[38].

In this paper Tucker observed that previous calculations of the α/β ratio were based on data prior to the advent of 3D conformal radiotherapy and that modern radiotherapy plans would deliver doses per fraction which would vary significantly across OARs. α/β values from clinical data were previously based on the total isoeffective target dose (the dose corresponding to a specific level of tissue injury). Inpatient variations in dose were disregarded and therefore these estimates are not accurate. Tucker et al used the LKB model for NTCP to allow them to take into account fractionation and volume effects at the same time. Data for the project was from a multicentre trial using dose escalation. The variations in dose were modest only going from 1.8-2 Gy but because this was 3DCRT, the variation in sub volumes across the rectum varied from 0 Gy to 2.2 Gy. It had been suggested by Wheldon et al[39] that due to the known effects of fractionation, the dose in a DVH should be adjusted to dose per fraction when NTCP models are fitted. Tucker also tried to determine whether LQ correction would lead to an improved fit compared with the total dose to the rectum. The study used the LKB model and the concept of effective dose but revised the LKB model to obtain the LQ corrected version of the model in which physical dose was replaced by EQD2 dose, the dose biologically equivalent to D_i if given in 2 Gy per fraction using equation 1.12.

$$\mathrm{LQ}_2 = D_i \frac{\frac{\alpha}{\beta} + d_i}{\frac{\alpha}{\beta} + 2} \qquad (1.12)$$

The dose per fraction was calculated by dividing D_i by the number of fractions received. The values of α/β were estimated as an unknown parameter when the LQ corrected LKB model was fitted to the data.

The paper was able to estimate the α/β ratio of 4.8 Gy, although only with a confidence of 68%. The LQ corrected model did not lead to a better fit of the LKB model for this cohort of patients. This is consistent with previous analysis of the same data in which volume effects

and not dose per fraction seem to explain increases in toxicity. The data used in this study had a very narrow range of doses per fraction, so perhaps would not be able to observe the difference and only parts of the rectum receiving the highest dose suffer this particular toxicity.

1.2.4.1 Modelling Lung Toxicity

Tucker et al[40] explored the superiority of effective dose over mean lung dose for predicting radiation induced pneumonitis (RP), using data from patients in the NCT00915005 trial which compared IMRT vs passive scattered protons. Using a dose of 66-74 Gy in 33-37 fractions with concurrent carboplatin/paclitaxel. A prior study from 10 years earlier used data from more than 500 patients looked at NTCP Grade 3 toxicity and above using the generalized Lyman model[41]. The model incorporated a flexible dosimetric factor, which is the effective dose and is also known as equivalent uniform dose. It allowed the exposed volumes of an organ at risk to be weighted differently than they are by mean lung dose[42]. The weighting of these volumes is governed by n, when n=1, each volume is weighted by the dose it receives and so the effective dose corresponds with the average. When n is less than 1, sub volumes receiving higher doses are weighted more heavily. The earlier analysis demonstrated a better fit to overall risk of RP from D_{eff} with n approx. 0.5 than with n =1. D_{eff}[41] is also known as equivalent uniform dose (EUD) and assumes that two dose distributions are the same if they cause the same radiobiological effect, it can be calculated directly from dose volume histograms. The Lyman model allowed the inclusion of factors outside dosimetric factors to influence toxicity and the earlier investigation revealed that smoking status had a significant effect on RP risk.

Subsequent NTCP modelling investigations[43] using 3DCRT found that RP incidence based on mean lung dose was associated with beam arrangement indicating the inadequacy of MLD as a risk predictor. D_{eff} was able to predict RP risk more effectively and do so independently of beam arrangement. It also derived an n parameter close to 0.5, an interpretation of this value is that organ subvolumes are weighted by the squared dose and D_{eff} is more conveniently described as the square root of the average squared dose to the organ at risk or the root mean squared dose (RMSD).

The study by Tucker evaluated D_{eff} verses MLD when using the Lyman model for describing RP risk. The analyses indicated that RP risk is best quantified using the RMSD to lung, which predicts risk equally well for IMRT. This highlighted that high doses to small volumes may have

a greater impact than low doses to large volumes for the risk of RP. The paper recommended that thought should be given to using RMSD dose over MLD and V20 Gy to predict lung toxicity more accurately.

1.2.5 Studies using LR

Huang et al[44,45] performed two studies for the prediction of oesophagitis using LR. The first modelled severe oesophagitis (grade 2 or higher) for patients from the RTOG 93-11 study of 374 patients in which 120 developed toxicity. They found that mean dose to the hottest 85% volume, mean oesophageal dose and V30Gy were strong predictors of toxicity using uni-variable LR analysis. Multi-variable statistical analysis with cross validation from the study suggested that a two variable logistic model based on mean dose and use of concurrent chemotherapy robustly predicted acute oesophagitis risk for the trial dataset. The second study tested the published two variable model on a new and independent data set and sought to update the model for clinical use. A total of 115 patients were analysed in which 94 patients developed grade 2 or higher toxicity. The model achieved an AUC of 0.78 and was found to be almost as predictive as a model built from the new data alone using a logistic function, suggesting the original model is generalisable to real world data.

Ryckman et al[46] performed a single institution retrospective review for patients who received SABR to the lung. The study dataset consisted of 93 patients of which 8 developed radiation induced pneumonitis of grade 2 or higher. The study found that mean lung dose > 6 Gy predicted 5 of the 8 patients, while a V20Gy > 10% captured only 2 patients that also had a MLD > 6Gy. They concluded that mean lung dose and V20 were the strongest dosimetric predictors of toxicity.

Makimoti et al[47] studied 111 patients with primary lung cancer who received radiotherapy or combined chemo-radiotherapy of which 17 patients developed severe pneumonitis. The study assessed age, gender, histology, clinical stage, pulmonary function tests, total radiation dose and a number of other clinical metrics. Logistic regression analysis found that pre radiotherapy interstitial changes and radiotherapy to the contralateral mediastinum of >40 Gy were significant risk factors associated with severe radiation pneumonitis and that these factors may be used to predict severe toxicity.

Palma et al[48] performed a study on 836 patients that underwent combine chemo-radiotherapy across Europe, north America, and Asia. Factors predicative of symptomatic pneumonitis of grade 2 or higher were evaluated using logistic regression. The overall rate of pneumonitis was 29.8% which equated to 249 patients. In the training data sets factors predictive of toxicity were lung volume receiving ≥20Gy and carboplatin/paclitaxel chemotherapy, with a trend for age. Other factors indicated by this study to be predictive include volumes of lung V5Gy, V13Gy, V25Gy and V30Gy, but these dose variables were found to be collinear.

Umberto et al[49] investigated predictors of radiation induced lung injury in SABR patients. The study cohort of 60 patients underwent SABR with a dose of 45Gy in 3 fractions. The following metrics were tested; the PTV volume, tumour location, primary vs metastatic tumour and mean lung dose (in EQD2). Grade 2 or higher toxicity was seen in 9 patients and logistic regression analysis showed a good correlation between the mean lung dose (EQD2) and incidence of toxicity, with no toxicity seen in patients with mean lung dose less than 12Gy.

1.2.6 Studies using ML

A recent review by Isaksson et al[8] found 53 papers using ML based models for the prediction of toxicity, of these models there were 15 studies for lung cancer and only one for the oesophageal cancer. The most popular cancer site was the prostate (16 studies).

For oesophageal cancer Hart et al[50] characterized the relationship between radiation induced pneumonitis and pulmonary metabolic activity on post treatment FDG-PET scans. The study dataset consisted of 101 patients with oesophageal cancer who underwent an FDG-PET scan 3-12 weeks after completing thoracic radiotherapy. Modelling was perfumed to determine the interaction of pulmonary metabolic radiation response (PMRR), mean lung dose and the percentage of the lung receiving >20Gy with the incidence of pneumonitis. From the study dataset 60 patients had a grade 2 pneumonitis, whilst 5 had grade 3. They found a combination of PMRR, and mean lung dose had the highest predictive accuracy and were able to achieve a sensitivity and accuracy of 53.3% and 62.5% respectively. The peak AUC achieved was 0.63.

Das et al[51,52] performed two studies to predict pneumonitis, the first used a dataset of 219 lung cancer patients treated with radiotherapy. For both studies the optimal models were

derived by fusing two or more single models. The models were trained for grade 2 or higher toxicity which occurred in 34 patients. The work used four common ML models, each model incorporated a small number of features from available set dose parameters and non-dose patient variables. Fusion was achieved by simple averaging of predictions from each patient from all four models. The AUC values achieved were 0.79 from cross validated results. The features arranged in order of importance were, chemotherapy, EUD, lung V20Gy, sex and squamous cell histology. The second model used a database of 234 patients of which 43 were diagnosed with pneumonitis. This model augmented the Lyman NTCP model by combining it with weighted non parametric decision trees for both dose and non dose inputs. This analysis achieved an AUC of 0.72, which was an improvement over the 0.63 achieved by the Lyman NTCP model alone.

Valdes et al[53] developed a patient specific big data clinical decision tool to predict pneumonitis in stage 1 NSCLC patients that were treated with SABR. They recorded 61 features for 201 patients in whom 8 developed pneumonitis. Feature selection highlighted that the most important features for pneumonitis were the diffusion capacity of the lungs for carbon monoxide and the dose to the heart, trachea, and bronchus. Su et al[54] attempted to predict radiation induced pneumonitis using artificial neural networks. The study was applied retrospectively to 142 patients treated with 3D-CRT of which 26 developed toxicity. Model inputs were limited to lung dose volume data only, a volume vector that described patient lung sub volumes receiving more than a set of threshold doses was used as the input variables. The optimal model achieved 73% sensitivity and 99% specificity with an AUC value of 0.85.

El Naqa et al[55] modelled the occurrence of oesophagitis in 166 NSCLC patients of which 45 developed grade 3 or higher toxicity using a Logistic Regression classifier. The authors concluded that model performance could be improved by mixing clinical and dosimetric factors as input parameters for modelling. In a second paper El Naqa[54] tested several different linear and non linear kernels to approximate treatment response. This was performed on a dataset of 166 patients and examined the occurrence of grade 2 or higher toxicity. The optimal oesophagitis model consisted of 5 dosimetric and clinical variables, the results showed the importance of concurrent chemotherapy and DVH cut offs of 30, 45, 55 and 85 Gy.

1.2.7 Summary of Models

Many of the LKB models discussed in this section were published around a decade ago and have used mature datasets based largely based on 3D conformal radiotherapy, which perhaps makes them less relevant to modern radiotherapy techniques that have largely transitioned to VMAT as the standard for lung radiotherapy. The studies by Tucker et al have used relatively large datasets in excess of 500 patients from clinical trials. Due to the nature of NTCP modelling, these studies are primarily focussed on using a single dose metric to quantify the risk of toxicity and do not take into account clinical or biological factors. For lung these studies have investigated equivalent uniform dose and mean dose as a surrogate to encapsulate the whole DVH. LKB studies often focus on deriving the variables in the LKB equation (TD50(1), m, and n) rather than predicting toxicity in patients, with importance placed on the TD50(1) which is the dose at which 50% of patients are likely to suffer toxicity. This is line with how dose constraints are currently used clinically, where the constraints are based on a fixed percentage of patients that are likely to suffer from a specific toxicity. The national SABR dose constraints by Diez et al[56] are a good example of this, as it stipulates the expected toxicity and rates of occurrence associated for each constraint. LKB studies often utilise the bootstrapping method to estimate the accuracy of their models and maximum likelihood estimation to work out model parameters. Some studies have also converted dose to EQD2 to compensate for the variation in dose deposition to OARs.

Logistic regression and machine learning papers have focus on creating models that can accurately predicting toxicity for the individual patient. This is in line with modern radiotherapy developments where treatments are becoming increasingly individualised to the patient. Both these methods of toxicity prediction commonly use a combination of dosimetric and clinical factors for modelling, which can maximise the information available for model training. For the oesophagus mean dose is a commonly used feature, but some papers have used novel metrics such as doses to partial volumes or pulmonary metabolic response from PET imaging. For the lung common features include mean dose, V20Gy and chemotherapy, but studies have used a range of other dose thresholds, age, sex, tumour location, tumour volume, DCLO and doses to nearby OARs amongst others. ML papers have frequently used multiple classifier types, common classifiers include logistic regression and naïve bayes models, although more complex kernel and neural networks have been used.

Whilst some papers such as those by Palma and Huang have used large patients' cohorts (836 and 374 patients respectively), many of the papers had small cohorts with either very high or very low rates of toxicity which make them unsuitable for modelling toxicity as they are dominated by a particular outcome. LR and ML papers will often report which features were most useful for modelling toxicity and the AUC values from ROC analysis for corresponding models as their endpoint. A small number of papers have been able to validate models on independent data sets, however most studies lack patient numbers to do this and often rely on methods such as k-fold cross validation or bootstrapping to assess the accuracy of their models.

Overall there is a trend towards more ML modelling being published rather than LKB or LR, with a focus on creating models to predict toxicity. This research is more often directed to lung toxicity rather than the oesophagus. Modelling routinely combines clinical and dosimetric data with newer studies also utilising more novel dosimetric features in an attempt to take into account the underlying structure of OARs, however there has been limited use of biomarkers or genetic information in toxicity modelling.

1.3 Project Specific

1.3.1 Thoracic Dose Constraints

Clinically significant symptomatic radiation pneumonitis (RP) occurs in approximately 5-50% of patients that are treated for lung cancer[57] making it one of the most common side effects of radiotherapy treatment. Mean Lung Dose (MLD) first proposed by Kwa et al[58] is often used to predict toxicity owing to its simplicity and reasonable effectiveness in predicting RP. Other methods have been shown to be more effective in predicting RP than MLD, although their increased complexity has meant that they are not currently as widely used[40]. Volume predictors such as the Lung V20 Gy first proposed by Purdy et al[59] are well correlated with toxicity and commonly used clinically. The dose constraints recommended by QUANTEC are a V20 ≤ 30-35% and MLD ≤ 23-23 Gy which should limit the risk of RP to ≤ 20% for patients with Non-Small Cell Lung Cancer (NSCLC).

Acute oesophagitis is another common side effect of radiotherapy for patients undergoing thoracic radiotherapy. Concurrent chemoradiotherapy or hyperfractionation results in 15-25% rate of severe (G3 or above) acute oesophagitis. There is no recommended dose

constraint for this, although a trend of exposures greater than 40-50 Gy increasing the risk of acute oesophagitis[60] has previously been demonstrated.

The risk of cardiac events is related to the dose and irradiated volume[57]. Excess deaths have been reported from exposures of 42-45 Gy[61], but the threshold could be as low as 30 Gy[62]. The risk of cardiac toxicity from radiation is still unclear due to the long lead time for toxicity and numerous other factors which can influence cardiac events, thus making it difficult to determine the exact threshold. Analysis of the IDEAL-CRT data showed that 38% of patients had ECG changes at 6 months , these were found to be highly correlated with left atrial wall receiving doses greater than 63 Gy[63].

1.4 Data Sources

Data from the IDEAL-CRT[10], a non-randomised clinical trial could provide valuable data for determining thoracic dose constraints. This dataset contains DICOM information relating to the radiotherapy treatment plan, including planning CT, structure sets and dose cubes as well as clinical information including occurrence and severity of radiation induced oesophagitis/pneumonitis, sex, age, disease staging, forced expiratory volume (FEV), Force Vital Capacity (FVC) and diffusing capacity of lung for carbon dioxide (DLCO).

Inclusion criteria for the trial were histologically/cytologically confirmed stage IIA-IIIB NSCLC, World Health Organization performance status (PS) 0 or 1, suitability for CRT agreed by multidisciplinary team, no prior anticancer therapy, FEV Predicted ≥1 L or 40%, DCLO ≥40% predicted, suitable for chemotherapy, and glomerular filtration rate 60 mL/min. Exclusion criteria were chronic liver disease or bilirubin >35 mmol/L, connective tissue disorders, and history of prior malignancy likely to interfere with the protocol treatment.

For the IDEAL-CRT[10,64] trial all patients had NSCLC and received radiotherapy with concurrent cisplatin and vinorelbine. Inclusion criteria for the trial were histologically/cytologically confirmed stage IIA-IIIB NSCLC, World Health Organisation performance status 0 or 1, suitability for CRT, FEV ≥1L or 40% predicted. RT doses between 63 Gy and 73 Gy in 30 fractions over 40 days in the first schedule (6 weeks, 5 fractions per week) and doses between 63Gy and 71Gy in 30 fractions over 33 days for the second schedule (5 weeks, 6 fractions per week, including 2 on the same day separated by a minimum of 6 hrs). Patients were given the highest tumour dose possible whilst maintaining OAR tolerances and target coverage limits.

Dose limits existed for lungs, spinal cord, brachial plexus, and heart, but not for oesophagus, so the dose limit was incrementally increased over the course of the trial. Patients were split into two non-randomised groups, for Group 1 the max dose deliverable was limited by oesophagus dose, this incrementally went up from 65 Gy to 71 Gy for the 6 week schedule and was 65Gy for the 5 week schedule. It was originally designed to go up to a final limit of 73 Gy, but this didn't happen in practice as tumour dose was not high enough to reach this threshold. Group 2 consisted of patients that were limited by other OAR dose constraints. Between September 2010 and March 2013, 120 patients from 10 UK centres were enrolled, with 118 starting treatment. Two patients did not start treatment, one patient due to deterioration and a second patient completed only one cycle of chemotherapy and did not finish RT owing to toxicity. There were two treatment related deaths in the study from 4 recorded grade 5 toxicities, although only two were at least possibly related to radiotherapy. One 6 week patient who received 71Gy to 1cc of the oesophagus experienced a grade 5 perforation, while the other fatality due to hemoptysis 14 months after RT was considered as possibly treatment related. Two more patients died from hemoptysis, but these were considered unrelated to treatment. Grade 2-5 RTPN toxicity was seen in 25% of patients and Grade 2-5 oesophagitis in 81% of patients. There were no ≥ grade 4 events for oesophagitis or pneumonitis.

1.5 Sample Size Calculation

Numerous predication models are published in medical literature each year, many are developed using datasets that are too small in terms of the total number of patients or outcome events. This can lead to inaccurate predictions that can lead to incorrect healthcare decisions. Riley et al [65] have provided guidance for calculating sample sizes. It is important that sample sizes are large enough to ensure that the results of modelling are applicable to new individuals in the target population.

For this the purposes of this study, the number of patients available for analysis is fixed, so these sample size calculation techniques can be used to determine the margin of error in the overall outcome proportion estimate.

A simple method is to calculate the sample size needed to precisely estimate the intercept in a model when no predictors are included. For binary outcomes, equation 1.13 may be used

where n is the number of patients, δ is the confidence interval and $\hat{\phi}$ is the anticipated outcome proportion.

$$n = \left(\frac{1.96}{\delta}\right)^2 \hat{\phi}(1-\hat{\phi}) \quad (1.13)$$

For the IDEAL-CRT patients, 67% presented with grade two or higher oesophagitis and 22% so presented with grade two or higher RTPN. Assuming 95% confidence interval, the study would require 340 patients to model oesophagitis and 264 patients to model RTPN. This is far excess of the number of patients available for analysis. Rearranging this equation gives us the ability to calculate the confidence interval for a set number of patients as shown in equation 1.14.

$$\delta = \hat{\phi} \pm 1.96\sqrt{\frac{\hat{\phi}(1-\hat{\phi})}{n}} \quad (1.14)$$

A total of 116 patients were available from the IDEAL trail, so using the previously discussed outcome proportions, the confidence intervals of models derived from the IDEAL data for oesophagitis and RTPN would be 91.4% and 92.5% respectively.

In addition to predicting the average outcome value precisely, the sample size for model development should also aim for precise predictions across the spectrum of predicted values. For binary outcomes van Smeden et al[66,67] use simulation across a wide range of scenarios to evaluate how the error of predicted outcome probabilities from a developed model depended on various characteristics of the development dataset sample from a target population. They found that total sample size, candidate predictor parameters and outcome proportion were the three main drivers of a model's predictive accuracy. This led to the development of equation 1.15.

$$\ln(MAPE) = -0.508 - 0.544\ln(n) + 0.259\ln(\phi) + 0.504\ln(P) \quad (1.15)$$

Where n is the sample size, ϕ is the anticipated outcome proportion and P is the number of candidate predictor parameters. MAPE denotes the Mean Absolute Prediction Error. Assuming the use of 3 predictor parameters for the 116 patients with ϕ of 0.67 and 0.22 for oesophagitis and RTPN respectively. The equation gives a MAPE of 92.9% for oesophagitis and 94.7% for RTPN.

Finally, we must consider the sample size required to minimize the problem of overfitting. Overfitting is when a developed model's output matches the sample dataset to closely, and

therefor is not generalisable to real world data. This notably occurs when the sample size is too small. Particularly when the number of candidate predictors is large relative to the sample size or the number of events within the sample group. The essential characteristic of an overfitted model is that the model's predictive performance is overly optimistic and when used with new data the model's performance will be much lower.

Shrinkage methods deal with the problem of overfitting by reducing the variability in models. The magnitude of shrinkage required is estimate from the development dataset and can fail when the sample size is too small. Riley et al[68,69] suggest identifying the sample size and number of candidate predictors that correspond to a small amount of shrinkage (<10%) during model development.

For binary outcomes Riley et al showed that the sample size (n) needed to achieve an expected uniform shrinkage factor of S can be expressed as below

$$n = \frac{P}{(S-1)ln\left(1-\frac{R^2_{CS}}{S}\right)} \tag{1.16}$$

Riley et al suggested targeting a shrinkage of ≤10% such that S ≥0.9. R^2_{CS} is the Cox-Snell R squared statistic, this value is important as it reflects the signal to noise ratio which has an impact on the estimate of multiple parameters and the potential for overfitting. When a low signal to noise ratio is present it becomes more difficult to identify true patterns and so models become naturally less reliable. Riley has suggested that in the absence of any other information sample sizes should be calculated an R^2_{CS} value of 0.15 which corresponds to a variance of 15%. With a sample size of 116 patients, this would allow the use of 2 candidate predictors assuming a shrinkage of ≤10% and a variance ≤15%.

1.6 Summary

There are numerous examples of NTCP modelling in the literature, although results tend to vary and there is not a consensus regarding which values to use clinically. Further studies are continuing to improve the accuracy of these models and take into account the changes due to modern radiotherapy techniques such as IMRT, VMAT and SABR. Some of this work is focused on using more sophisticated models such as the LKB model and machine learning for NTCP modelling and correcting the data for radiobiological effectiveness of the doses received

by OARs. NTCP modelling has its limitations, and the use of clinical trials data allows researchers to overcome some of these limits particularly with regards to the quality and consistency of the datasets. Researchers must still test the robustness of their data, as individual results are capable of data skewing the results of NTCP models. Gulliford et al presented a number of different methods that may be used to do this.

Data from the IDEAL-CRT trial has substantial variation in the doses received to OARs and a reasonably high occurrence of toxicity. This would infer that data from this trial would be suitable for NTCP modelling using concepts such as BED or EQD2 in an attempt to standardise the doses these patients received. The number of patients in this trial is relatively small and further analysis would be required to confirm any finding of this analysis.

1.7 Research hypothesis and objective

1.7.1 Primary objective:

This study aims to generate radiotherapy toxicity models for NSCLC patients treated with concurrent chemotherapy and isotoxic radiotherapy based on equivalent dose in 2Gy fractions (EQD2) to allow a single set of dose constraints to be used for non-standard lung radiotherapy treatment fractionations. Models will be generated for the prediction of radiation induced oesophagitis and pneumonitis using clinical and dosimetric information collected as part of the IDEAL-CRT trial.

1.7.2 Secondary objective/s:

Evaluation the performance and utility of Lyman Kutcher Burman NTCP modelling, Logistic Regression and Supervised Machine Learning for modelling radiotherapy induced toxicity.

1.7.3 Study Design

The study is a retrospective analysis of the radiotherapy planning, clinical and associated reported toxicity for patients from the IDEAL-CRT trial. IDEAL-CRT was a phase I/II trials of concurrent chemoradiation with dose escalated radiotherapy in patients with stage II/III non-small cell lung cancer. The study will utilise this data to correlate grade 2 or higher oesophagitis and pneumonitis with dosimetric and clinical parameters converted to EQD2 using radiobiological modelling. Data will be analysed using the Lyman Kutcher Burman NTCP model, Logistic Regression and Supervised Machine Learning.

1.7.4 Number of subjects

A total of 116 patients will be reviewed.

1.8 Author contribution to chapters

This study was designed by the author. For this study the author successfully completed a Health Research Authority (7.1) application to gain ethics approval for this project with the support of the IDEAL-CRT chief investigator and the CRUK & UCL clinical trials unit (7.2) which allowed the arrangement of a data sharing agreement between East and North Hertfordshire NHS trust and the UCL CTC. The author gathered all clinical and toxicity data from the trials unit and radiotherapy data from the national radiotherapy trials quality assurance group, then performed the processing and analysis of all data reported.

1.9 Thesis Rationale

This thesis analyses evidence from a non-randomised clinical trial of isotoxic radiotherapy for non-small cell lung cancer. It was written in journal format as the intention is to publish the results in an oncology journal, as the results have the potential to inform clinical decisions for a patient benefit and direct future research.

The structure of the thesis was designed so that the first paper which examined the statistical significance of data features available for analysis, the second paper used this information to inform toxicity modelling using supervised machine learning and the third paper compared the results of machine learning toxicity modelling with Lyman Kutcher Burman and logistic regression modelling to determine the best approach for future work.

2 Paper 1: Analysis of the significance of clinical and dosimetric factors for the prediction of radiation induced oesophagitis and pneumonitis

Rushil Patel[1], Karen Venables[1], Adam Aitkenhead[2], Laura Farrelly[3], Nicholas Counsell[3], David Landau[4]

1 Mount Vernon Cancer Centre, Northwood, UK

2 The Christie NHS Foundation Trust, Manchester, UK.

3 Cancer Research UK & UCL Cancer Trials Centre, London, UK

4 Guys & St Thomas NHS Trust, London, UK

2.1 Abstract

Background and Purpose: The lung and oesophagus are very sensitive to the effects of radiation. Radiation induced oesophagitis and pneumonitis can often be limiting factors in radiotherapy treatments for NSCLC patients. The study reported here uses data from the IDEAL-CRT trial to evaluate the utility of clinical and dosimetric factors for use as predictors of oesophagitis and pneumonitis.

Methods: Descriptive statistics were used to explore the dosimetric and clinical data from a total of 116 patients from the IDEAL-CRT trial, these patients then underwent uni-variable analyses using logistic regression to determine if there was a statistically significant correlation between any of the data features and the occurrence of grade ≥ 2 oesophagitis or pneumonitis.

Results: Of the 116 patients analysed, 27 (23%) and 78 (67%) suffered from grade ≥ 2 pneumonitis and oesophagitis respectively. Logistic regression analysis demonstrated a statistically significant corelation ($p<0.008$) between the Mean Dose, V35Gy, V50 Gy, D1cc dose metrics and the occurrence of grade ≥ 2 oesophagitis. Logistic regression analysis was unable to find a statistically significant link between any clinical or dosimetric factor for pneumonitis, FEV percent predicted, FVC percent predicted, V20Gy and Mean Dose all had P-values <0.1.

Conclusion: For organs such as the oesophagus there is no consensus in the literature as to the most appropriate predictors for toxicity. The results of this paper show an encouragingly

strong link between a number of dose metrics and the occurrence of oesophagitis. The results of this paper do not show a strong link between any of the predictors and pneumonitis, with lung function tests showing the strongest correlation. For both oesophagitis and pneumonitis, toxicity modelling using a wider range of input data and more complex modelling techniques such as machine learning may be able to build clinically useful toxicity models.

2.2 Introduction

Lung cancer is the primary cause of death by cancer in Europe with over 385,000 cases per year and an approximate death rate of 52.2/100,000 persons per year[70]. Chemo-radiotherapy is the gold standard therapeutic option for patients with locally advanced lung cancer that are ineligible for surgery[71]. Chemo-radiotherapy is a treatment that utilises both chemotherapy and radiotherapy. Chemotherapy uses anti-cancer drugs to destroy cancer cells, whilst radiotherapy uses x-rays to treat cancer cells. The chemotherapy drugs make cancer cells more sensitive to radiation and so giving these treatments concurrently is more effective than having either treatment on their own or delivered sequentially[71].

For the treatment of lung cancer, the two most significant side effects are oesophagitis and radiation induced pneumonitis (RTPN). The current chemoradiotherapy regimes result in a 15-25% rate of severe acute oesophagitis[60], with acute toxicity defined as occurring within 90 days of treatment and severe toxicity defined as grade 3 or higher according to the Radiotherapy Therapy Oncology Group (RTOG) scoring criteria. For the oesophagus late injury is less common, which may reflect more on the low survival time for lung cancer patients which results in less follow up data. Clinically significant RTPN develops in approximately 5-50% of patients that undergo radical radiotherapy for lung cancer[72].

The lung is sensitive to the effects of radiation. RTPN is an inflammation of the lung parenchyma caused by the delivery of radiotherapy to tumours within or in close proximity to lung tissue. The radiation dose to the chest results in the lungs producing less surfactant, a substance which helps keep the air passages open. A lack of surfactant can inhibit the lungs from fully expanding leading to symptoms such as cough, chest congestion, shortness of breath, and chest discomfort. There is a correlation between increased likelihood of developing pneumonitis with high doses of radiation, irradiating large volumes of lung tissue and certain chemotherapy drugs[73]. Radiation induced pneumonitis is one of the major

toxicities which limits the maximum radiation dose that can be safely delivered to tumours in the thoracic region. Its severity ranges from asymptomatic where cases can only be detected radiographically to clinically evident where patients have cough, shortness of breath, or fever. In the most severe cases there may be dense fibrotic lung changes and respiratory compromise which require patients to receive supplementary oxygen or assisted ventilation.

Radiation induced oesophagitis is the inflammation of the oesophagus due to radiation. The symptoms present 2 to 3 weeks after initial therapy and include throat pain and dysphagia. The cells that form the lining of the oesophagus are particularly vulnerable to chemotherapy and radiation, so patients undergoing chemo-radiotherapy are particularly susceptible to oesophagitis as a treatment side effect. Severe oesophagitis will affect a patient's ability to maintain nutrition and hydration, and so occurrences must be managed to ensure that oesophageal toxicity does not interrupt radiotherapy treatment, as delays in radiation delivery could impact upon patient outcomes.

The aim of radiotherapy is to achieve locoregional control of cancer. To do this we must balance the likelihood of tumour control and the risks associated with toxicity to ensure there is sufficient benefit to the patient. The more information we have regarding the effects of radiation exposure for an individual patient the better we are able to personalise their treatment to give them the optimum balance of tumour control and toxicity. To predict the probability of toxicity, we can look at selected clinical demographics about the patient and their health as well as dosimetric factors regarding their radiotherapy treatment plan. In routine clinical practice constraints are placed on doses to organs at risk (OARs) to limit the probability of side effects to acceptable levels. Clinical information from the patient may be used to influence treatment decisions but is not routinely used to predict toxicity. For the treatment of lung cancer there are a number of different treatment fractionation regimes available for patients, each of these having their own dose constraints[9,10]. With the advent of Isotoxic radiotherapy where treatment doses are escalated until an OAR constraint is met and the widespread use of Stereotactic Ablative Body Radiotherapy (SABR)[74] for the treatment of oligometastases, there is substantial variation in the radiation doses that OARs receive. Due to the variation in practice and the evidence base it can be difficult to establish which dose constraints to use for the respective OARs. Additionally, it would be useful to have OAR constraints based on Equivalent Dose in 2 Gy fractions (EQD2); this would allow a single set

of dose constraints to be used for all patients regardless of their treatment regime. Data from the IDEAL-CRT[10] trial, which utilised an Isotoxic fractionation scheme with a fixed number of fractions would provide a useful data set to generate these values.

The IDEAL-CRT trial was a non-randomised study of radiotherapy (RT) concurrent with two cycles of cisplatin and vinorelbine for patients with non small cell lung cancer (NSCLC), with inclusion criteria of confirmed stage IIA to IIIB and a number of other performance indicators. In total 120 patients were recruited from 8 treatment centres, of which 118 started treatment. IDEAL-CRT was an isotoxic trial and so tumour doses were prescribed to the highest possible dose to the tumour whilst meeting dose constraints for normal tissue and target coverage. This meant that doses to targets varied for each patient based on the size of their tumour and the proximity of critical structures. Previously established dose limits existed for lungs, spinal cord, brachial plexus, and heart, but not for oesophagus, so the trial incrementally increased the limit during the trial. Patients were split into two groups, Group 1 was limited by oesophageal dose, which was incrementally escalated from 65 Gy to 71 Gy. The trial planned to allow up to 73 Gy, but this was not achieved in practice as the tumour dose was not sufficiently high to reach this threshold. Group 2 were limited by other dose constraints. Patients were recruited between September 2010 and March 2015. Data was collected for the doses received by nearby organs, which included healthy lung tissue, heart, oesophagus, brachial plexus, and the spinal cord. Radiotherapy treatments were planned using 3D conformal, and intensity modulated radiotherapy (IMRT) techniques which were optimised to the individual patient, meaning that the doses to healthy organs varied depending on the specifics of each patient. The median follow up for toxicity was 50 months. Toxicity and additional clinical factors were collected using Case Report Forms (CRFs).

The aim of this analysis is to evaluate the utility of clinical and dosimetric factors for use as predictors of oesophagitis and RTPN for patients from the IDEAL-CRT trial.

2.3 Methods

2.3.1 Patients

Of the 120 patients enrolled in the IDEAL-CRT trial, 116 patients with available does volume histogram (DVH) data were available for analysis, 2 patients did not start treatment and 2 patients did not have a full data set available. All patients were treated as per trial protocol

with an Isotoxic regimen with the tumour receiving between 63 Gy and 73 Gy in 30 fractions over 40 days[10] delivered concurrently with 2 cycles of cisplatin and vinorelbine chemotherapy on days 1 and 29 of RT. Toxicity was graded according to the National Cancer Institute (NCI) Common Terminology Criteria for Adverse Events (CTCAE) Version 4[75].

2.3.2 EQD2 conversion

The dose to OARs for patients in the IDEAL-CRT trial was dependent on the level of dose escalation achieved and the location of the OAR with respect to the tumour location. To make the findings of this paper applicable to any dose fractionation, the Dose Volume Histograms (DVHs) were converted to EQD2. The basic Linear Quadratic (LQ) model describes the surviving fraction (SF) of clonogenic or stem cells as a function of radiation dose. The main parameters of this model are α and β which represent the intrinsic radio-sensitivity of an irradiated cell. The LQ model uses an α/β ratio, which is a measure of the fractionation sensitivity of the cells.

$$BED = N \times d \times \left[1 + \frac{d}{(\alpha/\beta)}\right] - k(T - T_{delay}) \tag{2.1}$$

This model allows the conversion of any dose treatment regimen into a Biologically Effective Dose (BED)[76] and is given in equation 2.1. For ease of comparison the LQ model allows any unusual fractionation to be expressed in 2 Gy fractions (EQD2) as per equation 2.2.

$$EQD2 = \frac{BED}{1 + \frac{2}{(\alpha/\beta)}} \tag{2.2}$$

For the purposes of this analysis, an α/β ratio of 4 was used for the lung, taken from the QUANTEC analysis[77,78]. For the oesophagus an α/β ratio of 10, commonly used in the literature[79,80] was used. The k term representing the dose recovered per day was set at 0.54Gy and 0.8Gy for the lung and oesophagus respectively, both values were from a reviews of clinical studies by Bentzen et al[81,82]. The T_{delay} was set at 28 days for both lung and oesophagus as per Fowler et al[83], with the T value of 33 days and 40 days for the 5 and 6 week treatment schedules respectively. Overall treatment time was strictly mandated in the trial guidelines.

Digitally Imaging and Communications in Medicine (DICOM) data was imported into the Eclipse v15.6 (Varian Medical Systems) treatment planning system, DVHs were calculated for the oesophagus and Lung-GTV (the whole lung volume excluding the GTV) structures. DVHs were calculated with 0.01 Gy dose bins to allow for high precision dose metrics to be calculated. This data was then exported to a text file using inbuilt functionality of Eclipse for all patients. This data was then processed using an in-house programme written in MATLAB v2019b (Appendix 5). The programme parses the DVH text file, converts the dose into EQD2 using the appropriate α/β ratio and then reports the desired dose metrics for the required structures. The software iterates through all DVH files and writes the dose metrics for each patient into a single text files which were imported into Microsoft Excel and combined with clinical and toxicity data reported from the IDEAL-CRT trial[10].

2.3.3 Statistics

Data were analysed with SPSS software (IBM) v28. Clinical factors collected through the IDEAL-CRT trial included sex, age, disease stage and Forced Expiratory Volume (FEV), Force Vital Capacity (FVC) and diffusing capacity of lung for carbon dioxide (DLCO). Dosimetric factors were generated using established constraints from QUANTEC[60,77] for the initial analysis, which included mean dose and V20 Gy for Lung-GTV and mean dose, V35 Gy, V50 Gy, and V70 Gy for the oesophagus. These factors were classified as continuous (dose metrics, age, FEV, FCV, DCLO) or categorical (sex, disease stage) data, with data classification determining the types of statistical tests that were performed. Oesophagitis and RTPN were analysed using the occurrence of CTCAE V4.0 grade 2 or 3 toxicity within 6 months of treatment as the outcome.

2.3.3.1 Data Characterisation

Descriptive statistics were calculated on the dataset to better explore the data. For categorical data, pie charts were plotted to show the distribution within the sample population. For continuous dosimetric data, scatter plots with data colour coded depending on their respective toxicity status and box and whisker in which dosimetric data was grouped by toxicity status were used to demonstrate the differences between the groups for key variables. Finally, Pearson's correlation coefficients were calculated to measure the strength of linear association between the dosimetric variables. Scatterplot matrices were also created to visually show the correlation between the different dosimetric parameters.

2.3.3.2 Data modelling

Data underwent binomial logistic regression to establish which factors were the strongest predictors of toxicity. Binary logistic regression predicts the probability that an observation falls into one of two categories of a dichotomous dependent variable based on one or more independent variables. In this case uni-variable analysis was performed to assess predicative ability of each variable. Due to the relatively small data set available for this research, it was not feasible to test all the predictive factors together but through uni-variable analysis it may be possible to determine which factors would be the most useful for further modelling. To interpret the results, we used the Wald test which determines the statistical significance for each independent variable and significance of the binary logistic regression test. These two values can help us determine which factors contribute to the prediction of toxicity and which can be removed for further modelling.

Binary logistic regression provides an estimate of the probability of an event. For the results of this analysis, if the estimated probability of the event occurring is equal to or greater than the cut-off value 0.5, SPSS classifies this event as occurring. If the probability is below 0.5 SPSS classifies the event as not occurring. This essentially provides the sensitivity and specificity of the observed model. The Wald value quantifies the magnitude of the impact of the variable on the model, and the p-value determines whether the result is statistically significant.

A bootstrapping method was applied to the data in which a thousand different cohorts of 116 patients were generated by resampling the original dataset with replacement. These cohorts were then fitted using logistic regression and combined by averaging the output. The bootstrapping method allows us to determine the 95% confidence intervals of the logistic regression analysis as well generating a bootstrapped average p value which informs us of the robustness of our p-value. A bias value is also generated, the smaller the bias the better the estimated coefficient.

As multiple statistical tests have been run for a number of variables in the same dataset, a multiple comparison correction must be applied. In this case a Bonferroni correction has been applied to generate a new p-value, using $p = {}^{\alpha}/_{n}$ where alpha is the original significance level and n is the number of tests performed.

2.3.4 Additional dosimetric parameters

Additional dosimetric parameters were also tested to determine whether they could provide an improvement in the correlation with toxicity. These included factors from that are commonly used in the literature. For the oesophagus, small volume doses such as the D1cc have been used in clinical trials[84,85]. For the Lung, V5 Gy[86] and Equivalent uniform dose (EUD)[40] have been used in the literature.

2.3.4.1 EUD Calculation

Previous studies have been published predicting RTPN using NTCP values based on the generalised Lyman model[29]. This model incorporates a flexible dosimetric parameter, which is known as Equivalent Uniform Dose (EUD)[85]. EUD allows sub volumes of an OAR that have been irradiated to be weighted differently. Sub volumes of different organs are weighted by the volume parameter n. When n=1 each sub volume is weighted equally and the EUD corresponds to the average dose which is the same as mean dose (MD). When n<1 sub volumes receiving a higher dose are weighted more heavily. Earlier analysis has shown that RTPN correlates better using EUD[40] with an n = 0.5 than with n = 1.

The EUD can be calculated as per equation 2.3

$$\mathrm{EUD} = \left(\sum \mathrm{D_i^{\frac{1}{n}}} \times \mathrm{v_i} \right)^{\mathrm{n}} \tag{2.3}$$

Where v_i is the volume of the dose bin receiving dose D_i, summed over all dose bins in the DVH. The sum is then raised to the power n to revert to dose units. The n parameter affects the relative weighting of the subvolumes.

2.4 Results

2.4.1 Incidence of Pneumonitis and Oesophagitis

Of the 116 patients available for analysis, 27 (23%) experienced CTCAE v4.0 grade ≥2 RTPN and 3 (3%) patients experienced grade ≥3 RTPN. For oesophagitis 78 (67%) patients experienced CTCAE v4.0 grade ≥2 toxicity and 7 (6%) patients experienced grade ≥3 oesophagitis. Due to the low incidence of grade ≥3 events the analysis will focus on grade ≥2 toxicity for both RTPN and oesophagitis.

2.4.2 Statistical Analysis

Figures 1-8 and Tables 1-10 contain the descriptive statistics performed on the IDEAL-CRT trial data, allowing exploration of the dataset and examining the correlations between the various dosimetric variables. Additional plots are available in Appendix 6.

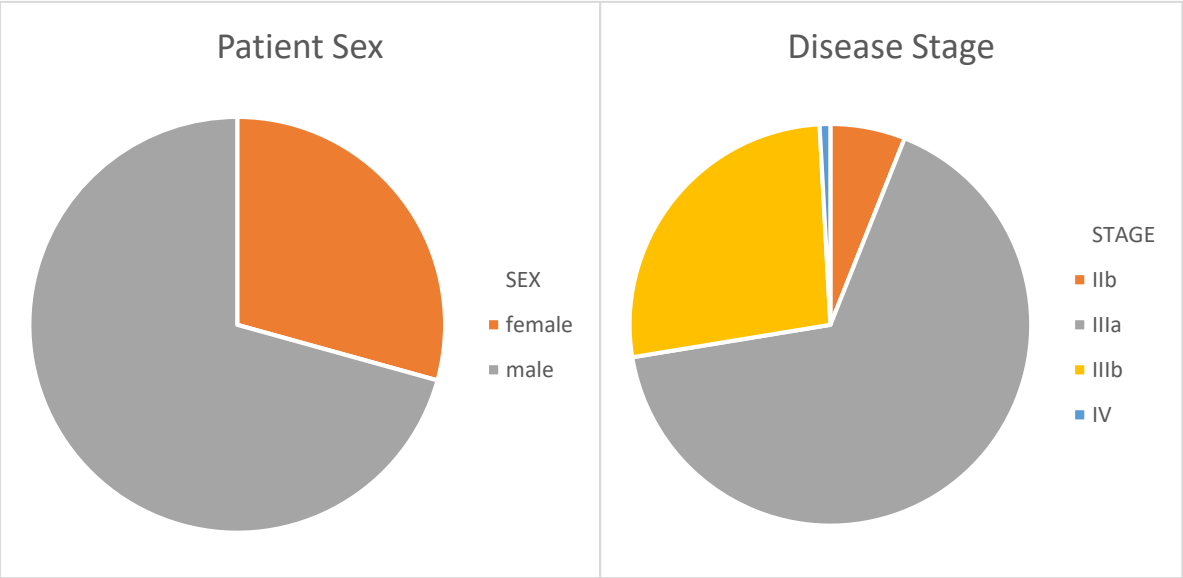
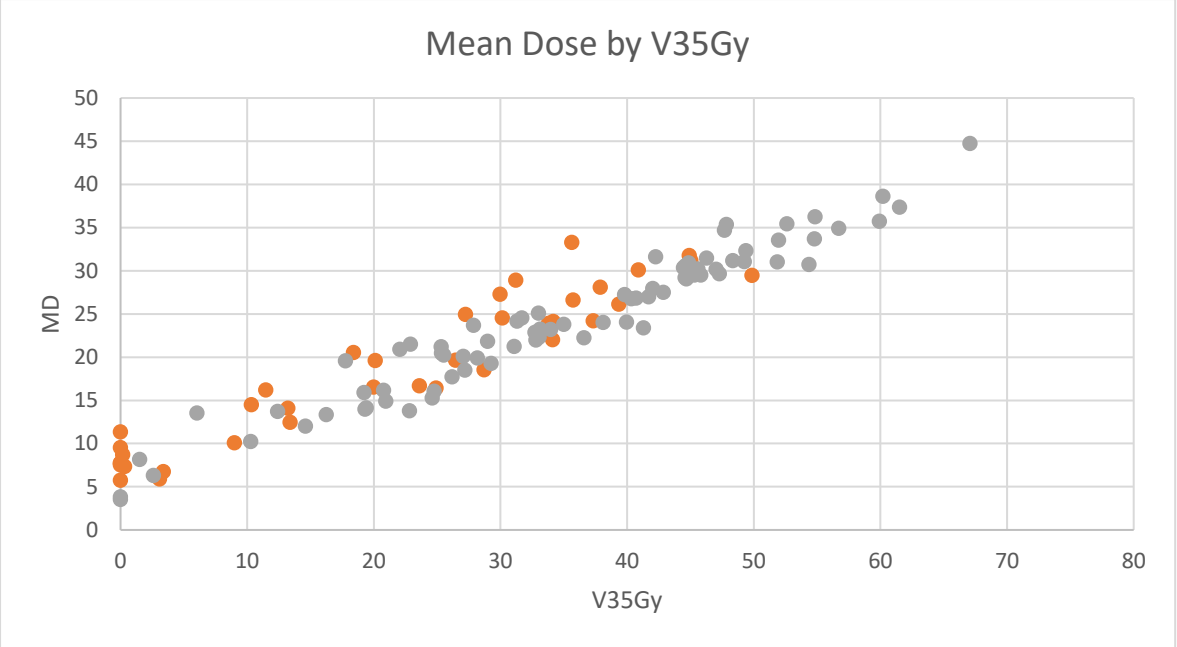


Figure 1 Pie charts of patient gender and disease stage for the IDEAL-CRT trial

Variables	Minimum	Maximum	Mean	Std. Deviation	Variance
AGE	42.60	83.70	65.64	7.67	58.88
D1cc	13.38	101.73	61.20	14.92	222.69
V35Gy	0.00	67.08	30.39	16.67	277.86
V50Gy	0.00	57.01	22.06	15.28	233.54
MD	3.55	44.74	22.32	8.83	78.01

Table 1 Descriptive statistics of all variables which may be used to predict Oesophagitis



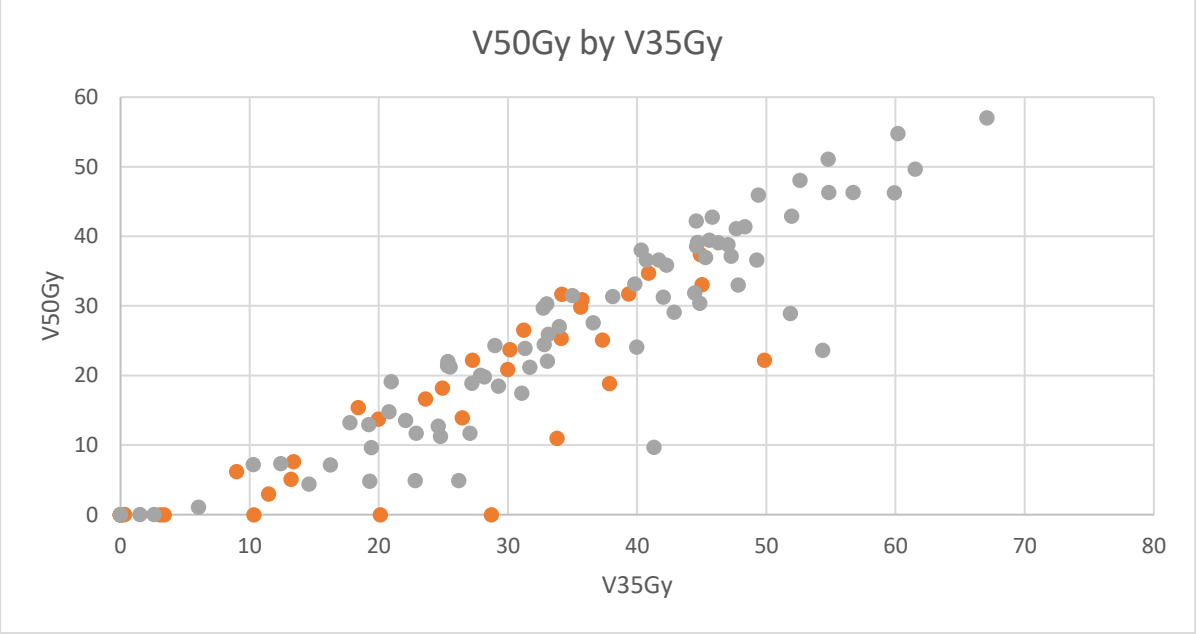
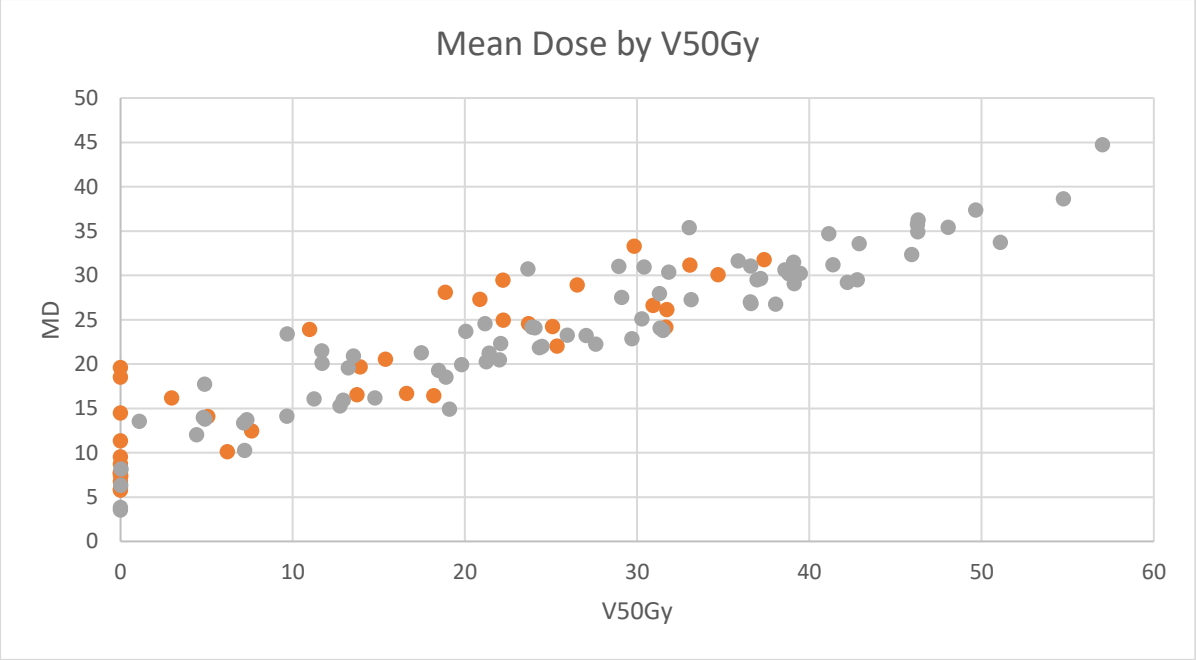


Figure 2 Scatterplots of Mean Dose and V50Gy to the Oesophagus tissue against dose V35Gy. Orange dots denote those patients that did not suffer ≥G2 Oesophagitis while grey dots denote those that did



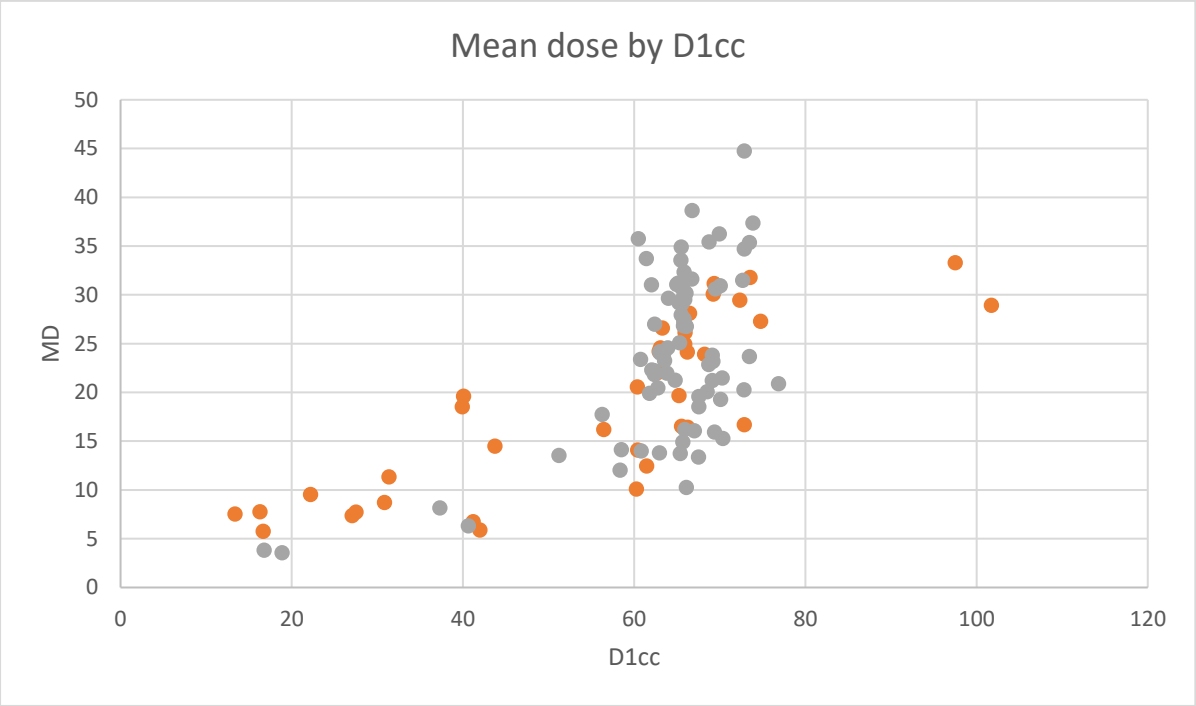


Figure 3 Scatterplots of Mean Dose to the Oesophagus tissue against dose V50Gy and D1cc. Orange dots denote those patients that did not suffer ≥G2 Oesophagitis while grey dots denote those that did

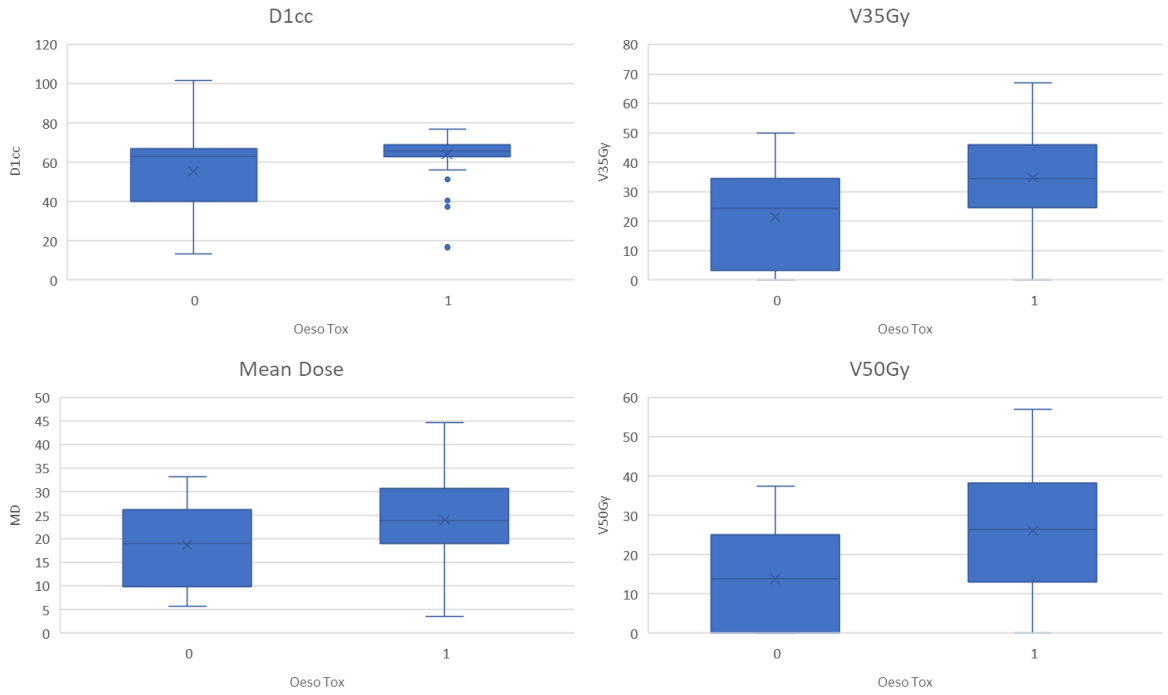


Figure 4 Box and whisker plots of the Oesophageal V35Gy, V50Gy, Mean Dose and D1cc, where box zero is for patients that did not suffer from ≥G2 Oesophagitis and one is for patients that did

Oesophagitis Correlations					
		D1cc	V35Gy	V50Gy	MD
D1cc	Pearson Correlation	1.00	0.657**	0.600**	0.679**
	Sig. (2-tailed)		0.00	0.00	0.00
V35Gy	Pearson Correlation	0.657**	1.00	0.923**	0.963**
	Sig. (2-tailed)	0.00		0.00	0.00
V50Gy	Pearson Correlation	0.600**	0.923**	1.00	0.923**
	Sig. (2-tailed)	0.00	0.00		0.00
MD	Pearson Correlation	0.679**	0.963**	0.923**	1.00
	Sig. (2-tailed)	0.00	0.00	0.00	

*Table 2 contains the results of a Pearson's correlation procedure run to determine the relationship between four commonly used dosimetric parameters which are commonly used to predict Oesophagitis. ** refers to results where the correlations are significant at the 0.01*

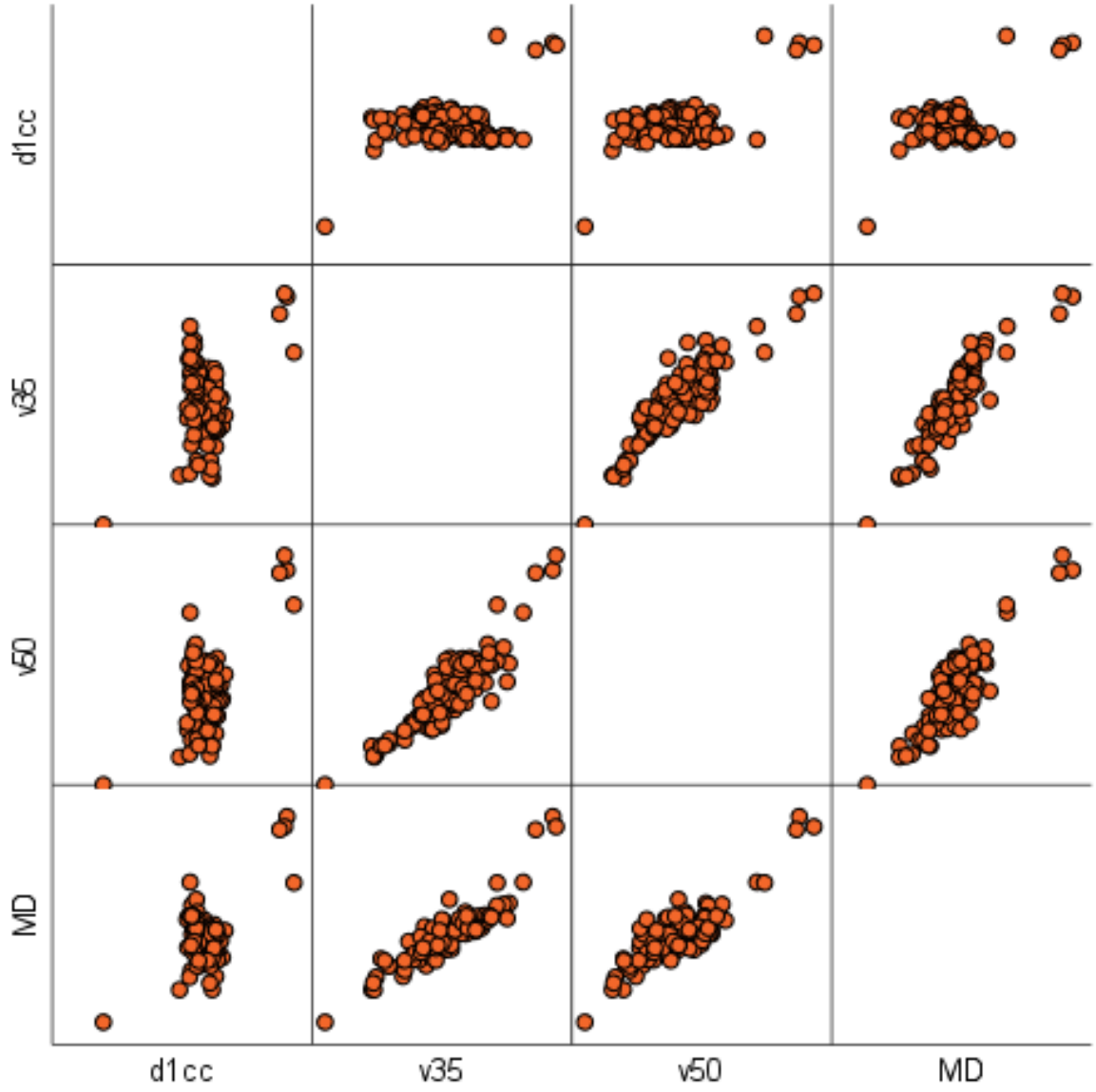
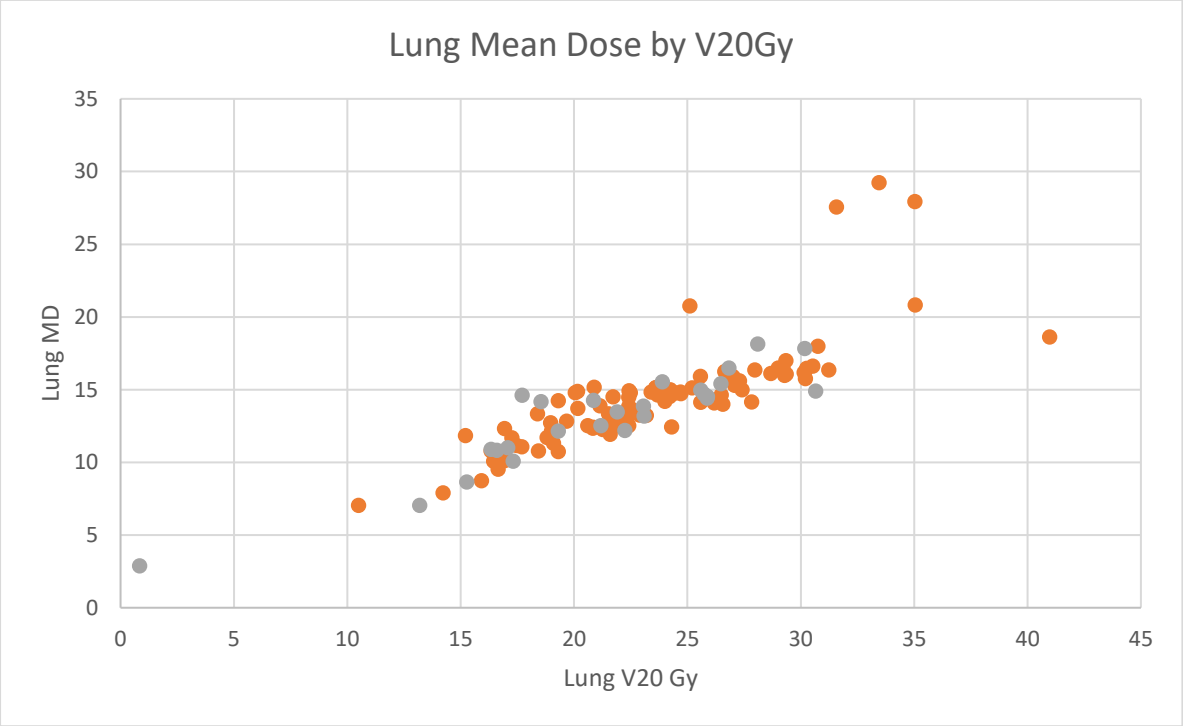


Figure 5 A scatterplot matrix which visually demonstrates the relationship between the various dosimetric predictors of Oesophagitis

Variable	Minimum	Maximum	Mean	Std. Deviation	Variance
AGE	42.60	83.70	65.64	7.67	58.88
FVC Percent Predicted	48.00	136.00	94.83	18.84	354.90
FEV Percent Predicted	36.80	147.00	76.65	20.25	409.94
DCLO Percent Predicted	41.70	111.00	70.52	17.69	312.90
EUD	6.52	46.59	24.60	4.83	23.28
V5Gy	26.61	73.46	50.16	10.17	103.37

V20Gy	0.85	40.98	23.03	5.51	30.39
MD	2.89	29.24	14.08	3.51	12.30

Table 3 Descriptive statistics of all variables which may be used to predict RTPN



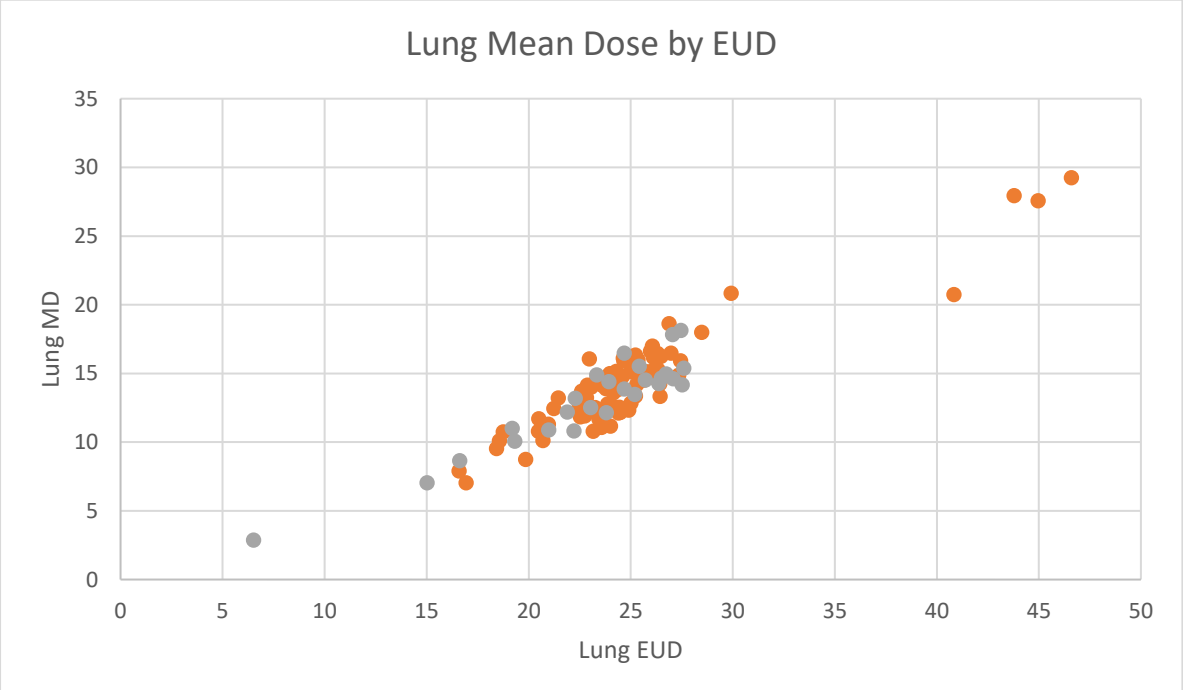


Figure 6 Scatterplots of Mean Dose to the healthy Lung tissue against dose V20Gy and Equivalent Uniform Dose. Orange dots denote those patients that did not suffer ≥G2 RTPN while grey dots denote those that did.

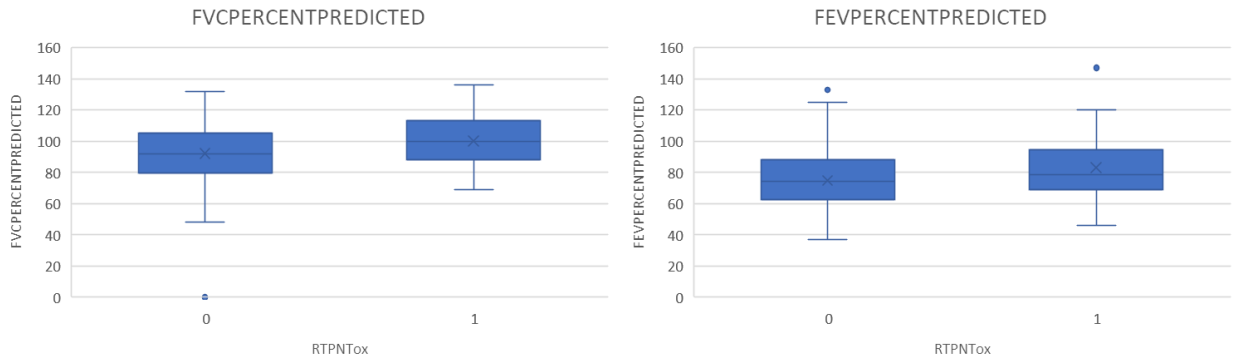


Figure 7 A box and whisker plot of patient lung function. These represent the Forced Vital Capacity and Forced Expiration Volume for patients compared with their expected values determined on baseline factors such as age, ethnicity, and height, where box zero is for patients that did not suffer from ≥G2 RTPN and box one is for patients that did

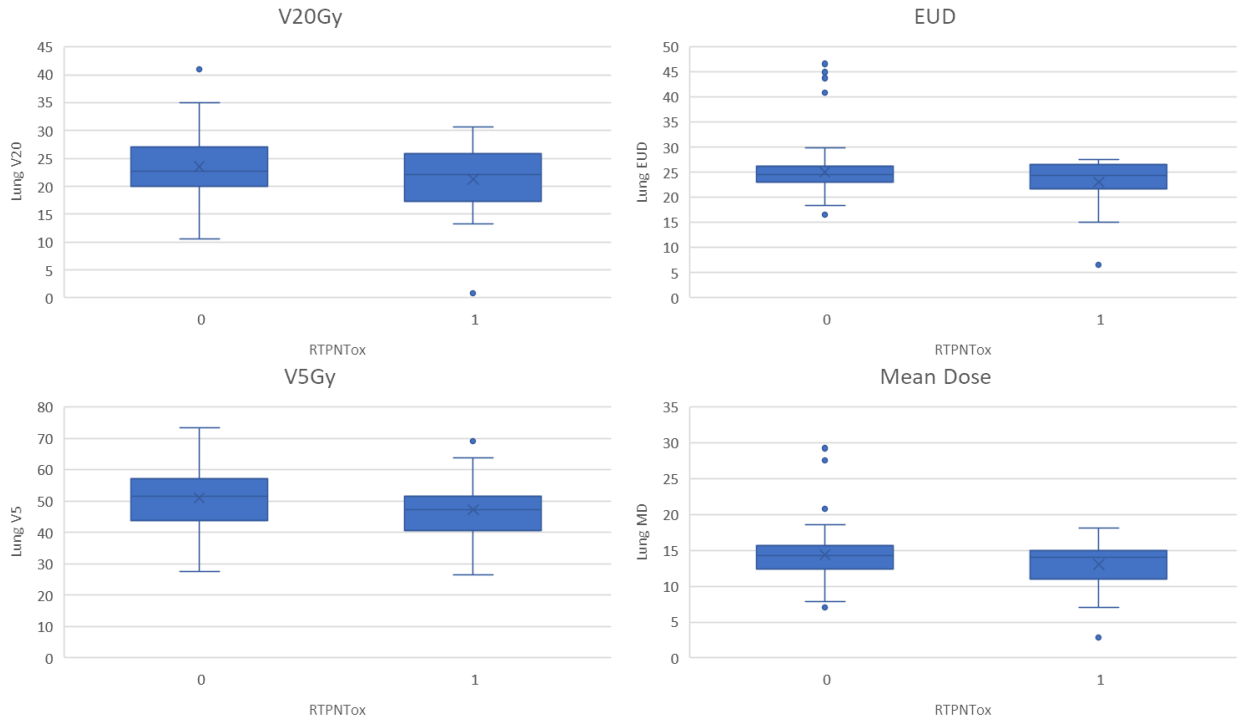


Figure 8 Box and whisker plots of the Lung V20Gy, V5Gy, Mean Dose and Equivalent Uniform Dose, where box zero is for patients that did not suffer from ≥G2 RTPN and box one is for patients that did

RTPN Correlations					
		Lung EUD	Lung V5	Lung V20	Lung MD
Lung EUD	Pearson Correlation	1.00	.512**	.653**	.930**
	Sig. (2-tailed)		0.00	0.00	0.00
Lung V5	Pearson Correlation	.512**	1.00	.735**	.746**
	Sig. (2-tailed)	0.00		0.00	0.00
Lung V20	Pearson Correlation	.653**	.735**	1.00	.831**
	Sig. (2-tailed)	0.00	0.00		0.00
Lung MD	Pearson Correlation	.930**	.746**	.831**	1.00

	Sig. (2-tailed)	0.00	0.00	0.00	

*Table 4 contains the results of a Pearson's correlation procedure run to determine the relationship between four commonly used dosimetric parameters which are commonly used to predict RTPN. ** refers to results where the correlations are significant at the 0.01 level (2-Tailed)*

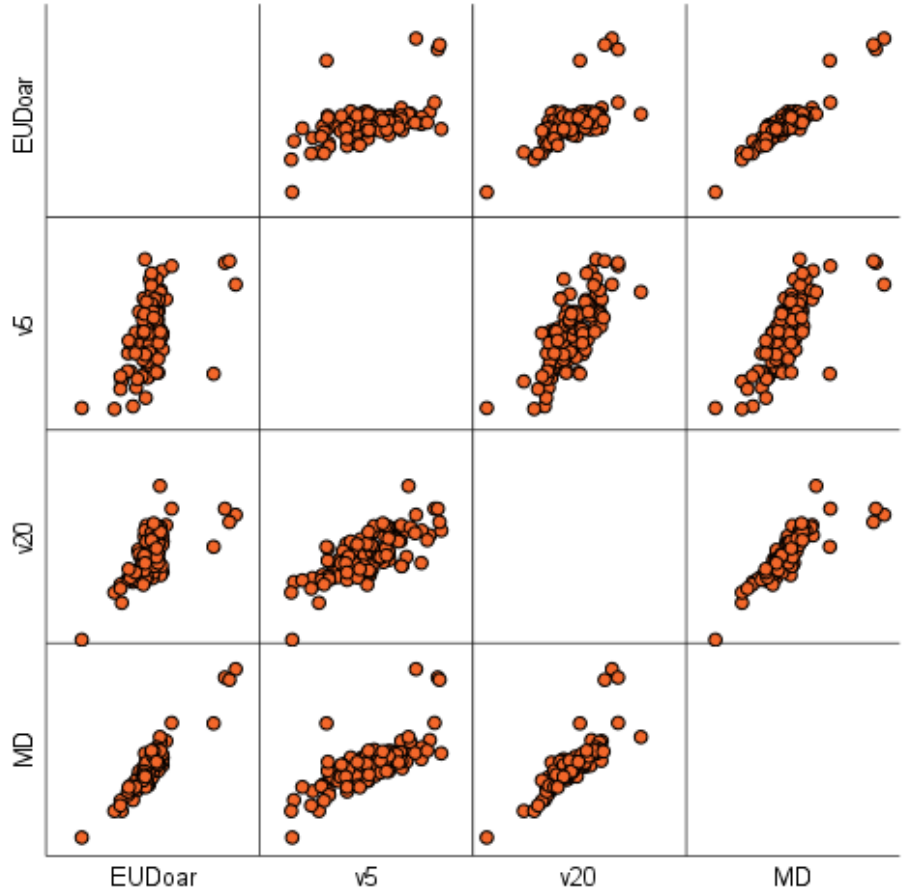


Figure 9 A scatterplot matrix which visually demonstrates the relationship between the various dosimetric predictors of RTPN

Table 5 shows the results of binary logistic regression analysis for oesophagitis, Wald value and p-value are reported for each tested variable. Table 6 shows the sensitivity, specificity, overall predictive accuracy, while Table 7 shows the results of the bootstrapping method. Boxplots of the four most significant predictive factors for oesophagitis are shown in Figure 4.

	B	Wald	Sig.
Sex	0.852	3.133	0.077
Age	-0.069	5.479	0.019
D1cc	0.039	7.576	0.006
V35Gy	0.053	14.609	0.000
V50Gy	0.061	14.839	0.000
MD	0.073	8.955	0.003

Table 5 Results of binary logistic regression analysis for factors predicting Oesophagitis

Predictive Accuracy (%)			
	TN	TP	Overall
Sex	0	100	67.2
Age	10.5	97.4	69
D1cc	28.9	94.9	73.3
V35Gy	39.5	89.7	73.3
V50Gy	44.7	84.6	71.6
MD	28.9	93.6	72.4

Table 6 The predictive accuracy associated with the binary logistic regression tests for oesophagitis

	B	Bias	Std. Error	Sig. (2-tailed)	95% Confidence Interval	
					Lower	Upper
Sex	0.852	0.028	0.515	0.067	-0.035	2.043
Age	-0.069	-0.002	0.030	0.018	-0.137	-0.014
D1cc	0.039	0.002	0.020	0.014	0.008	0.086
V35Gy	0.053	0.001	0.014	0.001	0.029	0.084
V50Gy	0.061	0.002	0.015	0.001	0.036	0.095
MD	0.073	0.002	0.026	0.002	0.031	0.131

Table 7 results of the bootstrapping method applied to logistic regression analysis

Table 8 shows the results of binary logistic regression analysis for pneumonitis, Wald value and p-value are reported for each tested variable. Table 9 shows the sensitivity, specificity, overall predictive accuracy, Table 10 shows the results of the bootstrapping method. Boxplots of the four most significant predictive factors for oesophagitis are shown in Figure 8.

	B	Wald	Sig.
Sex	0.542	1.337	0.247
Age	-0.023	0.645	0.422
FVC Percent Predicted	0.021	3.026	0.082
FEV Percent Predicted	0.020	3.395	0.065
DCLO Percent Predicted	0.006	0.212	0.645
EUD	-0.111	3.254	0.071
V5Gy	-0.037	2.567	0.109
V20Gy	-0.076	3.175	0.075
MD	-0.134	2.997	0.083

Table 8 Results of binary logistic regression analysis for factors predicting Pneumonitis

Predictive Accuracy (%)			
	TN	TP	Overall
Sex	100	0	77.6
Age	100	0	77.6
FVC Percent Predicted	100	0	77.6
FEV Percent Predicted	100	3.8	78.4

DCLO Percent Predicted	100	0	77.6
EUD	100	3.8	78.4
V5Gy	100	0	77.6
V20Gy	100	3.8	78.4
MD	100	3.8	78.4

Table 9 The predictive accuracy associated with the binary logistic regression tests for Pneumonitis

	B	Bias	Std. Error	Sig. (2-tailed)	95% Confidence Interval	
					Lower	Upper
Sex	0.542	-0.039	0.824	0.274	-0.452	1.474
Age	-0.023	0.002	0.031	0.439	-0.083	0.043
FVC Percent Predicted	0.021	0.000	0.011	0.035	0.002	0.044
FEV Percent Predicted	0.020	0.000	0.011	0.045	-0.002	0.043
DCLO Percent Predicted	0.006	0.000	0.014	0.673	-0.020	0.035
EUD	-0.111	-0.002	0.067	0.079	-0.256	0.016
V5Gy	-0.037	-0.001	0.026	0.126	-0.091	0.013
V20Gy	-0.076	0.002	0.045	0.070	-0.164	0.018
MD	-0.134	-0.006	0.075	0.062	-0.298	0.004

Table 10 results of the bootstrapping method applied to logistic regression analysis

2.5 Discussion

An important consideration for treatment plan optimisation is the ability to predict the likely risk of toxicity associated with a potential treatment plan. Any reduction in the risk of toxicity must be balanced against the risk of inadequate coverage of the tumour. The prediction of risk is a process that should be tailored to the individual patient, taking into account all relevant clinical and dosimetric data to ensure an optimum therapeutic ratio. The ability to accurately predict of the risk of toxicity is important to ensure treatment plans are optimal and the aim of this study is to demonstrate which predictors would be most useful for toxicity modelling.

When analysing the correlation between the various dosimetric parameters for RTPN, we can see that there is a statistically significant correlation between the four different metrics. It is a similar situation when analysing the parameters for oesophagitis, all four parameters are highly correlated. This would suggest that the benefit of using multiple dosimetric parameters for modelling may be of limited utility, although we must balance this against the fact that multiple dose metrics can give a more complete picture of the DVH which may be more beneficial when using advanced modelling techniques.

The Bonferroni corrected p-values for oesophagitis and pneumonitis were 0.008 and 0.006 respectively. In the case of oesophagitis, the results show that there is a statistically significant correlation between toxicity and Mean Dose, D1cc, V35Gy, and V50Gy dosimetry metrics in agreement with the literature[60,87]. The strongest correlation was seen between the V50Gy metric which had the highest Wald value and joint lowest P value with V35Gy. The overall predicative accuracy of this metric using binary logistic regression was 73.3%, although this model over predicated toxicity in patients that did not develop any. This was a common issue for all metrics tested for oesophagitis, which could potentially be improved by varying the cut-off value for the regression analysis. The age of the patients did not have a significant impact on toxicity in line with Pignon et al[88, 22]. The majority of patients did not receive doses as high as 75Gy and so this metric was not tested due to a lack of data. When the results of the bootstrapping are also analysed, the P-value of the D1cc is no longer significant. The V35Gy, V50Gy and mean dose remain statistically significant.

With regards to radiation induced pneumonitis, the dose metrics recommended by QUANTEC[60,78,89,90] (MD, V20 Gy) did not provide a strong correlation with toxicity for this dataset. Measures of lung function assessed using binary logistic regression performed well, with the FEV having the lowest P-value overall. The FVC performed similarly to the FEV, but the DCLO had very poor correlation with toxicity. The EUD, V20Gy and Mean Dose all had P values <0.1, suggesting that they may have some utility in toxicity modelling when combined with additional data. The EUD did show a slightly stronger correlation with RTPN than MLD, which was in line with the Tucker et al[40], although this also did not prove to be statistically significant. These results do not agree with publications in the literature[77] or with the dose constraints that are commonly used for clinical trials[85,91] where the V20Gy, MLD and V5Gy (only used for 3 of the 5 arms of the ADSCAN trial) constraints are commonly used. Analysis of the bootstrapping showed that FVC and FEV had the lowest P-values overall and that the Mean Dose, V20Gy and EUD all generated p-values less than 0.1 with V20Gy providing the lowest P-value of all dosimetric predictors.

For the IDEAL-CRT trial the dose constraint for the lung was an EQD2 mean of 18.2 Gy[10], with an expectation of a 20% rate of grade 2-5 RTPN[92], with the trial reporting an incidence rate of 23%. This low occurrence of toxicity and the relative homogeneity of the MLD and V20Gy due

to planning dose constraints may indicate that this dataset may not be well suited to determine the dosimetric factors with high correlation to RTPN.

2.6 Conclusion

For OARs such as the oesophagus, there is no consensus in the literature as to the most appropriate metrics to use to predict toxicity. With significant heterogeneity in the data, the mechanisms of toxicity may not be fully understood. The results of this paper show an encouragingly strong correlation with multiple dose metrics, which have the potential to be further refined using techniques such as machine learning[8] to allow for a more accurate prediction of oesophagitis.

The results for RTPN are less promising from this dataset. There were no statistically significant predictors of RTPN. Further analysis using multi-variable analysis and machine learning may be able to improve prediction for RTPN by virtue of using more predictors in the model and in the case of machine learning performing more complex data modelling[8]. The use of EUD also showed superiority over MLD. Further analysis using Lyman Kutcher Burman[36] NTCP modelling would be worth exploring to improve toxicity prediction for RTPN.

3 Paper 2: Predicting radiotherapy toxicity for NSCLC patients using Machine Learning Techniques

Paper 3: Rushil Patel[1], Karen Venables[1], Adam Aitkenhead[2], Laura Farrelly[3], Nicholas Counsell[3], David Landau[4]

1 Mount Vernon Cancer Centre, Northwood, UK

2 The Christie NHS Foundation Trust, Manchester, UK.

3 Cancer Research UK & UCL Cancer Trials Centre, London, UK

4 Guys & St Thomas NHS Trust, London, UK

3.1 Abstract

Background and Purpose: The accurate prediction of radiation induced toxicity enables clinicians to make more informed treatment decisions with regard to the potential risks and benefits of radiotherapy treatment. In this study we explore the use of machine learning using clinical and dosimetric data features to predict toxicity for NSCLC patients from the IDEAL-CRT trial.

Methods: Dosimetric and clinical data from 116 NSCLC patients from the IDEAL-CRT trial underwent supervised machine learning using the classification learner application in MATLAB 2022a. Data was modelled against outcomes of grade 2 or higher radiation induced pneumonitis and oesophagitis using the Decision Trees, Logistic Regression, Support Vector Machines, Ensemble, Neural Networks and Naïve Bayes classifiers. An MRMR feature selection method was used to select modelling features. 5-fold cross validation method was employed to determine the predictive accuracy, sensitivity, and specificity of each model. Receiver operator characteristic (ROC) analysis was performed to determine the area under the curve (AUC) to assess overall model performance.

Results: The predictive accuracies observed for all classifiers and MRMR selected feature combinations ranged from 69-74% for oesophagitis and 65-78% for pneumonitis. The Naïve Bayes model had the highest AUC of 0.79 for oesophagitis, with an overall accuracy of 78.4%, sensitivity of 93.6%, and specificity of 31.6%. The Naïve Bayes classifier utilised the oesophageal V50Gy and patient sex as inputs. For pneumonitis the Decision Tree classifier produced the model with the highest AUC of 0.53, the Ensemble and Neural Network models

provided a slight improvement in sensitivity over other models but overall the predictive performance of all models was poor.

Conclusion: The results of this study were comparable to those found in published literature. For radiation induced oesophagitis there is a clear correlation between volumetric dose constraints such as the V50 Gy that provided the machine learning classifiers with good predictive accuracy. For pneumonitis the models only achieved poor predictive accuracy. These models have the potential to be clinically useful and further analysis is required to determine if the use of additional input data can improve the performance. Larger datasets would also be useful as they would facilitate the use of more features for model training.

3.2 Introduction

Half of patients who undergo active treatment for cancer will receive radiotherapy as part of their treatment program. A particular challenge of radiotherapy is the close proximity of organs at risk (OARs) to the target regions, which can lead to radiotherapy induced toxicity. It is common practice when planning a radiotherapy treatment to adhere to predefined dose constraints for OARs to ensure an acceptable level of toxicity when treating patients. This allows the planner to balance the potential benefits of treatment in terms of tumour control against the possibility of damage to healthy organs and tissue. For the treatment of NSCLC two of the most common side effects are radiation induced oesophagitis and pneumonitis.

The treatment outcomes from radiotherapy are determined by complex interactions between treatment, anatomical, and patient related variables. These outcomes are traditionally modelled using information about the dose distribution and fractionation, but it is recognised that the response to radiation is multifactorial and can include clinical prognostics factors as well. The modelling of radiotherapy outcomes is conducted by two approaches: analytical modelling which employs biophysical understanding of irradiation effects such as the linear quadratic model and data driven models that rely on robust parameters gained from matched clinical and dosimetric data[93]. For a data driven model the observed treatment outcome can be considered as the result of mathematical mapping of several dosimetric, clinical or biological inputs. The treatment outcomes are provided by an experienced healthcare professional using a standardised scoring criterion and this data driven approach is a commonly used method in outcome modelling[94].

Data driven approaches increasingly utilise machine learning (ML). The theoretical framework for ML has been in existence since the 1950s[31], but it was not until recently that advances in technology have made it feasible for use in clinical settings[8]. ML is a subset of artificial intelligence (AI) and refers to algorithms that can learn to perform a specific task without implementation. ML can be supervised or unsupervised. For supervised learning a labelled dataset is used, where each data feature has been labelled with the outcome. The aim of supervised ML is to predict the right answer from a selection of features. In unsupervised learning, an unlabelled dataset is used in which the "correct" answer is unknown and the aim of this type of learning is to be able to structure the data.

Numerous studies[44,45,53,95–98] have used supervised ML to build models for toxicity prediction from radiotherapy. Supervised ML is an appropriate technique for this application because the endpoints (e.g. grade of toxicity) are known. It is a branch of artificial intelligence in which algorithms are used to learn from prior experience to process complex data. Supervised ML learning algorithms are used to train data with known outcomes using features to detect patterns and correlation through the learning process.

Data from clinical trials is ideal for supervised ML methods, as they provide high quality data with good patient follow up reported using a standardised methodology. This type of data also has high levels of curation and quality assurance which can reduce confounding effects and bias when used for analysis[99,100]. Chemo-radiotherapy is currently the gold standard treatment for non small cell lung cancer (NSCLC) in patients that are surgically inoperable[71] and data from the IDEAL-CRT[10] trial is well suited for toxicity analysis in NSCLC patients using ML methods.

There are many ML classifiers that would be suitable for analysing the IDEAL-CRT dataset, a brief description of relevant classifiers is given here. Many of the classifiers have sub types that can be separately trained and may also have an optimisable model which have multiple hyperparameters that can be used to further optimise the training of the model with the potential to improve the accuracy and generalisability of the subsequent results. Where multiple classifiers may be appropriate for training and it is not prohibitively computationally expensive, it is recommended to train multiple classifiers and assess predictive performance and robustness to determine the most appropriate one for a given dataset.

Decision Tree classifiers let you predict responses to data by following the decision from the root down to a leaf node. The tree consists of branching decision where the value of the predictor is compared with the training weight. The number of branches and the value of the weights are determined through the training process, additional modifications known as pruning can be used to simplify the model. The Logistic Regression classifier can predict the probability of a binary response. Due to its simplicity, it is commonly used. It uses a sigmoid function to return the probability of an outcome. It is widely used when the classification problem is binary. The sigmoid function generates a probability output by comparing the probability with a predefined threshold that is optimised during training.

SVM classifies data by finding the linear decision boundary (hyperplane) that separates all the data points of one class from those of another. The best boundaries for SVM models are the ones with the largest difference between two classes. If the data is not linearly separable, a loss function is used to penalize points on the wrong side. SVMs sometimes may also use a kernel transform to transform nonlinearly separable data into higher dimension where a linear boundary can be found. Ensemble classifiers combine several weaker decision trees into a stronger ensemble. For example, a bagged decision tree consists of trees that are trained independently on data that is bootstrapped from the input data. Boosting involves creating a strong learner by iteratively adding weak learners and adjusting the weight of each weak learner to focus on misclassified examples.

The Naïve Bayes classifier assumes the presence of a particular feature in a class is unrelated to the presence of any other feature. It classifies new data based on the highest probability of its belonging to a particular class. Neural Networks take inspiration from the learning process occurring in human brains. A neural network consists of layers, which convert an input into an output. Each unit takes an input, applies a function to it and passes it onto the next layer. Generally, these networks are designed to feed-forward, as a layer feeds its output to the next layer, there is not any feedback to the previous layer. Weightings are applied to the signals passing from one unit to another, these weightings are tuned during training to adapt to particular problem.

The accurate predication of clinical outcomes can enable clinicians to make more informed treatment decisions with regards to the risks and benefits of particular treatment choices. In

this paper we explore the use of clinical and dosimetric factors in combination with supervised ML to predict oesophagitis and pneumonitis in NSCLC patients.

3.3 Methods

3.3.1 Data pre-processing

As part of the IDEAL-CRT trial[10], clinical factors and radiotherapy treatment data were routinely collected for all trial patients. In total 120 patients were recruited to the IDEAL-CRT trial; 116 complete records were available for analysis. NCI CTCAE v4[76] Grade 2 to 5 radiotherapy induced pneumonitis (RTPN) was seen in 36 (30.5%) of patients that received the trial treatment, three of these events were grade 3 (3.7%), no patients had Grade 4 or 5 toxicity. Grade 2 to 5 oesophagitis rate was 82.9% overall with 5 grade 3 toxicities (6.1%) and no grade 4 or 5 toxicity. With regards to clinical factors, sex, age, and lung function measurements such as forced expiratory volume in one second (FEV), forced vital capacity (FVC) and diffusing capacity of lungs for carbon monoxide (DLCO) were collected. FEV measures how much air a person can exhale during a forced breath in one second, FVC is the total of air exhaled during a single breath and DLCO is a measure of the ability of the lungs to transfer gas from inspired air into the bloodstream.

IDEAL-CRT was an isotoxic trial with individual dose escalation for each patient. Doses to OARs were converted to the equivalent dose in 2 Gy fractions (EQD2) to allow comparison to conventional radiotherapy results in the literature, this conversion took into account overall treatment time. The dosimetric data for all patients was imported into the Eclipse Treatment planning system v15.6 (Varian Medical Systems) and DVH data was exported for each patient into an in-house MATLAB programme (Appendix 3) which converted the doses into EQD2T for analysis using the linear quadratic equation[11] with an α/β ratio of 4 for the lung and 10 for the oesophagus. The programme then calculated dosimetric parameters based on those from QUANTEC[3,60,77] and relevant clinical trials[10,84–86]. These included the V5Gy, V20Gy, Mean dose (MD), and Equivalent uniform dose (EUD) for Lung and V35Gy, V50Gy, D1cc, and Mean dose for the Oesophagus.

Previous statistical analysis[101] for this data set using binary logistic regression analysis showed weak statistical correlation between predictive dosimetric factors (V5Gy, V20Gy, MD and EUD) and RTPN using uni-variable analysis. That analysis showed a strong correlation ($P<0.01$)

between several predictive factors and oesophagitis. These consisted of all tested dosimetric factors. However, while these binary logistic regression models showed good sensitivity (85-100%), the specificity was poor (0-45%).

3.3.2 Supervised machine learning

Trial participants were mapped as either having grade 2 or greater toxicity or not for radiation induced oesophagitis and pneumonitis separately. This mapping along with dosimetric and clinical data features were imported into the Classification Learner Application in MATLAB v2019b to apply supervised machine learning techniques. This involved building and evaluating the ability of ML classifiers to correctly predict response based on the selected data features. For oesophagitis, the predictive features evaluated were, sex, age, MD, V35, V50, and D1cc. For pneumonitis, the factors available for modelling were, sex, age, FVC, FEV, MD, V20, and EUD. All factors were continuous data which are defined as numerical values that can be an infinite number of values between any two values, except for sex which was categorical as it contains a finite number of categories.

When training ML classifiers, the minimum number of samples per feature required is affected by the complexity of the data and model. The modelling of toxicity is complex and given the sample size of 116 patients, the number of predictors used for modelling should be minimized to aid the robustness of the model and prevent overfitting. Sample size calculations have shown that this should be a maximum of 2 features for this dataset to minimise shrinkage. Modern classification models are highly adaptive and capable of modelling complex relationships, but they can also easily overemphasise patterns that are not reproducible. A methodological approach is required to evaluate ML classifiers ensure they do not overfit the data, as this would lead to ML classifiers that are too specific to the original dataset and are not generalisable. Given that there is a fixed amount of data available, the samples will be split into sets used for modelling and sets used for evaluation. Ideally a model should be evaluated using samples that have not been used to create the model. For smaller datasets, the k-fold cross validation technique can be applied to test and validate the data. The data is split into k sets of roughly equal size. The model is trained using k-1 sets and is then evaluated using the set that was excluded from the training process. This procedure is performed k times, each time omitting a different sample from the training process and using it for evaluation. The k resampled estimates of performance are then summarised and

used to understand the relationship between tuning parameters and model utility. k =5 was used for this study to ensure a sufficient number of samples in each of the training data sets and it has been shown empirically to yield test error rate estimates that do not suffer from high bias or variance[102]. Within MATLAB, the sets are generated using a random seed when the data is imported into the classification learner, therefor the k sets are fixed for all modelling and can only be regenerated by starting a new modelling session.

Multiple classifiers are available for ML in the Classification Learner Application, with the suitability of the classifier determined by the nature of the dataset. The classifiers evaluated were: Decision Trees, Logistic Regression, Support Vector Machines (SVM), Ensemble, Neural Networks and Naïve Bayes. Nearest Neighbour and Discriminant Analysis methods were not evaluated since they do not support analysis of mixed continuous and categorical data types. Each of these different machine learning models have their own characteristics which can make them more or less appropriate for a give dataset. The characteristics of suitable ML models are given in Table 11.

Classifier	Interpretability	Types
Decision Trees	Easy	Coarse, Medium and Fine Trees
Logistic Regression	Easy	-
Naive Bayes	Easy	Gaussian, Kernel
SVM	Easy for Linear SVM, hard for others	Linear, Quadratic, Cubic, Gaussian
Ensembles	Hard	Boosted/Bagged/RUSboosted Trees, Subspace Discriminant, KNN
Neural Networks	Hard	Narrow, Medium, Wide, Bilayered, Trilayered

Table 11 A table of suitable ML models available in the classification learner application with the difficulty of interpretation and model sub-types available for training.

A large number of datasets and features can pose a problem when performing machine learning, both in terms of the efficiency of algorithms in dealing with the data and the performance of the resulting models in terms of accuracy and generalisability. To confront this problem, feature reductions techniques have been developed to reduce the number of features and therefor improve the performance of the learning process. To this end Minimum Redundancy Maximum Relevance (MRMR) algorithm was developed by Peng & Ding[103], it finds an optimal set of features that is mutually and maximally dissimilar and can represent the response variable effectively. The algorithm minimizes the redundancy of a feature set and maximises the relevance of a feature set to the response variable. The algorithm is able to quantify the redundancy and relevance using the mutual information of variables as a proxy for computing relevance and redundancy.

The goal of MRMR is to find an optimal set of features that maximises the relevance of the set with respect to the response variable. It is a filter-based feature selection approach. MRMR was developed for feature selection of microarray data. It tends to select a subset of features having the most correlation with a class (relevance) and the least correlation between themselves (redundancy)[104]. In this algorithm the features are ranked according to the minimal redundancy maximal relevance criteria. The relevance can be calculated using the F-statistic (for continuous data) or mutual information (for categorical data) and redundancy can be calculated by using Pearson correlation coefficient (for continuous data) or mutual information (categorical data). MRMR was used for feature reduction to enable efficient and optimal training for the algorithms. Prior to MRMR ranking a cut off was applied where groups of variables were highly correlated. For oesophagitis a cut off of $P \leq 0.01$ to the bootstrapped results of LR was applied for dose metrics which removed the D1cc metric. For pneumonitis a cut off of $P \leq 0.1$ to the bootstrapped results of LR was used for lung function and dose metrics, which removed the V5 Gy and DCLO metrics.

Each of the six classifiers were automatically trained to determine the most appropriate one(s) for further training using hyperparameter optimisation. The training utilised the features generated through MRMR feature reduction and evaluated the prediction accuracy, sensitivity, specificity, and the largest area under the curve (AUC) value from Receiver Operator Characteristic (ROC) analysis.

Models were trained using the clinical and dosimetric predictors mapped to a positive/negative score for Grade 2 or higher toxicity for either oesophagitis or pneumonitis. The predictive accuracy, sensitivity, specificity in addition to AUC values were used to evaluate the clinical potential of models. A desirable ML classification should balance both sensitivity and specificity rather than relying on overall accuracy alone.

3.4 Results

The results presented here are a representation of all ML classification models using the predictor combinations refined through MRMR. The predictive accuracies observed for all tested classifiers with MRMR selected feature combinations ranged from 69-74% for oesophagitis and 65-78% for pneumonitis. Overall predictive accuracy, sensitivity and specificity of the models are reported below

3.4.1 Oesophagitis

Table 12 shows the ranking of features using the MRMR algorithm. Table 13 gives an overview of the performance of the six classifiers using the MRMR selected features for oesophagitis. The top two features were selected on accordance with the results of sample size calculations. Figure 10 is a bar chart representing the predictive accuracy, sensitivity, and specificity of the trained models. Figure 11 shows the ROC curves for a simple and complex model, Table 14 shows the output of the same 2 models that have been retrained using different randomisation seeds for the K-fold validation.

Rank	Feature	MRMR
1	**V50**	0.0556
2	**Sex**	0.0202
3	V35	0
4	MD	0
5	Age	0

Table 12 Results of feature ranking using the MRMR algorithm for Oesophagitis, the top two features were used for training the ML classifiers

Model Features	Machine Learning Model Results					
	Decision Tree	Logistic Regression	Naïve Bayes	SVM	Ensemble	Neural Network

Total Accuracy	71.6	69	73.3	74.1	69.8	74.1
Sensitivity	92.3	84.6	93.6	94.9	82.1	89.7
Specificity	28.9	36.8	31.6	31.6	44.7	42.1
AUC	0.68	0.78	0.79	0.66	0.7	0.77
Model Sub Type	Coarse	N/A	Kernel	Fine Gaussian	Boosted Trees	Narrow

Table 13 Table showing the resulting 5-fold predictive accuracy, sensitivities, specificities, and Area under the Curve from ROC analysis for the trained ML classifiers against tested features for oesophagitis. Corresponding confusion matrices are plotted in Appendix 5

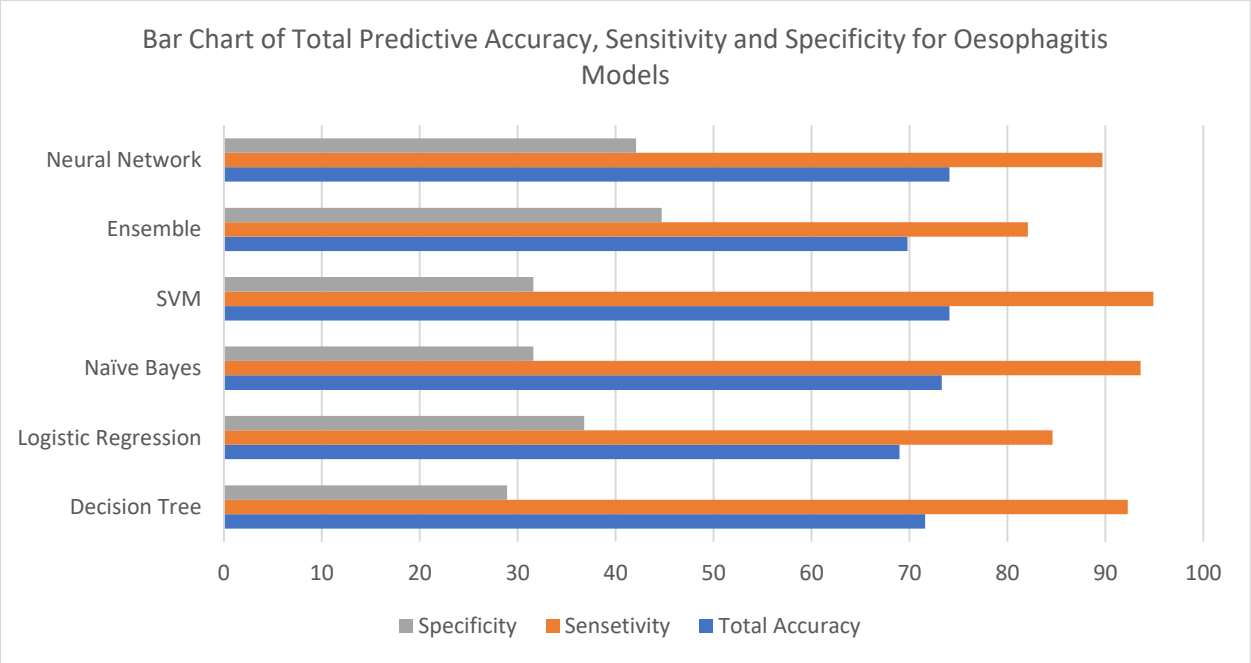


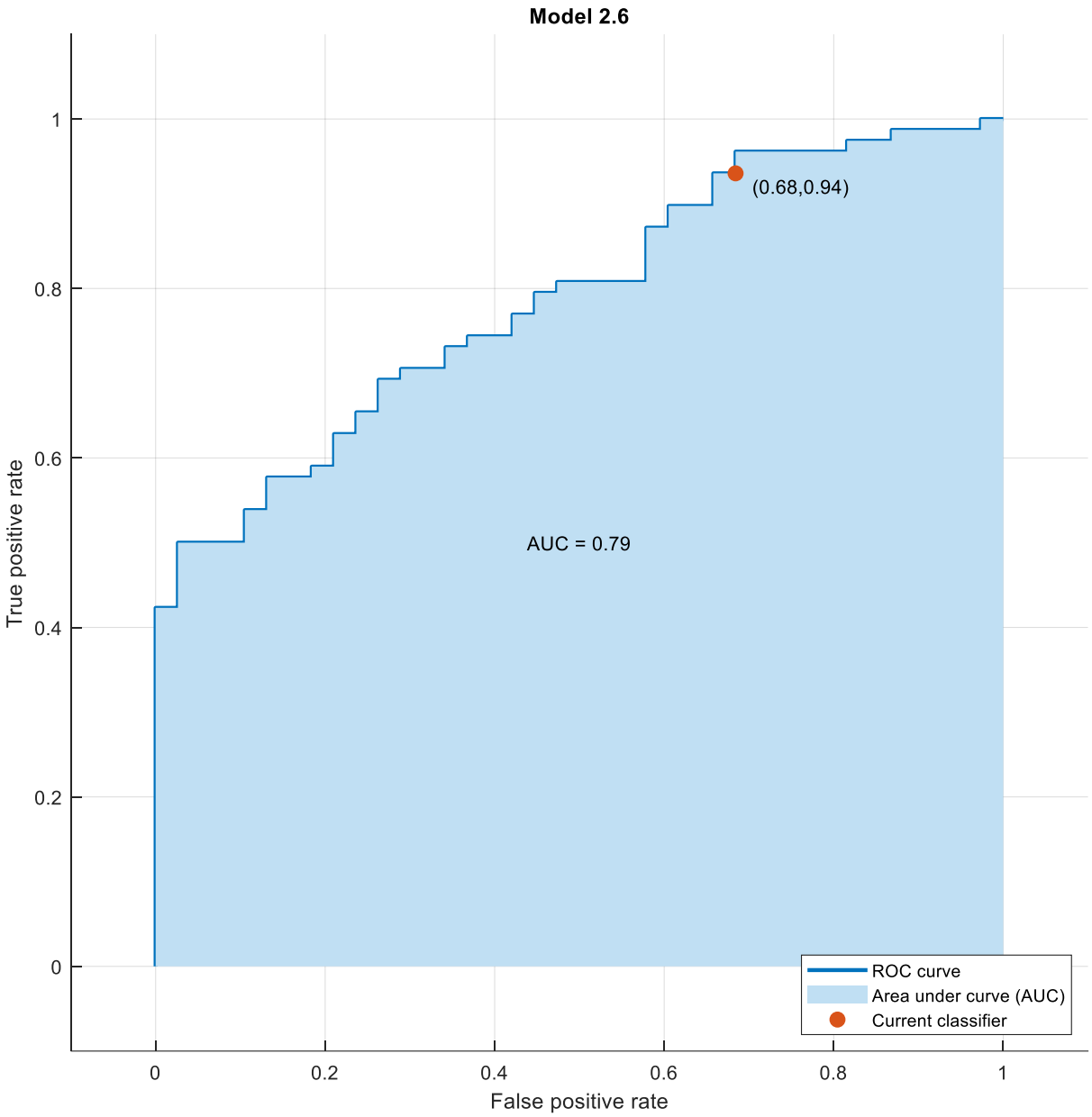
Figure 10 Bar chart demonstrating the median predictive accuracies, sensitivities and specificities obtained for the predictors and ML classifier stated in Table 7 for oesophagitis

Model	Features	0	1	2	3	4	5	Mean	S.D.
Naïve Bayes	**Total Accuracy**	73.3	74.1	73.3	72.4	74.1	74.1	73.6	0.63
	Sensitivity	93.6	94.9	93.6	92.3	94.9	94.9	94.0	0.97
	Specificity	31.6	31.6	31.6	31.6	31.6	31.6	31.6	0.00
	AUC	0.79	0.75	0.74	0.71	0.74	0.75	0.7	0.02

Neural Networks	**Total Accuracy**	74.1	71.6	67.2	69.8	68.1	98.1	74.8	10.66
	Sensitivity	89.7	84.6	78.2	87.2	83.3	80.8	84.0	3.82
	Specificity	42.1	44.7	44.7	34.2	36.8	42.1	40.8	3.94
	AUC	0.77	0.71	0.68	0.67	0.69	0.7	0.7	0.03

Table 14 Results of two models repeated with different randomisation seeds for K-fold validation

For the oesophagitis models in Table 13, the SVM and Neural Network models had the highest overall predictive accuracy (74.1%), the Neural Network model had the highest Specificity (44.7%). While the Neural network model had a sensitivity of 89.7%, the SVM (94.9%), Decision Tree (92.3%) and Naïve Bayes (93.6%) had superior performance. Receiver operator characteristics (ROC) curves for Naïve Bayes and Neural Network models are shown in Figure 11. The Area Under the Curve (AUC) provides an aggregate measure of the performance across all possible classification thresholds, the highest AUC was achieved by the Naïve Bayes Model (0.79), although the Logistic Regression (0.78) and Neural Network (0.77) models performed similarly. The Naïve Bayes and Neural Network models were further tested using a different randomisation seed for the K-fold cross validation as shown in Table 14, the standard deviation of the AUC for both models were similar at 0.2 and 0.3 respectively but the overall predictive accuracy varied more for the Neural Networks (10.66%) than the Naïve Bayes model (0.63%).


Model 2.6
True positive rate
False positive rate
(0.68,0.94)
AUC = 0.79
ROC curve
Area under curve (AUC)
Current classifier
0
0.2
0.4
0.6
0.8
1

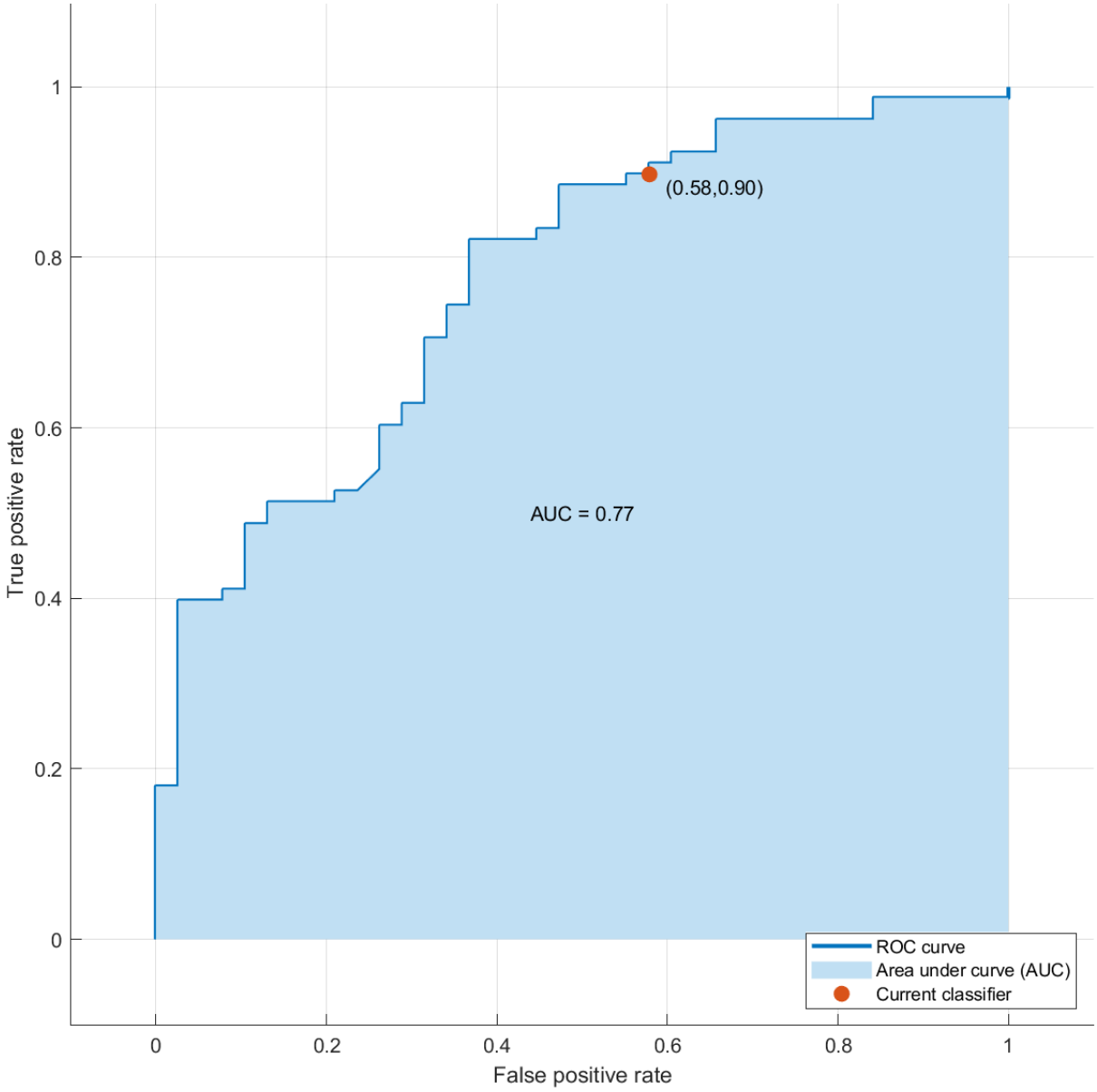


Figure 11 Receiver Operator Characteristic (ROC) curves for Oesophagitis Naive Bayes (Top) and Neural Networks (Bottom). The Area Under Curve (AUC) represents the model's overall ability to correctly classify structures into each category. The orange dot gives the optimal point on the curve that gives the highest overall predictive accuracy.

3.4.2 Pneumonitis

Table 15 shows the results of feature selection using the MRMR algorithm. Table 16 gives an overview of the performance of the six classifiers using the MRMR selected features for pneumonitis. Figure 12 is a bar chart representing the predictive accuracy, sensitivity, and specificity of the trained models. Figure 13 shows the ROC curves for a model and Table 16 shows the same model that has been retrained using different randomisation seeds for the K-fold validation.

Rank	Feature	MRMR Value
1	**Sex**	0.0056
2	**LungEUD**	0
3	V20	0
4	MD	0
5	FVC	0
6	FEV	0

Table 15 Results of feature ranking using the MRMR algorithm for RTPN, the top two features were used for training the ML classifiers

Model Features	Machine Learning Model Results					
	Decision Tree	Logistic Regression	Naïve Bayes	SVM	Ensemble	Neural Network
Total Accuracy	76.7	77.6	78.4	77.6	64.7	69.8
Sensitivity	3.8	0	3.8	0	23.1	11.5
Specificity	97.8	100	100	100	76.7	86.7
AUC	0.53	0.5	0.5	0.4	0.49	0.43
Model Sub Type	Coarse	N/A	Gaussian	Linear	Boosted Trees	Narrow

Table 16 A table showing the resulting 5-fold predictive accuracy, sensitivities, specificities, and Area under the Curve from ROC analysis for the trained ML classifiers against tested features for pneumonitis. Corresponding confusion matrices are plotted in Appendix 5

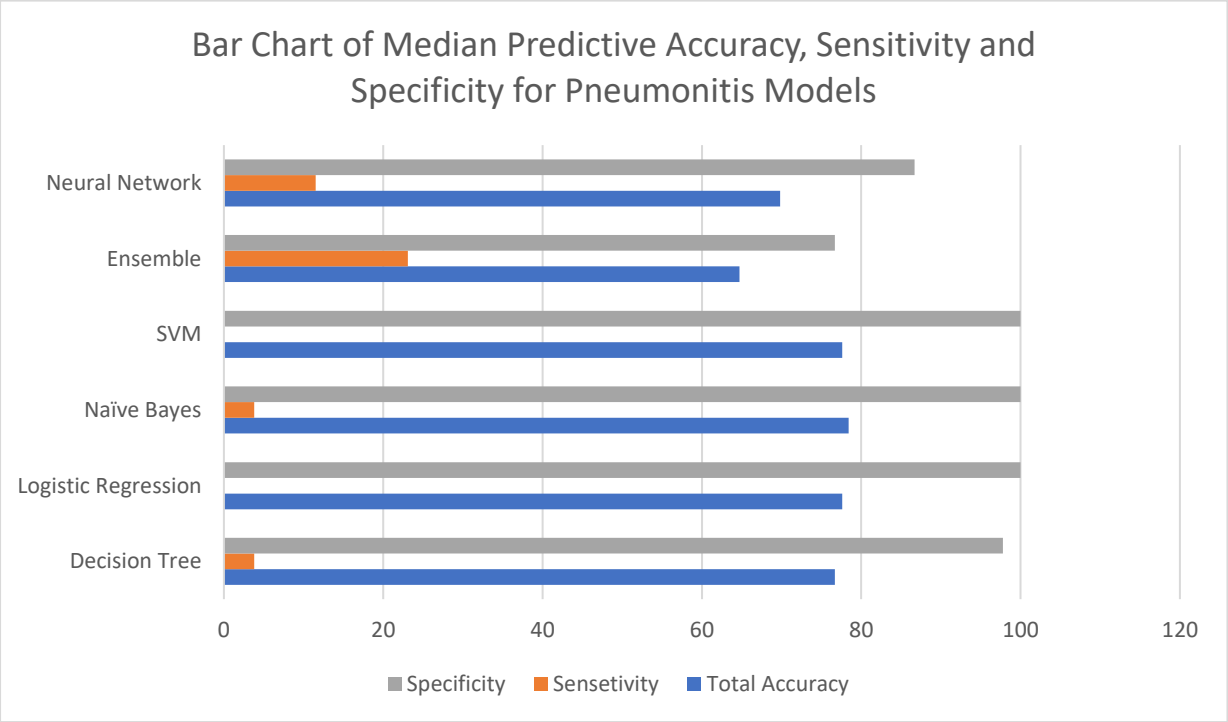


Figure 12 Bar chart demonstrating the median predictive accuracies, sensitivities and specificities obtained for the features and ML classifier stated in Table 7 for pneumonitis

Model	Features	0	1	2	3	4	5	Mean	S.D.
Decision Trees	Total Accuracy	76.7	73.3	75	74.1	75.9	76.7	75.3	1.28
	Sensitivity	3.8	7.7	15.4	23.1	23.1	34.6	18.0	10.34
	Specificity	97.8	92.2	92.2	88.9	91.1	87.8	91.7	3.19
	AUC	0.53	0.51	0.66	0.54	0.61	0.56	0.6	0.05

Table 17 Results of a model repeated with different randomisation seeds for K-fold validation

For pneumonitis, Naïve Bayes model had the highest overall predictive accuracy (78.4%), several models had a specificity of 100%. The ensemble model has the highest sensitivity (23.1%) and the lowest specificity (76.7%). Receiver operator characteristics (ROC) analysis was performed, the highest AUC for was 0.53 for the Decision Tree model.

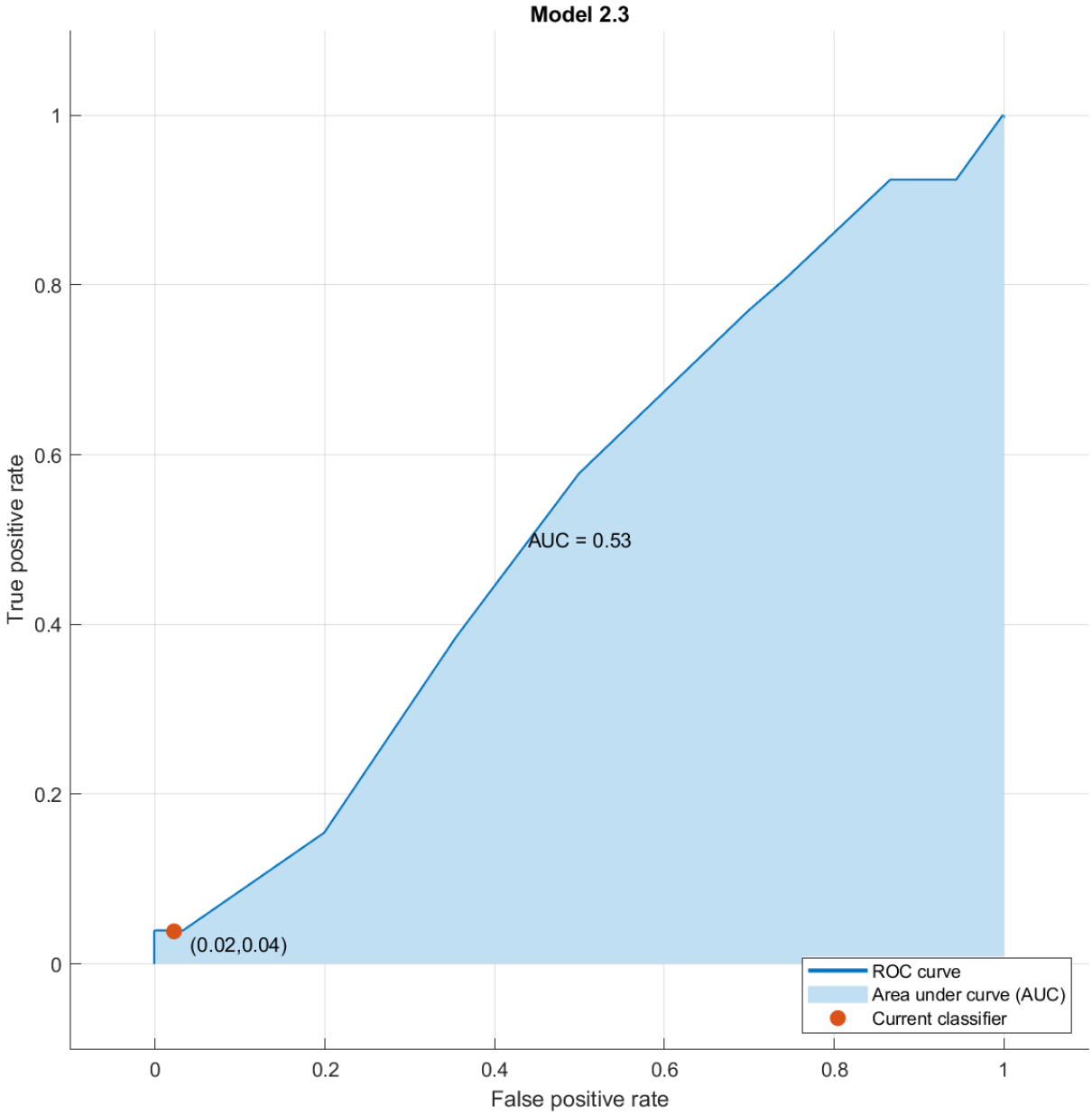


Figure 13 Receiver Operator Characteristic (ROC) curves for Pneumonitis Ensemble model 13. The Area Under Curve (AUC) represents the model's overall ability to correctly classify structures into each category. The current classifier gives the optimal point on the curve that produces the highest overall predictive accuracy.

3.5 Discussion

Developing predictive models describing the relationship between clinical parameters and outcomes in radiotherapy is complex. In order to perform effective modelling, all the important features must be contained within the dataset, it must be of sufficient size, and we must understand that many variables are highly correlated, particularly dosimetric ones. We must consider how different variables interact, which can be complex and non linear in determining the endpoint(s). Successful ML models for oesophagitis and pneumonitis have the potential to be clinically useful for patients with NSCLC. The accurate prediction of toxicity

would allow us to determine which patients would benefit most from an isotoxic treatment regime that has the potential to improve outcomes while maintaining acceptable toxicity levels. Conversely it would also enable us to determine which patients are likely to suffer from toxicity allowing us to attempt to mitigate potential treatment disruptions, which can negatively affect outcomes[105].

3.5.1 Oesophagitis

Training of the classifiers returned models with a high degree of sensitivity, with SVM model predicting 94.9% of all true positives, although this was at a cost of reduced specificity of only 31.6.% of true negatives. There was a general trend for all models to have low specificity. As 67% of all patients within this data set had grade 2 or higher toxicity, this may have led to the ML algorithms favouring sensitivity over specificity as this gave the highest overall predicative accuracy. For modelling oesophagitis, the SVM and Neural Network classifiers provided models with the highest predictive accuracy. Notably the Neural Network model was able to achieve this with the second highest specificity. The Naïve Bayes model had the highest AUC of 0.79, although the Neural Network model AUC of 0.77 was similar. When these models were repeated using a different randomisation seed for the K-fold cross validation, the performance of the Naïve Bayes model was consistent whereas the performance of the Neural Network model had substantially more variance. This may be due to the increased complexity of the Neural Network model in comparison to the Naïve Bayes model in combination with the relatively small size of the dataset. The small size of the dataset also limited the number of features used to two, including more features into the modelling dataset may be able to improve performance. Two other studies from El Naqa et al[54,55] have explored oesophagitis in lung cancer patients using a Logistic Regression classifier and found that performance could be improved by mixing clinical and dosimetric factors as input parameters. In the latter study they found that SVM provided superior performance to logistic regression and neural networks which is not consistent with the results presented here, they did not test the naïve bayes or ensemble classifier.

The Naïve Bayes model had the highest AUC with an overall accuracy of 73.3%, sensitivity of 93.6%, specificity of 31.6% and AUC of 0.79. This is comparable to results of Huang et al[44] whose most successful logistic regression model achieved an AUC of 0.83 using MD and concurrent chemotherapy. In a second paper they tested this model on independent data and

achieved an AUC of 0.78. While concurrent chemotherapy has been shown to have a significant impact on toxicity and would therefore be a useful feature for ML, all patients within the IDEAL-CRT trial underwent concurrent chemotherapy so this variable could not be used for modelling. Niedzielski et al[98] explored using CT imaging biomarkers to quantify the radiosensitivity of individual patients with the goal of predicting oesophagitis. They used ML techniques to produce models with an AUC of 0.75. This information could also be used to improve the accuracy of the model as Niedzielski found that models using radiosensitivity predictors outperformed those that did not for grade 3 toxicity.

3.5.2 Pneumonitis

For the modelling of pneumonitis 6 different variables were reduced to 2 using the MRMR algorithm. Classifier training produced toxicity models with negligible accuracy with each model favouring high specificity due the large percentage of patients within this cohort that did not suffer toxicities. Three of the tested models achieved 100% specificity with a fourth modelling achieving 97.8%, each of these four models achieved a sensitivity of less than 3.8%. the AUC values achieved by all models was between 0.4-0.53 which suggests these models are no better than random guesses. The decision trees classifier provided the highest AUC value of 0.53 with an overall accuracy of 76.7%, specificity of 3.8% and sensitivity of 97.8%. This model was tested a further 5 times with a different randomisation seed for the K-fold cross validation, the standard deviation of the AUC and overall accuracy was 0.05 and 10.34% respectively. Suggesting that even for a simple classifier the reproducibility of the model is poor and therefor it is unlikely to work well with new datasets. Here we found that the MRMR algorithm favoured EUD as the most appropriate dosimetric feature but we must take into account that patient gender had the highest MRMR score. This suggests that the features available for this dataset have weak correlation with RTPN and that additional features need to be analysed in order to improve model performance.

Das et al[52] used a parametric dose based Lyman NTCP model in conjunction with weighted non parametric decision trees to train a model with an AUC of 0.72, which was an improvement over the Lyman NTCP model alone which achieved an AUC of 0.62. A further study by Das et al[51] combing four classification models achieved an AUC of 0.79. Lee et al[96] used a Bayesian network approach using dose, clinical and blood biomarkers to achieve an AUC of 0.85, although this was using a relatively small patient sample of 54. Valdes et al[53]

used a larger dataset of 201 patients and found that the ensemble model had the best performance. They found that the dose to 15cc of the heart, dose to 4cc of the trachea or bronchus and ethnicity were the most important features in the model prediction, features not used for this project whilst ethnicity data is not available, doses to other OARS should be considered for future research. Similarly Luna et al[95] found that the oesophagus maximum dose was another potential new predictor.

3.6 Conclusion

There are numerous studies[51–55,96,98,106] that have reported associations between various parameters and toxicity to OARs and normal tissues. This information is used to attempt to improve patient care by providing clinicians and treatment planners with the tools to find the optimal balance between tumour control and toxicity. This study found that accurately predicting radiation induced oesophagitis was achievable using data from the IDEAL-CRT trial, but that predicting RTPN poses more of a challenge.

For radiation induced oesophagitis there is a clear correlation between various volumetric dose constraints, here the V50Gy metric provided the ML classifiers with high predictive accuracy. These models had good sensitivity (82-95%) but only moderate specificity (29-45%). The results compare well with the literature and indicate ML has the potential to create clinically useful models for predicting oesophagitis. For pneumonitis the models only achieved moderate predictive accuracy, there was no statistically significant link to any of the tested features. The results suggest that ML is unlikely to produce a clinically useful model for RTPN using the IDEAL-CRT dataset

ML techniques have the ability to utilise the complex data to predict radiotherapy induced toxicity and our results demonstrate what can be achieved using clinical and dosimetric features. There is scope to further improve these models by using doses to nearby OARs[53], imaging biomarkers[98] and genomic information[107]. Advances in technology make much of this data easier to obtain and clinical trials are increasingly collecting tissue samples for translational research for UK radiotherapy trials. With this additional information and the increasing ease of access to ML techniques, it is possible that highly accurate models can be produced for oesophagitis and pneumonitis.

4 Paper 3: Evaluating methods for predicting toxicity in NSCLC patients

4.1 Abstract

Background and Purpose: Accurate prediction of toxicity is essential to obtain a favourable trade-off between tumour control and toxicity (therapeutic ratio) in radiotherapy. Several trials[10,84,85] have attempted to utilise isotoxic lung radiotherapy, a technique that escalates and individualises RT doses to improve local control for these patients, however dose escalation is often constrained by the dose delivered to nearby healthy tissues, notably the oesophagus and lungs. Accurate prediction of toxicities could lead to an improvement in the therapeutic ratio. The study reported here aims to compare Lyman Kutcher Burman (LKB) NTCP modelling and multi-variable logistic regression (LR) analysis to previous machine learning (ML) and uni-variable LR approaches to assess methods of toxicity modelling.

Methods: Data from 116 NSCLC patients from the isotoxic IDEAL-CRT trial were analysed. LKB NTCP and multi-variable LR were compared with ML approaches for toxicity modelling. The overall predictive accuracy, sensitivity, specificity, and area under the curve (AUC) values from recover operating characteristics (ROC) analysis were compared for the three different techniques.

Results: A strong correlation was found using LKB NTCP analysis for oesophagitis using LR and Mann-Whitney U with both tests producing a P-value <0.01. Predictive accuracy of LKB NTCP values using LR was comparable with the results of uni-variable logistic regression. Statistical tests of LKB modelling for pneumonitis did not produce a statistically significant correlation. Multi-variable LR analysis offered a moderate improvement over uni-variable analysis for both oesophagitis and pneumonitis. ROC analysis of NTCP modelling for oesophagitis produced AUC values of 0.68, compared to 0.77 for multi-variable LR and 0.79 for ML. The corresponding AUC values for pneumonitis were 0.55 for LKB, 0.58 for multi-variable LR and 0.53 for ML.

Conclusion: Multi-variable LR analysis techniques can predict toxicity with similar accuracy to ML when there is good correlation between metrics and toxicity. When there is not, machine learning's ability to utilise more diverse data and customise parameters of learning classifiers enables superior toxicity models to be generated. Including imaging biomarkers and genetic

information with clinical and dosimetric information has the potential to improve the accuracy of toxicity models and has the potential to create clinically useful tools that can improve patient outcomes.

4.2 Introduction

Accurate prediction of toxicity is essential to obtain a favourable trade-off between tumour control and toxicity (therapeutic ratio) in radiotherapy (RT). Non small cell lung cancer (NSCLC) is often associated with poor outcomes and is the primary cause of death by cancer in Europe[108]. Several trials[10,84,85] have attempted to utilise isotoxic radiotherapy, a technique that escalates and individualises RT doses to improve local control for these patients, however dose escalation is often constrained by the dose delivered to nearby healthy tissues, notably the oesophagus and lungs. Accurate prediction of toxicities such as oesophagitis and pneumonitis for these patients may allow further dose optimisation, potentially leading to an improvement in the therapeutic ratio.

The occurrence of toxicity in radiotherapy patients is governed by complex relationships between multiple factors which include dosimetric, clinical and genetic parameters. When modelling toxicity, models can be classified as analytical or data driven. Analytical models are based on a simplified characterisation of the interaction between radiation and biological tissues explaining the underlying mechanisms with explicit algorithms. Common models such as the Lyman Kutcher Burman[29] (LKB) model use hand crafted rules with intricate exceptions that can often fail to predict the actual complications induced by RT. Data driven approaches are based on the assumption that the interaction between radiation and normal tissue is complex and cannot be properly represented deterministically. Commonly used models for predicting radiotherapy toxicity through statistical analysis include logistic regression (LR), normal tissue complication probabilities (NTCP) models such as LKB and artificial intelligence (AI) methods such as supervised machine learning (ML).

The LKB model is the most widely known NTCP model. It comprises of an empirical model of dose response as a function of irradiated volume, reducing the dose volume histogram to a single metric for each organ at risk tested (OAR)[94]. The dose distribution within an OAR is likely to be inhomogeneous in modern radiotherapy plans and, for these cases, this metric serves to translate the inhomogeneous dose distribution into the same equivalent dose

response as a homogeneous dose distribution to that OAR. The most commonly used metric is generalised equivalent uniform dose (EUD). This model has been used extensively in the Quantitative Analysis of Normal Tissue Effects in the Clinic (QUANTEC) publications[3,60,77] published in 2010 that presented evidence-based results made available from 3D treatment planning data. These summarised dosimetric and volumetric constraints for OARs after external beam radiotherapy for commonly reported side effects and these constraints are widely used in the radiotherapy community.

Alternative models consider the functional architecture of the OAR as to whether they can be considered as parallel or serial organs. Damage to a single functional unit in a serial organ can impair function of the whole organ, therefore the dose constraint for such structures should refer to the maximum dose delivered to any part of the organ. Parallel organs have a functional reserve, whereby a number of functional subunits may be damaged before there is any loss of function. In this case it is likely that mean dose (MD) is a more useful metric. In many cases the true organ architecture could be mixed, and the type of dose constraint to be used may be dependent on the specific toxicities being assessed. Traditionally the oesophagus is considered a serial organ and the lungs are considered parallel, QUANTEC recommends mean dose for oesophagitis and pneumonitis suggesting the tissue architecture for both these OARs is mixed.

When performing statistical analysis on the occurrence of toxicity in NSCLC patients, selection of the appropriate statistical test is critical. The data can be divided into two groups depending on whether they have or have not had a predefined grade of toxicity, and results in a categorical dependent variable. Independent variables can be continuous or categorical and can be chosen to test single or multiple variables at once. Statistical tests are routinely used to determine whether there is a statistically significant correlation between independent and dependent variables through the p-value, but this does not always translate into a predictive model. Binary Logistic Regression (LR), like all regression analyses, is predictive. LR is used to describe data and explain the relationship between a dependent binary variable such as toxicity and one or more nominal independent variables. Adding more independent variables to a LR model will increase the variance, too many and it can lead to overfitting, reducing the generalisability of the model. A cut-off is set for the LR value, standardly as 0.5, to predict whether a patient is classified as having had toxicity or not. This is then used to determine

overall predictive accuracy, sensitivity and specificity allowing comparison with ML derived models.

Supervised machine learning is an approach that has seen increasingly widespread use in radiotherapy. This is partly due to technological developments making ML more accessible in the clinic, as well as a move towards more digital healthcare allowing for greater access to patient data[8]. ML approaches are particularly well suited to modelling radiation induced toxicity as they are able to automatically create predictive models on new and unseen data and can easily handle large and diverse datasets. ML approaches will learn the parameters for a given toxicity model from the available data, using predefined artificial intelligence (AI) algorithms based on the type of ML classifier selected. There is risk of overfitting the model if the baseline dataset cannot be generalised to real world data. Previous publications using uni-variable statistical analysis[101] and ML techniques[109] have investigated predicting oesophagitis and pneumonitis using data from the IDEAL-CRT trial.

The study reported here aims to generate toxicity models using LKB NTCP modelling and multi-variable logistic regression for oesophagitis and pneumonitis in NSCLC patients using the IDEAL-CRT trial dataset. The efficacy of these models will be evaluated in the context of ML toxicity models on the same dataset by Patel et al[109] in order to assess the merits of these different approaches.

4.3 Methodology

Data was available for 116 patients from the IDEAL-CRT trial. Patients were treated with isotoxic radiotherapy with treatment doses ranging from 65-71Gy in 30 fractions, treated in either 5 or 6 weeks. All patients received concurrent chemotherapy. Full radiotherapy DICOM datasets were collected, and toxicity was reported using Common Terminology Criteria for Adverse Events (CTCAE), version 4.0. Clinical factors were collected through case report forms.

LKB and multi-variable logistic regression models were compared with ML toxicity models for overall predictive accuracy, sensitivity, specificity, and Area Under the Curve (AUC) from Receiver Operator Characteristics (ROC) analysis.

All doses were reported in terms of equivalent dose in 2 Gy daily treatment fractions (EQD2) corrected for overall treatment time to allow for straight-forward comparison of non-

standard treatment regimens (e.g. isotoxic) with more common treatment regimens such as the 60-64Gy in 30-32 fractions regime that is commonly used in the UK. Dose conversions were calculated using a custom MATLAB script (Appendix 3) that utilised the linear quadratic equation[110]. An α/β ratio of 4 was used for the lung, taken from the QUANTEC analysis[77,78]. For the oesophagus an α/β ratio of 10, commonly used in the literature[79,80] was used. The k term representing the dose recovered per day was set at 0.54Gy and 0.8Gy for the lung and oesophagus respectively, both values were from a reviews of clinical studies by Bentzen et al[81,82]. The T_{delay} was set at 28 days for both lung and oesophagus as per Fowler et al[83], with the T value of 33 days and 40 days for the 5 and 6 week treatment schedules respectively.

4.3.1 LKB Modelling

The LKB model is used for predicting NTCP for a radiotherapy treatment plan.

The LKB model from the original publication[29] is given in equation 4.1

$$NTCP = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-t^2/2}\, dt \tag{4.1}$$

Where

$$t = (D - TD50(V)/\sigma(V)) \tag{4.2}$$

$$u = \frac{D - TD50(V)}{m \times TD50(V)} \tag{4.3}$$

$$\mathrm{TD50(V)} = \frac{\mathrm{TD50(1)}}{V^n} \tag{4.4}$$

TD50(V) is the tolerance dose for a partial volume V. The parameter m is related to the standard deviation of TD50(1) and describes the steepness of the dose response curve, m is multiplied by TD50(V) approximates the standard deviation of volume V and n indicates the volume effect of the organ being assessed. A true serial organ, where maximum dose dominates outcome, is indicated by n=0; n=1 indicates a true parallel organ, where mean dose is related to outcome. D is the maximum dose of the DVH to ensure V<1. Histogram reduction can be performed to calculate the effective volume V according to the method described by Kutcher et al[26] shown in equation 4.5.

$$V = \sum_i \left(\frac{D_i}{D}\right)^{\frac{1}{n}} \Delta V_i \tag{4.5}$$

Where D_i is the dose defined for each bin in a differential DVH and D is the maximum dose to the organ. A Maximum Likelihood Estimation (MLE)[30] can be used to best fit values of the parameters TD50(1), m and n of the NTCP model for known binary outcomes y(i) of the available data by maximising the natural log of the likelihood (LLH) that the fitted model describes the data correctly.

There are numerous studies that have used the maximum likelihood method to find the LKB parameters that correlate with grade 2 or higher toxicity. IDEAL-CRT patients were treated with combined chemotherapy with dose escalation. In this study, the LKB parameters published by Belderbos et al[111] for oesophagitis, and those published by Lee et al[112] for pneumonitis have been applied, as these studies also used data from NSCLC patients treated with chemoradiotherapy and investigated toxicity of Grade 2 or higher. LKB values for each patient were generated using a custom MATLAB programme (Appendix 3). As NTCP is often used for ranking, the non parametric Mann Whitney U test was calculated using the SPSS Statistics v28 (IBM) software to confirm the correlation between toxicity and LKB value. The Mann-Whitney test ranks all of the data and then compare the sum of ranks for each group to determine whether the groups are the same or not.

Uni-variable LR was also performed to assess the efficacy of LKB modelling to predict toxicity. The sensitivity, specificity, overall predictive accuracy, and p-value from logistic regression were reported. The bootstrapping method was performed using 2000 samples generated by sampling with replacement to evaluate the robustness of the LR models generated. Receiver Operator Characteristic (ROC) curves were plotted and the Area Under the Curve (AUC) was reported to evaluate the overall performance of the model.

4.3.2 Logistic Regression

The factors available for logistic regression were, age, MD, V35, V50, D1cc, LKB, sex and age for oesophagitis and FVC, FEV, MD, V20, EUD, LKB, age and sex for pneumonitis. Uni-variable LR has been previously published[101]; a multi-variable analysis was performed within the work presented here for the most statistically significant factors with the aim of improving the accuracy of previously generated models. A minimum redundancy maximum relevance (MRMR)[103] feature selection method which uses mutual information as a proxy for computing relevance and redundancy among the selected variables was used to rank the relevance of features for modelling. The top two ranking features for oesophagitis and pneumonitis were

selected, as sample size calculation have determined that is the maximum permissible for the patients available for the IDEAL-CRT data set while assuming a shrinkage of ≤10% and a variance ≤15%. All analyses were performed in MATLAB 2022a and SPSS v28 (IBM).

As per previous uni-variable analysis, bootstrapping was performed to evaluate the robustness of LR models, and the sensitivity, specificity, overall predictive accuracy and p values of the variables were reported. ROC curves were also plotted to allow the AUC to be reported.

4.4 Results

4.4.1 LKB

LKB NTCP values generated through a custom MATLAB programme (Appendix 5) were imported into SPSS for analysis. The result of the Mann-Whitney U tests is given in Table 18.

4.4.1.1 Mann-Whitney U

	Lung_LKB	Oeso_LKB
Mann-Whitney U	1055	954
p-value	0..446	0.002

Table 18 A summary of results of Mann-Whitney U statistical test using LKB NTCP values for oesophagitis and pneumonitis. P-value reported is asymptomatic and 2 tailed.

4.4.1.2 Logistic Regression

Predictive accuracy, sensitivity and specificity were all calculated using a cut-off of 0.5 for the binary logistic regression analysis as reported in Table 19. LR and ROC analysis with AUC values are reported in Figure 14.

LKB-NTPC	Percentage Correctly Predicted			P Value	Sig (2-tailed)
	Specificity	Sensitivity	Overall Accuracy		
Oesophagitis	28.9%	94.9%	73.3%	0.004	0.004
Pneumonitis	100%	0%	77.6%	0.128	0.107

Table 19 Results of binary logistic regression analysis using LKB NTCP factors to predict oesophagitis and pneumonitis. The 2 tailed significance is generated thjrough LR of 2000 bootstrapped samples.

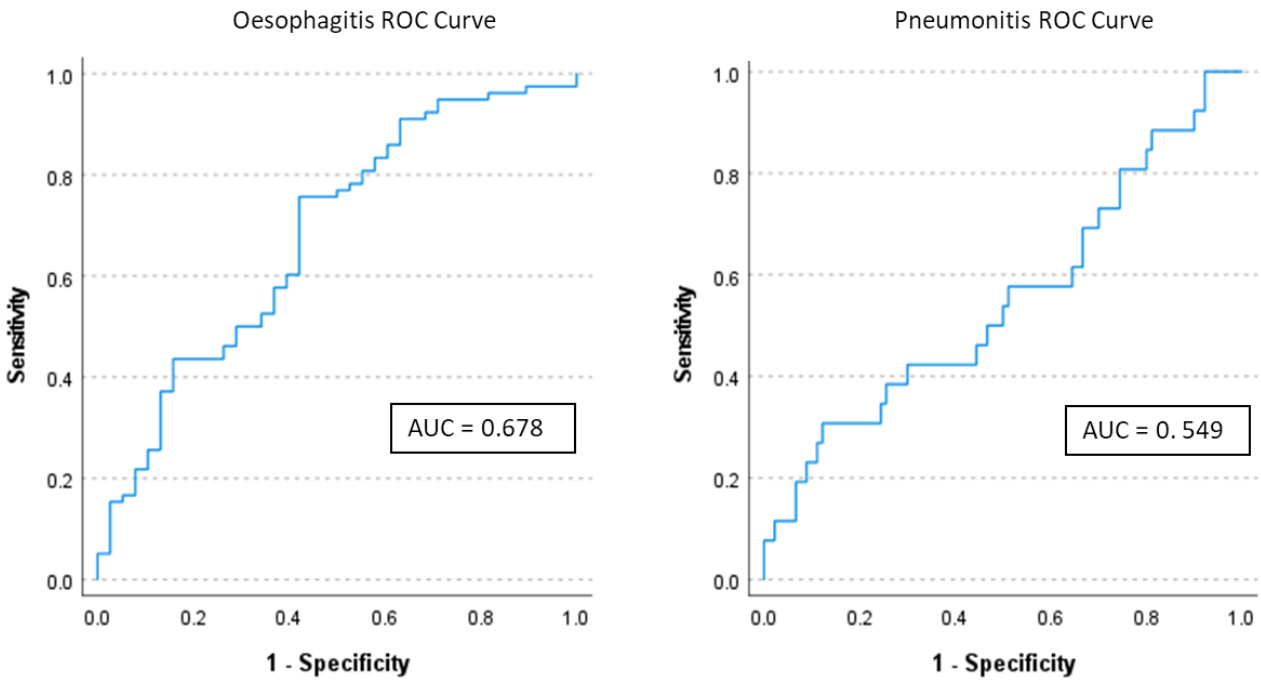


Figure 14 Receiver Operator Characteristic (ROC) curves for oesophagitis (left) and pneumonitis (right) from LBK NTCP analysis

4.4.2 Multi-variable Logistic Regression Analysis

The results from the LR models for oesophagitis and pneumonitis using the two highest ranking features from MRMR feature selection are given in Table 20, with the results of logistic regression with samples bootstrapped 2000 times given in Table 21 and Table 22 for oesophagitis and pneumonitis respectively. ROC analysis for the two models with the highest overall predictive accuracy for oesophagitis and pneumonitis are reported in Figure 15.

Toxicity	Variables	Sig.	Percentage Correctly Predicted		
			Specificity	Sensitivity	Overall Accuracy
Oesophagitis	Sex	0.02	42.1	85.9	71.6
	V50	0.00			
Pneumonitis	Sex	0.33	100	0	77.6
	EUD	0.09			

Table 20 Table showing the results of multi-variable binary logistic regression analysis for factors predicting oesophagitis and pneumonitis. Predictive accuracies are based on a cut-off value of 0.5

Variable	B	Bias	Std. Error	Sig. (2-tailed)	95% Confidence Interval	
					Lower	Upper
Sex	1.25	0.08	0.59	0.02	0.23	2.73
V50	0.07	0.00	0.02	0.00	0.04	0.11

Table 21 shows the results of multivariable logistic regression using bootstrapping for oesophagitis

Variable	B	Bias	Std. Error	Sig. (2-tailed)	95% Confidence Interval	
					Lower	Upper
Sex	0.46	-0.02	0.52	0.34	-0.52	1.39
EUD	-0.11	-0.01	0.06	0.06	-0.23	-0.01

Table 22 shows the results of multivariable logistic regression using bootstrapping for pneumonitis

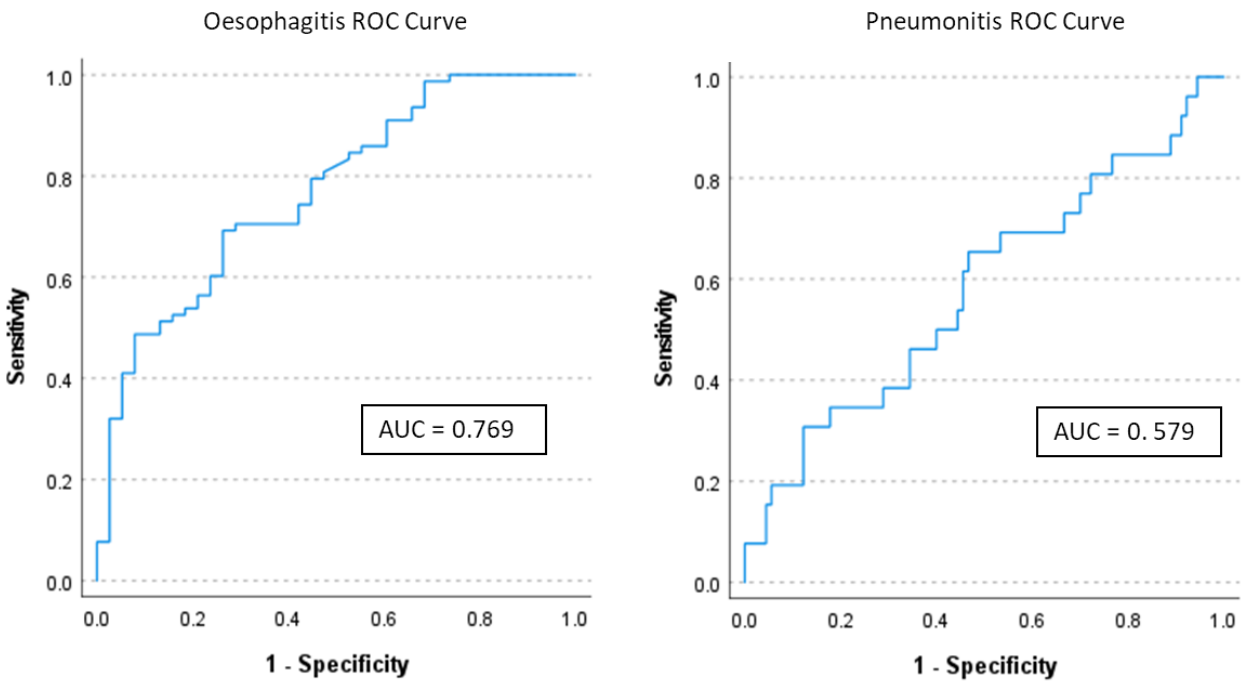


Figure 15 Receiver Operator Characteristic (ROC) curves for multi-variable Logistic Regression analysis of oesophagitis (left) and pneumonitis (right).

4.5 Discussion

With growing interest in the application of AI in all aspects of radiotherapy, some studies[8,51,52,54,55,97] have focussed on generating toxicity prediction models for oesophagitis and pneumonitis. One of clear advantages of these AI techniques is that they can easily be applied to large and diverse datasets and be used to model complex relationships, making them well suited to modelling toxicity in radiotherapy. A previously reported analysis[109] on the IDEAL-CRT trials data using machine learning techniques was able to produce good models for oesophagitis but struggled with pneumonitis. The study reported here compares established methods of LKB modelling and LR with ML to determine the advantages of these

different techniques for modelling oesophagitis and pneumonitis in patients from the IDEAL-CRT trial.

A strong correlation was found between the NTCP values and toxicity for oesophagitis using the Mann-Whitney U test for LKB modelling. This was also the case with LR analysis, with both statistical tests producing p-values <0.01. The p value of LR was robust when validated using the bootstrapped method. The LR predictive accuracy was similar to that achieved through previously reported uni-variable LR analyses[101], with high sensitivity (94.9%) but low specificity (28.9%). The low specificity would limit the use of the model clinically as it assumes most patients develop toxicity. The AUC was calculated as 0.678, which is lower than other models produced using the same data.

Neither the Mann-Whitney U test nor LR regression analysis showed a statistically significant correlation with pneumonitis. The predictive accuracy, sensitivity, and specificity were similar to uni-variable LR analysis.

Multi-variable LR was able to offer similar performance when compared to uni-variable analysis in terms of predictive accuracy, sensitivity and specificity for both oesophagitis and pneumonitis. The occurrence of oesophagitis had a strong correlation with various dose parameters but, as these metrics are interrelated, the benefits of using multiple dose metrics for LR was limited, hence MRMR was used to reduce the number of variables. V50Gy proved to be the most significant indicator of toxicity in the LR model for oesophagitis. Uni-variable LR analysis favoured specificity over sensitivity in predicting pneumonitis, a trend that was also apparent in multi-variable analysis. Here the EUD showed significant contribution to the LR model.

ROC analysis of multi-variable LR results produced AUC values for pneumonitis that were a slight improvement on those generated through ML, whilst the ML model AUC was slightly superior for oesophagitis. Results from the analysis by Patel et al[109] using machine learning for patients from the IDEAL-CRT trial are summarised in Table 23, for reference.

OAR	Variables	Percentage Correctly Predicted			ROC AUC
		Specificity	Sensitivity	Overall Accuracy	
Oesophagitis	Sex, V50	31.6%	93.6%	73.3%	0.79
Pneumonitis	Sex, EUD	3..8%	97.8%	76.7%	0.53

Table 23 The best performing ML models for oesophagitis and pneumonitis using the IDEAL-CRT dataset

Evaluating the three different approaches to modelling oesophagitis, a strong correlation is observed between dose metrics and the occurrence of oesophagitis. LKB modelling, which aims to distil the DVH to a single metric, shows a strong correlation with toxicity, however it does not offer an improvement on the standard dose metrics suggested by QUANTEC[60,77]. Multi-variable LR does offer an improvement over LBK modelling and uni-variable LR, providing higher specificity and leading to a higher overall accuracy and AUC. It also produces models comparable to ML using multiple classifiers

Previously reported uni-variable analyses using LR failed to identify a significant correlation between any of the tested dosimetric or clinical parameters and incidence of pneumonitis. The strongest clinical predictors were measures of lung function (FEV, FVC), with p-values of <0.05, and the strongest dosimetric factors were MD, V20Gy and EUD, with p-values <0.1. Given the weak correlation with dosimetric factors it is unsurprising that LBK analysis was also unable to show a strong statistical correlation with toxicity and resulted in a higher p-value than for MD, V20Gy and EUD. Multi-variable LR did not improve specificity and overall accuracy when compared with uni-variable LR, however ML techniques could offer a significantly higher specificity. Overall, results show the use of ML allows for more balanced models in terms of sensitivity and specificity.

A statistically significant correlation was observed between dosimetric factors and oesophagitis, and therefore ML and LR were able to produce predictive models with very similar performance. The LR model was able to achieve this producing a simpler model that is likely to be more robust. A simpler model is less likely to be affected by noise in the data and is more generalisable, being less likely to be overfitted to the original dataset. There was, however, little correlation seen for pneumonitis, and this is where ML outperformed LR. ML is able to utilise categorical and continuous data for analysis to produce predictive models, giving the user a wider choice of input data. ML models also offer more customisation than LR, allowing automatic model parameter optimisations and weighting of model parameters to improve sensitivity or specificity. This can produce more rounded models in data sets where one outcome dominates the overall predictive accuracy. The ML models also use K-fold cross validation, splitting up data sets into training and validation cohorts to ensure that

models do not overfit the data. LR uses the whole dataset for analysis and therefore would require a separate, independent dataset for validation of the results. Which is problematic given the data available for this study.

4.6 Conclusion

Predictive models require high specificity and selectivity in order to be clinically useful. Both LR and ML techniques can predict toxicity with similar accuracy when there is good correlation between metrics and toxicity. When there is not, machine learning's ability to utilise more diverse data and customise parameters of learning classifiers could enable superior toxicity models to be generated. Analyses reported here were limited to clinical and dosimetric parameters available from the IDEAL-CRT trial, however inclusion of dosimetric features from nearby OARs, spatial analysis of dose deposition within OARs, imaging biomarkers and genetic information in combination with ML have been applied successfully to create toxicity models. The addition of these variants has the potential to further improve toxicity models so they can become valuable tools, although validation on independent datasets is crucial before they can be translated to the clinic.

5 Critical Appraisal

Lung cancer is the leading cause of death from cancer in the UK. These patients have poor outcomes with lung cancer accounting for 21%[113] of all cancer deaths for men and women in the UK. Several recent radiotherapy clinical trials[10,84,85] for lung cancer have attempted to increase radiotherapy dose to the tumour using an isotoxic treatment model. The increased dose to the tumour improves the chance of tumour control but it also increases the risk of toxicity, as normal tissues are unavoidably irradiated during treatment. The intention of isotoxic radiotherapy is to escalate and individualise radiotherapy doses and find the ideal therapeutic ratio for each patient. In lung cancer, doses to organs at risk (OARs) such as the oesophagus and lungs are often the limiting factors in escalating tumour dose. Accurate prediction of oesophagitis and pneumonitis would allow treatment planners to safely escalate treatment doses for these patients improving treatment outcomes.

In the three papers presented in this study, clinical and dosimetric data for patients from the IDEAL-CRT trial were analysed to determine which data features have the strongest correlation with toxicity. This information was then utilised to develop predictive models for oesophagitis and pneumonitis to determine whether clinically useful toxicity models can be generated from the IDEAL-CRT trial data.

5.1 Oesophagitis

With regards to oesophagitis, uni-variable logistic regression (LR) analysis using the bootstrap method found a statistically significant ($p<0.01$) correlation between the occurrence oesophagitis and Mean Dose, V35Gy and V50Gy dose metrics. The most statistically significant factors were the V35 Gy and V50 Gy, which is in line with the recommendations of the QUANTEC paper[60] and suggests there may be a threshold dose for oesophageal toxicity. ADSCAN[85], a phase 3 platform clinical trial for isotoxic and hyperfractionated treatment regimens for non small cell lung cancer (NSCLC) applies a D0.1cc for 2 arms and a D1cc for one arm and no constraints for 2 arms. The LR analysis suggested that the three aforementioned dose metrics have superior correlation to the occurrence of toxicity than small volume constraints. It should be noted that all of the dose metrics tested are highly correlated and so the benefit of using multiple dose metrics for toxicity prediction may be limited.

Multi-variable LR provides a similar performance in terms of the overall predicative accuracy, sensitivity and specificity when compared to uni-variable analysis. The occurrence of oesophagitis is strongly correlated with dose and so the performance of LR analysis was dominated by the dose feature. When performing multivariable LR, MRMR feature selection ranked V50Gy and Sex as 1st and 2nd highest features respectively. The addition of patient sex to LR modelling had minimal impact on the overall performance of the model when compared to univariable modelling with the V50 Gy feature. The size of the IDEAL-CRT dataset has limited the number of features that can be used for modelling when mitigating the risk of overfitting. These results indicates that either additional clinical factors need to be explored or that a larger patient dataset (allowing more features in the model) would be required to create models with superior performance.

Lyman Kutcher Burman (LKB) toxicity modelling distils the entire DVH into a single dose parameter (EUD) which compensates for type of OAR and tissue. This type of modelling is entirely dependent on the dose information to predict toxicity, which would suggest it is well suited for modelling oesophagitis. While LR and Mann-Whitney U tests found that the correlation between LKB and oesophagitis were statistically significant, LR toxicity models using LKB data did not offer superior performance over uni-variable logistic regression with single dose parameters. This suggests the additional complication involved in generating LKB values was not worthwhile.

Toxicity models based on ML techniques performed similarly to multi-variable logistic regression. The dose metrics were the key data features for oesophagitis, with V50 Gy the best performing individual feature. Multiple ML models had strong performance and were able to produce high AUC values, with the simpler Naïve Bayes and more complex Neural Networks model providing similar performance. When evaluating the robustness of these models the simpler Naïve Bayes model demonstrated that it was more reproducible when changing the randomisation seed for the K-fold cross validation than the Neural Networks model. Suggesting that the more complex model was overfitting the data. This is another area where a larger dataset would be useful in producing more robust models and perhaps allowing separate training a validation cohorts which would provide a way to independently evaluate model performance.

5.2 Pneumonitis

Initial testing with uni-variable logistic regression using dose metrics recommended by QUANTEC[77] and relevant clinical trials[9,10,85,86] (V5 Gy, V20 Gy, Mean Dose and D1cc) and available clinical variables was unable to find a statistically significant correlation ($p<0.01$) with the occurrence of pneumonitis. The best performing data features were the FEV and FVC which achieved $p<0.05$. The best performing dose metric was MD, which is commonly used clinically along with the V20Gy metric. The V20 Gy and mean lung dose are widely used in clinical trials[85,91] but the results of this study indicate that baseline lung function may be a superior predictor of toxicity, additional dose metrics may improve correlation with toxicity. MRMR feature selection ranked Sex and EUD as 1st and 2nd highest features respectively, these features were used for multivariable LR and ML modelling. Multi-variable logistic regression had similar performance in terms of model sensitivity as uni-variable logistic regression, but the AUC was markedly improved.

LKB NTCP analysis did not demonstrate a benefit as it is based on dose, which was not found to correlate strongly with pneumonitis in previous analysis. Mann Whitney U and LR analysis was unable to find a statistically significant relationship between NTCP values and toxicity. ML techniques were able to produce toxicity models that were able to better balance specificity and sensitivity, where LR models largely favoured high specificity and poor sensitivity. The prediction of pneumonitis is perhaps where the benefits of ML approaches become apparent. The majority of patients in the study did not suffer from pneumonitis which meant that the dataset was unbalanced, and ML techniques can provide more complex methods to take this into account. Overall ML methods produced models with an AUC close to 0.5, which is similar performance as randomly guessing. Here the size of dataset which limits model training to two features and the lack of strong correlation of the tested features combines to give poor results. A larger dataset or features with a stronger correlation to toxicity are needed. The use of imaging biomarkers and genetic information to determine the radiosensitivity of the lungs prior to irradiation requires further investigation and has the potential to improve toxicity modelling. Palma et al[114] have also shown success predicting pneumonitis in patients by analysing spatial dose patterns, showing that the occurrence of toxicity is associated with dose to the lower portions of the lung and that it is possible to identify regional lung radiosensitivity. This work infers that refining dose features to sensitive parts of the lung or

adding tumour location (e.g. anterior, peripheral etc) would help further improve model performance using existing data.

5.3 Limits of Study

The IDEAL-CRT trial accrued patients using a mixture of 3D conformal radiotherapy and IMRT, this information could be extracted from the DICOM data and may have impacted toxicity. This is particular important as majority of UK centres have moved towards VMAT radiotherapy as the standard technique for lung patients and so the inclusion of 3DCRT may make this analysis less relevant for modern treatments. Doses to OARs in close proximity to the lung have been used as predictors for pneumonitis by Valdes et al[53], this data could have been extracted from the existing DICOM data set and may have improved toxicity models. It would have also been useful to gather information on patient smoking status, but this would require additional ethics approval.

All patients in the IDEAL-CRT trial were treated with concurrent chemotherapy which was standard practice at the time of the trial. The standard treatment of platinum doublet based chemo-radiotherapy for NSCLC has a 5 year survival of 15-32%[115]. The phase 3 Pacific trial[116] has shown immunotherapy agents such as Durvalumab significantly prolonged overall survival with long term analysis estimating 48-month OS rate was 49.6% for durvalumab verses 36.3% for placebo. The standard practice in the UK is now moving towards immunotherapy as standard of care for many lung cancer patients and the SARON trial has had a recent protocol amendment[117] to allow immunotherapy for trial patients. At this point the effects of immunotherapy on radiation sensitivity is unclear and results of this analysis may be less applicable to patients that are treated with it.

There was a mean dose limit of 18Gy in EQD2 for the lung for the IDEAL-CRT study, this was often a factor limiting in dose escalation and could have reduced variation in lung doses across the data cohort making it more difficult to model toxicity. The occurrence of grade 3 toxicity for pneumonitis and oesophagitis was limited to 3 and 5 patients respectively in the IDEAL-CRT trial. This means that the dataset could only be used to investigate grade 2 toxicity, for grade 2 toxicity 62.9% of patients suffered oesophagitis and 30.5% of patients suffered pneumonitis, the data sets were unbalanced which makes toxicity modelling more difficult as overall predicative accuracy is heavily influenced by the dominant outcome. Finally, the

number of patients available for this study was only 116 patients, ideally a larger patient dataset or an independent validation cohort would have been beneficial for toxicity modelling.

Data used in this study was from a phase 2 non randomised clinical trial, which had strict inclusion criteria for patient to be included. This study required a World Health Organisation performance status (WHO PS) of 0 or 1 and a FEV/DCLO value of ≥40%, which is not representative of the poor health status of most lung cancer patients, thus models derived from this data may not be generalisable to the real world. Also, whilst trial data is of high quality and has good data curation, the number of patients available for analysis is a limiting factor and it is here where the use of real-world data could be useful in producing more accurate and robust predictive models.

5.4 Conclusion

Data from randomized clinical trials provide high quality radiotherapy data with good patient follow up reported using a standardised methodology. Whilst this type of data generally has smaller patient numbers and cannot always be generalised to the real world due to patient selection criteria, it has high levels of curation and quality assurance which can reduce confounding effects and bias[99,100], which makes it useful for toxicity modelling. While technological advances since the IDEAL-CRT have led to an increased use of VMAT radiotherapy and immunotherapy as an alternative to chemotherapy, analysis of the IDEAL-CRT trial dataset has been worthwhile. The ADSCAN trial should have produced a more modern dataset of 360 patients in this space, but the trial has closed to recruitment early partially due to the effects on funding due covid but also as a result of the fast pace of change in the standard of lung cancer treatment effecting recruitment. It is unlikely that further randomised control trial data will be produced in this space in the near future with the prominent UK lung trials focusing on synchronous oligometastatic disease[74] and oligoprogressive disease[118].

Three different toxicity modelling techniques were assessed for oesophagitis and pneumonitis. The results of this study suggest that the utility of LKB NTCP modelling could now be an outdated approach as it is restricted to dosimetric data and cannot fully utilise the breadth of data that is available in modern radiotherapy datasets. This limitation is

demonstrated in the modelling of pneumonitis where dosimetric features had poor correlation with toxicity. The performance of LR and ML toxicity models in this study were similar in terms of overall predictive accuracy and AUC, but ML models were consistently able to produce more balanced models in terms of sensitivity and specificity. ML models offer greater scope for customisation through the weighting of model parameters to help achieve the desired specificity or sensitivity. The ML models also use K-fold cross validation, splitting up data sets into training and validation cohorts to ensure that models do not overfit the data while LR uses the whole dataset for analysis. This suggests that ML models are likely to be more generalisable to real world datasets, but this requires validation.

In terms of the toxicity models that have been produced in this study, the models for oesophagitis have been promising and consistently predict the patients that suffer from toxicity with high accurately but only attain moderate accuracy for detecting patients that do not. The pneumonitis models had poor balance in terms of sensitivity and specificity and do not provide the level of accuracy required to be clinically useful. In both cases it would be useful to test additional data features. Data features that could relatively easily be attained for the existing dataset would be smoking status, CT imaging biomarkers, spatial analysis of dose within organs and doses metrics for nearby OARs.

In conclusion ML approaches to toxicity modelling for oesophagitis and pneumonitis have shown encouraging results in building clinically useful toxicity models. With the proviso that these results need to be validated on an external dataset, analysis has shown a clear link between dosimetric factors and oesophagitis, although further exploration of data features is required for pneumonitis. ML models have the potential to be further refined through the incorporation of additional data features which can be derived from existing data and through the use of ML techniques it should be possible to develop clinically useful toxicity models that can improve patients' outcomes.

6 References

1. Lung cancer - NHS. Accessed September 13, 2019. https://www.nhs.uk/conditions/lung-cancer/

2. Emami B, Lyman J, Brown A, et al. Tolerance of normal tissue to therapeutic irradiation. *Int J Radiat Oncol Biol Phys*. 1991;21(1):109-122. doi:10.1016/0360-3016(91)90171-y

3. Marks LB, Yorke ED, Jackson A, et al. Use of Normal Tissue Complication Probability Models in the Clinic. *International Journal of Radiation Oncology Biology Physics*. 2010;76(3 SUPPL.). doi:10.1016/j.ijrobp.2009.07.1754

4. Vestergaard A, Muren LP, Søndergaard J, Elstrøm UV, Høyer M, Petersen JB. Adaptive plan selection vs. re-optimisation in radiotherapy for bladder cancer: a dose accumulation comparison. *Radiother Oncol*. 2013;109(3):457-462. doi:10.1016/j.radonc.2013.08.045

5. Papanikolaou N. Handbook of Radiotherapy Physics: Theory and Practice. *Medical Physics*. Published online 2008. doi:10.1118/1.2969650

6. Wideman TH, Zautra AJ, Edwards RR. Quantitative Analyses of Normal Tissue Effects in the Clinic (QUANTEC): An Introduction to the Scientific Issues. 2014;154(11):2262-2265. doi:10.1016/j.pain.2013.06.005.Re-Thinking

7. Burman C, Kutcher GJ, Emami B, Goitein AM. *FITTING OF NORMAL TISSUE TOLERANCE DATA TO AN ANALYTIC FUNCTION*. Vol 21. Pergamon Press plc; 1991.

8. Isaksson LJ, Pepa M, Zaffaroni M, et al. *Machine Learning-Based Models for Prediction of Toxicity Outcomes in Radiotherapy*. Vol 10. Frontiers Media S.A.; 2020:790. doi:10.3389/fonc.2020.00790

9. Hatton MQF, Hill R, Fenwick JD, et al. Continuous hyperfractionated accelerated radiotherapy – Escalated dose (CHART-ED): A phase I study. *Radiotherapy and Oncology*. 2016;118(3):471-477. doi:10.1016/j.radonc.2015.11.015

10. Landau DB, Hughes L, Baker A, et al. IDEAL-CRT: A Phase 1/2 Trial of Isotoxic Dose-Escalated Radiation Therapy and Concurrent Chemotherapy in Patients With Stage II/III

Non-Small Cell Lung Cancer. *International Journal of Radiation Oncology*Biology*Physics*. 2016;95(5):1367-1377. doi:10.1016/j.ijrobp.2016.03.031

11. Fowler JF. 21 Years of biologically effective dose. *British Journal of Radiology*. 2010;83(991):554-568. doi:10.1259/bjr/31372149

12. Elkind MM, Sutton H. Radiation Response of Mammalian Cells Grown in Culture: I. Repair of X-Ray Damage in Surviving Chinese Hamster Cells. *Radiation Research*. 1960;13(4):556. doi:10.2307/3570945

13. Fowler JF, Morgan RL, Silvester JA, Bewley DK, Turner BA. Experiments with Fractionated X-ray Treatment of the Skin of Pigs. I—Fractionation up to 28 Days. *The British Journal of Radiology*. 1963;36(423):188-196. doi:10.1259/0007-1285-36-423-188

14. Fowler JF, Kragt K, Ellis RE, Lindop PJ, Berry RJ. The Effect of Divided Doses of 15 MeV Electrons on the Skin Response of Mice. *International Journal of Radiation Biology and Related Studies in Physics, Chemistry and Medicine*. 1965;9(3):241-252. doi:10.1080/09553006514550291

15. Ellis F. Dose, time and fractionation: A clinical hypothesis. *Clinical Radiology*. 1969;20(1):1-7. doi:10.1016/S0009-9260(69)80043-7

16. Ellis F. Nominal standard dose and the ret. *The British Journal of Radiology*. 1971;44(518):101-108. doi:10.1259/0007-1285-44-518-101

17. Brenner DJ. The linear-quadratic model is an appropriate methodology for determining isoeffective doses at large doses per fraction. *Semin Radiat Oncol*. 2008;18(4):234-239. doi:10.1016/j.semradonc.2008.04.004

18. Douglas BG, Fowler JF. The Effect of Multiple Small Doses of X Rays on Skin Reactions in the Mouse and a Basic Interpretation. *Radiation Research*. 1976;66(2):401. doi:10.2307/3574407

19. Barendsen GW. Dose fractionation, dose rate and iso-effect relationships for normal tissue responses. *Int J Radiat Oncol Biol Phys*. 1982;8(11):1981-1997. doi:10.1016/0360-3016(82)90459-x

20. Withers, H., Thames, H. and Peters L. *Progress in Radio-Oncology II.*; 1982.

21. Guirado D, Aranda M, Ortiz M, et al. Low-dose radiation hyper-radiosensitivity in multicellular tumour spheroids. *Br J Radiol*. 2012;85(1018):1398-1406. doi:10.1259/bjr/33201506

22. Guerrero M, Li XA. Extending the linear–quadratic model for large fraction doses pertinent to stereotactic radiotherapy. *Physics in Medicine and Biology*. 2004;49(20):4825-4835. doi:10.1088/0031-9155/49/20/012

23. Park C, Papiez L, Zhang S, Story M, Timmerman RD, Timmerman R. UNIVERSAL SURVIVAL CURVE AND SINGLE FRACTION EQUIVALENT DOSE: USEFUL TOOLS IN UNDERSTANDING POTENCY OF ABLATIVE RADIOTHERAPY. doi:10.1016/j.ijrobp.2007.10.059

24. Curtis SB. Lethal and potentially lethal lesions induced by radiation--a unified repair model. *Radiat Res*. 1986;106(2):252-270.

25. Lyman JT. *Heavy Charged Particles in Research and Medicine*. Vol 8.

26. G. J. KUTCHER, PH.D. L, C. BURMAN, PH.D.,' L. BREWSTER, M.S.,' M. GOITEIN, PH.D.* AND R. MOHAN PHD l. HISTOGRAM REDUCTION METHOD FOR CALCULATING COMPLICATION PROBABILITIES FOR THREE-DIMENSIONAL TREATMENT PLANNING EVALUATIONS G. *Int J Radiarion Oncology Bid Phys Vol*. 1991;21:137-146.

27. Emami B, Lyman J, Brown ~ A, et al. *TOLERANCE OF NORMAL TISSUE TO THERAPEUTIC IRRADIATION*. Vol 21.

28. Gulliford SL, Foo K, Morgan RC, et al. Dose-Volume Constraints to Reduce Rectal Side Effects From Prostate Radiotherapy: Evidence From MRC RT01 Trial ISRCTN 47772397. *International Journal of Radiation Oncology Biology Physics*. 2010;76(3):747-754. doi:10.1016/j.ijrobp.2009.02.025

29. LYMAN JT. Complication Probability as Assessed from Dose-Volume Histograms Author (s): John T . Lyman Source : Radiation Research Supplement , Vol . 8 , Heavy Charged Particles in Research and Medicine . Proceedings of a Symposium Held at the Lawrence Berkeley La. 2017;8(May).

30. Cheung R, Tucker SL, Ye JS, et al. Characterization of rectal normal tissue complication probability after high-dose external beam radiotherapy for prostate cancer. *International Journal of Radiation Oncology*Biology*Physics*. 2004;58(5):1513-1519. doi:https://doi.org/10.1016/j.ijrobp.2003.09.015

31. TURING AM. I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*. 1950;LIX(236):433-460. doi:10.1093/mind/LIX.236.433

32. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. *European Urology, and Journal of Clinical Epidemiology*. 2015;162:55-63. doi:10.7326/M14-0697

33. Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH. MINIMAR (MINimum information for medical AI reporting): Developing reporting standards for artificial intelligence in health care. *Journal of the American Medical Informatics Association*. 2020;27(12). doi:10.1093/jamia/ocaa088

34. SYDES M, Stephens RJ, Moore AR, et al. Implementing the UK Medical Research Council (MRC) RT01 trial (ISRCTN 47772397): methods and practicalities of a randomised controlled trial of conformal radiotherapy in men with localised prostate cancer. *Radiotherapy and Oncology*. 2004;72(2):199-211. doi:10.1016/j.radonc.2004.04.007

35. Michalski JM, Gay H, Jackson A, Tucker SL, Deasy JO. RADIATION DOSE-VOLUME EFFECTS IN RADIATION-INDUCED RECTAL INJURY. *Radiation Oncology Biology*. 76:S123-S129. doi:10.1016/j.ijrobp.2009.03.078

36. Tucker SL, Thames HD, Michalski JM, et al. Estimation of α/β for late rectal toxicity based on RTOG 94-06. *International Journal of Radiation Oncology Biology Physics*. 2011;81(2):600-605. doi:10.1016/j.ijrobp.2010.11.080

37. Dutreix J. Fractionation in Radiotherapy. *International Journal of Radiation Biology and Related Studies in Physics, Chemistry and Medicine*. 1987;52(6):975-976. doi:10.1080/09553008714552561

38. Douglas BG, Fowler JF. The Effect of Multiple Small Doses of X Rays on Skin Reactions in the Mouse and a Basic. *Source: Radiation Research*. 2012;178(2). doi:10.1667/RRAV10.1

39. Wheldon TE, Deehan C, Wheldon EG, Barrett A. *The Linear-Quadratic Transformation of Dose-Volume Histograms in Fractionated Radiotherapy*.; 1998.

40. Tucker SL, Xu T, Paganetti H, et al. Validation of Effective Dose as a Better Predictor of Radiation Pneumonitis Risk than Mean Lung Dose: Secondary Analysis of a Randomized Trial. *International Journal of Radiation Oncology*Biology*Physics*. 2018;103(2):403-410. doi:10.1016/j.ijrobp.2018.09.029

41. Niemierko A. Reporting and analyzing dose distributions: A concept of equivalent uniform dose. *Medical Physics*. 1997;24(1):103-110. doi:10.1118/1.598063

42. Tucker SL, Liu HH, Liao Z, et al. Analysis of radiation pneumonitis risk using a generalized Lyman model. *Int J Radiat Oncol Biol Phys*. 2008;72(2):568-574. doi:10.1016/j.ijrobp.2008.04.053

43. Tucker SL, Mohan R, Liengsawangwong R, Martel MK, Liao Z. Predicting Pneumonitis Risk: A Dosimetric Alternative to Mean Lung Dose. *International Journal of Radiation Oncology*Biology*Physics*. 2013;85(2):522-527. doi:10.1016/j.ijrobp.2012.03.052

44. Huang EX, Bradley JD, El Naqa I, et al. Modeling the risk of radiation-induced acute esophagitis for combined Washington University and RTOG trial 93-11 lung cancer patients. *International Journal of Radiation Oncology Biology Physics*. 2012;82(5):1674-1679. doi:10.1016/j.ijrobp.2011.02.052

45. Huang EX, Robinson CG, Molotievschi A, Bradley JD, Deasy JO, Oh JH. Independent test of a model to predict severe acute esophagitis. *Advances in Radiation Oncology*. 2017;2(1):37-43. doi:10.1016/j.adro.2016.11.003

46. Ryckman JM, Baine M, Carmicheal J, et al. Correlation of dosimetric factors with the development of symptomatic radiation pneumonitis in stereotactic body radiotherapy. *Radiation Oncology*. 2020;15(1):33. doi:10.1186/s13014-020-1479-6

47. Makimoto T, Tsuchiya S, Hayakawa K, Saitoh R, Mori M. Risk factors for severe radiation pneumonitis in lung cancer. *Japanese Journal of Clinical Oncology*. 1999;29(4):192-197. doi:10.1093/jjco/29.4.192

48. Palma DA, Senan S, Tsujino K, et al. Predicting radiation pneumonitis after chemoradiation therapy for lung cancer: An international individual patient data meta-analysis. *International Journal of Radiation Oncology Biology Physics*. 2013;85(2):444-450. doi:10.1016/j.ijrobp.2012.04.043

49. Ricardi U, Filippi AR, Guarneri A, et al. Acta Oncologica Dosimetric predictors of radiation-induced lung injury in stereotactic body radiation therapy. Published online 2009. doi:10.1080/02841860802520821

50. Hart JP, McCurdy MR, Ezhil M, et al. Radiation Pneumonitis: Correlation of Toxicity With Pulmonary Metabolic Radiation Response. *International Journal of Radiation Oncology Biology Physics*. 2008;71(4):967-971. doi:10.1016/j.ijrobp.2008.04.002

51. Das SK, Chen S, Deasy JO, Zhou S, Yin FF, Marks LB. Combining multiple models to generate consensus: Application to radiation-induced pneumonitis prediction. *Medical Physics*. 2008;35(11):5098-5109. doi:10.1118/1.2996012

52. Das SK, Zhou S, Zhang J, Yin FF, Dewhirst MW, Marks LB. Predicting Lung Radiotherapy-Induced Pneumonitis Using a Model Combining Parametric Lyman Probit With Nonparametric Decision Trees. *International Journal of Radiation Oncology Biology Physics*. 2007;68(4):1212-1221. doi:10.1016/j.ijrobp.2007.03.064

53. Valdes G, Solberg TD, Heskel M, Ungar L, Simone CB. Using machine learning to predict radiation pneumonitis in patients with stage I non-small cell lung cancer treated with stereotactic body radiation therapy. *Physics in Medicine and Biology*. 2016;61(16):6105-6120. doi:10.1088/0031-9155/61/16/6105

54. El Naqa I, Bradley J, Blanco AI, et al. Multivariable modeling of radiotherapy outcomes, including dose-volume and clinical factors. *International Journal of Radiation Oncology Biology Physics*. 2006;64(4):1275-1286. doi:10.1016/j.ijrobp.2005.11.022

55. El Naqa I, Bradley JD, Lindsay PE, Hope AJ, Deasy JO. Predicting radiotherapy outcomes using statistical learning techniques. *Physics in Medicine and Biology*. 2009;54(18):S9. doi:10.1088/0031-9155/54/18/S02

56. Diez P, Hanna GG, Aitken KL, et al. UK 2022 Consensus on Normal Tissue Dose-Volume Constraints for Oligometastatic, Primary Lung and Hepatocellular Carcinoma Stereotactic Ablative Radiotherapy. *Clinical Oncology*. 2022;34(5):288-300. doi:10.1016/J.CLON.2022.02.010

57. Vogelius IS, Bentzen SM, Lebesque J V., et al. Radiation Dose–Volume Effects in the Lung. *International Journal of Radiation Oncology*Biology*Physics*. 2010;76(3):S70-S76. doi:10.1016/j.ijrobp.2009.06.091

58. Kwa SLS, Lebesque J v., Theuws JCM, et al. Radiation pneumonitis as a function of mean lung dose: an analysis of pooled data of 540 patients. *Int J Radiat Oncol Biol Phys*. 1998;42(1):1-9. doi:10.1016/S0360-3016(98)00196-5

59. Graham M v, Purdy JA, Emami B, et al. CLINICAL DOSE-VOLUME HISTOGRAM ANALYSIS FOR PNEUMONITIS AFTER 3D TREATMENT FOR NON-SMALL CELL LUNG CANCER (NSCLC). Published online 1999.

60. Werner-Wasik M, Yorke E, Deasy J, Nam J, Marks LB. Radiation Dose-Volume Effects in the Esophagus. *International Journal of Radiation Oncology Biology Physics*. 2010;76(3 SUPPL.):86-93. doi:10.1016/j.ijrobp.2009.05.070

61. Hancock SL, Donaldson SS, Hoppe RT. Cardiac disease following treatment of Hodgkin's disease in children and adolescents. *Journal of Clinical Oncology*. 1993;11(7):1208-1215. doi:10.1200/JCO.1993.11.7.1208

62. Adams MJ, Lipsitz SR, Colan SD, et al. Cardiovascular Status in Long-Term Survivors of Hodgkin's Disease Treated With Chest Radiotherapy. *Journal of Clinical Oncology*. 2004;22(15):3139-3148. doi:10.1200/JCO.2004.09.109

63. Vivekanandan S, Landau DB, Counsell N, et al. The Impact of Cardiac Radiation Dosimetry on Survival After Radiation Therapy for Non-Small Cell Lung Cancer. *International Journal of Radiation Oncology*Biology*Physics*. 2017;99(1):51-60. doi:10.1016/J.IJROBP.2017.04.026

64. Fenwick JD, Landau DB, Baker AT, et al. Long-Term Results from the IDEAL-CRT Phase 1/2 Trial of Isotoxically Dose-Escalated Radiation Therapy and Concurrent Chemotherapy for Stage II/III Non-small Cell Lung Cancer. *International Journal of Radiation Oncology Biology Physics*. 2020;106(4):733-742. doi:10.1016/j.ijrobp.2019.11.397

65. Riley RD, Ensor J, E Snell KI, et al. Calculating the sample size required for developing a clinical prediction model. Published online 2020. doi:10.1136/bmj.m441

66. van Smeden M, de Groot JAH, Moons KGM, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Medical Research Methodology*. 2016;16(1). doi:10.1186/s12874-016-0267-3

67. van Smeden M, Moons KGM, de Groot JAH, et al. Sample size for binary logistic prediction models: Beyond events per variable criteria. *Statistical Methods in Medical Research*. 2019;28(8). doi:10.1177/0962280218784726

68. Riley RD, Snell KIE, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: Part I – Continuous outcomes. *Statistics in Medicine*. 2019;38(7). doi:10.1002/sim.7993

69. Riley RD, Snell KIE, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Statistics in Medicine*. 2019;38(7). doi:10.1002/sim.7992

70. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. 394 CA: A Cancer Journal for Clinicians Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA CANCER J CLIN*. 2018;68:394-424. doi:10.3322/caac.21492

71. National Institute for Health and Care Excellence (NICE). Prostate cancer : diagnosis and management. NICE guideline. *National institute for health and care excellence*. 2019;(September 2017):2020.

72. Mehta V. RADIATION PNEUMONITIS AND PULMONARY FIBROSIS IN NON-SMALL-CELL LUNG CANCER: PULMONARY FUNCTION, PREDICTION, AND PREVENTION. Published online 2005. doi:10.1016/j.ijrobp.2005.03.047

73. Canadian Cancer Society. Radiation pneumonitis - Canadian Cancer Society. Published 2009. Accessed May 14, 2017. http://www.cancer.ca/en/cancer-information/diagnosis-and-treatment/radiation-therapy/side-effects-of-radiation-therapy/radiation-to-the-chest/radiation-pneumonitis/?region=on

74. Conibear J, Chia B, Ngai Y, et al. Study protocol for the SARON trial: a multicentre, randomised controlled phase III trial comparing the addition of stereotactic ablative radiotherapy and radical radiotherapy with standard chemotherapy alone for oligometastatic non-small cell lung cancer. *BMJ Open*. 2018;8(4):e020690. doi:10.1136/BMJOPEN-2017-020690

75. *Common Terminology Criteria for Adverse Events (CTCAE)*.

76. Jones B, Dale RG, Deehan C, Hopkins KI, Morgan DAL. The role of biologically effective dose (BED) in clinical oncology. *Clinical Oncology*. 2001;13(2):71-81. doi:10.1053/clon.2001.9221

77. Marks LB, Bentzen SM, Deasy JO, et al. Radiation Dose-Volume Effects in the Lung. *International Journal of Radiation Oncology Biology Physics*. 2010;76(3 SUPPL.):70-76. doi:10.1016/j.ijrobp.2009.06.091

78. Bentzen SM, Skoczylas JZ, Bernier J. Quantitative clinical radiobiology of early and late lung reactions. In: *International Journal of Radiation Biology*. Vol 76. Taylor and Francis Ltd; 2000:453-462. doi:10.1080/095530000138448

79. Hawkins PG, Boonstra PS, Hobson ST, et al. Prediction of Radiation Esophagitis in Non–Small Cell Lung Cancer Using Clinical Factors, Dosimetric Parameters, and Pretreatment Cytokine Levels. *Translational Oncology*. 2018;11(1). doi:10.1016/j.tranon.2017.11.005

80. Soni PD, Boonstra PS, Schipper MJ, et al. Lower Incidence of Esophagitis in the Elderly Undergoing Definitive Radiation Therapy for Lung Cancer. *Journal of Thoracic Oncology*. 2017;12(3):539-546. doi:10.1016/J.JTHO.2016.11.2227

81. Bentzen SM, Skoczylas JZ, Bernier J. Quantitative clinical radiobiology of early and late lung reactions. *International Journal of Radiation Biology*. 2000;76(4):453-462. doi:10.1080/095530000138448

82. Bentzen SM, Saunders MI, Dische S. From CHART to CHARTWELL in non-small cell lung cancer: Clinical radiobiological modelling of the expected change in outcome. *Clinical Oncology*. 2002;14(5):372-381. doi:10.1053/clon.2002.0117

83. Fowler JF, Tomé WA, Fenwick JD, Mehta MP. A challenge to traditional radiation oncology. *International Journal of Radiation Oncology*Biology*Physics*. 2004;60(4):1241-1256. doi:10.1016/J.IJROBP.2004.07.691

84. Haslett K, Franks K, Hanna GG, et al. Protocol for the isotoxic intensity modulated radiotherapy (IMRT) in stage III non-small cell lung cancer (NSCLC): A feasibility study. *BMJ Open*. 2016;6(4):e010457. doi:10.1136/bmjopen-2015-010457

85. Hatton MQF, Lawless CA, Faivre-Finn C, et al. Accelerated, Dose escalated, Sequential Chemoradiotherapy in Non-small-cell lung cancer (ADSCaN): A protocol for a randomised phase II study. *BMJ Open*. 2019;9(1):19903. doi:10.1136/bmjopen-2017-019903

86. Lester JF, Courtier N, Eswar C, et al. Initial results of the phase ib/II, I-START trial: Isotoxic accelerated radiotherapy for the treatment of stage II-IIIb NSCLC. *Journal of Clinical Oncology*. 2018;36(15_suppl):e20551-e20551. doi:10.1200/jco.2018.36.15_suppl.e20551

87. Thor M, Deasy J, Iyer A, et al. Toward personalized dose-prescription in locally advanced non-small cell lung cancer: Validation of published normal tissue complication probability models. *Radiotherapy and Oncology*. 2019;138:45-51. doi:10.1016/j.radonc.2019.05.011

88. Pignon T, Gregor A, Schaake Koning C, Roussel A, Van Glabbeke M, Scalliet P. Age has no impact on acute and late toxicity of curative thoracic radiotherapy. *Radiotherapy and Oncology*. 1998;46(3):239-248. doi:10.1016/S0167-8140(97)00188-6

89. Marks LB, ten Haken RK, Martel MK. Guest Editor's Introduction to QUANTEC: A Users Guide. *International Journal of Radiation Oncology Biology Physics*. 2010;76(3 SUPPL.):2009-2010. doi:10.1016/j.ijrobp.2009.08.075

90. Olsson CE, Jackson A, Deasy JO, Thor M. A Systematic Post-QUANTEC Review of Tolerance Doses for Late Toxicity After Prostate Cancer Radiation Therapy.

International Journal of Radiation Oncology Biology Physics. 2018;102(5). doi:10.1016/j.ijrobp.2018.08.015

91. Conibear J, Chia B, Ngai Y, et al. Study protocol for the SARON trial: A multicentre, randomised controlled phase III trial comparing the addition of stereotactic ablative radiotherapy and radical radiotherapy with standard chemotherapy alone for oligometastatic non-small cell lung cancer. *BMJ Open*. 2018;8(4). doi:10.1136/bmjopen-2017-020690

92. Van Der Kogel A. *Basic Clinical Radiobiology*.; 2009.

93. Deasy JO, Naqa I El. Image-Based Modeling of Normal Tissue Complication Probability for Radiation Therapy. In: Bentzen SM, Harari PM, Tomé WA, Mehta MP, eds. *Radiation Oncology Advances*. Springer US; 2008:211-252. doi:10.1007/978-0-387-36744-6_11

94. El Naqa I. *Machine Learning in Radiation Oncology*. (El Naqa I, Li R, Murphy MJ, eds.). Springer International Publishing; 2015. doi:10.1007/978-3-319-18305-3

95. Luna JM, Chao HH, Diffenderfer ES, et al. Predicting radiation pneumonitis in locally advanced stage II–III non-small cell lung cancer using machine learning. *Radiotherapy and Oncology*. 2019;133:106-112. doi:10.1016/j.radonc.2019.01.003

96. Lee S, Ybarra N, Jeyaseelan K, et al. Bayesian network ensemble as a multivariate strategy to predict radiation pneumonitis risk. *Medical Physics*. 2015;42(5):2421-2430. doi:10.1118/1.4915284

97. Bousabarah K, Temming S, Hoevels M, et al. Radiomic analysis of planning computed tomograms for predicting radiation-induced lung injury and outcome in lung cancer patients treated with robotic stereotactic body radiation therapy. *Strahlentherapie und Onkologie*. 2019;195(9):830-842. doi:10.1007/s00066-019-01452-7

98. Niedzielski JS, Yang J, Stingo F, et al. A Novel Methodology using CT Imaging Biomarkers to Quantify Radiation Sensitivity in the Esophagus with Application to Clinical Trials. *Scientific Reports*. 2017;7(1). doi:10.1038/s41598-017-05003-x

99. Zhong H, Men K, Wang J, et al. The impact of clinical trial quality assurance on outcome in head and neck radiotherapy treatment. *Frontiers in Oncology*. 2019;9(AUG):792. doi:10.3389/fonc.2019.00792

100. Thompson MK, Poortmans P, Chalmers AJ, et al. Practice-changing radiation therapy trials for the treatment of cancer: where are we 150 years after the birth of Marie Curie? *British Journal of Cancer*. 2018;119(4):389-407. doi:10.1038/s41416-018-0201-z

101. Patel R, Venables K, Aitkenhead A, Farrelly L, Counsell N, Landau D. Analysis of the significance of clinical and dosimetric factors for the prediction of radiation induced oesophagitis and pneumonitis. Published online 2021.

102. Casella G, Fienberg S, Olkin I. *An Introduction to Statistical Learning*.; 2013.

103. Peng H, Long F, Ding C. Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2005;27(8). doi:10.1109/TPAMI.2005.159

104. Radovic M, Ghalwash M, Filipovic N, Obradovic Z. Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. *BMC Bioinformatics*. 2017;18(1):1-14. doi:10.1186/S12859-016-1423-9/FIGURES/6

105. *The Timely Delivery of Radical Radiotherapy: Guidelines for the Management of Unscheduled Treatment Interruptions Fourth Edition*.; 2019.

106. Huang K, Dahele M, Senan S, et al. Radiographic changes after lung stereotactic ablative radiotherapy (SABR) - Can we distinguish recurrence from fibrosis? A systematic review of the literature. *Radiotherapy and Oncology*. 2012;102(3):335-342. doi:10.1016/j.radonc.2011.12.018

107. Kang J, Coates JT, Strawderman RL, Rosenstein BS, Kerns SL. Genomics models in radiotherapy: From mechanistic to machine learning. *Medical Physics*. 2020;47(5):e203-e217. doi:10.1002/mp.13751

108. Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *International Journal of Cancer*. 2010;127(12):2893-2917. doi:10.1002/ijc.25516

109. Patel R, Venables K, Aitkenhead A, Farrelly L, Counsell N, Landau D. Predicting radiotherapy toxicity for NSCLC patients using Machine Learning Techniques.

110. van Leeuwen CM, Oei AL, Crezee J, et al. The alfa and beta of tumours: A review of parameters of the linear-quadratic model, derived from clinical radiotherapy studies. *Radiation Oncology*. 2018;13(1):1-11. doi:10.1186/s13014-018-1040-z

111. Belderbos J, Heemsbergen W, Hoogeman M, Pengel K, Rossi M, Lebesque J. Acute esophageal toxicity in non-small cell lung cancer patients after high dose conformal radiotherapy. *Radiotherapy and Oncology*. 2005;75(2):157-164. doi:10.1016/j.radonc.2005.03.021

112. Lee S, Park S, Choi E, et al. Analysis of Radiation Pneumonitis on Consecutive CTs using LKB Model Parameters for Non-small Cell Lung Cancer. *International Journal of Radiation Oncology*Biology*Physics*. 2011;81(2):S609. doi:10.1016/j.ijrobp.2011.06.1146

113. Cancer statistics | World Cancer Research Fund UK. Published online 2014.

114. Palma G, Monti S, Xu T, et al. Spatial Dose Patterns Associated With Radiation Pneumonitis in a Randomized Trial Comparing Intensity-Modulated Photon Therapy With Passive Scattering Proton Therapy for Locally Advanced Non-Small Cell Lung Cancer. *International Journal of Radiation Oncology Biology Physics*. 2019;104(5). doi:10.1016/j.ijrobp.2019.02.039

115. Faivre-Finn C, Vicente D, Kurata T, et al. Four-Year Survival With Durvalumab After Chemoradiotherapy in Stage III NSCLC—an Update From the PACIFIC Trial. *Journal of Thoracic Oncology*. 2021;16(5). doi:10.1016/j.jtho.2020.12.015

116. Antonia SJ, Villegas A, Daniel D, et al. Durvalumab after Chemoradiotherapy in Stage III Non–Small-Cell Lung Cancer. *New England Journal of Medicine*. 2017;377(20). doi:10.1056/nejmoa1709937

117. McDonald F, Mak KM, Teague J, et al. SARON: Stereotactic Ablative Radiotherapy for Oligometastatic Non-small cell lung cancer (NSCLC): a randomised phase III trial. *Lung Cancer*. 2020;139:S91. doi:10.1016/s0169-5002(20)30243-9

118. McDonald F, Hanna GG. Oligoprogressive Oncogene-addicted Lung Tumours: Does Stereotactic Body Radiotherapy Have a Role? Introducing the HALT Trial. *Clinical Oncology*. 2018;30(1):1-4. doi:10.1016/j.clon.2017.10.013

7 Appendices

7.1 Appendix 1: HRA Approval Letter

Ymchwil Iechyd a Gofal Cymru
Health and Care Research Wales

NHS
Health Research Authority

Dr Karen Venables
Radiotherapy, Mount Vernon Cancer Centre
Rickmansworth Rd
Northwood
HA4 9TH

Email: hra.approval@nhs.net
HCRW.approvals@wales.nhs.uk

21 January 2020

Dear Dr Venables

HRA and Health and Care Research Wales (HCRW) Approval Letter

Study title:	**Generation of thoracic dose constraints based on biologically equivalent dose for patients from the IDEAL-CRT trial**
IRAS project ID:	**271980**
Protocol number:	**RD2019-76**
REC reference:	**20/HRA/0139**
Sponsor	**East and North Hertfordshire NHS Trust**

I am pleased to confirm that **HRA and Health and Care Research Wales (HCRW) Approval** has been given for the above referenced study, on the basis described in the application form, protocol, supporting documentation and any clarifications received. You should not expect to receive anything further relating to this application.

Please now work with participating NHS organisations to confirm capacity and capability, in line with the instructions provided in the "Information to support study set up" section towards the end of this letter.

How should I work with participating NHS/HSC organisations in Northern Ireland and Scotland?
HRA and HCRW Approval does not apply to NHS/HSC organisations within Northern Ireland and Scotland.

If you indicated in your IRAS form that you do have participating organisations in either of these devolved administrations, the final document set and the study wide governance report (including this letter) have been sent to the coordinating centre of each participating nation. The relevant national coordinating function/s will contact you as appropriate.

Please see IRAS Help for information on working with NHS/HSC organisations in Northern Ireland and Scotland.

How should I work with participating non-NHS organisations?
HRA and HCRW Approval does not apply to non-NHS organisations. You should work with your non-NHS organisations to obtain local agreement in accordance with their procedures.

What are my notification responsibilities during the study?

The *"After HRA Approval – guidance for sponsors and investigators"* document on the HRA website gives detailed guidance on reporting expectations for studies with HRA and HCRW Approval, including:

- Registration of Research
- Notifying amendments
- Notifying the end of the study

The HRA website also provides guidance on these topics and is updated in the light of changes in reporting expectations or procedures.

Who should I contact for further information?
Please do not hesitate to contact me for assistance with this application. My contact details are below.

Your IRAS project ID is **271980**. Please quote this on all correspondence.

Yours sincerely,
Barbara Cuddon

Approvals Specialist

Email: hra.approval@nhs.net

Copy to: *Prof Philip Smith*

List of Documents

The final document set assessed and approved by HRA and HCRW Approval is listed below.

Document	*Version*	*Date*
IRAS Application Form [IRAS_Form_08012020]		08 January 2020
IRAS Application Form XML file [IRAS_Form_08012020]		08 January 2020
IRAS Checklist XML [Checklist_08012020]		08 January 2020
Letter from sponsor [letter from sponser]	1	18 December 2019
Other [Support letter from IDEAL-CRT trials unit and CI]	1	24 September 2019
Other [REC not required confirmation]	1	17 December 2019
Research protocol or project proposal [RD2019-76 Generation of thoracic dose constraints based on BED study Protocol v1.0 05Dec19.pdf]	1	05 December 2019
Summary CV for Chief Investigator (CI) [CI CV]	1	05 December 2019
Summary CV for student [R Patel CV]	1	05 December 2019

IRAS project ID	271980

Information to support study set up

The below provides all parties with information to support the arranging and confirming of capacity and capability with participating NHS organisations in England and Wales. This is intended to be an accurate reflection of the study at the time of issue of this letter.

Types of participating NHS organisation	Expectations related to confirmation of capacity and capability	Agreement to be used	Funding arrangements	Oversight expectations	HR Good Practice Resource Pack expectations
This is a single site study sponsored by the participating NHS organisation therefore there is only one site type.	This is a single site study sponsored by the participating NHS organisation. You should work with your sponsor R&D office to make arrangements to set up the study. The sponsor R&D office will confirm to you when the study can start following issue of HRA and HCRW Approval.	This is a single site study sponsored by the participating NHS organisation therefore no agreements are expected.	No external study funding has been sought	It is expected that a Principal Investigator should be appointed at study sites	The sponsor has confirmed that local staff in participating organisations in England who have a contractual relationship with the organisation will undertake the expected activities. Therefore no honorary research contracts or letters of access are expected for this study.

Other information to aid study set-up and delivery

This details any other information that may be helpful to sponsors and participating NHS organisations in England and Wales in study set-up.
The applicant has indicated that they do not intend to apply for inclusion on the NIHR CRN Portfolio.

7.2 Appendix 2: Letter of Support from IDEAL-CRT Study

Cancer Research UK

University College London
90 Tottenham Court Road
London W1T 4TJ
e-mail: ctc.ideal-crt@ucl.ac.uk
website: http://www.ctc.ucl.ac.uk/

Health Research Authority (HRA)

To whom it may concern,

TRIAL: A phase I/II trial of concurrent chemoradiation with dose-escalated radiotherapy in patients with stage II or stage III Non-Small Cell Lung Cancer
REC reference number: 09/H0707/38
ISRCTN no: 12155469

The above study declared end of trial on 24/04/2017

Re: Generation of thoracic dose constraints based on biologically equivalent doses for patients from the IDEAL-CRT Trial

The above analyses has been discussed with the Chief Investigator and CR UK and UCL Cancer Trials Centre regarding data generated within the IDEAL-CRT Trial. This project is in line with the original objectives of the IDEAL-CRT trial and would be a suitable use of existing trial data. The outcome of this research project would provide useful information for current and future Lung Cancer trials and has the support of the IDEAL-CRT chief investigator David Landau and the CRUK and UCL Cancer Trials Centre.

A Data Sharing Agreement would be put in place prior to a data exchange.

Yours faithfully,

Laura Farrelly
Trials Group Lead (Lung/Gynae)
CRUK and UCL Cancer Trials Centre

David Landau
Chief Investigator - IDEAL CRT

7.3 Appendix 3: MATLAB Code

7.3.1 EQD2 conversion and dose reporting script

All MATLAB code is based on example code provided by Dr Colin Baker and have been adapted and rewritten for this project.

This section of code reads all DVH csv files generated by the Eclipse Treatment planning system from a folder. Converts the data into an array for the desired OAR, converts this to EQD2 and writes pre-set patients and dosimetric values to a text document for each file.

```
% Rush test for C2 project to read data from an absolute Eclipse DVH export

%looping through a folder and writing to file: think this should be in a
%separate file and will then call the function to go through each file and
%pull out the data
files = dir('G:\My Drive\HSST\Module C\C2\Matlab PUnC 2020\DVH-data\*.txt') ;    %
you are in folder of txt files
N = length(files) ;   % total number of files
% loop for each file


for i = 1:N
    thisfile = files(i).name ;
    % do what you want


    %OAR = 'Oesophagus';
    OAR = 'Lungs-GTV'; %name of the structure being analysed
    fidOAR=fopen(thisfile);  %location of the file
    ab_OAR = 4;  %alpha beta ratio used for dose conversion
    k = 0.54;
    t = 40;
    tdelay = 0;
    Nfract = 30; %number of fractions
    nLKB = 0.5;

        % runs the import eclipse function
    [DoseOAR, dDVHoar, ID] = ReadDVHEclipse(fidOAR,OAR);  %DoseOAR = dose value of
the bin, dDVHoar is the differential volume of the bin
        %fprintf('%s\n',t);
        %fprintf('%s\n',ID{6});
   for j = 0:4
        tdelay = j*7;

        if strcmp('1',ID{6})
            t=33;
        end
        %fprintf('%d\n',t);
        Dcor2(1)=DoseOAR(1)+0.5*(DoseOAR(2)-DoseOAR(1));
        nmaxD=0;
        for ibin=2:length(DoseOAR)-1
            Dcor2(ibin)=DoseOAR(ibin)+0.5*(DoseOAR(ibin+1)-DoseOAR(ibin));
            %also establish max dose bin
            if dDVHoar(ibin)>0
```

```
                nmaxD=ibin;
            end
        end

Dcor2(length(DoseOAR))=DoseOAR(length(DoseOAR))+0.5*(DoseOAR(length(DoseOAR))-
DoseOAR(length(DoseOAR)-1));
        %DoseOAR=Dcor2;

        %Normalise to relative volume - assume absolute volumes from DVH files
        OARvol=sum(dDVHoar);  %sum of the volumes
        dDVHoar=100*dDVHoar/OARvol; %dDVHoar is now relative

        %conversion factor to 2Gy fraction equivalence **********************
        lam = 0;
        for iDose=1:length(DoseOAR)
            lam(iDose)=(((DoseOAR(iDose)*(ab_OAR + DoseOAR(iDose)/Nfract))-(k*(t-
tdelay)))/(ab_OAR + 2)); %I've multiplied by the total dose to be safe

        end

        % working out the D1cc in Gy
        d1ccinpercent = 0;
        d1vol = 0;
        d1cc = 0;
        pos=0;
        d1ccinpercent = (1/OARvol)*100;
        for iDose=length(dDVHoar):-1:1
            if d1vol<=d1ccinpercent
                d1vol = d1vol + dDVHoar(iDose);
                pos = iDose;
            end
        end
        d1cc = lam(pos);

         %Normalise to relative volume - assume absolute volumes from DVH files
        OARvol=sum(dDVHoar);
        dDVHoar=100*dDVHoar/OARvol;
        % LKB
model**************************************************************
        sumeud=0;
        dDVHoar;
        for iDose=1:length(lam)
            eud(iDose)=0.01*dDVHoar(iDose)*(lam(iDose)^(1/nLKB));
            sumeud=sumeud+eud(iDose);
        end
        EUDoar=sumeud^nLKB;

        % working out the V5Gy as a percentage of the total lung volume
        fiveGy = 0;
        vfive = 0;
        for iDose=1:length(lam)
            if lam(iDose)>=5
                fiveGy = fiveGy + dDVHoar(iDose);
            end
        end
        vfive = (fiveGy/sum(dDVHoar))*100;

        %working out the V10Gy as a percentage of the total lung volume
        tenGy = 0;
```

```
        vten = 0;
        for iDose=1:length(lam)
            if lam(iDose)>=10
                tenGy = tenGy + dDVHoar(iDose);
            end
        end

        vtwenty = (twentyGy/sum(dDVHoar))*100;
%
        % working out the V35Gy as a percentage of the total lung volume
        v35a = 0;
        v35 = 0;
        for iDose=1:length(lam)
            if lam(iDose)>=35
                v35a = v35a + dDVHoar(iDose);
            end
        end
        v35 = (v35a/sum(dDVHoar))*100;

        % working out the V50Gy as a percentage of the total lung volume
        v50a = 0;
        v50 = 0;
        for iDose=1:length(lam)
            if lam(iDose)>=50
                v50a = v50a + dDVHoar(iDose);
            end
        end
        v50 = (v50a/sum(dDVHoar))*100;

        % working out the V70Gy as a percentage of the total lung volume
        v70a = 0;
        v70 = 0;
        for iDose=1:length(lam)
            if lam(iDose)>=70
                v70a = v70a + dDVHoar(iDose);
            end
        end
        v70 = (v70a/sum(dDVHoar))*100;

        % mean dose calculation
        meandose = 0;
        for iDose=1:length(lam)
            meandose = meandose + (lam(iDose)*(dDVHoar(iDose)/(sum(dDVHoar))));
        end

        %fprintf(1,'\n%s%5.2f','V20Gy [%] = ',vtwenty);
        %fprintf(1,'\n%s%5.2f','Mean Dose [Gy] = ',meandose);
        %fprintf(1, '\n%s%4.1f', 'OAR a/b [Gy] ', ab_OAR);
        %prints everthing into csv format
        fid = fopen('G:\My Drive\HSST\Module C\C2\Matlab PUnC
2020\tdelay\lung.txt','a'); %opens the file
        fprintf(fid,'\n'); %prints a new line
        fprintf(fid,'%s',ID{1:8},','); %adds the trials ID then comma
        fprintf(fid,'%s',OAR,','); % adds OAR name
        fprintf(fid,'%f%s',ab_OAR,','); %adds alpha-beta ratio
        fprintf(fid,'%f%s',k,','); %
        fprintf(fid,'%f%s',t,','); %
        fprintf(fid,'%f%s',tdelay,','); %
        fprintf(fid,'%f%s',d1cc,','); %adds D1cc
```

```
        fprintf(fid,'%f%s',EUDoar,','); %adds EUD OAR
        fprintf(fid,'%f%s',vfive,','); %adds V5
        fprintf(fid,'%f%s',vtwenty,','); %adds V20
        fprintf(fid,'%f%s',v35,','); %adds V35
        fprintf(fid,'%f%s',v50,','); %adds V50
        fprintf(fid,'%f%s',v70,','); %adds V70
        fprintf(fid,'%f%s',meandose,','); % adds mean dose
        fclose(fid); % closes file

    fclose('all');
    end
end
```

7.3.2 EUD Calculation

The code below generates the EUD value when used in the script in 7.5.1

```
dDVHoar;
        for iDose=1:length(lam)
            eud(iDose)=0.01*dDVHoar(iDose)*(lam(iDose)^(1/nLKB));
            sumeud=sumeud+eud(iDose);
        end
        EUDoar=sumeud^nLKB;
```

7.3.3 LKB Calculation

LKB NTCP calculation script assuming data import as in section 7.5.1

```
% LKB model*****************************************************************
        sumeud=0;
        dDVHoar;
        for iDose=1:length(lam)
            eud(iDose)=0.01*dDVHoar(iDose)*(lam(iDose)^(1/nLKB));
            sumeud=sumeud+eud(iDose);
        end
        EUDoar=sumeud^nLKB;
        tval=(EUDoar-TD50)/(mLKB*TD50);

        if tval>0
            NTCPLKB=100*(0.5*(1+erf(tval/sqrt(2))));
        else
            NTCPLKB=100*(0.5*(1-erf(-tval/sqrt(2))));
        end
```

7.4 Appendix 4: MINIMAR Compliance Table

Minimar AI Reporting		
Features	Description	Section
1. Study population and setting		
Population	Population from which study sample was drawn	1.4
Study setting	The setting in which the study was conducted (eg, academic medical left, community healthcare system, rural healthcare clinic)	1.4
Data source	The source from which data were collected	1.4
Cohort selection	Exclusion/inclusion criteria	1.4, 5.3
2. Patient demographic characteristics		
Age	Age of patients included in the study	2.4.2
Sex	Sex breakdown of study cohort	2.4.2
Race	Race characteristics of patients included in the study	Not Available
Ethnicity	Ethnicity breakdown of patients included in the study	Not Available
Socioeconomic status	A measure or proxy measure of the socioeconomic status of patients included in the study	Not Available
3. Model architecture		

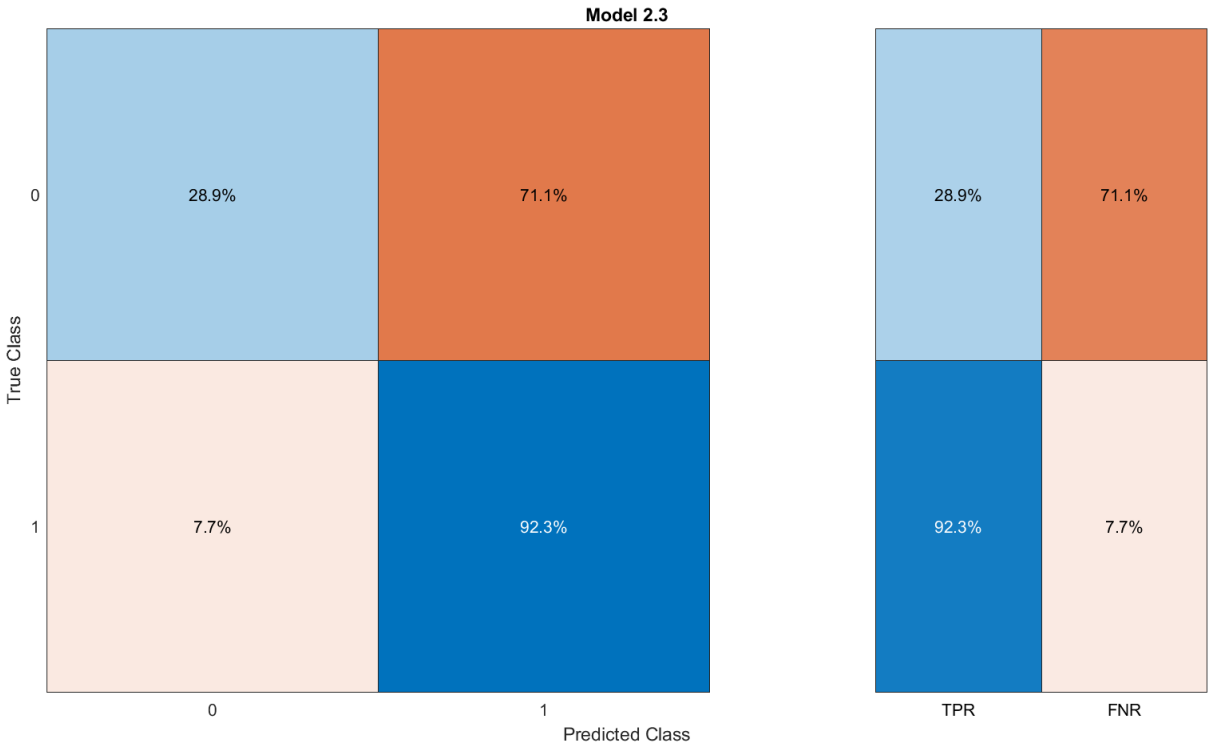
Model output	The computed result of the model	2.3, 3.3.2
Target user	The indented user of the model output (eg, clinician, hospital management team, insurance company)	5
Data splitting	How data were split for training, testing, and validation	3.3.2
Gold standard	Labeled data used to train and test the model	N/A
Model task	Classification or prediction	3.3.2
Model architecture	Algorithm type (eg, machine learning, deep learning, etc.)	3.3.2
Features	List of variables used in the model and how they were used in the model in terms of categories or transformation	3.3.1
Missingness	How missingness was addressed: reported, imputed, or corrected	N/A
4. Model evaluation		
Optimization	Model or parameter tuning applied	3.4
Internal model validation	Study internal validation	3.3.2
External model validation	External validation using data from another setting	N/A
Transparency	How code and data are shared with the community.	7.3

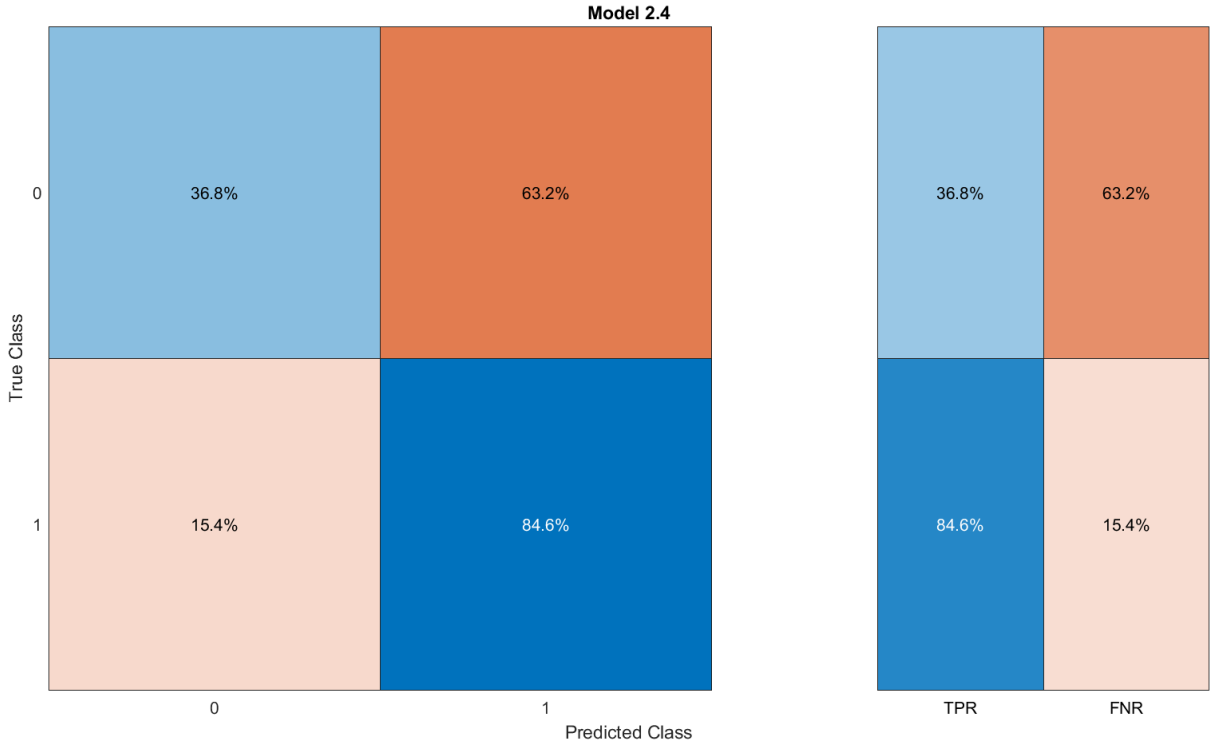
7.5 Appendix 5: Confusion Matrices

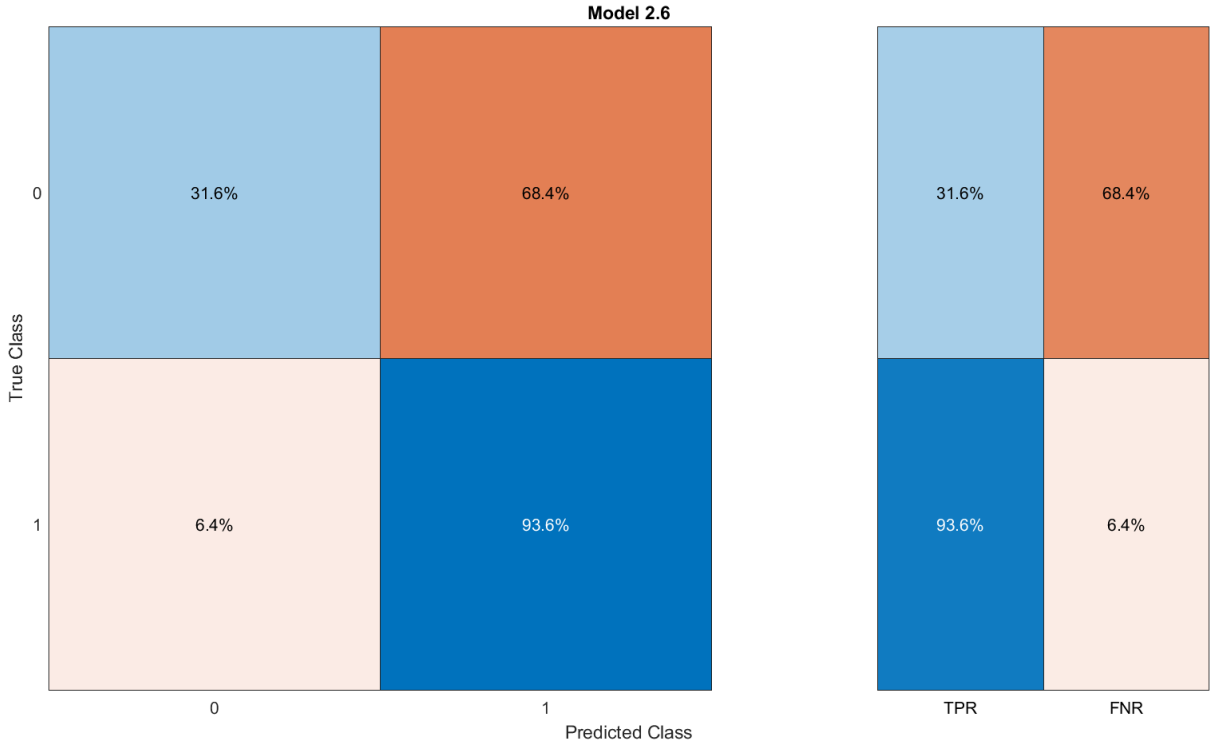
The confusion matrices below visually demonstrate ML model performance using the MRMR selected features. Here 0 and 1 represent no toxicity and toxicity respectively, TPR is the true positive rate and FVR is the False negative rate.

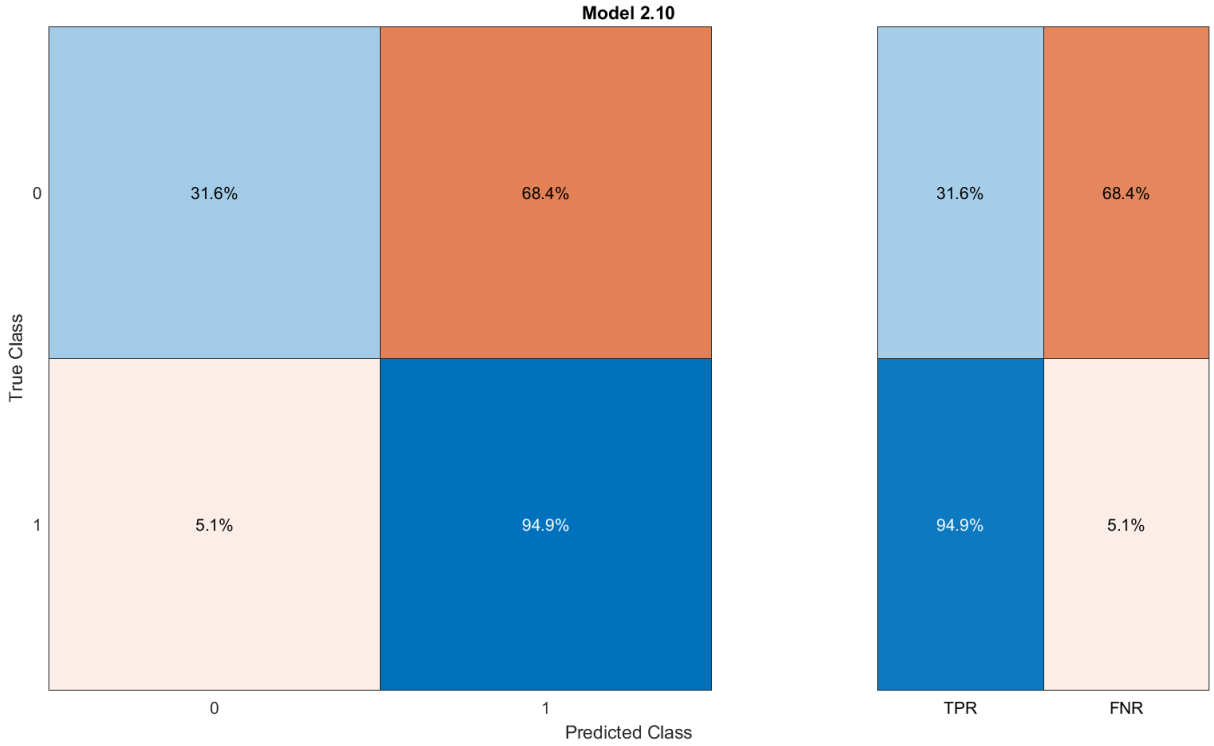
7.5.1 Oesophagitis

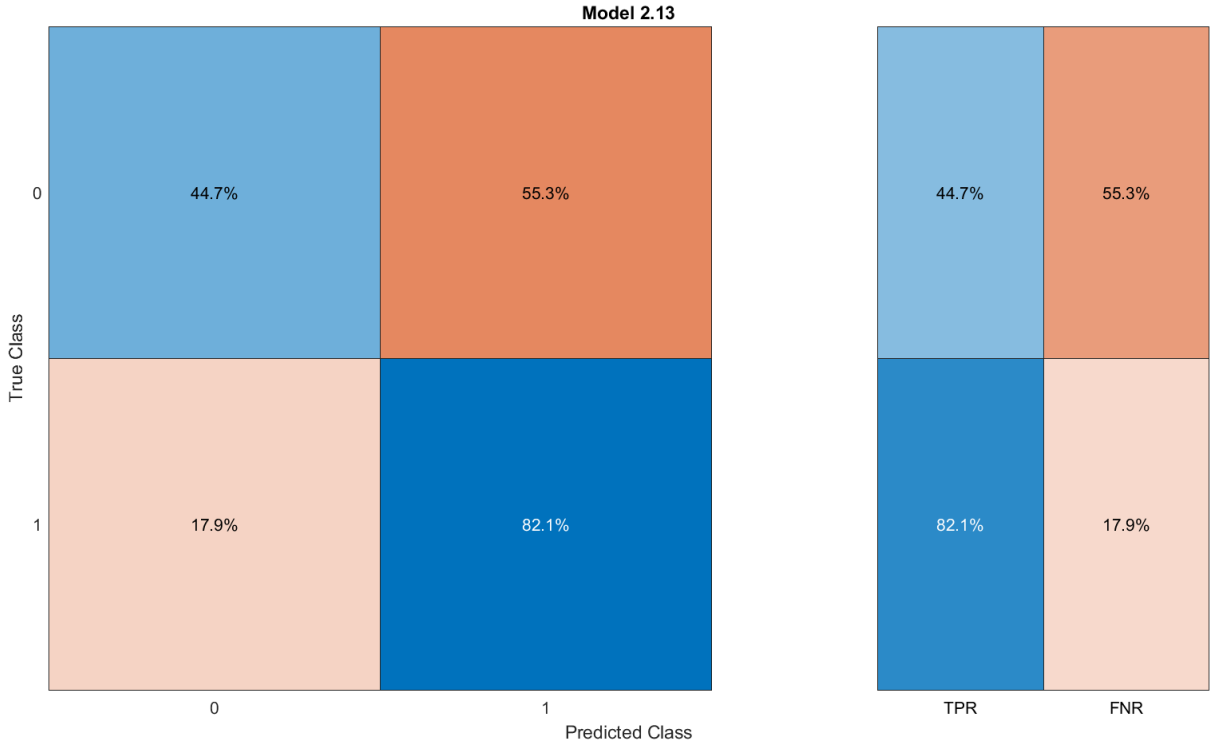
All models presented here use the V50Gy and patient sex as features. The classifier type and model numbers are as follows. Decision Trees (Model 2.3), Logistic Regression (2.4), Naïve Bayes (2.6), SVM (2.10), Ensemble (2.10) and Neural Network (2.16).

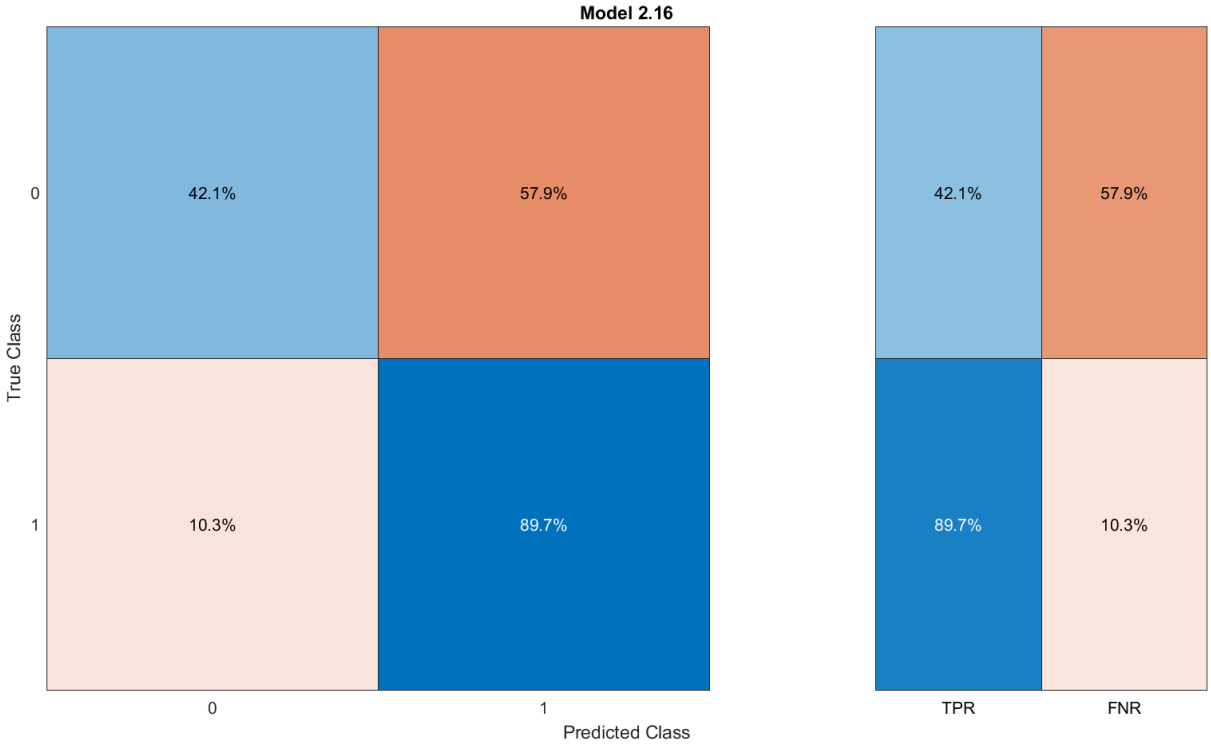



Model 2.4
36.8%
63.2%
15.4%
84.6%
True Class
0
1
0
1
Predicted Class
36.8%
63.2%
84.6%
15.4%
TPR
FNR


Model 2.6
31.6%
68.4%
6.4%
93.6%
True Class
0
1
0
1
Predicted Class
31.6%
68.4%
93.6%
6.4%
TPR
FNR

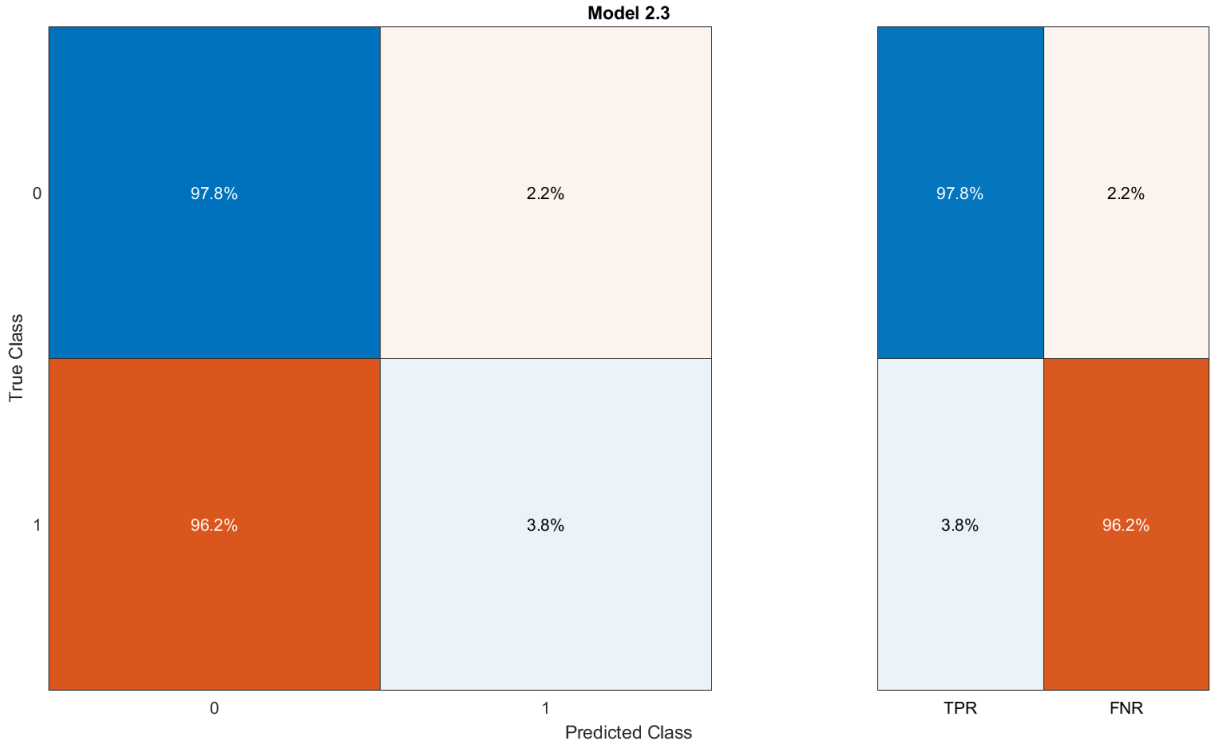

Model 2.10
0
1
True Class
31.6%
68.4%
5.1%
94.9%
0
1
Predicted Class
31.6%
68.4%
94.9%
5.1%
TPR
FNR

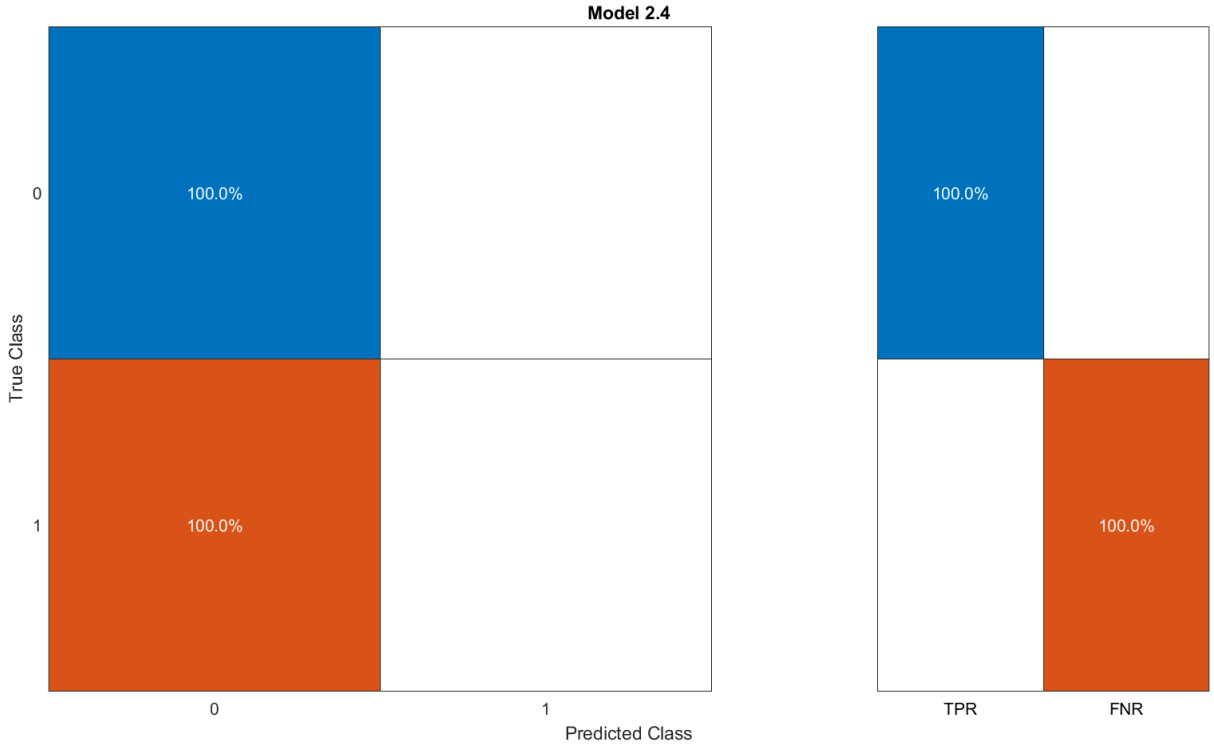

Model 2.13
0
1
True Class
44.7%
55.3%
17.9%
82.1%
0
1
Predicted Class
44.7%
55.3%
82.1%
17.9%
TPR
FNR

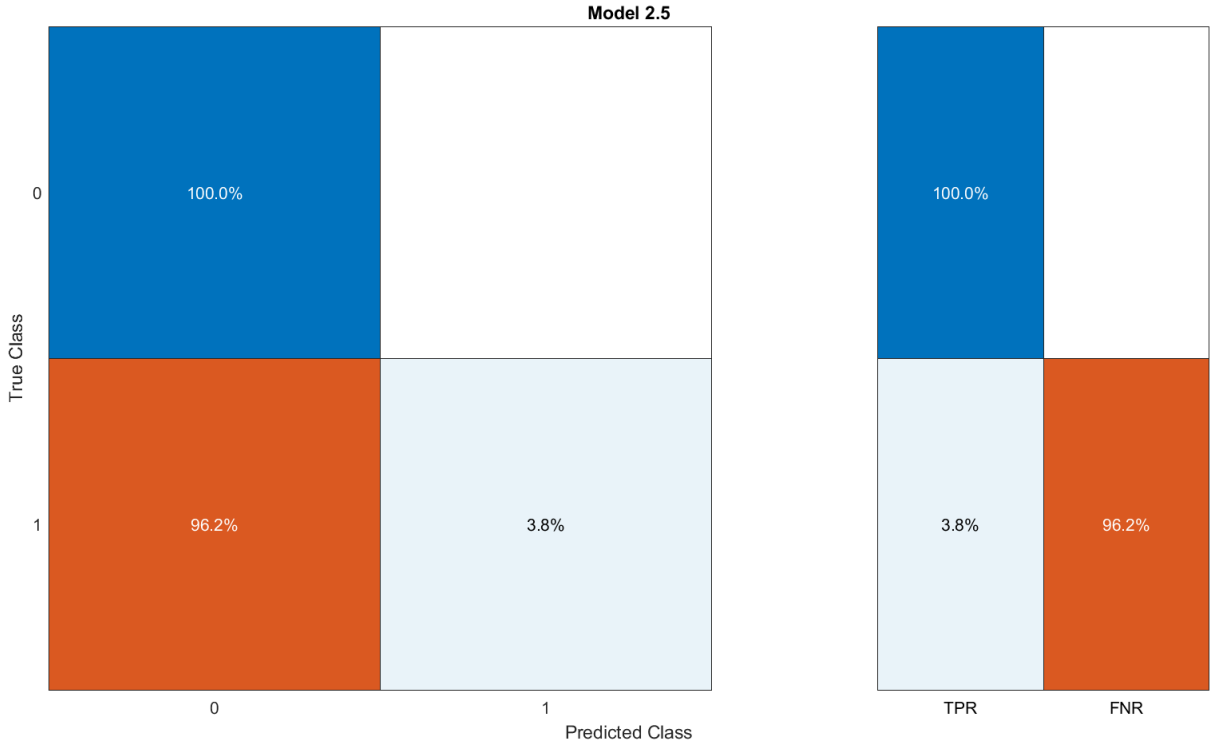


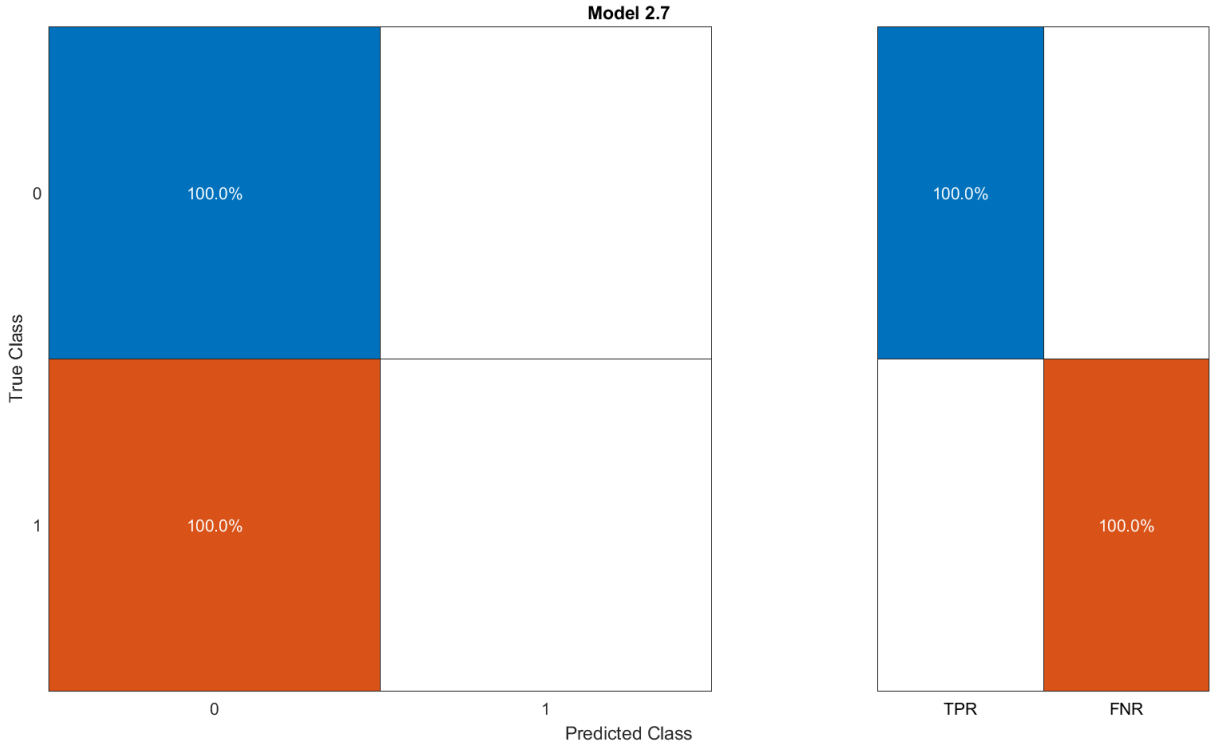
7.5.2 Pneumonitis

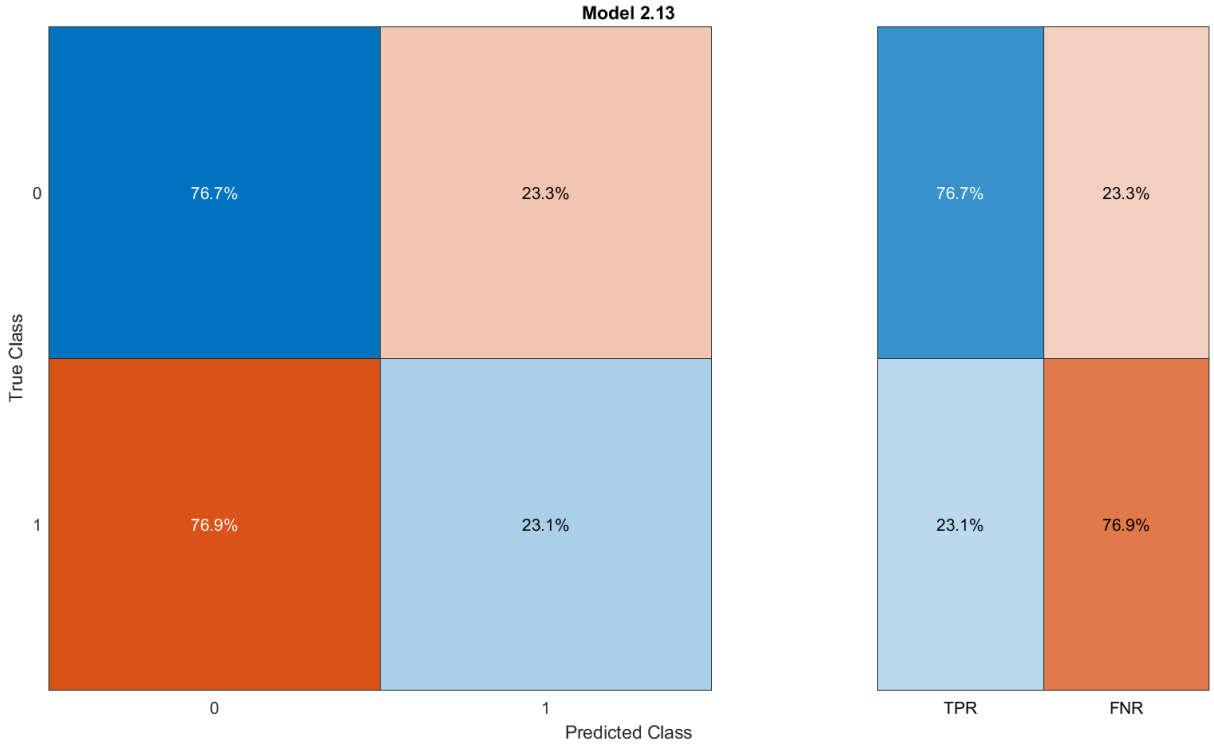
All models presented here use the EUD and patient sex as features. The classifier type and model numbers are as follows. Decision Trees (Model 2.3), Logistic Regression (2.4), Naïve Bayes (2.5), SVM (2.7), Ensemble (2.13) and Neural Network (2.16).

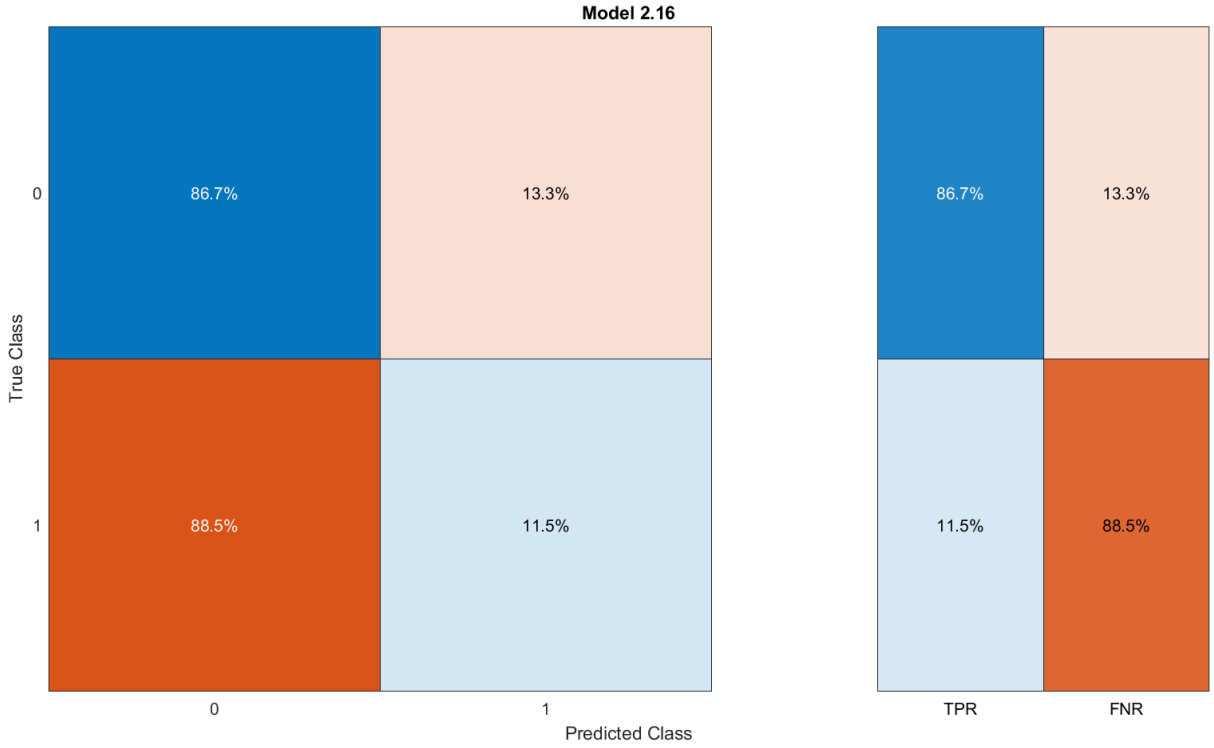

Model 2.3
0
1
True Class
97.8%
2.2%
96.2%
3.8%
0
1
Predicted Class
97.8%
2.2%
3.8%
96.2%
TPR
FNR


Model 2.4
0
1
True Class
100.0%
100.0%
0
1
Predicted Class
100.0%
100.0%
TPR
FNR


Model 2.5
0
1
True Class
100.0%
96.2%
3.8%
0
1
Predicted Class
100.0%
3.8%
96.2%
TPR
FNR


Model 2.7
0
1
True Class
100.0%
100.0%
0
1
Predicted Class
100.0%
100.0%
TPR
FNR


Model 2.13
0
1
True Class
76.7%
23.3%
76.9%
23.1%
0
1
Predicted Class
76.7%
23.3%
23.1%
76.9%
TPR
FNR

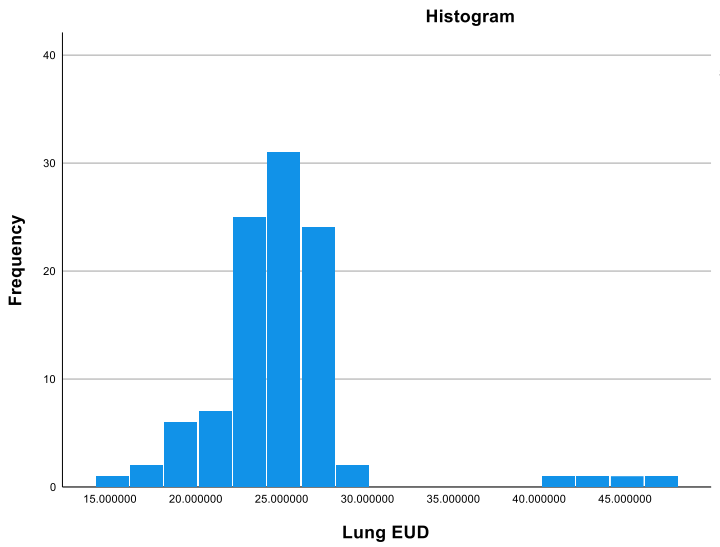

Model 2.16
0
1
True Class
86.7%
13.3%
88.5%
11.5%
0
1
Predicted Class
86.7%
13.3%
11.5%
88.5%
TPR
FNR

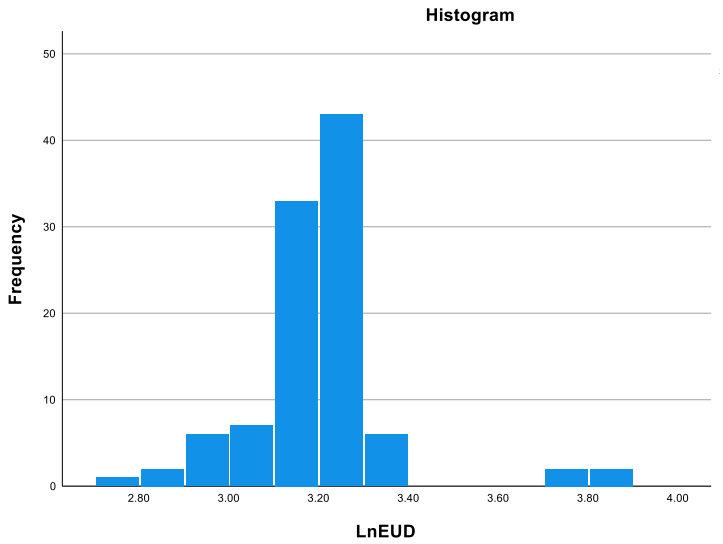
7.6 Appendix 6: Data Visualisation

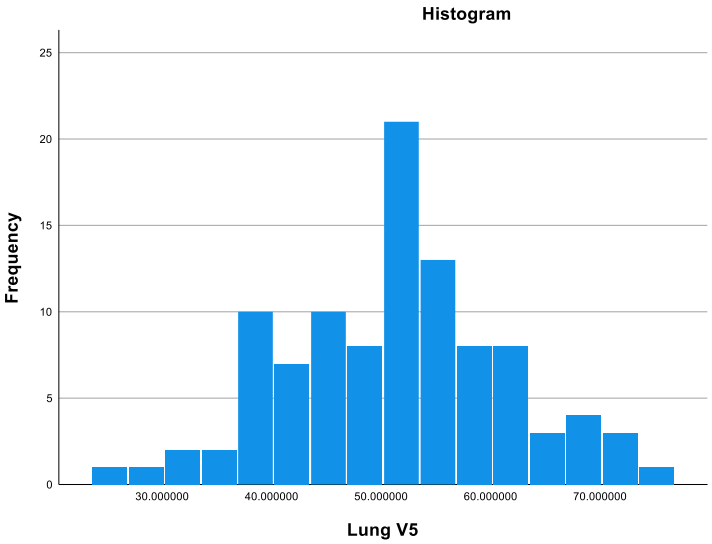
This section contains additional data plots to better visualise the data. Histograms and box plots have been plotted for the four lung and four oesophageal dose metrics used for modelling. For the histograms, the natural logarithm was also plotted to assess the impact of data transformation (variables prefixed with Ln). Finally Q-Q plots have been plotted for the dose metrics and measures of lung function to assess normality.

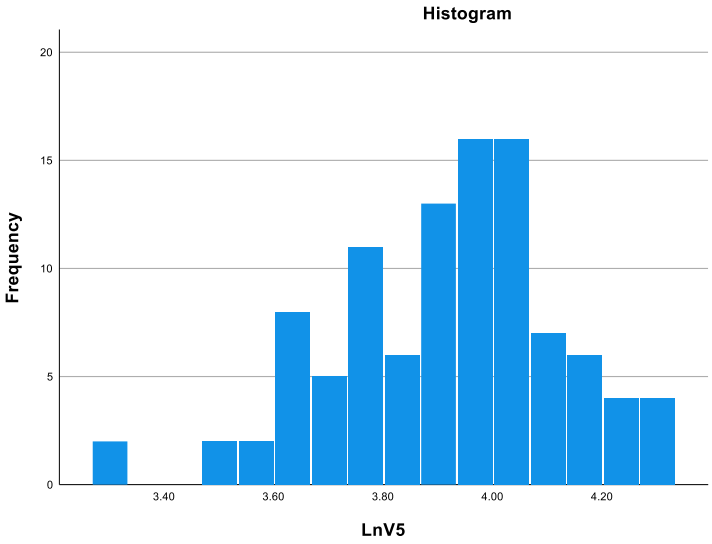
7.6.1 Lung

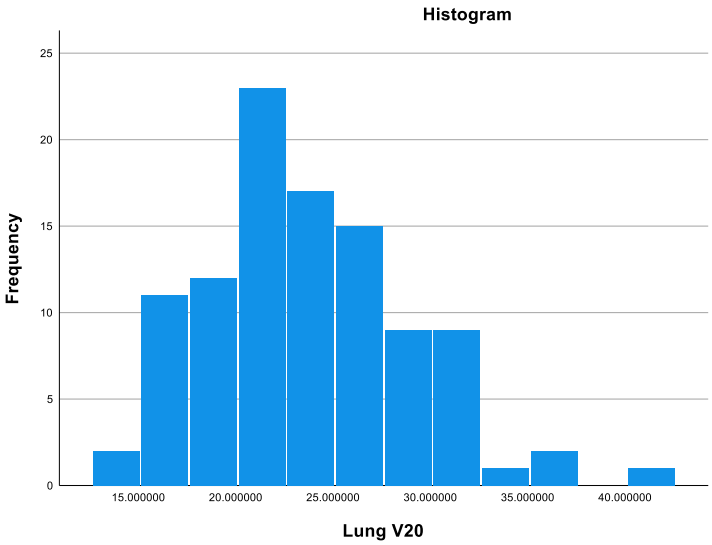
7.6.1.1 Histograms

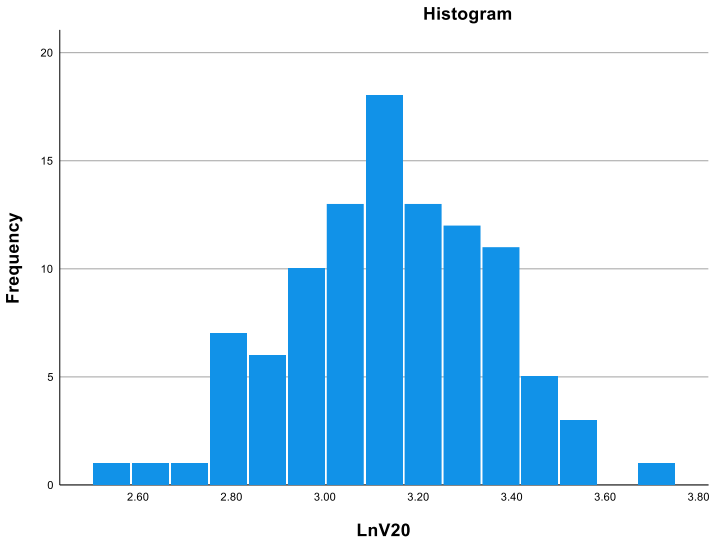


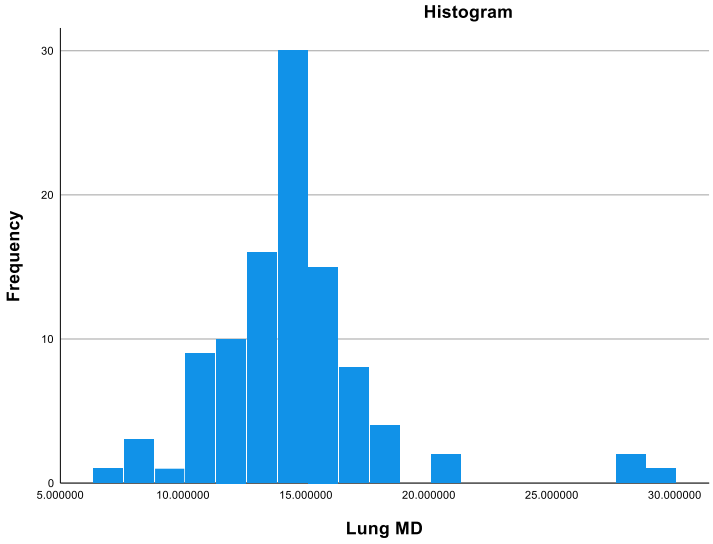


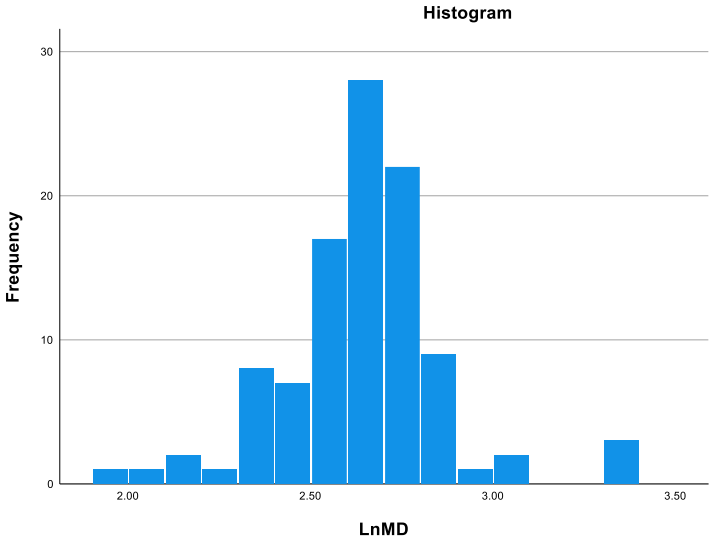

Histogram
Frequency
25
20
15
10
5
0
30.000000
40.000000
50.000000
60.000000
70.000000
Lung V5


Histogram
Frequency
20
15
10
5
0
3.40
3.60
3.80
4.00
4.20
LnV5

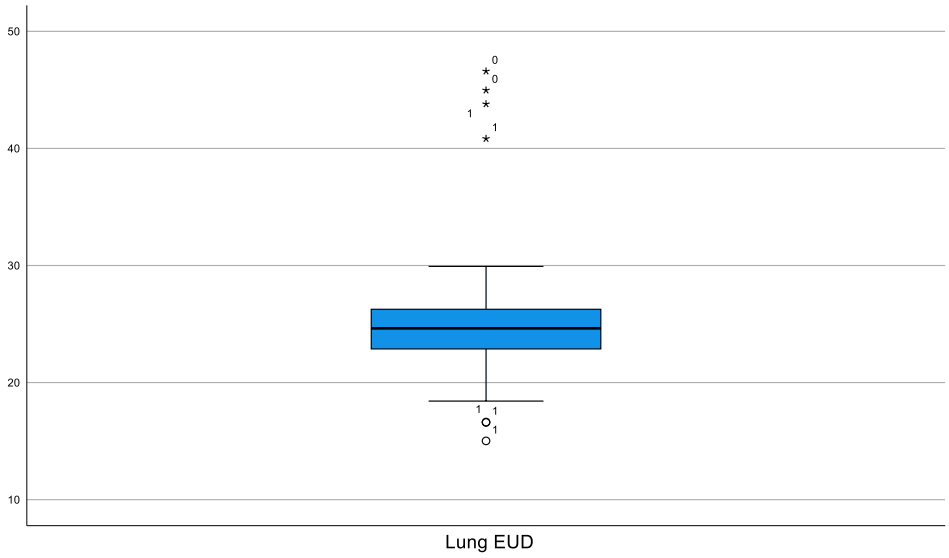

Histogram
Frequency
25
20
15
10
5
0
15.000000
20.000000
25.000000
30.000000
35.000000
40.000000
Lung V20

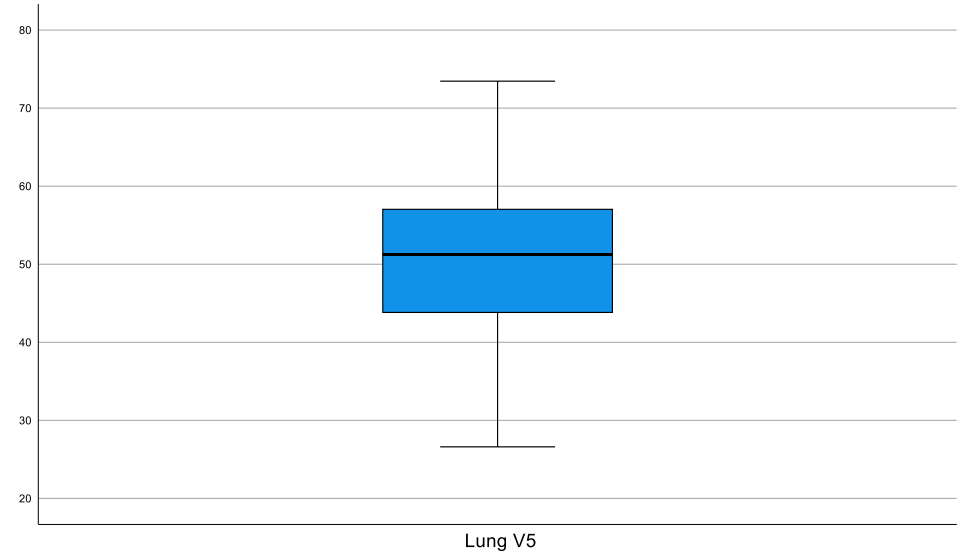

Histogram
Frequency
20
15
10
5
0
2.60
2.80
3.00
3.20
3.40
3.60
3.80
LnV20

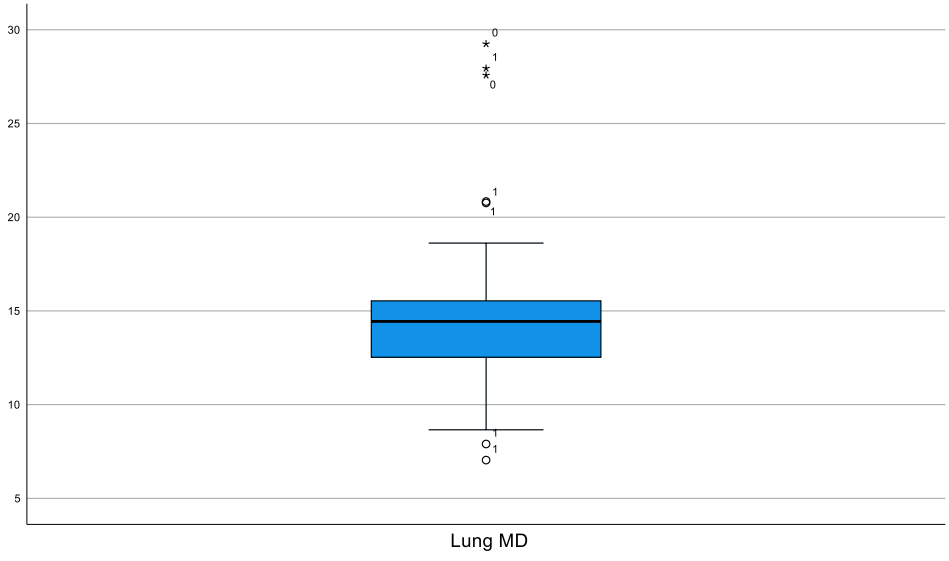

Histogram
Frequency
30
20
10
0
5.000000
10.000000
15.000000
20.000000
25.000000
30.000000
Lung MD

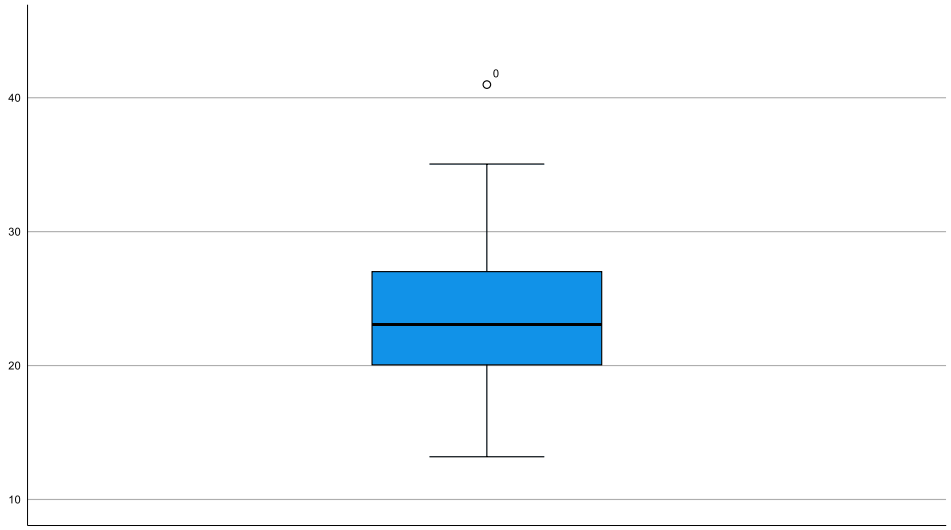

Histogram
Frequency
30
20
10
0
2.00
2.50
3.00
3.50
LnMD

7.6.1.2 Boxplots

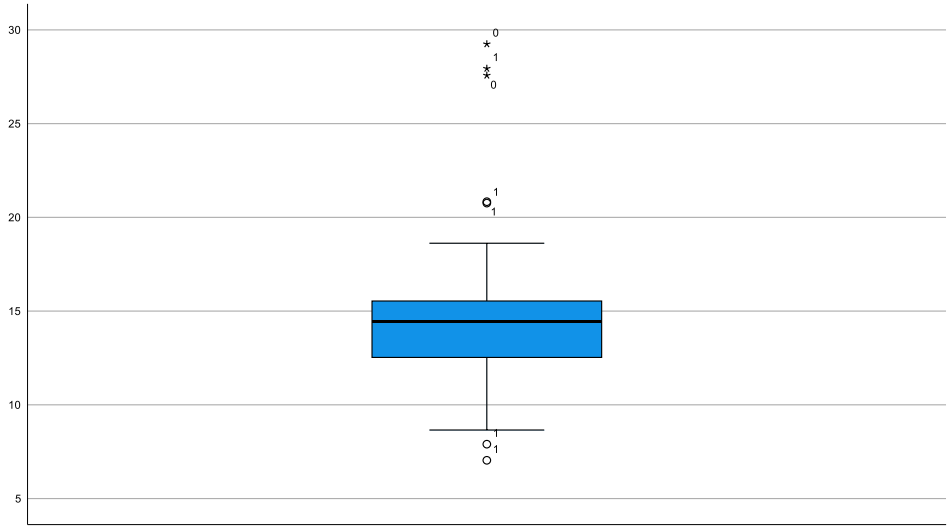






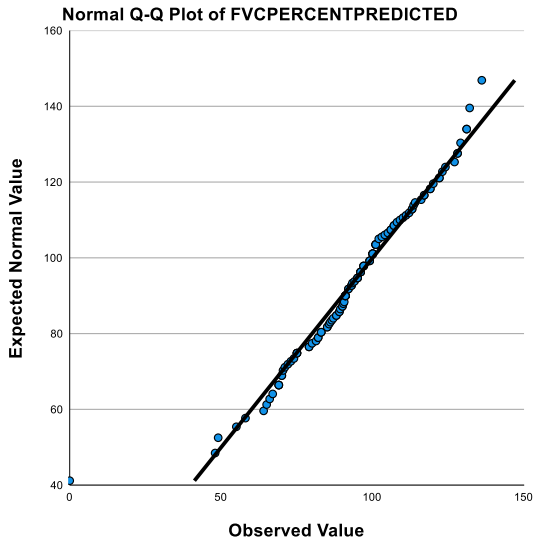


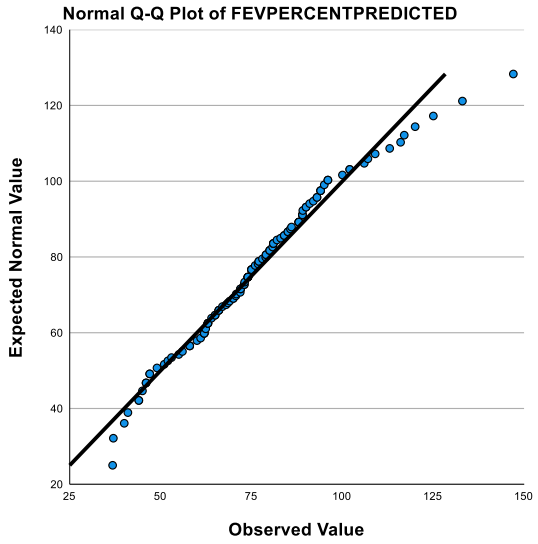
Lung V20

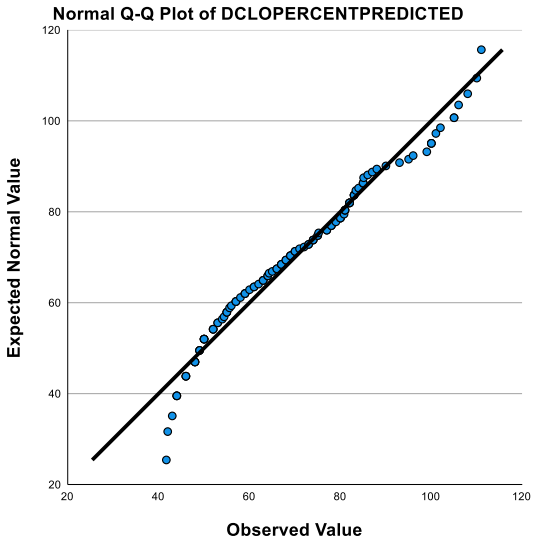


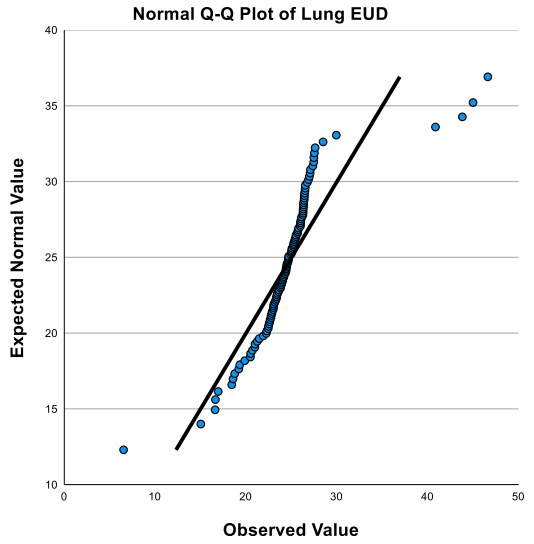
Lung MD

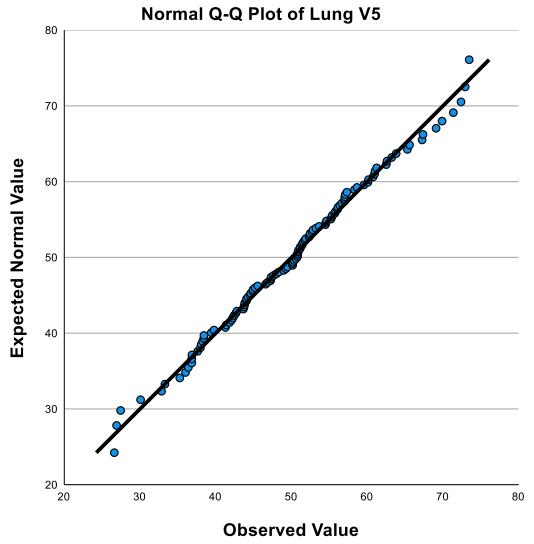
7.6.1.3 Q-Q plots

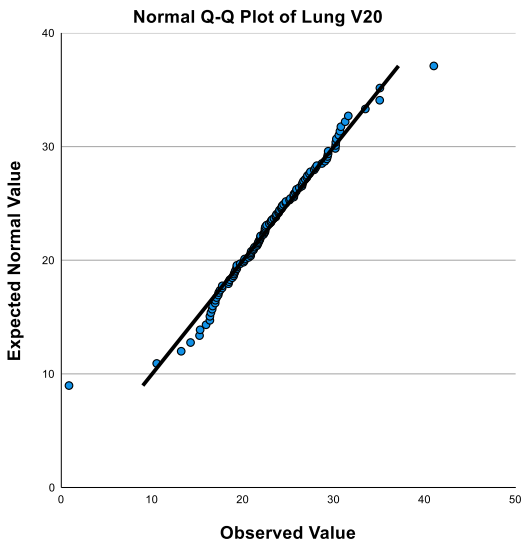


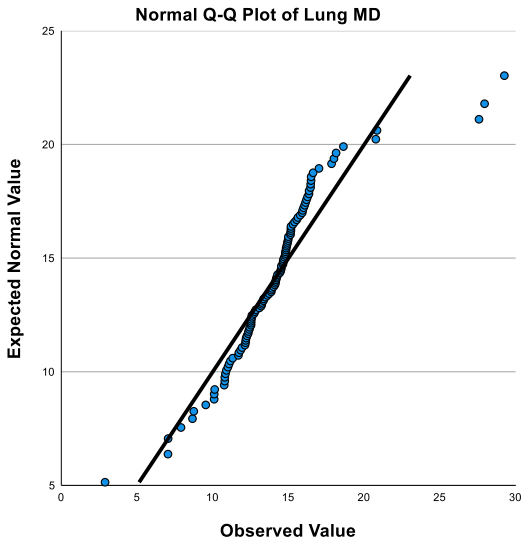



Normal Q-Q Plot of DCLOPERCENTPREDICTED
Expected Normal Value
Observed Value
120
100
80
60
40
20
20
40
60
80
100
120


Normal Q-Q Plot of Lung EUD
Expected Normal Value
Observed Value
40
35
30
25
20
15
10
0
10
20
30
40
50

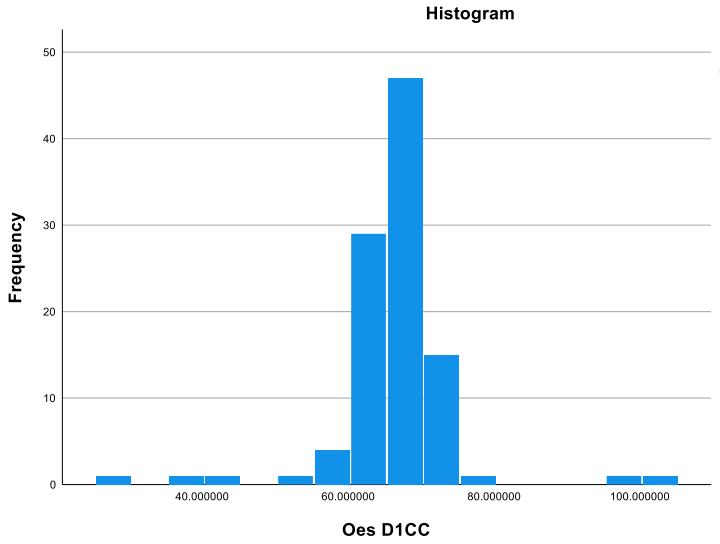

Normal Q-Q Plot of Lung V5
Expected Normal Value
Observed Value
80
70
60
50
40
30
20
20
30
40
50
60
70
80

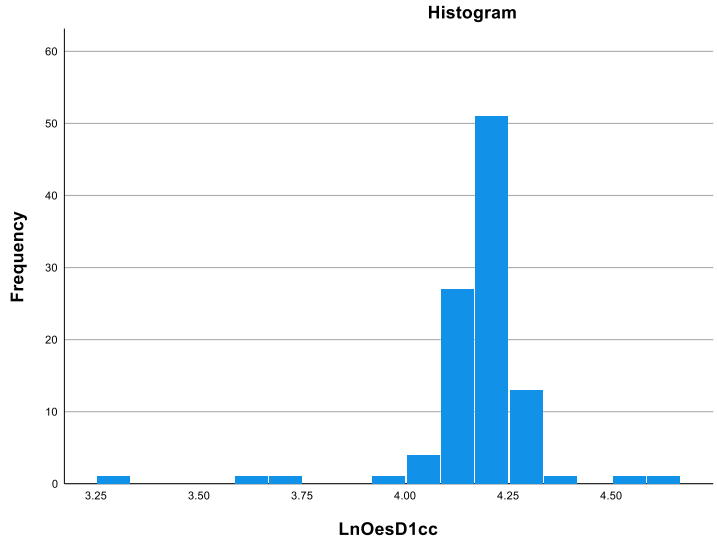

Normal Q-Q Plot of Lung V20
Expected Normal Value
Observed Value
0
10
20
30
40
50

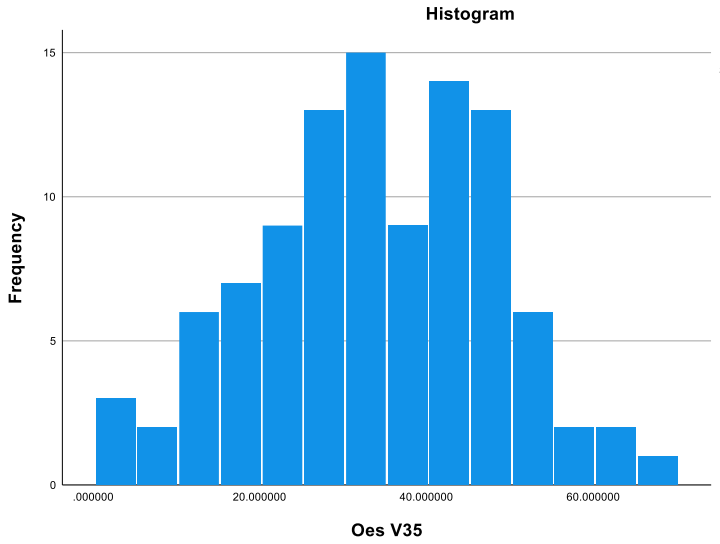

Normal Q-Q Plot of Lung MD
Expected Normal Value
Observed Value
5
10
15
20
25
30

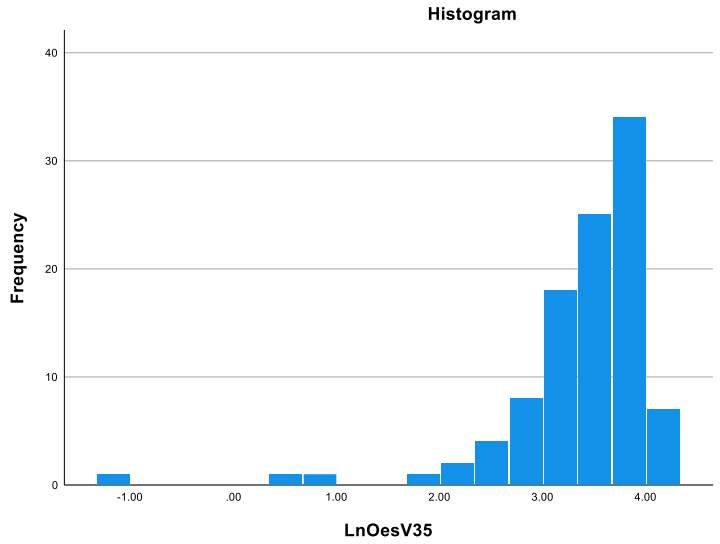
7.6.2 Oesophagus

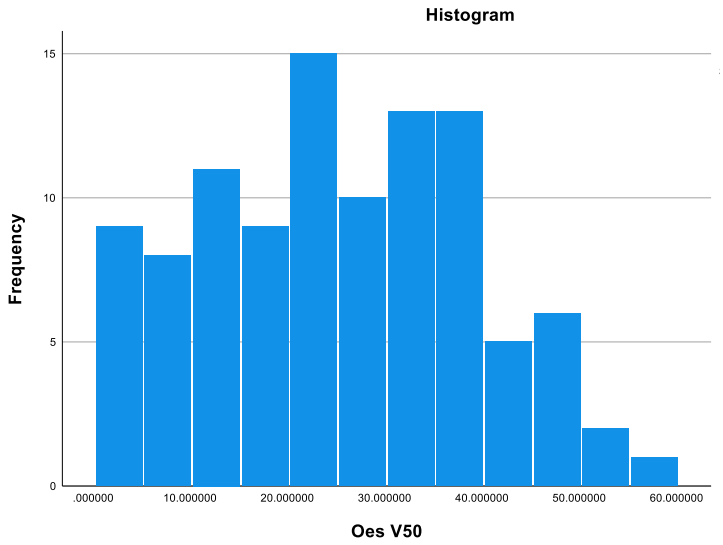
7.6.2.1 Histograms

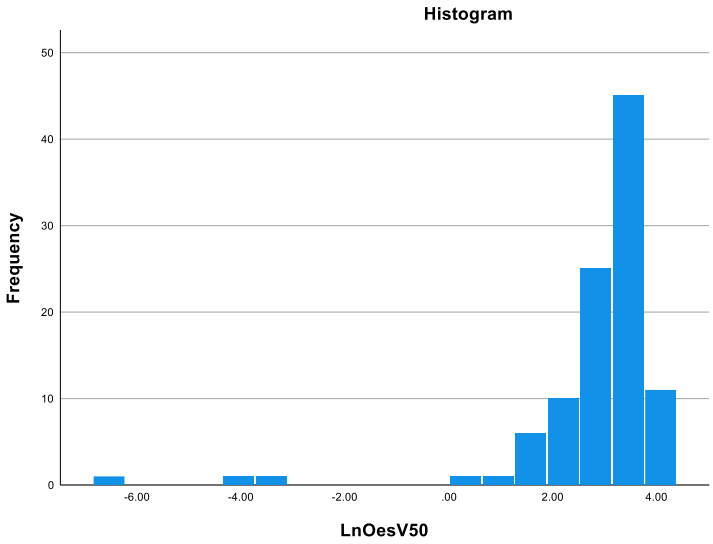


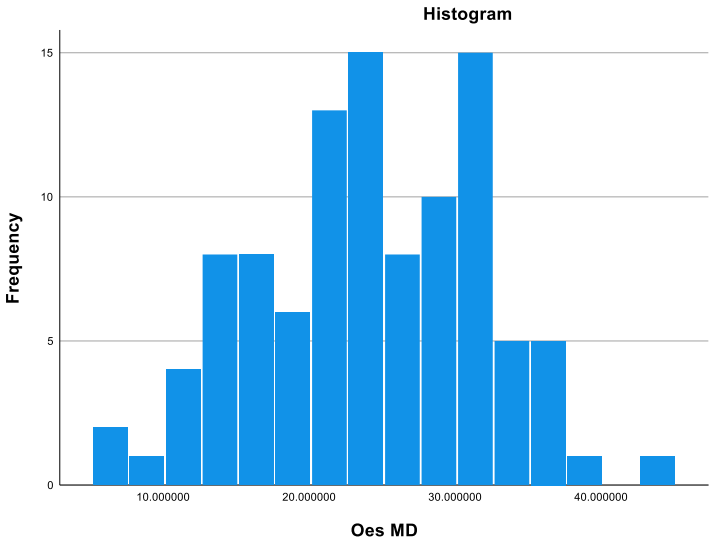


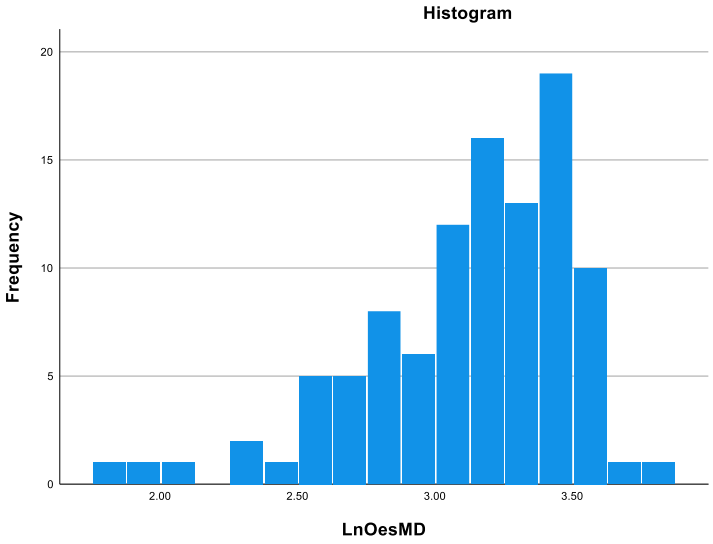

Histogram
Frequency
15
10
5
0
.000000
20.000000
40.000000
60.000000
Oes V35


Histogram
Frequency
40
30
20
10
0
-1.00
.00
1.00
2.00
3.00
4.00
LnOesV35

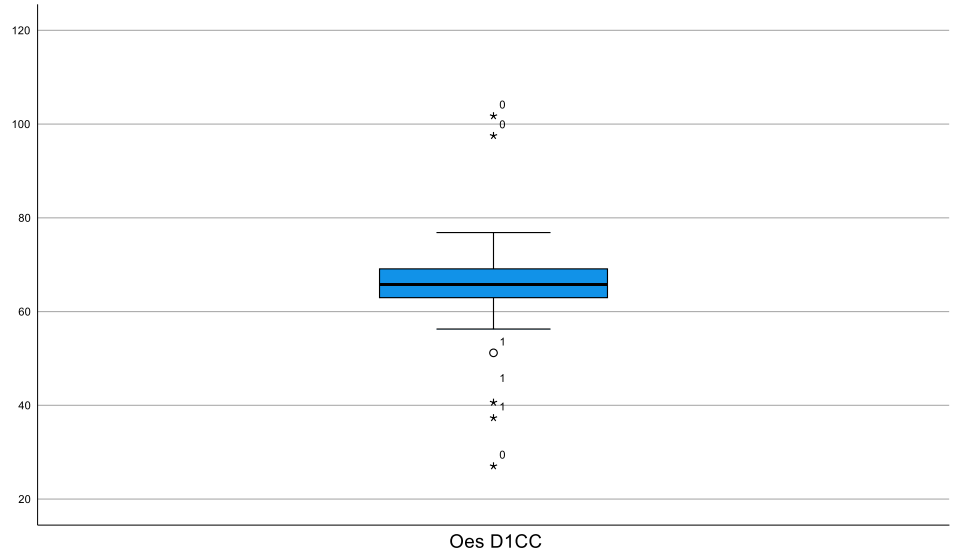

Histogram
Frequency
15
10
5
0
.000000
10.000000
20.000000
30.000000
40.000000
50.000000
60.000000
Oes V50

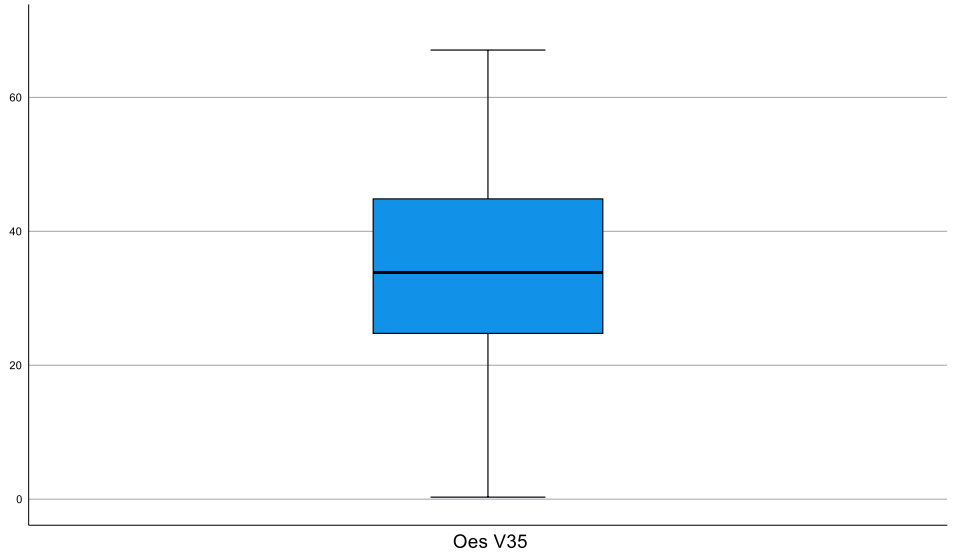

Histogram
Frequency
50
40
30
20
10
0
-6.00
-4.00
-2.00
.00
2.00
4.00
LnOesV50

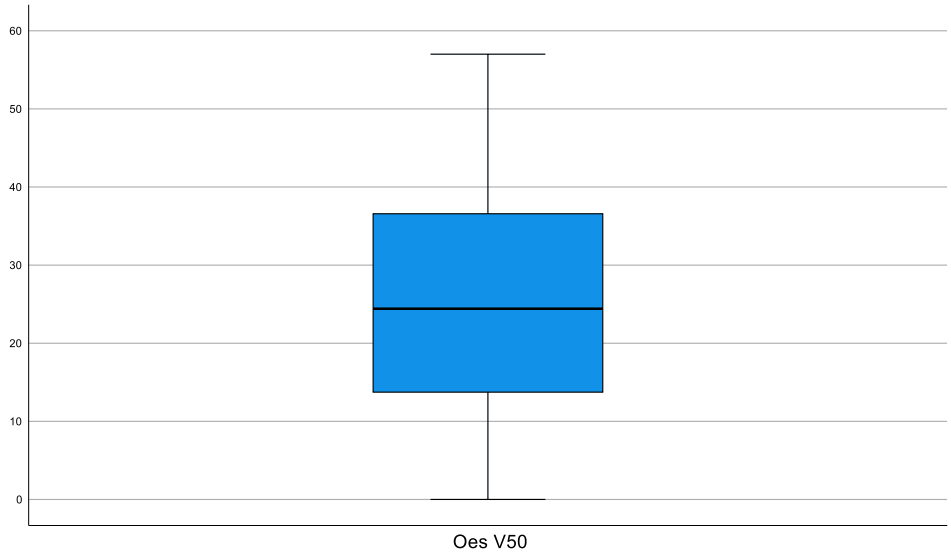

Histogram
Frequency
15
10
5
0
10.000000
20.000000
30.000000
40.000000
Oes MD

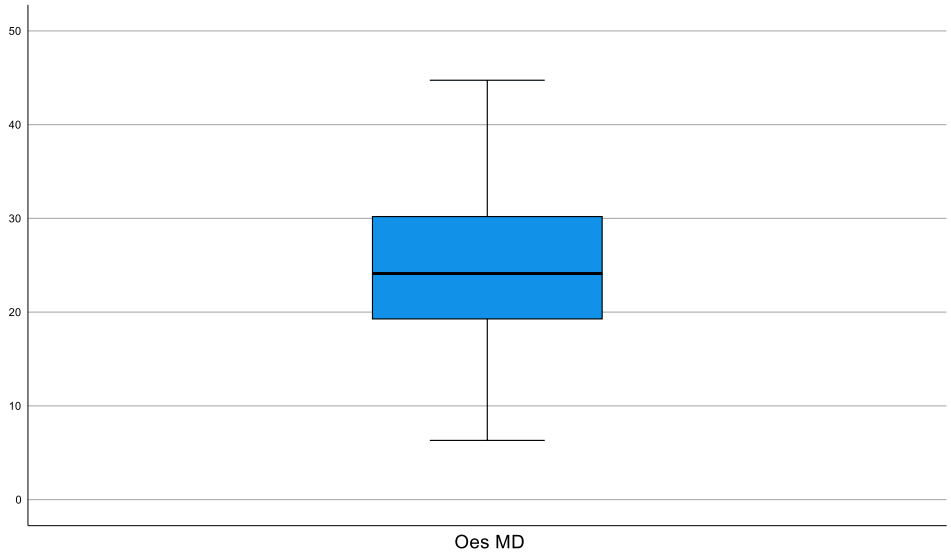

Histogram
Frequency
20
15
10
5
0
2.00
2.50
3.00
3.50
LnOesMD

7.6.2.2 Boxplots

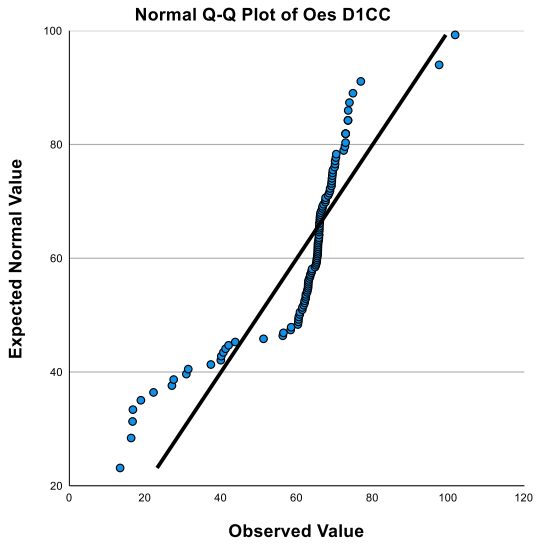


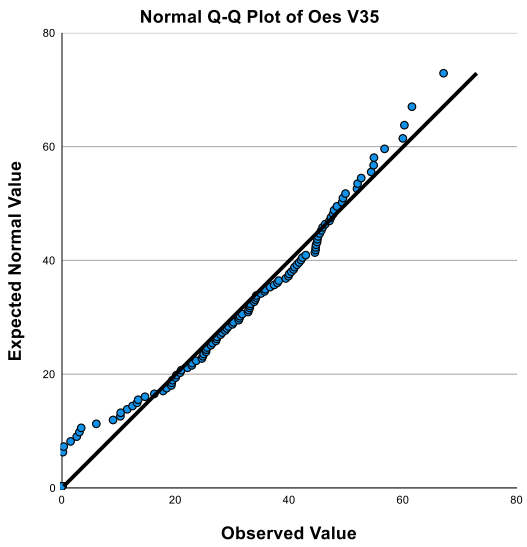


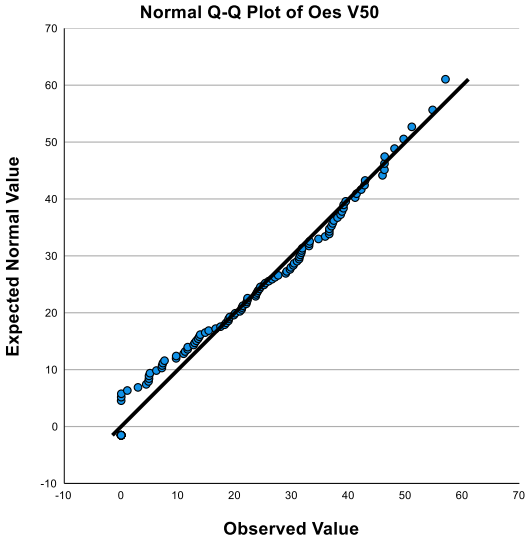


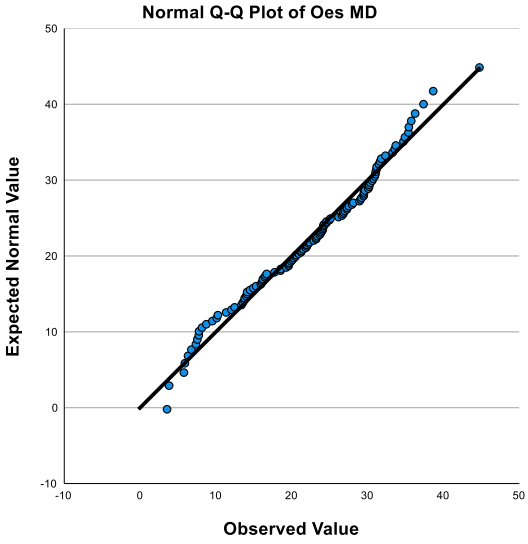


7.6.2.3 Q-Q Plots




Normal Q-Q Plot of Oes V35
Expected Normal Value
Observed Value
0
20
40
60
80


Normal Q-Q Plot of Oes V50
Expected Normal Value
Observed Value
-10
0
10
20
30
40
50
60
70


Normal Q-Q Plot of Oes MD
Expected Normal Value
Observed Value
-10
0
10
20
30
40
50