Application of Autoantibody Binding Curve Characteristics and Machine Learning Methods for Improving the Diagnostic Performance of an Early Detection Test for Lung Cancer

JARED ALLEN, MSci.

Thesis submitted to the University of Nottingham

For the degree of Doctor of Philosophy

JUNE 2023

Abstract

The EarlyCDT®-Lung test has been technically and clinically validated for the early detection of lung cancer with a sensitivity ~40% and a specificity of ~90% through measurement of a panel of seven serum autoantibodies.
The test generates curves of autoantibody binding to a titrated series of capture antigen concentrations thus providing patient-specific autoantibody profile titration curves. We postulated that the antibodies responsible for false positive results in healthy individuals exhibit different binding kinetics to specific autoantibodies present in cancer patients and that these differences may manifest themselves in the shape of the autoantibody-antigen titration curves.

The EarlyCDT®-Lung test result is currently a simple logic test combination of the results from the seven autoantibodies. The employment of machine learning models to combine the biomarker results, especially with the addition of a number of extra biomarker parameters, may allow improved clinical utility of the test through increased sensitivity and specificity.
A health economic analysis was undertaken to determine the current cost-effectiveness of the EarlyCDT®-Lung test for population screening for lung cancer compared to low-dose computed tomography, it showed that the current test performance was more cost-effective than LDCT screening at £37,679 per QALY, and quantified the performance needed to achieve cost-effectiveness at £30,000 per QALY was sensitivity of 39.8% at 99% specificity, 47.5% at 95% specificity, or 56.2% at 90% specificity respectively.

Serum autoantibodies from three case-control cohorts were measured on the EarlyCDT®-Lung test, as well as on an extended panel of autoantibodies. The titration binding curves returned by the test were analysed for signal magnitude, as well as curve characteristics including Slope, Intercept, Area Under Curve (AUC) and maximum slope obtained over the curve (SlopeMax). A range of unsupervised and supervised machine learning strategies for combining these biomarker results were explored, including principal components analysis, cluster analysis, logistic regression, decision tree analysis, naïve bayes, support vector machines, random forest, and extreme gradient boosting. The performance improvements of these optimised models was, however, modest and inconsistent across cohorts.

Finally, a simulated annealing based algorithm for multivariate panel optimisation was developed as an evolution of the Monte Carlo random search strategy previously used to establish panel cutoff thresholds. This algorithm was able to derive optimal panels that compared favourably to both the current commercial thresholds and to the best models derived by machine learning strategies.

Acknowledgements

I would like to sincerely thank my supervisors, Dr. Matthew Grainge and Dr. Caroline Chapman, who had no idea what they were signing up for when this project started, and who have both shown boundless patience, provided endless encouragement, and been invaluable sources of feedback and advice.

I would also like to thank Dr. Andrea Murray and Geoffrey Hamilton-Fairley, for inciting the analysis that inspired this work, and for giving me the opportunity to pursue this project, and Prof. John Robertson for inducting me into the field of autoantibodies for cancer diagnostics.

Huge thanks to all the Oncimmune staff that have played a part over the years, especially Dr. Celine Parsy-Kowalska, who has been a fantastic sounding board, confidante, and friend throughout this entire endeavour, Dr. Isabel Macdonald, who has steered the ship through the recent storms, Dr. Chris Welberry, my one time fellow student and lifelong fellow gastronome, Dr. Natalia Hudson, whose drive and love of science has been an inspiration, and Graham Healey, for imparting his statistical knowledge and years of experience.

I need to give my thanks to the 'Beauties', Jen, Adam, Natalie, Chris, Chris, Hayley and Leanna, whose friendship, countless gaming sessions, and weekend shenanigans may not have directly helped with the progress of this project, but without which this project would likely have ended prematurely.

To the Elder Isles Crew, Ant, Nikki, Jen, Ben, and Mike. Who are the very best of people and kept me sane by embracing my crazy. May our adventures never end.

To my Jitsu family, who gave me a physical outlet when I was mentally exhausted, and especially Mia for her endless energy and support.

To my sisters, Amber and Jade, I'm sure they helped somehow, not sure how, but they will be annoyed if I don't mention them.

To Heidi, my partner in crime, for your support, understanding, belief, patience, and love.

Finally I would like to dedicate this thesis to my parents, John and Heather Allen, for their lifelong support, encouragement, belief, and financial support. I couldn't have done any of this without them, and I am eternally grateful.

Contents

List of Figures

List of Tables

Glossary of terms and abbreviations

ACS	American Cancer Society
AFP	alpha-fetoprotein
ANN	Artificial neural network
AUC	Area Under the Curve
BART	Bayesian Additive Regression Trees
BIC	Bayesian Information Criterion
BMI	Body Mass Index
CA-125	cancer antigen 125
CAMs	cell-cell adhesion molecules
CART	Classification and Regression Trees
CEA	carcinoembryonic antigen
CEA	carcinoembryonic antigen
CLIA	Clinical Laboratory Improvement Amendments
COPD	Chronic obstructive pulmonary disease
CT	computed tomography
DES	Discrete Event Simulation
DKK1	Dickkopf-related protein 1
DNA	Deoxyribonucleic acid
ECLS	Early Detection of Cancer of the Lung Scotland
ELISA	Enzyme Linked Immunosorbent Assay.
FDA	US Food and Drug Administration
FN	False Negative
FP	False Positive
HRQoL	health related quality of life
ICBT	iterative combination of biomarkers and thresholds
ICER	incremental cost-effectiveness ratio
ID3	Iterative Dichotimiser 3
iNMB	incremental net monetary benefit
LASSO	Least absolute shrinkage and selection operator regression
LCDRAT	Lung Cancer Death Risk Assessment Tool
LCRAT	Lung Cancer Risk Assessment Tool
LDCT	low dose computed tomography
LLP	Liverpool Lung Project
NHIS	National Health Information Survey
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
NK	natural killer cells
NLST	National Lung Screening Trial
NSCLC	non-small cell lung cancer
NSE	neuron-specific enolase
NTLP	non-tumour lung pathologies
OD	Optical Density
PAS	population autoantibody study

PCA	Principal components analysis
PLCO	Prostate, Lung, Colon, Ovarian
PPV	Positive Predictive Value
ProGRP	pro–gastrin-releasing peptide
PSA	prostate specific antigen
QALY	quality-adjusted life year
ROC	receiver operating characteristic curve
SCC	squamous cell carcinoma–associated antigen
SCLC	small cell lung cancer
Sens / Sensitivity	True positive rate, the proportion of the case cohort that are correctly identified as positive for a diagnostic test.
Spec / Specificity	True negative rate, the proportion the control cohort that are correctly identified as negative for a diagnostic test.
STVR	Signal To Vol Ratio
SVM	Support vector machines
SW	Shapiro-Wilk
TAA	tumour-associated antigens
TGF	transforming growth factor
TN	True Negative
TNF	tumour necrosis factor
TP	True Positive
UKCTOCS	The United Kingdom Collaborative Trial of Ovarian Cancer Screening
VCOD	VOL Corrected Optical Density
VEGF	vascular endothelial growth factor
VOCs	Volatile organic compounds
XGBoost	Extreme gradient boosting

Glossary of Mathematical Symbols

$\prod x$	Product of all elements of x
$\sum x$	Summation of all elements of x
$x \in y$	x is an element of the set y
$x \subset y$	set x is a proper subset of set y
$\lceil x \rceil$	Ceiling of x (the smallest integer that is greater than or equal to x)

Chapter 1: Introduction

1.1 Global Cancer Burden

Globally, cancer mortality is a leading cause of death, and in 2020 there were an estimated 19.3 million new cases, and 10 million cancer deaths(1) (an increase from the 14.1 million cases and 8.2 million deaths at the commencement of this project in 2012(2)). This is predicted to continue to increase, despite the implementation of strategies designed to reduce cancer-related mortality, with estimates predicting up to 28.4 million new cases a year by 2040, due to a combination of factors including increased life expectancy, an aging population, population growth, and an increased adoption of cancer-associated lifestyle choices such as smoking, physical inactivity and “westernised” diets(1).

Advances in screening and treatments have improved the mortality of several cancers, such as breast cancer, which, while having the highest prevalence in women, shows mortality lower than lung, colorectal, stomach or liver cancer. Other cancers, such as prostate, show high incidence, but the tumours tend to be less aggressive and so less frequently result in mortality.

1.2 Cancer Burden in the UK

In 2020 in the United Kingdom, around 457,960 new cancers were diagnosed, and there were 179,648 cancer deaths(3). In 2014, direct costs of treatment and palliative care to the NHS were estimated to be around £6.7 billion a year to the NHS, representing 5% of the NHS budget, while the wider cost to the economy through loss of productivity and earnings was estimated to be in the region of an additional £7.6 billion a year(4).

1.3 Biology of Cancer

Cancer is a complex and diverse disease encompassing over 200 distinct disease types, with each potentially displaying multiple subtypes. It is a disease characterised by a loss of control of cellular growth resulting in an accelerated rate of cell proliferation and invasion of surrounding tissue. This loss of cellular control is brought about through genetic mutation in dividing cells, most commonly through DNA insertions, deletions or chromosomal translocations(5) which confer an advantage in the dividing cell that allows it to escape the normal control mechanisms which dictate cellular growth.

These changes are generally brought about through either gain-of-function mutations, which lead to the production of oncogenes, or loss-of-function mutations which result in the loss of tumour suppressor genes. Generally, a single mutation is not enough to cause cancer, and there must be a succession of genetic and epigenetic changes which gradually lead to an accumulation of cellular changes via a process analogous to Darwinian evolution, in which each successive change confers a competitive advantage in the mutated cells over the surrounding tissue, eventually leading to the progressive conversion of normal cells into a cancer(6).

For the cells to progress from preneoplastic to a malignant tumour growth, they must develop a set of hallmark traits initially described in 2000 by Hanahan and Weinberg(7). The initial six traits were;

- Self-sufficiency in growth signals - normal cells require mitogenic growth signals to stimulate them to move from a quiescent to a proliferative state, tumour cells circumvent this through various methods, including developing the ability to synthesise their own growth factors, upregulation of growth factor receptors, and alterations to the Ras signalling pathways that allow activation without upstream stimulation(8).

- Insensitivity to antigrowth signals - in normal tissue, anti-proliferative factors such as soluble, cell surface, and extracellular matrix bound growth inhibitors maintain cellular quiescence by either preventing cells from entering an active proliferative cycle, or inducing the cells to enter a postmitotic state. The most common way tumour cells seem to evade these antiproliferative signals is through disruption of the retinoblastoma protein signalling pathway in order to remain proliferative(9).
- Evasion of apoptosis - tumour cell populations expand by proliferating faster than the rate of cell attrition. Resistance towards apoptotic cell death allows the tumour cells to reduce the rate of attrition and allows accumulation of cells. This is most frequently accomplished through mutations involving the p53 suppressor gene, leading to inactivation of the p53 protein(10, 11).
- Limitless replicative potential - normal human cell types have been found to have the capacity for 60-70 doublings before shortening of telomeres causes them to enter senescence. Malignant tumour cells do not reach this limit, and instead have become immortalized, generally through upregulation of the telomerase enzyme, allowing them to continue replicating indefinitely(12).
- Sustained angiogenesis - in order to receive the oxygen and nutrients required for survival, normal cells are obligated to reside within 100µm

of a capillary blood vessel. Without the ability to stimulate new blood vessel growth, the neoplastic cells are unable to expand, therefore in order to progress to a malignant tumour, an incipient neoplasia must develop angiogenic ability, and it accomplishes this through the production or stimulation of angiogenic growth factors such as vascular endothelial growth factor (VEGF), angiogenin, transforming growth factor (TGF)-α, TGF-β, and tumour necrosis factor (TNF)-α, and down-regulation of angiogenic inhibitory factors such as angiostatin, endostatin, and interferon amongst others(13).

- Tissue invasion and metastasis - in the late stages of the development of a cancer, tumour cells develop the ability to escape from their site of origin, degrading the extracellular matrix and acquiring a more motile, invasive phenotype, allowing metastatic invasion and colonisation of adjacent tissues. These metastases are the predominant cause of cancer deaths, and the processes which contribute to this invasion are extremely complex, but generally include alterations to proteins involved in tethering of cells to their surroundings, such as cell-cell adhesion molecules (CAMs) and integrins. (14, 15)

1.4 Role of the Immune System

More recently the emerging importance of the role of the immune system in the development of cancer has been recognised as an additional hallmark(16, 17). The link between cancer and immune cells was first recognised over a century ago, and the concept of immunosurveillance, the elimination of developing cancers by the immune system, was proposed as early as 1957(18). However, this link was mostly disregarded until animal studies in the late 1990s and early 2000s, comparing tumour incidences in wild-type and immunodeficient mice upon the application of carcinogens and in tumour transplantation. These studies led to the recognition of a complex and varied role of the immune system during the development of cancer and resulted in the development of the immunoediting hypothesis of cancer(19, 20).

Immunoediting is now understood to be a dynamic process, comprised of three phases; elimination - the immune system regularly identifies and eliminates abnormal cells which have suffered genetic damage or developed mutations - equilibrium - mutations conferring poor immunogenicity or immunosuppressive qualities in preneoplastic cells evade elimination phase and are chronically maintained - and escape - further mutations confer advantages that allow the cancer to proliferate and progress to malignancy(20).

The contribution of the immune system in the development of a preneoplastic cell into a malignant cancer is highly complex and is influenced by the many factors, including the cellular type and mode of transformation of the original neoplastic cell, the anatomic location of the tumour, as well as the stromal response, cytokine production, and immunogenicity of the resulting cancer(19). In mammals, the immune system is comprised of two distinct components, the innate and the adaptive immune system, which communicate and coordinate in response to foreign pathogens. Both of these systems are implicated in the immune response to a cancer.

In the healthy immune system, the innate immune response is the first line of defence against a pathogenic insult, occurring within minutes to hours, it is a non-specific response and is enacted through leukocytes; including natural killer (NK) cells, mast cells, eosinophils, basophils; and phagocytic cells including macrophages, neutrophils, and dendritic cells. These cells rely on recognition of immunostimulants; conserved features of pathogens that are not present in the host, such as bacterial peptides and fungal cell wall molecules. The innate immune system cells constantly monitor their surrounding microenvironment and react quickly to the presence of these immunostimulant molecules, resulting in the activation of phagocytic cells and the production of an inflammatory response (21, 22).

In the tumour microenvironment, an abundance of infiltrating innate immune cells is generally associated with poor prognosis. This is because the innate immune system promotes an inflammatory microenvironment, generating free radicals which can cause DNA damage and additional mutation in the tumour cells, and promoting angiogenesis and tissue remodelling through the production of growth factors, cytokines, chemokines and matrix metalloproteinases. Finally, the innate immune response may promote cancer growth by suppressing the antitumour activity of the adaptive immune system(23).

In contrast to the innate immune system, the adaptive immune response is slow to respond but highly specific in its response, creating an immunological memory after initial pathogenic presentation by the dendritic cells of the innate immune system. In the case of infection by a foreign pathogen, the adaptive immune system is enacted through lymphocytes; T cells and B cells. When activated through antigen presentation, T cells differentiate to either cytotoxic T cells, which directly destroy cells presenting a specific antigen, or helper T cells which produce signal molecules that activate the bactericidal activity of macrophages, as well as secreting cytokines which the activate B cells. These activated B cells then produce antibodies which bind specifically to a foreign antigen, inactivating the pathogen by blocking their ability to bind to cell receptors, and marking the pathogen for destruction by phagocytic cells(21, 22).

In cancer, the presence of infiltrating lymphocytes in the tumour microenvironment is generally associated with a favourable prognosis, with the adaptive immune system inhibiting the growth of the tumour through cytotoxic T cell activity and cytokine-mediated lysis of tumour cells, however aspects of the adaptive immune system have also been shown to promote tumour growth, with regulatory T cells suppressing antitumour T cell activity, and humoral immune responses promoting a state of chronic inflammation which leads to further promotion of tumour growth by the innate immune system(23).

1.5 Lung Cancer

1.5.1 Natural History

Lung cancer is an uncontrolled growth of cells in the lungs. It is more common in smokers and ex-smokers, although environmental factors such as asbestos exposure can also increase the risk of developing lung cancer.

In the absence of treatment, lung cancer is almost universally fatal, and a systematic review of available data showed that without treatment, survival statistics for lung cancer are poor, with clinical stages I/II, III, and IV respectively having median survival times of 10, 5, and 3 months, one-year survival rates of 39%, 17%, and 9%, while only 2% of untreated early-stage lung cancer patients, and no late-stage patients survive to five years(24).

Estimates of tumour doubling times show a broad distribution, with estimates between 70 days(25) and 780 days(26), dependant on subtype and presentation method. This variation may be due in part to the heterogeneity of lung cancers, and may also in part due to the growth of tumours following a Gompertzian growth model(27), with a tumour doubling time that increases as the tumour size increases and outgrows the available resources.

Assuming an exponential growth model, a conservative estimate of doubling time of 158 days (based on radiography studies(28)), and a minimum size for imaging detection of 100,000 cells(29) gives an average growth time of around 7.2 years from first cell to clinically detectable cancer.

1.5.2 Subtypes

Lung cancer can be divided into two major categories, small cell lung cancer (SCLC), and non-small cell lung cancer (NSCLC). SCLC accounts for around 11% of the lung cancer cases in the UK and is highly aggressive, associated with rapid-onset symptoms and paraneoplastic syndromes, suggesting that SCLC is particularly immunogenic(30, 31). NSCLC accounts for 88% of UK lung cancer cases and can be further subdivided into three main pathological subtypes;

Adenocarcinoma – derived from lung secretory cells, this is the most common type of lung cancer in non-smokers, and is more common in women than men.

Squamous cell carcinoma – derived from flat cells which line the inside of the airways, often linked to a history of smoking.

Large cell carcinoma – can occur in any part of the lung and tends to be aggressive, growing and spreading quickly.

Several other subtypes of NSCLC exist, although they are far less common.

1.6 Detection of Cancer

Cancers which show the greatest mortality, such as lung cancer and hepatocellular carcinoma, often are not detected until they present symptomatically at late-stage disease, and while much research has focused on finding effective treatments for late-stage disease, it has been argued that early detection is a much more effective way of increasing cancer survival.

This is because prognostically, detection and treatment at an early stage invariably confers much greater survival rates, in lung cancer this is particularly evident with 1 year survival rates for lung cancer detected and treated at stage I being 87% compared to a one-year survival of only 19% when detected at stage IV (32). For this reason, early detection of cancer, before it is able to progress and metastasise, is an extremely important field of research, and the search for biomarkers to aid in the diagnosis and prognostic stratification of cancers has been at the forefront of cancer research since the development of the first screening test for cervical cancer by Herbert Traut and George Papanicolaou in 1943(33). While screening methods have shown success in several cancer types, such as breast, where mammography has been found to give an 18% reduction in mortality over 20 years(34), and colorectal cancer, in which faecal occult blood tests have been shown to give a 16% reduction in all-cause mortality(35), the majority of cancers still have no effective screening method, and are still not detected until late stage advanced disease. Cancer diagnosis rates in the US between 2005 and 2011 show several cancers where this is especially evident, with 57% of lung & bronchus cancers, 60% of ovarian cancers and 53% of pancreatic cancer diagnoses having distant metastases at time of diagnosis(36).

1.7 Early Detection of Lung Cancer

Lung cancer is currently the leading cause of cancer-related mortality(2), and is an attractive target for early detection, due to the high incidence, clearly defined high risk for smokers and ex-smokers, and the vastly improved prognosis from treatment at early stage compared to late stage with associated health economic benefit. While the initial interest in early detection of lung cancer was spurred by improvements to radiographic imaging technology, initial randomised controlled trials were not able to demonstrate reduced mortality from radiographic screening(37-40). Subsequent studies exploring chest X-ray combined with sputum cytology for early diagnosis of lung cancer were also unable to demonstrate a benefit for screening using these methods to detect abnormalities or lesions, which are referred to as nodules(41). More recently, the development of low dose helical computed tomography (LDCT) has allowed the detection of much smaller nodules (2-3mm) with much lower doses of radiation than chest X-ray. The improved sensitivity for detecting nodules using LDCT led to its assessment in a large randomised controlled trial(42), with results suggesting that LDCT performs with a high sensitivity (93.8%) but relatively low specificity (73.4%) for the detection of lung cancer, and that as a screening modality, LDCT was found to be able to reduce rates of lung cancer death (20%) and death from any cause (6.7%)(41, 43).

This impact of screening has resulted in the UK National Screening Committee updating their guidance for Lung Cancer in 2022 to now recommend lung cancer screening based on the benefits observed through the development of the targeted lung health check program in the NHS. This program recommends LDCT screening for smokers and ex-smokers based on their demographic risk factors, and more recently led to the launch of a national targeted lung cancer screening programme in the UK which has used mobile LDCT scanning equipment to increase the availability of LDCT screening, especially in rural or more deprived areas, with the aim of screening all high risk individuals ages between 55 and 75 with a history of smoking.

The current limitation of imaging technologies is the resolution, even using the most advanced imaging technologies and contrast agents, the detection size boundary currently remains limited to masses above 1mm^3, with most imaging technologies unable to visualise cancers smaller than 1cm^3. This equates to masses of around 3 million cells and 3 billion cells respectively(44), however at these sizes, malignant lesions are often difficult to distinguish from benign nodules, and the use of biomarkers may help to inform treatment decisions in these cases.

Blood-based biomarkers also have the potential to identify cancer risk in subjects who carry a malignancy that is still too small for detection by LDCT scanning, and also may introduce additional benefits with regards improve access to testing. As a blood-based biomarker test would need only a simple blood draw, and not specialised CT scanning equipment, this would allow for tests which can be available at a GP surgery at any time, which would be especially impactful in rural or deprived areas, giving a more accurate assessment of a subjects risk of cancer and need for further screening without the need to travel to a larger health centre, or wait upon the availability of mobile CT scanners. As smoking rates in the UK are around 4 times higher in the most deprived areas compared to the least deprived areas(45), this could increase access and uptake of screening testing in areas that need it most. Biomarker tests which are able to improve risk prediction prior to CT screening could also potentially reduce the number of LDCT scans needed, reducing any issues with waiting lists for access to scans, and potentially improving the health economics of a screening programme which integrates blood-based biomarker testing.

While disease biomarkers are widely used throughout medicine, very few have progressed through to be used for diagnosis and disease monitoring in cancer, and there is an outstanding need for a highly specific test for detecting lung cancer. In response to this, research has explored a vast range of potential biomarkers in addition to molecular imaging, such as volatile organic compounds, circulating micro-RNA, gene microarrays, serum antigens, and serum autoantibodies, the rationale and some recent examples for each potential biomarker is discussed here.

1.7.1 Volatile organic compounds (VOCs)

Exhaled breath is predominantly composed of nitrogen, oxygen, carbon dioxide, water, and inert gases. Volatile organic compounds (VOCs) are also present in the breath, in concentrations ranging from nmol/L–pmol/L. These compounds are either absorbed from the environment (exogenous) or generated within the body (endogenous). Endogenous volatile compounds are generated by biomolecular processes occurring within the body, therefore identification and measurement of endogenous VOCs in exhaled breath can reflect processes occurring within the body, and have been associated with conditions ranging from asthma, COPD, and liver disease, to transplant rejection and even schizophrenia(46).

The measurement of volatile organic compounds was first applied to cancer diagnostics in 1985 when analysis of 297 components of exhaled breath from 12 lung cancer patients and 17 controls reported a 93% accuracy for determining lung cancer using the presence of 3 compounds(47). To date 77 different VOCs have been identified over 50 studies as having discriminatory potential for lung cancer(48), with performance profiles ranging from 71.7% sensitivity and 66.7% specificity using 22 VOCs, to 54% sensitivity and 99% specificity using 2 VOCs(49). The main drawback of the use of VOCs is that most systems required to measure them are expensive and require expertise to use effectively, rendering them unsuitable for point of care diagnostics, and samples may easily be contaminated with ambient air.

1.7.2 Gene microarrays

DNA microarray-based gene expression profiling was developed in the mid-90s and utilises nucleic acid polymers, immobilised on a solid surface, to probe for complementary gene sequences in a sample(50). While the initial study looked at only 45 genes, the current precision with which the polymers can be coated to a plate now allows for the simultaneous measurement of the expression of thousands of genes from a single sample.

DNA microarray suffers from a variety of limitations that currently limit its suitability for early cancer screening or diagnostics, it requires relatively large amounts and high fractions of tumour cells in order to effectively determine an expression profile, and heterogeneity in cancer cells can cause different expression profiles, even from cells within the same tumour mass. Despite this, microarray has been successful at identifying subsets of genes expressed within a variety of cancer types such as ovarian, oral, melanoma, colorectal and prostate carcinomas(51, 52).It has also been of use identifying genes involved in cancer subtypes such as the gene PTK7 in lung adenocarcinoma(53), giving further information about the underlying cancer biology, and may have potential in predicting treatment sensitivity or prognostic outcome in diagnosed cancers.

1.7.3 Nucleic-Acid based Biomarkers

The discovery of free DNA in the blood plasma of cancer patients in the 1970s suggested that tumour cells release DNA into the bloodstream, and although sensitivity was relatively low, measurement of serum DNA may have potential for cancer detection(54). Subsequent advances in technology have resulted in the identification of several nucleic acid based biomarkers that have been associated with cancer, including cancer associated mutations to DNA, loss of heterozygosity and microsatellite instability, DNA methylation, and the presence of viral DNA(55).

Development of highly sensitive polymerase chain reaction (PCR) assays has allowed for detection and measurement of specific DNA mutations associated with cancer. While cancer heterogeneity initially prevented this from being useful on its own, advances in mutation specific ligation, mass spectroscopy, and the development of digital PCR techniques has allowed the identification of multiple mutations in a gene. The use of circulating tumour DNA is also limited due the detection threshold for most sequencing based methods still being too high to detect circulating tumour DNA in a large number of patients(56).

While sequencing technology continues to improve, most studies have focused on advanced stage cancers with relatively high concentrations of free DNA, to predict prognosis(57), recurrence(58), or response to treatment(59). There is still a shortage of studies which explore early-stage cancer and low concentrations of free DNA(60).

DNA methylation represents the most important epigenetic modification in mammals, it is able to selectively promote or silence gene expression without changes to the gene sequence, and plays a key role in maintaining chromosome stability. In healthy cells, DNA is normally hypermethylated, preventing the binding of transcription factors. During carcinogenesis, the normal methylation status is disrupted, leading to increased gene transcription, and a corresponding increase in the frequency of gene translocations, gene breaks, and gene mutations, eventually leading to cancer(61).

DNA methylation was recently shown to be able to detect cancer with extremely high specificity (>99%) across a range of cancer types, with sensitivities for all cancers types ranging from 18% in stage I, up to 93% in stage IV cancer(62), suggesting tumour burden is a potential driver of increased methylation levels.

1.7.4 Circulating micro-RNA (miRNA)

MicroRNAs (miRNAs) are small regulatory RNA molecules which moderate gene expression. They were first identified in 1993 and have subsequently been implicated in fundamental cellular processes such as development, differentiation, proliferation, apoptosis and stress responses(63). Dysregulation of miRNAs therefore result in a loss of gene expression regulation, which has been linked to the development of a vast number of different cancers including, but not limited to, prostate, pancreatic, colon, breast, liver, testicular, hepatocellular and lung cancers(64).

Discovery that circulating miRNAs were stable in human serum and plasma, and that the expression of tumour derived miRNAs was detectable in diffuse large B-cell lymphoma(65) as well as prostate(66), and lung and colon cancers(67) spurred an interest in the measurement of circulating miRNAs as biomarkers for the detection of cancer. Studies have subsequently illustrated the presence of miRNAs in prostate, ovarian, lung, breast, gastric, pancreatic, colorectal, and hepatocellular cancer(68). The use of miRNAs for diagnostics is still an emerging field and large case control studies are needed to fully elucidate their clinical potential.

1.7.5 Serum Protein Biomarkers

Serum protein biomarkers were first identified as having potential for cancer diagnosis in 1975, after elevated carcinoembryonic antigen (CEA) expression was observed in colorectal cancer cells(69) leading to the use of CEA serum measurements as a marker of disease progression. Research into these tumour associated proteins has identified a range of potential protein biomarkers in which normal expression has been altered, either through overexpression, aberrant expression, or expression of mutated forms of the protein. Some examples of these include overexpression of prostate specific antigen (PSA) in prostate cancer(70), cancer antigen 125 (CA-125) in ovarian cancer(71), and alpha-fetoprotein (AFP) in hepatocellular carcinoma(72). Aberrant expression of Livin/ML-IAP has been observed in lung cancer(73), aberrantly expressed MUC1 mucin has been observed in breast cancer(74), and aberrant expression of NY-ESO-1 was recorded in a range of cancers(75). Finally, expression of mutated forms of the p53 protein have been observed in gastrointestinal cancer(76).

The main limitation of these protein biomarkers is that they are invariably cancer associated, and not cancer specific, these proteins are expressed in normal body tissue, and may be expressed in benign conditions, which affects their specificity as a cancer marker. The overexpression of these proteins is also related to cancer burden, and therefore these proteins may not be measurable in early-stage disease, giving them poor sensitivity as screening tests, although they may still be useful as prognostic biomarkers. Finally, the heterogeneity of cancer results in protein overexpression occurring in only a subset of cancer patients, leading to poor overall sensitivity as diagnostic markers(55).

1.7.6 Serum Autoantibodies (AAbs)

Over the last 20 years, with the recognition of the role of the immune system in the development of cancers, has come the discovery of circulating antibodies to the mutated, aberrantly expressed, or overexpressed autologous tumour-associated antigens (TAAs) discussed previously(77). Serum autoantibodies (AAbs) have been shown to be present in a variety of cancers(78), and are the biomarkers currently used in the EarlyCDT®-Lung assay for the detection of lung cancer, which uses a panel of 7 serum AAb and detects lung cancer with a sensitivity ~40% and a specificity of ~90%(79).

Serum autoantibodies can be produced in response to limited exposure to TAAs, bringing about a measurable immune response while the quantities of the original TAA are too small to be detected with current assay techniques, representing an amplified ability to detect the TAA early in the disease state. A TAA-AAb response would also not be expected in normal body tissue, as can be the case with some TAAs, serum autoantibodies therefore represent a highly specific, amplified response to cancer, and should be detectable in disease at its earliest stages.

1.8 Combining Biomarkers

In the case of heterogeneous diseases such as cancers, any single biomarker test is likely to only detect a subset of the disease population, resulting in a low sensitivity and limited diagnostic value. Advancements in biomarker detection technologies have vastly increased the number of biomarkers available for assessment, and for early disease detection. Combinations of biomarkers have been found to provide improved discrimination over single marker tests. Various statistical methods have been developed to determine the best combinations of biomarkers from the high dimensional data which is increasingly becoming available, some examples of which are described here, with most studies utilising a combination of these approaches to come to a final panel or algorithm. These strategies are broadly separated into supervised methods in which models are trained to maximise accuracy of prediction to a class label, and unsupervised methods which examine patterns in data distributions in the absence of data labels.

1.8.1 Logic Rules

One of the simplest methods of combining biomarker tests is the application of cut-off rules which follow Boolean logic, in which a classification is applied to the data based on a set of simple logic based and-or criteria. The current EarlyCDT®-Lung test utilises a form of this, in logical terms:

For the seven AAb biomarker A, B, C, D, E, F, & G that constitute the EarlyCDT®-Lung panel with predetermined threshold cut-offs A_c, B_c, C_c, D_c, E_c, F_c, & G_c, and the corresponding results for a sample of interest A_s, B_s, C_s, D_s, E_s, F_s, G_s. A positive test result is given by:

$$A_s>A_c \text{ OR } B_s>B_c \text{ OR } C_s>C_c \text{ OR } D_s>D_c \text{ OR } E_s>E_c \text{ OR } F_s>F_c \text{ OR } G_s>G_c$$ (79)

Logic rules such as these were formally discussed and assessed for their application to improving diagnosis in prostate cancer, where a simple and-or combination of PSA with percent free PSA was able to improve the sensitivity from 33.6% to 34.3%, and increase specificity from 90.5% to 94.1%(80).

1.8.2 Monte Carlo Analysis

The Monte Carlo simulation method relies on repeated random sampling over a large number of repeats. For the optimisation of a panel for cancer detection, this involves setting discriminatory cut-offs at random for each diagnostic parameter, applying the cut-off rule, and then storing the resulting performance characteristics. This is repeated over a large number of times, with the resulting sensitivity and specificity characteristics being plotted out on a pseudo-ROC scatter plot. The final plot then shows optimal cut-off sets based on the desired performance. This method was used in 2011 by Boyle et al. in determining cut-offs against 7 autoantibodies in the validation of the EarlyCDT®-Lung test and resulted in a panel that performed reproducibly on two risk matched cohorts. The first group, comprised of 241 lung cancer and 241 healthy controls returned a sensitivity of 34% and a specificity of 91%, with a second cohort of 269 lung cancer and 269 healthy controls having sensitivity of 39% and specificity of 89%(79).

1.8.3 Logistic Regression

Logistic regression is a probabilistic modelling method, which attempts to find the best fitting model to predict an outcome based on a set of independent predictor variables(81). Several groups have utilised logistic regression modelling to develop a panel score for detection of lung cancer.

In 2006, Zhong et al. used fluorescent microarray to measure plasma tumour associated antibodies against a panel of 212 phage-expressed NSCLC-associated proteins in a cohort of 23 stage I lung cancer patients and 23 risk matched controls. Logistic regression was then used to select the five most predictive markers, resulting in a panel which performed with 91.3% sensitivity and 91.3% specificity. Validation of this panel on a set of 46 cancer and 57 healthy samples from the Mayo Clinic Lung Screening Trial showed a sensitivity of 82.6% and specificity of 87.5%(82).

Rom et al. used ELISA to measure serum autoantibody binding to 10 tumour associated antigens (TAAs) in 22 lung cancer patients and 36 healthy controls, as well as benign cohorts including high risk asbestos exposed patients with no nodules (n=35), with solid nodules (n=55), and with ground glass opacities (n=46). Using a logistic regression model which incorporated 6 of the 10 AAbs measured resulted in a logistic function which was 81% sensitive and 97% specific when identifying their cancer cohort(83).

1.8.4 Discriminant Analysis

Discriminant analysis is similar to regression, it attempts to find a linear combination of variables to explain a dependant variable, however it is limited to continuous independent variables which exhibit normal distribution. In lung cancer, Gao et al. used an antibody microarray to measure concentrations of 84 serum protein biomarkers. Of these 84, 7 were found to have significant discriminatory ability. The resulting discriminant function was based on 5 of these markers and gave 62.5% sensitivity and 100% specificity in a study involving 24 cancer patients, 24 healthy controls, and 32 patients with chronic obstructive pulmonary disease (COPD)(84).

The use of linear discriminant analysis in the construction of algorithms for four cancer datasets resulted in diagnostic accuracies of 94-100% in discriminating ovarian cancer from healthy controls, mesothelioma from adenocarcinoma, acute lymphocytic leukaemia from acute myeloid leukaemia, and metastatic from non-metastatic breast cancer(85).

1.8.5 Decision Trees

Decision tree analysis, such as Classification and Regression Tree (CART) modelling, attempts to map a classification based on series of logic observations about the independent variables. This results in a branched decision tree in which the branches represent groups of decisions, and the resulting leaves represent classification labels. Patz et al. utilised this technique, constructing a decision tree based on a panel of four serum protein biomarkers measured against a training set of 50 lung cancers and 50 healthy controls to obtain a CART classification algorithm that showed sensitivity of 89.3% and specificity of 84.7%. This was then validated in an independent set of 49 lung cancer and 48 matched controls with a resulting sensitivity of 77.8%, and specificity of 75.4%(86).

CART has also been explored for prediction of breast cancer recurrence risk category using pathological features and protein biomarkers to categorise samples as a cost-effective alternative to the Oncotype Dx gene expression test, and was found to correctly classify 69% of cases into their corresponding Oncotype Dx risk categories using just the expression of PR and Survivin, and the presence of nuclear pleomorphism(87).

1.8.6 Random Forest

Random forest analysis is an advancement of the classification tree analysis. It generates a large number of decision trees, and outputs a classification based on either an average or majority consensus of all the individual trees in order to reduce overfitting bias during training. This method has been used in lung cancer as a factor selection step prior to CART analysis to improve stability and reduce prediction errors. Borgia et al. applied a random forest algorithm to select a panel of 6 biomarkers from an initial pool of 15 for the prediction of lymph node metastases in lung cancer patients. These six biomarkers were then analysed using CART analysis to give a classification tree which predicted metastases with 88% sensitivity and 87% specificity in a cohort of 36 node positive and 71 node negative lung cancer(88).

Farlow et al. utilised the random forest algorithm to select an optimal panel of 6 analytes from an initial search of 21 tumour associated antigens, in their analysis of 117 NSCLC and 79 control subjects (31 osteoarthritis, 32 COPD & asthma patients, and 16 non-neoplastic nodule patients). The subsequent CART algorithm resulted in a sensitivity of 94.8% and specificity of 91.1% for detection of lung cancer(89).

1.8.7 Naive Bayes Classifiers

Naive Bayes is a probabilistic classification method, routinely used for tasks such as recognising spam e-mail, which examines all features as independent classifiers and predicts the probability of a sample belonging to a class based on the class probabilities of all the contributing features. Ostroff et al. utilised naive Bayes classification in their analysis of 813 proteins for detection of lung cancer. They examined a training set of 213 lung cancer cases and 772 healthy controls to select 44 potential biomarkers, and then, using a stepwise forward search algorithm and naive Bayes classification, identified a panel of 12 protein biomarkers which gave a sensitivity of 91% and a specificity of 84% in the training set, with a corresponding sensitivity of 89% and specificity of 83% in a verification set of 78 lung cancer cases and 263 healthy controls(90).

Models incorporating Bayes classifiers have also been found to be able to predict subtype in lung cancer to help guide treatment. A 2016 study by Pineda et al. used Bayes classification on DNA methylation data to determine between adenocarcinoma and squamous cell carcinoma with high classification (area under ROC >0.89) and allowed for construction of gene interaction networks(91).

1.8.8 Fuzzy Logic

Fuzzy logic employs many-valued logic where the truth value of variables can take any real number between 0 and 1, as opposed to Boolean 'true-false' logic. In this way the results for each biomarker can be examined in a more nuanced way for their ability to predict cancer. These methods were employed by Schneider et al. in an attempt to improve detection in a cohort of 175 lung cancer patients and 120 controls (17 healthy, 103 benign lung disease). The study examined 5 protein biomarkers, and use of fuzzy classification resulted in a final panel utilising 3 protein biomarkers which was able to predict disease progression with 92% sensitivity and 95% specificity(92).

1.8.9 Artificial Neural Networks

Artificial neural networks (ANNs) are a computational approach inspired by the way the brain solves problems, comprising a layer of neural units which, after training, 'learns' to identify patterns of associations to predict an output. They have been explored by several groups for their ability to identify lung cancer from biomarker data.

Wu et al. analysed 9 serum protein biomarkers, along with three metal ions. After back-propagation training of the ANN on a set of 35 lung cancer, 30 benign lung disease and 35 healthy normals, 6 markers were selected for inclusion in the ANN. The resulting network was able to predict lung cancer in a test group of 15 lung cancer, 10 benign and 15 healthy normal with a sensitivity of 100% and specificity of 100%(93).

O'Shea et al. used ANNs in the assessment of 5 sputum metabolite biomarkers, measured using flow-infusion electrospray-mass spectrometry in a group of 23 confirmed lung cancer, 11 symptomatic subjects, and 33 healthy controls. Training with leave-one-out cross validation resulted in an ANN which detected cancer with 96% sensitivity and 94% specificity(94).

1.8.10 Support Vector Machines

Support vector machines (SVMs) are supervised learning models which attempt to construct a hyperplane or set of hyperplanes which maximise the class separation of a training set. Leidinger et al. used SVM techniques in the analysis of 47 lung cancer sera, 26 non-tumour lung pathologies (NTLP) and 80 control sera, screened for antibodies against 1827 peptide clones, applying linear kernel SVMs, using 10 repetitions of 10-fold cross validation gave a final sensitivity of 97.6% and specificity of 97.0% against healthy controls, and specificity of 88.2% against both healthy and benign controls(95).

1.8.11 Cluster Analysis

Unlike the techniques described previously, cluster analysis is an unsupervised learning method, outcome data is not used in the training of the algorithm, rather the features of a dataset are explored, and objects are clustered based on a measure of their similarity. Au et al. performed a two-dimensional hierarchical cluster analysis on tissue microarray measurements from 18 protein biomarkers in a set of 284 lung cancer patients in an attempt to predict cancer subgroups. While the resulting clusters were not predictive of survival, the cluster groups did show good ability to predict cancer subtype, with one of the four groups containing 86% of the adenocarcinoma, and another containing 93% of the squamous cell carcinoma samples in the dataset(96).

1.8.12 Machine Learning Strategies for an Autoantibody Panel

The previously discussed machine learning strategies all have potential application in the development of models predictive for the incidence of lung cancer based on a panel of autoantibodies. Those of most interest in this particular study will be initially unsupervised methods, such as cluster analysis, as relationships and groupings that can be elucidated in the absence of disease class labels are likely to be reflective of differences in underlying biology and have the potential to identify relationships that supervised models may overlook. Of the supervised methods, decision trees represent an evolution of the logic rules model currently in commercial use for the test, allowing for a series of high specificity logic rules to be applied to stratify the subjects.

Random forest modelling then represents yet another evolution from decision trees, allowing for more highly fitted models through the training of large numbers of small decision trees, and may show improved performance due to being able to identify cancer immunotypes with smaller population in the dataset, and overcome some of the issues presented by the highly heterogeneous nature of cancer. In addition to these methods, logistic regression models will be considered due to their extensive use in existing diagnostic models, although the low sensitivity of individual autoantibody features suggests that this strategy may be better suited to biomarkers which show association with tumour burden, such as serum protein biomarkers or cell free DNA.

Support vector machines will also be considered under the assumption that the majority of samples will show only background signal, due to the high specificity and relatively low sensitivity of individual autoantibody biomarkers, therefore samples with no specific cancer response should group together in n-dimensional feature space, and outlier samples should represent those subjects with a cancer specific response, support vectors should then be able to define the border between the non-specific and specific signals.

Finally Naïve Bayes models will be explored due to its success in classification tasks, its relative simplicity, the ability to incorporate both continuous and discrete data into the models, its insensitivity to irrelevant features, and its treatment of the data as prior and posterior probabilities which may directly translate to incidence prediction.

1.9 Lung Cancer Demographic Risk Models

Previous lung cancer screening studies, such as the National Lung Screening Trial (NLST)(97), the NELSON study(98), and the PLCO cancer screening trial(99) have defined their screening eligibility through population demographic parameters, such as, in the case of the NLST study, ever-smokers aged 55-80 years with a smoking history of at least 30 pack-years and less than 15 years since cessation of smoking. As these large prospective studies have contributed greater amounts of data regarding the incidence of lung cancers in these high-risk groups, it has become possible to identify demographic risk predictors of lung cancer incidence and develop models which predict cancer incidence with a higher degree of accuracy than previously possible. As a result risk-models are now recommended to refer ever-smokers for screening rather than population demographic eligibility criteria(100). Two models have now been adopted by the NHS for LDCT screening eligibility, the PLCO$_{M2012}$ model(101) and the LLPv2 model(102), although there is evidence that two additional models, LCDRAT and LCRAT(103), may show greater accuracy of risk discrimination(104).

1.9.1 PLCO$_{M2012}$

The PLCO$_{M2012}$ model was derived from the PLCO study(99) based on a population of 36,286 ever-smokers and estimates a 6-year lung cancer risk using the demographic variables age, ethnic group, education, BMI, COPD, personal history of cancer, family history of lung cancer, and smoking history (status, intensity, duration, and quit time).

1.9.2 LLPv2

The LLPv2 model was developed using data from 579 lung cancer cases and 1157 age and sex-matched population-based controls(105), and has been validated using data from almost 76,000 individuals as part of the UKLS trial. The LLPv2 calculated 5-year risk is based on the variables age, gender, personal history of lung disease (pneumonia, emphysema, bronchitis, tuberculosis, or COPD), personal history of cancer, family history of lung cancer, asbestos exposure, and smoking duration.

1.9.3 LCRAT/LCDRAT

The Lung Cancer Risk Assessment Tool (LCRAT) and Lung Cancer Death Risk Assessment Tool (LCDRAT) models were trained using data from both the PLCO screening trial, and the NLST, as well as the US National Health Information Survey (NHIS). The LCRAT model was designed to predict 5-year lung cancer incidence, while the LCDRAT model was designed to predict 5-year lung cancer death. Both models included gender, race, education, emphysema, family history of lung cancer, smoking history (intensity and duration) and BMI as variables.

1.10 Multi-Cancer Early Detection Tests

Multi-cancer early detection (MCED) panels such as the Grail Galleri test have been explored recently for their ability to detect cancer using large DNA methylation panels, and initial large-scale trials have shown promising results, with the SYMPLIFY study(106) showing sensitivity of 66.3%, and specificity of 98.4% over all cancers. This study, however, explored only symptomatic subjects, which may lead to inflated diagnostic performance compared to use in a screening setting, and as a result showed higher sensitivity for later disease stages. While these tests appear to be a useful tool for identifying and potentially localising cancer that are causing generic symptoms, further work is needed to determine whether these tests perform with the same sensitivity in asymptomatic subjects in a screening setting, and would therefore be able to provide a stage-shift in diagnosis in over imaging that would lead to prognostic and health economic benefits.

1.11 EarlyCDT®-Lung

1.11.1 History of the EarlyCDT®-Lung test

The EarlyCDT®-Lung test is the first commercially validated test for the early detection of lung cancer. Developed in Nottingham, UK by Oncimmune Ltd, it is a quality-controlled, semi-automated, indirect enzyme-linked immunosorbent assay in which serum samples are reacted with a semi-log titration series of concentrations of tumour-associated antigens adsorbed to the surface of 96 well microtitre plates. The autoantibody assay method is described by Boyle et al, along with the initial validation in three groups of samples and showed that the test performed with estimated 40% sensitivity at 90% specificity(79). Post-validation by Lam et al. confirmed these findings in four additional sets of newly diagnosed lung cancer, including a set of small cell lung cancer samples, and a blinded matched cancer-control set from Vancouver, Canada. This study additionally demonstrated that the EarlyCDT®-Lung test is able to detect lung cancer at early stage as well as late stage, and that it detects both non-small cell lung cancers (NSCLC) as well as small cell lung cancers (SCLC)(107). Improvements to the test as described by Chapman et al. resulted in the expansion to the current commercial panel of 7 autoantibodies, which was confirmed to perform with a sensitivity of 41% and a specificity of 91% in an optimisation set of 235 lung cancer and 266 normal controls(108).

The EarlyCDT®-Lung test was first offered commercially in 2009, through private healthcare distributors in America, with the test being run centrally in a CLIA certified laboratory operated by Oncimmune LLC. The results of the first 1613 tests were reviewed by Jett et al., with all patients receiving 6 months follow up. The first 752 subjects were assessed with the 6AAb panel, detecting 12 out of 26 cancers (sensitivity 46%), and correctly classifying 599 out of 726 non-cancers with a negative EarlyCDT®-Lung result (specificity 83%). The subsequent 847 patients were assessed on the 7AAb panel, in which 13 out of the 35 subjects found to have lung cancer were detected with a positive EarlyCDT®-Lung test result (sensitivity 37%), and 742 of 812 non-cancer subjects correctly returning a negative EarlyCDT®-Lung result (specificity 91%).

An ongoing trial, the ECLS study, is currently assessing the EarlyCDT®-Lung test in a group of 12,000 high risk patients, enrolled through the Scottish NHS, to determine whether screening for lung cancer using EarlyCDT®-Lung will increase the number of patients being diagnosed with early-stage disease. Two year follow up data was published in 2019, and reported a sensitivity of 52.2% for early stage disease, and a sensitivity of 18.2% for late stage disease using the EarlyCDT®-Lung test, with a corresponding specificity of 90.3%, and achieved its primary end-point showing a significant decrease in presentation at late stage through the use of population screening with the EarlyCDT®-Lung test(109).

EarlyCDT®-Lung is currently used to assess risk of lung cancer incidence in high-risk individuals, evaluating the levels of seven autoantibodies using a logic rule-based assessment to give a result of either "High", "Moderate", or "No Significant Level". Clinical decision making based on this result then involves applying a diagnostic threshold model to the pre-test risk, according to the following formulae (where r = pre-test risk):

No Significant Level: No change to pre-test risk

Moderate: Post-test risk = $2.093r/(1+1.093r)$

High: Post-test risk = $13.421r/(1+12.421r)$

This increase in risk then instructs guidance to undergo LDCT screening or other enhanced surveillance(110).

1.11.2 EarlyCDT®-Lung Antigens

The seven autoantibodies that comprise the EarlyCDT®-Lung panel represent a combination of different classes of tumour associated antigen and are described in more detail here.

p53

p53 (Gene name TP53, Gene ID 7157) is a tumour suppressor protein which is ubiquitously expressed and has a role in co-ordinating cellular stress responses, regulating expression of genes to control arrest of the cell cycle, apoptosis, senescence, and DNA repair. It is the most frequently mutated gene in human cancer, with genome sequencing studies showing that approximately half of all cancers harbour a TP53 mutation, leading to inactivation or attenuation and loss of tumour suppression activity(111). Autoantibodies to mutated p53 have been reported in a range of cancers, with a systematic review of anti-p53 studies showing that across studies, anti-p53 autoantibodies show high specificity and sensitivity for oesophageal, head and neck, ovarian, colorectal, hepatocellular, bladder, lung, gastric, and breast cancer(112).

SOX2

SOX2 (Gene ID 6657) is a transcription factor with roles in regulation of gene expression and control of cell differentiation during embryonic development. Dysregulation and overexpression of SOX2 is associated with a variety of processes related to tumour progression including proliferation, epithelial-to-mesenchymal transition, migration, invasion, metastasis, colony formation, as well as resistance to apoptosis and cancer therapy(113). Autoantibodies to SOX2 have previously been identified as potential biomarkers in breast cancer, especially in early stage(114) and have specifically been linked to small-cell lung cancer(115).

CAGE

CAGE (Gene name DDX53, Gene ID 168400) is a cancer/testis antigen, initially identified as a cancer-associated gene expressed in hepatocellular carcinoma. While CAGE is widely expressed in various cancer tissues and cancer cell lines, it is rarely expressed in healthy tissue, and although its exact biological role is unknown, it has been determined to have a role related to the cell cycle and potentially plays a role in cell proliferation(116).

NY-ESO-1

NY-ESO-1 (Gene name CTAG1B, Gene ID 1485) is another cancer/testis antigen, normally expressed only by germ cells and placental cells during development, with some maintained expression in spermatogonia in adults. While it’s exact functions in healthy tissue are unknown, it is likely involved in cell cycle progression, growth, apoptosis, and differentiation. NY-ESO-1 has been reported in a wide range of cancer types, including bladder, oesophageal, hepatocellular, lung, ovarian, prostate and breast cancer, and autoantibodies to NY-ESO-1 have previously been explored for their biomarker potential in oesophageal, and colorectal cancer(117).

GBU 4-5

GBU 4-5 (Gene name TDRD12, Gene ID 91646) is an ATP-dependent RNA helicase normally expressed in the testis where it is primarily involved in spermatid development. The role of GBU 4-5 in cancer cells is still not fully understood, however recent work has suggested that GBU4-5 expression in cancer enables the proliferation of germ line tumour cells(118).

MAGE-A4

MAGE-A4 (Gene name MAGEA4, Gene ID 4103) is a cancer/testis antigen, which, along with other member of the MAGE-A family, is aberrantly expressed in multiple cancers, and its presence is associated with poor prognosis in hepatocellular carcinoma, lung squamous cell carcinoma, ovarian cancer, pancreatic ductal carcinoma, breast cancer, and thyroid carcinoma(119).

HuD

HuD (Gene name ELAVL4, Gene ID 1996) is a neuronal-specific RNA-binding protein whose roles involve regulation of neuronal development, survival, and plasticity through control of mRNA metabolism(120). In cancer, HuD was first recognised due to autoimmune responses to HuD in small-cell lung cancer resulting in paraneoplastic syndrome, and has previously been explored as a potential specific biomarker for small-cell lung cancer(121).

1.12 Diagnostic Pathway

The current recommended population for the EarlyCDT®-Lung assay corresponds to the US preventative services taskforce (USPTF) recommended population for lung cancer screening and includes adults aged 50-75 with at least a 20-pack year smoking, or a smoking history of less than 20 pack-years but immediate family history of lung cancer(122). The EarlyCDT®-Lung assay is a simple blood test therefore requires only access to a phlebotomist, rather than the specialised equipment required for LDCT screening, making it a much more accessible screening method than LDCT, and with lower up-front costs compared to LDCT screening. The test results would then be assessed by a clinician, with a positive test indicating an elevated risk of lung cancer incidence and triggering recommendation for LDCT screening to detect any nodules present, and enhanced surveillance with LDCT screening after 6 months intervals in the absence of a detectable nodule. After a nodule is identified, the patient would then receive the current standard of care, having had their cancer detected at a much earlier stage and therefore with vastly improved prognosis.

As a population screening test for a disease with a relatively low population prevalence, EarlyCDT®-Lung is designed to have high specificity, in order to minimise the number of false-positive results which may cause anxiety and unnecessary follow-up procedures to patients.

Proposed improvements to this diagnostic pathway involves an initial lung cancer risk assessment using one of the previously described demographic risk models, and screening high risk individuals. The risk score that should trigger the screening test was explored through health economic assessment of screening strategies, in a similar manner to that recently completed for LDCT screening for lung cancer(30). As the USPTF recommendation is annual repeat screening, and current UK national screening committee recommendations specify repeat screens every 2 years in high-risk individuals, the addition of the EarlyCDT®-Lung assay to these pathways would also allow repeat autoantibody measurement and appreciation of longitudinal changes in autoantibody biomarkers that may allow for refined algorithms including personalised baselines and more accurate risk prediction.

1.13 Discussion

A large number of biomarkers of various types have been explored for the detection of lung cancer. The first commercially available test validated for lung cancer, the EarlyCDT®-Lung assay measures the serum concentrations of a panel of autoantibodies and is able to detect early-stage lung cancer with ~40% sensitivity and ~90% specificity. While much research has been done exploring the presence of autoantibodies in cancer patients, the focus for their potential use as biomarkers has been on quantity. Antibodies bind to antigens with varying affinity, and this affinity increases over the lifetime of the immune response through affinity maturation. The binding affinity of an antibody biomarker could therefore give additional useful information with regards detecting a specific cancer response and the nature of that response. Although the binding affinity of the autoantibodies cannot be calculated directly using the current EarlyCDT®-Lung assay, the shape of the autoantibody-antigen titration curves generated by the assay may reflect the underlying binding kinetics, and represent additional biomarker parameters which may allow for improved cancer normal discrimination, with a preliminary study into the potential of these binding curve characteristics showing the ability to increase the specificity of the EarlyCDT®-Lung test.

The heterogeneity of cancer as a disease results in single biomarkers generally showing limited sensitivity for the detection of lung cancer. To obtain clinically useful sensitivities, the EarlyCDT®-Lung test uses a panel of biomarkers, which contribute to a panel result with much greater sensitivity than any single biomarker. The exploitation of multiple biomarkers as a diagnostic panel has been utilised by many groups, and a diverse range of techniques for combining the individual marker results into a panel score have been developed. The method currently used in the EarlyCDT®-Lung test is a simple logic rule whereby a positive result in one or more of the constituent biomarkers returns a positive result for the panel. The use of more sophisticated techniques for combining the biomarkers, including the additional binding curve characteristic biomarkers, has the potential to improve the sensitivity and specificity of the test in early detection of lung cancer, and an analysis of these various techniques and the improvements that they may bring about could lead to an improved clinical utility for the test.

When determining clinical improvement brought about by early detection, it is important to recognise and adjust for the potential impact of lead-time bias and length bias when quantifying the benefits of early detection. Lead-time bias coming about where disease is detected earlier, but survival is not affected, leading to an increase in apparent time with disease, and length bias whereby early detection results in the diagnosis of a greater number of slower progressing milder disease cases.

1.14 Hypothesis

Cancer cells express altered proteins which may elicit a non-self autoimmune reaction resulting in an amplified autoantibody response to the cancer antigen. This response is measurable in cancer patients and is currently being exploited in the EarlyCDT®-Lung test for early detection of lung cancer.

There is limited evidence that autoantibodies may be present in healthy individuals, Li et al. found autoantibodies to alpha-enolase and heterogeneous nuclear ribonucleoprotein L in over 50% of a cohort of 36 healthy control group analysed by mass spectroscopy(123), while Nolen et al. reported 4.39% of a cohort of 200 healthy patients showed elevated levels in over 40% of tumour associated AAbs tested(124). These studies have demonstrated that there may be pre-existing AAbs, produced in response to non-malignant autoimmune disorders, and not associated with the tumours in cancer patients, which may reduce the specificity of AAb tests in the detection of cancer.

The antigen-antibody reaction is a reversible chemical reaction:

$$\textit{antigen} + \textit{antibody} \rightleftarrows \textit{antigen-antibody complex}$$

and the forces complexing the antigen and antibody are ‘weak’ non-covalent bond interactions. The strength of these binding interactions and the resulting complex differs depending on the structures of the antigen and antibody, with the affinity of the bond being defined by the equilibrium constant (K_{eq}) and related dissociation constant (K_d). During an immune response, the affinity of the produced antibodies also progressively increases, as repeated exposure of the immune system to an antigen causes maturation of the response and preferential production of higher affinity antibodies, resulting in antibodies with affinities up to 100 times greater than those produced during the initial immune response(125, 126).

For this reason, it is hypothesised that the relative binding affinity of the response measured by the EarlyCDT®-Lung test may give additional information about the nature of the underlying immune response, and may allow the preferential identification of highly cancer-specific responses over the responses observed in a small percentage of the normal population. Identification and reclassification of false positive subjects would then result in improvements to the specificity of the EarlyCDT®-Lung test, through quantification of the curve characteristics related to the relative binding affinity and their use as a secondary metric to exclude normal subjects. Determination of the binding affinities may also identify additional subjects with high affinity autoantibodies that are currently at serum titres which are too low to allow distinction from healthy controls and has the potential to increase the number of true positive subjects identified and give improvement to the sensitivity of the test.

1.15 Aims

The aim of this project was to assess the potential of AAb-TAA binding curve characteristics as a complementary parameter to the magnitude of the AAb response in the EarlyCDT®-Lung early detection test, to increase the specificity and clinical utility of the test. In addition, this project set out to explore the health economic benefits of the current EarlyCDT®-Lung test, along with the potential increase in economic benefits that would be derived from changes to the diagnostic performance of the test.

This project also aimed to examine all of the generated biomarker parameters using a range of machine learning strategies for combining biomarker data to generate models and algorithms which could maximise the diagnostic performance of the test.

This project utilised three case-control cohorts, the first of which was partitioned into training and test sets for initial assessment of parameters, as well as biomarker selection and model building. The remaining sets were used for validation of panels and models created and trained on the training set, in order to confirm findings in an independent sample set and therefore reduce the effect of overfitting biases.

Chapter 2: Pilot Study

2.1 Aims

During clinical validation and subsequent commercial release of the EarlyCDT®-Lung test, I made the observation that the shape of the antigen-specific binding curves varied between individuals, even those showing the same resultant magnitude, suggesting that differences in the autoantibody-antigen binding interactions are reflected in the titration curve data, and the following study set out to explore whether metrics based on the shape of the titration curve may allow discrimination and reclassification of false positive signals.

2.2 Introduction

The EarlyCDT®-Lung test has been technically and clinically validated for the early detection of lung cancer with a sensitivity ~40% and a specificity of ~90%. Due to the relatively low incidence of lung cancer, the positive predictive value (PPV) of the test is primarily driven by the false positive rate (FPR). Identification of false positives and reclassification as true negatives would therefore increase the PPV and hence clinical utility of EarlyCDT®-Lung.

The test generates curves of autoantibody (AAb) binding to a titrated series of capture antigen concentrations thus providing patient-specific autoantibody profile titration curves. We postulated that the antibodies responsible for false positive results in healthy individuals would exhibit different binding characteristics to the specific autoantibodies present in cancer patients and that these differences may be extrapolated in the shape of the autoantibody-antigen titration curves. Here we demonstrate the exploitation of differences in titration curve shape characteristics for the reclassification of subjects identified as false positive under the current test.

2.3 Method

2.3.1 Autoantibody Biomarkers

The autoantibody biomarkers investigated in this pilot study are summarised in Table 2-1, the initial seven being those currently assessed in the EarlyCDT®-Lung test, along with an expanded panel of additional 13 investigatory autoantibodies.

Table 2-1: Autoantibody Biomarkers under investigation.

Panel	Autoantibody	Cancer Expression
EarlyCDT®-Lung Commercial	p53	Mutated Form
EarlyCDT®-Lung Commercial	SOX2	Overexpressed
EarlyCDT®-Lung Commercial	CAGE	Aberrantly Expressed
EarlyCDT®-Lung Commercial	NY-ESO-1	Aberrantly Expressed
EarlyCDT®-Lung Commercial	GBU 4-5	Unassigned Tumour Associated Antigen
EarlyCDT®-Lung Commercial	MAGE A4	Aberrantly Expressed
EarlyCDT®-Lung Commercial	HuD	Mutated Form
Expanded Panel	p62	Aberrantly Expressed
Expanded Panel	ALDH1	Overexpressed
Expanded Panel	p16-C	Mutated Form
Expanded Panel	GRP78	Overexpressed
Expanded Panel	SSX1	Aberrantly Expressed
Expanded Panel	P53-95	Mutated Form
Expanded Panel	Alpha enolase	Overexpressed
Expanded Panel	P53-C term	Mutated Form
Expanded Panel	KOC	Aberrantly Expressed
Expanded Panel	K-Ras	Mutated Form
Expanded Panel	CK8	Overexpressed
Expanded Panel	CK20	Overexpressed
Expanded Panel	Lmyc2	Overexpressed

2.3.2 Assay

Autoantibodies were measured using the EarlyCDT®-Lung test, a quality-controlled, semi-automated indirect enzyme-linked immunosorbent assay in which samples were reacted with a series of concentrations of tumour associated antigens, as described in Chapman et al(108). Liquid-handling steps were carried out using an automated system. The same assay technique was then employed for the assessment of the thirteen additional antigens. Optical density data for the ELISA assays was determined spectrophotometrically at 650nm and exported to Microsoft Excel for assessment. After semi-automated and visual assessment for anomalous results, the optical density data was collated into .dta files for further assessment in Stata 14.0.

2.3.3 Sample Sets

Sample cohorts consisted of a development set, a confirmation set, and a large normal control cohort.

The development set consisted of serum samples from 337 lung cancer patients, collected at or shortly after histopathological confirmation of lung cancer. These samples were obtained from lung cancer centres and sample biobanks in North America, Ukraine, and the UK, and 415 normal controls, obtained from biobanks in North America, and a UK sample collection. This set represents a subset of the samples described as Groups 1-4 when described by Lam et al.(107), for which data on the additional autoantibodies MAGE-A4 and HuD was obtained. These comprised:

- 33 lung cancer samples from patients with SCLC presenting to a single centre in the United Kingdom (described by Lam et al. as Group 1).
- 161 lung cancer and 193 healthy control samples were obtained from patients with lung cancer collected in multiple European centres. The lung cancer patients were originally matched for age, sex, and smoking history with samples from normal populations in Europe and the United States however matching was not maintained in this subset (described by Lam et al. as Group 2).
- 120 patients with lung cancer treated at a single center in Vancouver, Canada, who were matched to 113 control samples from high-risk individuals who did not have lung cancer previously (described by Lam et al. as Group 3).
- 23 patients with lung cancer treated in the UK, and 109 healthy controls from normal populations in Europe and the United States (described by Lam et al. as Group 4).

The confirmation set consisted of serum samples from 235 lung cancer patients, again collected at or shortly after histopathological confirmation of lung cancer, also obtained from lung cancer centres and sample biobanks in North America, Ukraine, and the UK, and 266 normal controls obtained from a North American biobank, 235 of which were matched as closely as possible to the cancer cohort for age, gender and smoking history. This cohort was previously described as the `Optimisation Set` when published by Chapman et al(108).

Finally the Normal set was comprised of serum samples from 2110 healthy normal subjects collected as part of a population autoantibody study in the UK midlands. In obtaining samples for the cancer cohorts, early-stage disease was prioritised in order to ensure that results were applicable to detection of early-stage disease, and not simply a reflection of cancer burden in late-stage disease.

Additionally a subset of 151 lung cancers and 104 healthy normal samples from the development set were assessed against an expanded panel of autoantibodies, and are referred to subsequently as the "expanded panel set".

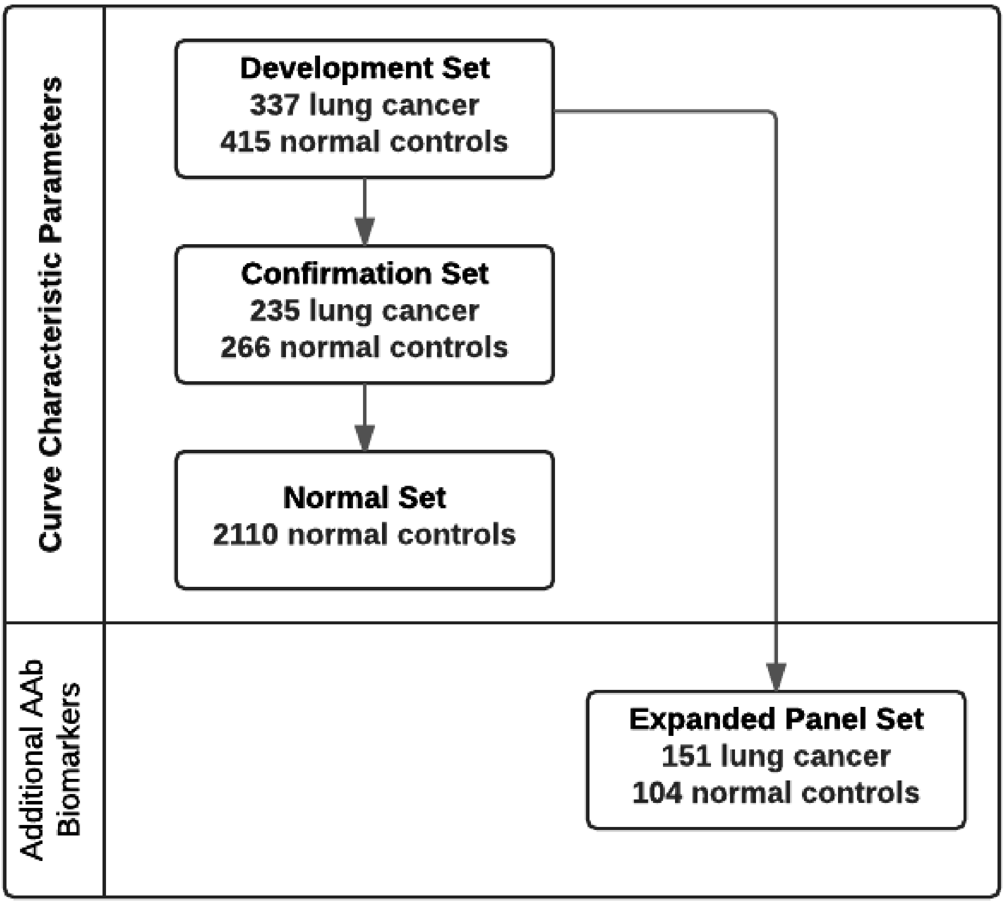


Figure 2-1: Preliminary study configuration.

2.3.4 Curve Characteristic Calculations

Linear regression was performed on both raw and log-scaled titration curve data obtained in order to generate values for the Slope, and Intercept, while Area Under the Curve (AUC) was calculated using the trapezoid method, and SlopeMax (representing the slope at the steepest point of the titration curve) of each antigen was calculated through calculation of slope (slope = $\frac{\Delta y}{\Delta x}$) for each set of adjacent point. These features are illustrated in Figure 2-2.

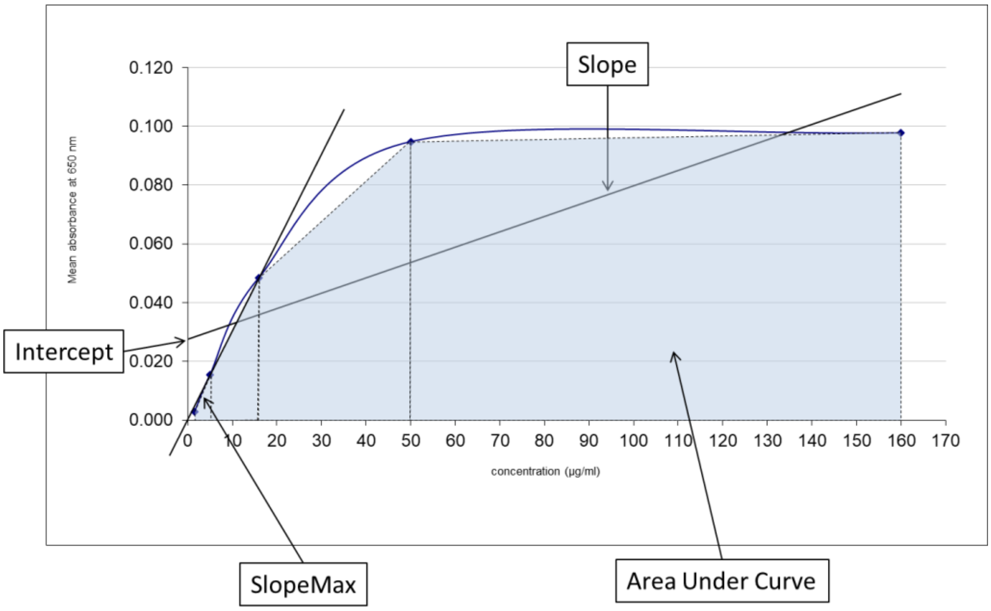


Figure 2-2: Diagram of curve characteristic parameters on example autoantibody titration curve

These secondary curve characteristics were then investigated in addition to the standard metric (the magnitude of the signal at the two highest concentrations of the curve) to determine whether the curve characteristics could selectively reclassify false positive signals as true negatives. The panel was then extended in a subset of the development set consisting of 151 lung cancer patients and 104 normal controls, with the inclusion of additional autoantibody measurements in order to restore sensitivity and optimise the test performance characteristics.

2.3.5 Panel Optimisation

The panel was optimised through manual assessment of cut-off thresholds based on visual assessment of paired scatter plots, and selection for each antigen of the curve characteristic feature which gave the greatest specificity improvement while maintaining sensitivity for each antigen.

2.4 Results

2.4.1 Cohort Demographics

Demographic and histological information for the examined cohorts are summarised in Table 2-2 and Table 2-3, showing that, with the exception of the large cohort of normal samples (Normal Set), the cohorts were predominantly male. All cancer cohorts were mainly composed of current and ex-smokers, and the age profile of all the cohorts was relatively wide, with minimum ages lower than would generally be included in lung cancer screening. Cancer cohorts intentionally showed a higher proportion of early-stage disease, as elevated immune responses are expected early in tumour development, and this study was concerned with identifying these early signals. The cancer cohorts were also primarily non-small cell cancer subtypes, reflecting cancer incidence in the population, although an increased proportion of small cell lung cancer was included in the cohort to ensure that any test derived from this analysis is also relevant to these more aggressive cancers.

Table 2-2: Demographic profile of studied datasets

	Development Set		Confirmation Set		Normal Set	Expanded Set	
	Cancer Case	Healthy Control	Cancer Case	Healthy Control	Healthy Control	Cancer Case	Healthy Control
Samples (n)	337	415	235	266	2110	151	104
Gender[1]							
Male	219 (65.0%)	265 (63.9%)	171 (72.8%)	185 (69.5%)	352 (16.7%)	125 (82.8%)	90 (86.5%)
Female	118 (35.0%)	148 (35.7%)	64 (27.2%)	81 (30.5%)	1730 (82.0%)	26 (17.2%)	14 (13.5%)
Unknown	0 (0.0%)	2 (0.5%)	0 (0.0%)	0 (0.0%)	28 (1.3%)	0 (0.0%)	0 (0.0%)
Age (years)[2]	63 (23,90)	62 (23,87)	65 (42,85)	65 (38,86)	53 (17,88)	61 (39, 86)	60 (39,81)
Smoking Status[1]							
Current Smoker	176 (52.2%)	78 (18.8%)	108 (46.0%)	93 (35.0%)	325 (15.4%)	92 (60.9%)	8 (7.7%)
Former Smoker	112 (33.2%)	237 (57.1%)	67 (28.5%)	139 (52.3%)	616 (29.2%)	36 (23.8%)	39 (37.5%)
Never Smoker	43 (12.8%)	99 (23.9%)	24 (10.2%)	26 (9.8%)	1134 (53.7%)	21 (13.9%)	57 (54.8%)
Unknown	6 (1.8%)	1 (0.2%)	36 (15.3%)	8 (3.0%)	35 (1.7%)	2 (1.3%)	0 (0.0%)
Country of Origin[1]							
UK	55 (16.3%)	217 (52.3%)	47 (20.0%)	0 (0.0%)	2110 (100%)	14 (9.3%)	104 (100%)
US	0 (0.0%)	85 (20.5%)	36 (15.3%)	266 (100%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Canada	120 (35.6%)	113 (27.2%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Ukraine	133 (39.5%)	0 (0.0%)	102 (43.4%)	0 (0.0%)	0 (0.0%)	111 (74.5%)	0 (0.0%)
Unknown	29 (8.6%)	0 (0.0%)	50 (21.2%)	0 (0.0%)	0 (0.0%)	26 (17.2%)	0 (0.0%)

1 n (%), 2 Median (Min, Max)

Table 2-3: Histologic profile of case cohorts in studied datasets

	Development Set Cases	Confirmation Set Cases	Expanded Set Cases
Stage I	162 (48.1%)	99 (42.1%)	95 (62.9%)
Stage II	41 (12.2%)	56 (23.8%)	13 (8.6%)
Stage III	44 (13.1%)	12 (5.1%)	16 (10.6%)
Stage IV	21 (6.2%)	9 (3.8%)	15 (9.9%)
Stage Unknown	69 (20.5%)	59 (25.1%)	12 (7.9%)
Adenocarcinoma	136 (40.4%)	61 (26.0%)	46 (30.5%)
Squamous	56 (16.6%)	82 (34.9%)	29 (19.2%)
Adenosquamous	0 (0.0%)	2 (0.9%)	0 (0.0%)
Small Cell	37 (11.0%)	46 (19.6%)	4 (2.6%)
Large Cell	6 (1.8%)	5 (2.1%)	1 (0.7%)
Other	65 (20.5%)	14 (6.0%)	35 (23.2%)
Subtype Unknown	37 (11.0%)	25 (10.6%)	36 (23.8%)

2.4.2 Diagnostic Performance after Panel Optimisation

Table 2-4: Effect of application of a secondary curve parameter cut-off on assay performance characteristics in three different patient cohorts. Also, the effect of expanding the EarlyCDT®-Lung panel to include additional autoantibody measurements, along with their secondary curve parameters to optimise test performance. N/A = Not Applicable

Cut-off	Specificity (%)	Sensitivity (%)	Positive Predictive Value (%)
Development set: Standard EarlyCDT®-Lung	90.1	29.7	5.8
Plus secondary curve parameter cut-off	98.1	22.0	19.1
Confirmation Set: Standard EarlyCDT®-Lung	90.6	41.5	8.3
Plus secondary curve parameter cut-off	97.0	27.9	16.0
2110 Normal Controls: Standard EarlyCDT®-Lung	86.6	N/A	N/A
Plus secondary curve parameter cut-off	97.7	N/A	N/A
Expanded Panel Set: 7Ag Standard EarlyCDT®-Lung	90.4	31.8	6.3
7Ag Panel plus secondary curve parameter cut-off	98.1	23.8	20.4
18AAb Expanded Panel plus secondary curve parameter cut-off	99.0	50.3	50.7

Application of cut-offs based on one of the four curve characteristic parameters was able to increase specificity for each of the panel autoantibodies and resulted in a reproducible panel specificity of around 98%, these cut-offs are summarised in plots shown in Appendix 2A. Although a corresponding reduction in sensitivity was observed in each case, the overall performance showed an increased PPV in every dataset.

Applying the same rationale for application of cut-offs to both magnitude and simultaneously a curve characteristic parameter in order to incorporate additional autoantibodies to the panel in the expanded panel set resulted in a final panel which assessed 18 autoantibodies and performed with 99.0% specificity and 50.3% sensitivity.

2.5 Chapter Conclusions

The curve characteristic features based on the shape of the antigen titration curves generated by the EarlyCDT®-Lung test were able to give additional useful curve-parameter metrics. When cut-offs were applied to the curve parameter, in combination with cut-off thresholds to the current standard magnitude of signal-based measurements, it was possible to improve the specificity and hence PPV of the test, and these improvements were maintained in three independent datasets, including one large normal set, with only a moderate reduction in sensitivity.

It was then possible to restore and improve upon the initial sensitivity of the test by inclusion of additional autoantibodies to the panel, however this expanded panel set was relatively small and represent the results of training only, additional cohorts are required to confirm the performance improvements that may be provided through addition of further autoantibody biomarkers.

True calculation of the affinity of the protein interactions would require either quantification of the concentration of free antibody in the serum sample, or a competitive binding assay against a sample of known antibody concentration. As the true concentration of free antibody is not currently quantified, and the current EarlyCDT®-Lung test does not use a competitive binding assay, a comparative estimation of the protein binding will be explored by examining parameters describing the shape of the titration curve. The initial parameters explored here are the slope, intercept, area under the curve and maximum slope of the curve. As easily calculable attributes of the titration curve, these have shown potential for improving the specificity of the autoantibody panel in this pilot study, however a more accurate assessment of the binding kinetics of the autoantibodies being measured may lead to greater insight into whether differences in these protein interactions are indicative of specific immune responses to cancer-related aberrant or upregulated protein production, or are non-specific or cross-reactive immune responses to non-cancer related causes.

The availability of samples necessitated the use of samples from across both Europe and the US in these analyses. As immunological differences across geographies have not been properly explored in cancer autoimmune responses this may have introduced bias into the results, as cancer immunology may differ based on regional genetics or differences in carcinogens and environmental risk factors, and the datasets were not balanced across cases and controls, or between cohorts, for sample country of origin. The exploration of these differences were outside the scope of this study, and the impact of this bias was limited by the inclusion of both European and US samples within both case and control cohorts in the development set, and the lack of European controls in the confirmation set is offset by the application of models to the large cohort of normal control samples from a European collection.

Although further investigation is warranted, these methods were able to give significant improvement in the performance characteristics and potential clinical utility of the EarlyCDT®-Lung test.

2.6 Chapter Discussion

The pilot study showed good potential for reclassification of false positive signal and improvement to the test specificity and suggested that the, in the absence of direct measurements enabling the calculation of dissociation constants, observed differences in the shape of the autoantibody binding curves may act as surrogate features indicating strength or specificity of autoantibody binding. These initial findings suggested that addition of further features could then result in increases to sensitivity, however I decided that quantifying the diagnostic performance improvements required would instruct exactly how many additional features would be required for any resultant algorithms, which inspired the subsequent investigation into the health economics behind early lung cancer screening, and the presence of autoantibodies prior to imaging presentation.

Chapter 3: Lung Cancer Associated Autoantibody Responses are Detectable Years Before Clinical Presentation

3.1 Aims

A potential issue with the use of autoantibodies to detect cancer at its earliest stages, is that the autoantibody response may return a positive result for a cancer that is too small for current radiology to identify, which would then mistakenly be classified as a false positive. To try and address this and instruct the length of follow-up required following a positive test result, I undertook an analysis on a longitudinal data set that was obtained in a collaboration with UCL and Abcodia to identify how long prior to CT presentation elevated autoantibodies can be detected.

3.2 Introduction

Globally, lung cancer has the highest mortality rate of all cancers and was estimated to be responsible for nearly 1.8 million deaths in 2020(127). Lung cancer is generally not detected until symptomatic presentation, at which point it is usually advanced stage, with extremely poor prognosis. For this reason, detecting lung cancer at an early stage can vastly improve 5-year survival, with stage 1 detected lung cancers having a 5-year survival of 62.7%, compared to only 4.3% for those cancers diagnosed at stage 4(128). In vitro estimates of imaging detection limits have suggested that a lung cancer does not become detectable by current imaging modalities until it has reached a population of at least 100,000 cells(29), assuming exponential growth, this would suggest that on average a malignant cell population would need to double 16.6 times ($\log_2(100000)$) before becoming detectable by imaging. Radiography studies have determined a mean doubling time in malignant lung cancer of 158 days(28), which would denote that, on average, a lung cancer is present and potentially able to elicit an autoimmune response for at least 7.2 years before it can be confirmed by imaging. This is supported by studies which have shown evidence of detectable levels of tumour associated autoantibodies in individuals prior to presenting with cancer, including p53 responses a median of 3.5 years prior to lung cancer imaging detection(129), and the detection of p53 and Her2 in pre-diagnostic breast cancer samples(130). Imaging surveillance of subjects after a positive autoantibody test response should allow for detection of a cancer at the very earliest stages, vastly improving prognosis.

The EarlyCDT®-Lung test(107, 108) is a simple blood test which detects elevated levels of a panel of tumour-associated autoantibodies generated in response to abnormal tumour cells. This test has been validated in case-control studies, and in commercial practice(131), and it’s use as a screening test in a high risk population has been shown to result in a stage shift in diagnosis favouring early stage detection.(109) Previous validation studies have focused on diagnostic sensitivity and specificity at time of testing, and the length of time prior to clinical presentation of lung cancer (lead time) that elevated autoantibodies can be detected using the EarlyCDT®-Lung antigen panel has not previously been established.

The United Kingdom Collaborative Trial of Ovarian Cancer Screening (UKCTOCS)(132) was a large prospective trial which aimed to quantify the benefits of an ovarian cancer screening program. In this trial 202,638, postmenopausal women aged 50 and above were recruited through thirteen UK centers between 2001 and 2005. The multimodal screening arm of this study contained 50,640 women for whom annual blood samples were taken, all of whom were followed up for development of cancer. Within this cohort a number of women went on to develop a lung cancer, analysis of their longitudinal blood samples compared to a matched control cohort will allow assessment of when elevated autoantibodies were first detectable in their blood samples, compared to the time at which their cancer was detected. The primary aim of this study was to assess the diagnostic performance of the EarlyCDT®-Lung panel of autoantibodies over the years before diagnosis, in order to estimate how early before symptomatic presentation elevated autoantibodies can be detected using the EarlyCDT®-Lung test.

3.3 Materials and Methods

3.3.1 Patient Cohorts

Case and control cohorts were identified from samples collected as part of the UKCTOCS trial of post-menopausal women aged 50-74 years. The case cohort was comprised of 142 subjects who presented with lung cancer during the course of the UKCTOCS study follow up, and who had at least three serial samples and a known date of lung cancer diagnosis. Subjects were matched for age at trial entry (+/- 5 years), smoking history, and trial entry date (+/- 2 years) to a control cohort of 142 subjects who had no evidence of developing lung cancer during the UKCTOCS follow-up period. All cases had between 4 and 8 longitudinal samples (median 7), covering a period of between 3.1 and 8.9 years (median 6.2), while controls had between 3 and 8 longitudinal samples (median 6) covering a period of between 2.9 and 9.2 years (median 6.3). Demographic details of the cohorts are outlined in Table 3-1. Histological subtypes in the cancer cases were 49% Adenocarcinoma, 17% Squamous cell carcinoma, 16.4% Unspecified Non-small cell carcinoma, 13% Small cell carcinoma, with the remaining 5.6% comprised of carcinoid tumour, large cell carcinoma, and neuroendocrine carcinoma.

Table 3-1: Cohort Demographics

Variable	N	CASE, N = 142	CONTROL, N = 142	p-value
Age at Sample Collection[1]	284	64 (59, 69)	64 (59, 69)	0.8[3]
Smoking status	258			0.9[4]
Non-Smoker[2]		27 (21%)	28 (22%)	
Smoker[2]		102 (79%)	101 (78%)	
Unknown		13	13	

1 Median (IQR), 2 n (%)

3 Wilcoxon rank sum test, 4 Pearson's Chi-squared test

All samples were received blinded and assessed on the EarlyCDT®-Lung test for autoantibody responses to a panel of seven tumour associated antigens, these responses were compared to pre-determined commercial cut-off thresholds to return a panel assessment of either “negative” referring to no elevated risk of lung cancer, “moderate positive” relating to an elevated risk of lung cancer, or “high positive” referring to a highly elevated risk of lung cancer.

Samples were unblinded, and positivity assessed by subject and longitudinal timepoint. For case samples, assay positivity was compared to date of cancer detection to determine how early prior to current detection methods a detectable autoantibody response was present.

Additionally, time to detection has been assessed by histological subtype for the three most prevalent subtypes in the dataset, adenocarcinoma (69 subjects, 49% of the cohort), squamous cell carcinoma (24 subjects, 17% of the cohort), and small cell carcinoma (18 subjects, 13% of the cohort).

3.4 Results

3.4.1 Commercial performance

EarlyCDT®-Lung commercial performance for the study cohorts was assessed over all samples, with subjects being assigned the highest level of positivity returned from any of their longitudinal samples, the results of which are summarised in Table 3-2, and show that these samples returned a sensitivity of 26.1%, and specificity of 88.7%.

Table 3-2: 2x2 Contingency Table summarising diagnostic performance (moderate and high positive) over all samples.

	EarlyCDT Positive	EarlyCDT Negative	Total
Cases	37	105	142
Controls	16	126	142
Total	53	231	284

3.4.2 Cancer case time to detection - moderate and high positive

Table 3-3: Earliest detection and median time to detection by antigen (EarlyCDT moderate and high positive)

Antigen	Number of Cases Positive	Earliest Pre-Dx Time to Detection (months)	Median Pre-Dx Time to Detection (months)
p53	15	101.0	53.3
SOX-2	2	73.6	73.2
CAGE	6	78.6	24.1
NY-ESO-1	6	95.5	31.0
GBU 4-5	3	77.5	58.1
MAGE-A4	6	81.2	48.1
HuD	0	NA	NA
Panel	37	101.0	49.9

Examining all positive responses (both moderate positive, and high positive) as summarized in Table 3-3, autoantibodies were detected up to 101.0 months (8.4 years) prior to clinical presentation using current detection methods, with positive responses being present a median of 49.9 months (4.2 years) before presentation of lung cancer. The earliest autoantibody responses were observed against p53, which also showed the highest sensitivity in this study, however responses were observed in all autoantibodies other than HuD at time points in excess of 73 months (6.1 years) prior to clinical presentation with current detection methods.

3.4.3 Cancer case time to detection - high positive

Table 3-4: Earliest detection and median time to detection by antigen (EarlyCDT high positive only)

Antigen	Number of Cases Positive	Earliest Pre-Dx Time to Detection (months)	Median Pre-Dx Time to Detection (months)
p53	4	31.2	9.9
SOX-2	1	60.0	60.0
CAGE	4	33.4	8.8
NY-ESO-1	4	50.6	33.4
GBU 4-5	0	NA	NA
MAGE-A4	1	14.9	14.9
HuD	0	NA	NA
Panel	14	60.0	15.6

EarlyCDT®-Lung high positive responses are summarized in Table 3-4 and show high positive results were detected up to 60 months (5 years) prior to clinical presentation, with a median time of 15.6 months (1.3 years) from high positive autoantibody test to presentation.

3.4.4 Cancer case time to detection by subtype - moderate and high positive

Adenocarcinoma

Table 3-5: Adenocarcinoma earliest detection and median time to detection by antigen (EarlyCDT moderate and high positive)

Antigen	Number of Cases Positive	Earliest Pre-Dx Time to Detection (months)	Median Pre-Dx Time to Detection (months)
p53	5	86.6	50.1
SOX-2	1	72.9	72.9
CAGE	2	40.9	39.8
NY-ESO-1	3	75.2	48.6
GBU 4-5	0	NA	NA
MAGE-A4	4	81.2	48.2
HuD	0	NA	NA
Panel	14	86.6	47.4

Squamous Cell Carcinoma

Table 3-6: Squamous cell carcinoma earliest detection and median time to detection by antigen (EarlyCDT moderate and high positive)

Antigen	Number of Cases Positive	Earliest Pre-Dx Time to Detection (months)	Median Pre-Dx Time to Detection (months)
p53	5	81.5	75.4
SOX-2	1	73.6	73.6
CAGE	3	9.6	4.0
NY-ESO-1	1	95.5	95.5
GBU 4-5	2	58.1	38.0
MAGE-A4	2	49.9	34.3
HuD	0	NA	NA
Panel	14	95.5	51.6

Small Cell Carcinoma

Table 3-7: Small cell carcinoma earliest detection and median time to detection by antigen (EarlyCDT moderate and high positive)

Antigen	Number of Cases Positive	Earliest Pre-Dx Time to Detection (months)	Median Pre-Dx Time to Detection (months)
p53	0	NA	NA
SOX-2	0	NA	NA
CAGE	0	NA	NA
NY-ESO-1	1	13.4	13.4

GBU 4-5	1	77.5	77.5
MAGE-A4	0	NA	NA
HuD	0	NA	NA
Panel	2	77.5	45.5

3.5 Chapter Conclusions

Analysis of the cohort of subjects that went on to develop cancer within the UKCTOCS study shows that EarlyCDT®-Lung was able to identify tumour associated autoantibody responses up to a maximum of 8.4 years before CT presentation, with detectable elevated autoantibody responses presenting a median of 4.2 years before detection by CT in subjects that went on to develop lung cancer. This is comparable to the work published by Li et al(123) which reported elevated p53 autoantibodies were detectable an average of 3.5 years prior to clinical presentation in a cohort of 49 high risk subjects who subsequently developed a lung cancer, and supports the theory that the immune system responds to the presence of tumour associated antigens early in cancer development, while the cancer is still comprised of a relatively small number of cells, and detection and monitoring of these autoantibodies represents a unique opportunity to identify a malignancy at its earliest stages.

The presence of these autoantibody responses at such an early stage prior to presentation may occur for other reasons, while have theorized that the long timespan likely represents the initial presentation of a small number of malignant tumour cells to the immune system, they may also be part of an immune response to a larger malignancy, suppressing its growth until it is able to mutate to develop immunosuppressive or immunoevasive traits.

Observation of the change in autoantibody signal over time (as shown in Appendix 3A) reveals that many of the state changes from negative to positive between time points are associated with a drastic increase in autoantibody signal. This reflects the natural amplification of an immune response to a malignancy and lends further credence to these being highly specific responses to antigen mutation or overexpression as a result of tumour growth.

The UKCTOCS collection was primarily concerned with identifying ovarian cancer, the examined cohort was identified from incident lung cancers discovered during the follow-up period of the study and represented an opportunity to examine longitudinal pre-diagnostic autoantibody responses in lung cancer. The cohort is therefore not the intended screening population for the EarlyCDT®-Lung test, being entirely female and with a large proportion of never smokers. As cancer incidence and mortality is higher in males, and smoking history is a major risk factor for lung cancer this would result in the cohort having a lower pre-existing cancer risk, and may account in part for the observed sensitivity of 26.1% in this cohort being lower than the 37.1% previously observed during clinical use, although it is still within 95% confidence intervals for the test(131). Additionally, while the EarlyCDT®-Lung test has been validated in case-control studies to have a specificity around 90%, evidence of elevated autoantibodies up to 8.4 years prior to clinical presentation as presented by this analysis raises the possibility that at least a portion of those presenting as false positives in these case-control studies represent latent cancers which may subsequently present, or are potentially indicative of a successful immune response to mutated cells which will resolve without malignant presentation.

Finally, median times to detection observed did show some differences by antigen, and this was explored further through analysis by histological subtype. Tumour doubling times have previously been shown to differ based on histology, with adenocarcinoma typically being associated with longer doubling times than squamous cell carcinoma or small cell carcinoma(133), however median time to detection in this study was comparable between adenocarcinoma, squamous cell carcinoma, and small cell carcinoma, at 47.4, 51.6, and 45.5 months respectively. Within adenocarcinoma and squamous cell carcinoma subjects, CAGE autoantibody responses were evident closer to diagnosis than other panel autoantibodies, especially in squamous cell carcinoma, and this may be indicative that CAGE overexpression or mutation either develop later in disease progression or are possibly related to more aggressive malignancies with shorter doubling times. Further work is needed on larger, more representative cohorts to understand these relationships, although this raises the possibility that assessment of autoantibodies may be useful in assessing the aggressiveness of a malignancy prior to its clinical presentation.

3.6 Chapter Discussion

While the presence of elevated autoantibodies in response to the development of a malignancy is expected to occur prior to the tumour being large enough to be visualized by imaging screening, this study is able to confirm the presence of these autoantibodies up to 8.4 years prior to CT screen presentation. The presence of autoantibodies at this early stage may represent a unique opportunity for surveillance and potentially disease stratification in the time before a cancer can be confirmed and treated, although training and confirmation of models such as that proposed for a longitudinal changepoint based test similar to that attempted by the ROCA test(134), or the assessment of personalized baselines, would require large scale longitudinal trials in high risk populations to gather a large enough case cohort.

Chapter 4: Health Economic Assessment of an Autoantibody Screening Test for Lung Cancer compared to CT Screening.

4.1 Aims

Having shown that autoantibody responses are present and measurable prior to symptomatic presentation, and with the overarching aim of this project being to improve the diagnostic performance of the EarlyCDT®-Lung test, I undertook a health economics analysis to quantify firstly exactly how health economically beneficial the current format of the EarlyCDT®-Lung test is, in comparison to no screening and CT screening, and from that, identify exactly what improvements, in terms of diagnostic sensitivity and specificity, would lead to health economic improvements which would accelerate acceptance and adoption of the EarlyCDT®-Lung test.

4.2 Introduction

The American Cancer Society (ACS) currently recommends annual lung cancer screening by low dose computed tomography (LDCT) in 55-74 year old current or previous smokers with a pack year history of 30 years or greater. These recommendations are based upon the results of the National Lung Screening Trial, which showed that screening with LDCT showed a 20.0% decrease in lung cancer mortality compared to screening by radiography(43), which was deemed to be cost-effective at $81,000 per quality-adjusted life year (QALY) gained when compared to no screening(135).

The EarlyCDT®-Lung test is an autoantibody blood test (AABT) that detects the presence of a panel of cancer related autoantibodies. Elevated levels of one or more of these autoantibodies above a pre-determined cut-off threshold indicates an increased risk of lung cancer in high-risk individuals. Clinical validation and subsequent audit of the EarlyCDT®-Lung test has shown it to have a specificity of 91%, with sensitivity of 37-41% for the early detection of lung cancer(108, 131), while it has also been shown in the Early Diagnosis of Lung Cancer in Scotland (ECLS) trial to increase cancer detection at an earlier stage(109). EarlyCDT®-Lung has been the subject of a National Institute for Health and Care Excellence (NICE) diagnostics guidance for risk assessment in indeterminate pulmonary nodules which concluded that while EarlyCDT®-Lung has the potential to identify nodules that require immediate treatment or biopsy and may result in improved treatment options and patient outcomes, further research is still needed to confirm the accuracy of EarlyCDT®-Lung and the models through which the assessment of risk in positive samples is calculated(136).

While it is not currently recommended by the ACS, the EarlyCDT®-Lung test has been shown in a decision-analytical modelling assessment to be cost effective as a screening test in a hypothetical cohort of 100,000 high risk individuals (at a cost of $300 per test), when used in combination with computed tomography (CT) (at a cost of $301 per test) using a strategy of confirming positive Early-CDT-Lung test with a follow up CT. This analysis determined an estimated cost of $20,044 per quality-adjusted life year (QALY) compared to no screening(137), Additionally this strategy was shown in a separate analysis to have increased cost effectiveness in a high risk smoking population, with estimated costs of $9,549 per QALY (138).

The cost-effectiveness of the EarlyCDT®-Lung test has also been established in directing treatment of incidentally detected pulmonary nodules, whereby a decision analytical model showed a cost per QALY of $24,330 when compared to CT surveillance alone (139). These studies were based on data obtained from the National Lung Screening Trial (NLST)(42) and focused on healthcare costs in a US setting. In the UK, the National Screening Committee has only recently (June 2022) recommended that targeted lung screening using low-dose computed tomography be undertaken in high risk individuals(140), although further work is recommended to better establish the effectiveness of different implementation strategies. This recommendation was instructed by an extensive health technology assessment which utilised a discrete event simulation model to explore several strategies for screening with low dose CT in high-risk individuals. The assessment was able to show that a screening programme in smokers aged 60–75 years with a ≥ 3% risk of lung cancer using a single screen of low-dose CT would be cost effective in the NHS at an estimated cost of around £28,000 per QALY gained(30). As a combination screening strategy of EarlyCDT®-Lung and CT screening was shown to be more cost-effective than screening with CT alone in the US, a similar strategy should prove more cost effective than CT alone in an NHS setting.

A health economic assessment based upon the discrete event simulation model defined by Snowsill et al.(30) has been undertaken in order to quantify the health economic benefits of screening strategies using the EarlyCDT®-Lung autoantibody test as a primary screening tool, compared to those obtained by current best practise - low-dose CT screening - in a UK NHS setting. Additionally this model has been used to give an estimate of the potential diagnostic lead times associated with EarlyCDT®-Lung, in comparison to those already observed in LDCT screening, as well as estimate the health economic benefit associated with the stage shift in clinical presentation and resultant mortality reduction previously observed with lung cancer screening(42), and the EarlyCDT®-Lung test specifically(109).

An appreciation of the relative impact of all model parameters on the health economic outcomes has been undertaken through the use of univariate sensitivity analysis, giving evidence for which factors have the greatest influence on the cost-effectiveness of the screening test. This was then extended to estimate the potential health economic benefits of improvements to diagnostic performance (sensitivity and specificity) of the EarlyCDT®-Lung test, assuming no change in test cost, compared to the current commercial test, in order to quantify the performance levels necessary to reach the various health economic thresholds currently advised by the NHS.

4.3 Methods

4.3.1 Model Development

An individual patient simulation model was constructed in R v4.0.3 with a parallel model constructed in Excel for Microsoft 365 (version 2101), using a Discrete Event Simulation (DES) framework based upon that developed by

Snowsill et al, for the comparison of screening strategies based on the use of the EarlyCDT®-Lung autoantibody test (AABT) against both low dose CT (LDCT) screening, and no screening (current practise) in a UK NHS setting. The decision was made to follow the approach used by Snowsill et al, due to the development and inclusion of a natural history model for simulated lung cancers, allowing for an economic appreciation of the stage shift that has been demonstrated through the use of the EarlyCDT®-Lung test.

The model was developed for a UK NHS setting, with invitation to screening and determination of pre-existing cancer risk for purpose of screening eligibility undertaken at a primary care level. Screening AABT and LDCT would be performed in a secondary or tertiary care setting. Differences in access to equipment necessary for AABT and LDCT screening has not been factored into this model.

The target population for screening are people at high risk of lung cancer, specifically people aged 55 to 80 years with a history of smoking. Further exploration of subgroups of this population was undertaken through restrictions to the age range and cancer risk during the modelling of population strategies. All individuals were modelled from a minimum age of entry into screening of 55 years until death.

The cost perspective was based on adoption by the NHS for screening and early detection of lung cancer, and therefore does not include costs for affected individuals such as revenue, productivity, or out of pocket expenses such as transport costs.

The health perspective focused on direct health effects for affected individuals contacted through the screening programme. Direct health effects on family or carers were not considered in this analysis. Effects of screening on smoking behaviour was also not considered within the context of this model.

4.3.2 Model Details

The model broadly followed the approach established by Snowsill et al.(30). By using the natural history model and population parameters published by Snowsill et al., we aimed to generate cohorts with similar baseline characteristics and comparable results to complement and expand upon their findings with additional screening methodologies. Additional starting characteristics were explored, allowing for an assessment of screening in populations with pre-test cancer risk of 0%-2%, in addition to the 3%-5% previously explored. This is to reflect both the accessibility of the autoantibody test - as a simple blood test it does not require access to CT screening equipment - and the high specificity of the autoantibody test, which is designed to minimise false positive test results, and as such it may be more suitable for testing a larger proportion of people with a smoking history.

Individual patients were generated with randomised age, sex, baseline disease state, and underlying cancer risk. Age and sex were sampled from probability distributions based on participants returning a questionnaire in the UKLS trial(141), with age at study entry truncated between 55 and 80 years, reflecting the widest age range based on eligibility criteria. The baseline disease state was determined by sampling the age of preclinical lung cancer incidence for each individual based on the incidence of lung cancer in England in 2014, adjusted for estimated smoking population(30), and comparing to sampled age at entry, with an age of preclinical incidence lower than age of entry resulting in the patient entering the study with an occult cancer, the stage of which was estimated based on probabilities proportional to the expected time spent in each stage in the absence of screening. Finally, the underlying risk was estimated from a statistical model constructed by Snowsill et al.(30) based on the performance of the LLPv2 risk model in the UKLS trial, which assumed all invited subjects were smokers or ex-smokers, and included coefficients for age, gender, and the presence of cancer in the simulated patient within 3 years of entering the study. The parameter values describing the underlying probability distributions are described in Appendix 4A.

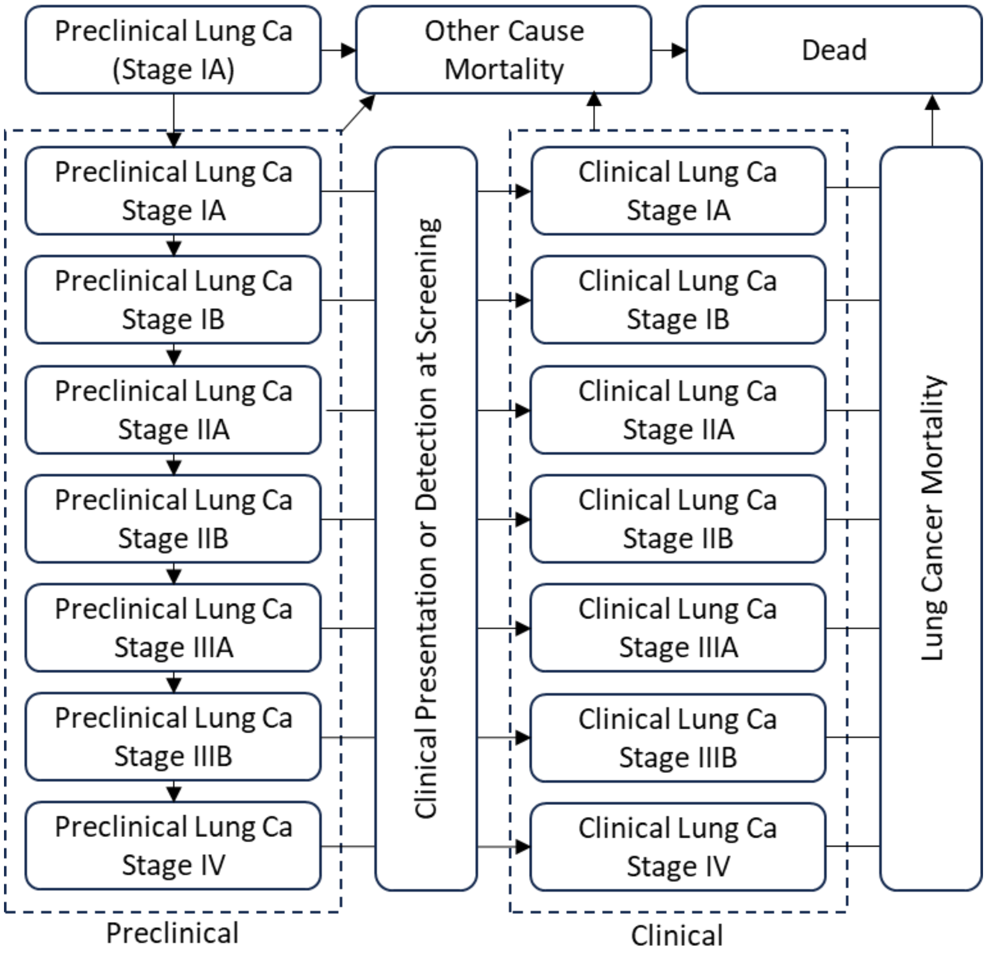

Preclinical Lung Ca (Stage IA)
Other Cause Mortality
Dead
Preclinical Lung Ca Stage IA
Preclinical Lung Ca Stage IB
Preclinical Lung Ca Stage IIA
Preclinical Lung Ca Stage IIB
Preclinical Lung Ca Stage IIIA
Preclinical Lung Ca Stage IIIB
Preclinical Lung Ca Stage IV
Clinical Presentation or Detection at Screening
Clinical Lung Ca Stage IA
Clinical Lung Ca Stage IB
Clinical Lung Ca Stage IIA
Clinical Lung Ca Stage IIB
Clinical Lung Ca Stage IIIA
Clinical Lung Ca Stage IIIB
Clinical Lung Ca Stage IV
Lung Cancer Mortality
Preclinical
Clinical

Figure 4-1: Markov chain diagram showing patient simulation through natural history model of lung cancer.

Screening uptake was estimated for simulated subjects estimated from screening uptake observed in the UKLS trial, and subjects deemed to accept invitation and fulfil age and risk criteria were then progressed through the natural history model developed by Snowsill et al.(30) – as demonstrated in Figure 4-1, which simulated preclinical progression, clinical presentation, and lung cancer survival – based on data from the NLST study(43), as well as other cause mortality (based on interim life tables for England and Wales for the years 2010-2012(30)), in three screening intervention arms – LDCT screening alone, AABT screening alone, and LDCT screening confirmed by AABT screening – as well as a control arm. While it is anticipated that the nature of the EarlyCDT AABT test may lead to higher uptake, as it only requires access to a phlebotomist, and not specialised CT scanning equipment, differences in uptake between LDCT, AABT, and combined AABT+LDCT strategies have not been explored here due to a lack of information about the impact of the improved accessibility on test uptake, and in an effort to present the most conservative comparison between LDCT and AABT. Similarly the capital costs of equipment for analysing AABT has been excluded as the test is designed to be run as an ELISA test on 96 well plates and the majority of NHS diagnostic labs should already have access to all required equipment. Each individual was run through all screening arms and the control arm in order to reduce stochastic variation. This was repeated to generate 100,000 individuals, who then defined a cohort for the purpose of modelling population strategies.

Population strategies were defined in terms of age for entering the screening programme (minimum and maximum), and minimum risk threshold. Simulated individuals meeting the population criteria would undergo screening intervention, while individuals not meeting these criteria would undergo no screening.

Three screening programme designs were compared with no screening in 24 population alternatives representing all combinations of the conditions: minimum age for screening (55 or 60), maximum age for screening (75 or 80), and minimum pre-existing risk for screening (0%, 1%, 2%, 3%, 4%, or 5%) which is assumed to have been calculated prior to screening using the Liverpool Lung Project (version 2) (LLPv2) risk prediction tool(105). This resulted in 72 intervention strategies and one control (no screening) representing current practice.

4.3.3 Health Outcomes

Primary health outcomes measured in this analysis were quality adjusted life years (QALYs) representing health related quality of life (HRQoL), as well as life-years attained under each strategy, compared to no screening. Screening strategies were compared by calculating incremental cost effectiveness ratios which corresponds to the change in cost per quality adjusted life year gained. Costs and QALYs were discounted at 3.5% per year, derived from the UK Treasury discount rate, as is conventional for technology appraisal in England(142).

Secondary health outcomes examined were:

- Diagnosis lead time - calculated as the difference between age at diagnosis in the screening arm compared to the age at diagnosis in the same cohort under no screening and is therefore additional time spent with a known diagnosis of lung cancer that would have remained unknown in the absence of screening.

- Additional lung-cancer survival – life-years gained in in the screening cohort in simulated subjects who develop lung cancer screening compared to the same individuals when they undergo no screening.
- 5-year lung-cancer survival – percentage of diagnosed lung cancers surviving at least 5 years after diagnosis.
- Stage distribution at diagnosis – percentage distribution of cancer stages at presentation.
- Average age at diagnosis, lung cancer mortality (percentage of cohort dying as a results of a lung cancer), and other cause mortality (percentage of cohort dying from non lung-cancer related causes)

4.3.4 Analysis methods

The economic evaluation focuses on a cost-effectiveness analysis, in which the costs and QALYs for each strategy are estimated and a cost-effectiveness frontier constructed by eliminating strategies that are dominated. Cost effectiveness for each alternative is then assessed through calculation of its incremental cost-effectiveness ratio (ICER).

- Main analysis: cost-effectiveness analysis of LDCT and AABT strategies.
- Secondary analysis: sensitivity analysis, using the most cost-effective strategy determined in the main analysis, focused on adjustments to the sensitivity and specificity of the screening test, to determine the theoretical minimum sensitivity and specificity required by the AABT in a screening setting to achieve cost-effectiveness.

- Secondary analysis: deterministic sensitivity analysis, using the most cost-effective strategy determined in the main analysis, and examining the effects of increases and decreases to modelling assumptions and parameters.

4.4 Results

4.4.1 Base case

Seventy-two hypothetical screening programmes were modelled, representing single screening with LDCT which was determined to be the most cost-effective strategy in the analysis by Snowsill et al. single screening with the current commercial EarlyCDT®-Lung autoantibody panel, and single screening with a theoretical improved EarlyCDT®-Lung autoantibody panel, over a number of combinations of age and risk eligibility. A cohort of 100,000 individuals was simulated allowing for a greater level of accuracy than the analysis undertaken by Snowsill et al.

4.4.2 Cost-effectiveness - Current AABT

Examining LDCT and AABT strategies, three of the modelled strategies were on the cost-effectiveness frontier, along with “No Screening” as the least costly and least effective option. These are demonstrated in Figure 4-3 and include AABT screening individuals with minimum 1% pre-test cancer risk, aged between 60 and 75 years old or between 60 and 80 years old. The third and fourth strategies on the cost-effectiveness frontier were CT screening strategies, including subjects between 60 and 80 years old with at least a 1% pre-test risk of cancer, and 55 and 80 years old with at least a 1% pre-test risk of cancer, respectively.

Table 4-1: Proportion of smokers joining and not joining screening:

Population Criteria	% population invited and joined screening	% population invited and declined screening	% population invited but ineligible: Low risk	% population invited but didn't respond to invitation	% population not invited to screening
No Screening	0.0%	0.0%	0.0%	0.0%	100.0%
55-80-0%	14.3%	16.4%	0.0%	69.3%	0.0%
55-80-1%	10.9%	12.6%	7.2%	69.3%	0.0%
55-80-2%	6.5%	7.5%	16.6%	69.3%	0.0%
55-80-3%	4.1%	4.7%	22.0%	69.3%	0.0%
55-80-4%	2.6%	3.0%	25.1%	69.3%	0.0%
55-80-5%	1.8%	2.0%	26.9%	69.3%	0.0%
60-80-0%	10.6%	12.2%	0.0%	51.5%	25.7%
60-80-1%	9.2%	10.5%	3.1%	51.5%	25.7%
60-80-2%	6.1%	7.0%	9.8%	51.5%	25.7%
60-80-3%	3.9%	4.5%	14.4%	51.5%	25.7%
60-80-4%	2.6%	3.0%	17.3%	51.5%	25.7%
60-80-5%	1.7%	2.0%	19.1%	51.5%	25.7%
55-75-0%	13.3%	15.3%	0.0%	64.7%	6.7%
55-75-1%	10.0%	11.5%	7.2%	64.7%	6.7%
55-75-2%	5.7%	6.5%	16.5%	64.7%	6.7%
55-75-3%	3.3%	3.7%	21.7%	64.7%	6.7%
55-75-4%	2.0%	2.2%	24.5%	64.7%	6.7%
55-75-5%	1.2%	1.4%	26.1%	64.7%	6.7%
60-75-0%	9.7%	11.1%	0.0%	46.8%	32.4%
60-75-1%	8.2%	9.4%	3.1%	46.8%	32.4%
60-75-2%	5.2%	5.9%	9.7%	46.8%	32.4%
60-75-3%	3.1%	3.6%	14.1%	46.8%	32.4%
60-75-4%	1.9%	2.2%	16.7%	46.8%	32.4%
60-75-5%	1.2%	1.4%	18.2%	46.8%	32.4%

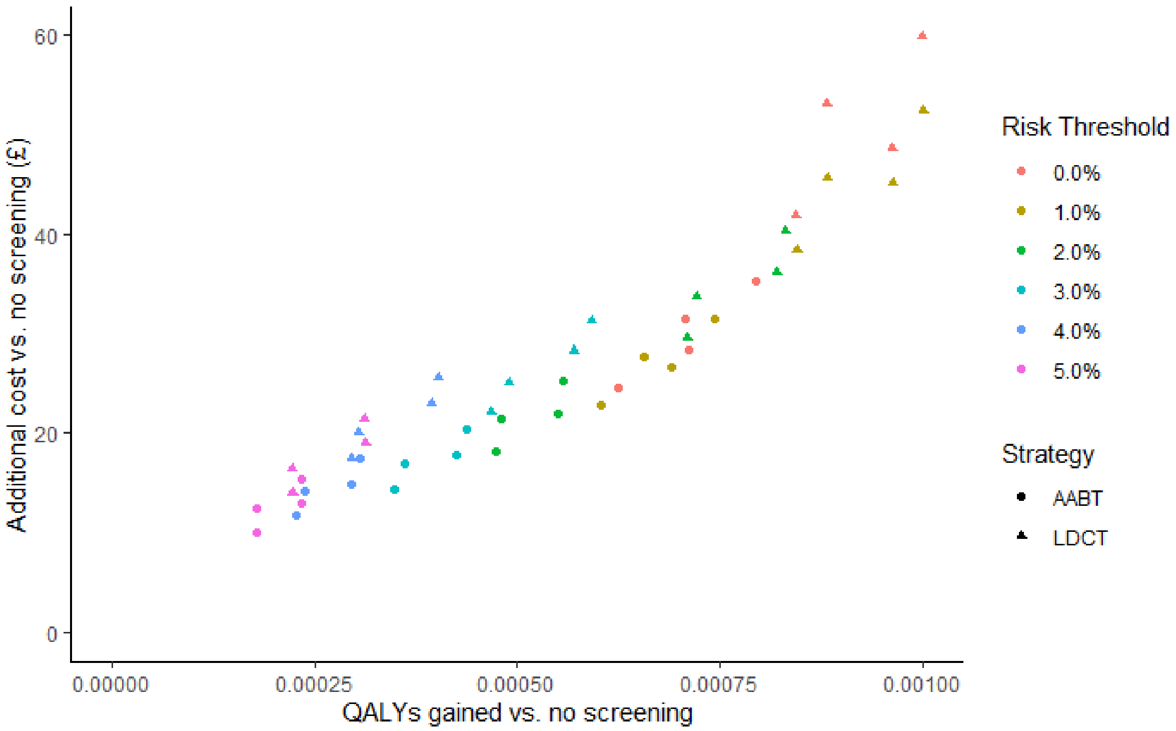


Figure 4-2:Cost-effectiveness plane for base-case results

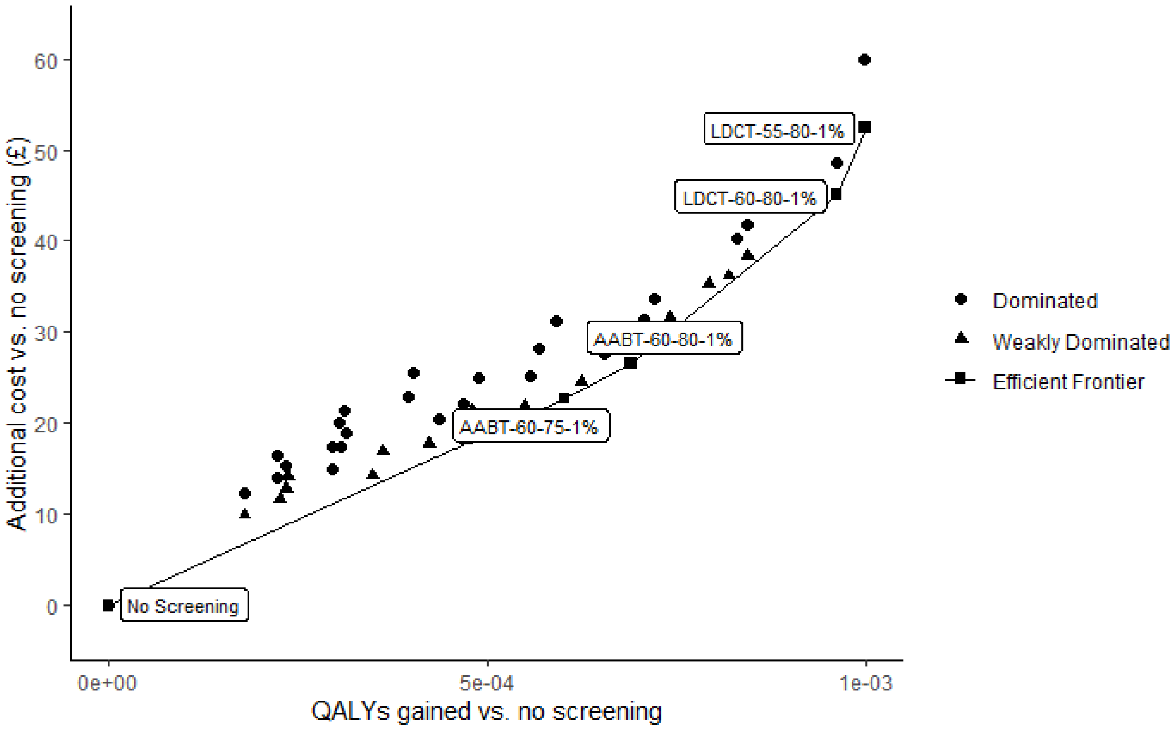


Figure 4-3: Cost-effectiveness frontier for base-case results

Table 4-2: Base-case cost-effectiveness results - LDCT and current performance ECDT strategies on the cost-effectiveness frontier

Strategy	Costs	QALYs	ICER (vs.	Incremental	Incremental	ICER (vs.

	(£)		no screening) (£)	costs (vs. previous)	QALYs (vs. previous)	previous) (£)
No Screening	1028.18	8.520	NA	NA	NA	NA
AABT-60-75-1%	1050.90	8.521	37679.00	22.72	$6.03\text{x}10^{-4}$	37679.00
AABT-60-80-1%	1054.72	8.521	38466.49	3.82	$8.70\text{x}10^{-5}$	43926.66
LDCT-60-80-1%	1073.27	8.521	46851.71	18.55	$2.73\text{x}10^{-4}$	68079.12
LDCT-55-80-1%	1080.59	8.521	52409.19	7.32	$3.76\text{x}10^{-5}$	194793.67

4.4.3 Univariate Sensitivity Analysis – Impact of changes to model parameters

Univariate sensitivity analyses were undertaken by re-running the base case with a single parameter adjusted each time with either a 10% increase, or 10% decrease in the parameter value.

Each re-run simulated the entire cohort of 100,000 individuals, the same cohort was generated for each iteration with only the parameter of concern changing to prevent variation in the generation of cohorts from contributing to, or masking, the effect of varying each parameter. The impact on cost-effectiveness of adjusting the model parameters was assessed through calculation of the incremental net monetary benefit (iNMB) of the screening strategy determined in base case analysis to be the most cost effective - autoantibody test, in individuals between the ages of 60 and 75 with at least a 1% pre-existing risk of lung cancer versus no screening.

The results of this analysis are summarised in Figure 4-4, a tornado plot displaying the effect of varying each of the parameters described in Appendix 4A. In this analysis, the screening test specificity was shown to have a large effect on the cost-effectiveness of the model, as reducing the number of false positive results is able to reduce the costs associated with unnecessary follow-up. In addition to this, it can be seen that the effect of the sensitivity of the screening test is far less influential, this is due to the low incidence of cancer, changes to sensitivity affect the diagnosis in only a small number of the screened individuals whereas changes to specificity affect a vastly higher proportion of individuals.

The largest impact to the cost-effectiveness of the model was changes to the mu_AB parameter, representing the lognormal parameter (location) for pre-clinical incidence of lung cancer. Reducing this parameter leads to a decrease in the mean age at which preclinical lung cancer occurs in the model, and leads to a higher number of simulated individuals presenting with cancer at screening, allowing screening to have a far higher benefit, this would also explain the presence of the sigma_AB parameter high on the tornado plot, as this is also concerned with the age at which pre-clinical lung cancer occurs.

The second largest effect was elicited by changes to the gamma_ocm_F parameter for the other cause mortality distribution in simulated female individuals. Increases to this parameter lead to a reduction in the age at which simulated females in the model experience non-cancer related mortality, reducing the likelihood of a preclinical lesion presenting and the costs associated with such, while decreases to this parameter increase the age at which the simulated individual will experience non-cancer related mortality, increasing the benefit possible from detecting a preclinical lesion at an early stage.

The mean age for simulated individuals also has a large effect on the model, as simulating younger subjects results in a greater number of subjects for whom early diagnosis is able to confer a greater benefit to, as they are less likely to experience non-cancer related mortality prior to the benefit of screening manifesting.

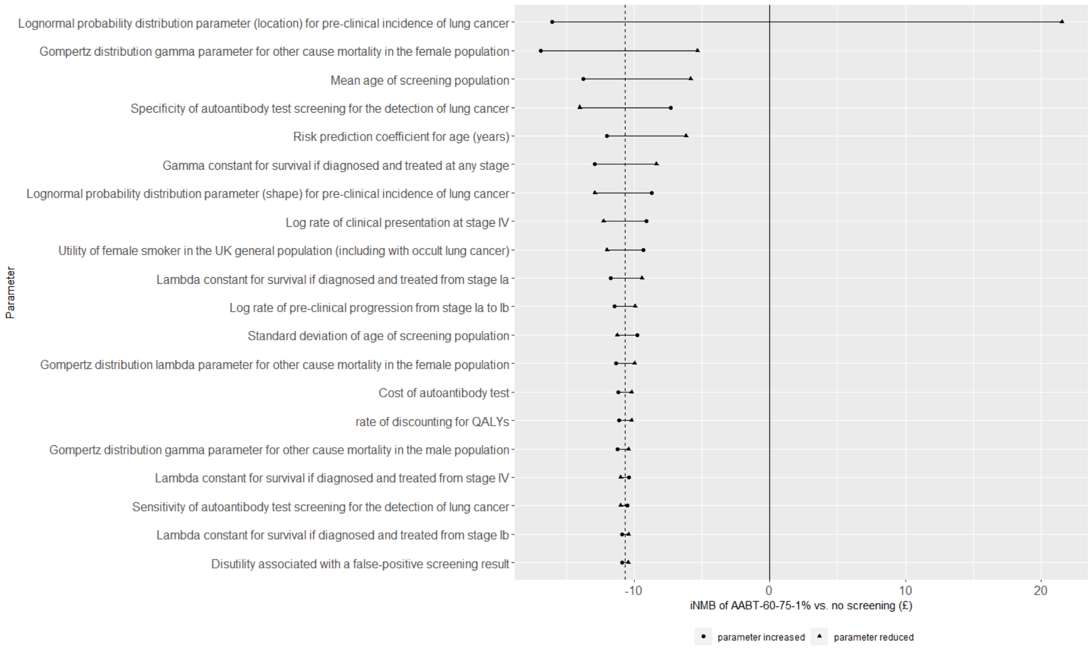


Figure 4-4: Tornado diagram for univariate sensitivity analysis demonstrating effect on incremental net monetary benefit (iNMB) of controlled changes to individual model parameters while maintaining base case values for remaining parameters, in otherwise identical cohorts. 20 parameters with greatest effect on iNMB shown.

Table 4-3: Clinical outcomes for participants of strategies on the cost-effectiveness frontier. Strategies are coded by screening test (autoantibody test (AABT) or low dose CT (LDCT)), minimum screening age (years), maximum screening age (years), and minimum pre-existing risk of cancer.

	Strategy			
	AABT-60-75-1%	AABT-60-80-1%	LDCT-60-80-1%	LDCT-55-80-1%
Per participant				
Number of screens	1	1	1	1
Number of false positives	0.093	0.093	0.348	0.351
Lead time (years)	0.083	0.085	0.151	0.140
Life-years gained	0.014	0.014	0.024	0.022
Additional lung cancer survival (%)	0.124	0.130	0.223	0.196
Additional 5-year lung-cancer survival (%)	7.152	7.404	11.466	12.369
Additional survival time with lung cancer (years)	0.808	0.833	1.352	1.444
Change in age at lung cancer diagnosis	-0.760	-0.761	-1.246	-1.358
Change in age at death from lung cancer	0.016	0.016	0.025	0.027
	Per 100,000 participants			
Proportion of diagnoses arising	20.048	21.558	35.361	38.377

	Strategy			
Per participant	AABT-60-75-1%	AABT-60-80-1%	LDCT-60-80-1%	LDCT-55-80-1%
from screening (%)				
Number of screen-detected cases	2638	2873	4792	5266
Additional lung cancer diagnoses	272	368	661	765
Lung cancer deaths averted	124	130	196	223

4.4.4 Lead time

The lead times for strategies on the cost-effectiveness frontier are summarised in Table 4-3 above. The average lead time for LDCT strategies was 1.44 years, while AABT strategies gave an average lead time of 0.81 years.

4.4.5 Lung cancer survival

For strategies based on the autoantibody test, the average number of lung cancer deaths, per 100,000 participants, was 13,429, compared to 13,337 lung cancer deaths for low dose CT strategies, and 13,607 lung cancer deaths associated with no screening in the same populations.

Across the strategies explored, a reduction in lung cancer mortality of 1.4% to 2.4% was predicted (RR) for individuals participating in LDCT strategies versus no screening, with corresponding reductions of 0.9% to 1.2% for AABT strategies versus no screening.

Additional lung cancer survival for strategies on the cost-effectiveness frontier is summarised in table 3, and shows that for every 100,000 participants in screening, between 124 and 223 lung cancer deaths can be averted, depending on the screening strategy. 5-year survival in the no-screening arm was estimated to be 4.2%, with average 5-year survival associated with screening programmes of 17.7% for LDCT strategies, and 12.1% for AABT strategies.

4.4.6 Stage Distribution

As illustrated in Table 4-4, the use of screening increased the probability of diagnosis at early stage (I and II). The average odds ratios of early diagnosis (geometric mean) by screening compared to no screening were predicted to be 1.95 for LDCT strategies, and 1.52 for AABT strategies. Table 4-4 demonstrates a stage shift at diagnosis with screening that is most evident in increased diagnoses at stage IA and reduced diagnoses at stage IV.

Table 4-4: Average stage distributions for screening programmes

	Lung cancer stage						
Screening Programme	IA	IB	IIA	IIB	IIIA	IIIB	IV
No screening	0.06	0.02	0.02	0.02	0.05	0.03	0.80
LDCT	0.12	0.03	0.03	0.03	0.06	0.03	0.69
AABT	0.09	0.03	0.02	0.02	0.06	0.03	0.75

4.4.7 Age at diagnosis and death

Screening strategies were associated with a reduction in the age at diagnosis of lung cancer. The average age at diagnosis for no-screening, LDCT, and AABT was 75.9, 74.4, and 75.1 respectively.

Screening strategies were also associated with a moderate increase in the age of death from lung cancer, with average age of death for no-screening, LDCT, and AABT calculated at 76.38, 76.44, and 76.74 years respectively.

Average age at other cause mortality was also increased in the screening arms. Although this effect was minimal and may reflect some older participants dying from other causes after screening and treatment in the screening arm, where they would be included in the lung cancer mortality in the non-screening arm.

4.4.8 Costs

The cost per participant of AABT screening is estimated at around £66 compared to £105 for LDCT screening, with costs for confirmatory testing assumed to be comparable between the different testing methods. The costs for the strategies on the cost-effectiveness frontier are summarised in table 5 and show that lung cancer costs, excluding end of life costs, are higher for LDCT screening, most likely due to the costs incurred by the higher incidence of false positive results, which is reduced in AABT strategies due to the high specificities associated with AABT testing.

The costs for the screening programmes are also summarised in table 5, and show predicted costs increases of £1,506M to 3,938M for screening of a population of 13million smokers aged 55-80 years.

Table 4-5: Costs for programmes on the cost-effectiveness frontier. Strategies are coded by screening test (autoantibody test (AABT) or low dose CT (LDCT)), minimum screening age (years), maximum screening age (years), and minimum pre-existing risk of cancer.

Costs			Strategy		
	No screening	AABT-60-75-1%	AABT-60-80-1%	LDCT-60-80-1%	LDCT-55-80-1%
Costs for each participant (£)					
Screening Test		66.06	66.06	104.75	104.75
Lung cancer costs (excluding end of life)		869.69	869.58	1064.54	1017.54
End of life		395.67	403.08	399.21	389.91
Total cost		1331.42	1365.72	1568.50	1512.20
Population of 13 million smokers aged 55-80 years (lifetime costs, £M)					
Screening administration					
Screening Test	0.0	493.1	550.1	872.2	1042.4
Lung cancer costs (excluding end of life)	8666.9	9718.5	9960.0	11358.5	11644.6
End of life	4699.4	4660.7	4654.6	4622.4	4618.2
Total cost	13366.4	14872.3	15164.7	16853.1	17305.2
Additional cost vs. no screening		1506.0	1798.3	3486.7	3938.8

4.4.9 Sensitivity Analysis - Effect of increasing the sensitivity and specificity of the AAb test

The previous analysis demonstrates that the AAb test is more cost effective than low dose CT scanning as a screening modality, and that the greatest benefit can be obtained through screening with the autoantibody test in at-risk individuals between 60 and 75 with at least a 1% pre-existing risk of lung cancer. In this analysis, the most cost-effective strategy was above the NICE guideline threshold of £20,000 per QALY. This strategy was also above the less conservative threshold of £30,000 per QALY, therefore a sensitivity analysis was undertaken to explore the sensitivity and specificity characteristics required for cost effectiveness for the AABT test, assuming all other parameters remained constant. The original cohort of 100,000 simulated subjects was modelled using the strategy previously determined to be the most cost effective (smokers between 60 and 75 with at least 1% pre-existing risk of lung cancer) and repeated through sensitivities of 20% to 80% for specificities of 90% to 99%. These results are illustrated in Figure 4-5, and demonstrate that increases to the sensitivity and specificity of the AAb test would lead to it becoming cost-effective at a threshold of £30,000 per QALY. The analysis also demonstrated that even at the highest levels of specificity and sensitivity, the test did not show cost effectiveness at a threshold of £20,000 per QALY.

Regression modelling was undertaken on the cost-effectiveness curves for each specificity, in order to calculate the sensitivity required for the test to be considered cost-effective at a threshold of £30,000 per QALY, as shown in Figure 4-6.

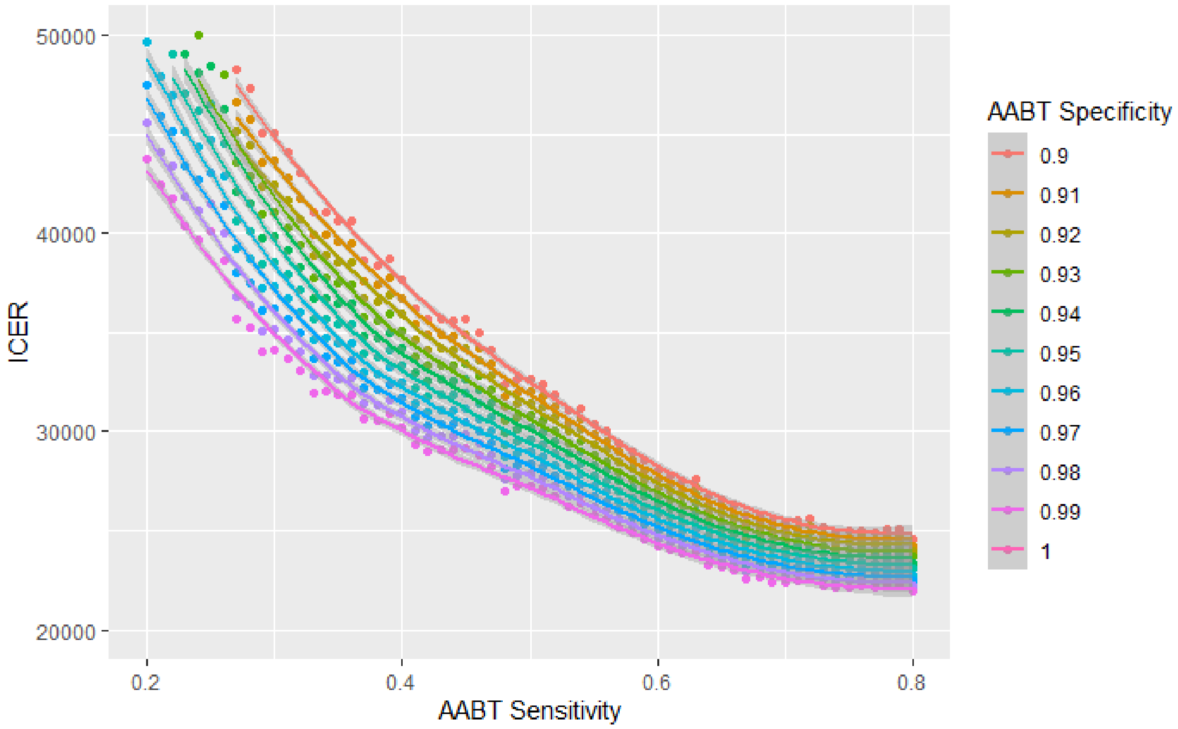


Figure 4-5: Exploration of autoantibody test (AABT) performance required to achieve cost-effectiveness.

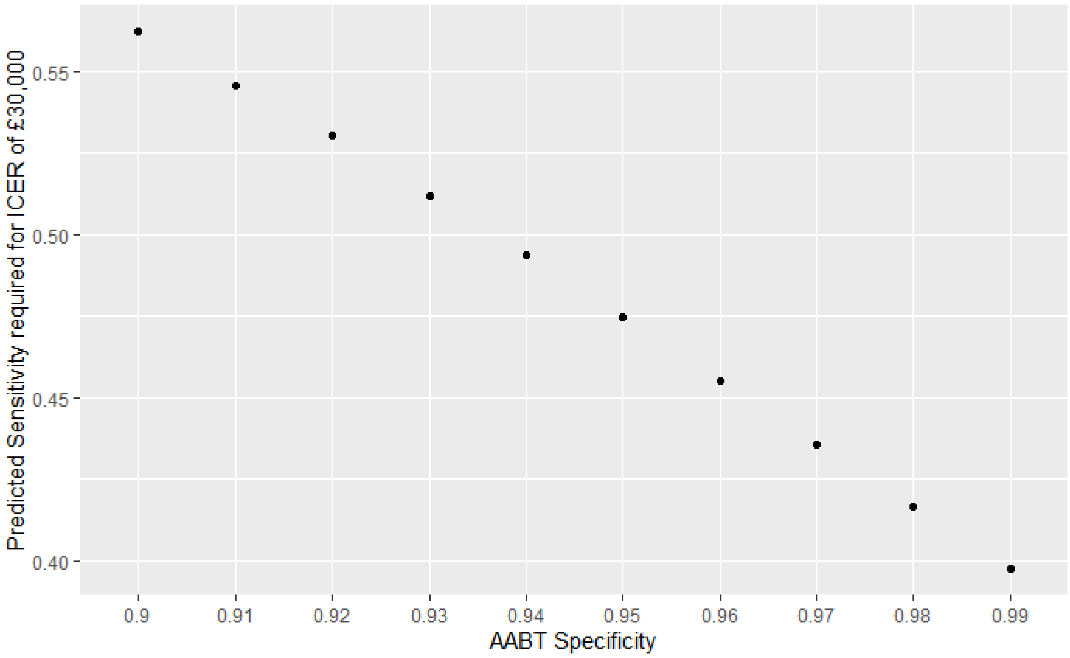


Figure 4-6: Predicted autoantibody test (AABT) sensitivity levels required at each specificity for cost effectiveness at a threshold of £30,000 per quality adjusted life-year (QALY)

4.5 Chapter Conclusions

Introduction of lung cancer screening is predicted to result in earlier diagnosis, with a stage shift at diagnosis, as has already been demonstrated in the ECLS trial(109), leading to better prognosis and increased survival for participants. The introduction of a screening programme to detect asymptomatic lung cancer does, however, result in additional costs to the NHS. In this analysis, the most cost-effective strategies for LDCT and AABT would not be considered cost-effective at the NICE recommended threshold of £20,000 per QALY(143), with predicted costs of £41,705, and £37,679 per QALY gained respectively, however, an exploration of the effect of increasing the sensitivity and specificity of the autoantibody test showed that the test becomes cost-effective at a threshold of £30,000 per QALY threshold, at sensitivities of 39.8% at 99% specificity, 47.5% at 95% specificity, or 56.2% at 90% specificity respectively.

Increases to the performance of the EarlyCDT®-Lung test could be brought about through a variety of methods, including expanding the autoantibody panel with additional autoantibodies that have shown high specificity discriminatory potential in lung cancer, many of which have been described in studies, such as autoantibodies to Dickkopf-related protein 1 (DKK1)(144), Cyclin B1, and Survivin(145), or the addition of complementary antigenic biomarkers to the panel, such as neuron-specific enolase (NSE), carcinoembryonic antigen (CEA), CYFRA 21-1, squamous cell carcinoma–associated antigen (SCC), CA15.3, and pro–gastrin-releasing peptide (ProGRP)(146). Additionally, the pilot study outlined in Chapter 2 established the possibility of using metrics derived from the antigen-autoantibody binding curve to selectively reduce false positive (type I) errors and established that further specificity may be derived from scrutinising the binding characteristics of tumour associated autoantibodies.

The cost effectiveness of the test may also be affected by exploring additional screening strategies. The analysis undertaken here looks only at a single screening round, therefore only pre-existing malignancies are detected, and those which develop after the screening time-point are not detected. The accessibility of the EarlyCDT®-Lung test, with a lack of a need for CT scanning equipment and its associated radiation exposure, increases the potential for repeated screening strategies. It has been shown that autoantibodies are detectable up to 4 years before current clinical diagnosis(147), therefore repeated screening every 3-5 years may allow for a greater number of malignancies to be detected. Although as the univariate sensitivity analysis demonstrated that the QALY estimates derived from this model are dependent on the model parameters, and changes to the underlying risk and age group examined have a large effect on the health economic benefits, subsequent screening rounds would be expected to have a lower risk, due to the removal of prevalent malignancies in the initial screening round, therefore repeat screening strategies would not be expected to lead to improved cost effectiveness.

Future work would also benefit from better understanding the differences in screening uptake between a blood test and LDCT based screening, as well as how this uptake would be affected by combination strategies. While increased test uptake was included in the univariate sensitivity analysis, the compounded effect of both low response to initial questionnaires (30.7%) and relatively low uptake (46.5%) potentially reduced the impact of 10 percent increases to response and uptake, and an improved understanding on the benefits of any improved uptake attributable to a screening blood test over LDCT imaging would lead to a more accurate assessment of the benefits that could be brought about through adoption of the EarlyCDT autoantibody test.

The Snowsill health economic model that this analysis is based on includes a well calibrated natural history model for lung cancer, trained on NLST data, which allows for appreciation of stage progression, and calculation of stage shift in diagnosis between screening methods, however the model does not currently consider presence of a latent cancer which is not yet detectable by CT. A future refinement of this model could include an appreciation of latent cancer presence to enable more accurate assessment of autoantibody screening accuracy.

The model described here was able to improve upon the model reported by Snowsill et al., through refinement to the deterministic sensitivity analysis. Unlike the Snowsill model, which is run in Microsoft Excel, the model trained here was created in the R programming language, which allowed for greater control and adjustment of simulated populations. In the original model, during deterministic sensitivity analysis a new sub-population would be generated each time, with a single parameter adjusted. Due to the variability in the generated populations, differences due to random variation in variables other than the variable of interest had a large influence on the results. In the R model described here, the simulated population was able to be kept identical, with only the variable of interest being changed, and health costs and QALYs generated were then directly comparable leading to much higher accuracy in the analysis.

4.6 Chapter Discussion

Assessment of the health economic benefits associated with an autoantibody screening test that exhibits high specificity showed that even at its current diagnostic performance, the EarlyCDT®-Lung autoantibody test shows a lower cost per QALY than low-dose CT for lung cancer screening, with improvements to the specificity and sensitivity of the test having the potential to bring it within the NICE cost-effectiveness guideline of £30,000 per QALY. While future analyses would benefit from a greater investigation into whether the accessibility of a blood test would result in improved screening test uptake, and a better understanding of how that improved uptake would influence the health economics, as well as whether repeat screening strategies, allowing longitudinal monitoring of autoantibodies, could be cost-effective, these are outside of the scope of this project. This analysis has identified target diagnostic performance values which will instruct the subsequent machine learning analyses.

Chapter 5: Data Collection

5.1 Aims

The pilot study showed that curve characteristics showed potential for improving the diagnostic performance of the test, and the health economic analysis showed that improvements to both sensitivity and specificity are required to increase the health economic benefit of the test. In order to achieve these improvements curve characteristic data along with additional autoantibody features was collected on datasets that had been obtained for EarlyCDT development studies, and these datasets are described here.

5.2 Introduction

To explore the potential utility of curve characteristics as biomarkers, as well as the performance of an expanded autoantibody biomarker panel, a series of case-control studies were conducted. The structure of the cohorts included in these studies, the details of data collection using the EarlyCDT®-Lung test, and the methods of calculation of the curve characteristic features are outlined here. The initial cohort was separated into training and hold-out test sets, with model training being completed on the training set, and model performance being assessed on the test set and subsequent validation sets to try and reduce overfitting of models.

5.3 Methods

5.3.1 Study Sample Cohorts

All of the samples used in these studies were obtained as part of a series of case-control studies from a combination of sample banks in the UK, the US and Ukraine, along with healthy normal samples collected in the UK as part of a population autoantibody study (PAS), and samples provided as part of a collaborative study with a group in Cleveland. All cancer samples were obtained after diagnosis but prior to any anticancer treatment.

Sample Sources

NHS Biobank

Samples obtained from subjects with known lung cancer were procured from the Nottingham NHS Biobank, these samples were collected between 2011 and 2013 from subjects attending Nottingham University Hospitals in the UK.

Asterand

Samples obtained from subjects with known lung cancer were purchased from the Asterand biobank, these samples were collected between 2007 and 2013 (with the exception of one sample collected in 2005) from subjects in Eastern Europe (Romania, Ukraine, Russia, Bulgaria, and Moldova) and the USA.

Kiev

Samples obtained from subjects with known lung cancer, along with healthy normal subjects with no personal history of cancer, were collected from subjects in the Ukraine. These samples were collected between 2008 and 2010.

Nottingham PM

Samples were obtained from subjects with known small-cell lung cancer as part of a collaboration exploring autoantibodies in patients presenting with

Lamber-Eaton myasthenic syndrome to a clinic in the UK. Remaining samples from this study were included in validation studies for the EarlyCDT-test to provide data on SCLC subjects. These samples were collected between 2006 and 2012.

PAS Collection

Serum samples were collected from healthy subjects with no personal history of cancer or autoimmune disease as part of a study into autoantibody profiles in the UK general population. These samples were collected between 2007 and 2010.

Springfield

Serum samples from healthy subjects with no personal history of cancer were purchased from a US biobank for use as a control cohort in EarlyCDT Lung validation studies. These samples were collected in 2012.

Cleveland

Serum samples provided as part of a collaborative study with Peter Mazzone of the Cleveland Clinic, OH, USA. These samples included both known lung cancer samples, obtained at point of diagnosis before commencement of treatment, and healthy control samples obtained from subjects with no personal history of cancer. These samples were collected between 2011 and 2014.

Training/Test Cohort

The training/test cohort was comprised of 335 cancer samples from the NHS Biobank, Asterand, Nottingham PM, and Kiev collections, and 331 healthy control samples obtained from the PAS and Springfield collections. Samples were assayed in 2014 and 2015 on the EarlyCDT®-Lung format ELISA, with samples run over 6 days, along with 2 plates of additional autoantibody features, each run over 4 days. Cancer samples were matched as closely as possible to healthy normal samples by age, smoking-history, and gender, with first priority being to match for age within 3 years, followed by smoking history, and finally by gender

Validation 1

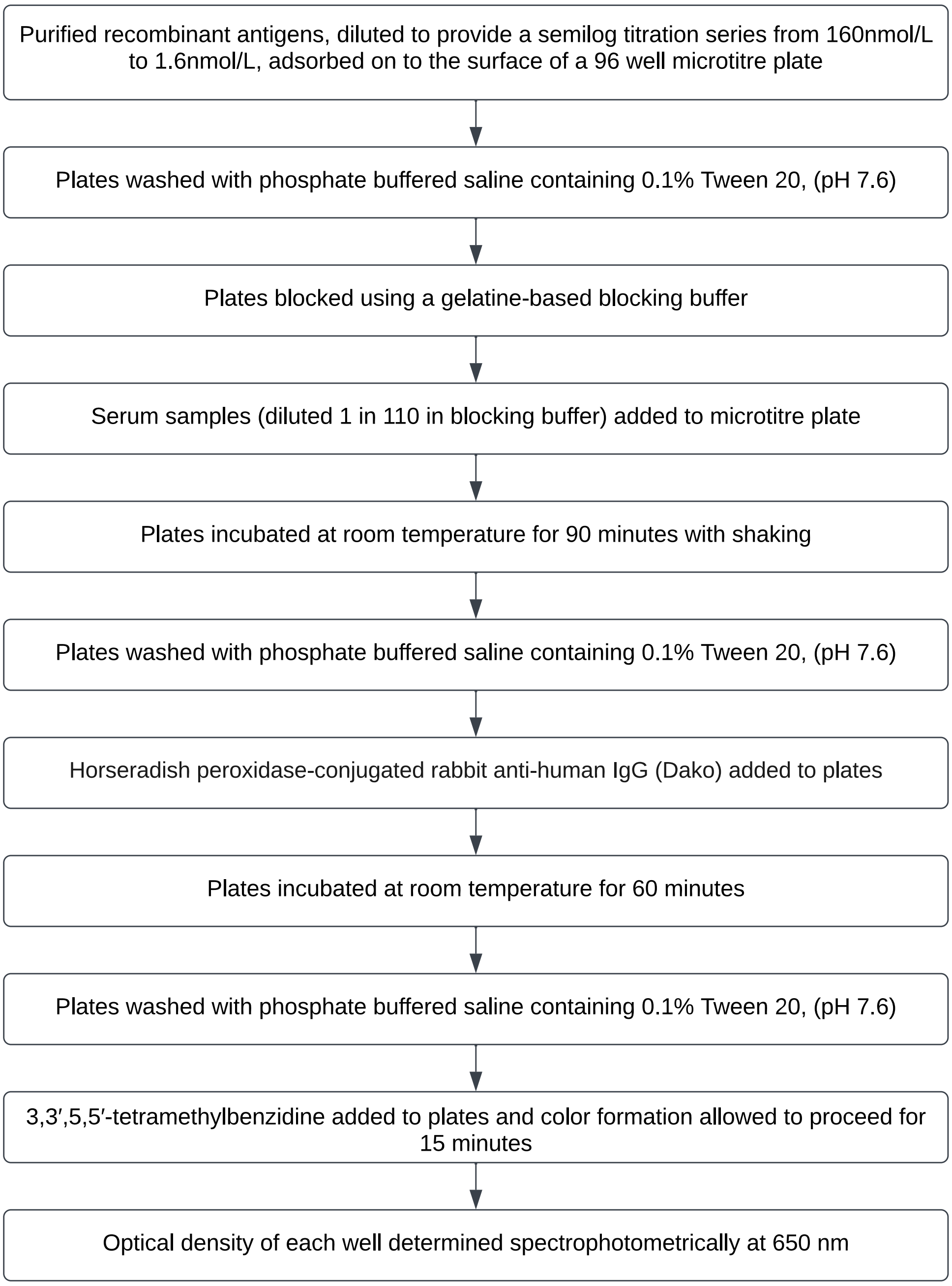
The validation 1 cohort was comprised of 93 cancer samples from the Asterand sample bank, and 96 healthy control samples obtained during the UK PAS collection, assayed in 2015 on the EarlyCDT®-Lung format ELISA, with samples run over 1 days, along with 2 plates of additional autoantibody features, each run over 1 days. Cancer samples were matched as closely as possible to healthy normal samples by age, smoking-history, and gender, with first priority being to match for age within 3 years, followed by smoking history, and finally by gender

Validation 2

The validation 2 cohort was comprised of samples provided through an academic collaboration with Dr Peter Mazzone and the Cleveland Clinic Foundation. These samples were received and assayed as blinded samples, with subsequent demographic unblinding after reciprocal submission of assessed data. These samples were assayed in 2015 on the EarlyCDT®-Lung format ELISA, with samples run over 3 days, along with 2 plates of additional autoantibody features, each run over 3 days also. After transfer of laboratory data to collaborators in Cleveland, samples were unblinded and were determined to be comprised of 215 cancer samples, and 321 healthy control samples.

5.3.2 Autoantibody Assays

Autoantibody levels were determined through the use of a semi-automated direct ELISA in which samples were reacted with a titration series of antigen concentrations. This assay technique is described fully in Lam et al(107), and is summarised in Figure 5-1 below:



Purified recombinant antigens, diluted to provide a semilog titration series from 160nmol/L to 1.6nmol/L, adsorbed on to the surface of a 96 well microtitre plate

Plates washed with phosphate buffered saline containing 0.1% Tween 20, (pH 7.6)

Plates blocked using a gelatine-based blocking buffer

Serum samples (diluted 1 in 110 in blocking buffer) added to microtitre plate

Plates incubated at room temperature for 90 minutes with shaking

Plates washed with phosphate buffered saline containing 0.1% Tween 20, (pH 7.6)

Horseradish peroxidase-conjugated rabbit anti-human IgG (Dako) added to plates

Plates incubated at room temperature for 60 minutes

Plates washed with phosphate buffered saline containing 0.1% Tween 20, (pH 7.6)

3,3′,5,5′-tetramethylbenzidine added to plates and color formation allowed to proceed for 15 minutes

Optical density of each well determined spectrophotometrically at 650 nm

Figure 5-1: Autoantibody ELISA process

5.3.3 Correction for Non-Specific Binding

The EarlyCDT®-Lung test was designed with the inclusion of a non-specific protein tag referred to as ‘BirA’ or ‘VOL’ on each antigen protein. This protein is also included as a control on each plate, and the optical density (OD) returned by the VOL control for each sample is subtracted from the raw OD to account for non-specific binding present in the patient sample, giving a value referred to subsequently as a VOL Corrected OD (VCOD). During initial investigations, it was suggested that re-expressing the raw OD signal as a ratio of specific to non-specific binding may give a more accurate assessment of a subjects immune response, and to investigate whether this is the case, an alternate non-specific binding correction method has been investigated by dividing the antigen specific OD signal by the signal returned by the VOL control antigen. This correction method is subsequently referred to as the Signal To Vol Ratio (STVR).

5.3.4 Calculation of Curve Characteristic Features

The curve characteristic features to be investigated in the subsequent analyses were calculated using custom R functions within RStudio(v2022.07.02) to generate the following features for raw OD, VCOD, and STVR data:

Slope and Intercept

Slope and Intercept features were calculated using the ‘lm’ function from the R ‘stats’ package to fit a linear regression of the OD (raw or corrected) with the formula $y = mx + b$ against the plated antigen concentration using quantile regression, returning values for both the slope m and intercept b of the subsequent fitted linear regression line.

Area Under the Curve (AUC)

AUC features were calculated using the ‘trapz’ function from the R ‘caTools’ package to derive the area under the titration curve using trapezoid rule integration whereby the area of the region under the curve is approximated through the summation of the areas of the trapezoids constructed by adjacent points.

SlopeMax

The SlopeMax feature was defined as the slope ($\frac{dy}{dx}$) at the steepest point of the titration curve, and was calculated through a custom R function which assessed slope values for all adjacent sets of points on the curve, and returned the greatest value.

5.4 Results

5.4.1 Study Sample Cohorts

The demographics data for the full dataset is summarised in Table 5-1, with cancer subtype and stage distribution as shown in

Table 5-2. In addition, two further independent validation cohorts of samples have been assessed on all features, and resultant models have been applied to these cohorts as a further means of assessing model performance reproducibility. The demographic details of these cohorts are summarised in Table 5-1 to Table 5-6

Table 5-1: Training/Test Cohort demographic summary

Characteristic	Cancer, N = 335[1]	Matched Normal, N = 331[1]
Age	64 (57, 71)	64 (57, 72)
Gender		
Female	95 (29%)	143 (43%)
Male	230 (71%)	186 (57%)
Unknown	10	2
Smoking History		
Current smoker	132 (45%)	105 (35%)
Ex-smoker	118 (41%)	120 (40%)
Non-smoker	41 (14%)	72 (24%)
Unknown	44	34
Sample Source		
NHS Biobank	123 (37%)	0 (0%)
Asterand	177 (53%)	0 (0%)
Kiev	19 (5.7%)	4 (1.2%)
Nottingham PM	16 (4.8%)	0 (0%)
PAS Collection	0 (0%)	173 (52%)
Springfield	0 (0%)	154 (47%)
Sample Origin		
UK	139 (42%)	173 (52%)
USA	14 (4.2%)	154 (47%)
Ukraine	40 (12%)	4 (1.2%)
Romania	98 (30%)	0 (0%)
Russia	40 (12%)	0 (0%)
Unknown	4	0

[1] n (%); Median (IQR)

Table 5-2: Training/Test Cancer sample Subtype and Stage summary

Characteristic	N = 335[1]
Subtype	
Adenocarcinoma	133 (40%)
Squamous cell carcinoma	134 (40%)
Adenosquamous carcinoma	6 (1.8%)
Carcinoid	7 (2.1%)
Leiomyosarcoma	2 (0.6%)
NSCLC	1 (0.3%)
SCLC	35 (11%)
Other	15 (4.5%)
Unknown	2
Stage	
IA	52 (25%)
IB	65 (31%)
IIA	28 (13%)
IIB	28 (13%)
IIIA	31 (15%)
IIIB	3 (1.4%)
IV	2 (1.0%)
Unknown	12

[1] n (%)

Table 5-3: Validation Cohort 1 demographic summary

Characteristic	Cancer, N = 93[1]	Matched Normal, N = 96[1]
Age	61 (54, 67)	60 (54, 67)
Gender		
Female	26 (28%)	25 (26%)
Male	67 (72%)	71 (74%)
Smoking History		
Current smoker	40 (48%)	38 (40%)
Ex-smoker	31 (37%)	35 (36%)
Non-smoker	12 (14%)	23 (24%)
Unknown	10	0
Sample Source		
Asterand	93 (100%)	0 (0%)
PAS Collection	0 (0%)	96 (100%)
Sample Origin		
Ukraine	49 (53%)	0 (0%)
Romania	30 (32%)	0 (0%)
Russia	4 (4.3%)	0 (0%)
Moldova	2 (2.2%)	0 (0%)
Bulgaria	1 (1.1%)	0 (0%)
USA	7 (7.5%)	0 (0%)
UK	0 (0%)	96 (100%)

[1] n (%); Median (IQR)

Table 5-4: Validation Cohort 1 Cancer sample Subtype and Stage summary

Characteristic	N = 93[1]
Subtype	
Adenocarcinoma	25 (28%)
Squamous cell carcinoma	48 (54%)
Adenosquamous carcinoma	2 (2.2%)
Large Cell	7 (7.9%)
NSCLC	1 (1.1%)
SCLC	4 (4.5%)
Unknown	4
Stage	
IA	17 (18%)
IB	21 (23%)
IIA	14 (15%)
IIB	12 (13%)
III	3 (3.2%)
IIIA	18 (19%)
IIIB	6 (6.5%)
IV	2 (2.2%)

[1] n (%)

Table 5-5: Validation Cohort 2 demographic summary

Characteristic	Cancer, N = 215[1]	Matched Normal, N = 321[1]
Age	68 (61, 74)	66 (60, 70)
Gender		
Female	89 (41%)	159 (50%)
Male	126 (59%)	162 (50%)
Smoking History		
Current smoker	29 (13%)	114 (36%)
Ex-smoker	173 (80%)	194 (61%)
Non-smoker	13 (6.0%)	10 (3.1%)
Unknown	0	3
Sample Source		
Cleveland	215 (100%)	321 (100%)
Sample Origin		
USA	215 (100%)	321 (100%)

[1] n (%); Median (IQR)

Table 5-6: Validation Cohort 2 Cancer sample Subtype and Stage summary

Characteristic	N = 215[1]
Subtype	
Adenocarcinoma	111 (52%)
Squamous cell carcinoma	67 (31%)
Adenosquamous carcinoma	3 (1.4%)
Large Cell	1 (0.5%)
Neuroendocrine	1 (0.5%)
NSCLC	13 (6.1%)
SCLC	18 (31%)
Unknown	1
Stage	
I	3 (1.4%)
IA	40 (19%)
IB	24 (11%)
IIA	11 (5.1%)
IIB	16 (7.4%)
IIB/IV	1 (0.5%)
IIIA	61 (28%)
IIIB	14 (6.5%)
IV	45 (21%)

[1] n (%)

Cohort partitioning was carried out to split the data into training and test cohorts, with 70% of the data (235 cancer cases, 232 normal controls) used for training, and the remaining 30% (100 cancer cases, 99 normal controls) being used as a hold-out test set for confirming model performance. A split of 70:30 is common practice in machine learning experiments, and has been proven previously to give the best performance for training models(148). The commercial cutoff thresholds for the EarlyCDT®-Lung test have been applied to these cohorts, and diagnostic performance calculated and been summarised in Table 5-7.

Table 5-7: Summary of Commercial performance in study cohorts

Cohort	TP	FN	FP	TN	Sens	Spec
Training	72	163	23	209	30.6%	90.1%
Test	23	77	12	87	23.0%	87.8%
Validation 1	19	74	3	93	20.4%	96.9%
Validation 2	74	134	38	271	35.6%	87.7%

Model parameter optimisation was carried out on the training cohort, using 10-fold cross-validation during parameter optimisation to reduce the impact of overfitting on model optimisation.

5.5 Chapter Conclusions

The datasets run show differences in the composition of both subtype and staging which may contribute to the differences in commercial performance between the four cohorts. While the training cohort showed equal representation of adenocarcinoma and squamous cell carcinoma, along with 11% SCLC which is comparable to expected population prevalence, the Validation 1 cohort showed a higher proportion of squamous cell carcinoma and SCLC was underrepresented at only 4.5%. Contrary to this the Validation 2 cohort had a higher proportion of adenocarcinoma and SCLC was overrepresented comprising 31% of the cohort. Additionally, there were large differences in the stage distributions, with the training cohort being comprised of an extremely high proportion of early-stage disease (82%) compared to 69% early-stage in the validation 1 cohort, and only 43.9% in the validation 2 cohort. Future studies would benefit from greater efforts to ensure that datasets are balanced, and subtype and stage distributions are reflective of the target population.

5.6 Chapter Discussion

Datasets were sourced by the company without full consideration of the effect of confounding variables in the datasets. The training/test cohort and validation 1 cohort resultantly showed large imbalance between cases and controls regarding the country of origin of the samples, which may contribute to poor reproducibility of trained models, in addition to imbalance observed between the histological subtypes represented between cohorts. Unfortunately, I had little influence in the sourcing of samples for these studies, as the purchasing of samples was done without consultation as to their suitability for a machine learning project.

Chapter 6: Data Transformations and Unsupervised Analysis

6.1 Aims

Initial exploration of the data applying rules developed in the pilot study was unable to reproduce the performance improvements illustrated in that study. As availability of tools and techniques had advanced in the time between the initial pilot study and the analysis of the cohorts described previously, I decided to explore the data in much greater depth using machine learning strategies in an attempt to develop models incorporating the curve characteristics that are able to improve the performance and health economic benefits of the EarlyCDT®-Lung test. Initially this required assessment of the data distributions to identify whether transformations were necessary to approximate normal distributions and potentially reduce inter-assay variability, along with an initial focus on exploration of unsupervised strategies, as I wished to see whether the underlying patterns in autoantibody reactivity are able to separate or classify the data, and potentially highlight relationships that may be missed in supervised machine learning analyses.

6.2 Introduction

With the ultimate goal of exploring a variety of modelling techniques and machine learning strategies to develop models which are able to contribute to early cancer diagnostics, an investigation was undertaken to explore and characterise the distributions of the biomarker and titration curve derived features. This was necessary as a number of quantitative modelling techniques, such as regression models and linear discriminant analysis, assume that data being modelled follow a bivariate or multivariate normal distribution, and therefore ensuring that features are normally distributed - or applying a transformation that allows the data to approximate the normal distribution - may lead to improved model fitting and better performance in resultant models.

To ensure that the generated features are suitable for parametric modelling techniques, the following analysis will confirm that the features approximate a normal distribution, as deviations from normality such as skewness and kurtosis can lead to inaccuracies in data modelling and reduce the generalizability of trained models. Features which do not approximate normality will undergo additional transformation prior to inclusion in modelling.

6.3 Methods

6.3.1 Exploration of Distributions

Signal distribution was visualised for each feature as a histogram, along with an overlaid normal distribution function using the mean and standard deviation of the featureset, and quantile-quantile plots were also constructed to compare the distribution of the data to the theoretical normal distribution (data not shown).

To statistically assess the distributions, a series of tests for normality were undertaken using the ‘stat.desc’ function from the R ‘pastecs’ package, these tested for skewness, kurtosis, and normality.

Skewness is a measure of the asymmetry of a dataset, and was assessed through calculation of the skewness coefficient, in which a value of 0 is indicative of symmetrical data, with positive values indicating a greater proportion of data is lower than the mean, and negative values indicating that the majority of the data is above the mean.

Kurtosis is a measure of how much of the data is distributed at extreme values of the distribution, with higher values for the kurtosis (referred to as leptokurtic) indicating a higher degree of data at the edges of the distribution and potentially higher quantities of outliers. Lower values for the kurtosis (referred to as platykurtic) indicates a greater amount of the data is close to the distribution centre.

The Shapiro-Wilk (SW) test was used to test for normality in the datasets, and has been previously recommended in the case of suspected asymmetric data(149). SW measures the differences between the measured data and data that would be expected based on the mean and standard deviation of the dataset, to generate a test statistic W between 0 and 1, with 1 indicating perfect conformity to a normal distribution. Shapiro-Wilk also produces an associated p-value for the null hypothesis that the data is normally distributed, with values <0.05 indicating that the data is significantly not normally distributed.

6.3.2 Transformation Techniques

The data for each feature underwent transformation through the following strategies: log, square root, exponential, arcsinh, Box Cox, and Yeo-Johnson, according to the following strategies:

log transformation:

$$g(x) \ = \ log(x + a) \text{ where } a \ = \ max(0, -\min(x) + e)$$

square root:

$$g(x) = \ \sqrt{x + a} \text{ where } a \ = \ max(0, -\min(x) + e)$$

exponential:

$$g(x) = exp(x)$$

arcsinh:

$$g(x) = \log(x + \sqrt{x^2 + 1})$$

Box Cox:

As proposed by Box and Cox in 1964(150), whereby lambda is estimated by maximum likelihood.

$$g(x;\lambda) = 1_{(\lambda\neq 0)}\frac{x\lambda - 1}{\lambda} + 1_{(\lambda=0)}logx$$

Yeo-Johnson:

As proposed by Yeo-Johnson in 2000(151), whereby the value of lambda is found which minimizes the Kullback-Leibler distance between the normal and transformed distributions through estimation by maximum likelihood.

$$g(x;\lambda) = 1_{(\lambda\neq 0, x\geq 0)}\frac{(x+1)^{\lambda} - 1}{\lambda}$$

$$+ 1_{(\lambda=0, x\geq 0)}\log(x+1)$$

$$+ 1_{(\lambda\neq 2, x<0)}\frac{(1-x)^{2-\lambda} - 1}{\lambda - 2}$$

$$+ 1_{(\lambda=2, x<0)} - \log(1-x)$$

These strategies were selected for investigation due to their relative simplicity and transferability to new datasets.

The transformed features then each had their distribution assessed according to a goodness of fit test based on calculation of the Pearson P statistic – divided by its degrees of freedom. This assessment of normality is the default for the bestNormalise package due to its interpretability, the ratio is comparable between different transformations representing an absolute measure of the departure from normality, with ratio values close to 1 signifying proximity to a normal distribution.

6.3.3 Principal Components Analysis

To explore the structure of the data, and assess relationships between features, principal component analyses(152) were undertaken, initially on the unadjusted OD data from the commercial panel antigens at all concentrations, and subsequently on data corrected to compensate for the effect of non-specific binding, before assessing the curve characteristic features.

Principal components analysis (PCA) helps identify underlying structure and relationships within the data and can provide insights into the most important factors driving the variation in the dataset. It is a dimensionality reduction technique that transforms data, based on the covariance of the features, to a series of eigenvalues, representing the variance explained, and eigenvectors representing the direction of these eigenvalues. The eigenvector with the highest eigenvalue represents the first principal component, which explains the most variance in the dataset. Subsequent eigenvectors represent subsequent principal components, each explaining a decreasing amount of variance. The original dataset is then projected onto the principal components to obtain the transformed dataset. Each observation in the transformed dataset is represented by its scores along the principal components. The principal components can then be interpreted based on the feature contributions to the corresponding eigenvectors, and may represent patterns or combinations of the original variables.

PCA analyses were undertaken on the datasets using the ‘prcomp’ function from the R ‘stats’ package.

6.3.4 Identification of most informative antigen concentration feature

Curve Characteristic features – calculated as outlined in section 5.2.4 were then also explored using principal component analysis. For these analyses, a single representative magnitude feature for each antigen was included in the analysis. To determine the most informative concentration for each antigen a brief investigation was undertaken as follows:

From the training set data, the relative utility of each antigen concentration was determined by the following method:

The training data for each antigen concentration underwent cutpoint optimisation using the R ‘cutpointr’ package, identifying the cutpoint which maximised sensitivity constrained to specificities above 95% to ensure that the concentration being selected showed high specificity. The antigen concentration which showed the greatest sensitivity was then selected as the most informative.

6.3.5 Feature Correlation

Correlation was assessed using the ‘cor’ function from the R ‘stats’ package to determine sources of high positive or negative correlation existing between features.

6.3.6 Cluster Analysis

Each feature in the transformed training set for the commercial antigen panel was converted to a standard score (by dividing the transformed value from the mean of the population, and then dividing by the standard deviation) to remove any residual differences in feature scale, then an investigation was undertaken using k-means clustering, firstly identifying an optimal number of clusters to train to, then after assigning all samples to a cluster, retrieving disease class information for the samples to see if the clustering of the data results in groups which could be informative about disease status. Considering the high influence of no-specific binding observed in the uncorrected featureset during principal component analysis, this analysis was undertaken only on the vol corrected features, and the features converted to a signal-to-vol ratio, in both cases including a single magnitude feature determined to be the optimal concentration for each antigen.

6.4 Results

6.4.1 Exploration of Distributions

Uncorrected OD Features

As is demonstrated in Table 6-1 to Table 6-8, features other than area under the curve show high positive skewness, as represented by skewness values greater than 1, while area under the curve features show a high negative skewness for all antigens, prior to any correction for non-specific binding. All features show high kurtosis, indicating features are highly leptokurtic, especially slopemax features. Normality testing also shows that all features are significantly non-normally distributed.

p53

Table 6-1: Descriptive statistics for uncorrected optical density values in p53 features.

	OD	od_intercept	od_slope	od_auc	od_slopemax
median	0.183	0.134	6.730E+05	0.000	3.788E+06
mean	0.256	0.181	1.030E+06	0.000	1.136E+07
95% Confidence Interval	0.017	0.011	8.969E+04	0.000	2.633E+06
Standard Deviation	0.217	0.140	1.179E+06	0.000	3.460E+07
Coefficient of Variance	0.846	0.774	1.144	-0.942	3.046
skewness	3.204	2.639	2.905	-2.529	7.735
skew.2SE[1]	16.914	13.931	15.339	-13.353	40.840
kurtosis	13.267	8.339	10.442	10.827	67.824
kurt.2SE[2]	35.074	22.047	27.605	28.624	179.311
Shapiro-Wilk normality statistic	0.654	0.702	0.694	0.790	0.265
Shapiro-Wilk test p-value	0.000	0.000	0.000	0.000	0.000

SOX2

Table 6-2: Descriptive statistics for uncorrected optical density values in SOX2 features.

	OD	od_intercept	od_slope	od_auc	od_slopemax
median	0.158	0.123	6.381E+05	0.000	2.375E+06
mean	0.204	0.160	8.747E+05	0.000	5.060E+06
95% Confidence Interval	0.011	0.009	7.372E+04	0.000	7.092E+05
Standard Deviation	0.149	0.120	9.689E+05	0.000	9.321E+06
Coefficient of Variance	0.729	0.749	1.108	-0.945	1.842
skewness	3.709	2.878	5.110	-4.208	6.242
skew.2SE[1]	19.585	15.193	26.977	-22.214	32.954
kurtosis	22.874	10.099	39.613	29.310	54.656
kurt.2SE[2]	60.473	26.699	104.727	77.488	144.498
Shapiro-Wilk normality statistic	0.666	0.679	0.600	0.678	0.443
Shapiro-Wilk test p-value	0.000	0.000	0.000	0.000	0.000

CAGE

Table 6-3: Descriptive statistics for uncorrected optical density values in CAGE features.

	OD	od_intercept	od_slope	od_auc	od_slopemax
median	0.175	0.126	5.569E+05	0.000	3.009E+06
mean	0.223	0.165	7.705E+05	0.000	6.195E+06
95% Confidence Interval	0.011	0.009	6.761E+04	0.000	9.824E+05
Standard Deviation	0.148	0.118	8.886E+05	0.000	1.291E+07
Coefficient of Variance	0.666	0.717	1.153	-1.032	2.084
skewness	2.497	2.802	3.869	-2.101	9.177
skew.2SE[1]	13.185	14.791	20.428	-11.092	48.453
kurtosis	8.245	9.918	22.269	11.389	109.318
kurt.2SE[2]	21.799	26.222	58.874	30.108	289.009
Shapiro-Wilk normality statistic	0.747	0.695	0.661	0.766	0.345
Shapiro-Wilk test p-value	0.000	0.000	0.000	0.000	0.000

NY-ESO-1

Table 6-4: Descriptive statistics for uncorrected optical density values in NY-ESO-1 features.

	OD	od_intercept	od_slope	od_auc	od_slopemax
median	0.176	0.128	6.561E+05	0.000	4.069E+06
mean	0.252	0.184	1.017E+06	0.000	1.443E+07
95% Confidence Interval	0.020	0.012	1.028E+05	0.000	3.832E+06
Standard Deviation	0.265	0.162	1.351E+06	0.000	5.036E+07
Coefficient of Variance	1.051	0.878	1.328	-0.960	3.491
skewness	4.168	2.982	3.342	-2.107	7.253
skew.2SE[1]	22.003	15.743	17.645	-11.125	38.291
kurtosis	20.308	10.233	13.038	6.606	59.158
kurt.2SE[2]	53.688	27.053	34.468	17.465	156.398
Shapiro-Wilk normality statistic	0.526	0.641	0.626	0.818	0.242
Shapiro-Wilk test p-value	0.000	0.000	0.000	0.000	0.000

[1] skewness divided by 2 Standard Errors – if greater than 1 data is significantly skewed (p<0.05)

[2] kurtosis divided by 2 Standard Errors – if greater than 1 data is significantly skewed (p<0.05)

GBU 4-5

Table 6-5: Descriptive statistics for uncorrected optical density values in GBU 4-5 features.

	OD	od_intercept	od_slope	od_auc	od_slopemax
median	0.312	0.143	1.937E+06	0.000	6.625E+06
mean	0.368	0.183	2.290E+06	0.000	1.143E+07
95% Confidence Interval	0.016	0.010	1.166E+05	0.000	1.650E+06
Standard Deviation	0.212	0.126	1.532E+06	0.000	2.169E+07
Coefficient of Variance	0.575	0.690	0.669	-0.617	1.898
skewness	1.943	2.832	1.389	-1.233	11.625
skew.2SE[1]	10.258	14.952	7.334	-6.510	61.378
kurtosis	6.386	10.521	2.512	2.643	191.844
kurt.2SE[2]	16.882	27.815	6.640	6.987	507.188
Shapiro-Wilk normality statistic	0.846	0.710	0.898	0.926	0.332
Shapiro-Wilk test p-value	0.000	0.000	0.000	0.000	0.000

VOL

Table 6-6: Descriptive statistics for uncorrected optical density values in VOL features.

	OD	od_intercept	od_slope	od_auc	od_slopemax
median	0.122	0.119	3.149E+04	0.000	1.432E+06
mean	0.164	0.157	1.474E+05	0.000	3.183E+06
95% Confidence Interval	0.009	0.009	5.022E+04	0.000	5.966E+05
Standard Deviation	0.121	0.118	6.601E+05	0.000	7.841E+06
Coefficient of Variance	0.735	0.747	4.478	-4.290	2.464
skewness	2.682	2.964	6.486	-7.487	15.294
skew.2SE[1]	14.157	15.648	34.245	-39.528	80.745
kurtosis	8.852	11.094	57.333	74.799	312.866
kurt.2SE[2]	23.401	29.331	151.575	197.749	827.141
Shapiro-Wilk normality statistic	0.698	0.678	0.471	0.413	0.286
Shapiro-Wilk test p-value	0.000	0.000	0.000	0.000	0.000

MAGE-A4

Table 6-7: Descriptive statistics for uncorrected optical density values in MAGE-A4 features.

	OD	od_intercept	od_slope	od_auc	od_slopemax
median	0.198	0.136	4.966E+05	0.000	3.375E+06
mean	0.252	0.176	7.502E+05	0.000	6.658E+06
95% Confidence Interval	0.013	0.009	8.113E+04	0.000	8.279E+05
Standard Deviation	0.176	0.124	1.066E+06	0.000	1.088E+07
Coefficient of Variance	0.699	0.703	1.421	-1.632	1.634
skewness	2.634	2.644	2.876	-2.972	5.464
skew.2SE[1]	13.905	13.961	15.186	-15.689	28.850
kurtosis	8.976	8.706	11.849	15.595	40.566
kurt.2SE[2]	23.731	23.016	31.325	41.228	107.247
Shapiro-Wilk normality statistic	0.725	0.711	0.727	0.751	0.493
Shapiro-Wilk test p-value	0.000	0.000	0.000	0.000	0.000

HuD

Table 6-8: Descriptive statistics for uncorrected optical density values in HuD features.

	OD	od_intercept	od_slope	od_auc	od_slopemax
median	0.204	0.128	1.100E+06	0.000	3.563E+06
mean	0.261	0.170	1.390E+06	0.000	6.700E+06
95% Confidence Interval	0.014	0.010	9.346E+04	0.000	1.075E+06
Standard Deviation	0.185	0.126	1.228E+06	0.000	1.413E+07
Coefficient of Variance	0.709	0.738	0.884	-0.747	2.109
skewness	3.094	2.701	3.290	-2.082	12.439
skew.2SE[1]	16.333	14.260	17.371	-10.992	65.676
kurtosis	14.611	9.128	16.714	8.551	219.742
kurt.2SE[2]	38.628	24.132	44.188	22.607	580.942
Shapiro-Wilk normality statistic	0.710	0.706	0.717	0.841	0.314
Shapiro-Wilk test p-value	0.000	0.000	0.000	0.000	0.000

VOL subtracted OD Features

After correction for non-specific binding by VOL subtraction, the feature distributions remain highly non-normal, as shown in Table 6-9 to Table 6-15, with all features other than area under the curve retaining high positive skewness, while area under the curve features remain highly negatively skewed. All features retain high kurtosis, indicating features are highly leptokurtic. Normality testing again shows that all features remain significantly non-normally distributed.

p53

Table 6-9: Descriptive statistics for VOL corrected optical density values in p53 features.

	VCOD	vcod_intercept	vcod_slope	vcod_auc	vcod_slopemax
median	0.046	0.008	5.967E+05	0.000	3.390E+06
mean	0.094	0.027	8.803E+05	0.000	1.134E+07
95% Confidence Interval	0.014	0.006	8.151E+04	0.000	2.688E+06
Standard Deviation	0.178	0.078	1.071E+06	0.000	3.533E+07
Coefficient of Variance	1.885	2.851	1.217	-1.069	3.115
skewness	5.008	6.484	3.412	-2.406	7.770
skew.2SE[1]	26.440	34.232	18.012	-12.704	41.020
kurtosis	29.817	50.653	15.150	15.636	68.175
kurt.2SE[2]	78.827	133.914	40.052	41.338	180.239
Shapiro-Wilk normality statistic	0.444	0.350	0.662	0.800	0.265
Shapiro-Wilk test p-value	0.000	0.000	0.000	0.000	0.000

SOX2

Table 6-10: Descriptive statistics for VOL corrected optical density values in SOX2 features.

	VCOD	vcod_intercept	vcod_slope	vcod_auc	vcod_slopemax
median	0.026	0.002	5.551E+05	0.000	2.305E+06
mean	0.045	0.007	7.504E+05	0.000	4.944E+06
95% Confidence Interval	0.007	0.002	6.863E+04	0.000	7.428E+05
Standard Deviation	0.094	0.026	9.020E+05	0.000	9.762E+06
Coefficient of Variance	2.070	3.685	1.202	-1.042	1.975
skewness	9.673	11.868	5.469	-4.354	6.857
skew.2SE[1]	51.070	62.657	28.872	-22.988	36.201
kurtosis	134.989	202.014	43.157	29.621	62.965
kurt.2SE[2]	356.878	534.075	114.096	78.311	166.463
Shapiro-Wilk normality statistic	0.365	0.353	0.565	0.663	0.414
Shapiro-Wilk test p-value	0.000	0.000	0.000	0.000	0.000

CAGE

Table 6-11: Descriptive statistics for VOL corrected optical density values in CAGE features.

	VCOD	vcod_intercept	vcod_slope	vcod_auc	vcod_slopemax
median	0.041	0.006	4.443E+05	0.000	2.803E+06
mean	0.062	0.011	6.430E+05	0.000	5.820E+06
95% Confidence Interval	0.007	0.002	6.128E+04	0.000	9.854E+05
Standard Deviation	0.093	0.019	8.054E+05	0.000	1.295E+07
Coefficient of Variance	1.504	1.789	1.253	-1.172	2.225
skewness	6.058	5.719	4.829	-2.403	10.186
skew.2SE[1]	31.983	30.192	25.498	-12.688	53.779
kurtosis	50.045	44.060	33.466	13.544	133.777
kurt.2SE[2]	132.307	116.484	88.475	35.807	353.673
Shapiro-Wilk normality statistic	0.490	0.507	0.591	0.776	0.323
Shapiro-Wilk test p-value	0.000	0.000	0.000	0.000	0.000

NY-ESO-1

Table 6-12: Descriptive statistics for VOL corrected optical density values in NY-ESO-1 features.

	VCOD	vcod_intercept	vcod_slope	vcod_auc	vcod_slopemax
median	0.039	0.007	5.629E+05	0.000	4.588E+06
mean	0.091	0.030	8.916E+05	0.000	1.457E+07
95% Confidence Interval	0.018	0.008	9.632E+04	0.000	3.830E+06
Standard Deviation	0.237	0.109	1.266E+06	0.000	5.034E+07
Coefficient of Variance	2.610	3.699	1.420	-0.989	3.455
skewness	5.625	6.339	3.889	-1.950	7.236
skew.2SE[1]	29.698	33.469	20.534	-10.295	38.204
kurtosis	32.767	42.091	17.680	5.339	58.630
kurt.2SE[2]	86.628	111.277	46.742	14.116	155.003
Shapiro-Wilk normality statistic	0.307	0.252	0.570	0.836	0.239
Shapiro-Wilk test p-value	0.000	0.000	0.000	0.000	0.000

GBU 4-5

Table 6-13: Descriptive statistics for VOL corrected optical density values in GBU 4-5 features.

	VCOD	vcod_intercept	vcod_slope	vcod_auc	vcod_slopemax
median	0.158	0.018	1.814E+06	0.000	6.717E+06
mean	0.204	0.029	2.117E+06	0.000	1.065E+07
95% Confidence Interval	0.013	0.004	1.036E+05	0.000	1.585E+06
Standard Deviation	0.172	0.050	1.362E+06	0.000	2.083E+07
Coefficient of Variance	0.842	1.732	0.643	-0.658	1.956
skewness	2.401	9.606	1.255	-0.670	12.909
skew.2SE[1]	12.677	50.717	6.624	-3.536	68.156
kurtosis	10.400	143.280	1.654	2.507	226.169
kurt.2SE[2]	27.495	378.796	4.372	6.628	597.934
Shapiro-Wilk normality statistic	0.803	0.443	0.907	0.947	0.298
Shapiro-Wilk test p-value	0.000	0.000	0.000	0.000	0.000

MAGE-A4

Table 6-14: Descriptive statistics for VOL corrected optical density values in MAGE-A4 features.

	VCOD	vcod_intercept	vcod_slope	vcod_auc	vcod_slopemax
median	0.056	0.012	4.307E+05	0.000	3.460E+06
mean	0.090	0.022	6.211E+05	0.000	6.805E+06
95% Confidence Interval	0.009	0.003	6.211E+04	0.000	8.382E+05
Standard Deviation	0.120	0.035	8.163E+05	0.000	1.102E+07
Coefficient of Variance	1.331	1.580	1.314	-2.035	1.619
skewness	3.611	4.982	3.293	-1.757	5.463
skew.2SE[1]	19.067	26.305	17.384	-9.276	28.844
kurtosis	15.795	34.114	14.840	23.303	40.570
kurt.2SE[2]	41.758	90.190	39.232	61.608	107.258
Shapiro-Wilk normality statistic	0.606	0.537	0.693	0.746	0.501
Shapiro-Wilk test p-value	0.000	0.000	0.000	0.000	0.000

HuD

Table 6-15: Descriptive statistics for VOL corrected optical density values in HuD features.

	VCOD	vcod_intercept	vcod_slope	vcod_auc	vcod_slopemax
median	0.067	0.008	1.017E+06	0.000	4.188E+06
mean	0.097	0.015	1.243E+06	0.000	7.407E+06
95% Confidence Interval	0.010	0.002	7.830E+04	0.000	1.065E+06
Standard Deviation	0.128	0.031	1.029E+06	0.000	1.399E+07
Coefficient of Variance	1.315	2.080	0.828	-0.725	1.889
skewness	6.194	7.574	3.717	-1.434	12.244
skew.2SE[1]	32.704	39.986	19.625	-7.573	64.646
kurtosis	54.477	84.483	23.877	5.060	216.203
kurt.2SE[2]	144.024	223.351	63.125	13.376	571.587
Shapiro-Wilk normality statistic	0.494	0.477	0.717	0.903	0.345
Shapiro-Wilk test p-value	0.000	0.000	0.000	0.000	0.000

Signal to Vol Ratio Features

After correction for non-specific binding by re-expression of values as a ratio of specific to non-specific binding, the feature distributions again remain highly non-normal. Table 6-16 to Table 6-22 show that all features other than area under the curve still show high positive skewness, while area under the curve features are again highly negatively skewed. All features retain high kurtosis, indicating features are still highly leptokurtic after correction. Normality testing once more shows that all features remain significantly non-normally distributed.

p53

Table 6-16: Descriptive statistics for signal to VOL ratio values in p53 features.

	STVR	stvr_intercept	stvr_slope	stvr_auc	stvr_slopemax
median	1.382	1.059	4.578E+06	0.000	4.214E+07
mean	1.727	1.205	6.789E+06	0.000	1.084E+08
95% Confidence Interval	0.115	0.054	7.019E+05	0.000	2.402E+07
Standard Deviation	1.509	0.704	9.225E+06	0.000	3.157E+08
Coefficient of Variance	0.874	0.584	1.359	-1.294	2.914
skewness	6.158	6.961	3.972	-2.807	8.155
skew.2SE[1]	32.511	36.751	20.971	-14.822	43.054
kurtosis	49.453	56.386	24.361	15.462	78.623
kurt.2SE[2]	130.741	149.070	64.405	40.877	207.858
Shapiro-Wilk normality statistic	0.412	0.324	0.648	0.789	0.267
Shapiro-Wilk test p-value	0.000	0.000	0.000	0.000	0.000

SOX2

Table 6-17: Descriptive statistics for signal to VOL ratio values in SOX2 features.

	STVR	stvr_intercept	stvr_slope	stvr_auc	stvr_slopemax
median	1.198	1.001	4.332E+06	0.000	3.254E+07
mean	1.365	1.036	5.909E+06	0.000	5.606E+07
95% Confidence Interval	0.068	0.018	5.809E+05	0.000	7.520E+06
Standard Deviation	0.898	0.231	7.635E+06	0.000	9.883E+07
Coefficient of Variance	0.658	0.223	1.292	-1.101	1.763
skewness	8.625	10.056	5.457	-4.210	9.462
skew.2SE[1]	45.538	53.093	28.811	-22.228	49.954
kurtosis	105.197	155.262	48.210	35.678	132.170
kurt.2SE[2]	278.114	410.475	127.455	94.325	349.425
Shapiro-Wilk normality statistic	0.396	0.406	0.608	0.727	0.406
Shapiro-Wilk test p-value	0.000	0.000	0.000	0.000	0.000

CAGE

Table 6-18: Descriptive statistics for signal to VOL ratio values in CAGE features.

	STVR	stvr_intercept	stvr_slope	stvr_auc	stvr_slopemax
median	1.307	1.038	3.638E+06	0.000	4.002E+07
mean	1.489	1.067	5.023E+06	0.000	6.411E+07
95% Confidence Interval	0.067	0.014	5.627E+05	0.000	9.173E+06
Standard Deviation	0.886	0.181	7.396E+06	0.000	1.206E+08
Coefficient of Variance	0.595	0.170	1.472	-1.390	1.880

	STVR	stvr_intercept	stvr_slope	stvr_auc	stvr_slopemax
skewness	7.271	6.649	6.304	-4.339	10.179
skew.2SE[1]	38.389	35.102	33.284	-22.907	53.744
kurtosis	70.277	63.387	60.051	38.499	134.808
kurt.2SE[2]	185.794	167.579	158.760	101.783	356.400
Shapiro-Wilk normality statistic	0.442	0.508	0.536	0.713	0.347
Shapiro-Wilk test p-value	0.000	0.000	0.000	0.000	0.000

NY-ESO-1

Table 6-19: Descriptive statistics for signal to VOL ratio values in NY-ESO-1 features.

	STVR	stvr_intercept	stvr_slope	stvr_auc	stvr_slopemax
median	1.302	1.055	4.410E+06	0.000	5.585E+07
mean	1.711	1.217	6.833E+06	0.000	1.401E+08
95% Confidence Interval	0.157	0.073	8.507E+05	0.000	3.621E+07
Standard Deviation	2.069	0.955	1.118E+07	0.000	4.759E+08
Coefficient of Variance	1.209	0.785	1.636	-1.302	3.396
skewness	6.733	7.844	6.088	-1.935	9.835
skew.2SE[1]	35.550	41.416	32.142	-10.215	51.928
kurtosis	51.492	75.851	51.020	14.969	124.124
kurt.2SE[2]	136.133	200.530	134.884	39.575	328.153
Shapiro-Wilk normality statistic	0.291	0.233	0.498	0.759	0.220
Shapiro-Wilk test p-value	0.000	0.000	0.000	0.000	0.000

GBU 4-5

Table 6-20: Descriptive statistics for signal to VOL ratio values in GBU 4-5 features.

	STVR	stvr_intercept	stvr_slope	stvr_auc	stvr_slopemax
median	2.186	1.130	1.293E+07	0.000	7.379E+07
mean	2.618	1.225	1.616E+07	0.000	1.073E+08
95% Confidence Interval	0.111	0.032	9.448E+05	0.000	1.443E+07
Standard Deviation	1.452	0.421	1.242E+07	0.000	1.896E+08
Coefficient of Variance	0.555	0.344	0.769	-0.910	1.767
skewness	1.926	7.644	1.328	-0.688	10.913
skew.2SE[1]	10.168	40.359	7.009	-3.635	57.618
kurtosis	4.907	94.231	2.797	3.575	152.011
kurt.2SE[2]	12.972	249.124	7.395	9.452	401.880
Shapiro-Wilk normality statistic	0.826	0.512	0.908	0.934	0.316
Shapiro-Wilk test p-value	0.000	0.000	0.000	0.000	0.000

MAGE-A4

Table 6-21: Descriptive statistics for signal to VOL ratio values in MAGE-A4 features.

	STVR	stvr_intercept	stvr_slope	stvr_auc	stvr_slopemax
median	1.469	1.093	3.413E+06	0.000	4.098E+07
mean	1.684	1.150	4.764E+06	0.000	6.542E+07
95% Confidence Interval	0.071	0.018	5.042E+05	0.000	5.918E+06
Standard Deviation	0.935	0.238	6.627E+06	0.000	7.778E+07
Coefficient of Variance	0.555	0.207	1.391	-2.329	1.189
skewness	4.805	3.664	4.032	-3.098	3.369
skew.2SE[1]	25.371	19.346	21.289	-16.355	17.788
kurtosis	34.298	19.269	26.170	26.735	16.045
kurt.2SE[2]	90.675	50.942	69.187	70.681	42.418
Shapiro-Wilk normality statistic	0.608	0.676	0.687	0.763	0.663
Shapiro-Wilk test p-value	0.000	0.000	0.000	0.000	0.000

HuD

Table 6-22: Descriptive statistics for signal to VOL ratio values in HuD features.

	STVR	stvr_intercept	stvr_slope	stvr_auc	stvr_slopemax
median	1.525	1.057	7.541E+06	0.000	4.463E+07
mean	1.718	1.098	9.361E+06	0.000	6.677E+07
95% Confidence Interval	0.075	0.022	6.461E+05	0.000	7.556E+06
Standard Deviation	0.983	0.291	8.491E+06	0.000	9.931E+07

	STVR	stvr_intercept	stvr_slope	stvr_auc	stvr_slopemax
Coefficient of Variance	0.572	0.265	0.907	-0.891	1.487
skewness	8.136	14.553	3.579	-1.754	8.400
skew.2SE[1]	42.953	76.837	18.893	-9.258	44.347
kurtosis	98.008	289.688	22.310	6.013	99.454
kurt.2SE[2]	259.108	765.863	58.981	15.897	262.933
Shapiro-Wilk normality statistic	0.461	0.351	0.737	0.885	0.435
Shapiro-Wilk test p-value	0.000	0.000	0.000	0.000	0.000

6.4.2 Transformation Techniques

For each feature type, all transformations were undertaken (see appendix 6A for full results), and transformation that best transformed the antigens explored was identified by calculating the mean Pearson/df ratio over all the antigens explored. This transformation was then applied across all antigens for each feature. In the case of log and square root features, the maximum value of ‘a’ across the features was identified in order to apply the transformation to additional datasets, similarly in the case of box cox and yeo-johnson transformed features, the mean value of lambda across the antigens was retained for transforming additional datasets.

Table 6-23: Selected transformation for each feature.

Feature	Chosen Transform	Max(a)	mean(lambda)
OD	boxcox		-0.52
od_intercept	boxcox		-0.69
od_slope	log_x	1839250.22	
od_auc	arcsinh_x		
od_slopemax	log_x	411764.71	
VCOD	log_x	0.00	
vcod_intercept	log_x	0.03	
vcod_slope	log_x	1005852.14	
vcod_auc	arcsinh_x		
vcod_slopemax	yeojohnson		0.17
STVR	yeojohnson		-1.98
stvr_intercept	yeojohnson		-4.76
stvr_slope	log_x	13639735.04	
stvr_auc	arcsinh_x		
stvr_slopemax	log_x	1573220.86	

All features for the dataset were then transformed according to the selected transformations summarised above, with listed *a* and *lambda* values as relevant, prior to analysis with unsupervised learning techniques.

6.4.3 Principal Component Analysis on Full Titration Curve Data

Unadjusted OD Data

For the commercial antigens, the first component accounts for 76.7% of the variance observed in the dataset, compared to only 7.1% accounted for by the second component. Examination of the first component loadings highlights that the first component is influenced primarily by the lower concentrations and 0nM values, especially compared to the second component which shows higher contributions from the specific binding observed at the top end of the concentration ranges. This would suggest that the uncorrected data is being heavily influenced by non-specific binding in the assay, and correction is necessary to reduce this influence prior to modelling.

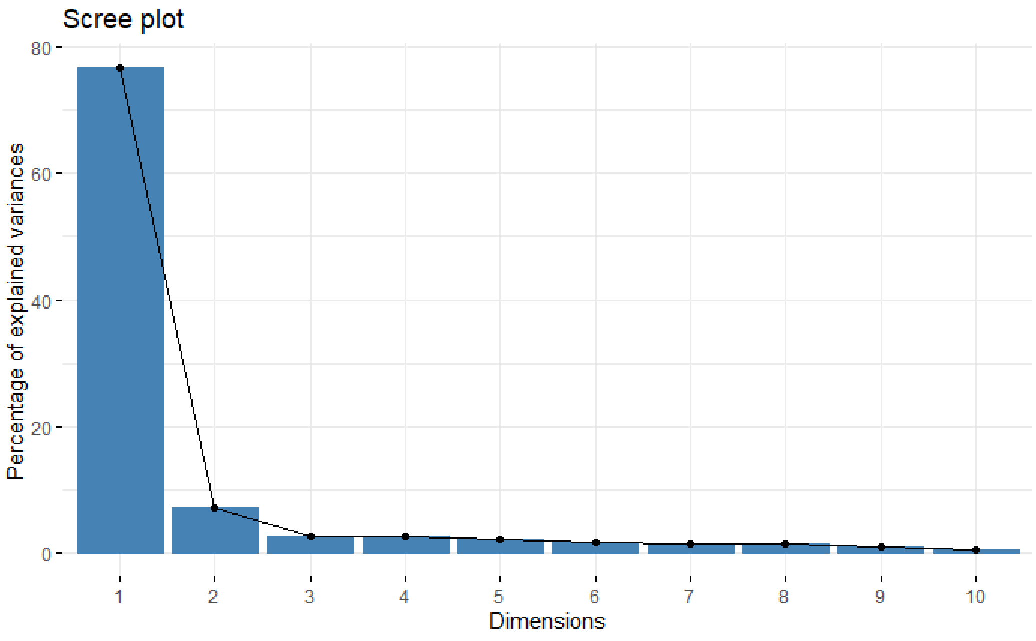


Figure 6-1: Scree plot showing the contribution of the first 10 principal dimensions to the data variance in the EarlyCDT®-Lung commercial antigen panel.

Table 6-24: PCA Feature Loadings for first 10 principal dimensions – unadjusted OD data for EarlyCDT®-Lung commercial antigen panel.

Feature	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7	Dim.8	Dim.9	Dim.10
p53_160	1.156	6.411	13.972	1.057	8.890	0.415	0.017	0.235	0.363	2.807
p53_50	1.669	2.683	12.862	2.139	7.804	0.100	0.000	0.351	0.592	0.001
p53_16	2.003	0.320	9.834	1.690	7.521	0.099	0.044	0.258	0.756	1.219
p53_5	2.229	0.020	5.988	0.914	5.048	0.378	0.065	0.067	0.210	1.710
p53_1.6	2.316	0.569	3.046	0.246	2.877	0.541	0.152	0.013	0.115	0.356
p53_0	2.398	1.479	0.099	0.401	0.086	0.031	0.109	0.014	1.632	2.012
SOX_2_160	1.469	3.662	1.310	0.029	1.391	21.144	1.263	12.841	0.007	1.275
SOX_2_50	1.990	0.395	0.323	0.378	1.075	13.879	0.808	14.986	0.418	0.680
SOX_2_16	2.311	0.202	0.183	0.024	0.507	5.890	1.062	7.845	0.003	0.272
SOX_2_5	2.444	0.790	0.099	0.054	0.198	1.621	0.524	2.755	0.413	0.007
SOX_2_1.6	2.461	1.494	0.008	0.193	0.030	0.465	0.156	0.346	0.653	0.402
SOX_2_0	2.464	1.634	0.001	0.313	0.002	0.028	0.018	0.006	1.332	1.763
CAGE_160	1.543	3.142	0.178	0.249	1.700	2.857	28.324	9.080	0.464	0.351
CAGE_50	1.956	0.966	0.035	0.152	1.456	2.228	21.722	6.954	0.109	1.534
CAGE_16	2.410	0.206	0.003	0.003	0.626	1.148	6.531	2.349	0.097	1.018
CAGE_5	2.456	0.703	0.002	0.054	0.282	0.843	1.609	0.734	0.333	0.041
CAGE_1.6	2.440	1.513	0.002	0.202	0.020	0.203	0.124	0.031	0.410	0.051
CAGE_0	2.469	1.710	0.009	0.230	0.007	0.068	0.037	0.066	0.975	1.327
NY_ESO_1_160	1.143	6.876	13.822	5.885	2.410	0.741	0.102	0.448	0.156	4.651
NY_ESO_1_50	1.712	2.001	11.166	9.574	1.213	0.785	0.015	0.819	0.847	0.021
NY_ESO_1_16	2.037	0.001	7.307	8.614	0.864	1.339	0.003	1.236	0.144	1.510
NY_ESO_1_5	2.144	0.137	6.149	5.455	0.899	1.609	0.063	0.550	0.031	0.633
NY_ESO_1_1.6	2.260	0.525	3.963	2.120	0.792	1.289	0.212	0.487	0.825	0.001
NY_ESO_1_0	2.414	1.983	0.000	0.173	0.001	0.059	0.112	0.144	0.689	1.252
GBU_4_5_160	0.669	11.473	1.167	18.857	2.203	3.165	1.613	0.015	6.295	0.240
GBU_4_5_50	1.095	9.081	1.896	17.118	1.625	1.904	0.494	0.007	0.283	1.041
GBU_4_5_16	1.872	2.541	1.287	9.794	0.536	0.432	0.017	0.176	8.365	3.877
GBU_4_5_5	2.367	0.003	0.465	4.284	0.034	0.267	0.000	0.212	4.014	1.933
GBU_4_5_1.6	2.469	0.526	0.139	1.845	0.016	0.177	0.006	0.178	0.525	0.241
GBU_4_5_0	2.422	1.773	0.003	0.391	0.003	0.145	0.127	0.040	0.803	1.507
VOL_160	1.962	1.357	0.192	0.262	2.502	0.004	0.204	0.851	22.860	27.479
VOL_50	2.350	0.044	0.017	0.127	1.725	0.411	0.320	0.700	11.879	8.692
VOL_16	2.557	0.666	0.010	0.188	0.153	0.023	0.000	0.113	1.198	0.247
VOL_5	2.531	1.327	0.009	0.308	0.031	0.011	0.046	0.019	0.031	0.350
VOL_1.6	2.459	1.629	0.001	0.422	0.000	0.049	0.071	0.014	0.123	0.319
VOL_0	2.419	1.794	0.004	0.331	0.011	0.136	0.107	0.000	0.908	1.675
MAGE_A4_160	1.137	6.432	1.503	1.566	13.592	4.394	13.358	1.351	4.845	0.589
MAGE_A4_50	1.522	4.050	1.446	1.951	15.421	1.359	9.113	0.866	0.046	0.271
MAGE_A4_16	2.128	0.520	0.902	1.453	8.035	0.110	3.107	0.036	3.672	7.121
MAGE_A4_5	2.486	0.250	0.215	0.072	1.926	0.123	0.880	0.017	0.760	4.457
MAGE_A4_1.6	2.531	1.031	0.056	0.033	0.546	0.121	0.259	0.003	0.049	0.518
MAGE_A4_0	2.456	1.779	0.021	0.223	0.012	0.131	0.015	0.021	0.591	0.184
HuD_160	1.218	7.478	0.180	0.077	1.827	9.098	2.318	11.568	18.253	0.637
HuD_50	1.753	3.198	0.001	0.045	2.687	12.262	3.588	11.007	0.443	0.505
HuD_16	2.274	0.037	0.023	0.042	1.101	6.141	0.953	7.369	0.827	8.036
HuD_5	2.479	0.683	0.040	0.037	0.229	1.401	0.162	1.983	0.022	3.262
HuD_1.6	2.491	1.309	0.001	0.205	0.063	0.315	0.034	0.724	0.397	0.443
HuD_0	2.457	1.598	0.060	0.222	0.021	0.059	0.138	0.114	1.236	1.482

The influence of non-specific binding evident in the first principal component in these analyses confirms a need for correction methods to reduce or remove the effect of non-specific binding prior to modelling on these features. Two methods of correction for non-specific binding have been applied to the data, as described previously, and are explored in the following analyses.

VOL-Corrected OD Data

After correction for non-specific binding by subtraction of VOL signal, the first component accounts for 19.9% of the variance observed in the dataset, compared to 10.3% accounted for by the second component. The lower proportion of variance in the first principal component suggests that the impact of non-specific binding has been reduced after VOL subtraction.

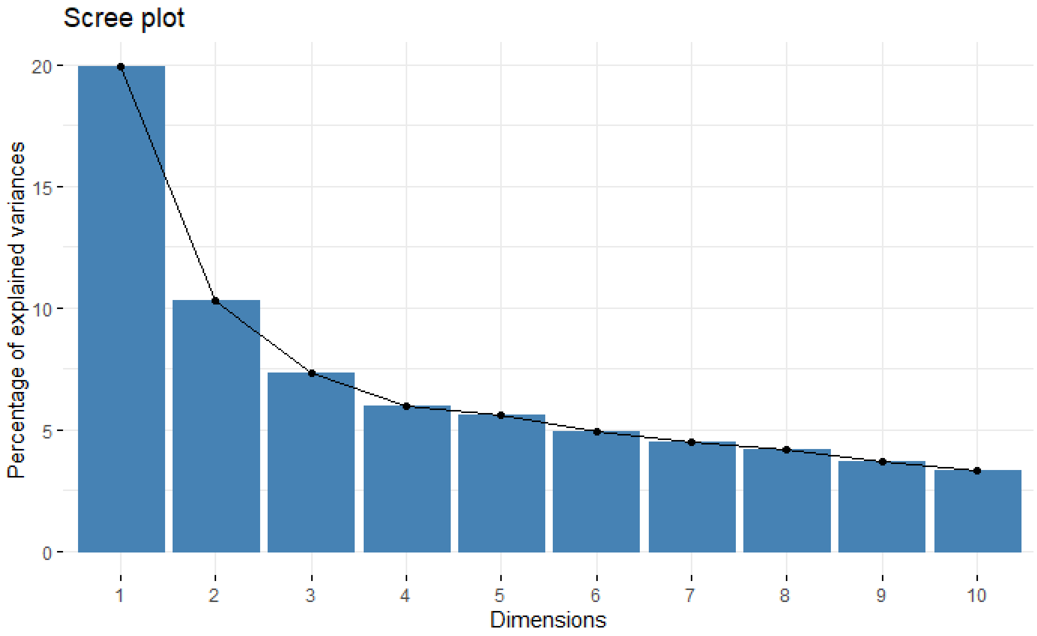


Figure 6-2: Scree plot showing the contribution of the first 10 principal dimensions to the data variance in the EarlyCDT®-Lung commercial antigen panel after VOL subtraction.

Table 6-25: PCA Feature Loadings for first 10 principal dimensions – VOL subtracted OD data for EarlyCDT®-Lung commercial antigen panel.

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7	Dim.8	Dim.9	Dim.10
p53_160	4.691	1.532	1.098	0.068	5.706	0.000	4.806	0.271	3.139	0.646
p53_50	5.166	0.351	0.014	0.018	11.149	1.361	0.574	0.520	3.582	0.047
p53_16	4.028	0.045	1.017	0.387	11.465	2.740	0.057	3.025	1.393	0.205
p53_5	3.106	1.559	0.916	0.024	12.714	5.604	0.295	0.702	1.577	0.189

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7	Dim.8	Dim.9	Dim.10
p53_1.6	1.988	3.188	0.592	0.160	6.062	8.028	4.909	0.092	0.171	1.351
p53_0	0.509	6.923	3.398	4.515	0.065	0.089	0.043	0.000	0.692	0.937
SOX_2_160	2.376	1.790	8.345	0.413	2.490	0.079	2.877	2.771	1.301	0.195
SOX_2_50	2.291	2.438	7.347	0.774	0.004	0.880	0.003	3.580	2.416	0.299
SOX_2_16	1.829	0.370	4.964	0.279	0.288	2.484	1.153	2.779	14.851	1.809
SOX_2_5	1.587	0.215	1.632	0.414	0.078	3.940	5.049	6.044	9.228	4.192
SOX_2_1.6	0.564	4.198	0.847	3.787	0.852	5.737	2.970	0.026	9.338	0.310
SOX_2_0	0.402	5.915	3.176	5.760	0.069	2.873	0.198	0.111	0.720	0.295
CAGE_160	2.603	1.588	4.916	1.665	0.529	1.032	0.511	6.832	0.029	8.343
CAGE_50	3.564	1.471	3.131	0.670	2.611	0.376	0.817	10.407	0.004	2.948
CAGE_16	2.191	0.195	0.279	1.018	1.880	0.125	6.733	11.663	4.003	1.729
CAGE_5	1.960	0.483	0.001	1.504	1.946	0.000	13.849	6.845	1.923	1.101
CAGE_1.6	0.618	3.952	0.080	2.585	0.663	0.888	1.808	9.309	0.907	0.837
CAGE_0	0.474	6.246	2.715	5.111	0.251	3.163	0.048	0.003	0.014	0.086
NY_ESO_1_160	4.284	2.104	3.001	0.000	4.652	0.688	0.039	0.401	3.921	2.077
NY_ESO_1_50	4.962	0.836	1.958	0.860	1.881	1.141	1.697	0.255	7.252	2.744
NY_ESO_1_16	4.429	0.194	0.319	3.581	2.749	0.914	5.212	0.048	5.670	2.267
NY_ESO_1_5	3.285	1.670	0.043	4.661	1.320	0.326	4.537	0.002	7.323	3.683
NY_ESO_1_1.6	1.645	3.861	0.000	4.780	5.179	0.303	0.493	1.539	4.493	2.484
NY_ESO_1_0	0.338	6.178	1.397	1.192	0.084	3.775	0.134	0.098	0.352	0.066
GBU_4_5_160	3.211	1.965	0.052	8.945	3.879	1.490	1.424	2.972	0.021	0.172
GBU_4_5_50	3.188	1.275	1.340	9.755	2.828	2.740	0.124	2.300	0.077	0.873
GBU_4_5_16	3.096	0.359	3.351	9.115	4.096	2.582	2.667	0.360	0.197	0.343
GBU_4_5_5	2.681	0.075	5.348	3.428	2.043	2.984	4.303	0.062	0.010	1.497
GBU_4_5_1.6	1.525	2.403	2.221	0.341	3.286	7.975	0.946	3.131	0.131	0.105
GBU_4_5_0	0.145	3.446	1.790	5.063	0.005	0.948	0.013	0.233	0.011	1.905
MAGE_A4_160	2.578	3.380	0.173	0.260	1.111	3.745	4.516	1.216	1.266	8.398
MAGE_A4_50	3.031	2.874	2.804	0.463	0.777	5.219	1.075	0.936	3.199	8.801
MAGE_A4_16	1.982	0.466	9.871	0.048	0.019	7.785	0.003	0.217	5.763	1.974
MAGE_A4_5	2.483	0.256	8.630	0.992	0.515	6.651	0.089	0.488	1.797	0.710
MAGE_A4_1.6	1.131	2.846	4.639	3.735	0.254	2.493	6.764	1.126	1.874	0.656
MAGE_A4_0	0.375	5.864	1.478	2.677	0.737	3.377	0.042	0.039	0.518	0.402
HuD_160	4.158	1.890	0.203	0.779	1.027	0.145	7.710	0.193	0.024	6.503
HuD_50	5.000	0.551	0.332	0.014	0.988	0.540	1.225	2.911	0.017	9.664
HuD_16	2.744	0.195	2.908	0.001	1.266	1.464	1.214	9.750	0.467	11.470
HuD_5	2.758	2.763	2.045	1.543	0.032	0.617	0.831	6.069	0.013	6.969
HuD_1.6	0.721	5.284	0.652	5.008	2.064	0.259	7.174	0.118	0.028	0.720
HuD_0	0.302	6.805	0.974	3.607	0.387	2.438	1.070	0.559	0.290	0.001

Examination of the feature loadings in Table 6-25 confirm that correction through subtraction of VOL has reduced the impact of non-specific binding, with the first principal component loadings now being spread across the upper concentrations of all the specific antigen features. The second principal component is now showing high loadings from the 0nM concentrations of each of the specific antigen biomarkers, suggesting that there may be non-specific binding present above that removed by VOL subtraction which is still having a large influence on the variance of the dataset.

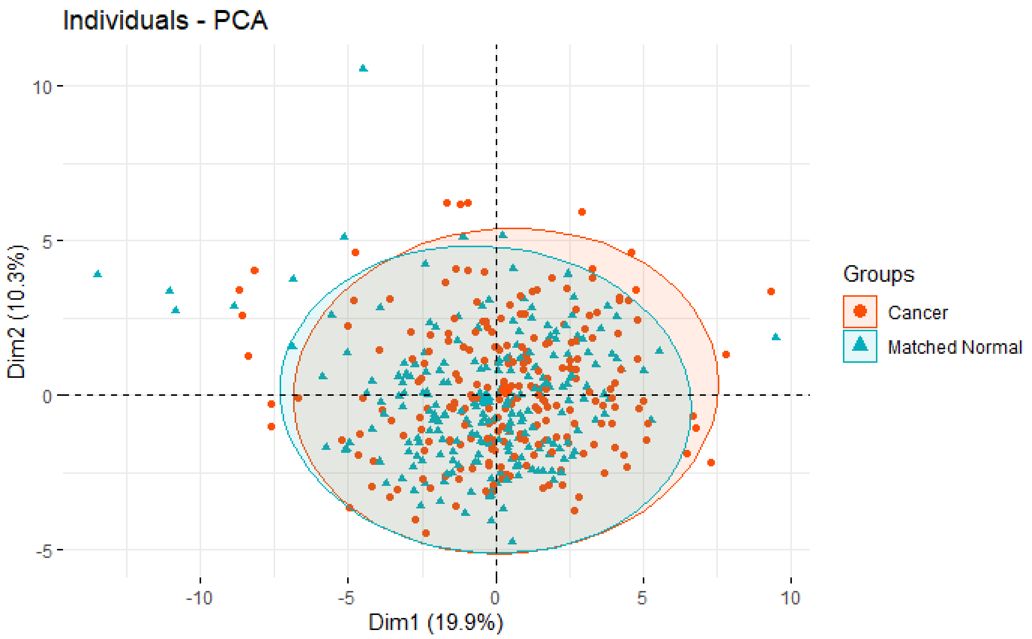


Figure 6-3: Plot of individuals by their component scores on the first two principal dimensions on VOL corrected data.

Examination of distribution of samples by their first two principal dimensions in this analysis of VOL corrected data shows very little separation of samples by their disease class, indicating that the primary sources of variance in the dataset are not related to disease class.

Signal To VOL Ratio Data

For the commercial antigens, after correction through re-expression as a ratio of specific to non-specific binding, the first component accounts for 22.4% of the variance observed in the dataset, compared to 12.1% accounted for by the second component. Examination of the component loadings in Table 6-26 highlights that the first component is now divided amongst the specific antigen features, while – similarly to the previous analysis of VOL subtracted data – the second principal component is heavily influenced by the 0nM concentration features, again suggesting that the data still shows a degree of variance that is attributable to non-specific binding.

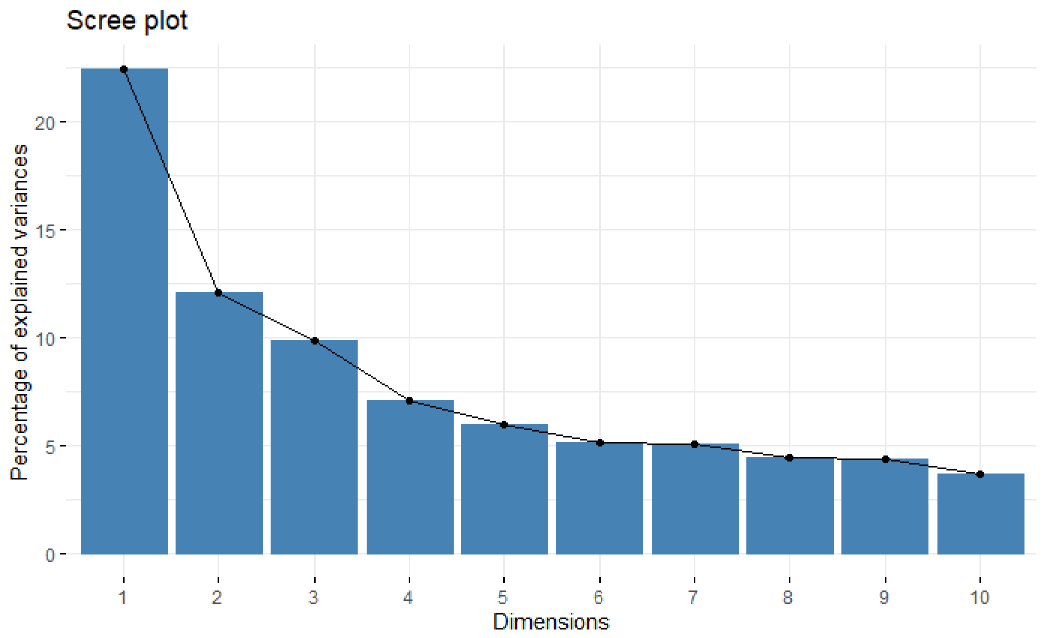


Figure 6-4: Scree plot showing the contribution of the first 10 principal dimensions to the data variance in the commercial antigen panel after re-expression as a Signal To VOL Ratio.

Table 6-26: PCA Feature Loadings for first 10 principal dimensions – Signal To VOL Ratio data for the commercial antigen panel.

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7	Dim.8	Dim.9	Dim.10
p53_160	5.235	0.477	1.348	0.289	6.887	0.002	0.584	1.436	0.841	0.534
p53_50	4.622	0.061	0.059	3.029	8.923	1.655	0.204	1.430	1.883	0.001
p53_16	2.851	0.187	1.335	6.658	8.563	2.374	1.771	2.268	0.875	0.045
p53_5	2.004	0.570	2.429	5.809	9.111	3.670	1.883	1.509	0.846	0.089
p53_1.6	1.136	1.767	5.568	0.011	13.482	0.029	0.335	0.732	2.573	0.005
p53_0	0.271	10.639	2.780	0.043	0.113	0.012	0.958	0.029	0.005	0.172
SOX_2_160	3.774	1.031	3.453	4.220	0.670	1.601	0.021	0.766	3.707	0.001
SOX_2_50	3.864	0.906	1.477	2.242	1.631	0.131	3.138	0.001	5.096	2.742
SOX_2_16	2.130	0.087	0.090	1.390	0.386	0.308	10.761	0.879	14.233	4.794
SOX_2_5	1.709	0.021	0.031	0.266	0.218	0.695	12.318	1.382	12.517	4.183
SOX_2_1.6	0.406	1.072	4.767	8.841	2.912	4.159	1.299	0.566	2.168	0.433
SOX_2_0	0.209	11.387	3.591	0.022	0.000	0.014	0.022	0.461	0.115	0.001
CAGE_160	3.872	1.024	2.598	2.580	0.708	0.009	1.963	3.796	0.461	4.497
CAGE_50	4.569	0.674	0.424	0.415	0.239	1.639	0.094	11.122	2.042	2.924
CAGE_16	2.223	0.131	0.231	0.072	0.270	4.036	0.169	22.787	0.601	4.408
CAGE_5	1.439	0.014	1.100	0.011	0.601	4.701	0.563	18.729	0.268	3.389
CAGE_1.6	0.471	1.649	4.655	5.219	1.374	2.230	0.060	2.879	1.401	0.700
CAGE_0	0.216	12.688	3.153	0.026	0.000	0.025	0.147	0.016	0.021	0.000
NY_ESO_1_160	5.329	0.583	2.985	3.061	0.595	0.132	0.439	2.426	0.335	0.019
NY_ESO_1_50	5.392	0.297	0.287	2.512	2.345	3.432	0.002	3.110	1.251	0.413
NY_ESO_1_16	3.015	0.040	0.652	3.406	6.400	10.574	0.001	3.776	0.810	0.180
NY_ESO_1_5	2.144	0.312	1.503	3.217	6.502	10.996	0.004	3.710	1.190	0.829
NY_ESO_1_1.6	1.327	1.037	3.611	10.562	1.818	1.088	0.100	2.956	3.638	1.111
NY_ESO_1_0	0.072	11.173	2.870	0.006	0.059	0.065	0.566	0.045	0.023	0.066
GBU_4_5_160	4.706	0.763	2.545	0.047	0.307	8.537	0.144	0.007	2.420	0.117
GBU_4_5_50	4.449	0.381	0.381	1.999	1.924	5.882	2.762	0.039	4.367	0.762
GBU_4_5_16	2.757	0.001	0.148	5.232	6.681	4.491	6.955	0.044	1.862	0.158
GBU_4_5_5	1.564	0.058	1.634	5.054	7.715	2.957	8.016	0.071	0.951	0.004
GBU_4_5_1.6	0.590	2.014	5.913	0.368	0.986	9.695	2.578	0.125	2.842	0.000
GBU_4_5_0	0.213	8.449	2.520	0.190	0.089	0.017	0.000	0.272	0.090	0.577
MAGE_A4_160	4.815	0.884	0.684	0.131	0.254	1.224	9.302	0.402	0.013	5.047
MAGE_A4_50	4.173	0.421	0.109	2.651	0.000	0.121	7.380	1.175	0.014	10.763
MAGE_A4_16	2.125	0.007	3.118	5.475	1.436	0.173	7.880	1.093	2.603	6.822
MAGE_A4_5	1.224	0.182	4.646	3.635	0.802	0.876	7.018	1.039	3.853	3.408
MAGE_A4_1.6	0.513	2.056	9.904	1.078	0.265	2.618	5.654	0.787	0.037	1.460
MAGE_A4_0	0.183	11.036	2.452	0.278	0.024	0.210	0.386	0.790	0.071	0.017
HuD_160	5.265	0.765	2.400	0.195	0.113	3.872	2.208	0.668	0.465	2.913
HuD_50	5.270	0.100	0.004	0.583	0.883	0.789	0.482	1.209	2.694	6.404
HuD_16	2.178	0.298	1.261	2.486	3.373	0.019	0.101	2.565	9.238	11.440

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7	Dim.8	Dim.9	Dim.10
HuD_5	1.260	0.769	2.005	1.674	0.811	0.155	0.096	2.016	10.548	15.666
HuD_1.6	0.253	1.927	7.026	4.849	0.344	4.789	1.342	0.825	1.025	2.905
HuD_0	0.183	12.063	2.252	0.166	0.186	0.000	0.293	0.063	0.010	0.001

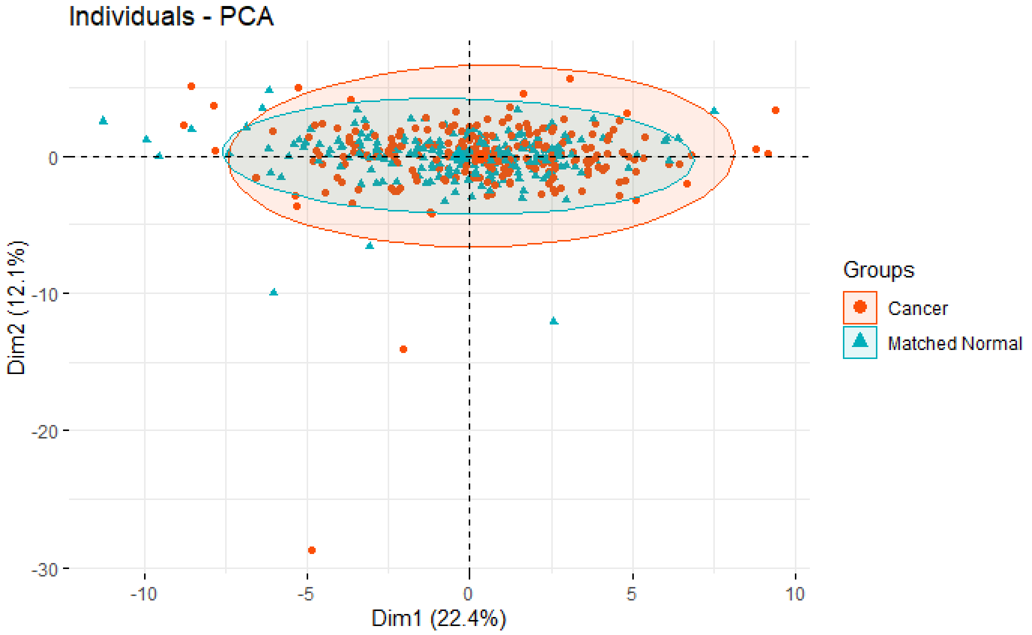


Figure 6-5: Plot of individuals by their component scores on the first two principal dimensions on Signal to Vol Ratio expressed data.

Once again, plotting individuals by their component scores on the first two principal dimensions as shown in Figure 6-5 shows very little separation between cancer and normal disease class.

6.4.4 Identification of most informative antigen concentration feature

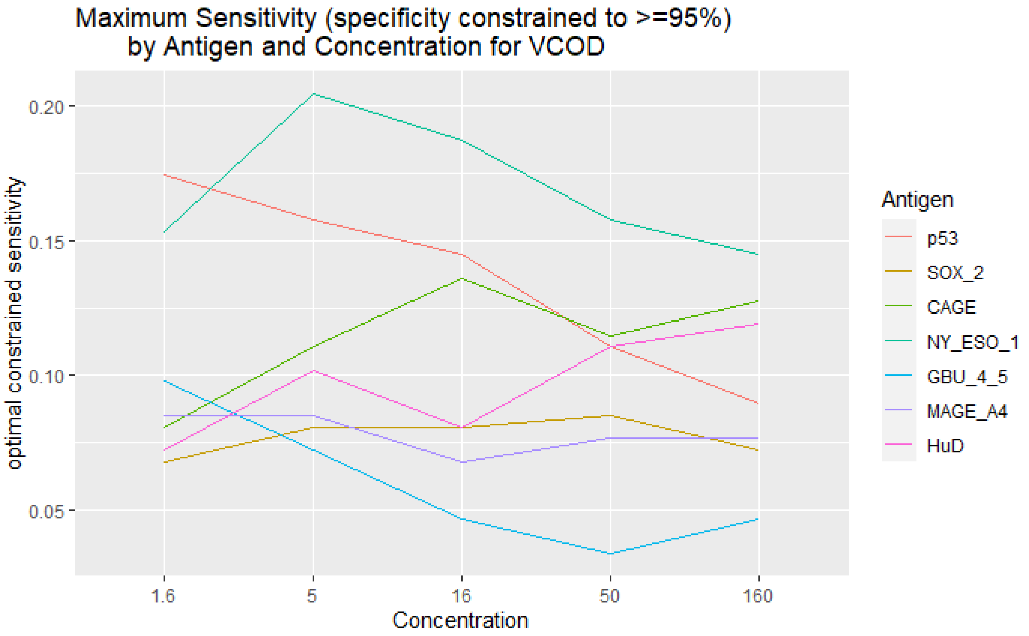


Figure 6-6: Maximum sensitivities returned from cutpoint optimisation for specificities constrained to >=95% for VCOD features.

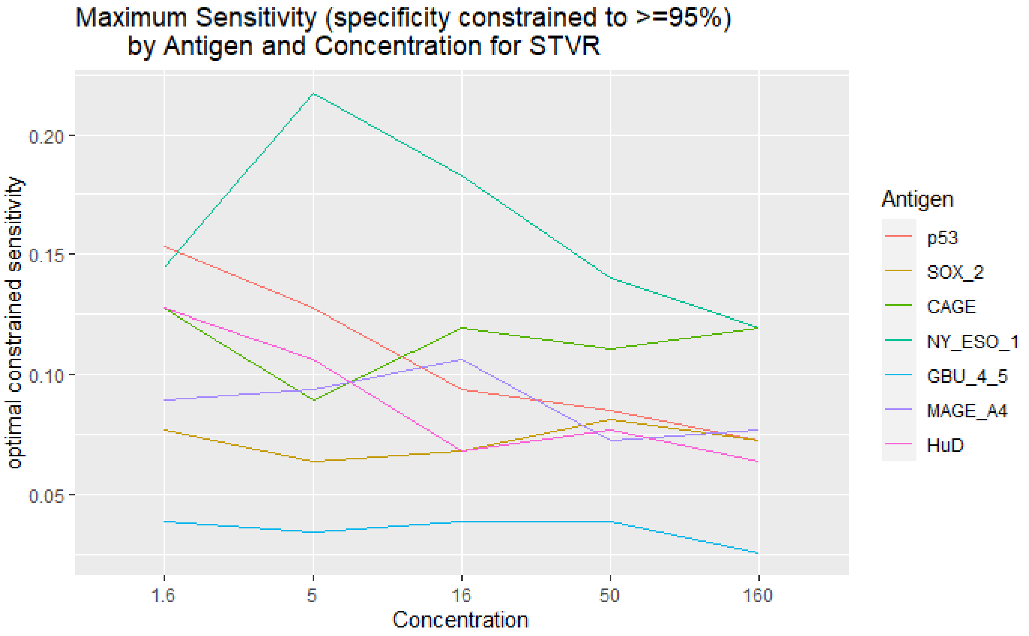


Figure 6-7: Maximum sensitivities returned from cutpoint optimisation for specificities constrained to >=95% for STVR features.

Table 6-27: Concentrations selected for each Antigen for VCOD and STVR features

Antigen	VCOD concentration showing greatest constrained sensitivity	STVR concentration showing greatest constrained sensitivity

p53	1.6	1.6
SOX_2	50	50
CAGE	16	1.6
NY_ESO_1	5	5
GBU_4_5	1.6	50
MAGE_A4	5	16
HuD	160	1.6

6.4.5 Principal Component Analysis on Curve Characteristic Features

Curve Characteristic Features from VOL-Corrected OD Data

Examining the features derived from the curves based on VOL subtracted data, the first component accounts for 19.3% of the variance observed in the dataset, compared to 12.6% accounted for by the second component. Examination of the component loadings in Table 6-28 show the first principal dimension shows high contribution from NY-ESO-1 and HuD derived features, while the second principal dimension has high loading contributions from MAGE-A4 features.

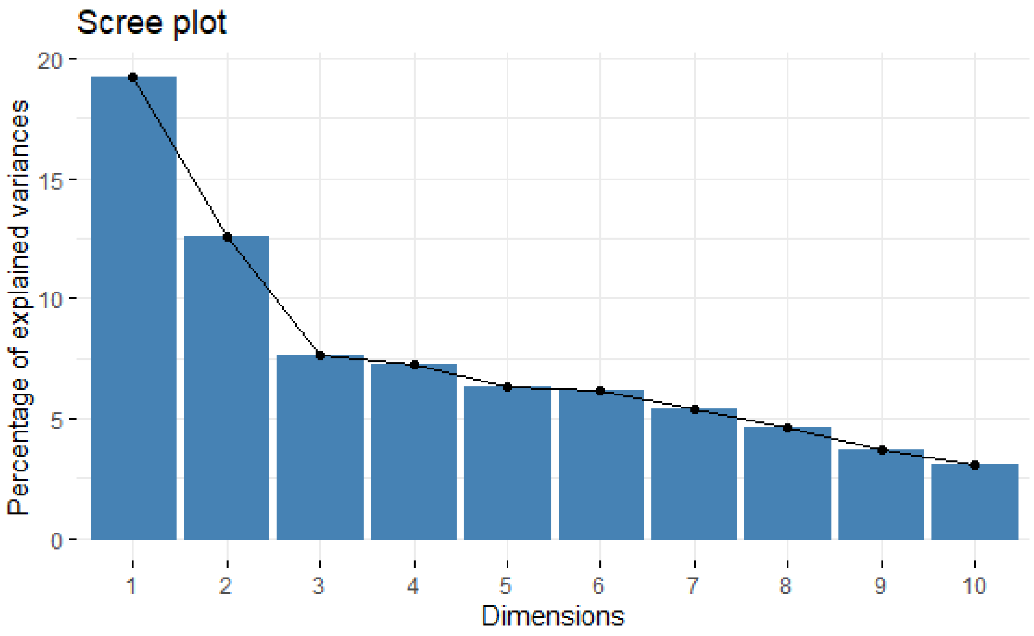


Figure 6-8: Scree plot showing the contribution of the first 10 principal dimensions to the data variance in curve characteristics derived from VOL corrected OD values for the commercial antigen panel.

Table 6-28: PCA Feature Loadings for first 10 principal dimensions – Curve characteristics derived from VOL corrected OD values for the commercial antigen panel.

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7	Dim.8	Dim.9	Dim.10
VCOD_p53	1.923	2.986	0.333	4.137	4.708	4.223	9.680	0.062	0.110	3.992
VCOD_SOX_2	2.153	1.369	0.028	6.597	6.289	1.249	0.730	4.457	2.162	3.987
VCOD_CAGE	2.454	0.403	10.359	0.411	0.332	2.440	3.725	0.047	1.085	4.267
VCOD_NY_ESO_1	3.746	2.751	0.368	5.393	3.232	0.537	6.850	0.234	0.018	0.018
VCOD_GBU_4_5	2.419	1.719	5.777	1.439	3.699	3.828	0.343	0.036	0.120	8.891
VCOD_MAGE_A4	2.341	4.055	0.042	0.996	0.291	13.126	0.224	3.007	3.254	3.272
VCOD_HuD	5.303	2.220	0.917	0.657	1.254	1.625	1.301	10.395	0.465	0.236
vcod_intercept_p53	1.891	3.801	0.162	4.124	5.236	4.006	10.461	0.158	0.146	0.255
vcod_intercept_SOX_2	0.830	0.805	0.026	1.792	5.184	2.438	0.778	1.411	22.113	0.971
vcod_intercept_CAGE	2.218	2.921	8.915	0.159	0.161	3.785	5.927	0.240	2.426	3.623
vcod_intercept_NY_ESO_1	3.002	2.566	1.692	8.797	4.853	0.652	6.974	0.314	0.420	0.444
vcod_intercept_GBU_4_5	2.906	0.961	6.999	2.063	8.406	2.892	2.233	1.977	0.195	1.383
vcod_intercept_MAGE_A4	2.298	5.371	0.230	2.169	0.207	11.872	1.235	5.487	0.950	0.924
vcod_intercept_HuD	1.602	6.449	1.918	0.895	0.603	0.658	2.148	5.530	0.365	3.583
vcod_slope_p53	0.601	4.094	0.588	0.245	2.842	0.172	3.397	0.947	0.751	7.510
vcod_slope_SOX_2	3.207	2.579	1.647	11.095	4.239	0.290	1.431	4.418	5.139	0.375
vcod_slope_CAGE	3.328	0.906	15.151	0.257	0.363	2.163	1.967	0.338	7.408	1.131
vcod_slope_NY_ESO_1	5.281	2.355	0.001	3.126	5.656	0.427	5.153	0.017	0.101	1.382
vcod_slope_GBU_4_5	4.134	4.107	2.569	3.024	4.545	0.983	0.234	3.665	0.305	3.587
vcod_slope_MAGE_A4	2.933	1.592	2.898	3.961	0.231	7.516	0.089	9.200	4.909	5.865
vcod_slope_HuD	5.686	3.105	1.535	1.537	0.980	1.517	0.897	13.832	1.229	0.256
vcod_auc_p53	2.148	3.930	0.191	0.523	2.501	0.004	3.393	0.107	7.093	9.829
vcod_auc_SOX_2	2.200	1.456	1.940	9.149	2.509	0.061	1.983	3.143	12.831	3.109
vcod_auc_CAGE	1.619	1.093	14.600	0.256	0.525	2.220	1.423	0.489	11.904	5.743
vcod_auc_NY_ESO_1	4.543	3.556	0.406	0.856	5.246	0.087	2.049	0.089	0.253	3.513
vcod_auc_GBU_4_5	2.532	6.842	0.683	1.163	1.183	0.372	0.057	2.600	0.356	1.969
vcod_auc_MAGE_A4	1.449	4.286	3.881	0.936	0.252	2.547	0.781	4.062	7.552	10.555
vcod_auc_HuD	4.023	5.069	1.141	2.322	0.840	0.913	0.193	11.174	1.239	2.021
vcod_slopemax_p53	2.731	3.090	0.002	4.560	5.163	3.702	10.533	0.027	0.922	0.004
vcod_slopemax_SOX_2	2.542	0.214	0.639	5.498	7.419	2.122	0.892	0.844	1.578	3.223
vcod_slopemax_CAGE	2.594	2.657	5.537	0.049	0.047	2.459	2.635	0.743	1.436	0.091
vcod_slopemax_NY_ESO_1	4.141	1.926	0.545	8.086	3.421	0.555	4.896	0.588	0.315	1.444
vcod_slopemax_GBU_4_5	3.669	1.077	7.225	1.603	5.952	3.469	2.243	0.798	0.048	2.280
vcod_slopemax_MAGE_A4	2.782	3.774	0.033	1.203	0.368	13.562	1.158	1.636	0.789	0.060
vcod_slopemax_HuD	2.770	3.915	1.024	0.922	1.264	1.529	1.986	7.928	0.015	0.207

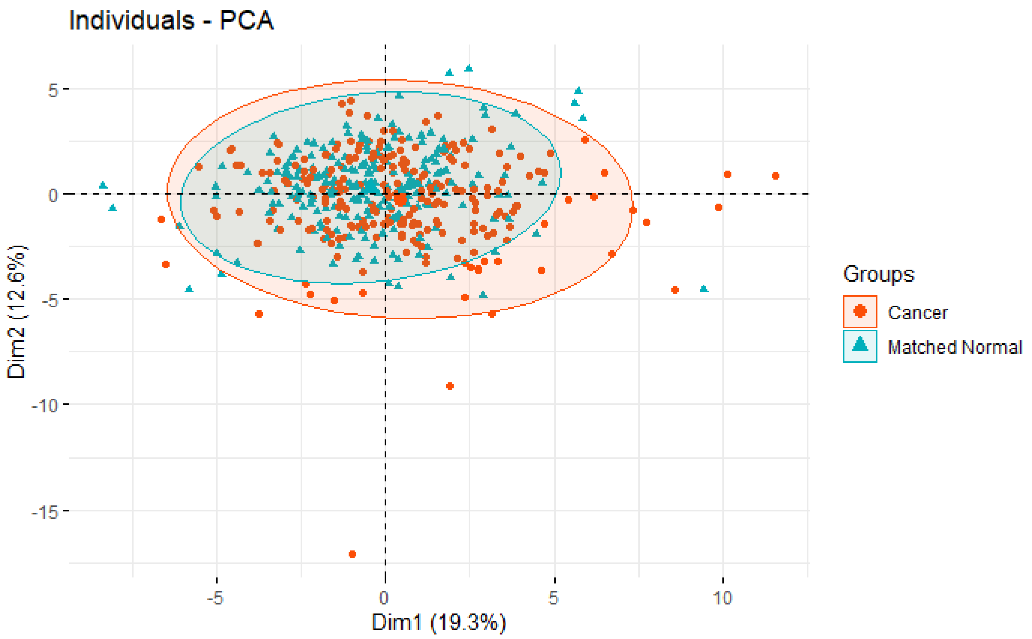
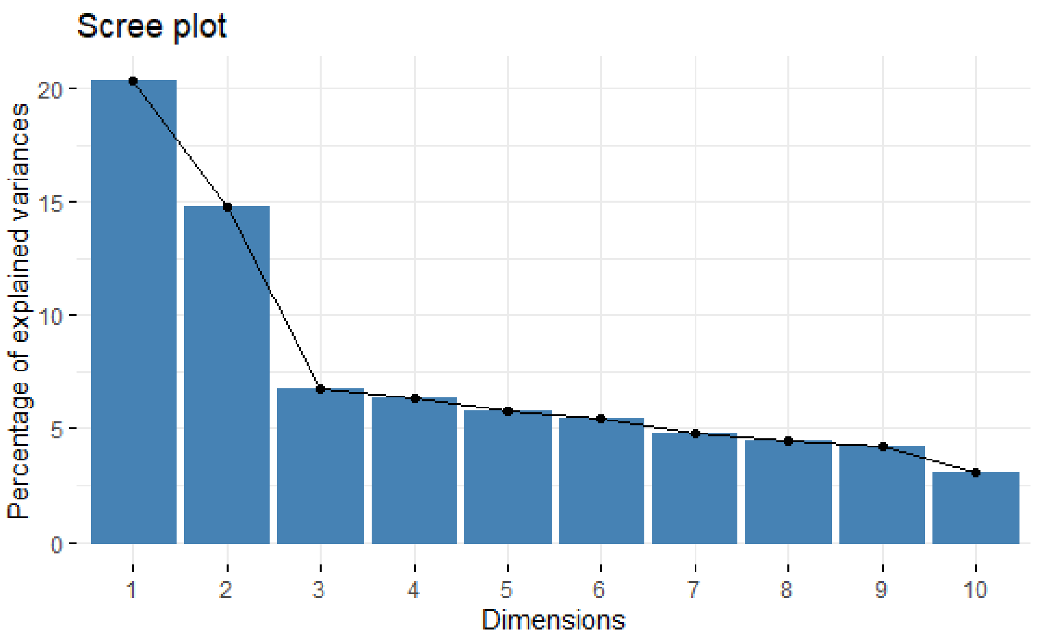


Figure 6-9: Plot of individuals by their component scores on the first two principal dimensions on Vol Corrected curve characteristic data.

Examination of the individuals in Figure 6-9 plotted by their scores on the first two principal dimensions again shows that there is little separation between the cancer and normal samples, although once more the cancer samples distribution is spread further along both dimensions than the distribution of the normal samples.

Curve Characteristic Features from Signal to Vol Ratio expressed OD Data

Examining the features derived from the curves based on Signal to VOL ratio data, the first component accounts for 20.4% of the variance observed in the dataset, compared to 14.8% accounted for by the second component. Examination of the first component loadings highlights that with this correction method, the first principal component has higher loadings for features SOX-2 and GBU 4-5, along with slope and auc features, while the second principal component is comprised of high loadings for the remaining antigens, along with intercept and slopemax features.



	Dim. 1	Dim. 2	Dim. 3	Dim. 4	Dim. 5	Dim. 6	Dim. 7	Dim. 8	Dim. 9	Dim. 10
STVR_p53	0.069	4.489	10.297	2.124	3.963	5.541	0.067	0.648	7.427	0.009
STVR_SOX_2	4.963	0.187	1.139	1.193	1.078	2.303	0.997	11.496	10.950	0.448
STVR_CAGE	0.045	2.822	11.930	0.107	3.906	0.123	0.473	6.571	1.487	3.219
STVR_NY_ESO_1	0.794	5.855	0.080	17.754	0.175	2.725	0.030	0.316	5.218	0.852
STVR_GBU_4_5	4.546	1.276	0.004	3.184	14.804	0.008	6.809	0.084	0.007	0.142
STVR_MAGE_A4	0.252	6.896	1.403	6.859	0.224	7.620	10.765	0.007	0.768	0.480
STVR_HuD	0.003	3.472	2.284	1.168	0.991	7.743	2.646	18.737	0.045	2.641
stvr_intercept_p53	0.013	7.029	5.849	3.075	0.003	2.958	0.419	6.093	4.368	0.707
stvr_intercept_SOX_2	0.212	2.326	6.866	0.521	0.234	0.481	0.302	12.494	6.406	6.636
stvr_intercept_CAGE	0.377	4.913	10.018	0.002	0.429	4.414	2.970	0.466	5.613	0.010
stvr_intercept_NY_ESO_1	0.072	8.222	0.965	14.014	0.218	1.797	0.329	0.029	1.953	0.308
stvr_intercept_GBU_4_5	0.173	5.028	0.000	2.528	18.131	0.074	6.113	0.369	0.053	0.761
stvr_intercept_MAGE_A4	0.012	9.310	0.171	5.563	0.516	5.369	9.689	0.097	0.758	0.934
stvr_intercept_HuD	0.319	7.716	0.191	0.134	1.715	2.397	2.864	8.196	0.391	2.018
stvr_slope_p53	0.469	0.212	0.036	3.425	2.750	0.236	1.182	0.033	2.470	15.126
stvr_slope_SOX_2	7.628	0.206	0.000	1.358	0.723	5.093	2.279	0.839	6.875	5.427
stvr_slope_CAGE	5.311	0.028	0.781	0.012	6.526	6.436	5.644	0.593	3.014	5.958
stvr_slope_NY_ESO_1	7.187	0.040	0.023	6.021	0.748	1.706	0.404	0.209	7.222	1.063
stvr_slope_GBU_4_5	8.722	0.202	0.035	1.336	4.865	0.052	2.870	0.283	0.271	1.719
stvr_slope_MAGE_A4	6.249	0.074	0.063	3.108	0.662	6.010	5.493	1.379	0.015	8.363
stvr_slope_HuD	7.778	0.016	0.003	0.491	0.054	1.488	0.956	4.368	0.061	0.897
stvr_auc_p53	5.248	0.914	0.120	1.230	0.885	0.498	0.247	0.014	1.856	2.903
stvr_auc_SOX_2	4.868	0.750	0.172	1.033	0.733	4.945	1.838	0.999	3.141	11.653

	Dim. 1	Dim. 2	Dim. 3	Dim. 4	Dim. 5	Dim. 6	Dim. 7	Dim. 8	Dim. 9	Dim. 10
stvr_auc_CAGE	3.715	0.530	0.865	0.047	9.172	5.776	5.212	1.953	1.765	6.832
stvr_auc_NY_ESO_1	5.730	0.574	0.004	2.115	0.827	0.410	0.534	0.008	5.027	0.743
stvr_auc_GBU_4_5	8.174	1.902	0.047	0.481	0.186	0.350	0.247	0.529	1.062	1.925
stvr_auc_MAGE_A4	6.196	0.987	0.049	0.003	2.946	1.517	2.077	0.875	0.677	11.581
stvr_auc_HuD	7.371	0.861	0.078	0.478	0.152	1.474	0.654	3.278	0.763	1.747
stvr_slopemax_p53	0.196	4.226	0.619	4.596	6.040	4.281	1.577	6.345	9.871	1.397
stvr_slopemax_SOX_2	0.695	1.010	6.355	0.607	4.234	8.330	0.291	7.502	6.174	0.303
stvr_slopemax_CAGE	0.495	1.704	2.364	0.007	5.231	0.505	7.708	0.002	2.678	0.979
stvr_slopemax_NY_ESO_1	0.976	4.606	5.795	11.098	0.821	0.420	1.626	0.116	1.115	0.077
stvr_slopemax_GBU_4_5	0.636	3.054	8.204	0.898	1.084	1.021	12.380	0.484	0.023	0.905
stvr_slopemax_MAGE_A4	0.151	4.911	12.285	3.013	2.096	0.602	2.288	0.036	0.276	0.030
stvr_slopemax_HuD	0.353	3.653	10.905	0.417	2.878	5.298	0.018	4.552	0.199	1.205

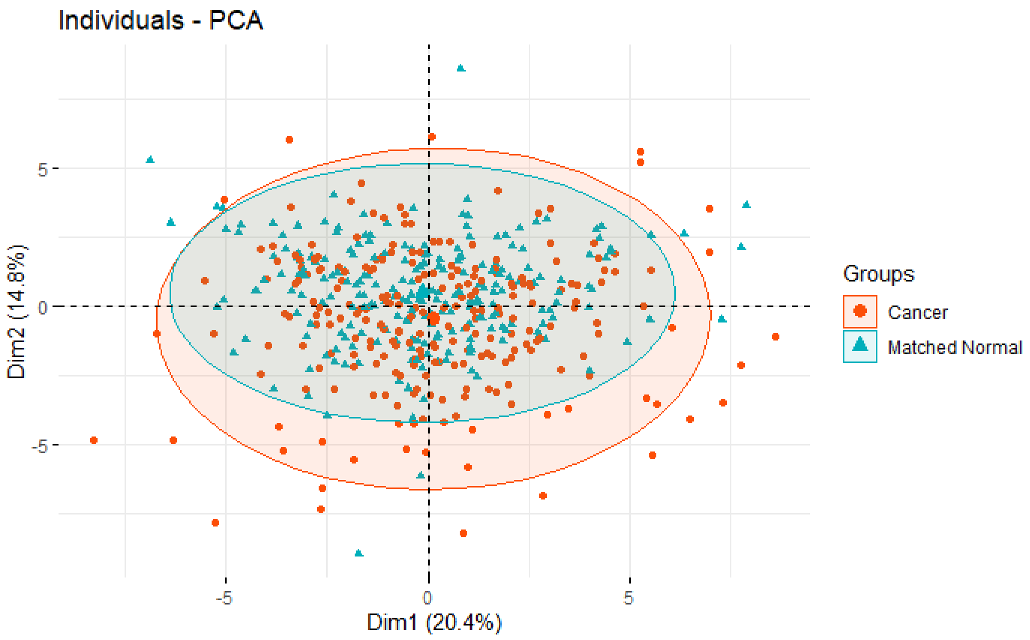


Figure 6-10: Plot of individuals by their component scores on the first two principal dimensions on Signal to Vol Ratio expressed curve characteristic data.

Once more, assessment of the individuals in plotted by their scores on the first two principal dimensions as shown in Figure 6-10 shows that there is little separation between the cancer and normal samples, although once more the cancer samples distribution is wider than that of the normal, especially in the negative direction of the second principal dimension.

6.4.6 Feature Correlation

VOL-Corrected OD Data – Commercial Panel

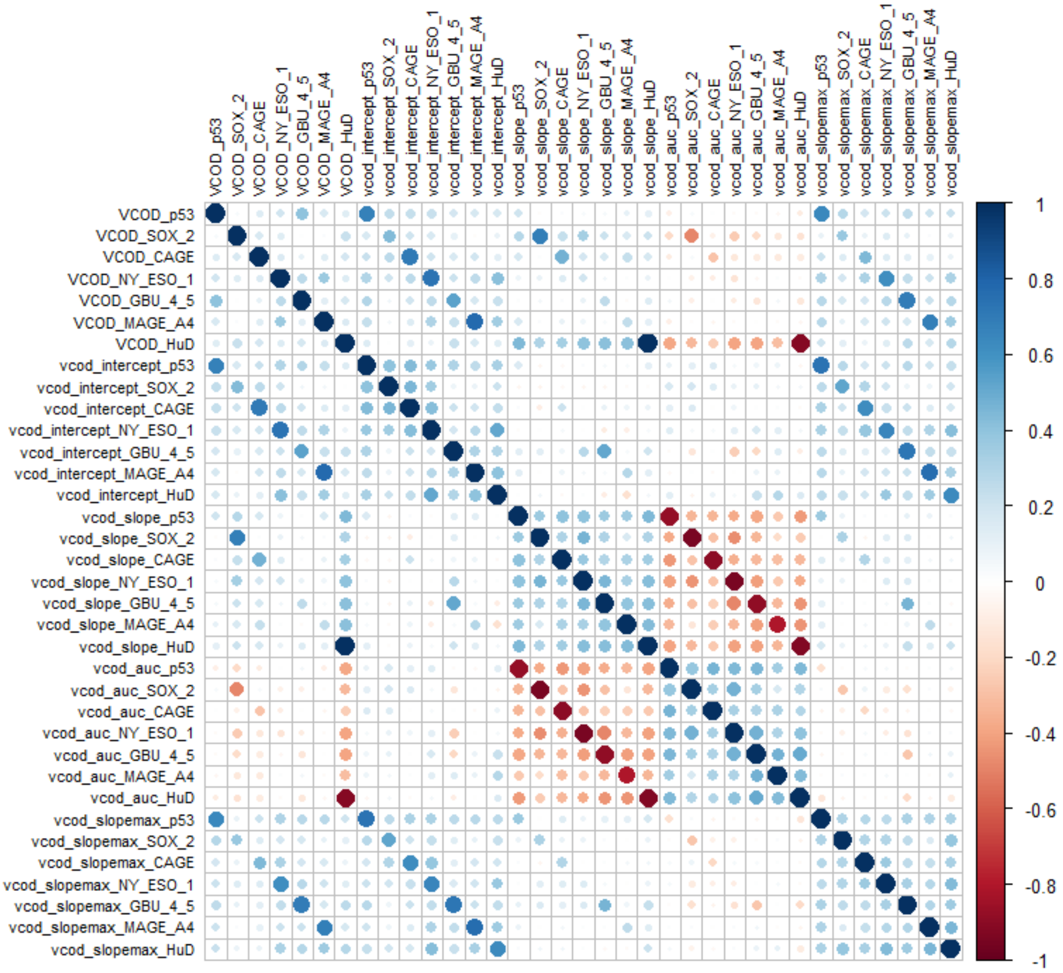


Figure 6-11: Feature correlation in features corrected by VOL subtraction in commercial antigen panel

As illustrated in Figure 6-11, high correlation was shown between corresponding magnitude (VCOD) and intercept features, and magnitude (VCOD) and slope features, however there was a lack of strong correlation between intercept and slope features. Negative correlation was observed between the magnitude (VCOD) features, and area under the curve (auc) features, with particularly high negative correlation being observed between slope and auc features, this suggests that samples with high area under the curve tend to be elevated at a lower magnitude than samples with high slope values, but with maintained elevation along the entire concentration range, while samples with particularly high slope values may have a tendency to show low autoantibody binding at lower concentrations.

Signal To VOL Ratio Data – Commercial Panel

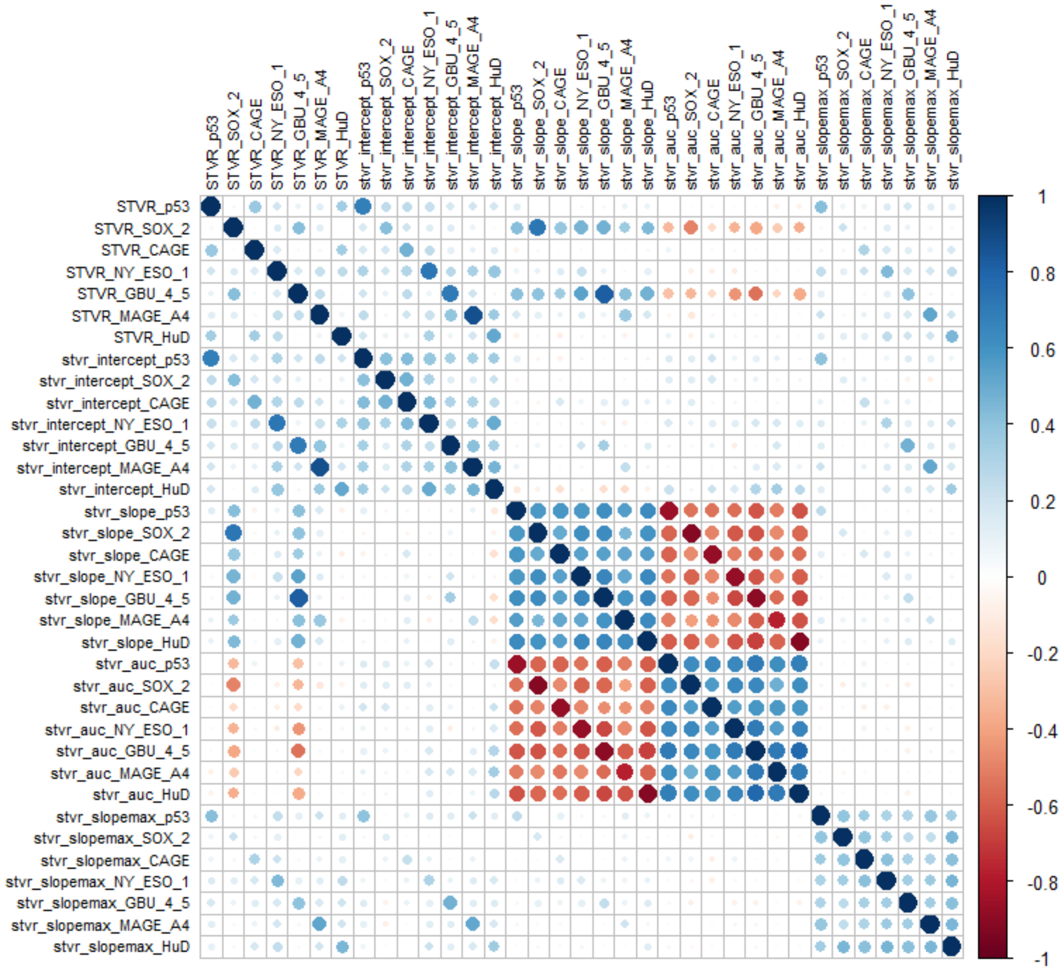


Figure 6-12: Feature correlation in features corrected by re-expression as ratio of specific signal to non-specific signal (Signal To VOL Ratio (STVR))

As illustrated in Figure 6-12, high correlation was shown between corresponding magnitude (VCOD) and intercept features, and magnitude (VCOD) and slope features, however there was a lack of strong correlation between intercept and slope features. Negative correlation was observed between the magnitude (VCOD) features, and area under the curve (auc) features, with particularly high negative correlation being observed between slope and auc features, this suggests that samples with high area under the curve tend to be elevated at a lower magnitude than samples with high slope values, but with maintained elevation along the entire concentration range, while samples with particularly high slope values may have a tendency to show low autoantibody binding at lower concentrations.

6.4.7 Cluster Analysis

VCOD Cluster Analysis

The optimal number of clusters to separate the dataset was determined by iterating through clustering a values of K from 1 to 20, and calculating a value for the Bayesian Information Criterion(153) (BIC) at each value of K. BIC was selected as it is known to outperform the normally used `Elbow Method`(154) for identifying the optimal number of clusters for k-means. The k number identified was then used for k-means clustering, with the resultant clusters then undergoing unblinding with respects to disease classification, to establish whether the clustering corresponds to disease presence. Figure 6-13 illustrates the result of this analysis and shows that a k value of 13 clusters was selected as being most informative for this dataset.

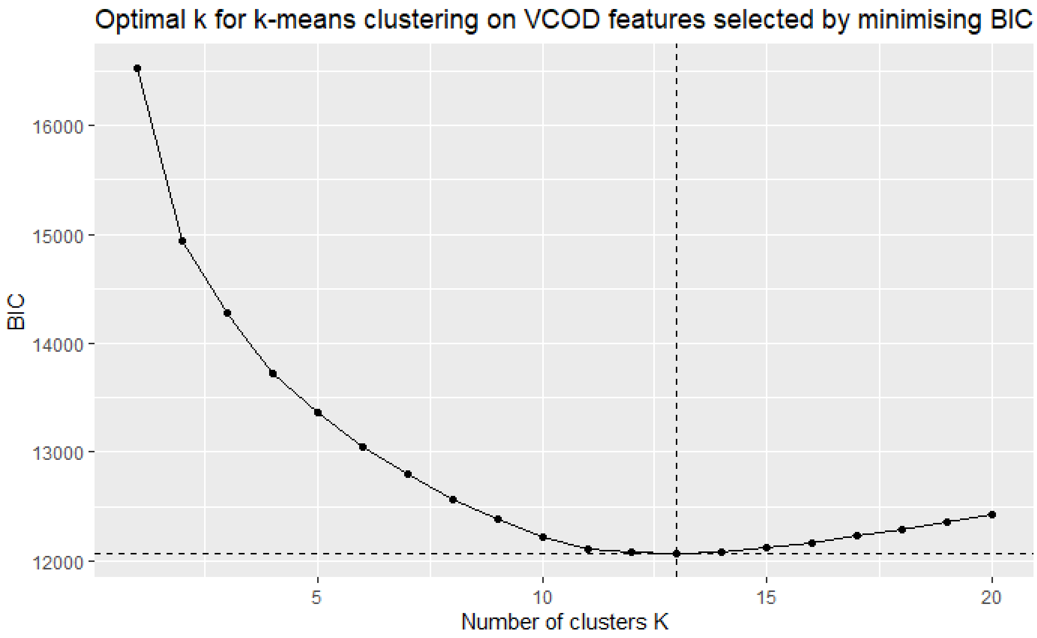


Figure 6-13: Bayesian Information Criterion scores for k-means clustering on VCOD features with k values from 1 to 20

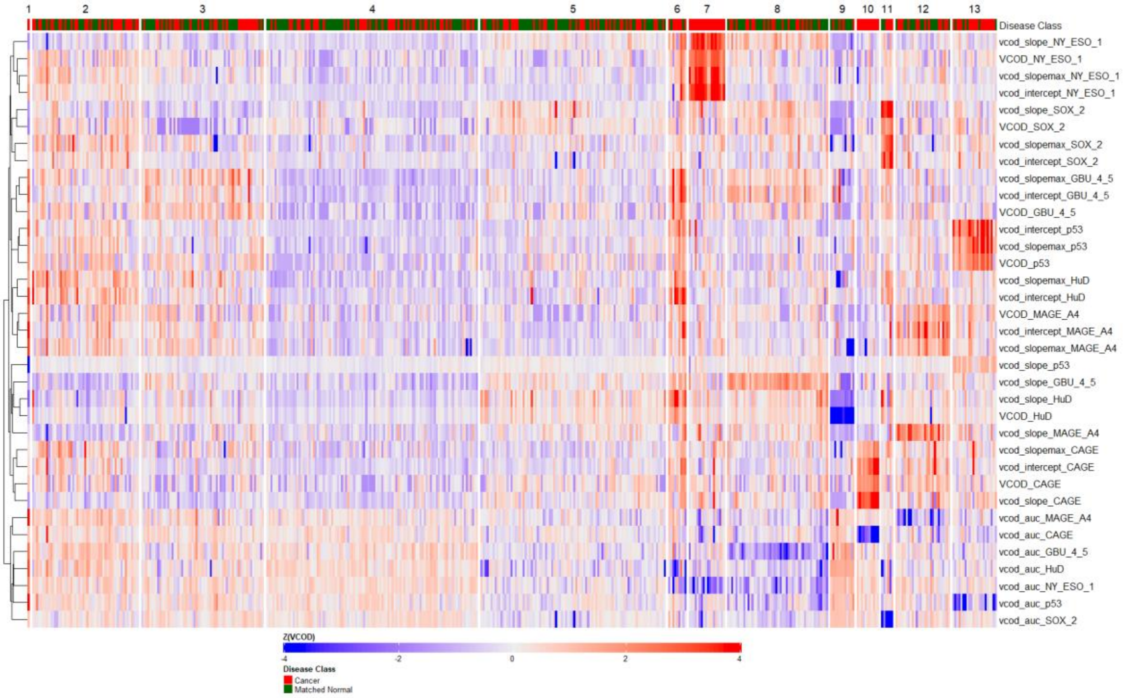


Figure 6-14: Heatmap showing K Means clustering results for VCOD features with k=13 clusters.

The heatmap in Figure 6-14 shows several clusters have high proportions of cancer samples contained within them, cluster 7 is comprised entirely of cancer samples that are characterised by high signal in NY-ESO-1 features, similarly cluster 10 identifies cancer samples with high signal in CAGE features, and clusters 6 and 13 identify samples with high GBU 4-5 and p53 signal respectively, although a number of normal samples are also captured in these clusters. Combining the clusters that are comprised of greater than 60% cancer samples (Clusters referred to as 1, 6, 7, 10, 11, and 13 in Table 6-29) in order to give a diagnostic assessment would correctly classify 58 cancer samples, and 223 normal samples, for a sensitivity of 24.7%, and specificity of 96.1%, therefore would not reach levels of diagnostic accuracy that could be considered an improvement on the current EarlyCDT test classification method.

Table 6-29: Disease class composition of VCOD feature clusters determined by k-means

Cluster Number	Cancers Present	Normals Present	Cancer %
1	1	0	100%
2	30	23	57%
3	29	32	48%
4	45	60	43%
5	38	54	41%
6	7	2	78%
7	18	0	100%
8	19	31	38%
9	4	8	33%
10	11	0	100%
11	4	2	67%
12	12	15	44%
13	17	5	77%

STVR Cluster Analysis

The same strategy using BIC to identify the optimal value for k was then followed for the STVR features. Figure 6-15 illustrates the result of this analysis and shows that a k value of 16 clusters was selected as being most informative for the STVR dataset.

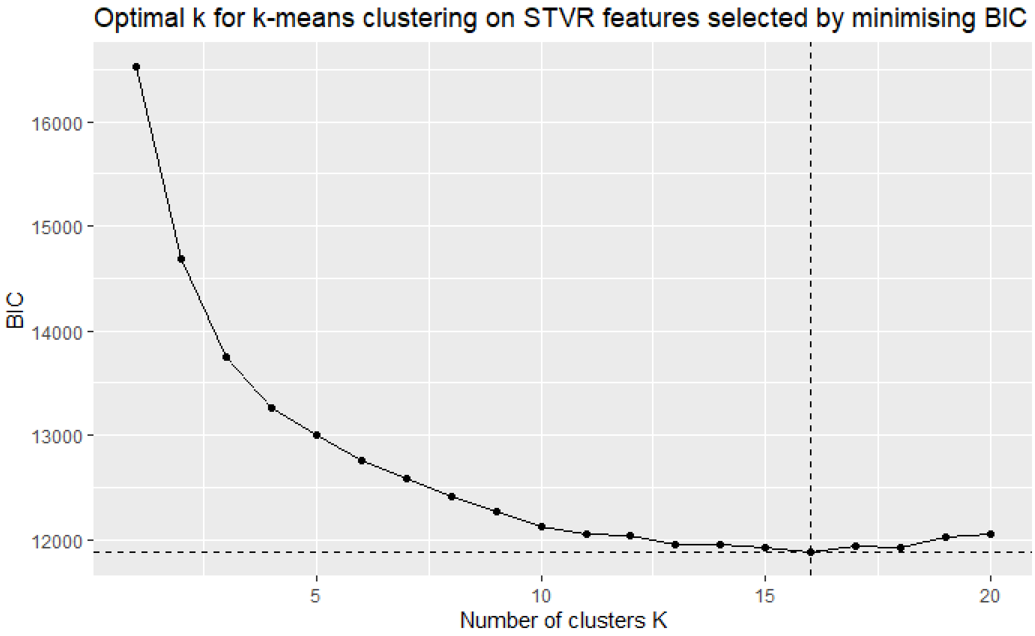


Figure 6-15: Bayesian Information Criterion scores for k-means clustering on STVR features with k values from 1 to 20

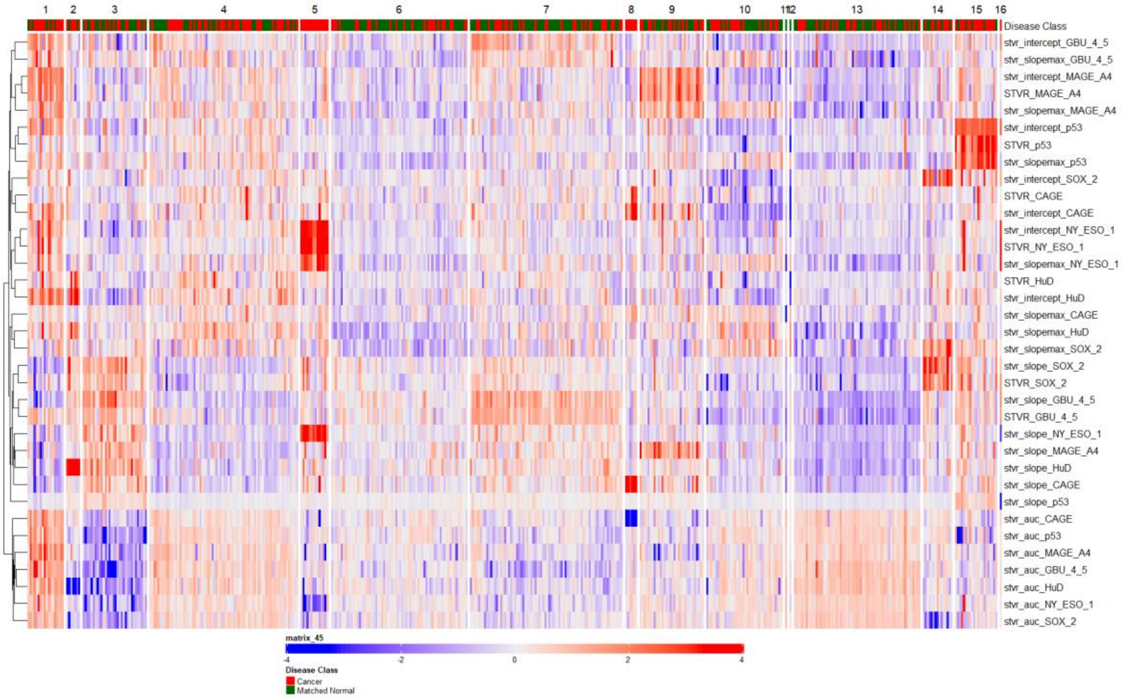


Figure 6-16: Heatmap showing K Means clustering results for VCOD features with k=16 clusters.

The heatmap in Figure 6-16 again shows several clusters have high proportions of cancer samples contained within them, although the numbers of cancer samples identified is smaller in this case. It can be seen that cluster 5 is comprised of cancer samples that are characterised by high signal in NY-ESO-1 features, much like cluster 7 in the VCOD cluster analysis, however only 14 samples are identified in the cluster in this case, compared to 18 in the VCOD analysis. Similarly, cluster 8 in this analysis identifies cancer samples with high signal in CAGE features, paralleling cluster 10 in the previous analysis, and again in this case fewer samples are identified, with 6 cancer samples being identified in the STVR analysis compared to 11 in the VCOD clusters. Once again combining the clusters that are comprised of greater than 60% cancer samples (Clusters referred to as 1, 2, 5, 8, 14, 15, and 16 in Table 6-30Table 6-29) in order to give a diagnostic assessment would correctly classify 64 cancer samples, and 214 normal samples, giving a sensitivity of 27.2%, and specificity of 92.2%, again not reaching levels of diagnostic accuracy that would be considered an improvement on the current EarlyCDT test classification method.

Table 6-30: Disease class composition of STVR feature clusters determined by k-means

Cluster Number	Cancer	Matched Normal	Cancer %
1	13	5	72%
2	5	2	71%
3	14	18	44%
4	42	32	57%
5	14	0	100%
6	25	43	37%
7	29	47	38%
8	6	0	100%
9	16	16	50%
10	19	19	50%
11	0	1	0%
12	0	1	0%
13	26	37	41%
14	9	6	60%
15	16	5	76%
16	1	0	100%

6.5 Chapter Conclusions

Investigation into the distributions of the autoantibody biomarker data revealed high degrees of non-normality and high kurtosis in all tested features, along with high positive skewness for all features, with the exception of auc features, which showed high negative skewness. As most modelling strategies are based on the assumption of normality in the examined data, the features underwent normalisation and scaling to remove any influence of feature distribution or magnitude from having a detrimental effect on model training.

PCA analysis of the uncorrected data showed that high levels of variance in the dataset were attributable to non-specific binding, as both ODs returned in response to the control protein VOL, and ODs returned by the low and 0nM antigen wells constituted the first principal dimension. As it is not possible to fully separate specific and non-specific binding in this data, analysis of uncorrected data is likely to train models only on artefacts in the non-specific binding which would be unreproducible in subsequent datasets, therefore due to this high influence of non-specific binding, uncorrected data will not be considered in subsequent machine learning analyses. Corrected datasets still showed a small influence of non-specific binding during PCA assessment, evidenced by high loadings for the 0nM autoantibody values in the second principal dimension for both vol subtracted and signal to vol ratio adjusted data, although they no longer had high loadings in the first principal dimension, therefore non-specific binding was no longer the greatest source of variance.

K-means cluster analysis illustrated that the data did show clustering based on autoantibody profiles, however a relatively large number of clusters were required to categorise the data, possibly due to the heterogeneity of lung cancer and the associated immune response, and while it would be possible to define a model based on these clusters, the performance is inferior to that returned by the current EarlyCDT®-Lung test.

6.6 Chapter Discussion

Initial unsupervised modelling was not able to show underlying patterns in the data distributions that could be exploited for diagnostic purposes. While this was disappointing, it was not entirely unexpected given the low sensitivity of individual autoantibody features. Cluster analysis did show some potential for disease discrimination and although it was unable to improve upon current commercial performance it does suggest that supervised strategies may be successful at disease discrimination.

Demographic features were not included in the principal components analysis, and therefore the potential confounding effect of variables such as age, gender, sample supplier, and sample country of origin have not been assessed here. Any future reanalysis of this data should include an appreciation of the effect of these confounders through use of factor analysis of mixed data (FAMD) analysis to determine whether these variables are influencing the data.

Chapter 7: Exploration of Supervised Machine Learning Strategies for Early Lung Cancer Diagnosis.

7.1 Aims

With unsupervised strategies not showing ability to classify data, I proceeded to explore supervised machine learning strategies for their ability to train predictive models. I wanted to apply a variety of established strategies and try to determine which showed the greatest potential for predicting disease class based on autoantibody data. Modelling strategies that were able to outperform the current logic-based model strategy would then be developed further in an effort to construct a commercial diagnostic algorithm.

7.2 Introduction

Having determined that unsupervised learning techniques are unable to adequately separate cancer and normal subjects based on their autoantibody magnitude values and derived curve characteristics, an investigation was carried out into whether supervised learning techniques may be applied to the magnitude and curve characteristic data generated from the EarlyCDT®-Lung panel, and used to create a model which is able to distinguish cancer samples from normal controls at a higher sensitivity and specificity than the current commercial strategy, whereby a positive response is defined as having an autoantibody response above a predetermined cut-off threshold for any one of the panel autoantibody measures.

Machine learning algorithms to explore have been selected to cover a range of different popular techniques which have been used previously for disease classification problems, and include logistic regression, support vector machine learning, naïve Bayes, decision trees, random forest, and extreme gradient boosted regression trees, these are discussed in more detail in their relevant sections, however, in brief logistic regression was explored due to its extensive use historically for binary classification problems, and proven utility in previous diagnostic tests(82, 83), similarly Naïve Bayes was explored due to it's previous successful use in diagnostic tests(90, 91), and the added benefit of the output of probability score which may be integrated into ensemble models. Support Vector Machines were considered as the expectation regarding the autoantibody panel data is that, due to high individual autoantibody specificities, the majority of sample autoantibody responses would present as non-specific binding, and therefore be relatively similar in n-dimensional featurespace, specific autoantibody responses would therefore present as outliers to the cloud of non-specific signal which could then potentially be distinguished through the use of a support vector hyperplane. Decision Tree models were explored due to their similarity to the current logic rule based algorithm, as the current test could be expressed as an extremely simple decision tree, with random forest models being explored as a development from decision tree models, whereby bagging and averaging of large numbers of small trees may allow for more accurate models, and similarly extreme gradient boosted trees then being explored as a refinement of random forest models thanks to the addition of boosting whereby additional models are trained on the errors from previous modelling to build an ensemble model.

7.3 Features

7.3.1 Magnitude Features

The EarlyCDT®-Lung test collected autoantibody response values for a panel of 7 autoantibodies; p53, SOX2, CAGE, NY-ESO-1, GBU 4-5, MAGE-A4, and HuD, over a dilution range 1.6nM, 5nM, 16nM, 50nM, and 160nM. A previous analysis determined the single dilution value which showed the greatest discriminatory ability for each autoantibody from these values, and the concentration referred to as magnitude features in the subsequent analyses are as follows: for vol corrected OD (VCOD) data; p53 at 1.6nM, SOX2 at 50nM, CAGE at 16nM, NY-ESO-1 at 5nM, GBU4-5 at 1.6nM, MAGE-A4 at 5nM, and HuD at 160nM. For signal to vol ratio (STVR) data; p53 at 1.6nM, SOX2 at 50nM, CAGE at 1.6nM, NY-ESO-1 at 5nM, GBU4-5 at 150nM, MAGE-A4 at 16nM, and HuD at 1.6nM. All magnitude features underwent transformation and scaling as described previously to ensure data approximated a normal distribution, and to prevent differences in feature scale from detrimentally influencing the modelling.

7.3.2 Curve Characteristic Features

Curve characteristic features refer to features derived from the shape of the overall titration curve for each antigen, as described previously, including linear regression derived features slope and intercept, area under the curve as calculated using sum of trapezoids, and the slopemax, defined as the slope at the steepest point of the titration curve. The Curve Characteristic feature set in the following analyses also includes the magnitude features. This is to ensure curve characteristic features – which require more data points, and a higher level of data processing – are not included if the magnitude feature is able to provide the same discriminatory value.

All Curve characteristic features underwent transformation and scaling as described in section 6.3 to ensure data approximated a normal distribution, and to prevent differences in feature scale from detrimentally influencing the modelling.

7.4 Individual Feature Discriminatory Performance

The discriminatory ability of the magnitude and curve characteristic features in both VCOD and STVR feature sets has been summarised by identifying an optimal cutpoint which maximised the Youden index for the training cohort, constrained to specificities above 90%, using the R cutpointr package (v1.1.2) to iterate over all features, after which the optimal cutpoint was applied to the hold-out test cohort to return performance characteristics.

7.4.1 Vol Subtracted (VCOD) Features

Summary values for accuracy in the training cohort (training accuracy), area under the ROC curve in the training cohort (training AUC), and sensitivity and specificity at the optimal cutoff threshold in both training cohort (training sensitivity and training specificity respectively), and test cohort (test sensitivity and test specificity respectively) are summarised in Table 7-1, showing that for the majority of features, specificity is maintained in the test cohort, with a mean decrease in the test cohort of only 0.6%, and the largest change being in the HuD magnitude feature in which the specificity decreased to 82.7% compared to 90.9% in the training cohort. Sensitivity does show a small drop off in the test cohort, with a mean decrease of 3.7% across all features, and the largest reduction being shown in the CAGE slopemax feature in which the sensitivity reduction is 10.7%. Summary ROC plots were constructed to allow comparison of the features discriminatory ability over the dynamic range of the assay, and these are included in appendix 1.

Table 7-1: Discriminatory ability of individual Vol corrected magnitude and curve characteristic features, optimised over training cohort and applied to test cohort.

Feature	optimal cutpoint	training accuracy	training AUC	training sens	training spec	test sens	test spec
VCOD_p53	1.016	0.559	0.537	21.7%	90.5%	13.0%	90.8%
VCOD_SOX_2	0.929	0.505	0.503	11.1%	90.5%	14.0%	85.7%
VCOD_CAGE	0.960	0.537	0.527	17.0%	90.9%	11.0%	90.8%
VCOD_NY_ESO_1	0.663	0.582	0.607	26.8%	90.1%	25.0%	85.7%
VCOD_GBU_4_5	1.039	0.525	0.514	15.3%	90.1%	8.0%	89.8%
VCOD_MAGE_A4	1.039	0.527	0.547	15.7%	90.1%	13.0%	90.8%
VCOD_HuD	0.717	0.533	0.533	16.2%	90.9%	9.0%	82.7%
vcod_intercept_p53	0.754	0.550	0.549	19.6%	90.9%	14.0%	93.9%
vcod_intercept_SOX_2	0.722	0.512	0.535	12.8%	90.1%	16.0%	89.8%
vcod_intercept_CAGE	0.731	0.542	0.554	18.7%	90.1%	13.0%	88.8%
vcod_intercept_NY_ESO_1	0.347	0.574	0.610	23.4%	91.8%	19.0%	90.8%
vcod_intercept_GBU_4_5	1.209	0.505	0.497	11.5%	90.1%	6.0%	89.8%
vcod_intercept_MAGE_A4	0.913	0.540	0.564	18.3%	90.1%	11.0%	88.8%
vcod_intercept_HuD	0.850	0.529	0.563	16.2%	90.1%	13.0%	86.7%
vcod_slope_p53	0.432	0.512	0.504	12.8%	90.1%	15.0%	88.8%
vcod_slope_SOX_2	1.122	0.499	0.491	10.2%	90.1%	12.0%	88.8%
vcod_slope_CAGE	0.807	0.542	0.550	18.7%	90.1%	13.0%	82.7%
vcod_slope_NY_ESO_1	0.855	0.542	0.515	18.7%	90.1%	12.0%	93.9%
vcod_slope_GBU_4_5	1.493	0.493	0.465	8.9%	90.1%	8.0%	93.9%
vcod_slope_MAGE_A4	1.092	0.525	0.541	15.3%	90.1%	7.0%	88.8%
vcod_slope_HuD	0.996	0.525	0.527	14.5%	90.9%	9.0%	81.6%
vcod_auc_p53	0.785	0.507	0.499	11.9%	90.1%	11.0%	96.9%
vcod_auc_SOX_2	0.723	0.520	0.514	14.0%	90.5%	14.0%	93.9%
vcod_auc_CAGE	0.758	0.516	0.444	13.6%	90.1%	14.0%	89.8%
vcod_auc_NY_ESO_1	0.900	0.514	0.483	12.8%	90.5%	12.0%	94.9%
vcod_auc_GBU_4_5	0.989	0.520	0.542	13.6%	90.9%	9.0%	89.8%
vcod_auc_MAGE_A4	0.865	0.497	0.457	9.8%	90.1%	10.0%	91.8%
vcod_auc_HuD	0.892	0.497	0.489	9.4%	90.5%	11.0%	83.7%
vcod_slopemax_p53	0.869	0.546	0.555	19.1%	90.5%	14.0%	91.8%
vcod_slopemax_SOX_2	1.119	0.514	0.537	13.2%	90.1%	13.0%	92.9%
vcod_slopemax_CAGE	1.029	0.542	0.563	18.7%	90.1%	8.0%	86.7%

Feature	optimal cutpoint	training accuracy	training AUC	training sens	training spec	test sens	test spec
vcod_slopemax_NY_ESO_1	0.697	0.548	0.566	19.6%	90.5%	13.0%	91.8%
vcod_slopemax_GBU_4_5	1.100	0.518	0.507	13.6%	90.5%	8.0%	87.8%
vcod_slopemax_MAGE_A4	1.065	0.518	0.556	14.0%	90.1%	10.0%	92.9%
vcod_slopemax_HuD	1.059	0.533	0.562	16.6%	90.5%	6.0%	93.9%

7.4.2 Signal to Vol Ratio (STVR) Features

Summary values for accuracy in the training cohort (training accuracy), area under the ROC curve in the training cohort (training AUC), and sensitivity and specificity at the optimal cutpoint in both training cohort (training sensitivity and training specificity respectively), and test cohort (test sensitivity and test specificity respectively) for STVR features are summarised in Table 7-2, showing that for the majority of features, specificity is maintained in the test cohort, with a mean decrease in the test cohort of only 0.4%, and the largest change being in the HuD slope feature in which the specificity decreased to 81.6% compared to 90.9% in the training cohort. Sensitivity does show a small drop off in the test cohort, with a mean decrease of 2.3% across all features, and the largest reduction being shown in the p53 magnitude feature in which the sensitivity reduction is 9.0%. Summary ROC plots were constructed to allow comparison of the features discriminatory ability over the dynamic range of the assay, and these are included in appendix 2.

Table 7-2: Discriminatory ability of individual Signal to Vol Ratio magnitude and curve characteristic features, optimised over training cohort and applied to test cohort.

Feature	optimal cutpoint	training accuracy	training AUC	training sens	training spec	test sens	test spec
STVR_p53	0.491	0.565	0.538	23.0%	90.5%	14.0%	87.8%
STVR_SOX_2	0.933	0.522	0.508	14.9%	90.1%	18.0%	89.8%
STVR_CAGE	0.683	0.548	0.549	19.6%	90.5%	17.0%	84.7%
STVR_NY_ESO_1	0.277	0.585	0.600	27.2%	90.1%	28.0%	86.7%
STVR_GBU_4_5	1.229	0.497	0.476	9.4%	90.5%	15.0%	86.7%

Feature	optimal cutpoint	training accuracy	training AUC	training sens	training spec	test sens	test spec
STVR_MAGE_A4	1.065	0.531	0.528	16.6%	90.1%	10.0%	91.8%
STVR_HuD	0.625	0.546	0.545	19.6%	90.1%	12.0%	89.8%
stvr_intercept_p53	1.040	0.540	0.539	18.3%	90.1%	14.0%	91.8%
stvr_intercept_SOX_2	0.746	0.520	0.503	14.5%	90.1%	15.0%	90.8%
stvr_intercept_CAGE	0.893	0.522	0.550	14.0%	90.9%	12.0%	94.9%
stvr_intercept_NY_ESO_1	0.558	0.582	0.605	26.4%	90.5%	19.0%	91.8%
stvr_intercept_GBU_4_5	1.304	0.495	0.488	9.4%	90.1%	10.0%	91.8%
stvr_intercept_MAGE_A4	1.170	0.527	0.558	15.7%	90.1%	9.0%	89.8%
stvr_intercept_HuD	0.857	0.550	0.560	20.4%	90.1%	13.0%	84.7%
stvr_slope_p53	0.380	0.499	0.503	10.2%	90.1%	12.0%	90.8%
stvr_slope_SOX_2	0.904	0.520	0.498	14.5%	90.1%	20.0%	86.7%
stvr_slope_CAGE	0.817	0.531	0.544	16.6%	90.1%	15.0%	84.7%
stvr_slope_NY_ESO_1	0.675	0.542	0.514	18.7%	90.1%	17.0%	89.8%
stvr_slope_GBU_4_5	1.385	0.482	0.462	6.4%	90.5%	13.0%	86.7%
stvr_slope_MAGE_A4	1.216	0.507	0.525	11.9%	90.1%	11.0%	91.8%
stvr_slope_HuD	1.017	0.525	0.506	14.5%	90.9%	13.0%	81.6%
stvr_auc_p53	0.734	0.518	0.518	13.6%	90.5%	6.0%	96.9%
stvr_auc_SOX_2	0.756	0.512	0.511	12.8%	90.1%	11.0%	92.9%
stvr_auc_CAGE	0.676	0.518	0.461	14.0%	90.1%	12.0%	92.9%
stvr_auc_NY_ESO_1	0.686	0.518	0.512	14.0%	90.1%	10.0%	96.9%
stvr_auc_GBU_4_5	0.886	0.518	0.533	14.0%	90.1%	9.0%	90.8%
stvr_auc_MAGE_A4	0.807	0.512	0.488	12.8%	90.1%	7.0%	93.9%
stvr_auc_HuD	0.922	0.505	0.516	11.5%	90.1%	9.0%	86.7%
stvr_slopemax_p53	0.918	0.552	0.571	20.4%	90.5%	17.0%	94.9%
stvr_slopemax_SOX_2	1.375	0.501	0.555	10.6%	90.1%	8.0%	94.9%
stvr_slopemax_CAGE	0.729	0.540	0.578	18.3%	90.1%	10.0%	90.8%
stvr_slopemax_NY_ESO_1	0.786	0.550	0.581	20.4%	90.1%	18.0%	88.8%
stvr_slopemax_GBU_4_5	1.182	0.505	0.537	11.5%	90.1%	9.0%	93.9%
stvr_slopemax_MAGE_A4	1.096	0.527	0.548	15.7%	90.1%	12.0%	92.9%
stvr_slopemax_HuD	0.922	0.542	0.574	18.7%	90.1%	13.0%	91.8%

7.5 Boruta Feature Selection

7.5.1 Introduction

While the majority of the modelling strategies explored in this investigation have internal feature selection to optimise the final model feature set, an additional feature selection method has been explored in order to identify a reduced feature set for modelling, to determine whether further improvements can be obtained by removing uninformative features prior to model training. The Boruta feature selection algorithm(155) has been selected as it is able to work with classification problems and identifies all features which are relevant to the outcome variable, rather than aiming to minimise the feature set error and reducing the feature set to the smallest subset of features. Boruta is a wrapper method around a random forest modelling algorithm, determining feature importance by comparing features to permuted ‘shadow’ features. In this investigation, feature importance was determined over 100 iterations of the random forest model, and features whose importance was significantly higher than the maximum importance of the shadow features – as determined by a Mean Decrease Accuracy measure – were selected as important and added to a `Boruta Feature Set` to be assessed alongside the magnitude features, and the full feature set of magnitude and curve characteristic features.

7.5.2 Methods

Boruta analysis was undertaken in R (v4.2.1) using the Boruta library, with all features showing a significantly higher importance than that of the maximum importance of the shadow features being included in the subsequent Boruta feature sets. This was completed for both VCOD features, and STVR features. Any features which showed importance that was not significantly different to the importance of the shadow features were considered as tentative candidates. These were reclassified by comparing their median importance over the entire run compared to the median value of the maximum importance of the shadow features, features showing greater importance than the shadow features were retained.

7.5.3 Results

VCOD Features

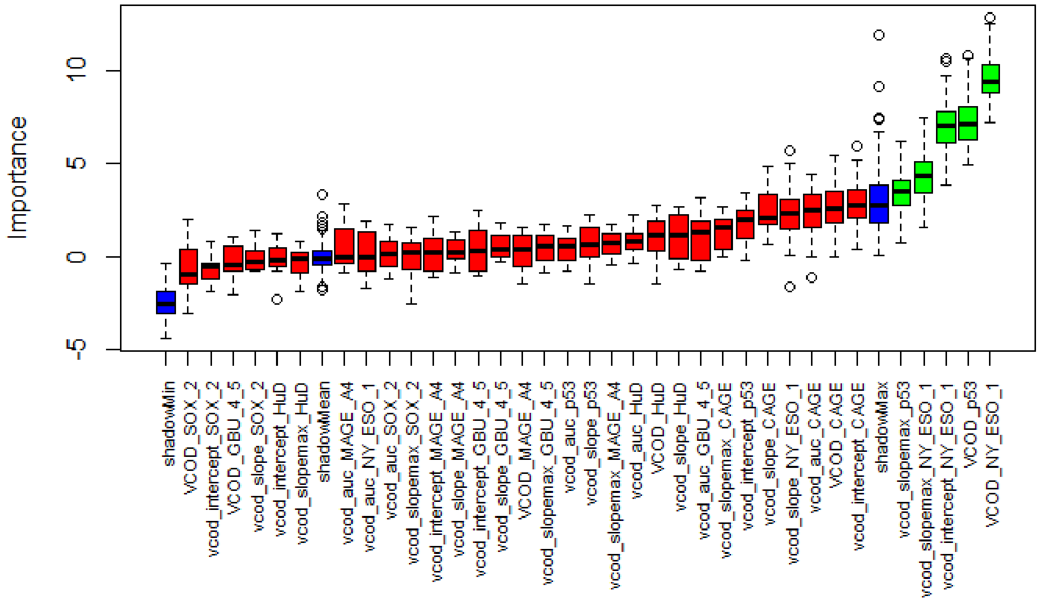


Figure 7-1: Boruta feature selection undertaken on VCOD magnitude and curve characteristic features, Importance was calculated as z-score of the mean accuracy decrease.

Boruta selection on the VCOD feature set identified only 5 features which were considered important after comparison to shadow features, the full results are displayed in Figure 7-1, which shows that the magnitude features for NY-ESO-1 and p53, the intercept feature for NY-ESO-1, and the slopemax features for NY-ESO-1 and p53 were classified as important and will be examined as distinct feature set in the subsequent modelling analyses.

STVR Features

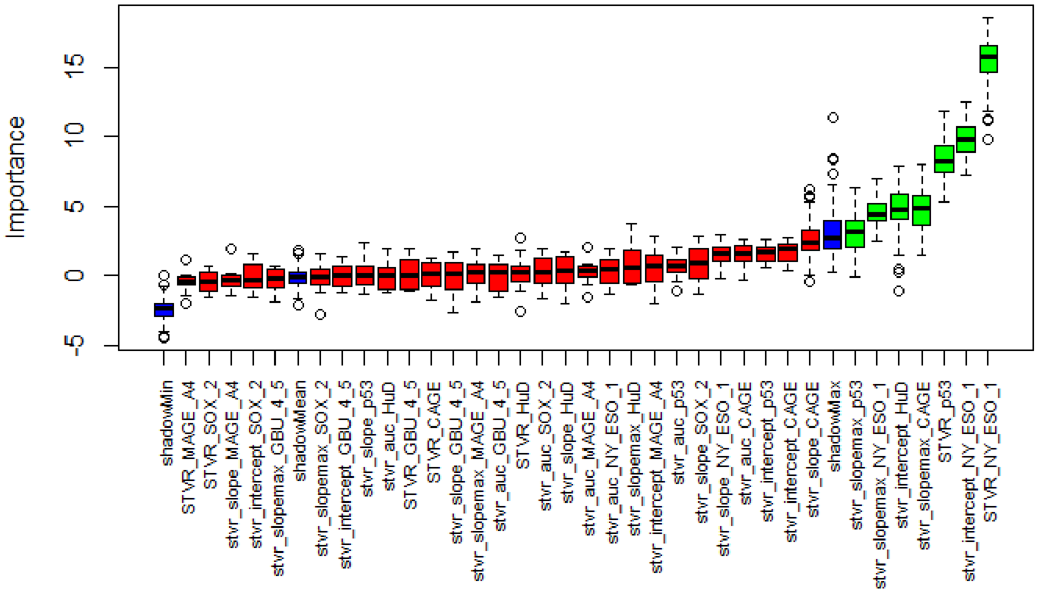


Figure 7-2: Boruta feature selection undertaken on STVR magnitude and curve characteristic features, Importance was calculated as z-score of the mean accuracy decrease.

Boruta selection on the STVR feature set identified 7 features which were considered important after comparison to shadow features, the full results are displayed in Figure 7-2, which shows that similar to the analysis on the VCOD features, the magnitude features for NY-ESO-1 and p53, the intercept feature for NY-ESO-1, and the slopemax features for NY-ESO-1 and p53 were classified as important, in addition the slopemax feature for CAGE, and the intercept feature for HuD were also deemed to be important in this analysis. This set of 7 features will be examined as distinct feature set in the subsequent modelling analyses.

7.6 Penalised Classification Search Function

For the subsequent supervised modelling analysis, a two-class penalised classification search function has been used which is designed to penalise a model for the presence of false positive results, and therefore maximise sensitivity constrained to models which exhibit high specificity. The formula is detailed here:

$$score = \left(\frac{TN + TP}{TN + TP + FN + FP}\right) - \left(\frac{(FP * penalty) + FN}{TN + TP + FN + FP}\right)$$

A penalty value of 3 was determined to return models optimised for specificity between 90% - 95% and has been used in the model optimisation described hereafter.

7.7 Binary Logistic Regression Modelling

7.7.1 Introduction

Logistic regression(156) is a regression strategy that allows the inclusion of both categorical and continuous variables to predict the probability of a binary outcome, initially developed in the 1940s for bioassay, and subsequently refined and extended, regression models have now been used for predicting the risk of lung cancer in both indeterminate nodules(157, 158), and in high-risk populations(105).

Least absolute shrinkage and selection operator (LASSO) regression(159) was developed to improve accuracy and interpretability of regression models and is a sparse penalized regression approach which constrains the absolute value of the regression coefficients and is able to reduce co-efficients to zero in order to exclude unnecessary variables from the regression model. Combined with cross-validation, this allows for the identification of an optimal set of coefficients which maximise the performance of the resulting model and has been shown to be superior to stepwise regression for identifying an optimal feature subset(160). LASSO regression and has been used previously in the development of diagnostic models, including those for breast cancer(161).

7.7.2 Methods

Least absolute shrinkage and selection operator (LASSO) regression was undertaken in R (v4.2.1) using the `glmnet` training method within the caret library to fit a binary logistic regression model via penalized maximum likelihood.

Parameter Tuning of Binary Logistic Models

To optimise the model fitting, a LASSO approach was used, employing values for the regularization constant of 0.001 to 0.1, at intervals of 0.001, and using 10-fold cross-validation to tune parameters to reduce any effect of overfitting. During cross validation, model performance was summarised and compared using a penalised classification metric designed to prioritise high specificity, as described previously.

7.7.3 Results

Performance of the LASSO regression models trained on the training cohort for each of the candidate feature sets are summarised in Figure 7-3 and potential diagnostic model performance is summarised in Table 7-3, showing model performance with probability thresholds optimised based on maximising sensitivity for specificity greater than 90% in the training cohort. These show that the trained regression models were unable to exceed the current commercial performance for the majority of the explored feature sets and cohorts, with the most promising being the model trained on the full feature set of STVR magnitude and curve characteristic features, however this only showed marginal improvements in the training and test cohorts which did not transfer to the validation cohorts, and led to specificities in the validation cohorts that would be too low for a screening modality.

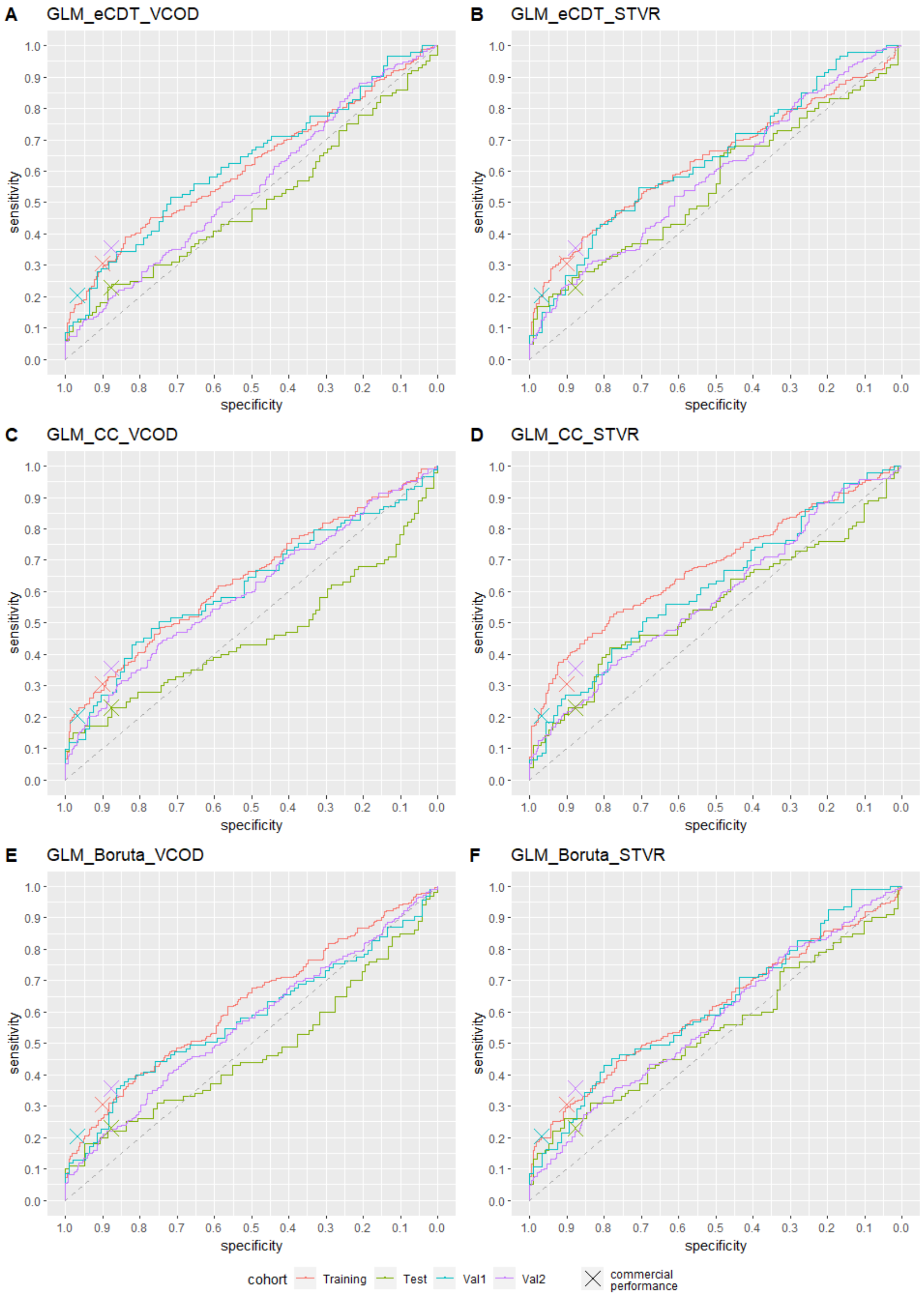


Figure 7-3: Comparison of diagnostic performance of LASSO regression models. A and B) EarlyCDT antigen panel – magnitude features only. C and D) Full feature set of magnitude and curve characteristics derived from EarlyCDT panel. E and F) Boruta selected features. A, C, and E) Subtraction of VOL for correction of non-specific binding. B, D, and F) Ratio of antigen to VOL signal for correction of non-specific binding.

Table 7-3: Summary of diagnostic performance of LASSO models on explored feature sets. Bold italicized values represent an improvement over performance obtained by the current commercial test, as defined by an increase in the Youden index.

NSB Correction	Feature set	Cohort	TP	FN	FP	TN	Sens	Spec
RU	Commercial	Training	72	163	23	209	30.6%	90.1%
RU	Commercial	Test	23	77	12	87	23.0%	87.8%
RU	Commercial	Validation 1	19	74	3	93	20.4%	96.9%
RU	Commercial	Validation 2	74	134	38	271	35.6%	87.7%
VCOD	eCDT Panel	Training	70	165	23	209	29.8%	90.1%
VCOD	eCDT Panel	Test	22	78	11	87	22.0%	88.8%
VCOD	eCDT Panel	Validation 1	***27***	***66***	***10***	***86***	***29.0%***	***89.6%***
VCOD	eCDT Panel	Validation 2	57	151	63	246	27.4%	79.6%
VCOD	All features	Training	67	168	23	209	28.5%	90.1%
VCOD	All features	Test	17	83	9	89	17.0%	90.8%
VCOD	All features	Validation 1	25	68	11	85	26.9%	88.5%
VCOD	All features	Validation 2	57	151	40	269	27.4%	87.1%
VCOD	Boruta	Training	62	173	23	209	26.4%	90.1%
VCOD	Boruta	Test	18	82	8	90	18.0%	91.8%
VCOD	Boruta	Validation 1	25	68	11	85	26.9%	88.5%
VCOD	Boruta	Validation 2	49	159	44	265	23.6%	85.8%
STVR	eCDT Panel	Training	***76***	***159***	***22***	***210***	***32.3%***	***90.5%***
STVR	eCDT Panel	Test	***24***	***76***	***10***	***88***	***24.0%***	***89.8%***
STVR	eCDT Panel	Validation 1	27	66	12	84	29.0%	87.5%
STVR	eCDT Panel	Validation 2	64	144	51	258	30.8%	83.5%
STVR	All features	Training	***92***	***143***	***23***	***209***	***39.1%***	***90.1%***
STVR	All features	Test	25	75	15	83	25.0%	84.7%
STVR	All features	Validation 1	28	65	16	80	30.1%	83.3%
STVR	All features	Validation 2	61	147	57	252	29.3%	81.6%
STVR	Boruta	Training	70	165	23	209	29.8%	90.1%
STVR	Boruta	Test	***28***	***72***	***16***	***82***	***28.0%***	***83.7%***
STVR	Boruta	Validation 1	23	70	10	86	24.7%	89.6%
STVR	Boruta	Validation 2	58	150	50	259	27.9%	83.8%

7.8 Support Vector Machine Models

7.8.1 Introduction

Support vector machines (SVMs) were developed as a binary classification tool(162), and are based on determining an optimal hyperplane in n-dimensional feature space which gives the greatest margin of separation between two classes. This hyperplane is then defined by identifying points lying on the boundaries of this margin, which are subsequently referred to as support vectors, and can be used to classify subsequent data points. One of the advantages of support vector machines is the ability to use kernels to transform the data into higher dimension feature spaces in order to create linear separators for data that would otherwise not be linearly separable. SVMs have been explored previously for their potential use in diagnosis of several cancers, including gastric lymph node cancer based on histological imaging data(163), prostate cancer based on MRI imaging(164), breast cancer recurrence based on clinicopathological features(165), and lung cancer based on analysis of CT scan data and demographic features(166), where they have shown the potential to form highly accurate models.

7.8.2 Methods

Support vector machine learning was undertaken in R (v4.2.1) using the `svmLinear` and `svmRadial` training methods within the caret library to explore both linear kernel and radial kernel support vector classification rules.

Parameter Tuning of Support Vector Machine Models

Linear support vector machines were optimised through exploring values of the misclassification cost ‘C’ between 0.5 and 4, smaller values of which allows for higher rates of misclassification in order to give separating hyperplanes with larger margins, while higher values result in lower training misclassification and hyperplanes with smaller margins, which may result in overfit models and poorer test set performance. 10-fold cross-validation was used to determine the optimal value of C for the final model.

7.8.3 Results - Linear Kernel

Performance of the linear kernel support vector models trained on the training cohort for each of the candidate feature sets are summarised in Figure 7-4 and potential diagnostic model performance is summarised in Table 7-4, showing model performance with probability thresholds optimised based on maximising sensitivity for specificity greater than 90% in the training cohort. These models showed improvements over the current commercial test in only the training cohort for all feature sets other than the full set of STVR magnitude and curve characteristic features, which showed improvement in the test cohort also, but not in the validation cohorts. This suggests that the improvements in the training set are the result of overfitting only, especially considering the reductions in specificity in the validation cohorts observed across all fitted models.

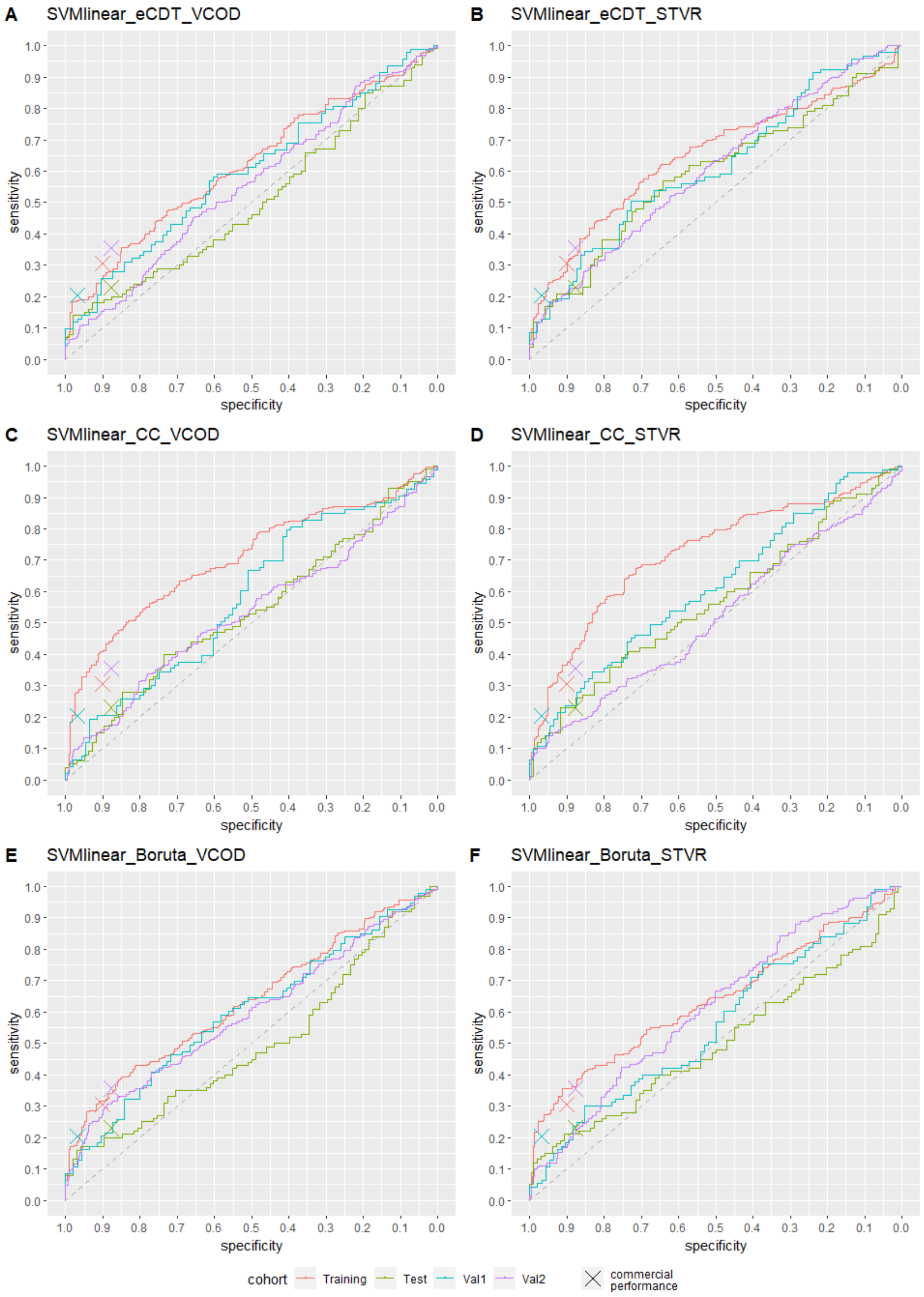


Figure 7-4: Comparison of diagnostic performance of linear kernel support vector machine models. A and B) EarlyCDT antigen panel – magnitude features only. C and D) Full feature set of magnitude and curve characteristics derived from EarlyCDT panel. E and F) Boruta selected features. A, C, and E) Subtraction of VOL for correction of non-specific binding. B, D, and F) Ratio of antigen to VOL signal for correction of non-specific binding.

Table 7-4: Summary of diagnostic performance of linear kernel support vector machine models on explored feature sets. Bold italicized values represent an improvement over performance obtained by the current commercial test, as defined by an increase in the Youden index.

NSB Correction	Feature set	Cohort	TP	FN	FP	TN	Sens	Spec
RU	Commercial	Training	72	163	23	209	30.6%	90.1%
RU	Commercial	Test	23	77	12	87	23.0%	87.8%
RU	Commercial	Validation 1	19	74	3	93	20.4%	96.9%
RU	Commercial	Validation 2	74	134	38	271	35.6%	87.7%
VCOD	eCDT Panel	Training	63	172	23	209	26.8%	90.1%
VCOD	eCDT Panel	Test	20	80	12	86	20.0%	87.8%
VCOD	eCDT Panel	Validation 1	26	67	12	84	28.0%	87.5%
VCOD	eCDT Panel	Validation 2	48	160	57	252	23.1%	81.6%
VCOD	All features	Training	***97***	***138***	***23***	***209***	***41.3%***	***90.1%***
VCOD	All features	Test	21	79	14	84	21.0%	85.7%
VCOD	All features	Validation 1	32	61	25	71	34.4%	74.0%
VCOD	All features	Validation 2	75	133	81	228	36.1%	73.8%
VCOD	Boruta	Training	***74***	***161***	***22***	***210***	***31.5%***	***90.5%***
VCOD	Boruta	Test	19	81	10	88	19.0%	89.8%
VCOD	Boruta	Validation 1	24	69	15	81	25.8%	84.4%
VCOD	Boruta	Validation 2	66	142	42	267	31.7%	86.4%
STVR	eCDT Panel	Training	***73***	***162***	***23***	***209***	***31.1%***	***90.1%***
STVR	eCDT Panel	Test	21	79	11	87	21.0%	88.8%
STVR	eCDT Panel	Validation 1	26	67	12	84	28.0%	87.5%
STVR	eCDT Panel	Validation 2	59	149	49	260	28.4%	84.1%
STVR	All features	Training	***88***	***147***	***23***	***209***	***37.4%***	***90.1%***
STVR	All features	Test	***23***	***77***	***9***	***89***	***23.0%***	***90.8%***
STVR	All features	Validation 1	26	67	12	84	28.0%	87.5%
STVR	All features	Validation 2	47	161	56	253	22.6%	81.9%
STVR	Boruta	Training	**84**	**151**	**21**	**211**	**35.7%**	**90.9%**
STVR	Boruta	Test	22	78	13	85	22.0%	86.7%
STVR	Boruta	Validation 1	28	65	15	81	30.1%	84.4%
STVR	Boruta	Validation 2	71	137	63	246	34.1%	79.6%

7.8.4 Results - Radial Kernel

Performance of the radial kernel support vector models trained on the training cohort for each of the candidate feature sets are summarised in Figure 7-5 and potential diagnostic model performance is summarised in Table 7-5, showing model performance with probability thresholds optimised based on maximising sensitivity for specificity greater than 90% in the training cohort. Examination of the ROC plots suggests that support vector machine models using a radial kernel STVR correction for non-specific binding show a higher degree of overfitting to the training cohort than VCOD corrected feature sets, as the ROC curves for the test and validation sets in the STVR models trend closer to the line of equality than those shown by models trained on VCOD feature sets. Radial kernel support vector machine models show some potential for returning an improved diagnostic model with the Boruta selected VCOD feature set showing improvements in the training, test, and validation 1 cohorts, however this model also returned a specificity of only 80.9% in the validation 2 cohort.

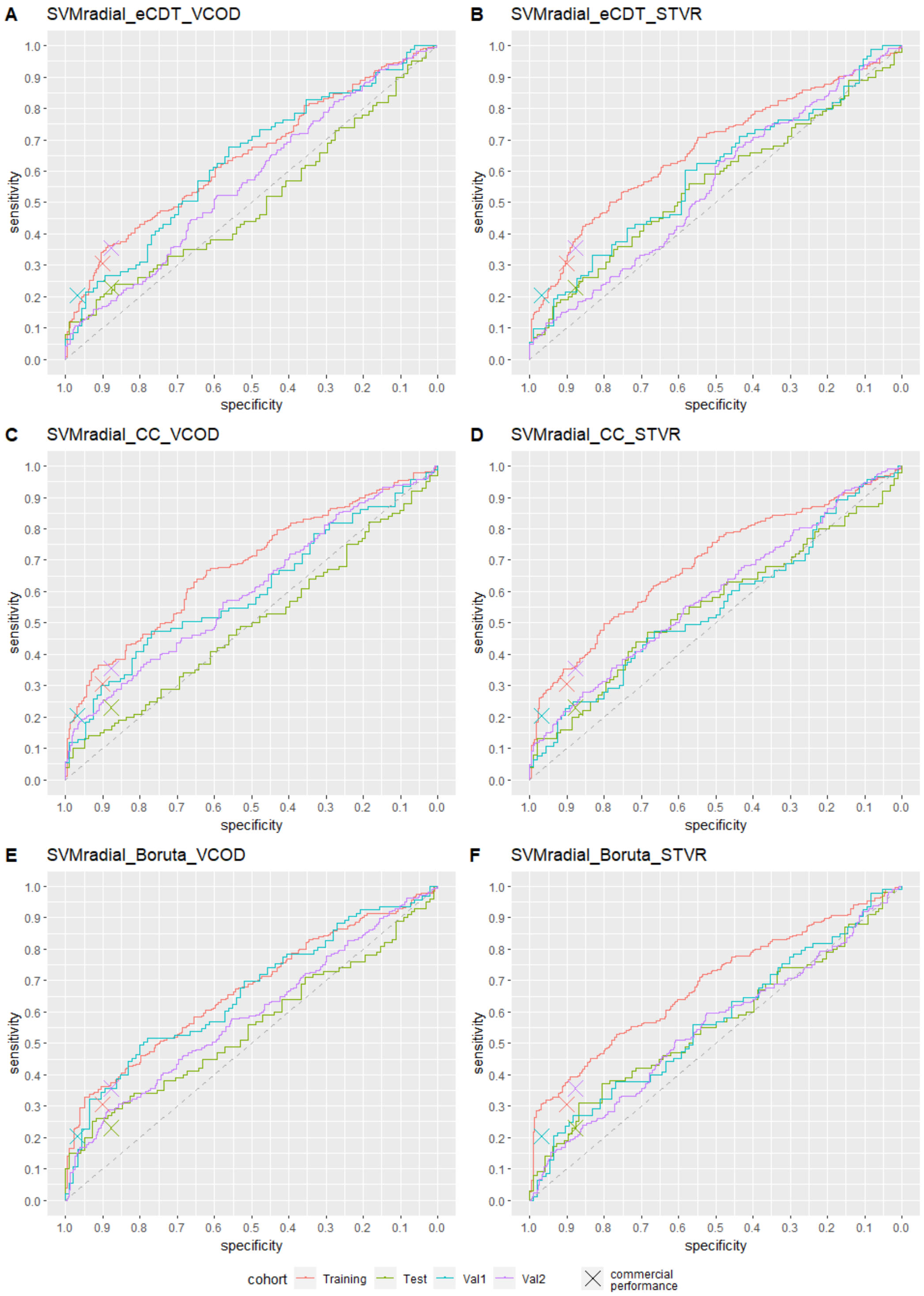


Figure 7-5: Comparison of diagnostic performance of radial kernel support vector machine models. A and B) EarlyCDT antigen panel – magnitude features only. C and D) Full feature set of magnitude and curve characteristics derived from EarlyCDT panel. E and F) Boruta selected features. A, C, and E) Subtraction of VOL for correction of non-specific binding. B, D, and F) Ratio of antigen to VOL signal for correction of non-specific binding.

Table 7-5: Summary of diagnostic performance of radial kernel support vector machine models on explored feature sets. Bold italicized values represent an improvement over performance obtained by the current commercial test, as defined by an increase in the Youden index.

NSB Correction	Feature set	Cohort	TP	FN	FP	TN	Sens	Spec
RU	Commercial	Training	72	163	23	209	30.6%	90.1%
RU	Commercial	Test	23	77	12	87	23.0%	87.8%
RU	Commercial	Validation 1	19	74	3	93	20.4%	96.9%
RU	Commercial	Validation 2	74	134	38	271	35.6%	87.7%
VCOD	eCDT Panel	Training	***81***	***154***	***23***	***209***	***34.5%***	***90.1%***
VCOD	eCDT Panel	Test	23	77	13	85	23.0%	86.7%
VCOD	eCDT Panel	Validation 1	27	66	18	78	29.0%	81.3%
VCOD	eCDT Panel	Validation 2	54	154	68	241	26.0%	78.0%
VCOD	All features	Training	***86***	***149***	***22***	***210***	***36.6%***	***90.5%***
VCOD	All features	Test	15	85	10	88	15.0%	89.8%
VCOD	All features	Validation 1	***30***	***63***	***14***	***82***	***32.3%***	***85.4%***
VCOD	All features	Validation 2	64	144	50	259	30.8%	83.8%
VCOD	Boruta	Training	***85***	***150***	***23***	***209***	***36.2%***	***90.1%***
VCOD	Boruta	Test	***26***	***74***	***10***	***88***	***26.0%***	***89.8%***
VCOD	Boruta	Validation 1	***33***	***60***	***10***	***86***	***35.5%***	***89.6%***
VCOD	Boruta	Validation 2	67	141	59	250	32.2%	80.9%
STVR	eCDT Panel	Training	***78***	***157***	***23***	***209***	***33.2%***	***90.1%***
STVR	eCDT Panel	Test	22	78	12	86	22.0%	87.8%
STVR	eCDT Panel	Validation 1	26	67	16	80	28.0%	83.3%
STVR	eCDT Panel	Validation 2	55	153	73	236	26.4%	76.4%
STVR	All features	Training	***83***	***152***	***22***	***210***	***35.3%***	***90.5%***
STVR	All features	Test	20	80	12	86	20.0%	87.8%
STVR	All features	Validation 1	24	69	19	77	25.8%	80.2%
STVR	All features	Validation 2	64	144	60	249	30.8%	80.6%
STVR	Boruta	Training	***89***	***146***	***23***	***209***	***37.9%***	***90.1%***
STVR	Boruta	Test	***24***	***76***	***12***	***86***	***24.0%***	***87.8%***
STVR	Boruta	Validation 1	30	63	20	76	32.3%	79.2%
STVR	Boruta	Validation 2	69	139	81	228	33.2%	73.8%

7.9 Naïve Bayes Models

7.9.1 Introduction

Naïve Bayes is a collective term for classification models which are based on Bayes Theorem, which work by considering all features as independent and calculating a conditional probability for an outcome based on the contributions of all features. The strength of Naïve Bayes for diagnostic modelling is its relative simplicity, as well as the ability to incorporate both continuous and discrete data into the models, and its insensitivity to irrelevant features.

Naïve Bayes models have previously been explored for diagnosis of brain tumours from segmented MRI images(167), diagnosis of breast cancer based on histological features(168), and lung cancer survival based on histological and demographic data(169).

7.9.2 Methods

Naïve Bayes modelling was undertaken in R (v4.2.1) using the `NaiveBayes` function within the `klaR` package (v4.2.3), and training using the functions included in the caret library to explore Naïve Bayes classification rules.

Parameter Tuning of Naïve Bayes Models

Naïve Bayes models were optimised through exploration of the use of kernel function, if used, adjustment of the bandwidth of the kernel function density, with values of 0 to 5 at intervals of 1, and through use of Laplace smoothing at values of 0, 0.001, 0.1, 1, 10, and 100. These hyperparameters were tested using 10-fold cross-validation to identify the optimal set of hyperparameters for the Naïve Bayes models.

7.9.3 Results

Performance of the Naïve Bayes models trained on the training cohort for each of the candidate feature sets are summarised in Figure 7-6 and potential diagnostic model performance is summarised in Table 7-6, showing model performance with probability thresholds optimised based on maximising sensitivity for specificity greater than 90% in the training cohort. These models showed the ability to improve or match the current clinical test in the training cohort for each feature set, however specificity was generally not maintained in the validation cohorts suggesting that these models displayed a higher degree of overfitting than the commercial panel, and none of the modelling on the explored feature sets showed consistent improvement over current commercial performance. Examination of the ROC plots suggests that STVR correction results in a higher degree of overfitting in the explored feature sets.

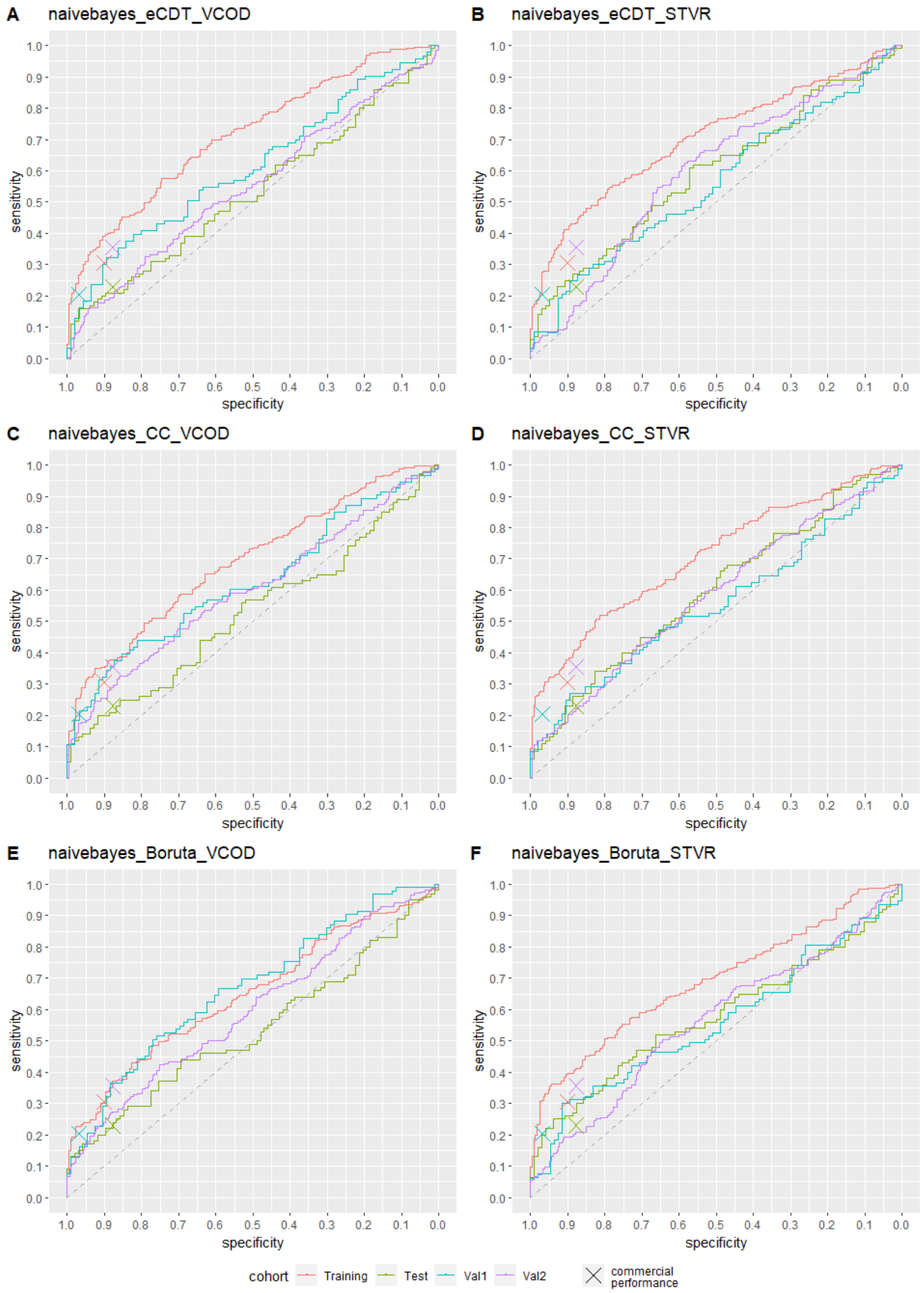


Figure 7-6: Comparison of diagnostic performance of Naïve Bayes models. A and B) EarlyCDT antigen panel – magnitude features only. C and D) Full feature set of magnitude and curve characteristics derived from EarlyCDT panel. E and F) Boruta selected features. A, C, and E) Subtraction of VOL for correction of non-specific binding. B, D, and F) Ratio of antigen to VOL signal for correction of non-specific binding.

Table 7-6: Summary of diagnostic performance of Naïve Bayes models on explored feature sets. Bold italicized values represent an improvement over performance obtained by the current commercial test, as defined by an increase in the Youden index.

NSB Correction	Feature set	Cohort	TP	FN	FP	TN	Sens	Spec
RU	Commercial	Training	72	163	23	209	30.6%	90.1%
RU	Commercial	Test	23	77	12	87	23.0%	87.8%
RU	Commercial	Validation 1	19	74	3	93	20.4%	96.9%
RU	Commercial	Validation 2	74	134	38	271	35.6%	87.7%
VCOD	eCDT Panel	Training	***92***	***143***	***22***	***210***	***39.1%***	***90.5%***
VCOD	eCDT Panel	Test	25	75	16	82	25.0%	83.7%
VCOD	eCDT Panel	Validation 1	41	52	30	66	44.1%	68.8%
VCOD	eCDT Panel	Validation 2	88	120	105	204	42.3%	66.0%
VCOD	All features	Training	***84***	***151***	***23***	***209***	***35.7%***	***90.1%***
VCOD	All features	Test	20	80	9	89	20.0%	90.8%
VCOD	All features	Validation 1	***38***	***55***	***16***	***80***	***40.9%***	***83.3%***
VCOD	All features	Validation 2	81	127	70	239	38.9%	77.3%
VCOD	Boruta	Training	72	163	23	209	30.6%	90.1%
VCOD	Boruta	Test	18	82	8	90	18.0%	91.8%
VCOD	Boruta	Validation 1	***27***	***66***	***9***	***87***	***29.0%***	***90.6%***
VCOD	Boruta	Validation 2	60	148	46	263	28.8%	85.1%
STVR	eCDT Panel	Training	***97***	***138***	***22***	***210***	***41.3%***	***90.5%***
STVR	eCDT Panel	Test	***29***	***71***	***15***	***83***	***29.0%***	***84.7%***
STVR	eCDT Panel	Validation 1	28	65	18	78	30.1%	81.3%
STVR	eCDT Panel	Validation 2	66	142	70	239	31.7%	77.3%
STVR	All features	Training	***89***	***146***	***23***	***209***	***37.9%***	***90.1%***
STVR	All features	Test	27	73	16	82	27.0%	83.7%
STVR	All features	Validation 1	34	59	23	73	36.6%	76.0%
STVR	All features	Validation 2	88	120	90	219	42.3%	70.9%
STVR	Boruta	Training	***93***	***142***	***23***	***209***	***39.6%***	***90.1%***
STVR	Boruta	Test	***30***	***70***	***14***	***84***	***30.0%***	***85.7%***
STVR	Boruta	Validation 1	***33***	***60***	***17***	***79***	***35.5%***	***82.3%***
STVR	Boruta	Validation 2	75	133	87	222	36.1%	71.8%

7.10 C5.0 Decision Tree Modelling

7.10.1 Introduction

Decision tree or recursive partitioning methods stratify a dataset by organising it into groups through a series of variable interaction rules. Decision tree modelling was first introduced in 1984 with the CART (Classification and Regression Trees) algorithm(170), and since its inception, several different tree modelling algorithms have been developed, including CART, BART (Bayesian Additive Regression Trees)(171), ID3 (Iterative Dichotimiser 3)(172) and its successors C4.5(173) and C5.0. For the following analysis the C5.0 algorithm was selected, over other methods such as the ID3 and CART algorithms, due to its ability to incorporate both continuous and discrete features, therefore allowing the potential development of improved models with the addition of demographic risk features such as gender and smoking history, and the incorporation of post-pruning in the algorithm to remove branches and nodes that contribute little to the classification accuracy.

7.10.2 Methods

C5.0 model training was undertaken in R (v4.2.1), using the `C5.0` training method within the caret library.

Parameter Tuning of C5.0 Decision Tree Models

10-fold cross-validation was used to determine whether applying winnowing – removal of uninformative features before training the tree model - was applied. During cross validation, model performance was summarised and compared using a penalised classification metric designed to prioritise high specificity, as described previously.

7.10.3 Results

Performance of the c5.0 decision tree models trained on the training cohort for each of the candidate feature sets are summarised in Figure 7-7 and potential diagnostic model performance is summarised in

Table 7-7, showing model performance with probability thresholds optimised based on maximising sensitivity for specificity greater than 90% in the training cohort. The ROC plots suggest that the STVR correction method led to a higher degree of overfitting using c5.0 decision tree modelling, especially when all magnitude and curve characteristic features were included in the modelling (Figure 7D). VCOD correction showed a lower degree of overfitting, although this could be due to the relatively small tree formed, as the same decision tree resulted from modelling of both magnitude features and the full magnitude and curve characteristic feature set and was comprised of only the magnitude features for NY-ESO-1 and p53.

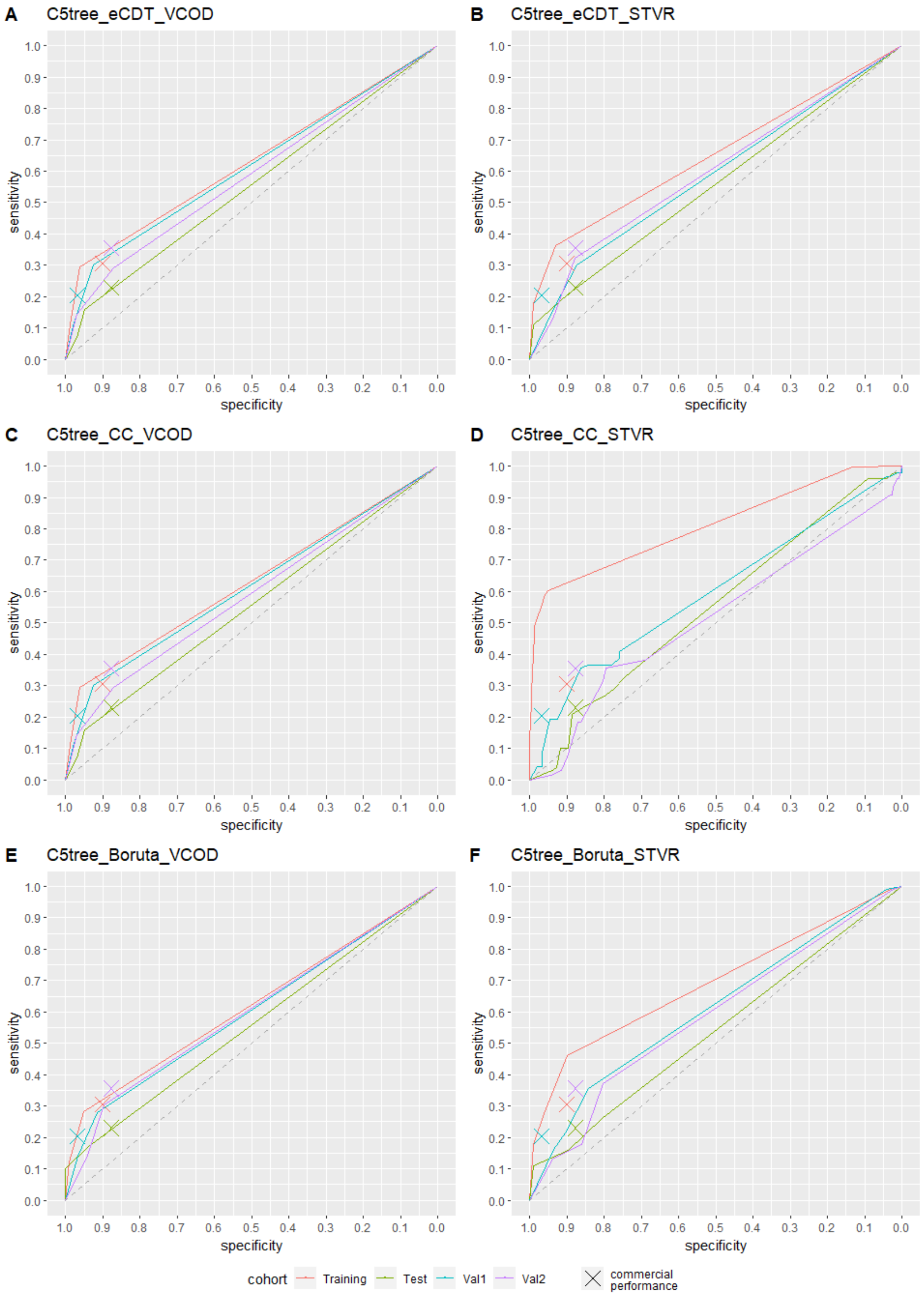


Figure 7-7: Comparison of diagnostic performance of C5.0 decision tree models. A and B) EarlyCDT antigen panel – magnitude features only. C and D) Full feature set of magnitude and curve characteristics derived from EarlyCDT panel. E and F) Boruta selected features. A, C, and E) Subtraction of VOL for correction of non-specific binding. B, D, and F) Ratio of antigen to VOL signal for correction of non-specific binding.

Table 7-7: Summary of diagnostic performance of C5.0 decision tree models on explored feature sets. Bold italicized values represent an improvement over performance obtained by the current commercial test, as defined by an increase in the Youden index.

NSB Correction	Feature set	Cohort	TP	FN	FP	TN	Sens	Spec
RU	Commercial	Training	72	163	23	209	30.6%	90.1%
RU	Commercial	Test	23	77	12	87	23.0%	87.8%
RU	Commercial	Validation 1	19	74	3	93	20.4%	96.9%
RU	Commercial	Validation 2	74	134	38	271	35.6%	87.7%
VCOD	eCDT Panel	Training	***69***	***166***	***9***	***223***	***29.4%***	***96.1%***
VCOD	eCDT Panel	Test	***16***	***84***	***5***	***93***	***16.0%***	***94.9%***
VCOD	eCDT Panel	Validation 1	***28***	***65***	***7***	***89***	***30.1%***	***92.7%***
VCOD	eCDT Panel	Validation 2	61	147	40	269	29.3%	87.1%
VCOD	All features	Training	***69***	***166***	***9***	***223***	***29.4%***	***96.1%***
VCOD	All features	Test	***16***	***84***	***5***	***93***	***16.0%***	***94.9%***
VCOD	All features	Validation 1	***28***	***65***	***7***	***89***	***30.1%***	***92.7%***
VCOD	All features	Validation 2	61	147	40	269	29.3%	87.1%
VCOD	Boruta	Training	**66**	**169**	**11**	**221**	**28.1%**	**95.3%**
VCOD	Boruta	Test	17	83	6	92	17.0%	93.9%
VCOD	Boruta	Validation 1	**26**	**67**	**8**	**88**	**28.0%**	**91.7%**
VCOD	Boruta	Validation 2	63	145	32	277	30.3%	89.6%
STVR	eCDT Panel	Training	**85**	**150**	**16**	**216**	**36.2%**	**93.1%**
STVR	eCDT Panel	Test	21	79	10	88	21.0%	89.8%
STVR	eCDT Panel	Validation 1	**28**	**65**	**12**	**84**	**30.1%**	**87.5%**
STVR	eCDT Panel	Validation 2	67	141	37	272	32.2%	88.0%
STVR	All features	Training	**142**	**93**	**11**	**221**	**60.4%**	**95.3%**
STVR	All features	Test	33	67	25	73	33.0%	74.5%
STVR	All features	Validation 1	38	55	23	73	40.9%	76.0%
STVR	All features	Validation 2	82	126	100	209	39.4%	67.6%
STVR	Boruta	Training	**108**	**127**	**23**	**209**	**46.0%**	**90.1%**
STVR	Boruta	Test	26	74	19	79	26.0%	80.6%
STVR	Boruta	Validation 1	**33**	**60**	**15**	**81**	**35.5%**	**84.4%**
STVR	Boruta	Validation 2	77	131	61	248	37.0%	80.3%

7.11 Random Forest Modelling

7.11.1 Introduction

Random Forest modelling is an extension of classification and regression tree (CART) modelling whereby a large number of tree-structured classifiers are trained on the data, either from randomised initial splits, or with each tree being trained on a random bagged subset of the initial training data, and a classification probability is then derived from the overall consensus of this large number of trees(174). This technique has been previously used to train models with extremely high clinical performance in prediction of prostate cancer progression(175), as well as in the prediction of Alzheimer's disease conversion(176).

For the following investigation, the ranger(177) implementation of the random forest algorithm was selected due to it being optimised for high dimensional data, giving much lower runtime and memory usage than other random forest implementations, but without appreciable reduction in the predictive performance of the resultant models.

7.11.2 Methods

Random Forest modelling was undertaken in R (v4.2.1) using the ranger library, applying the `ranger` training method to the data within the caret library Train function.

Parameter tuning of Random Forest Models

Random Forest model tuning was done over the following parameters: number of variables to possibly split at each node (mtry), splitting rule, and minimum node size.

Values for mtry were either 1 to the number of features for magnitude feature sets and Boruta selected feature sets, or odd values from 1 to 35 for the combined curve characteristic and magnitude feature set.

Options for the splitting rule were either selection by minimisation of Gini impurity, the extra-trees algorithm(178), or Hellinger distance(179).

Values explored for minimal node size were even numbers from 2 to 10.

Model performance for each combination of parameters was summarised over 10-fold cross validation and compared using a penalised classification metric designed to prioritise high specificity, as described previously.

7.11.3 Results

Performance of the random forest models trained on the training cohort for each of the candidate feature sets are summarised in Figure 7-8 and potential diagnostic model performance is summarised in

Table 7-8, the high overfitting observed in the training cohort meant that probability thresholds based on the training cohort showed a lack of ability to transfer to the test and validation cohorts, therefore model performance was optimised based on maximising sensitivity for specificity greater than 90% in the test cohort. Review of the ROC plots shows that random forest modelling suffers from an extremely high degree of overfitting, with the training cohort approaching perfect discrimination in almost every case. These models did not maintain their performance on test and validation cohorts, and in almost all cases – with only the exception of the VCOD Boruta feature set – the test and validation performance are inferior to that obtained by the commercial test.

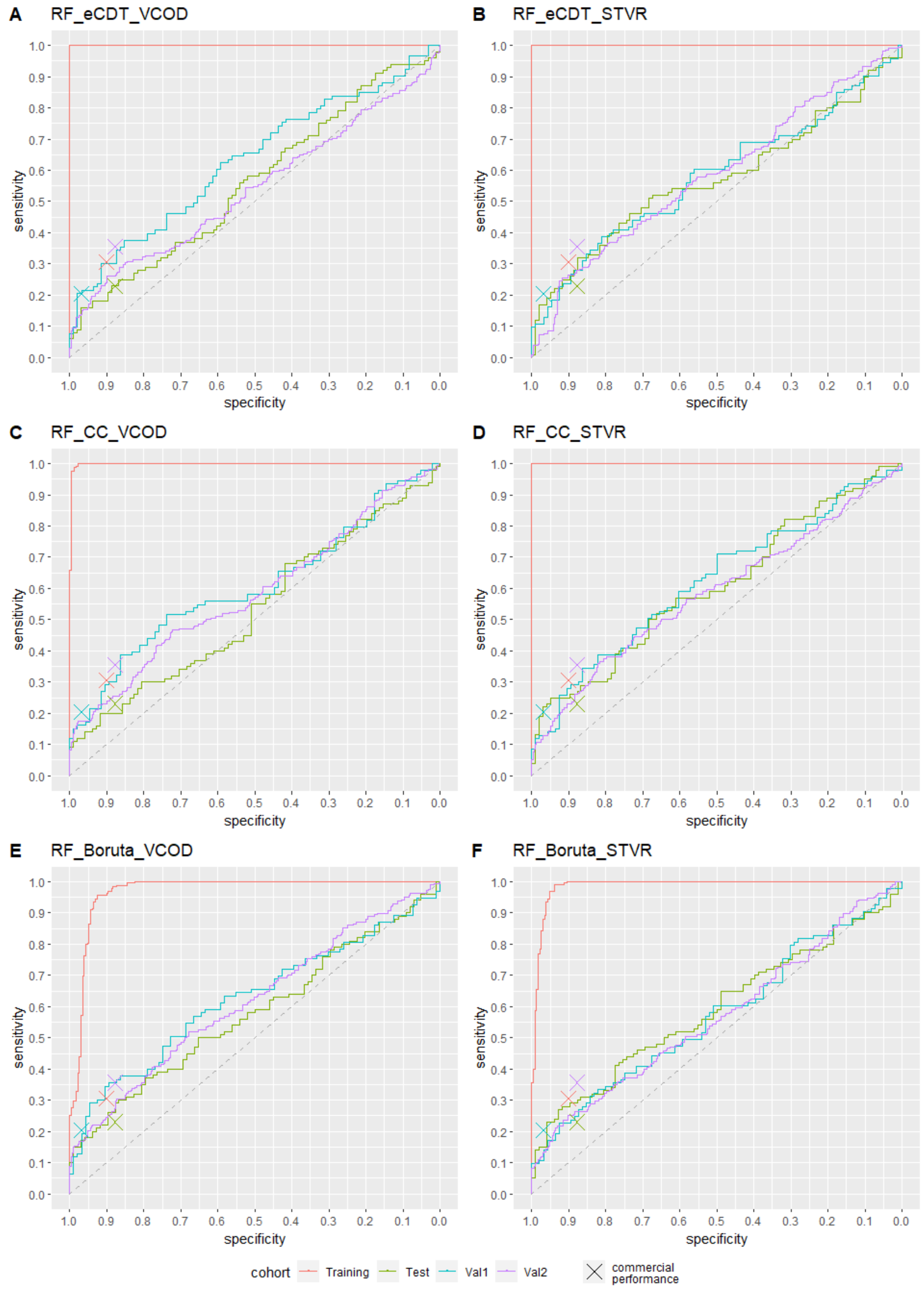


Figure 7-8: Comparison of diagnostic performance of random forest models. A and B) EarlyCDT antigen panel – magnitude features only. C and D) Full feature set of magnitude and curve characteristics derived from EarlyCDT panel. E and F) Boruta selected features. A, C, and E) Subtraction of VOL for correction of non-specific binding. B, D, and F) Ratio of antigen to VOL signal for correction of non-specific binding.

Table 7-8: Summary of diagnostic performance of random forest models on explored feature sets. Bold italicized values represent an improvement over performance obtained by the current commercial test, as defined by an increase in the Youden index.

NSB Correction	Feature set	Cohort	TP	FN	FP	TN	Sens	Spec
RU	Commercial	Training	72	163	23	209	30.6%	90.1%
RU	Commercial	Test	23	77	12	87	23.0%	87.8%
RU	Commercial	Validation 1	19	74	3	93	20.4%	96.9%
RU	Commercial	Validation 2	74	134	38	271	35.6%	87.7%
VCOD	eCDT Panel	Training	***222***	***13***	***0***	***232***	***94.5%***	***100.0%***
VCOD	eCDT Panel	Test	18	82	7	91	18.0%	92.9%
VCOD	eCDT Panel	Validation 1	32	61	12	84	34.4%	87.5%
VCOD	eCDT Panel	Validation 2	67	141	62	247	32.2%	79.9%
VCOD	All features	Training	***194***	***41***	***1***	***231***	***82.6%***	***99.6%***
VCOD	All features	Test	20	80	8	90	20.0%	91.8%
VCOD	All features	Validation 1	30	63	12	84	32.3%	87.5%
VCOD	All features	Validation 2	70	138	61	248	33.7%	80.3%
VCOD	Boruta	Training	***123***	***112***	***7***	***225***	***52.3%***	***97.0%***
VCOD	Boruta	Test	22	78	8	90	22.0%	91.8%
VCOD	Boruta	Validation 1	***33***	***60***	***10***	***86***	***35.5%***	***89.6%***
VCOD	Boruta	Validation 2	66	142	47	262	31.7%	84.8%
STVR	eCDT Panel	Training	***226***	***8***	***0***	***232***	***96.2%***	***100.0%***
STVR	eCDT Panel	Test	25	75	9	89	25.0%	90.8%
STVR	eCDT Panel	Validation 1	31	62	14	82	33.3%	85.4%
STVR	eCDT Panel	Validation 2	55	153	32	277	26.4%	89.6%
STVR	All features	Training	***235***	***0***	***0***	***232***	***100.0%***	***100.0%***
STVR	All features	Test	25	75	7	91	25.0%	92.9%
STVR	All features	Validation 1	29	64	13	83	31.2%	86.5%
STVR	All features	Validation 2	59	149	44	265	28.4%	85.8%
STVR	Boruta	Training	***159***	***76***	***4***	***228***	***67.7%***	***98.3%***
STVR	Boruta	Test	28	72	8	90	28.0%	91.8%
STVR	Boruta	Validation 1	32	61	21	75	34.4%	78.1%
STVR	Boruta	Validation 2	72	136	68	241	34.6%	78.0%

7.12 Extreme Gradient Boosted Trees Modelling

7.12.1 Introduction

Extreme gradient boosted trees is a technique which combines regression trees, gradient descent, and boosting(180), and since it's development has become an extremely popular modelling method, being used in high performing models for a wide variety of different problems, most relevantly to this investigation it has been successfully used to create models for disease prediction(181) and diagnosis(182), as well as a recently developed questionnaire based screening model for lung cancer(183).

XGBoost is named extreme due to its combination of several complementary algorithms. As a tree algorithm, data is split at each decision node in a similar way to the C5.0 algorithm, although XGBoost is also a boosting algorithm, meaning it is an ensemble method which iteratively builds new models by training on the residuals of the existing models, and boosting attributes that led to misclassifications of previous iterations. XGBoost also includes a regularization step which reduces the influence of individual ensemble models during the model building to reduce overfitting.

7.12.2 Methods

Random Forest modelling was undertaken in R (v4.2.1) using the ranger library and applying the `ranger` training method to the data within the caret library Train function.

Parameter tuning of XGBoost Models

XGBoost modelling includes many parameters over which the model can be tuned. Tuning was again accomplished over 10-fold cross-validation, and the tuning parameters and their explored values are detailed here:

eta – shrinkage rate used in the regularization step to reduce overfitting. Values searched: 0.01, 0.05, 0.1, 0.2, and 0.3.

max_depth – Maximum depth (number of edges from the root node to the furthest leaf node) of a tree. Larger values make the model more complex and more likely to overfit. Values searched: 2, 4, 6, and 8.

gamma – Minimum loss reduction required to make a further partition on a leaf node of the tree. Values searched: 0, and 5.

subsample – Subsample ratio of the training instances. Values searched: 0.5, 0.75, and 1.

colsample_bytree –Subsample ratio of columns when constructing each tree. Values searched: 0.5, 0.75, and 1.

min_child_weight – Minimum sum of instance weight needed in a child; larger values result in more conservative models. Values searched: 2, 4, 6, 8, and 10.

During cross validation, model performance was summarised and compared using a penalised classification metric designed to prioritise high specificity, as described previously.

7.12.3 Results

Performance of the extreme gradient boosted trees models trained on the training cohort for each of the candidate feature sets are summarised in Figure 7-9 and potential diagnostic model performance is summarised in Table 7-9, showing model performance with probability thresholds optimised based on maximising sensitivity for specificity greater than 90% in the training cohort. These models showed improved performance over the current commercial panel for the training and validation 1 cohort in each case, however these models showed generally poor specificity in the Validation 2 and test cohorts.

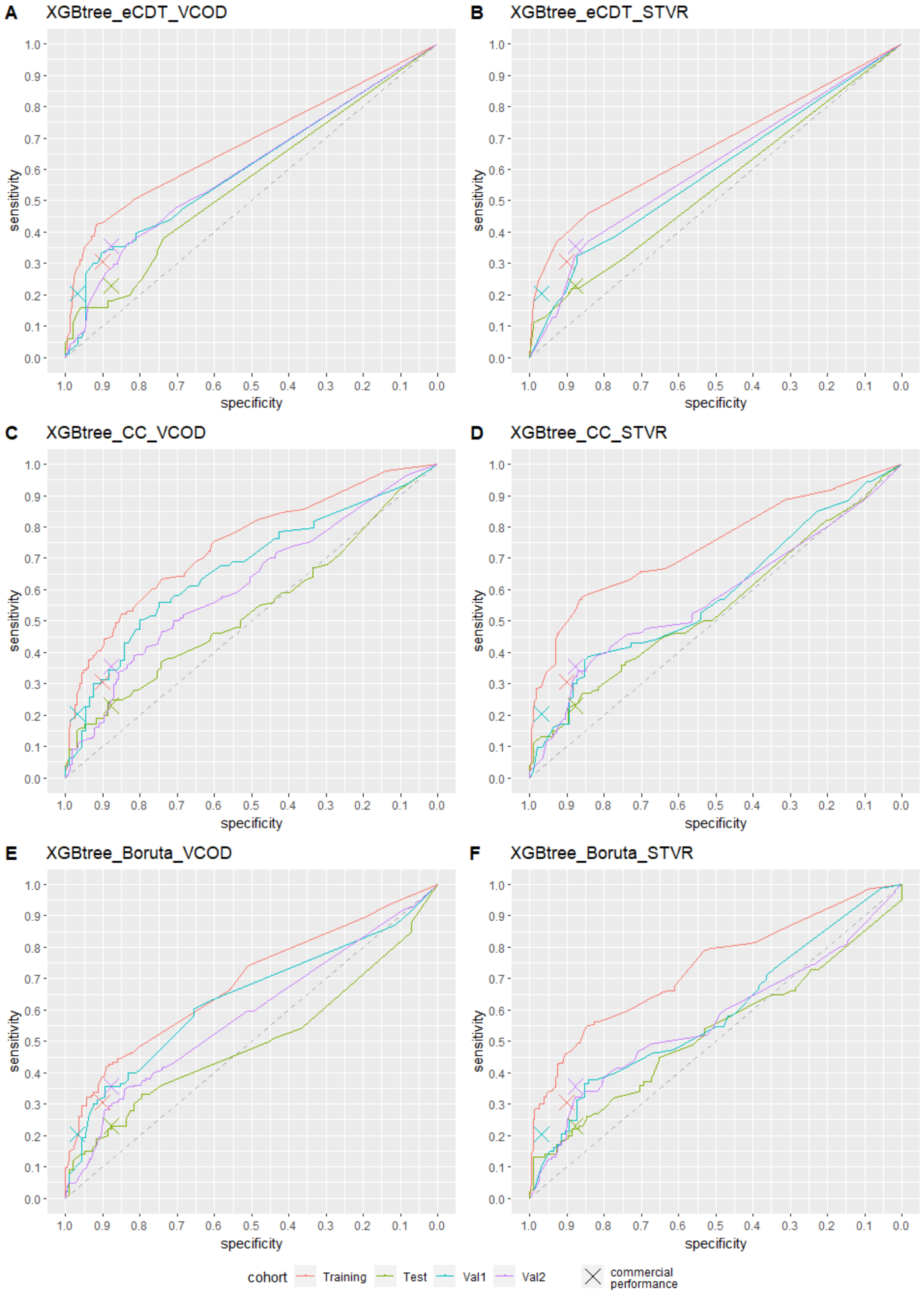


Figure 7-9: Comparison of diagnostic performance of extreme gradient boosted tree models. A and B) EarlyCDT antigen panel – magnitude features only. C and D) Full feature set of magnitude and curve characteristics derived from EarlyCDT panel. E and F) Boruta selected features. A, C, and E) Subtraction of VOL for correction of non-specific binding. B, D, and F) Ratio of antigen to VOL signal for correction of non-specific binding.

Table 7-9: Summary of diagnostic performance of extreme gradient boosted tree models on explored feature sets. Bold italicized values represent an improvement over performance obtained by the current commercial test, as defined by an increase in the Youden index.

NSB Correction	Feature set	Cohort	TP	FN	FP	TN	Sens	Spec
RU	Commercial	Training	72	163	23	209	30.6%	90.1%
RU	Commercial	Test	23	77	12	87	23.0%	87.8%
RU	Commercial	Validation 1	19	74	3	93	20.4%	96.9%
RU	Commercial	Validation 2	74	134	38	271	35.6%	87.7%
VCOD	eCDT Panel	Training	***101***	***134***	***23***	***209***	***43.0%***	***90.1%***
VCOD	eCDT Panel	Test	22	78	18	80	22.0%	81.6%
VCOD	eCDT Panel	Validation 1	***37***	***56***	***18***	***78***	***39.8%***	***81.3%***
VCOD	eCDT Panel	Validation 2	88	120	76	233	42.3%	75.4%
VCOD	All features	Training	***96***	***139***	***22***	***210***	***40.9%***	***90.5%***
VCOD	All features	Test	20	80	11	87	20.0%	88.8%
VCOD	All features	Validation 1	***32***	***61***	***13***	***83***	***34.4%***	***86.5%***
VCOD	All features	Validation 2	81	127	57	252	38.9%	81.6%
VCOD	Boruta	Training	***91***	***144***	***23***	***209***	***38.7%***	***90.1%***
VCOD	Boruta	Test	21	79	11	87	21.0%	88.8%
VCOD	Boruta	Validation 1	***33***	***60***	***12***	***84***	***35.5%***	***87.5%***
VCOD	Boruta	Validation 2	76	132	62	247	36.5%	79.9%
STVR	eCDT Panel	Training	***89***	***146***	***19***	***213***	***37.9%***	***91.8%***
STVR	eCDT Panel	Test	22	78	13	85	22.0%	86.7%
STVR	eCDT Panel	Validation 1	***30***	***63***	***12***	***84***	***32.3%***	***87.5%***
STVR	eCDT Panel	Validation 2	73	135	44	265	35.1%	85.8%
STVR	All features	Training	***108***	***127***	***17***	***215***	***46.0%***	***92.7%***
STVR	All features	Test	28	72	18	80	28.0%	81.6%
STVR	All features	Validation 1	***36***	***57***	***15***	***81***	***38.7%***	***84.4%***
STVR	All features	Validation 2	85	123	66	243	40.9%	78.6%
STVR	Boruta	Training	***108***	***127***	***23***	***209***	***46.0%***	***90.1%***
STVR	Boruta	Test	27	73	18	80	27.0%	81.6%
STVR	Boruta	Validation 1	***35***	***58***	***16***	***80***	***37.6%***	***83.3%***
STVR	Boruta	Validation 2	80	128	62	247	38.5%	79.9%

7.13 Summary

Table 7-10: Summary of diagnostic performance achieved by modelling strategies in explored feature sets.

Modelling Strategy	Metric	Training		Test		Validation 1		Validation 2	
		Sens	Spec	Sens	Spec	Sens	Spec	Sens	Spec
Commercial	Commercial	30.6%	90.1%	23.0%	87.8%	20.4%	96.9%	35.6%	87.7%
GLM LASSO	VCOD Magnitude	29.8%	90.1%	22.0%	88.8%	29.0%	89.6%	27.4%	79.6%
GLM LASSO	VCOD All	28.5%	90.1%	17.0%	90.8%	26.9%	88.5%	27.4%	87.0%
GLM LASSO	VCOD Boruta	26.4%	90.1%	18.0%	91.8%	26.9%	88.5%	23.6%	85.8%
GLM LASSO	STVR Magnitude	32.3%	90.5%	24.0%	89.8%	29.0%	87.5%	30.8%	83.5%
GLM LASSO	STVR All	39.1%	90.1%	25.0%	84.7%	30.1%	83.3%	29.3%	81.6%
GLM LASSO	STVR Boruta	29.8%	90.1%	28.0%	83.7%	24.7%	89.6%	27.9%	83.8%
SVM Linear	VCOD Magnitude	26.8%	90.1%	20.0%	87.8%	28.0%	87.5%	23.1%	81.6%
SVM Linear	VCOD All	41.3%	90.1%	21.0%	85.7%	34.4%	74.0%	36.1%	73.8%
SVM Linear	VCOD Boruta	31.5%	90.5%	19.0%	89.8%	25.8%	84.4%	31.7%	86.4%
SVM Linear	STVR Magnitude	31.1%	90.1%	21.0%	88.8%	28.0%	87.5%	28.4%	84.1%
SVM Linear	STVR All	37.4%	90.1%	23.0%	90.8%	28.0%	87.5%	22.6%	81.9%
SVM Linear	STVR Boruta	35.7%	90.9%	22.0%	86.7%	30.1%	84.4%	34.1%	79.6%
SVM Radial	VCOD Magnitude	34.4%	90.1%	23.0%	86.7%	29.0%	81.3%	26.0%	78.0%
SVM Radial	VCOD All	36.6%	90.5%	15.0%	89.8%	32.3%	85.4%	30.8%	83.8%
SVM Radial	VCOD Boruta	36.2%	90.1%	26.0%	89.8%	35.5%	89.6%	32.2%	80.9%
SVM Radial	STVR Magnitude	33.2%	90.1%	22.0%	87.8%	28.0%	83.3%	26.4%	76.4%
SVM Radial	STVR All	35.3%	90.5%	20.0%	87.8%	25.8%	80.2%	30.8%	80.6%
SVM Radial	STVR Boruta	37.9%	90.1%	24.0%	87.8%	32.3%	79.2%	33.2%	73.8%
Naïve Bayes	VCOD Magnitude	39.1%	90.5%	25.0%	83.7%	44.1%	68.8%	42.3%	66.0%
Naïve Bayes	VCOD All	35.7%	90.1%	20.0%	90.8%	40.9%	83.3%	38.9%	77.3%
Naïve Bayes	VCOD Boruta	30.6%	90.1%	18.0%	91.8%	29.0%	90.6%	28.8%	85.1%
Naïve Bayes	STVR Magnitude	41.3%	90.5%	29.0%	84.7%	30.1%	81.3%	31.7%	77.3%
Naïve Bayes	STVR All	37.9%	90.1%	27.0%	83.7%	36.6%	76.0%	42.3%	70.9%
Naïve Bayes	STVR Boruta	39.6%	90.1%	30.0%	85.7%	35.5%	82.3%	36.0%	71.8%
C5.0 Tree	VCOD Magnitude	29.4%	96.1%	16.0%	94.9%	30.1%	92.7%	29.3%	87.0%
C5.0 Tree	VCOD All	29.4%	96.1%	16.0%	94.9%	30.1%	92.7%	29.3%	87.0%
C5.0 Tree	VCOD Boruta	28.1%	95.3%	17.0%	93.9%	28.0%	91.7%	30.3%	89.6%
C5.0 Tree	STVR Magnitude	36.2%	93.1%	21.0%	89.8%	30.1%	87.5%	32.2%	88.0%
C5.0 Tree	STVR All	60.4%	95.3%	33.0%	74.5%	40.9%	76.0%	39.4%	67.6%
C5.0 Tree	STVR Boruta	46.0%	90.1%	26.0%	80.6%	35.5%	84.4%	37.0%	80.3%
Random Forest	VCOD Magnitude	94.5%	100.0%	18.0%	92.9%	34.4%	87.5%	32.2%	79.9%
Random Forest	VCOD All	82.6%	99.6%	20.0%	91.8%	32.3%	87.5%	33.7%	80.3%
Random Forest	VCOD Boruta	52.3%	97.0%	22.0%	91.8%	35.5%	89.6%	31.7%	84.8%
Random Forest	STVR Magnitude	96.2%	100.0%	25.0%	90.8%	33.3%	85.4%	26.4%	89.6%
Random Forest	STVR All	100.0%	100.0%	25.0%	92.9%	31.2%	86.5%	28.4%	85.8%
Random Forest	STVR Boruta	67.6%	98.3%	28.0%	91.8%	34.4%	78.1%	34.6%	78.0%
XGBoost	VCOD Magnitude	43.0%	90.1%	22.0%	81.6%	39.8%	81.3%	42.3%	75.4%
XGBoost	VCOD All	40.9%	90.5%	20.0%	88.8%	34.4%	86.5%	38.9%	81.5%
XGBoost	VCOD Boruta	38.7%	90.1%	21.0%	88.8%	35.5%	87.5%	36.5%	79.9%
XGBoost	STVR Magnitude	37.9%	91.8%	22.0%	86.7%	32.3%	87.5%	35.0%	85.8%
XGBoost	STVR All	46.0%	92.7%	28.0%	81.6%	38.7%	84.4%	40.9%	78.6%
XGBoost	STVR Boruta	46.0%	90.1%	27.0%	81.6%	37.6%	83.3%	38.5%	79.9%

Performance of all explored models has been summarised, as shown in Table 7-10, in order to compare the modelling strategies both to each other, and to the performance obtained using the commercial assessment of the sample results. This shows that the test cohort also showed a reduced performance using the current commercial test, and that the majority of modelling methods compare unfavourably to the current standard, with LASSO models giving much lower specificities than would be required for a screening test, Random Forest modelling techniques showed the greatest propensity to heavily overfit to the training data, and test cohort results show that specificity is not maintained in these models. c5.0 decision tree models show performance that is comparable to the current commercial performance, but do not display obvious or substantial gains in performance that would justify the redevelopment of the commercial test to analyse the data using these modelling strategies.

7.14 Discussion

A variety of supervised machine learning strategies have been explored against both tumour associated autoantibody magnitude features, and features derived from the binding curves of these autoantibodies to their antigens at a range of serum concentrations.

7.14.1 Logistic Regression Modelling

While logistic regression modelling has been proven to be effective for risk models such as those for determining risk of malignancy in indeterminate pulmonary nodules(158) which utilised demographic risk factors, as well as radiological characteristics, as well as models which utilise levels of tumour associated antigens as features, such as PSA in prostate cancer(184), LASSO models generated in this investigation, based upon the autoantibody features, and features derived from the binding curves generated during the EarlyCDT®-Lung assay, showed limited ability to determine cases from controls.

Logistic regression modelling also required additional intervention to ensure high specificity in the output, with predictions based on a 50% probability threshold for cancer giving specificities that were far too low to be commercially useful, as the false positive rate for the test would be far too high leading to large numbers of unnecessary follow up investigation, most likely in the form of CT screening.

The performance of logistic regression modelling using autoantibody data may be limited by the high specificity and low sensitivity of the individual autoantibody biomarkers, whereby the majority of the results are representative of non-specific binding and background noise. While this method was explored due to its popularity as a technique, the poor performance was unsurprising given that it require summation of a large number of expectantly negative features, with relatively variable signals, which, even after applying co-efficients, lead to non-specific assay noise having a large influence on the model result. For this reason decision tree methods which apply cut-off thresholds to categorise the data were expected to show higher accuracy for this data.

7.14.2 Support Vector Machine Models

Support vector machine models have been explored here as they are able to emulate the multivariate threshold assessment that was applied in a previous pilot study, and the immune amplification of an autoantibody response to relatively few cells should lead to a small population of clearly elevated autoantibody signals for each antigen, each of which should be distinct from the non-specific signal, and able to be discriminated by the application of a support vector machine threshold. The poor performance in these models is potentially due to an aggregation of non-specific signals over the multiple features, as the models trained on all features showed lower specificity than those trained on the current commercial autoantibody panel, or the Boruta selected feature panels. This suggests that this modelling technique is of limited utility in the early diagnosis of lung cancer, where we expect large feature sets of highly specific autoantibody features, each contributing small sensitivities, due to the extremely heterogeneous nature of cancers as a disease.

7.14.3 Naïve Bayes Models

Naïve Bayes, when applied to the Boruta selected features, returned a model that approached the current commercial performance, however it seems to be unsuitable as an approach when incorporating larger numbers of features, this is potentially due to an aggregation of false positives which leads to reductions in specificity that render it unsuitable for large autoantibody feature sets. Naïve Bayes is also better suited for categorical input variables and may potentially perform better on autoantibody data that has already had a diagnostic threshold applied to allow it to be considered as a binary categorical variable rather than a continuous autoantibody level as was explored here.

7.14.4 C5.0 Decision Tree Modelling

Of the strategies applied, decision tree modelling is most like the current commercial strategy applied by the EarlyCDT®-Lung test, whereby an autoantibody magnitude value above a threshold cut-off for any of the autoantibody features determines a positive result, in fact the strategy for the commercial EarlyCDT®-Lung result could be expressed in the form of a tree model. This may be the reason that the C5.0 Tree models for the VCOD features seem to show discriminatory ability most similar to the current commercial output, albeit with a higher specificity as enforced by the penalised classification search function that has been used in the modelling. Tree models may be more suitable for assessing autoantibody features, due to the nature of the immune response to cancer being extremely heterogeneous, therefore the preponderance of low values returned by assay are more likely to be due to non-specific binding. with decision tree models applying a binary decision threshold at each decision node, rather than a co-efficient multiplier to the magnitude, the influence of non-specific responses are greatly diminished.

In this investigation, decision tree modelling on the VCOD magnitude features gave the best results, with curve characteristics not being included in the model while magnitude features were available. This model also showed minimal reduction in specificity between training and validation, suggesting a lower degree of overfitting, however this may be due to only two magnitude features being incorporated in the decision tree.

The models generated in this analysis have not shown discriminatory ability that is a sufficient improvement over the current commercial strategy to warrant redevelopment of the test to apply a tree model algorithm, however, further investigation is warranted into whether inclusion of demographic risk variables is able to further improve the performance of these models to reach improved sensitivities. Additionally expansion of the model from a simple binary output to a model which can determine probability of histological subtype would potentially allow for greater contribution of features other than NY-ESO-1 and p53, such as SOX-2 and HuD which are known to be associated to small cell lung cancers, which had a lower representation in the explored data sets, at only 11% of the cancer cohort, compared to 40% of the cohort for adenocarcinoma, and 40% for squamous cell carcinoma. While this reflects the incidence distribution of lung cancer in the population, it may in this case have resulted in models which are overfitted to non-small cell lung carcinomas, therefore the features which have been excluded in the decision tree models trained here may be of greater contribution in a multivariate decision tree model.

7.14.5 Random Forest Modelling

Random forest modelling on these feature sets resulted in highly overfit models which were not applicable to the hold-out test set. Examination of the feature importance for these models also shows that the random forest modelling tended to include all but one feature each time, and the random forest model trained on the VCOD magnitude features was the only mode which gave high importance to the HuD and SOX-2 features. Considering HuD and SOX-2 are known to be related to small-cell lung cancer which comprises only 11% of the cancers in the training data this suggests that the random forest model is training to non-specific signal, which is resulting in extremely poor performance in the test data. This suggests that random forest modelling is unsuitable for training models on autoantibody data.

7.14.6 Extreme Gradient Boosting Modelling

XGBoost models have been found to be extremely versatile and show discriminatory ability comparable to both the current commercial, and to the C5.0 decision tree models, although again display no significant increase in performance over the current commercial strategy that would result in redevelopment of the commercial assay. The feature importance in the XGBoost models shows that the modelling did not try to incorporate large numbers of features which may be why it was less prone to overfitting that random forest modelling, and the inclusion of HuD at a lower importance in the VCOD magnitude model suggests that XGboost modelling may benefit from being trained against histological subtype as opposed to just cancer presence. As XGBoost has already been used in the development of a lung cancer risk screening model(183), the inclusion of additional demographic risk factors to the autoantibody screening test may also result in a diagnostic risk model with improved performance over that currently obtained by the EarlyCDT®-Lung test, or by risk models that include only demographic risk factors.

7.15 Chapter Conclusions

A number of well-established supervised machine learning techniques have been applied to the autoantibody magnitude and curve characteristic data generated from the results of the EarlyCDT®-Lung test. None of the techniques explored show appreciable improvement in diagnostic performance over that achieved by the current commercial output of the test, however C5.0 decision trees, extreme gradient boosted trees, and potentially also radial kernel support vector machines, show performance comparable to the current commercial test, while random forest models showed extremely high degrees of overfitting to the training data, and is likely unsuitable for these applications. Throughout this exploration, however, the derived curve characteristics do not show the ability to greatly improve the specificity or sensitivity of the biomarker panel, and improvements to the diagnostic performance are more likely to be achieved through other means.

Additional investigations (not shown) were undertaken to determine whether the addition of risk modelling based upon the available demographic variable information could contribute to improved predictive performance for the trained models. This analysis, however, was limited by the lack of detailed smoking history or additional risk factors that have previously contributed to established demographic risk models, with only age, sex, and categorical smoking history (current smoker, ex-smoker or never-smoker) being captured in the majority of the cohort. Of the available demographic risk models that have been established in the literature, four have been previous determined to show the highest accuracy for cancer risk prediction(100) (Bach(185), $\text{PLCO}_{\text{M2012}}$(101), LCRAT, and LCDRAT(103)), and amongst them required demographic information including smoking duration and intensity, asbestos exposure, as well as education levels, BMI, history of related disease such as emphysema, and family history of lung cancer. Even the simplest risk model, the Pittsburgh Predictor(186) required information on smoking duration and intensity along with age and smoking status. Future studies would benefit from ensuring that this level of demographic detail is captured to enable demographic risk features to be incorporated with biomarker results to develop ensemble models, especially as such demographic risk should feed into public health policy regarding the target population that would benefit from a screening test for cancer.

While these models currently do not show improvements in diagnostic performance over the current commercial thresholds, models with comparable performance to the EarlyCDT threshold tests, and which can return a percentage likelihood score rather than a binary decision, may lend themselves more easily to inclusion in ensemble models, with the incorporation of additional diagnostic biomarkers such as antigenic biomarkers - which have been explored previously, and are mostly associated with late stage disease, demographic risk features - which have already shown use in models such as Bach(185), and PLCO$_{M2012}$(101), genetic risk features, such as circulating tumour DNA which has shown potential for directing disease management, or DNA methylation markers(187). Along with an increased variety of diagnostic biomarkers, longitudinal testing methods, with maintained history of screening test results may also provide an avenue for improved diagnostic performance, with application of personalised baselines and surveillance of changes in biomarker levels, such as that attempted by the ROCA test(134) providing another avenue for additional data which could feed into ensemble models.

7.16 Chapter Discussion

It is massively disappointing that none of the explored modelling strategies was able to train a model that could consistently outperform the current commercial test. While some of the results were unsurprising, the lack of consistent improvement using either support vector machines or xgboost strategies was frustrating. The inconsistency of the trained models' performance may indicate a higher impact of confounding variables than was appreciated prior to undertaking this research, and future explorations would benefit from more carefully curated sample cohorts, balanced for demographic risk factors between cases and controls, and also between training, test and validation cohorts.

Also disappointing was discovering the lack of detailed smoking history and demographic information accompanying the samples that would have allowed for the application of risk models such as PLCOM2012 and LLPv2 for an accurate calculation of each subjects demographic risk. The availability of this demographic information would have allowed for exploration of ensemble models incorporating the pre-test risk score and may have been able to improve the accuracy of resultant models.

Chapter 8: Exploration of Supervised Machine Learning Strategies for Early Lung Cancer Diagnosis using an Extended Panel of Tumour Associated Autoantibodies.

8.1 Aims

With the supervised machine learning strategies explored being unable to show consistent improvement in diagnostic performance above that returned by the current commercial panel strategy, I repeated the analysis including data from additional autoantibody biomarkers to try and add more sources of sensitivity, again aiming to train models with diagnostic performance superior to that of the current panel assay.

8.2 Introduction

Investigations into machine learning strategies applied to the EarlyCDT panel of autoantibodies, incorporating derived curve characteristics based upon the autoantibody binding curves generated during the EarlyCDT®-Lung assay resulted in only marginal performance improvements over the current commercial EarlyCDT®-Lung panel strategy. This suggests that there is limited additional diagnostic utility to be returned by the current commercial panel above that currently being returned by the commercial test method. Due to the highly heterogeneous nature of lung cancer, additional autoantibodies representing alternate biological pathways which may be affected during tumorigenesis should be able to provide additional sensitivity to detect further lung cancers.

A high throughput biomarker discovery strategy was employed to assess a large number of potential antigens for their ability to bind tumour-associated autoantibody biomarkers(188), from which 5 additional antigens were identified as having the potential to add sensitivity to the commercial panel. These initial 5 additional antigens are: alpha-enolase, p53-C-term, cytokeratin 20, cytokeratin 8, and L-myc-2, and their discriminatory performance as determined in the biomarker discovery study is summarised in Table 8-1.

Alpha-enolase

Alpha-enolase (Gene name ENO1, Gene ID 2023) is a glycolytic enzyme expressed in most tissues, with multiple functions dependant on location, including catalysing glycolysis, roles in transcription, apoptosis regulation and cell differentiation, as well as acting as a strong receptor and activator of plasminogen when expressed at the cell-surface(189). Alpha-enolase is over-expressed in multiple human cancer types, contributing to increased glycolysis and tumor growth, and this overexpression has been shown to elicit an autoantibody response in Liver cancer(190), Pancreatic cancer(191), and Lung cancer(192), with autoantibodies to alpha-enolase being associated with more aggressive tumours and poor prognosis(193).

p53-C-term

p53-C-term is a 35kDa fragment representing the carboxy-terminus of the p53 protein (Gene name TP53, Gene ID 7157), which is involved in regulation of the cell cycle, apoptosis, and genomic stability, which has been described previously in section 1.10.2.

Cytokeratin 20

Cytokeratin 20 (CK20) (Gene name KRT20, Gene ID 54474) is a polypeptide normally expressed in the gastric and intestinal epithelium, urothelium, and

Merkel cells, with elevated expression having been observed in colorectal carcinoma and colorectal adenoma(194), as well as adenocarcinomas of the stomach, gall bladder and bile ducts(195).

Cytokeratin 8

Cytokeratin 8 (CK8) (Gene name KRT8, Gene ID 3856) is a structural protein involved in the formation of filaments within cell cytoplasm which generate a stabilizing framework and facilitate the movement of signalling molecules and metabolites within the cell. Autoantibodies to CK8 have previously been reported for the detection of breast cancer(196).

L-Myc-2

L-Myc-2 (Gene name MYCL, Gene ID 4610) is a transcription factor which has key roles in cell proliferation, growth, differentiation and apoptosis. Amplification and overexpression of L-Myc has previously been observed in ovarian cancer(197) and small-cell lung cancer(198)

Table 8-1: Discriminatory performance of proposed additional lung antigens in discovery study(188).

Protein	Discovery Study Sensitivity	Discovery Study Specificity
Alpha-enolase	15%	98%
p53-C-term	14%	98%
Cytokeratin 20	10%	98%
Cytokeratin 8	4%	100%
L-Myc-2	10%	98%

In addition, 7 autoantibody biomarkers were identified from a further high throughput analysis and have been included here to explore whether they are able to contribute to increased sensitivity for lung cancer through the application of machine learning strategies. These autoantibodies are p16-C, KOC, ALDH1, p62, SSX1, p53-95, and K-ras G13C/Q91H.

p16-C

p16-C (Gene name CDKN2A, Gene ID 1029) is a cyclin-dependent kinase (CDK) inhibitor which is involved in down-regulation of the cell cycle through inactivation of CDK4 and CDK6 activities during the G1 growth phase of the cell cycle. It acts as a tumour-suppressor gene, with loss of p16 activity having been linked to many cancers(199), with autoantibodies to p16 having been identified as having diagnostic potential in lung cancer(200, 201), breast cancer(201-203) as well as HCC, colorectal, esophageal, pharyngeal, uterine(201), hepatocellular and nasopharyngeal cancers(203).

KOC

KOC (also known as IMP3, Gene name IGF2BP3, Gene ID 10643) is a messenger RNA binding protein, normally only expressed in the placenta, but known for being overexpressed in cancers (in fact KOC is an abbreviation of "K homology domain containing protein Overexpressed in Cancer"). KOC was initially discovered in pancreatic cancer(204), and studies have subsequently shown overexpression of KOC in lung cancers(205), renal cell carcinoma(206), endometrioid adenocarcinoma(207), and melanoma(208), as well as expression that was associated with lung cancer histologic grade(209). Immune reactivity to KOC has been previously established with autoantibodies to KOC being identified in esophageal, lymphoma, pharyngeal cancers (202), HCC, gastric, breast (202, 210), and lung cancer(205, 210).

ALDH1

ALDH1 (Gene name LDH1A1, Gene ID 216) is an aldehyde dehydrogenase, responsible for oxidation of acetaldehyde to acetic acid, and plays roles in both gene expression and tissue differentiation. In healthy tissue, ALDH1 exists primarily in the cytoplasm of liver cells, however studies have associated aberrant ALDH1 expression with a number of solid tumours including lung(211), breast(212), and colorectal cancer(213), with elevated ALDH1 expression being associated with poor cancer prognosis and malignant tumour progression(214, 215).

p62

p62 (also known as IMP2, Gene name SQSTM1, Gene ID 8878), is a multidomain protein which acts as a signalling hub and serves critical roles in a number of cellular functions including cell survival and apoptosis, with evidence that p62 accumulation is an important promotor of tumorigenesis(216, 217). Autoantibodies to p62 have been identified in a variety of tumour types, including breast(202), esophageal, colorectal, lung, pharyngeal and uterine cancers(201).

SSX1

SSX1 (Gene ID 6756) is a cancer testis antigen, whose expression is normally restricted to testis germline cells. SSX1 expression has been detected in a wide range of cancer cells including bladder, breast, colorectal, hepatocellular, myeloma, and lung cancer(218)

p53-95 is a fragment comprised of the first 95 amino acids of the p53 protein (Gene name TP53, Gene ID 7157), which has been described previously in section 1.10.2.

K-ras G13C/Q61H

K-ras G13C/Q61H (Gene name KRAS, Gene ID 3845) is a version of the K-ras protein with two single nucleotide polymorphisms which result in the 13th amino acid being swapped from glycine to cysteine, and the 61st amino acid being swapped from glutamine to histidine. In healthy tissue, Ras proteins are small GTPase proteins, whose activation leads to the activation of proteins involved in cell growth, differentiation and survival. Cancer associated Ras genes are generally characterised by single base missense mutations, with 99% of these being found at residues G12, G13, or Q61(219). K-ras is the most frequently mutated Ras isoform and K-ras mutants have been associated with a wide range of cancers including pancreatic(220), colorectal(221), and lung cancer(222).

As discussed previously in sections 7.6 to 7.11, the machine learning techniques explored consist of logistic regression, support vector machine learning, naïve Bayes, decision trees, random forest, and extreme gradient boosted regression trees.

8.3 Datasets

Assay data results from the commercial EarlyCDT®-Lung test panel from two case-control studies was combined, and divided into training and test cohorts. The demographics for the full dataset is summarised in Table 5-1, with cancer subtype and stage distribution as shown in

Table 5-2. In addition, the trained models have been applied to two additional independent cohorts of samples to assess model performance reproducibility.

8.4 Features

As described previously, the EarlyCDT®-Lung test collected autoantibody response values over a range of dilutions (1.6nM, 5nM, 16nM, 50nM, and 160nM), the additional autoantibodies assessed in this analysis were also assessed over this dilution range, and the single dilution value which showed the greatest discriminatory ability for each autoantibody was determined. These concentrations are as follows: for vol corrected OD (VCOD) data for the commercial panel - as reported previously; p53 at 1.6nM, SOX2 at 50nM, CAGE at 16nM, NY-ESO-1 at 5nM, GBU4-5 at 1.6nM, MAGE-A4 at 5nM, and HuD at 160nM. For VCOD data for the initial 5 extended panel autoantibodies: alpha-enolase at 16nM, p53-C-term at 5nM, cytokeratin 20 at 160nM, cytokeratin 8 at 16nM, and L-myc-2 at 5nM. Finally for VCOD values for the additional 7 autoantibodies: p16-C, KOC, ALDH1, p62, and K-ras G13C/Q91H were assessed at 1.6nM while SSX1 and p53-95 were assessed at 16nM.

For signal to vol ratio (STVR) data; as reported previously, p53 at 1.6nM, SOX2 at 50nM, CAGE at 1.6nM, NY-ESO-1 at 5nM, GBU4-5 at 50nM, MAGE-A4 at 16nM, and HuD at 1.6nM. For STVR data for the initial 5 extended panel autoantibodies: alpha-enolase at 5nM, p53-C-term at 5nM, cytokeratin 20 at 16nM, cytokeratin 8 at 16nM, and L-myc-2 at 50nM. Finally for STVR values for the additional 7 autoantibodies: KOC, ALDH1, p62, and

K-ras G13C/Q91H were assessed at 1.6nM while p16-C, SSX1 and p53-95 were assessed at 5nM.

All features underwent transformation and scaling as described previously to ensure data approximated a normal distribution, and to prevent differences in feature scale from detrimentally influencing the modelling.

8.4.1 Individual Feature Discriminatory Performance

The discriminatory ability of the magnitude features in both VCOD and STVR feature sets has been summarised by identifying an optimal cutpoint which maximised the Youden index for the training cohort, constrained to specificities above 90%, using the R cutpointr package (v1.1.2) to iterate over all features, after which the optimal cutpoint was applied to the hold-out test cohort to return performance characteristics.

Vol Subtracted (VCOD) Features

Summary values for accuracy in the training cohort (training accuracy), area under the ROC curve in the training cohort (training AUC), and sensitivity and specificity at the optimal cutpoint in both training cohort (training sensitivity and training specificity respectively), and test cohort (test sensitivity and test specificity respectively) are summarised in Table 8-2, showing that that for the majority of features, specificity is maintained in the test cohort, with a mean decrease in the test cohort of only 2.2%, and the largest change being in HuD, in which the specificity decreased to 82.7% compared to 90.9% in the training cohort. Sensitivity does show a small drop off in the test cohort, with a mean decrease of 4.3% across all features, and the largest reduction being shown in Lmyc2 in which the sensitivity reduction is 9.4%. Summary ROC plots were constructed to allow comparison of the features discriminatory ability over the dynamic range of the assay, and these are included in appendix 1.

Table 8-2: Discriminatory ability of individual Vol corrected magnitude and curve characteristic features, optimised over training cohort and applied to test cohort.

Feature	optimal cutpoint	training accuracy	training AUC	training sens	training spec	test sens	test spec
VCOD_p53	1.016	0.559	0.537	21.7%	90.5%	13.0%	90.8%
VCOD_SOX_2	0.929	0.505	0.503	11.1%	90.5%	14.0%	85.7%
VCOD_CAGE	0.960	0.537	0.527	17.0%	90.9%	11.0%	90.8%
VCOD_NY_ESO_1	0.663	0.582	0.607	26.8%	90.1%	25.0%	85.7%
VCOD_GBU_4_5	1.039	0.525	0.514	15.3%	90.1%	8.0%	89.8%
VCOD_MAGE_A4	1.039	0.527	0.547	15.7%	90.1%	13.0%	90.8%
VCOD_HuD	0.717	0.533	0.533	16.2%	90.9%	9.0%	82.7%
VCOD_p16-C	1.161	0.544	0.583	18.7%	90.5%	15.0%	87.9%
VCOD_KOC	0.988	0.540	0.508	17.4%	90.9%	11.0%	91.9%
VCOD_ALDH1	1.211	0.493	0.481	8.9%	90.1%	10.0%	87.9%
VCOD_p62	1.035	0.544	0.544	18.7%	90.5%	10.0%	85.9%
VCOD_SSX1	1.127	0.512	0.506	12.8%	90.1%	13.0%	89.9%
VCOD_P53 C-term	0.978	0.546	0.551	19.6%	90.1%	14.0%	87.9%
VCOD_P53_95	0.886	0.567	0.572	23.4%	90.5%	24.0%	82.8%
VCOD_Kras G13C/Q61H	1.134	0.527	0.503	15.7%	90.1%	11.0%	90.9%
VCOD_CK8	0.972	0.533	0.540	16.6%	90.5%	8.0%	91.9%
VCOD_CK20	0.824	0.537	0.518	17.9%	90.1%	11.0%	81.8%
VCOD_Alpha-enolase	1.288	0.499	0.484	10.2%	90.1%	12.0%	90.9%
VCOD_Lmyc2	1.035	0.535	0.561	17.4%	90.1%	8.0%	89.9%

Signal to Vol Ratio (STVR) Features

Summary values for accuracy in the training cohort (training accuracy), area under the ROC curve in the training cohort (training AUC), and sensitivity and specificity at the optimal cutpoint in both training cohort (training sensitivity and training specificity respectively), and test cohort (test sensitivity and test specificity respectively) for STVR features are summarised in Table 8-3, showing that for the majority of features, specificity is once again maintained in the test cohort, with a mean decrease in the test cohort of only 1.1%, the largest change being in CAGE, in which the specificity decreased to 84.8% compared to 90.5% in the training cohort. Sensitivity does show a small drop off in the test cohort, with a mean decrease of 1.7% across all features, and the largest reduction being shown in p53 in which the sensitivity reduction is 9.0%. Summary ROC plots were again constructed to allow comparison of the features discriminatory ability over the dynamic range of the assay, and these are included in appendix 2.

Table 8-3: Discriminatory ability of individual Signal to Vol Ratio magnitude and curve characteristic features, optimised over training cohort and applied to test cohort.

Feature	optimal cutpoint	training accuracy	training AUC	training sens	training spec	test sens	test spec
STVR_p53	0.477	0.565	0.538	23.0%	90.5%	14.0%	87.9%
STVR_SOX_2	0.929	0.522	0.508	14.9%	90.1%	18.0%	89.9%
STVR_CAGE	0.682	0.548	0.549	19.6%	90.5%	17.0%	84.8%
STVR_NY_ESO_1	0.257	0.585	0.600	27.2%	90.1%	28.0%	86.9%
STVR_GBU_4_5	1.249	0.497	0.476	9.4%	90.5%	15.0%	86.9%
STVR_MAGE_A4	1.056	0.531	0.528	16.6%	90.1%	10.0%	91.9%
STVR_HuD	0.625	0.546	0.545	19.6%	90.1%	12.0%	89.9%
STVR_p16-C	0.515	0.550	0.531	20.4%	90.1%	22.0%	88.9%
STVR_KOC	0.947	0.522	0.520	14.9%	90.1%	10.0%	89.9%
STVR_ALDH1	1.123	0.499	0.489	9.8%	90.5%	11.0%	88.9%
STVR_p62	0.573	0.546	0.545	18.7%	90.9%	11.0%	90.9%
STVR_SSX1	0.837	0.514	0.520	13.2%	90.1%	15.0%	91.9%
STVR_p53 C-term	0.617	0.542	0.556	18.7%	90.1%	13.0%	88.9%
STVR_p53_95	0.307	0.565	0.546	23.0%	90.5%	21.0%	86.9%
STVR_Kras G13C/Q61H	0.823	0.525	0.500	14.9%	90.5%	13.0%	91.9%
STVR_CK8	0.894	0.537	0.531	17.0%	90.9%	14.0%	89.9%
STVR_CK20	0.967	0.527	0.513	15.3%	90.5%	13.0%	87.9%
STVR_Alpha-enolase	1.108	0.497	0.494	9.4%	90.5%	14.0%	94.9%
STVR_Lmyc2	1.187	0.520	0.532	14.5%	90.1%	16.0%	86.9%

8.5 Boruta Feature Selection

8.5.1 Methods

Boruta analysis was undertaken in R (v4.2.1) using the Boruta library, with all features showing a significantly higher importance than that of the maximum importance of the shadow features being included in the subsequent Boruta feature sets. This was completed for both VCOD features, and STVR features. Any features which showed importance that was not significantly different to the importance of the shadow features were considered as tentative candidates. These were reclassified by comparing their median importance over the entire run compared to the median value of the maximum importance of the shadow features, features showing greater importance than the shadow features were retained.

8.5.2 Results

VCOD Features

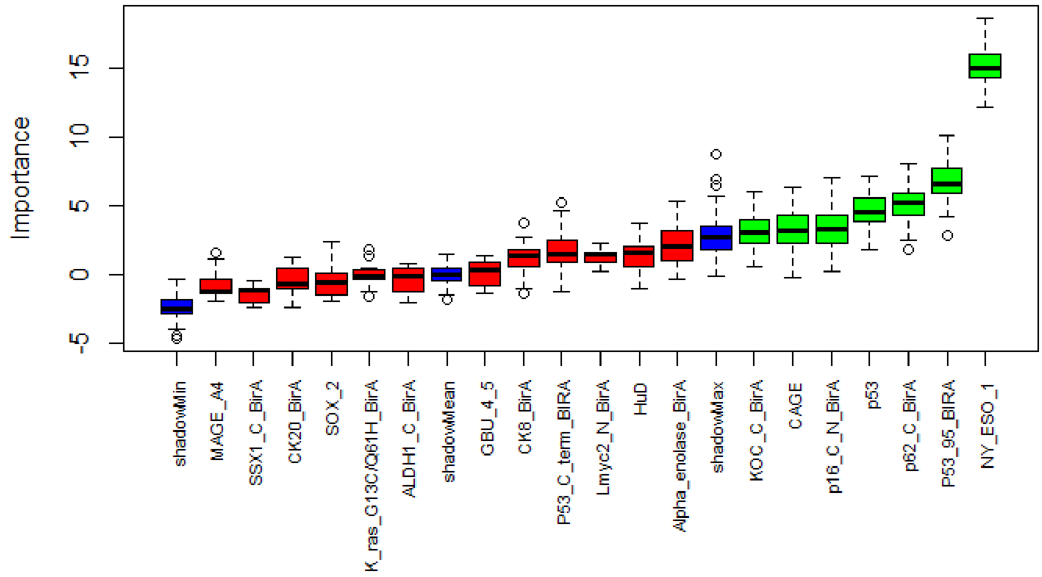


Figure 8-1: Boruta feature selection undertaken on VCOD corrected features, Importance was calculated as z-score of the mean accuracy decrease.

Boruta selection on the VCOD feature set identified 7 features which were considered important after comparison to shadow features, the full results are displayed in Figure 8-1, which shows that the EarlyCDT commercial autoantibodies NY-ESO-1, p53, and CAGE were deemed important, along with the extended panel autoantibodies p53_95, p62, p16, and KOC. This set of autoantibody features were examined as distinct feature set in the subsequent modelling analyses.

STVR Features

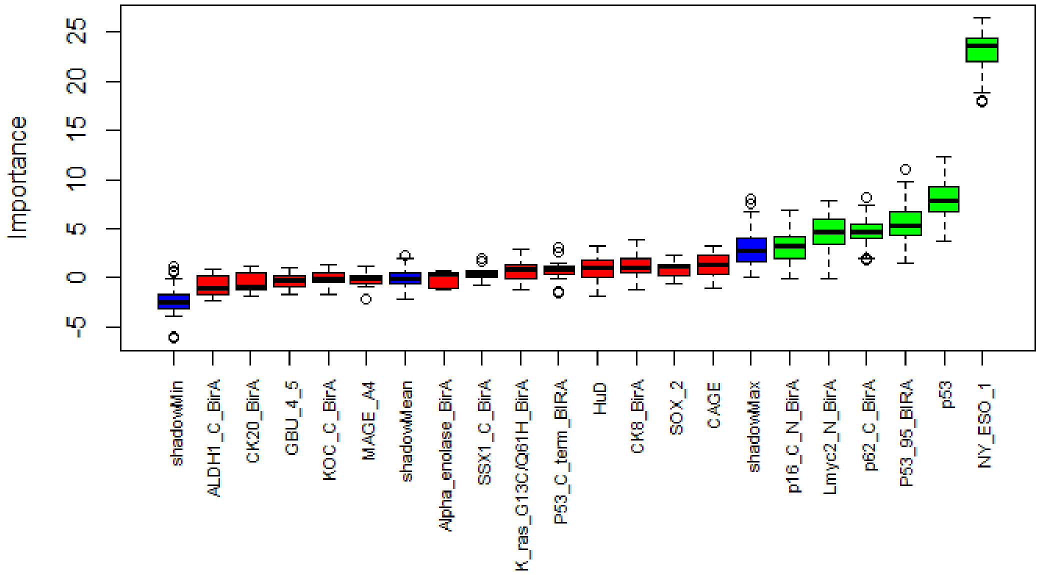


Figure 8-2: Boruta feature selection undertaken on STVR corrected features, Importance was calculated as z-score of the mean accuracy decrease.

Boruta selection on the STVR feature set identified 6 features which were considered important after comparison to shadow features, the full results are displayed in Figure 8-2, which shows that similar to the analysis on the

VCOD features, the EarlyCDT autoantibodies NY-ESO-1 and p53 were deemed important, along with the additional autoantibodies p53_95, p62, L-Myc-2 and p16-C. This set of 6 features was examined as a distinct feature set in the subsequent modelling analyses.

8.6 Penalised Classification Search Function

For the subsequent supervised modelling analysis, a two-class penalised classification search function was again used, as described in section 7.6.

8.7 Binary Logistic Regression Modelling

8.7.1 Methods

Least absolute shrinkage and selection operator (LASSO) regression was undertaken in R (v4.2.1) using the `glmnet` training method within the caret library to fit a binary logistic regression model via penalized maximum likelihood.

Parameter Tuning of Binary Logistic Models

To optimise the model fitting, a LASSO approach was used, employing values for the regularization constant of 0.001 to 0.1, at intervals of 0.001, and using 10-fold cross-validation to tune parameters to reduce any effect of overfitting. During cross validation, model performance was summarised and compared using a penalised classification metric designed to prioritise high specificity, as described previously.

8.7.2 Results

Performance of the LASSO regression models trained on the training cohort for each of the candidate feature sets are summarised in Figure 7-3 and potential diagnostic model performance is summarised in Table 7-3, showing model performance with probability thresholds optimised based on maximising sensitivity for specificity greater than 90% in the training cohort. These show that the trained regression models were unable to exceed the current commercial performance for the majority of the explored feature sets and cohorts, with the most promising being the model trained on the Boruta identified feature set of STVR corrected features, however the high specificity obtained from the training cohort was not maintained in the test and validation cohorts, resulting in specificity that would be too low for a screening modality.

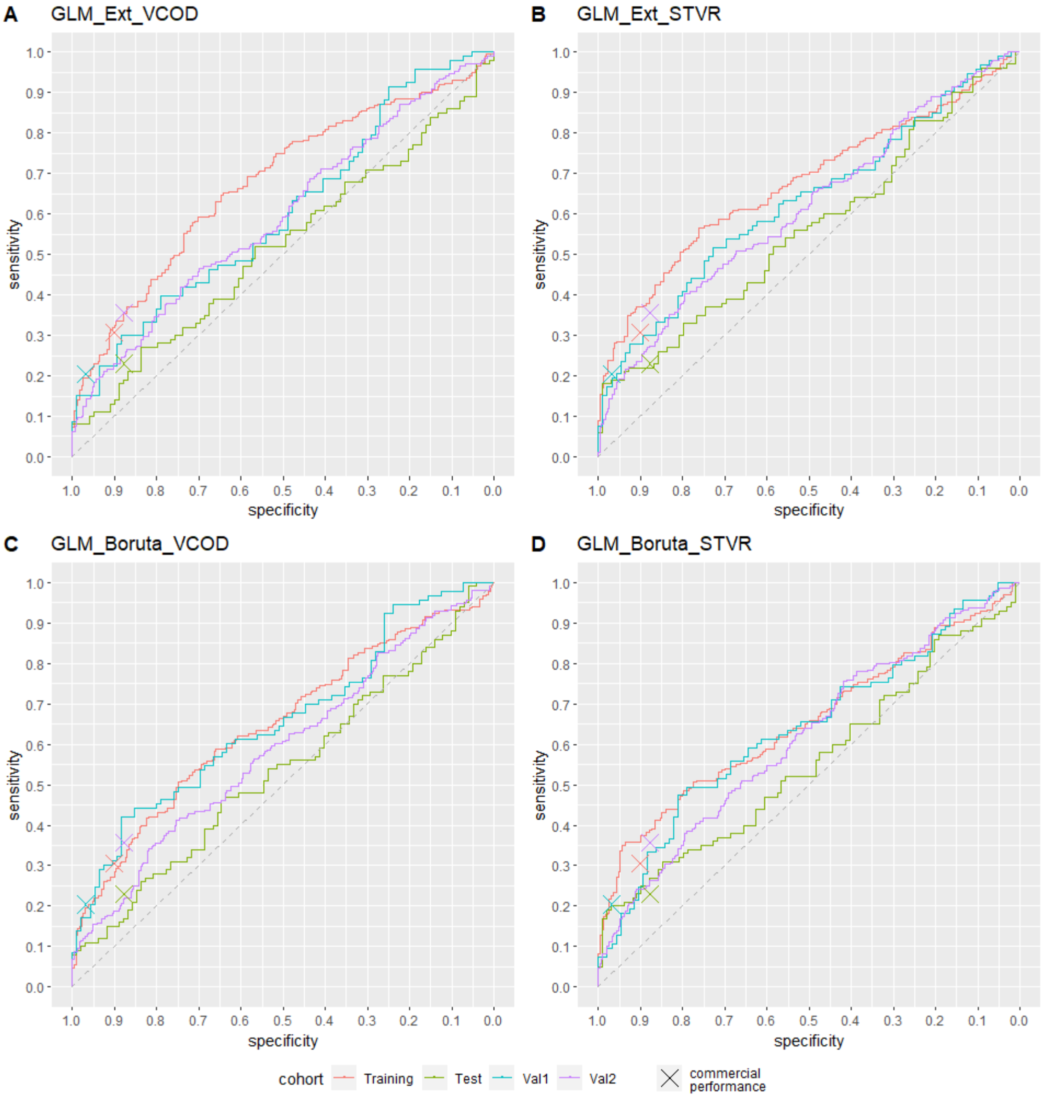


Figure 8-3: Comparison of diagnostic performance of LASSO regression models. A and B) Full extended antigen panel. C and D) Boruta selected features. A, and C) Subtraction of VOL for correction of non-specific binding. B, and D) Ratio of antigen to VOL signal for correction of non-specific binding.

Table 8-4: Summary of diagnostic performance of LASSO models on explored feature sets. Bold italicized values represent an improvement over performance obtained by the current commercial test, as defined by an increase in the Youden index.

NSB Correction	Feature set	Cohort	TP	FN	FP	TN	Sens	Spec
RU	Commercial	Training	72	163	23	209	30.6%	90.1%
RU	Commercial	Test	23	77	12	87	23.0%	87.8%
RU	Commercial	Validation 1	19	74	3	93	20.4%	96.9%
RU	Commercial	Validation 2	74	134	38	271	35.6%	87.7%
VCOD	All features	Training	***76***	***159***	***23***	***209***	***32.3%***	***90.1%***
VCOD	All features	Test	24	76	16	83	24.0%	83.8%
VCOD	All features	Validation 1	***28***	***65***	***12***	***84***	***30.1%***	***87.5%***
VCOD	All features	Validation 2	71	137	60	249	34.1%	80.6%
VCOD	Boruta	Training	67	168	23	209	28.5%	90.1%
VCOD	Boruta	Test	20	80	14	85	20.0%	85.9%
VCOD	Boruta	Validation 1	***25***	***68***	***6***	***90***	***26.9%***	***93.8%***
VCOD	Boruta	Validation 2	55	153	49	260	26.4%	84.1%
STVR	All features	Training	***87***	***148***	***22***	***210***	***37.0%***	***90.5%***
STVR	All features	Test	24	76	14	85	24.0%	85.9%
STVR	All features	Validation 1	**32**	**61**	**16**	**80**	**34.4%**	**83.3%**
STVR	All features	Validation 2	73	135	56	253	35.1%	81.9%
STVR	Boruta	Training	***88***	***147***	***23***	***209***	***37.4%***	***90.1%***
STVR	Boruta	Test	***27***	***73***	***14***	***85***	***27.0%***	***85.9%***
STVR	Boruta	Validation 1	***31***	***62***	***13***	***83***	***33.3%***	***86.5%***
STVR	Boruta	Validation 2	68	140	58	251	32.7%	81.2%

8.8 Support Vector Machine Models

8.8.1 Methods

Support vector machine learning was undertaken in R (v4.2.1) using the \`svmLinear\` and \`svmRadial\` training methods within the caret library to explore both linear kernel and radial kernel support vector classification rules.

Parameter Tuning of Support Vector Machine Models

Linear support vector machines were optimised through exploring values of the misclassification cost ‘C’ between 0.5 and 4, smaller values of which allows for higher rates of misclassification in order to give separating hyperplanes with larger margins, while higher values result in lower training misclassification and hyperplanes with smaller margins, which may result in overfit models and poorer test set performance. 10-fold cross-validation was used to determine the optimal value of C for the final model.

8.8.2 Results - Linear Kernel

Performance of the linear kernel support vector models trained on the training cohort for each of the candidate feature sets are summarised in Figure 7-4 and potential diagnostic model performance is summarised in Table 7-4, showing model performance with probability thresholds optimised based on maximising sensitivity for specificity greater than 90% in the training cohort. These models showed improvements over the current commercial test in only the training cohort for all feature sets other than the full set of STVR corrected features, which showed improvement in the test cohort and Validation 1 cohort also. The increased performance was at the cost of specificity in the test and validation sets, with these sets showing specificities around 85% which would be insufficient for a screening modality, especially as the performance gains were only in the range of around 2%-4%.

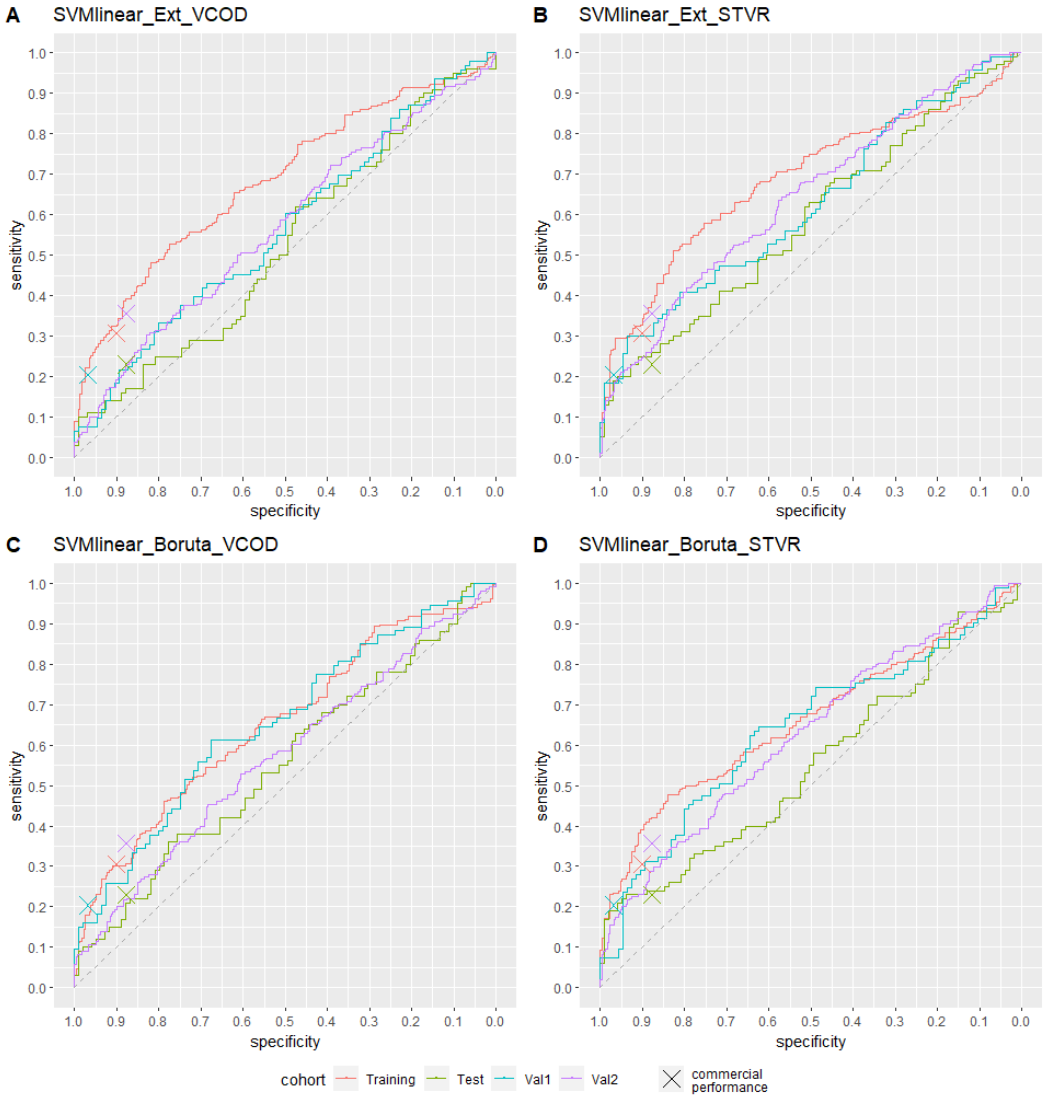


Figure 8-4: Comparison of diagnostic performance of linear kernel support vector machine models. A and B) Full extended antigen panel. C and D) Boruta selected features. A, and C) Subtraction of VOL for correction of non-specific binding. B, and D) Ratio of antigen to VOL signal for correction of non-specific binding.

Table 8-5: Summary of diagnostic performance of linear kernel support vector machine models on explored feature sets. Bold italicized values represent an improvement over performance obtained by the current commercial test, as defined by an increase in the Youden index.

NSB Correction	Feature set	Cohort	TP	FN	FP	TN	Sens	Spec
RU	Commercial	Training	72	163	23	209	30.6%	90.1%
RU	Commercial	Test	23	77	12	87	23.0%	87.8%
RU	Commercial	Validation 1	19	74	3	93	20.4%	96.9%
RU	Commercial	Validation 2	74	134	38	271	35.6%	87.7%
VCOD	All features	Training	***77***	***158***	***23***	***209***	***32.8%***	***90.1%***
VCOD	All features	Test	24	76	19	80	24.0%	80.8%
VCOD	All features	Validation 1	23	70	14	82	24.7%	85.4%
VCOD	All features	Validation 2	58	150	51	258	27.9%	83.5%
VCOD	Boruta	Training	***71***	***164***	***22***	***210***	***30.2%***	***90.5%***
VCOD	Boruta	Test	21	79	13	86	21.0%	86.9%
VCOD	Boruta	Validation 1	24	69	12	84	25.8%	87.5%
VCOD	Boruta	Validation 2	55	153	49	260	26.4%	84.1%
STVR	All features	Training	***81***	***154***	***23***	***209***	***34.5%***	***90.1%***
STVR	All features	Test	***27***	***73***	***14***	***85***	***27.0%***	***85.9%***
STVR	All features	Validation 1	***33***	***60***	***15***	***81***	***35.5%***	***84.4%***
STVR	All features	Validation 2	66	142	46	263	31.7%	85.1%
STVR	Boruta	Training	***92***	***143***	***22***	***210***	***39.1%***	***90.5%***
STVR	Boruta	Test	26	74	17	82	26.0%	82.8%
STVR	Boruta	Validation 1	30	63	16	80	32.3%	83.3%
STVR	Boruta	Validation 2	75	133	57	252	36.1%	81.6%

8.8.3 Results - Radial Kernel

Performance of the radial kernel support vector models trained on the training cohort for each of the candidate feature sets are summarised in Figure 7-5 and potential diagnostic model performance is summarised in Table 7-5, showing model performance with probability thresholds optimised based on maximising sensitivity for specificity greater than 90% in the training cohort. Examination of the ROC plots suggests that support vector machine models using a radial kernel STVR correction for non-specific binding show a higher degree of overfitting to the training cohort than VCOD corrected feature sets, as the ROC curves for the test and validation sets in the STVR models trend closer to the line of equality than those shown by models trained on VCOD feature sets. Radial kernel support vector machine models trained on this data show little evidence of being able to return diagnostic performance that is superior to that returned by the current commercial panel strategy.

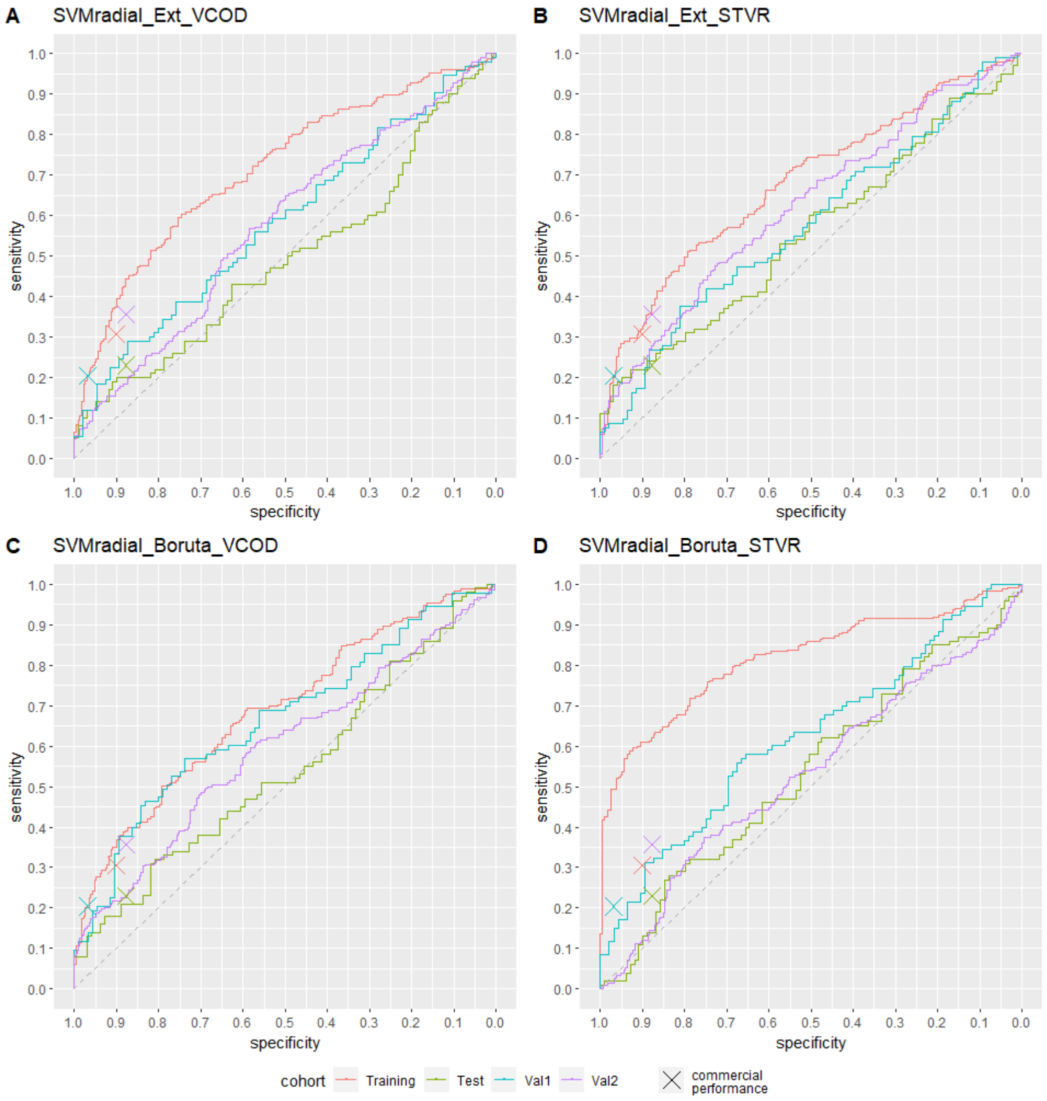


Figure 8-5: Comparison of diagnostic performance of radial kernel support vector machine models. A and B) Full extended antigen panel. C and D) Boruta selected features. A, and C) Subtraction of VOL for correction of non-specific binding. B, and D) Ratio of antigen to VOL signal for correction of non-specific binding.

Table 8-6: Summary of diagnostic performance of radial kernel support vector machine models on explored feature sets. Bold italicized values represent an improvement over performance obtained by the current commercial test, as defined by an increase in the Youden index.

NSB Correction	Feature set	Cohort	TP	FN	FP	TN	Sens	Spec
RU	Commercial	Training	72	163	23	209	30.6%	90.1%
RU	Commercial	Test	23	77	12	87	23.0%	87.8%
RU	Commercial	Validation 1	19	74	3	93	20.4%	96.9%
RU	Commercial	Validation 2	74	134	38	271	35.6%	87.7%
VCOD	All features	Training	***93***	***142***	***23***	***209***	***39.6%***	***90.1%***
VCOD	All features	Test	20	80	18	81	20.0%	81.8%
VCOD	All features	Validation 1	28	65	18	78	30.1%	81.3%
VCOD	All features	Validation 2	59	149	69	240	28.4%	77.7%
VCOD	Boruta	Training	***87***	***148***	***23***	***209***	***37.0%***	***90.1%***
VCOD	Boruta	Test	21	79	16	83	21.0%	83.8%
VCOD	Boruta	Validation 1	***30***	***63***	***9***	***87***	***32.3%***	***90.6%***
VCOD	Boruta	Validation 2	63	145	51	258	30.3%	83.5%
STVR	All features	Training	***77***	***158***	***23***	***209***	***32.8%***	***90.1%***
STVR	All features	Test	***23***	***77***	***11***	***88***	***23.0%***	***88.9%***
STVR	All features	Validation 1	25	68	13	83	26.9%	86.5%
STVR	All features	Validation 2	69	139	54	255	33.2%	82.5%
STVR	Boruta	Training	***143***	***92***	***23***	***209***	***60.9%***	***90.1%***
STVR	Boruta	Test	37	63	31	68	37.0%	68.7%
STVR	Boruta	Validation 1	41	52	27	69	44.1%	71.9%
STVR	Boruta	Validation 2	91	117	114	195	43.8%	63.1%

8.9 Naïve Bayes Models

8.9.1 Methods

Naïve Bayes modelling was undertaken in R (v4.2.1) using the `NaiveBayes` function within the `klaR` package (v4.2.3), and training using the functions included in the caret library to explore Naïve Bayes classification rules.

Parameter Tuning of Naïve Bayes Models

Naïve Bayes models were optimised through exploration of the use of kernel function, if used, adjustment of the bandwidth of the kernel function density, with values of 0 to 5 at intervals of 1, and through use of Laplace smoothing at values of 0, 0.001, 0.1, 1, 10, and 100. These hyperparameters were tested using 10-fold cross-validation to identify the optimal set of hyperparameters for the Naïve Bayes models.

8.9.2 Results

Performance of the Naïve Bayes models trained on the training cohort for each of the candidate feature sets are summarised in Figure 7-6 and potential diagnostic model performance is summarised in Table 7-6, showing model performance with probability thresholds optimised based on maximising sensitivity for specificity greater than 90% in the training cohort. These models showed the ability to improve or match the current clinical test in the training cohort for each feature set, however specificity was generally not maintained in the validation cohorts, dropping to as low as 74.4% in the validation 2 cohort, suggesting that these models experienced a higher degree of overfitting than the commercial panel, and none of the modelling on the explored feature sets showed consistent improvement over current commercial performance.

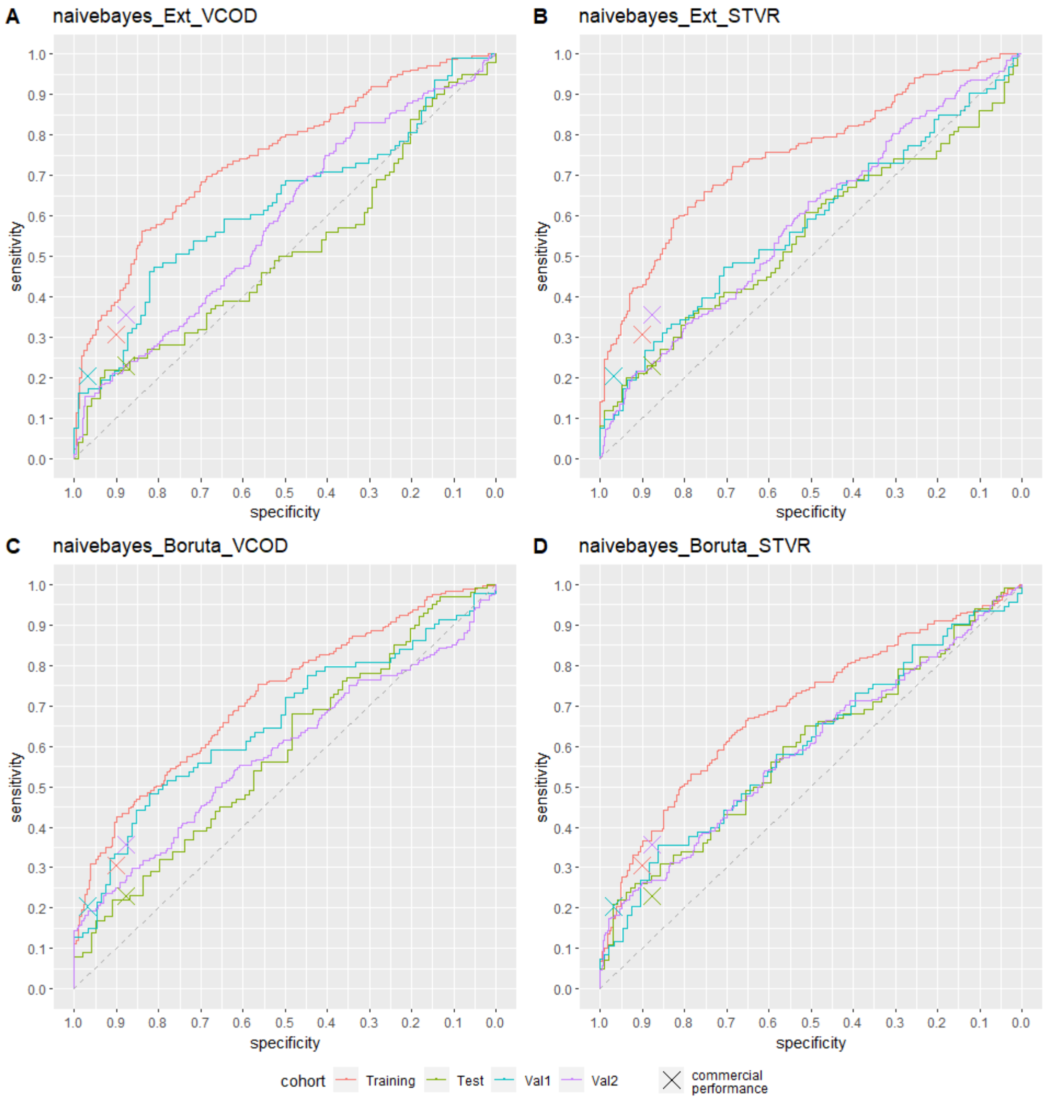


Figure 8-6: Comparison of diagnostic performance of Naïve Bayes models. A and B) Full extended antigen panel. C and D) Boruta selected features. A, and C) Subtraction of VOL for correction of non-specific binding. B, and D) Ratio of antigen to VOL signal for correction of non-specific binding.

Table 8-7: Summary of diagnostic performance of Naïve Bayes models on explored feature sets. Bold italicized values represent an improvement over performance obtained by the current commercial test, as defined by an increase in the Youden index.

NSB Correction	Feature set	Cohort	TP	FN	FP	TN	Sens	Spec
RU	Commercial	Training	72	163	23	209	30.6%	90.1%
RU	Commercial	Test	23	77	12	87	23.0%	87.8%
RU	Commercial	Validation 1	19	74	3	93	20.4%	96.9%
RU	Commercial	Validation 2	74	134	38	271	35.6%	87.7%
VCOD	All features	Training	***91***	***144***	***22***	***210***	***38.7%***	***90.5%***
VCOD	All features	Test	24	76	14	85	24.0%	85.9%
VCOD	All features	Validation 1	***43***	***50***	***17***	***79***	***46.2%***	***82.3%***
VCOD	All features	Validation 2	68	140	79	230	32.7%	74.4%
VCOD	Boruta	Training	***100***	***135***	***23***	***209***	***42.6%***	***90.1%***
VCOD	Boruta	Test	24	76	16	83	24.0%	83.8%
VCOD	Boruta	Validation 1	***41***	***52***	***16***	***80***	***44.1%***	***83.3%***
VCOD	Boruta	Validation 2	69	139	62	247	33.2%	79.9%
STVR	All features	Training	***101***	***134***	***23***	***209***	***43.0%***	***90.1%***
STVR	All features	Test	***26***	***74***	***14***	***85***	***26.0%***	***85.9%***
STVR	All features	Validation 1	32	61	19	77	34.4%	80.2%
STVR	All features	Validation 2	75	133	78	231	36.1%	74.8%
STVR	Boruta	Training	***86***	***149***	***23***	***209***	***36.6%***	***90.1%***
STVR	Boruta	Test	***30***	***70***	***14***	***85***	***30.0%***	***85.9%***
STVR	Boruta	Validation 1	***29***	***64***	***13***	***83***	***31.2%***	***86.5%***
STVR	Boruta	Validation 2	70	138	68	241	33.7%	78.0%

8.10 C5.0 Decision Tree Modelling

8.10.1 Methods

C5.0 model training was undertaken in R (v4.2.1), using the `C5.0` training method within the caret library.

Parameter Tuning of C5.0 Decision Tree Models

10-fold cross-validation was used to determine whether applying winnowing – removal of uninformative features before training the tree model - was applied. During cross validation, model performance was summarised and compared using a penalised classification metric designed to prioritise high specificity, as described previously.

8.10.2 Results

Performance of the c5.0 decision tree models trained on the training cohort for each of the candidate feature sets are summarised in Figure 7-7 and potential diagnostic model performance is summarised in
Table 7-7Table 7-6, showing model performance with probability thresholds optimised based on maximising sensitivity for specificity greater than 90% in the training cohort. The ROC plots suggest that the STVR correction method results in a higher degree of overfitting using c5.0 decision tree modelling. VCOD correction showed a lower degree of overfitting, although this could be due to the relatively small tree formed, as the same decision tree resulted from modelling of both the full feature set, and the Boruta selected feature set and was comprised of only NY-ESO-1 and p53.

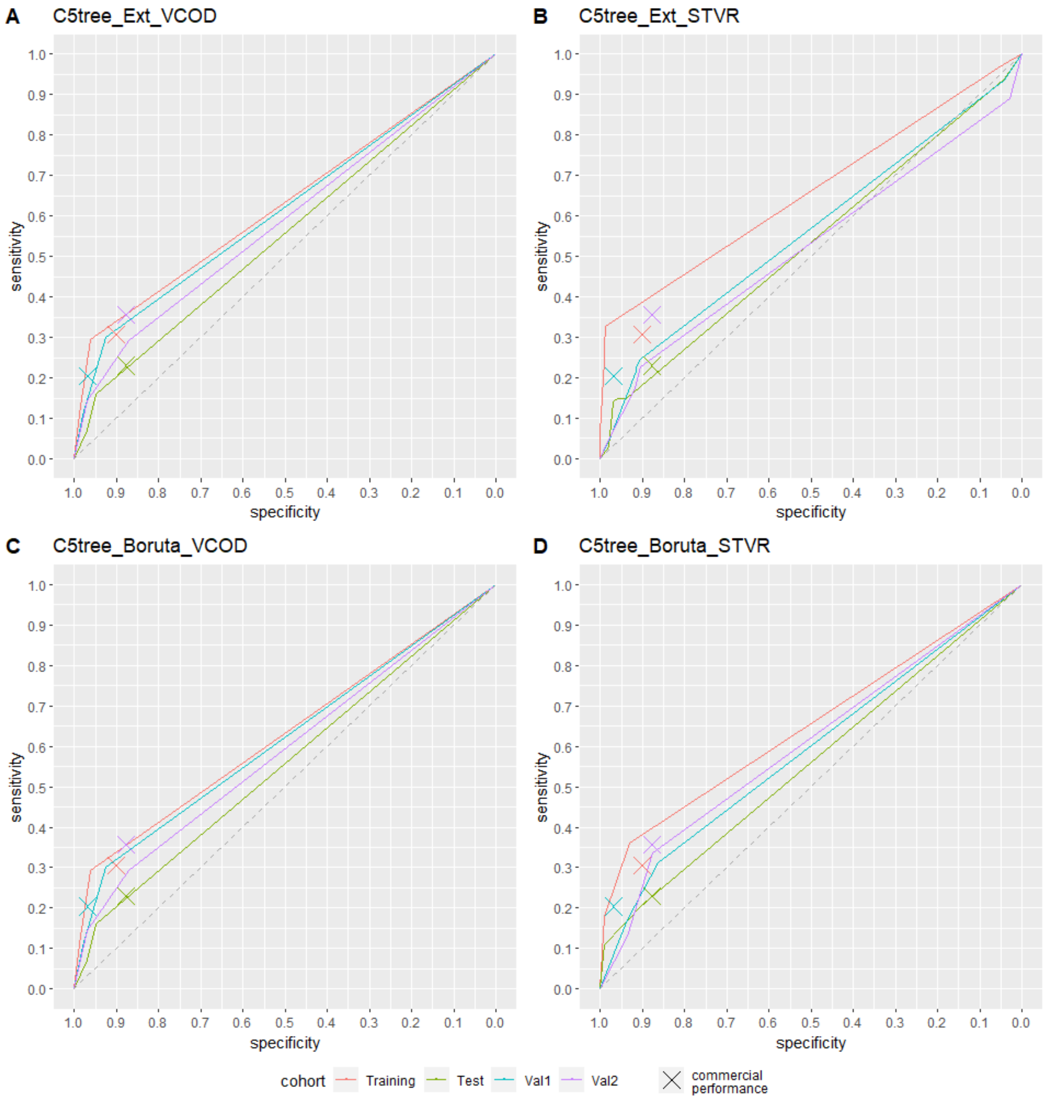


Figure 8-7: Comparison of diagnostic performance of C5.0 decision tree models. A and B) Full extended antigen panel. C and D) Boruta selected features. A, and C) Subtraction of VOL for correction of non-specific binding. B, and D) Ratio of antigen to VOL signal for correction of non-specific binding.

Table 8-8: Summary of diagnostic performance of C5.0 decision tree models on explored feature sets. Bold italicized values represent an improvement over performance obtained by the current commercial test, as defined by an increase in the Youden index.

NSB Correction	Feature set	Cohort	TP	FN	FP	TN	Sens	Spec
RU	Commercial	Training	72	163	23	209	30.6%	90.1%
RU	Commercial	Test	23	77	12	87	23.0%	87.8%
RU	Commercial	Validation 1	19	74	3	93	20.4%	96.9%
RU	Commercial	Validation 2	74	134	38	271	35.6%	87.7%
VCOD	All features	Training	***69***	***166***	***9***	***223***	***29.4%***	***96.1%***
VCOD	All features	Test	***16***	***84***	***5***	***94***	***16.0%***	***94.9%***
VCOD	All features	Validation 1	***28***	***65***	***7***	***89***	***30.1%***	***92.7%***
VCOD	All features	Validation 2	61	147	40	269	29.3%	87.1%
VCOD	Boruta	Training	***69***	***166***	***9***	***223***	***29.4%***	***96.1%***
VCOD	Boruta	Test	***16***	***84***	***5***	***94***	***16.0%***	***94.9%***
VCOD	Boruta	Validation 1	***28***	***65***	***7***	***89***	***30.1%***	***92.7%***
VCOD	Boruta	Validation 2	61	147	40	269	29.3%	87.1%
STVR	All features	Training	***77***	***158***	***3***	***229***	***32.8%***	***98.7%***
STVR	All features	Test	15	85	6	93	15.0%	93.9%
STVR	All features	Validation 1	23	70	9	87	24.7%	90.6%
STVR	All features	Validation 2	47	161	29	280	22.6%	90.6%
STVR	Boruta	Training	***85***	***150***	***16***	***216***	***36.2%***	***93.1%***
STVR	Boruta	Test	***21***	***79***	***10***	***89***	***21.0%***	***89.9%***
STVR	Boruta	Validation 1	***29***	***64***	***13***	***83***	***31.2%***	***86.5%***
STVR	Boruta	Validation 2	70	138	38	271	33.7%	87.7%

8.11 Random Forest Modelling

8.11.1 Methods

Random Forest modelling was undertaken in R (v4.2.1) using the ranger library, applying the `ranger` training method to the data within the caret library Train function.

Parameter tuning of Random Forest Models

Random Forest model tuning was done over the following parameters: number of variables to possibly split at each node (mtry), splitting rule, and minimum node size.

Values for mtry were either 1 to the number of features for magnitude feature sets and boruta selected feature sets, or odd values from 1 to 35 for the combined curve characteristic and magnitude feature set.

Options for the splitting rule were either selection by minimisation of gini impurity, the extra-trees algorithm(178), or Hellinger distance(179).

Values explored for minimal node size were even numbers from 2 to 10.

Model performance for each combination of parameters was summarised over 10-fold cross validation and compared using a penalised classification metric designed to prioritise high specificity, as described previously.

8.11.2 Results

Performance of the random forest models trained on the training cohort for each of the candidate feature sets are summarised in Figure 7-8 and potential diagnostic model performance is summarised in

Table 7-8Table 7-6, the high overfitting observed in the training cohort again meant that probability thresholds based on the training cohort showed poor ability to transfer to test and validation cohorts, therefore model performance was again optimised based on maximising sensitivity for specificity greater than 90% in the test cohort. Review of the ROC plots shows that random forest modelling suffers from an extremely high degree of overfitting, with the training cohort approaching perfect discrimination in almost every case, and the models being unable to maintain their performance on test and validation cohorts. While the test and validation cohorts do not maintain the performance of the training cohort, they show some improvements over the commercial test, however the increases to the performance are modest, and not present across all cohorts.

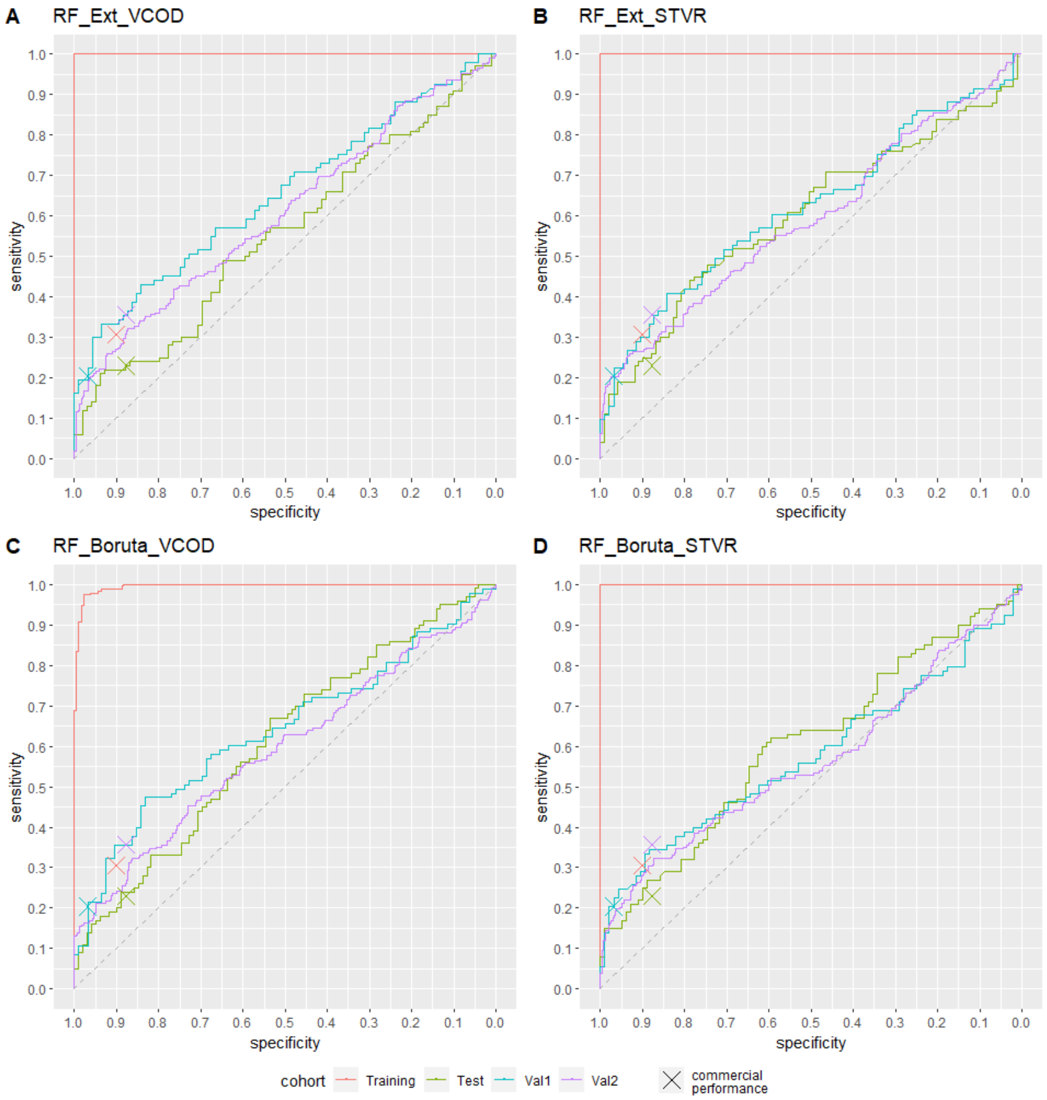


Figure 8-8: Comparison of diagnostic performance of random forest models. A and B) Full extended antigen panel. C and D) Boruta selected features. A, and C) Subtraction of VOL for correction of non-specific binding. B, and D) Ratio of antigen to VOL signal for correction of non-specific binding.

Table 8-9: Summary of diagnostic performance of random forest models on explored feature sets. Bold italicized values represent an improvement over performance obtained by the current commercial test, as defined by an increase in the Youden index.

NSB Correction	Feature set	Cohort	TP	FN	FP	TN	Sens	Spec
RU	Commercial	Training	72	163	23	209	30.6%	90.1%
RU	Commercial	Test	23	77	12	87	23.0%	87.8%
RU	Commercial	Validation 1	19	74	3	93	20.4%	96.9%
RU	Commercial	Validation 2	74	134	38	271	35.6%	87.7%
VCOD	All features	Training	***235***	***0***	***0***	***232***	***100.0%***	***100.0%***
VCOD	All features	Test	***22***	***78***	***8***	***91***	***22.0%***	***91.9%***
VCOD	All features	Validation 1	***38***	***55***	***15***	***81***	***40.9%***	***84.4%***
VCOD	All features	Validation 2	71	137	50	259	34.1%	83.8%
VCOD	Boruta	Training	***131***	***104***	***0***	***232***	***55.7%***	***100.0%***
VCOD	Boruta	Test	19	81	9	90	19.0%	90.9%
VCOD	Boruta	Validation 1	***30***	***63***	***8***	***88***	***32.3%***	***91.7%***
VCOD	Boruta	Validation 2	53	155	36	273	25.5%	88.3%
STVR	All features	Training	***234***	***1***	***0***	***232***	***99.6%***	***100.0%***
STVR	All features	Test	***24***	***76***	***9***	***90***	***24.0%***	***90.9%***
STVR	All features	Validation 1	***31***	***62***	***11***	***85***	***33.3%***	***88.5%***
STVR	All features	Validation 2	61	147	40	269	29.3%	87.0%
STVR	Boruta	Training	***198***	***37***	***0***	***232***	***84.3%***	***100.0%***
STVR	Boruta	Test	***22***	***78***	***9***	***90***	***22.0%***	***90.9%***
STVR	Boruta	Validation 1	***25***	***68***	***8***	***88***	***26.9%***	***91.7%***
STVR	Boruta	Validation 2	60	148	32	277	28.8%	89.6%

8.12 Extreme Gradient Boosted Trees Modelling

8.12.1 Methods

Random Forest modelling was undertaken in R (v4.2.1) using the ranger library, and applying the `ranger` training method to the data within the caret library Train function.

Parameter tuning of XGBoost Models

XGBoost modelling includes many parameters over which the model can be tuned. Tuning was again accomplished over 10-fold cross-validation, and the tuning parameters and their explored values are detailed here:

eta – shrinkage rate used in the regularization step to reduce overfitting. Values searched: 0.01, 0.05, 0.1, 0.2, and 0.3.

max_depth – Maximum depth (number of edges from the root node to the furthest leaf node) of a tree. Larger values make the model more complex and more likely to overfit. Values searched: 2, 4, 6, and 8.

gamma – Minimum loss reduction required to make a further partition on a leaf node of the tree. Values searched: 0, and 5.

subsample – Subsample ratio of the training instances. Values searched: 0.5, 0.75, and 1.

colsample_bytree –Subsample ratio of columns when constructing each tree. Values searched: 0.5, 0.75, and 1.

min_child_weight – Minimum sum of instance weight needed in a child, larger values result in more conservative models. Values searched: 2, 4, 6, 8, and 10.

During cross validation, model performance was summarised and compared using a penalised classification metric designed to prioritise high specificity, as described previously.

8.12.2 Results

Performance of the extreme gradient boosted trees models trained on the training cohort for each of the candidate feature sets are summarised in Figure 7-9Figure 7-7 and potential diagnostic model performance is summarised in Table 7-9Table 7-6, showing model performance with probability thresholds optimised based on maximising sensitivity for specificity greater than 90% in the training cohort. These models again showed improved performance over the current commercial panel for the training and validation 1 cohort in each case, and once again these models showed generally poor specificity in the Validation 2 and test cohorts.

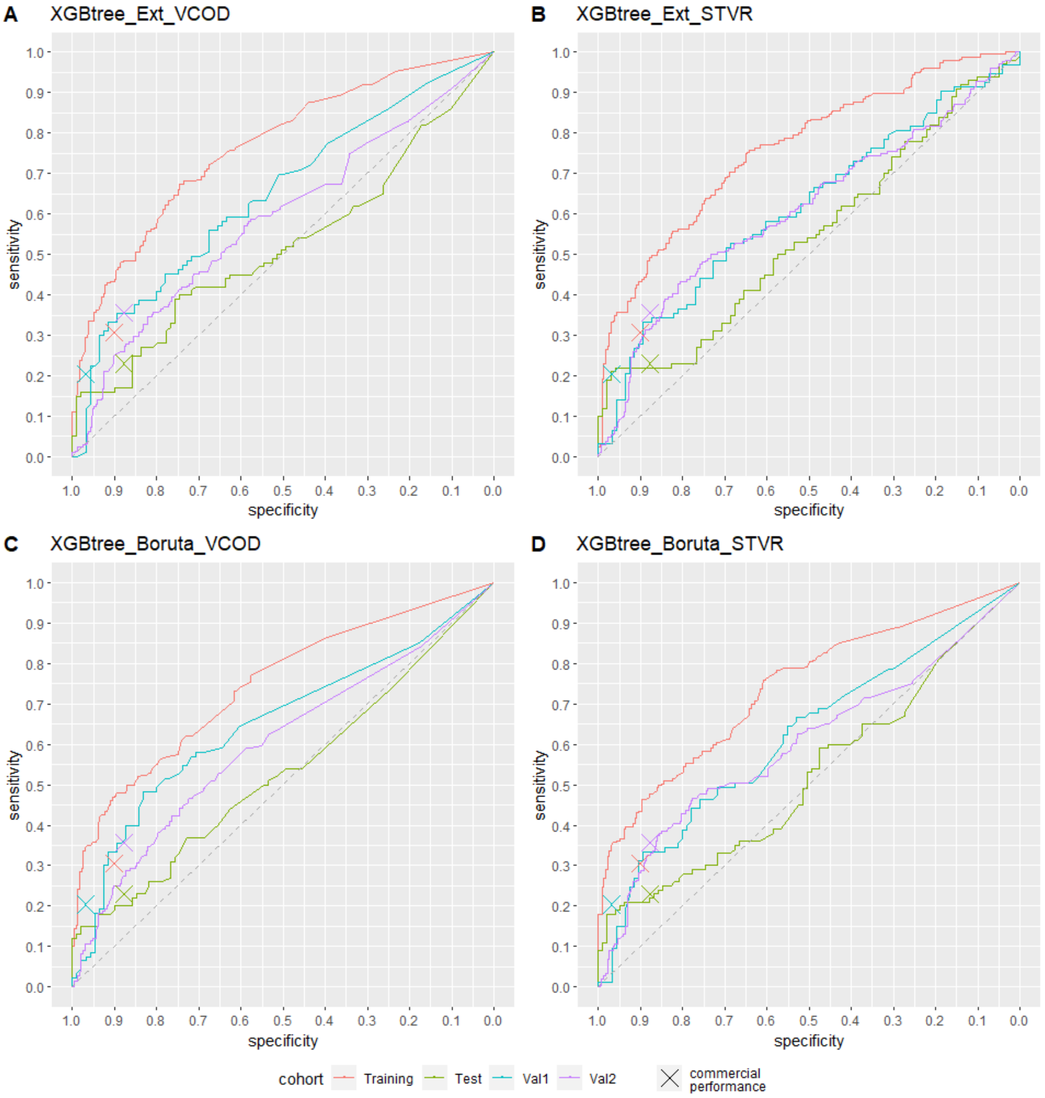


Figure 8-9: Comparison of diagnostic performance of extreme gradient boosted tree models. A and B) Full extended antigen panel. C and D) Boruta selected features. A, and C) Subtraction of VOL for correction of non-specific binding. B, and D) Ratio of antigen to VOL signal for correction of non-specific binding.

Table 8-10: Summary of diagnostic performance of extreme gradient boosted tree models on explored feature sets. Bold italicized values represent an improvement over performance obtained by the current commercial test, as defined by an increase in the Youden index.

NSB Correction	Feature set	Cohort	TP	FN	FP	TN	Sens	Spec
RU	Commercial	Training	72	163	23	209	30.6%	90.1%
RU	Commercial	Test	23	77	12	87	23.0%	87.8%
RU	Commercial	Validation 1	19	74	3	93	20.4%	96.9%
RU	Commercial	Validation 2	74	134	38	271	35.6%	87.7%
VCOD	All features	Training	***103***	***132***	***23***	***209***	***43.8%***	***90.1%***
VCOD	All features	Test	32	68	23	76	32.0%	76.8%
VCOD	All features	Validation 1	***36***	***57***	***16***	***80***	***38.7%***	***83.3%***
VCOD	All features	Validation 2	80	128	73	236	38.5%	76.4%
VCOD	Boruta	Training	***111***	***124***	***23***	***209***	***47.2%***	***90.1%***
VCOD	Boruta	Test	26	74	20	79	26.0%	79.8%
VCOD	Boruta	Validation 1	***38***	***55***	***15***	***81***	***40.9%***	***84.4%***
VCOD	Boruta	Validation 2	85	123	73	236	40.9%	76.4%
STVR	All features	Training	***102***	***133***	***22***	***210***	***43.4%***	***90.5%***
STVR	All features	Test	23	77	19	80	23.0%	80.8%
STVR	All features	Validation 1	***33***	***60***	***17***	***79***	***35.5%***	***82.3%***
STVR	All features	Validation 2	81	127	49	260	38.9%	84.1%
STVR	Boruta	Training	***102***	***133***	***22***	***210***	***43.4%***	***90.5%***
STVR	Boruta	Test	25	75	15	84	25.0%	84.8%
STVR	Boruta	Validation 1	***32***	***61***	***16***	***80***	***34.4%***	***83.3%***
STVR	Boruta	Validation 2	80	128	49	260	38.5%	84.1%

8.13 Summary

Table 8-11: Summary of diagnostic performance achieved by modelling strategies in explored feature sets.

Modelling Strategy	Metric	Training		Test		Validation 1		Validation 2	
		Sens	Spec	Sens	Spec	Sens	Spec	Sens	Spec
Commercial	Commercial	30.6%	90.1%	23.0%	87.8%	20.4%	96.9%	35.6%	87.7%
GLM LASSO	VCOD All	32.3%	90.1%	24.0%	83.8%	30.1%	87.5%	34.1%	80.6%
GLM LASSO	VCOD Boruta	28.5%	90.1%	20.0%	85.9%	26.9%	93.8%	26.4%	84.1%
GLM LASSO	STVR All	37.0%	90.5%	24.0%	85.9%	34.4%	83.3%	35.1%	81.9%
GLM LASSO	STVR Boruta	37.4%	90.1%	27.0%	85.9%	33.3%	86.5%	32.7%	81.2%
SVM Linear	VCOD All	32.8%	90.1%	24.0%	80.8%	24.7%	85.4%	27.9%	83.5%
SVM Linear	VCOD Boruta	30.2%	90.5%	21.0%	86.9%	25.8%	87.5%	26.4%	84.1%
SVM Linear	STVR All	34.5%	90.1%	27.0%	85.9%	35.5%	84.4%	31.7%	85.1%
SVM Linear	STVR Boruta	39.1%	90.5%	26.0%	82.8%	32.3%	83.3%	36.1%	81.6%
SVM Radial	VCOD All	39.6%	90.1%	20.0%	81.8%	30.1%	81.3%	28.4%	77.7%
SVM Radial	VCOD Boruta	37.0%	90.1%	21.0%	83.8%	32.3%	90.6%	30.3%	83.5%
SVM Radial	STVR All	32.8%	90.1%	23.0%	88.9%	26.9%	86.5%	33.2%	82.5%
SVM Radial	STVR Boruta	60.9%	90.1%	37.0%	68.7%	44.1%	71.9%	43.8%	63.1%
Naïve Bayes	VCOD All	38.7%	90.5%	24.0%	85.9%	46.2%	82.3%	32.7%	74.4%
Naïve Bayes	VCOD Boruta	42.6%	90.1%	24.0%	83.8%	44.1%	83.3%	33.2%	79.9%
Naïve Bayes	STVR All	43.0%	90.1%	26.0%	85.9%	34.4%	80.2%	36.1%	74.8%
Naïve Bayes	STVR Boruta	36.6%	90.1%	30.0%	85.9%	31.2%	86.5%	33.7%	78.0%
C5.0 Tree	VCOD All	29.4%	96.1%	16.0%	94.9%	30.1%	92.7%	29.3%	87.1%
C5.0 Tree	VCOD Boruta	29.4%	96.1%	16.0%	94.9%	30.1%	92.7%	29.3%	87.1%
C5.0 Tree	STVR All	32.8%	98.7%	15.0%	93.9%	24.7%	90.6%	22.6%	90.6%
C5.0 Tree	STVR Boruta	36.2%	93.1%	21.0%	89.9%	31.2%	86.5%	33.7%	87.7%
Random Forest	VCOD All	100.0%	100.0%	22.0%	91.9%	40.9%	84.4%	34.1%	83.8%
Random Forest	VCOD Boruta	55.7%	100.0%	19.0%	90.9%	32.3%	91.7%	25.5%	88.3%
Random Forest	STVR All	99.6%	100.0%	24.0%	90.9%	33.3%	88.5%	29.3%	87.0%
Random Forest	STVR Boruta	84.3%	100.0%	22.0%	90.9%	26.9%	91.7%	28.8%	89.6%
XGBoost	VCOD All	43.8%	90.1%	32.0%	76.8%	38.7%	83.3%	38.5%	76.4%
XGBoost	VCOD Boruta	47.2%	90.1%	26.0%	79.8%	40.9%	84.4%	40.9%	76.4%
XGBoost	STVR All	43.4%	90.5%	23.0%	80.8%	35.5%	82.3%	38.9%	84.1%
XGBoost	STVR Boruta	43.4%	90.5%	25.0%	84.8%	34.4%	83.3%	38.5%	84.1%

Performance of all explored models has been summarised, as shown in Table 7-10, to allow comparison of modelling strategies both to each other, and to the performance obtained using the commercial assessment of the sample results. This shows while machine learning models are able train models that give comparable performance to the current EarlyCDT®-Lung test cutoff threshold method, the methods explored here are unable to give consistent improvements over the current commercial test which would justify the additional work required to implement them in place of the current commercial analysis strategy.

8.14 Chapter Conclusions

A variety of supervised machine learning strategies have been explored against an extended panel of autoantibody features to determine whether they are able to add sensitivity to the commercial EarlyCDT®-Lung panel, while maintaining a high panel specificity. While several of the modelling strategies are able to return models with comparable diagnostic performance to the current commercial autoantibody panel, none were able to show consistent diagnostic performance improvements, even with the inclusion of additional autoantibody measurements.

8.15 Chapter Discussion

Again the lack of appreciable and consistent improvements over the current commercial test assessment in models trained on this expanded autoantibody panel was disappointing and surprising. It is possible that there is a law of diminishing returns when it comes to autoantibody features, and a proportion of lung cancers are able to develop without triggering detectable responses from the adaptive immune system, in which case additional biomarkers such as cell free DNA, and methylation markers, may be necessary to lend additional sensitivity in ensemble models. It is also possible that the immune response to cancer is attenuated over time, through either B-cell exhaustion or malignancies developing to escape immune control, in which case the cancer samples explored in this project, almost all taken at point of diagnostic presentation, may have been collected after an autoantibody response has become diminished, unfortunately to explore this would require an extremely large longitudinal autoantibody study and is therefore understandably outside of the scope of this project.

Chapter 9: Development of a Simulated Annealing Gradient Ascent Algorithm for Optimisation of High Specificity Multivariate Panels

9.1 Aims

In tandem with exploring methods to improve the diagnostic performance of the EarlyCDT®-Lung test, and in the absence of machine learning algorithms which showed performance improvements over that of the current logic-based panel assessment, I set about improving the method by which the panel cut-off thresholds were determined, from a time consuming Monte Carlo random search strategy, to a more focused strategy which incorporated a gradient ascent strategy towards a predetermined search function, using a simulated annealing algorithm to allow a rapid wide search of potential panel solutions while reducing the risk of getting trapped in suboptimal local maxima.

9.2 Introduction

Lung cancer is responsible for the greatest number of cancer-related deaths worldwide, representing 18.0% of all cancer deaths in 2020(1). Survival rates for lung cancer are particularly poor as the majority of patients are not diagnosed until late-stage disease, with 57% of patients presenting after the cancer has metastasized, giving a 5-year survival rate of only 4.7%, and only 16% of lung cancers are detected while the disease is still localised, with a 5-year survival of 56.3%(223). Early diagnosis of lung cancer would allow for more patients to be detected while the disease is still in its earliest stages, and vastly improve prognosis.

The presence of autoantibodies in the peripheral blood of patients with cancer has been established in a number of studies(79, 108, 129, 130, 224-226). In these studies, individual autoantibodies show high specificity, but relatively low sensitivity, reflecting the heterogeneous nature of cancer as a disease, and the high specificity of the immune response to the presence of mutated or overexpressed proteins. Due to the low sensitivities of these autoantibodies, they are individually not suitable as screening tests, however their high specificity can be exploited in the formation of biomarker panels, whereby the sensitivities of the combination of multiple autoantibody markers can reach levels which become cost effective for cancer screening or nodule stratification. This methodology has been used in the development of the EarlyCDT®-Lung test(79, 108, 227).

The EarlyCDT®-Lung autoantibody test is a panel biomarker test which uses a logic rule-based classifier, whereby each autoantibody has an associated cut-off threshold, and a signal above the cut-off in any of the panel autoantibodies is regarded as a positive response and indicates increased risk of cancer in the subject.

In its initial incarnations, the calculation of cut-off thresholds for the EarlyCDT®-Lung autoantibody panel was performed using a ‘brute force’ Monte Carlo(228) random search method to iterate through large numbers of random combinations of potential cut-off threshold values for each antigen, applying each set of thresholds to a training cohort, and returning performance statistics for the resultant panel. The panels with adequate performance on the training set were then applied to a hold-out test set to determine which of the candidate cut-off sets that showed optimal performance in the training set were reproducible.

The Monte Carlo random search method is time-consuming and necessitated a large number of iterations, to sample from a large search space which includes a high proportion of non-optimal search combinations. To reduce the time taken, and ensure that the search focused on cut-off sets that give optimised performance in the training set, an approach was developed using a simulated annealing(229, 230) approach which would allow for an iterative search over optimal threshold combinations, utilitising an adapted Metropolis algorithm(231), with the simulated annealing algorithm allowing for non-optimal movements with higher likelihood in the early stages of the search, which prevents the algorithm from resolving in panels which represent sub-optimal local maxima of the search function score.

This approach was inspired by the iterative combination of biomarkers and thresholds (ICBT) method(232), and allows for iterative improvement to a panel from a randomised start position until a set of optimal panel cut-off thresholds is found. This method differs from the ICBT method, however, as ICBT performs an exhaustive search of all combinations of biomarkers, the combinatorial complexity of which escalates vastly with increasing panel size, and number of thresholds to test. Assuming the use of all available biomarkers, each with a differing number of potential thresholds, the number of possible threshold combinations I for a panel of n biomarkers can be expressed in its simplest form as:

$$I = \prod T$$

Equation 1

Where T is a vector containing the number of thresholds of all biomarkers in the panel, and a threshold is included which is set at a value above the greatest value for the feature, which upon selection results in that feature being excluded from the panel assessment.

Due to the exponential increase in threshold combinations as either panel, or number of thresholds in each marker, increases, exhaustive search methods are limited to relatively small panels, and sampling from a small number of optimal thresholds. Additionally, as panel size increases, Monte Carlo random search methods inevitably become less specific as additional features continue to contribute greater numbers of false positives. This could be addressed by limiting the panel size, however when data from large numbers of biomarkers is available, this results in the need to either introduce a feature selection step prior to the Monte Carlo search, or iteration through the possible combinations of features which vastly increases the complexity of the search, as for a panel size of m out of n biomarkers, the number of combinations of panels C to search becomes:

$$C = \frac{n!}{m!\,(n-m)!}$$

Equation 2

This increases the number of possible threshold combinations I to test, to:

$$I = \sum_{j=1}^{C} \left(\prod T_j \right)$$

Equation 3

Where j represents a single biomarker combination within C.

With the understanding that cancers are extremely heterogenous, large panels of markers may be necessary to obtain high sensitivities, and limiting panel size through pre-filtering feature selection may exclude potentially informative biomarkers and could limit the sensitivity detectable. The following simulated annealing approach was developed to allow discovery of optimal threshold sets for larger panels than would be feasible with an exhaustive search, without the need to pre-filter, as the simulated annealing search allows for uninformative features to be removed during the iterative improvement of the panel, and with higher specificities than would be found with an untargeted Monte Carlo random search of the feature space.

9.3 Methods

9.3.1 Simulated Annealing Panel Optimisation Algorithm

For a panel of biomarkers, a patient is considered positive if the concentration of any individual biomarker exceeds a threshold for that biomarker.

This can be expressed as:

$$O_p = J\left(\left(\sum_{i=1}^{n} I\left(X_{ip} \geq T_i\right)\right) \geq 1\right)$$

Equation 4

Where O_p is the predicted outcome for patient p, n is the number of biomarkers in the panel, X_{ip} is the concentration of the i^{th} biomarker in patient p, T_i is the threshold for the i^{th} biomarker, $I(x)$ is an indicator function which takes the value of 1 for x=true, and 0 otherwise, and $J(x)$ is an indicator function which takes the value of 1 for x = true, and 0 otherwise.

ROC local maxima identification

The threshold search space was reduced by assessing candidate thresholds and removing redundant values by selecting only those values which represent local maxima of the ROC curve. For potential thresholds sorted in descending order, a local maxima was defined as a threshold at which sensitivity was greater than or equal to that obtained from the next consecutive threshold, and specificity was greater than or equal to that obtained from the previous consecutive threshold, which can be summarised as the following:

$$T_{locmax} = (SE_i \geq SE_{i+1} \,\&\, SP_i \ \geq SP_{i-1})$$

Equation 5

Whereby SE_i is sensitivity at the current threshold, SE_{i+1} is sensitivity at the threshold immediately below the current threshold, SP_i is specificity at the current threshold, SE_{i+1} is specificity at the threshold immediately above the current threshold.

Thresholds which do not represent local ROC maxim are excluded as they are by definition inferior in either sensitivity, specificity, or both, to a local ROC maxima threshold.

Search Function

A search function is established to define the target performance of the final optimised panel. During supervised panel training, this could be any metric that can be derived from comparing the predicted disease outcome to the actual disease outcome, broken down into numbers of true positives (TP), false positives (FP), false negatives (FN) and true negatives (TN), and includes (but is not limited to):

Accuracy: $(TP + TN)/(TP + FP + FN + TN)$

Youden Index: $\left(\left(\frac{TP}{TP+FN}\right) + \left(\frac{TN}{TN+FP}\right)\right) - 1$

F1 Score: $(2 * TP)\ /\ (2\ *\ TP\ +\ FP\ +\ FN)$

In addition, a custom search function has been designed to prioritise panels with maximised sensitivity for specificity around 95%:

High Specificity Function: $((TP/(TP + FN)) - ((0.95 - (TN/(TN + FP)))\text{^}2))$

Initial `Hill Climb` optimisation

Features are added iteratively in a random order to the panel, with each additional feature having its initial threshold selected to maximise a search function relevant to the panel being optimised. To prioritise high specificity, once all features have been added to the panel, they are randomly iterated through and their thresholds increased, provided increasing the threshold has no effect on the score returned by the search function.

Simulated Annealing optimisation

The elements of the simulated annealing panel optimisation search are as follows:

1: The panel of biomarkers $i_1, i_2 \dots i_n$.

2: For each biomarker i, the set of potential cut-off thresholds $T_i \in T_1, T_2 \dots T_t$.

3: The search function C, giving a score c for the measured combination of biomarker thresholds, where $\dot{c} \subset c$ representing scores at a search maxima.

4: For each biomarker threshold T_i, a set of neighbouring thresholds T_{i-k}: T_{i+k}.

5: A cooling schedule, defined as a nonincreasing function $R: N \to (0, \infty)$, whereby N is a set of positive integers, updated at each Markov chain event transition via a cooling rate according to the following formula:

$$R_{t+1} = R_t * \nabla_t$$

Equation 6

Where R_t is the system temperature at time t, and ∇_t is the cooling rate.

6: An initial state assigned by the initial hill climb optimisation.

The simulated annealing panel optimisation then consists of a discrete-time Markov chain -representing a random walk - with the following transition process:

1: Retrieve system temperature R_t. From current temperature determine current value of cost function.

2: Randomly select panel biomarker i_X with threshold T_i.

3: Determine maximum threshold movement $m = \lceil \sqrt{R_t} \rceil$. (Greater system temperature allows larger movements).

4: Randomly select a movement from list: T_{i-m}: T_{i-1}, T_{i+1}: T_{i+m}.

5: Implement movement of biomarker threshold and recalculate search function score. Score is compared to previous value, if search function score is increased, movement is accepted, comprising a gradient ascent step. If search function score is not improved, movement may be accepted conditional to current value of cost function according to the probability:

$$P(accept) = e^{\left(\frac{c - c_{prev}}{R_t}\right)}$$

Equation 7

Where c is the current search function score, c_{prev} is the previous search function score, R_t is the current system temperature, to allow the algorithm to escape from local gradient ascent maxima.

6: If movement is accepted, update current panel, otherwise reject movement and return to previous panel.

7: Update system temperature according to the cooling rate.

The score according to the search function is stored at each event, and the simulated annealing optimisation is then allowed to run until the system temperature reaches a pre-determined lower limit, known as the exit temperature.

Additionally, a restart function has been included to reduce the likelihood of the simulated annealing algorithm becoming trapped in a sub-optimal local maxima, whereby if the search function score is not improved over a set number of Markov Chain events, the panel will revert to the last best panel found during the search, and the search will continue again from that panel, while maintaining the current system temperature.

9.3.2 Monte Carlo Random Search Method

To allow direct comparison of the Simulated annealing algorithm to a Monte Carlo direct search strategy, as used previously for the discovery of optimal biomarker panel thresholds, panel optimisation was undertaken using both techniques.

Cut-off thresholds were selected at random for each biomarker from a candidate pool of thresholds. These thresholds were then applied to training data to return a panel result as described in Equation 4, from which panel sensitivity and specificity was calculated. Finally the thresholds were applied to the hold-out test data.

The Monte Carlo direct search was restricted to ROC local maxima thresholds, as described earlier, derived from the training cohort, with the target thresholds limited to only those with individual biomarker specificity greater than 95%.

With the expectation that much larger Monte Carlo searches would be required to adequately cover the search space, searches were undertaken with 1000, 5000, and 10000 iterations respectively to allow for the estimation of a minimal run time required to return satisfactory threshold combinations.

9.3.3 Algorithm Benchmarking

A representative sample cohort has been assessed on the EarlyCDT®-Lung platform, and autoantibody binding values against a panel of seven tumour associated antigens, each at two concentrations, have been collected for a cohort of 335 lung cancer cases, and 330 matched normal controls, divided into training and test cohorts as described in Chapter 5. In addition to these 14 features, to assess the ability of the simulated annealing methodology to scale to larger panels of biomarkers, data from two investigatory panels of additional biomarkers have been included in this assessment, each comprised of seven autoantibody features measured at two concentrations. These have been used to give overall panels of 14, 28, and 42 autoantibody features respectively for assessing the ability of the simulated annealing algorithm to identify optimal panels and thresholds.

The cohort was divided into training (70% of the data) and test (30% of the data) cohorts, and training data underwent panel optimisation through both the simulated annealing algorithm described previously and using a Monte Carlo random search approach. A range of search sizes were undertaken for both methods, and the performance of the optimal panel discovered on both training and test cohorts, along with the time taken for the optimisation to run, have been summarised.

9.3.4 Implementation

All searches were undertaken in RStudio v2022.07.2, using R version 4.2.1. Searches were undertaken on a high-end desktop workstation to show the run times and performance levels capable without necessitating access to cloud computing or supercomputers. The testing scenario used a Microsoft Windows PC with Intel(R) Xeon(R) Quad-Core CPU at 3.70GHz and 32GB of RAM, as representative of the computing power available to lab-based researchers without needing to access web-based multi-thread or GPU processors.

9.3.5 Application of Simulated Annealing to Curve Characteristic Dataset

This simulated annealing algorithm was applied as described here to both the VCOD corrected and STVR corrected curve characteristic data sets explored previously in Chapter 7:, using the penalised classification search function described in section 7.6 as the search function, and iterating over 250 individual searches to find optimal panel cutoff thresholds for the commercial panel, the full panel of magnitude and curve characteristic features, and the features selected by Boruta analysis as described in section 7.5.

9.3.6 Application of Simulated Annealing to Extended Panel of Tumour Associated Autoantibodies

This simulated annealing algorithm was then applied to both the VCOD corrected and STVR corrected extended lung data sets explore previously in Chapter 8:, again using the penalised classification search function described in chapter 7.6 as the search function, and iterating over 250 individual searches to find optimal panel cutoff thresholds for the commercial panel, the full extended autoantibody panel, and the autoantibody features selected by Boruta analysis as described in chapter 8.5.

9.4 Results

9.4.1 Algorithm Benchmarking

The simulated annealing algorithm was able to return panels with higher performance characteristics than the Monte Carlo direct search method over all features, the highest performing panel, determined by ranking all panels by Training Youden Index, Training Specificity, Test Youden Index, and Test Specificity in that order, has been summarised in Table 9-1 and Table 9-2, showing that while panels optimised by simulated annealing were able to maintain Training specificities above 95%, even when optimising over 42 features, Monte Carlo derived panels showed reduced specificity reaching only 90.5% in the 14 biomarker panel, and reducing to 78.0% in the 42 biomarker panel, as the lack of any targeted optimisation resulted in accumulation of false positives from included features.

Table 9-1: Optimal panels as discovered by a Monte Carlo random search over all possible features

Panel Size	Iterations	Monte Carlo ‘Optimal’ Panel				Run Time
		Training		Test		
		Spec	Sens	Spec	Sens	

14	1000	86.2%	40.0%	73.5%	26.0%	105.175 secs
14	5000	88.8%	39.1%	76.5%	28.0%	562.143 secs
14	10000	90.5%	37.4%	77.6%	28.0%	1190.124 secs
28	1000	84.9%	43.4%	76.5%	40.0%	127.178 secs
28	5000	87.5%	42.1%	82.7%	33.0%	583.254 secs
28	10000	87.9%	42.1%	77.6%	34.0%	1267.126 secs
42	1000	74.1%	55.3%	62.2%	50.0%	143.283 secs
42	5000	72.0%	58.7%	61.2%	51.0%	651.261 secs
42	10000	78.0%	52.8%	67.3%	45.0%	1438.727 secs

Table 9-2: Optimal panels as discovered by simulated annealing algorithm optimisation over all possible features

Panel Size	Iterations	Simulated Annealing `Optimal` Panel				Run Time
		Training		Test		
		Spec	Sens	Spec	Sens	
14	250	97.4%	29.8%	93.9%	17.0%	3061.936 secs
14	500	97.4%	29.8%	93.9%	17.0%	7283.542 secs
14	1000	97.4%	29.8%	93.9%	17.0%	14936.549 secs
28	250	97.4%	34.0%	93.9%	21.0%	6735.586 secs
28	500	97.4%	34.0%	94.9%	20.0%	12253.133 secs
28	1000	97.4%	34.0%	93.9%	22.0%	25108.771 secs
42	250	95.3%	38.7%	91.8%	23.0%	9854.406 secs
42	500	95.3%	38.7%	91.8%	23.0%	18876.882 secs
42	1000	95.3%	38.7%	91.8%	23.0%	40486.890 secs

While Monte Carlo run times are lower in the scenario tested, the Monte Carlo search implemented explored only the full combination of biomarkers in each case, in a manner as described in Equation 1. To fully determine optimal panels using the Monte Carlo approach would require the exploration of all combinations of biomarkers at all possible panel sizes up to the number of biomarkers explored, as detailed in Equation 3, and would vastly increase the computational intensity of the task. Extrapolating from the run times returned, described in Table 9-1, is able to give an estimate based on a linear regression. Conservatively assuming 1,000 Monte Carlo iterations for each combination of biomarkers in a 14 biomarker panel, would require 16,383,000 iterations. On a system with comparable power to that used for the examples detailed here, it is estimated that such an analysis would require 3,581,846 seconds, or 41.5 days to complete.

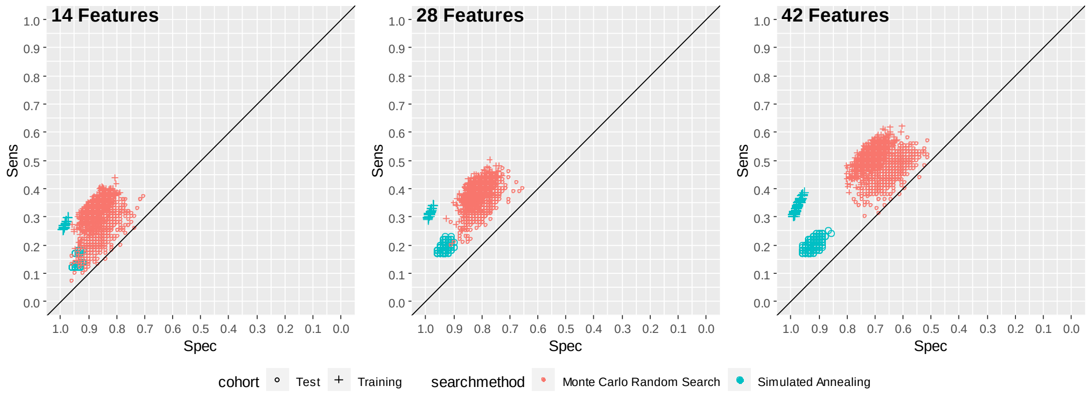


Figure 9-1: ROC scatter plots showing discovered panels after 1000 iterations of both Monte Carlo random search and Simulated Annealing panel optimisation on panels of 14, 28 and 42 autoantibody biomarkers

9.4.2 Curve Characteristic Panel Optimisation

Simulated annealing search of the curve characteristic data set explored in Chapter 7: gave panels as summarised in Figure 9-2 and Figure 9-3, showing again a propensity for the simulated annealing algorithm to overfit the training data. The best performing panel for each search – as defined by the greatest mean search function score over the 4 cohorts – has been summarised in Table 9-3 to Table 9-8.

VCOD Curve Characteristic Features

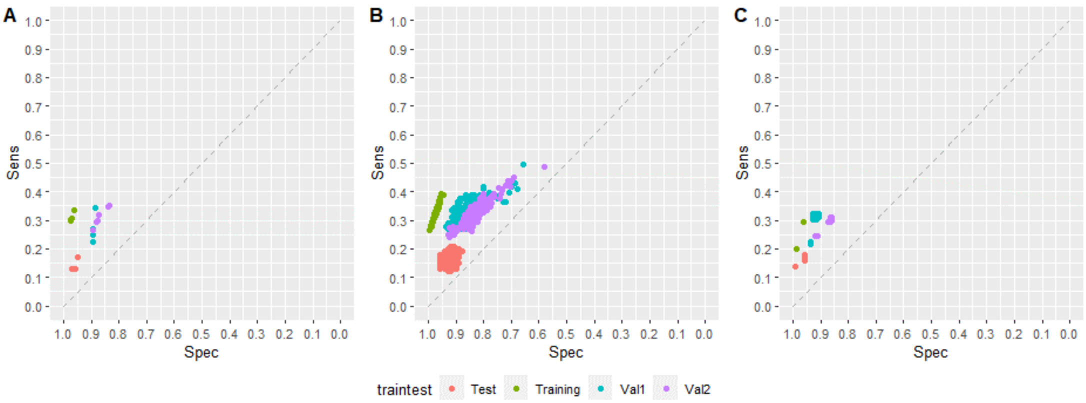


Figure 9-2: ROC Scatter summary of performance of simulated annealing derived optimal panel performance for VCOD corrected data on A) Commercial EarlyCDT®-Lung panel, B) Full feature set of magnitude and curve characteristic features, and C) Boruta selected magnitude and curve characteristic features.

Table 9-3: Best performing panel from simulated annealing search of Commercial EarlyCDT®-Lung panel autoantibody panel

cohort	TN	FP	FN	TP	Sens	Spec
Training	223	9	156	79	33.6%	96.1%

Test	93	5	83	17	17.0%	94.9%
Val1	85	11	61	32	34.4%	88.5%
Val2	259	50	136	72	34.6%	83.8%

Table 9-4: Best performing panel from simulated annealing search of full magnitude and curve characteristic feature set

cohort	TN	FP	FN	TP	Sens	Spec
Training	221	11	144	91	38.7%	95.3%
Test	90	8	79	21	21.0%	91.8%
Val1	84	12	58	35	37.6%	87.5%
Val2	249	60	129	79	38.0%	80.6%

Table 9-5: Best performing panel from simulated annealing search of Boruta selected magnitude and curve characteristic features

cohort	TN	FP	FN	TP	Sens	Spec
Training	229	3	188	47	20.0%	98.7%
Test	97	1	86	14	14.0%	99.0%
Val1	90	6	72	21	22.6%	93.8%
Val2	284	25	157	51	24.5%	91.9%

STVR Curve Characteristic Features

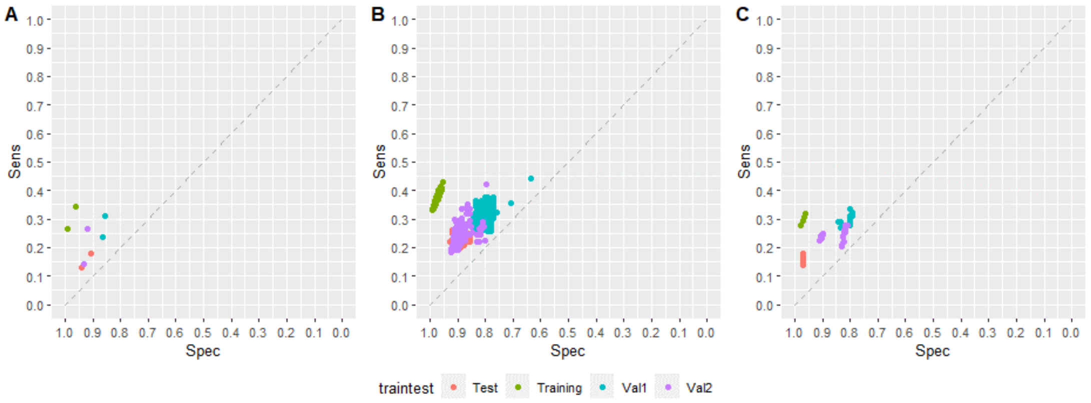


Figure 9-3: ROC Scatter summary of performance of simulated annealing derived optimal panel performance for STVR corrected data on A) Commercial EarlyCDT®-Lung panel, B) Full feature set of magnitude and curve characteristic features, and C) Boruta selected magnitude and curve characteristic features.

Table 9-6: Best performing panel from simulated annealing search of STVR corrected Commercial EarlyCDT®-Lung panel autoantibody panel

cohort	TN	FP	FN	TP	Sens	Spec
Training	223	9	154	81	34.5%	96.1%
Test	89	9	82	18	18.0%	90.8%
Val1	82	14	64	29	31.2%	85.4%
Val2	284	25	153	55	26.4%	91.9%

Table 9-7: Best performing panel from simulated annealing search of full magnitude and curve characteristic feature set

cohort	TN	FP	FN	TP	Sens	Spec
Training	223	9	142	93	39.6%	96.1%
Test	87	11	77	23	23.0%	88.8%
Val1	80	16	59	34	36.6%	83.3%
Val2	280	29	148	60	28.8%	90.6%

Table 9-8: Best performing panel from simulated annealing search of Boruta selected magnitude and curve characteristic features

cohort	TN	FP	FN	TP	Sens	Spec
Training	224	8	163	72	30.6%	96.6%
Test	95	3	83	17	17.0%	96.9%
Val1	81	15	66	27	29.0%	84.4%
Val2	279	30	157	51	24.5%	90.3%

9.4.3 Extended Lung Autoantibody Panel Optimisation

Simulated annealing search of the curve characteristic data set explored in Chapter 8:Chapter 7: gave panels as summarised in , showing again a propensity for the simulated annealing algorithm to overfit the training data. The best performing panel for each search – as defined by the greatest mean search function score over the 4 cohorts – has been summarised in Table 9-3 to Table 9-5.

VCOD Curve Characteristic Features

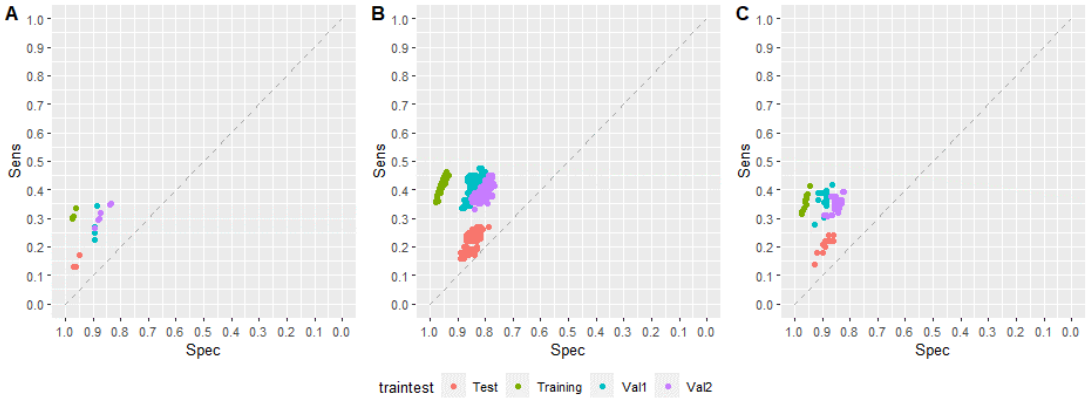


Figure 9-4: ROC Scatter summary of performance of simulated annealing derived optimal panel performance for VCOD corrected data on A) Commercial EarlyCDT®-Lung panel, B) Full extended panel of autoantibody features, and C) Boruta selected autoantibody features.

Table 9-9: Best performing panel from simulated annealing search of extended autoantibody panel

cohort	TN	FP	FN	TP	Sens	Spec
Training	222	10	136	99	42.1%	95.7%
Test	86	13	78	22	22.0%	86.9%
Val1	83	13	53	40	43.0%	86.5%
Val2	248	61	118	90	43.3%	80.3%

Table 9-10: Best performing panel from simulated annealing search of Boruta selected magnitude and curve characteristic features

cohort	TN	FP	FN	TP	Sens	Spec
Training	223	9	150	85	36.2%	96.1%
Test	91	8	82	18	18.0%	91.9%
Val1	88	8	57	36	38.7%	91.7%
Val2	264	45	133	75	36.1%	85.4%

STVR Curve Characteristic Features

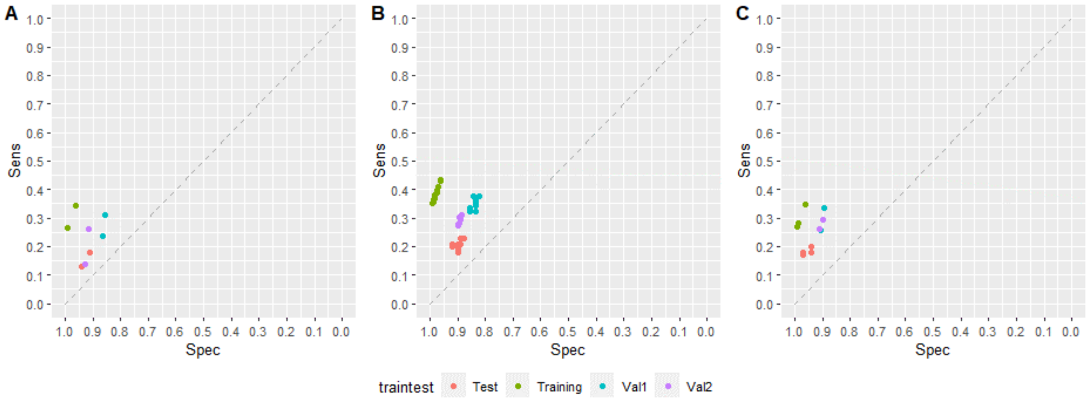


Figure 9-5: ROC Scatter summary of performance of simulated annealing derived optimal panel performance for STVR corrected data on A) Commercial EarlyCDT®-Lung panel, B) Full extended panel of autoantibody features, and C) Boruta selected autoantibody features.

Table 9-11: Best performing panel from simulated annealing search of extended autoantibody panel

cohort	TN	FP	FN	TP	Sens	Spec
Training	223	9	134	101	43.0%	96.1%
Test	88	11	77	23	23.0%	88.9%
Val1	81	15	58	35	37.6%	84.4%
Val2	275	34	144	64	30.8%	89.0%

Table 9-12: Best performing panel from simulated annealing search of Boruta selected magnitude and curve characteristic features

cohort	TN	FP	FN	TP	Sens	Spec
Training	223	9	153	82	34.9%	96.1%
Test	93	6	80	20	20.0%	93.9%
Val1	86	10	62	31	33.3%	89.6%
Val2	278	31	147	61	29.3%	90.0%

9.5 Chapter Conclusions

Detection of extremely heterogeneous diseases such as cancers pose a particular problem, as they may arise through a wide variety of different mechanisms and pathways, each of which may require different biomarkers for detection. A large panel of high specificity biomarkers represents one potential strategy for screen detecting disease at a clinically useful sensitivity, however optimising diagnostic or prognostic thresholds for large panels becomes extremely computationally intensive. The simulated annealing algorithm defined here represents an effective way of finding optimal combinations of thresholds for large panels of biomarkers, with internal feature selection allowing for the removal of redundant biomarkers during the random walk/gradient ascent Markov chain process.

9.5.1 Algorithm Benchmarking

The benchmarking data analysed here shows that it outperforms a Monte Carlo random search strategy in the identification of high specificity threshold combinations when a simplistic, all feature, Monte Carlo approach is applied, and that, even after 10,000 iterations, Monte Carlo random search did not identify the highest performing panels discovered during only 250 iterations of the simulated annealing algorithm.

Increasing the complexity of the Monte Carlo random search to evaluate all combinations of different sized panel, in an attempt to reduce the influence of less informative or redundant biomarkers vastly increases the computational intensity of the Monte Carlo strategy, and in the example outlined in this study, assessing all combinations for panel size of 14 biomarkers resulted in the need to evaluate 16383 combinations of biomarker, and would be estimated to take over 40 days to complete on a high-end desktop PC. While this may be computationally feasible with access to more powerful cloud based computing, applying the same strategy to the panel of 28 markers would result in 268,435,455 potential panel combinations to try and discover an optimal panel, and extending to the 42 biomarker panel would require assessment of 4,398,046,511,103 potential biomarker combinations, which demonstrates how the Monte Carlo random search technique that has been used successfully for smaller panels, becomes unsuitable once panel sizes start to increase.

Both simulated annealing and Monte Carlo strategies displayed a degree of overfitting in the analysed data. It is expected that the simulated annealing strategy will potentially have a higher degree of overfitting, however several strategies may be employed, dependant on the panels being explored, to reduce this effect. One advantage of the simulated annealing strategy as outlined here is the ability to return statistics on both the frequency with which the biomarkers are incorporated into optimal panels, along with the frequency of each potential threshold value. This could be used for feature ranking and subsequent feature selection, as the biomarkers with the greatest discriminatory ability would have a greater propensity to be included in optimal panels. This also allows for determination of the most commonly incorporated thresholds for each biomarker, and panels using the median threshold for each optimal biomarker may give panels with a lower degree of overfitting than any single optimised panel, at a cost to specificity and sensitivity in the training cohort.

9.5.2 Application of Simulated Annealing to Curve Characteristic Dataset

While overfitting is evident, the optimised panels discovered for the full feature set show better performance than the machine learning models explored previously, for example using VCOD corrected data, the XGBoost models trained in Chapter 7: gave sensitivities and specificities of 40.9% and 90.5%, 20.0% and 88.8%, 34.4% and 86.5%, and 38.9% and 81.5% for the training, test, validation 1 and validation 2 cohorts respectively when trained on all VCOD corrected features, using the same data the simulated annealing algorithm was able to return a higher specificity in the training set (95.3%) for only a small loss in sensitivity (38.7%), and showed increases in both sensitivity and specificity across all other cohorts.

9.5.3 Application of Simulated Annealing to Extended Panel of Tumour Associated Autoantibodies

Again a degree of overfitting is evident in the simulated annealing derived panels determined on the extended panel of autoantibodies, however the simulated annealing strategy was again able to outperform the machine learning strategies. The C5.0 tree model trained on the Boruta selected features or the VCOD corrected extended lung panel as described in Chapter 8: gave sensitivities and specificities of 29.4% and 96.1%, 16.0% and 94.9%, 30.1% and 92.7%, and 29.3% and 87.1% for the training, test, validation 1 and validation 2 cohorts respectively. In comparison, the highest performing panel based on the same features using the simulated annealing based algorithm gave sensitivities and specificities of 36.2% and 96.1% for the training cohort, an increase in sensitivity of 6.8%, 18.0% and 91.9% in the test cohort, a 2% improvement in the sensitivity with a corresponding loss of 3% in specificity, 38.7% and 91.7% in the validation 1 cohort, a sensitivity increase of 8.6% for a specificity loss of only 1.0%, and finally sensitivity of 36.1% and specificity of 85.4% in the validation 2 cohort, a sensitivity increase of 6.8% for a specificity loss of only 1.7%.

9.6 Chapter Discussion

The simulated annealing algorithm represents an improvement over Monte Carlo random search for the discovery of optimal sets of panel cutoff thresholds. Until a greater understanding can be reached of the limitations of the supervised machine learning strategies explored previously in this project, the logic test strategy of the current commercial EarlyCDT®-Lung continues to return the best commercial performance. Establishing the cutoff thresholds for this panel would previously require between hours and days dependant on the size of the panel, and return a vast number of non-optimal panel sets. The simulated annealing strategy is able to return only optimal panel threshold sets, in much shorter times, and also has the potential to incorporate bagging strategies to reduce overfitting.

Chapter 10: Discussion

This research project was initiated to determine whether it was possible to improve the diagnostic performance of the EarlyCDT®-Lung test for early detection of lung cancer, based on a re-analysis of data that was already generated by the EarlyCDT®-Lung commercial assay format to try and identify and remove false positive signal. Pilot study analysis showed a great deal of promise with regards to selectively identifying and removing false positive signal in the assay, resulting in increases to the test specificity which then allowed additional autoantibody biomarkers to be incorporated to further increase sensitivity while maintaining high panel specificity.

10.1 Health Economic Assessment

In order to determine the diagnostic performance required to prove health economic cost-effectiveness of the EarlyCDT®-Lung test for population screening of individuals at a high risk of lung cancer in an NHS setting, a health economic analysis was undertaken, based on that published by Snowsill et al(30), which showed that EarlyCDT®-Lung was more cost-effective than LDCT for screening, and suggested that at a specificity above 95%, and sensitivity above 47.5%, the EarlyCDT Test reached the NICE recommended threshold of £30,000 per QALY that is considered cost-effective. This £30,000 per QALY threshold was established in 2013, but is widely recognised as having no empirical foundation, and arguments have been made that the threshold should be as high as £50,000 per QALY for technologies that offer health benefits where the burden of disease is high, such as lung cancer(233).

A future refinement of this health economics model would be to improve the natural history model to include the length of time that a malignancy is present prior to the point where it is detectable by CT, which would allow for a more accurate appreciation of the costs of enhanced surveillance that would be attributable to the autoantibody test when detecting malignancies in their earliest stages, as well as allowing more accurate calculation of the additional malignancies that would be discoverable at their earliest stages thanks to the autoantibody test. Along with these refinements, a more extensive search of potential screening strategies would then also be undertaken, including rescreening at intervals such as 3 and 5 years to allow detection of malignancies that develop in the interim time between screenings.

10.2 Machine Learning Analysis

A range of machine learning strategies were then explored to determine whether they had potential utility in improving the diagnostic performance of the EarlyCDT®-Lung autoantibody screening test. The methods explored initially focused on white-box techniques, where the calculations underlying the diagnostic decision can be extracted and scrutinized. Due to the transparency of the calculations, white-box models are more readily accepted by regulatory bodies than black-box models(234). The black-box techniques of random forest and extreme gradient boosted trees were included in the exploration because of their recent use in disease prediction models(175, 176, 181, 183), and increased adoption and regulatory acceptance of black-box models including neural networks and deep learning techniques, symbolised by the publication of the HMA-EMA Joint Big Data

Taskforce summary report in 2019(235) suggests that more sophisticated machine learning and deep learning strategies are becoming more acceptable, with AI based diagnostic models now being approved by the FDA, such as the deep-learning based IDx-DR for diagnosis of diabetic retinopathy(236).

The results of these machine learning analyses also suggested that, in spite of assessing an additional 12 autoantibody features, little additional sensitivity was evident from the additional markers. This may be due to a number of causes, firstly the case cohort were confirmed cancer samples, which means that the tumours are already at a stage where they are large enough to be found by CT screening. From the examination of longitudinal samples in Chapter 3 it was evident that autoantibody responses can precede CT presentation by an average of 4 years, and work by Bruno et al. showed that with constant epitope presentation, B cells can become exhausted, and exhausted tumour infiltrating B cells become less functional, expressing less antibody. This exhausted B cell phenotype may also attenuate T-cell responses and dampen antitumor response(237) which may result in a diminished immune response. Additionally, these cohorts were a mixture of different cancer subtypes, and the heterogeneous nature of cancer may necessitate that much larger cohorts are examined, with specific autoantibody panels and models dedicated to various histological subtypes.

10.3 Future Studies

In order to assess the machine learning strategies more accurately would require a large-scale prospective screening trial with longitudinal collection to identify autoantibody responses at their earliest point, with longitudinal assessment allowing greater investigation into the stability and behaviour of autoantibody responses, including whether autoantibody responses become depleted over time, or whether malignancies reach a stage where they are able to escape immunosurveillance and the corresponding effect this would have on related autoantibody profiles. The greater adoption of LDCT screening would also allow enhanced surveillance in subjects that show autoantibody reactivity, allowing identification of cancer at its earliest stage. Ideally such a study would also explore additional detection strategies in parallel to autoantibody and LDCT screening, and include a reinforcement learning strategy to refine predictive models during the course of the study as new data becomes available and would follow patients for a minimum of ten years to ensure that all malignancies present during the study are identified. Such a study would require large numbers of subjects, with around 7,750 high risk (2% or greater 5-year risk) individuals required to identify 155 cancer cases in order to give the study statistical power (based on example sensitivity of 40% and the use of exact methods(238)).

While EarlyCDT®-Lung has undergone a large prospective study in the form of the Early Detection of Cancer of the Lung Scotland (ECLS) study(239), which has recruited over 12,000 subjects and aims to follow up all subjects for 10 years, the ECLS study was designed to confirm diagnostic stage shift from late to early stage, and the data has limited use for model training, as only half of the study cohort comprised the intervention group and collected EarlyCDT results, and while the study recruited subjects at high risk, the incidence was lower than expected, limiting the number of positive cases – at 2 year follow up there were only 56 confirmed cancers in the intervention arm, although it is expected that additional cancers will present over the remaining follow-up years, and given the study has already demonstrated stage-shift to earlier diagnosis, these additional cancers would be expected to present at a higher rate in the test positive subjects in the intervention arm, potentially increasing the appreciable specificity of the test, returned by the study. In addition, the study involved only a single EarlyCDT®-Lung test, with LDCT follow-up on only the positive subjects, this lack of detailed follow up on the test-negative arm limits the utility of this dataset for training new models.

With the greater availability and acceptance of deep learning strategies for medical diagnostics, the future development of predictive cancer screening tests will likely benefit greatly from inclusion of multiple ‘omics’ measurements, such as protein biomarkers, methylation signature, cell free DNA, and CT imaging data, along with autoantibody biomarkers and demographic risk factors to create ensemble or deep learning models which have a greater ability to detect cancers in spite of their heterogeneous development and presentation. In such a test, autoantibodies would likely represent the best opportunity to identify cancer at its earliest stage. The combination of multiple testing modalities also raises the potential for greater understanding of underlying cancer phenotypes, and the use of such a test for companion diagnostics and treatment stratification could lead to new avenues in personalised medicine.

These tests are already in development(240), and autoantibodies represent a potential key contributor in these future tests. One potential future source of data is the currently running iDx Lung trial(241), which is recruiting patients undergoing an NHS Targeted Lung Health Check and is collecting CT scan data alongside blood biomarker tests including the EarlyCDT®-Lung test, as well as protein biomarker panel, a circulating tumour DNA test, and an RNA-Seq analysis. While these testing modalities are being assessed separately, the combination of the data could represent an avenue for the exploration of multi-omics testing for lung cancer detection.

In addition to the inclusion of further test modalities, a testing strategy that includes repeated measurements at set intervals may allow for the development of personalised diagnostic baselines, with deviation from baseline potentially revealing disease presence more accurately than a cross-sectional test. This strategy has been explored for protein biomarker testing for ovarian cancer previously(242) and some initial work was undertaken to determine whether such an approach would be beneficial for autoantibodies in lung cancer (147).

10.4 Simulated Annealing Algorithm

In tandem to the machine learning experiments, refinement of the previous Monte Carlo random search strategy was undertaken with the development of a simulate annealing based multivariate panel optimisation algorithm, with the aim of discovering optimal sets of panel cutoff thresholds without searching a large amount of suboptimal feature space. This proved again to give comparable performance to the machine learning strategies, and this algorithm could be further refined with the inclusion of cross-validation to reduce overfitting, and the potential use of model bagging(243) of multiple high specificity overfit models to give a consensus output.

10.5 Bias

While analysing the various study results, a number of limitations of this programme of research, and potential sources of bias, became evident. The focus of these analyses has been the diagnostic performance as summarised by sensitivity and specificity measurements. These metrics were used as they are more widely understood and easier to interpret, as well as allowing comparison to other biomarker tests in the literature. These measures do have associated drawbacks, as described previously(244) which include that they are influenced by the population under investigation, and so can be affected by the composition of the sample cohort, which was evident from the particularly low sensitivity of the test cohort. Use of these metrics did however allow for comparison of high specificity diagnostic cut-off thresholds applied to multiple cohorts, which would not have been possible if reporting a metric such as the area under the ROC curve, and from the outcomes of the health economic sensitivity analysis in Chapter 4, we were aware that high specificity was a major driver of cost-effectiveness and therefore was a focus in the modelling strategies explored.

The training, test, and validation 1 datasets explored in the machine learning analyses were matched for age, sex, and smoking history, however this matching was not perfect with regards gender, and data availability on smoking intensity and duration was not available for a majority of the samples examined, therefore a cancer risk score was not able to be calculated for the subjects, and these variables were not considered for matching. As smoking duration and intensity are both contributors to pre-existing cancer risk, this may have resulted in risk imbalance between the case and control cohorts, and also between the training, test, and validation cohorts, which may have impacted the model fitting.

These datasets also showed different distributions of both subtype and staging which may have influenced how well model performance transferred to the validation cohorts, with the training/test cohort showing an even split of adenocarcinoma and squamous, with SCLC comprising 11% of the cohort, the Validation 1 cohort shows a majority of samples with Squamous cell carcinoma, and SCLC is under-represented in this dataset at only 4.5%, which could result in reduced overall sensitivity, contrary to this, the Validation 2 cohort shows a majority of samples with Adenocarcinoma, and 31% SCLC, which is a far larger proportion than in the training data. Differences in staging are also large between the datasets, with the training/test cohort having predominantly early stage (82%), while the Validation 1 cohort shows a slightly lower proportion of early stage (69%), and the Validation 2 contains a majority of late stage (55%), which could have a major influence on the model performance. Future studies would benefit from ensuring a wider range of demographic variables were obtained on all samples, and a relevant demographic cancer risk model, such as the PLCOM$_{2012}$, applied to all samples so pre-test risk could also be used in the calculation of a propensity score for case-control matching, and also as a partitioning criterion during the splitting of the data into training and test cohorts to ensure balanced datasets for modelling. Appreciation of cancer subtype and staging should also be considered during study design and data partitioning to ensure that datasets are balanced and representative of the test target population.

The control cohort in these studies also harbours two potential sources of verification bias, as it was not possible to verify the true disease status of all examined control samples, and the assumption was made that they were all disease free. However, given that the cohort was intended to be high risk, it is likely that a proportion of latent cancers were harboured within the normal cohort, especially as we have shown evidence in Chapter 3 that autoantibodies can be detectably elevated several years before clinical presentation. Assuming a 2% 5-year risk in this cohort with a conservative 35% sensitivity estimate would lead to such a latent cancer population reducing the specificity of a predictive model by about 0.7%, for this reason future studies would benefit from follow-up of the normal population with cancer registries for several years after testing, such as has been the case in the ECLS study(245), to ensure the most accurate estimates of test specificity and sensitivity. In addition to this, several of the autoantibody markers explored have shown sensitivity in multiple cancer types, this has minimal impact when exploring a case-control study, as the incidence of these cancers in the control cohort would be extremely low, and would contribute only a small amount to the false-positive results, if at all.

While the commercial EarlyCDT®-Lung test uses on plate controls and between run calibration to compensate for inter-assay variability, the calibration is designed only for the top two antigen concentrations of the commercial 7 autoantibody panel, therefore could not be used for reducing inter-assay variability in curve characteristic features or the expanded panel of autoantibodies. As the assays were run over 26 days total, and without control for assay operators. This may have introduced additional variability to the study results and reduced the effectiveness of some of the modelling analyses.

10.6 Summary

Early detection of lung cancer represents the best avenue for improving prognosis and increasing survival, and the EarlyCDT®-Lung panel of tumour associated autoantibodies is able to facilitate this detection up to several years prior to the current best practise of LDCT screening, as well as showing favourable health economics in comparison to LDCT screening. Unfortunately the sensitivity of autoantibody screening is currently limited, due to either the heterogeneity of lung cancer, or other aspects of the biological immune response to cancer, and investigations into machine learning strategies in an effort to overcome these limitations was unable to train models which performed consistently better than the current EarlyCDT®-Lung commercial test, although this may have been exacerbated by imbalance and confounding variables in the explored datasets. Large scale longitudinal studies using balanced datasets with collection of demographic risk factors and appreciation of confounding variables are needed to develop a more full understanding of the immune response to lung cancer, and the addition of other modalities of biomarker such as antigenic, DNA and methylation based markers would likely be the best strategy for developing a more accurate early detection test.

Chapter 11: References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: A Cancer Journal for Clinicians. 2021;71(3):209-49.
2. Torre LA, Freddie B, Siegel RL, Jacques F, Joannie L-T, Ahmedin J. Global cancer statistics, 2012. CA Cancer J Clin. 2015;65(2):87-108.
3. Observatory TGC. 2021 [Available from: https://gco.iarc.fr/today/data/factsheets/populations/826-united-kingdom-fact-sheets.pdf.
4. Hilhorst S, Lockey A. Cancer costs: a'ripple effect'analysis of cancer’s wider impact. 2020.
5. Nambiar M, Mridula N, Vijayalakshmi K, Raghavan SC. Chromosomal translocations in cancer. Biochimica et Biophysica Acta (BBA) - Reviews on Cancer. 2008;1786(2):139-52.
6. Foulds L. The experimental study of tumor progression: a review. Cancer Res. 1954;14(5):327-39.
7. Hanahan D, Douglas H, Weinberg RA. The Hallmarks of Cancer. Cell. 2000;100(1):57-70.
8. Witsch E, Sela M, Yarden Y. Roles for growth factors in cancer progression. Physiology. 2010;25(2):85-101.
9. Nevins JR. The Rb/E2F pathway and cancer. Hum Mol Genet. 2001;10(7):699-703.
10. Greenblatt MS, Bennett WP, Hollstein M, Harris CC. Mutations in the p53 tumor suppressor gene: clues to cancer etiology and molecular pathogenesis. Cancer Res. 1994;54(18):4855-78.
11. Harris CC. COMMENTARY: p53 Tumor suppressor gene: from the basic research laboratory to the clinic—an abridged historical perspective. Carcinogenesis. 1996.
12. Shay JW, Wright WE. Role of telomeres and telomerase in cancer. Semin Cancer Biol. 2011;21(6):349-53.
13. Nishida N, Yano H, Nishida T, Kamura T, Kojiro M. Angiogenesis in cancer. Vasc Health Risk Manag. 2006;2(3):213-9.
14. Bendas G, Borsig L. Cancer Cell Adhesion and Metastasis: Selectins, Integrins, and the Inhibitory Potential of Heparins. Int J Cell Biol. 2012;2012.
15. Okegawa T, Pong R-C, Li Y, Hsieh J-T. The role of cell adhesion molecule in cancer progression and its application in cancer therapy. Acta Biochim Pol. 2004;51(2):445-57.
16. Hanahan D, Douglas H, Weinberg RA. Hallmarks of Cancer: The Next Generation. Cell. 2011;144(5):646-74.
17. Cavallo F, Federica C, De Giovanni C, Patrizia N, Guido F, Pier-Luigi L. 2011: the immune hallmarks of cancer. Cancer Immunol Immunother. 2011;60(3):319-26.
18. Burnet M. Cancer–A Biological Approach: I. The Processes Of Control. II. The Significance of Somatic Mutation. BMJ. 1957;1(5022):779-86.
19. Dunn GP, Old LJ, Schreiber RD. The immunobiology of cancer immunosurveillance and immunoediting. Immunity. 2004;21(2):137-48.
20. Vesely MD, Kershaw MH, Schreiber RD, Smyth MJ. Natural innate and adaptive immunity to cancer. Annu Rev Immunol. 2011;29:235-71.
21. Warrington R, Watson W, Kim HL, Antonetti FR. An introduction to immunology and immunopathology. Allergy Asthma Clin Immunol. 2011;7(1):1-8.
22. Alberts B, Johnson A, Lewis J, Morgan D, Raff M, Roberts K, et al. Molecular Biology of the Cell, Sixth Edition: Garland Science; 2014 2014/11//.

23. de Visser KE, Eichten A, Coussens LM. Paradoxical roles of the immune system during cancer development. Nat Rev Cancer. 2006;6(1):24-37.
24. Detterbeck FC, Gibson CJ. Turning gray: the natural history of lung cancer over time. Journal of Thoracic Oncology. 2008;3(7):781-92.
25. Spratt Jr JS, Spratt TL. Rates of growth of pulmonary metastases and host survival. Annals of Surgery. 1964;159(2):161.
26. Lindell RM, Hartman TE, Swensen SJ, Jett JR, Midthun DE, Tazelaar HD, et al. Five-year lung cancer screening experience: CT appearance, growth rate, location, and histologic features of 61 lung cancers. Radiology. 2007;242(2):555-62.
27. Norton L. A Gompertzian model of human breast cancer growth. Cancer research. 1988;48(24_Part_1):7067-71.
28. Kanashiki M, Tomizawa T, Yamaguchi I, Kurishima K, Hizawa N, Ishikawa H, et al. Volume doubling time of lung cancers detected in a chest radiograph mass screening program: Comparison with CT screening. Oncol Lett. 2012;4(3):513-6.
29. Fischer BM, Olsen MW, Ley CD, Klausen TL, Mortensen J, Højgaard L, et al. How few cancer cells can be detected by positron emission tomography? A frequent question addressed by an in vitro study. Eur J Nucl Med Mol Imaging. 2006;33(6):697-702.
30. Snowsill T, Yang H, Griffin E, Long L, Varley-Campbell J, Coelho H, et al. Low-dose computed tomography for lung cancer screening in high-risk populations: a systematic review and economic evaluation. Health Technol Assess. 2018;22(69):1-276.
31. Van Meerbeeck JP, Fennell DA, De Ruysscher DK. Small-cell lung cancer. The Lancet. 2011;378(9804):1741-55.
32. England PH. National Cancer Intelligence Network: Cancer survival in England by stage. 2014.
33. Traut HF, Papanicolaou GN. Cancer of the Uterus: The Vaginal Smear in Its Diagnosis. Cal West Med. 1943;59(2):121-2.
34. Tabar L, Yen M-F, Vitak B, Chen H-HT, Smith RA, Duffy SW. Mammography service screening and mortality in breast cancer patients: 20-year follow-up before and after introduction of screening. Lancet. 2003;361(9367):1405-10.
35. Logan RFA, Patnick J, Nickerson C, Coleman L, Rutter MD, von Wagner C, et al. Outcomes of the Bowel Cancer Screening Programme (BCSP) in England after the first 1 million tests. Gut. 2012;61(10):1439-46.
36. Siegel RL, Miller KD, Ahmedin J. Cancer statistics, 2016. CA Cancer J Clin. 2016;66(1):7-30.
37. Flehinger BJ, Melamed MR, Zaman MB, others. Early Lung Cancer Detection: Results of the Initial (Prevalence) Radiologic and Cytologic Screening in the Memorial Sloan-Kettering Study 1–3. American Review of. 1984.
38. Robert SF, et al. Early Lung Cancer Detection: Results of the Initial (Prevalence) Radiologic and Cytologic Screening in the Mayo Clinic Study. Am Rev Respir Dis. 1984;130(4):561-5.
39. Kubík AK, Maxwell Parkin D, Petr Z. Czech study on lung cancer screening. Cancer. 2000;89(S11):2363-8.
40. Frost M.D. et J. Early Lung Cancer Detection: Results of the Initial (Prevalence) Radiologic and Cytologic Screening in The Johns Hopkins Study1–3. Am Rev Respir Dis. 1984;130(4):549-54.
41. Bach PB, Kelley MJ, Tate RC, McCrory DC. Screening for lung cancer: a review of the current literature. Chest. 2003;123(1 Suppl):72S-82S.
42. National Lung Screening Trial Research T, Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med. 2011;365(5):395-409.
43. National Lung Screening Trial Research T, Church TR, Black WC, Aberle DR, Berg CD, Clingan KL, et al. Results of initial low-dose computed tomographic screening for lung cancer. N Engl J Med. 2013;368(21):1980-91.

44. Weissleder R, Nahrendorf M. Advancing biomedical imaging. Proc Natl Acad Sci U S A. 2015;112(47):14424-8.

45. (ONS) OfNS. Deprivation and the impact on smoking prevalence, England and Wales: 2017 to 2021. ONS website. 2023.

46. Miekisch W, Schubert JK, Noeldge-Schomburg GFE. Diagnostic potential of breath analysis–focus on volatile organic compounds. Clin Chim Acta. 2004;347(1-2):25-39.

47. Gordon SM, Szidon JP, Krotoszynski BK, Gibbons RD, O'Neill HJ. Volatile organic compounds in exhaled air from patients with lung cancer. Clin Chem. 1985;31(8):1278-82.

48. Saalberg Y, Wolff M. VOC breath biomarkers in lung cancer. Clin Chim Acta. 2016;459:5-9.

49. Mazzone PJ. Analysis of volatile organic compounds in the exhaled breath for the diagnosis of lung cancer. J Thorac Oncol. 2008;3(7):774-80.

50. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 1995;270(5235):467-70.

51. Ramaswamy S, Golub TR. DNA microarrays in clinical oncology. J Clin Oncol. 2002;20(7):1932-41.

52. Kim H. Role of microarray in cancer diagnosis. Cancer Res Treat. 2004;36(1):1-3.

53. Chen R, Khatri P, Mazur PK, Polin M, Zheng Y, Vaka D, et al. A meta-analysis of lung cancer gene expression identifies PTK7 as a survival gene in lung adenocarcinoma. Cancer Res. 2014;74(10):2892-902.

54. Leon SA, Shapiro B, Sklaroff DM, Yaros MJ. Free DNA in the serum of cancer patients and the effect of therapy. Cancer Res. 1977;37(3):646-50.

55. Sidransky D. Emerging molecular markers of cancer. Nat Rev Cancer. 2002;2(3):210-9.

56. Newman AM, Bratman SV, To J, Wynne JF, Eclov NCW, Modlin LA, et al. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. Nat Med. 2014;20(5):548-54.

57. Nygaard AD, Garm Spindler K-L, Pallisgaard N, Andersen RF, Jakobsen A. The prognostic value of KRAS mutated plasma DNA in advanced non-small cell lung cancer. Lung Cancer. 2013;79(3):312-7.

58. Sozzi G, Conte D, Leon M, Ciricione R, Roz L, Ratcliffe C, et al. Quantification of free circulating DNA as a diagnostic marker in lung cancer. J Clin Oncol. 2003;21(21):3902-8.

59. Bean J, Brennan C, Shih J-Y, Riely G, Viale A, Wang L, et al. MET amplification occurs with or without T790M mutations in EGFR mutant lung tumors with acquired resistance to gefitinib or erlotinib. Proc Natl Acad Sci U S A. 2007;104(52):20932-7.

60. Qin Z, Ljubimov VA, Zhou C, Tong Y, Liang J. Cell-free circulating tumor DNA in cancer. Chin J Cancer. 2016;35:36.

61. Li Y, Fan Z, Meng Y, Liu S, Zhan H. Blood-based DNA methylation signatures in cancer: A systematic review. Biochim Biophys Acta Mol Basis Dis. 2023;1869(1):166583.

62. Liu MC, Oxnard GR, Klein EA, Swanton C, Seiden MV. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. Ann Oncol. 2020;31(6):745-59.

63. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell. 2004;116(2):281-97.

64. Croce CM. Causes and consequences of microRNA dysregulation in cancer. Nat Rev Genet. 2009;10(10):704-14.

65. Lawrie CH, Gal S, Dunlop HM, Pushkaran B, Liggins AP, Pulford K, et al. Detection of elevated levels of tumour-associated microRNAs in serum of patients with diffuse large B-cell lymphoma. Br J Haematol. 2008;141(5):672-5.
66. Mitchell PS, Parkin RK, Kroh EM, Fritz BR, Wyman SK, Pogosova-Agadjanyan EL, et al. Circulating microRNAs as stable blood-based markers for cancer detection. Proc Natl Acad Sci U S A. 2008;105(30):10513-8.
67. Chen X, Ba Y, Ma L, Cai X, Yin Y, Wang K, et al. Characterization of microRNAs in serum: a novel class of biomarkers for diagnosis of cancer and other diseases. Cell Res. 2008;18(10):997-1006.
68. Kosaka N, Iguchi H, Ochiya T. Circulating microRNA in body fluid: a new potential biomarker for cancer diagnosis and prognosis. Cancer Sci. 2010;101(10):2087-92.
69. Gold P, Freedman SO. Tests for carcinoembryonic antigen. Role in diagnosis and management of cancer. JAMA. 1975;234(2):190-2.
70. Catalona WJ, Smith DS, Ratliff TL, Dodds KM, Coplen DE, Yuan JJ, et al. Measurement of prostate-specific antigen in serum as a screening test for prostate cancer. N Engl J Med. 1991;324(17):1156-61.
71. Jacobs I, Bast RC. The CA 125 tumour-associated antigen: a review of the literature. Hum Reprod. 1989.
72. Chen DS, Sung JL. Serum alphafetoprotein in hepatocellular carcinoma. Cancer. 1977;40(2):779-83.
73. Hariu H, Hirohashi Y, Torigoe T, Asanuma H, Hariu M, Tamura Y, et al. Aberrant expression and potency as a cancer immunotherapy target of inhibitor of apoptosis protein family, Livin/ML-IAP in lung cancer. Clin Cancer Res. 2005;11(3):1000-9.
74. Croce MV, Isla-Larrain MT, Demichelis SO, Gori JR, Price MR, Segal-Eiras A. Tissue and serum MUC1 mucin detection in breast cancer patients. Breast Cancer Res Treat. 2003;81(3):195-207.
75. Chen YT, Scanlan MJ, Sahin U, Türeci O, Gure AO, Tsang S, et al. A testicular antigen aberrantly expressed in human cancers detected by autologous antibody screening. Proc Natl Acad Sci U S A. 1997;94(5):1914-8.
76. Attallah AM, Abdel-Aziz MM, El-Sayed AM, Tabll AA. Detection of serum p53 protein in patients with different gastrointestinal cancers. Cancer Detect Prev. 2003;27(2):127-31.
77. Tan EM. Autoantibodies as reporters identifying aberrant cellular mechanisms in tumorigenesis. J Clin Invest. 2001;108(10):1411-5.
78. Nesterova M, Johnson N, Cheadle C, Cho-Chung YS. Autoantibody biomarker opens a new gateway for cancer diagnosis. Biochim Biophys Acta. 2006;1762(4):398-403.
79. Boyle P, Chapman CJ, Holdenrieder S, Murray A, Robertson C, Wood WC, et al. Clinical validation of an autoantibody test for lung cancer. Ann Oncol. 2011;22(2):383-9.
80. Etzioni R, Kooperberg C, Pepe M, Smith R, Gann PH. Combining biomarkers to detect disease with application to prostate cancer. Biostatistics. 2003;4(4):523-38.
81. Hosmer DW, Lemeshow S. Applied Logistic Regression: Wiley; 2004 2004.
82. Zhong L, Coe SP, Stromberg AJ, Khattar NH, Jett JR, Hirschowitz EA. Profiling tumor-associated antibodies for early detection of non-small cell lung cancer. J Thorac Oncol. 2006;1(6):513-9.
83. Rom WN, Goldberg JD, Addrizzo-Harris D, Watson HN, Khilkin M, Greenberg AK, et al. Identification of an autoantibody panel to separate lung cancer from smokers and nonsmokers. BMC Cancer. 2010;10:234.
84. Gao W-M, Kuick R, Orchekowski RP, Misek DE, Qiu J, Greenberg AK, et al. Distinctive serum protein profiles involving abundant proteins in lung cancer patients based upon antibody microarray analysis. BMC Cancer. 2005;5:110.

85. Mamtani MR, Thakre TP, Kalkonde MY, Amin MA, Kalkonde YV, Amin AP, et al. A simple method to combine multiple molecular biomarkers for dichotomous diagnostic classification. BMC Bioinformatics. 2006;7(1):442.
86. Patz EF, Jr., Campa MJ, Gottlin EB, Kusmartseva I, Guan XR, Herndon JE, 2nd. Panel of serum biomarkers for the diagnosis of lung cancer. J Clin Oncol. 2007;25(35):5578-83.
87. Ingoldsby H, Webber M, Wall D, Scarrott C, Newell J, Callagy G. Prediction of Oncotype DX and TAILORx risk categories using histopathological and immunohistochemical markers by classification and regression tree (CART) analysis. Breast. 2013;22(5):879-86.
88. Borgia JA, Basu S, Faber LP, Kim AW, Coon JS, Kaiser-Walters KA, et al. Establishment of a multi-analyte serum biomarker panel to identify lymph node metastases in non-small cell lung cancer. J Thorac Oncol. 2009;4(3):338-47.
89. Farlow EC, Patel K, Basu S, Lee B-S, Kim AW, Coon JS, et al. Development of a multiplexed tumor-associated autoantibody-based blood test for the detection of non-small cell lung cancer. Clin Cancer Res. 2010;16(13):3452-62.
90. Ostroff RM, Bigbee WL, Franklin W, Gold L, Mehan M, Miller YE, et al. Unlocking Biomarker Discovery: Large Scale Application of Aptamer Proteomic Technology for Early Detection of Lung Cancer. PLoS One. 2010;5(12):e15003.
91. Pineda AL, Ogoe HA, Balasubramanian JB, Rangel Escareño C, Visweswaran S, Herman JG, et al. On Predicting lung cancer subtypes using 'omic' data from tumor and tumor-adjacent histologically-normal tissue. BMC Cancer. 2016;16:184.
92. Schneider J, Bitterlich N, Velcovsky H-G, Morr H, Katz N, Eigenbrodt E. Fuzzy logic-based tumor-marker profiles improved sensitivity in the diagnosis of lung cancer. Int J Clin Oncol. 2002;7(3):145-51.
93. Wu Y, Wu Y, Wang J, Yan Z, Qu L, Xiang B, et al. An optimal tumor marker group-coupled artificial neural network for diagnosis of lung cancer. Expert Syst Appl. 2011;38(9):11329-34.
94. O'Shea K, Cameron SJS, Lewis KE, Lu C, Mur LAJ. Metabolomic-based biomarker discovery for non-invasive lung cancer screening: A case study. Biochim Biophys Acta. 2016;1860(11 Pt B):2682-7.
95. Leidinger P, Keller A, Heisel S, Ludwig N, Rheinheimer S, Klein V, et al. Identification of lung cancer with high sensitivity and specificity by blood testing. Respir Res. 2010;11:18.
96. Au N, Cheang M, Huntsman DG, Yorida E, Coldman A, Elliott WM, et al. Evaluation of immunohistochemical markers in non-small cell lung cancer by unsupervised hierarchical clustering analysis: a tissue microarray study of 284 cases and 18 markers. J Pathol. 2004;204(1):101-9.
97. Lung Screening Trial Research Team N. The national lung screening trial: overview and study design1. Radiology. 2011.
98. van den Bergh KA, Essink-Bot ML, Borsboom GJ, Scholten ET, van Klaveren RJ, de Koning HJ. Long-term effects of lung cancer computed tomography screening on health-related quality of life: the NELSON trial. Eur Respir J. 2011;38:154-61.
99. Oken MM, Hocking WG, Kvale PA, Andriole GL, Buys SS, Church TR, et al. Screening by chest radiograph and lung cancer mortality: the Prostate, Lung, Colorectal, and Ovarian (PLCO) randomized trial. JAMA. 2011;306(17):1865-73.
100. Katki HA, Kovalchik SA, Petito LC, Cheung LC, Jacobs E, Jemal A, et al. Implications of nine risk prediction models for selecting ever-smokers for computed tomography lung cancer screening. Annals of internal medicine. 2018;169(1):10-9.
101. Tammemägi MC, Katki HA, Hocking WG, Church TR, Caporaso N, Kvale PA, et al. Selection criteria for lung-cancer screening. New England Journal of Medicine. 2013;368(8):728-36.

102. Marcus MW, Chen Y, Raji OY, Duffy SW, Field JK. LLPi: Liverpool Lung Project Risk Prediction Model for Lung Cancer Incidence. Cancer Prev Res (Phila). 2015;8(6):570-5.

103. Katki HA, Kovalchik SA, Berg CD, Cheung LC, Chaturvedi AK. Development and validation of risk models to select ever-smokers for CT lung cancer screening. JAMA. 2016;315(21):2300-11.

104. Robbins HA, Alcala K, Swerdlow AJ, Schoemaker MJ, Wareham N, Travis RC, et al. Comparative performance of lung cancer risk models to define lung screening eligibility in the United Kingdom. Br J Cancer. 2021;124(12):2026-34.

105. Cassidy A, Myles JP, Van Tongeren M, Page RD, Liloglou T, Duffy SW, et al. The LLP risk model: an individual risk prediction model for lung cancer. British Journal of Cancer. 2008;98(2):270-6.

106. Nicholson BD, Oke J, Virdee PS, Harris DA, O'Doherty C, Park JES, et al. Multi-cancer early detection test in symptomatic patients referred for cancer investigation in England and Wales (SYMPLIFY): a large-scale, observational cohort study. The Lancet Oncology. 2023;24(7):733-43.

107. Lam S, Boyle P, Healey GF, Maddison P, Peek L, Murray A, et al. EarlyCDT-Lung: an immunobiomarker test as an aid to early detection of lung cancer. Cancer Prev Res. 2011;4(7):1126-34.

108. Chapman CJ, Healey GF, Murray A, Boyle P, Robertson C, Peek LJ, et al. EarlyCDT®-Lung test: improved clinical utility through additional autoantibody assays. Tumor Biology. 2012;33(5):1319-26.

109. Sullivan FM, Mair FS, Anderson W, Armory P, Briggs A, Chew C, et al. Earlier diagnosis of lung cancer in a randomised trial of an autoantibody blood test followed by imaging. European Respiratory Journal. 2020;57(1):2000670.

110. Healey GF, Lam S, Boyle P, Hamilton-Fairley G, Peek LJ, Robertson JF. Signal stratification of autoantibody levels in serum samples and its application to the early detection of lung cancer. J Thorac Dis. 2013;5(5):618-25.

111. Kastenhuber ER, Lowe SW. Putting p53 in context. Cell. 2017;170(6):1062-78.

112. Suppiah A, Greenman J. Clinical utility of anti-p53 auto-antibody: systematic review and focus on colorectal cancer. World J Gastroenterol. 2013;19(29):4651-70.

113. Novak D, Hüser L, Elton JJ, Umansky V, Altevogt P, Utikal J, editors. SOX2 in development and cancer biology. Seminars in cancer biology; 2020: Elsevier.

114. Sun Y, Zhang R, Wang M, Zhang Y, Qi J, Li J. SOX2 Autoantibodies As Noninvasive Serum Biomarker for Breast CarcinomaSOX2 Autoantibodies in Breast Cancer. Cancer epidemiology, biomarkers & prevention. 2012;21(11):2043-7.

115. Maddison P, Thorpe A, Silcocks P, Robertson JF, Chapman CJ. Autoimmunity to SOX2, clinical phenotype and survival in patients with small-cell lung cancer. Lung Cancer. 2010;70(3):335-9.

116. Cho B, Lim Y, Lee D-Y, Park S-Y, Lee H, Kim WH, et al. Identification and Characterization of a Novel Cancer/Testis Antigen Gene CAGE. Biochemical and Biophysical Research Communications. 2002;292(3):715-26.

117. Thomas R, Al-Khadairi G, Roelands J, Hendrickx W, Dermime S, Bedognetti D, et al. NY-ESO-1 based immunotherapy of cancer: current perspectives. Frontiers in immunology. 2018;9:947.

118. Alsulami M. Characterisation of human TDRD12 and LKAAEAR1 as potential oncogenic cancer testis antigen genes with clinical potential: Bangor University; 2019.

119. Poojary M, Jishnu PV, Kabekkodu SP. Prognostic Value of Melanoma-Associated Antigen-A (MAGE-A) Gene Expression in Various Human Cancers: A Systematic Review and Meta-analysis of 7428 Patients and 44 Studies. Mol Diagn Ther. 2020;24(5):537-55.

120. Silvestri B, Mochi M, Garone MG, Rosa A. Emerging Roles for the RNA-Binding Protein HuD (ELAVL4) in Nervous System Diseases. International Journal of Molecular Sciences. 2022;23(23):14606.

121. Feifei W, Jianyi LU, Shentao LI, Xueyun HUO, Xin LIU, Xiaoyan DU, et al. Application of Serum ELAVL4 (HuD) Antigen Assay for Small Cell Lung Cancer Diagnosis. Anticancer Research. 2017;37(8):4515.

122. Sullivan F, Farmer E, Mair FS, Treweek S, Kendrick D, Jackson C, et al. Detection in blood of autoantibodies to tumour antigens as a case-finding method in lung cancer using the EarlyCDT®-Lung Test (ECLS): study protocol for a randomized controlled trial. BMC Cancer. 2017;17(1):1-10.

123. Li W-H, Zhao J, Li H-Y, Liu H, Li A-L, Wang H-X, et al. Proteomics-based identification of autoantibodies in the sera of healthy Chinese individuals from Beijing. Proteomics. 2006;6(17):4781-9.

124. Nolen B, Winans M, Marrangoni A, Lokshin A. Aberrant tumor-associated antigen autoantibody profiles in healthy controls detected by multiplex bead-based immunoassay. J Immunol Methods. 2009;344(2):116-20.

125. Foote J, Eisen HN. Kinetic and affinity limits on antibodies produced during immune responses. Proc Natl Acad Sci U S A. 1995;92(5):1254-6.

126. Reverberi R, Reverberi L. Factors affecting the antigen-antibody reaction. Blood Transfus. 2007;5(4):227-40.

127. Ferlay J, Colombet M, Soerjomataram I, Parkin DM, Piñeros M, Znaor A, et al. Cancer statistics for the year 2020: An overview. International journal of cancer. 2021;149(4):778-89.

128. UK CR. Lung cancer statistics 2023 [Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/lung-cancer.

129. Li Y, Karjalainen A, Koskinen H, Hemminki K, Vainio H, Shnaidman M, et al. p53 autoantibodies predict subsequent development of cancer. International Journal of Cancer. 2005;114(1):157-60.

130. Lu H, Ladd J, Feng Z, Wu M, Goodell V, Pitteri SJ, et al. Evaluation of Known Oncoantibodies, HER2, p53, and Cyclin B1, in Prediagnostic Breast Cancer Sera. Cancer Prevention Research. 2012;5(8):1036-43.

131. Jett JR, Peek LJ, Fredericks L, Jewell W, Pingleton WW, Robertson JFR. Audit of the autoantibody test, EarlyCDT®-Lung, in 1600 patients: An evaluation of its performance in routine clinical practice. Lung Cancer. 2014;83(1):51-5.

132. Jacobs IJ, Menon U, Ryan A, Gentry-Maharaj A, Burnell M, Kalsi JK, et al. Ovarian cancer screening and mortality in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): a randomised controlled trial. The Lancet. 2016;387(10022):945-56.

133. Usuda K, Saito Y, Sagawa M, Sato M, Kanma K, Takahashi S, et al. Tumor doubling time and prognostic assessment of patients with primary lung cancer. Cancer. 1994;74(8):2239-44.

134. Skates SJ, Greene MH, Buys SS, Mai PL, Brown P, Piedmonte M, et al. Early detection of ovarian cancer using the risk of ovarian cancer algorithm with frequent CA125 testing in women at increased familial risk–combined results from two screening trials. Clinical Cancer Research. 2017;23(14):3628-37.

135. Black WC, Gareen IF, Soneji SS, Sicks JD, Keeler EB, Aberle DR, et al. Cost-effectiveness of CT screening in the National Lung Screening Trial. N Engl J Med. 2014;371(19):1793-802.

136. National Institute for H. EarlyCDT Lung for assessing risk of lung cancer in solid lung nodules 2022 [Available from: https://www.nice.org.uk/guidance/indevelopment/gid-dg10041.

137. Weycker D, Jett JR, Detterbeck FC, Miller DL, Khuu A, Kennedy TC, et al. Cost-effectiveness of an autoantibody test (AABT) as an aid to diagnosis of lung cancer. J Clin Oncol. 2010;28(15_suppl):7030-.

138. Weycker D, Boyle P, Lopez A, Jett JR, Detterbeck F, Kennedy TC, et al. Cost-Effectiveness Of Screening Older Adult Smokers For Lung Cancer With An Autoantibody Test (AABT). B38 CAN WE IMPROVE QUALITY AND REDUCE COST OF CARE? American Thoracic Society International Conference Abstracts: American Thoracic Society; 2011. p. A2937-A.
139. Edelsberg J, Weycker D, Atwood M, Hamilton-Fairley G, Jett JR. Cost-effectiveness of an autoantibody test (EarlyCDT-Lung) as an aid to early diagnosis of lung cancer in patients with incidentally detected pulmonary nodules. PLoS One. 2018;13(5):e0197826.
140. Health SfP. Targeted screening for lung cancer in individuals at increased risk. In: Committee UNS, editor. 2022.
141. Field JK, Duffy SW, Baldwin DR, Brain KE, Devaraj A, Eisen T. The UK Lung Cancer Screening Trial: a pilot randomised controlled trial of low-dose computed tomography screening for the early detection of lung cancer. Health Technol Assess. 2016;20(40).
142. National Institute for H, Care E. Guide to the Methods of Technology Appraisal. 2013.
143. Excellence NIfHaC. Developing NICE Guidelines: The Manual. Developing NICE Guidelines: The Manual. NICE Process and Methods Guides. London2015.
144. Shen L, Wu X, Tan J, Gu M, Teng Y, Wang Z, et al. Combined detection of dickkopf-1 subtype classification autoantibodies as biomarkers for the diagnosis and prognosis of non-small cell lung cancer. Oncotargets and Therapy. 2017;Volume 10:3545-56.
145. Li P, Shi J-X, Xing M-T, Dai L-P, Li J-T, Zhang J-Y. Evaluation of serum autoantibodies against tumor-associated antigens as biomarkers in lung cancer. Tumor Biology. 2017;39(10):101042831771166.
146. Molina R, Marrades RM, Augé JM, Escudero JM, Viñolas N, Reguart N, et al. Assessment of a Combined Panel of Six Serum Tumor Markers for Lung Cancer. American Journal of Respiratory and Critical Care Medicine. 2016;193(4):427-37.
147. Jett J, Healey G, Macdonald I, Parsy-Kowalska C, Peek L, Murray A. P2.13-013 Determination of the Detection Lead Time for Autoantibody Biomarkers in Early Stage Lung Cancer Using the UKCTOCS Cohort. Journal of Thoracic Oncology. 2017;12(11):S2170.
148. Nguyen QH, Ly H-B, Ho LS, Al-Ansari N, Le HV, Tran VQ, et al. Influence of Data Splitting on Performance of Machine Learning Models in Prediction of Shear Strength of Soil. Mathematical Problems in Engineering. 2021;2021:4832864.
149. Yap BW, Sim CH. Comparisons of various types of normality tests. Journal of Statistical Computation and Simulation. 2011;81(12):2141-55.
150. Box GEP, Cox DR. An Analysis of Transformations. Journal of the Royal Statistical Society Series B (Methodological). 1964;26(2):211-52.
151. Yeo IK, Johnson RA. A new family of power transformations to improve normality or symmetry. Biometrika. 2000;87(4):954-9.
152. Wold S, Esbensen K, Geladi P. Principal component analysis. Chemometrics and Intelligent Laboratory Systems. 1987;2(1):37-52.
153. Gideon S. Estimating the Dimension of a Model. The Annals of Statistics. 1978;6(2):461-4.
154. Hofmeyr DP. Degrees of freedom and model selection for k-means clustering. Computational Statistics & Data Analysis. 2020;149:106974.
155. Kursa MB, Rudnicki WR. Feature Selection with the Boruta Package. Journal of Statistical Software. 2010;36(11):1 - 13.
156. Bewick V, Cheek L, Ball J. Statistics review 14: Logistic regression. Crit Care. 2005;9(1):112-8.
157. Deppen SA, Blume JD, Aldrich MC, Fletcher SA, Massion PP, Walker RC, et al. Predicting Lung Cancer Prior to Surgical Resection in Patients with Lung Nodules. Journal of Thoracic Oncology. 2014;9(10):1477-84.

158. Swensen SJ, Silverstein MD, Ilstrup DM, Schleck CD, Edell ES. The probability of malignancy in solitary pulmonary nodules: application to small radiologically indeterminate nodules. Archives of internal medicine. 1997;157(8):849-55.

159. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological). 1996;58(1):267-88.

160. Hastie T, Tibshirani R, Tibshirani RJ. Extended comparisons of best subset selection, forward stepwise selection, and the lasso. arXiv preprint arXiv:170708692. 2017.

161. Kim SA-O, Kim YA-O, Jeong KA-O, Jeong HA-O, Kim JA-O. Logistic LASSO regression for the diagnosis of breast cancer using clinical demographic data and the BI-RADS lexicon for ultrasonography. (2288-5919 (Print)).

162. Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995;20(3):273-97.

163. Ishikawa T, Takahashi J, Takemura H, Mizoguchi H, Kuwata T, editors. Gastric lymph node cancer detection using multiple features support vector machine for pathology diagnosis support system. The 15th International Conference on Biomedical Engineering: ICBME 2013, 4th to 7th December 2013, Singapore; 2014: Springer.

164. Shah V, Turkbey B, Mani H, Pang Y, Pohida T, Merino MJ, et al. Decision support system for localizing prostate cancer based on multiparametric magnetic resonance imaging. Medical physics. 2012;39(7Part1):4093-103.

165. Kim W, Kim KS, Lee JE, Noh D-Y, Kim S-W, Jung YS, et al. Development of novel breast cancer recurrence prediction model using support vector machine. Journal of breast cancer. 2012;15(2):230-8.

166. Sun T, Wang J, Li X, Lv P, Liu F, Luo Y, et al. Comparative evaluation of support vector machines for computer aided diagnosis of lung cancer in CT based on a multi-dimensional data set. Computer Methods and Programs in Biomedicine. 2013;111(2):519-24.

167. Zaw HT, Maneerat N, Win KY, editors. Brain tumor detection based on Naïve Bayes Classification. 2019 5th International Conference on Engineering, Applied Sciences and Technology (ICEAST); 2019 2-5 July 2019.

168. Karabatak M. A new classifier for breast cancer detection based on Naïve Bayesian. Measurement. 2015;72:32-6.

169. Dimitoglou G, Adams JA, Jim CM. Comparison of the C4. 5 and a Naïve Bayes classifier for the prediction of lung cancer survivability. arXiv preprint arXiv:12061121. 2012.

170. Breiman L. Classification and regression trees: Routledge; 2017.

171. Chipman HA, George EI, McCulloch RE. BART: Bayesian additive regression trees. 2010.

172. Quinlan J. Induction of decision trees. mach. learn. 1986.

173. Quinlan JR. C4. 5: programs for machine learning: Elsevier; 2014.

174. Breiman L. Machine Learning. 2001;45(1):5-32.

175. Toth R, Schiffmann H, Hube-Magg C, Büscheck F, Höflmayer D, Weidemann S, et al. Random forest-based modelling to detect biomarkers for prostate cancer progression. Clinical Epigenetics. 2019;11(1):148.

176. Velazquez M, Lee Y. Random forest model for feature-based Alzheimer's disease conversion prediction from early mild cognitive impairment subjects. PLoS One. 2021;16(4):e0244773.

177. Wright MN, Ziegler A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. Journal of Statistical Software. 2017;77(1):1 - 17.

178. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Machine Learning. 2006;63(1):3-42.

179. Aler R, Valls JM, Boström H. Study of Hellinger Distance as a splitting metric for Random Forests in balanced and imbalanced classification datasets. Expert Systems with Applications. 2020;149:113264.
180. Chen T, Guestrin C, editors. XGBoost. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 2016-08-13: ACM.
181. Srinivas P, Katarya R. hyOPTXg: OPTUNA hyper-parameter optimization framework for predicting cardiovascular disease using XGBoost. Biomedical Signal Processing and Control. 2022;73:103456.
182. Sharma A, Verbeke WJMI. Improving Diagnosis of Depression With XGBOOST Machine Learning Model and a Large Biomarkers Dutch Dataset (n = 11,081). Frontiers in Big Data. 2020;3.
183. Pan Z, Zhang R, Shen S, Lin Y, Zhang L, Wang X, et al. OWL: an optimized and independently validated machine learning prediction model for lung cancer screening based on the UK Biobank, PLCO, and NLST populations. eBioMedicine. 2023;88.
184. Carlson GD, Calvanese CB, Partin AW. An algorithm combining age, total prostate-specific antigen (PSA), and percent free PSA to predict prostate cancer: results on 4298 cases. Urology. 1998;52(3):455-61.
185. Bach PB, Kattan MW, Thornquist MD, Kris MG, Tate RC, Barnett MJ, et al. Variations in lung cancer risk among smokers. Journal of the National Cancer Institute. 2003;95(6):470-8.
186. David OW, Joel W. A simple model for predicting lung cancer occurrence in a lung cancer screening program: The Pittsburgh Predictor. Lung Cancer. 2015;89(1):31-7.
187. Hong Y, Kim WJ. DNA Methylation Markers in Lung Cancer. (1389-2029 (Print)).
188. Macdonald IK, Murray A, Healey GF, Parsy-Kowalska CB, Allen J, McElveen J, et al. Application of a High Throughput Method of Biomarker Discovery to Improvement of the EarlyCDT®-Lung Test. PLOS ONE. 2012;7(12):e51002.
189. Benjamin T, Nicolas D, Philippe G, Amélie S, Loïc G, Luc M. Alpha-enolase: A target of antibodies in infectious and autoimmune diseases. Autoimmunity Reviews. 2007;6(3):176-82.
190. Zhang L, Lu T, Yang Y, Hu L. &alpha;-enolase is highly expressed in liver cancer and promotes cancer cell invasion and metastasis. Oncol Lett. 2020;20(5):152.
191. Sun L, Guo C, Cao J, Burnett J, Yang Z, Ran Y, et al. Over-Expression of Alpha-Enolase as a Prognostic Biomarker in Patients with Pancreatic Cancer. Int J Med Sci. 2017;14(7):655-61.
192. Zang R, Li Y, Jin R, Wang X, Lei Y, Che Y, et al. Enhancement of diagnostic performance in lung cancers by combining CEA and CA125 with autoantibodies detection. OncoImmunology. 2019;8(10):e1625689.
193. Cancemi P, Buttacavoli M, Roz E, Feo S. Expression of Alpha-Enolase (ENO1), Myc Promoter-Binding Protein-1 (MBP-1) and Matrix Metalloproteinases (MMP-2 and MMP-9) Reflect the Nature and Aggressiveness of Breast Tumors. International Journal of Molecular Sciences. 2019;20(16):3952.
194. Liu Z, Tang H, Zhang W, Wang J, Wan L, Li X, et al. Coupling of serum CK20 and hyper-methylated CLIP4 as promising biomarker for colorectal cancer diagnosis: from bioinformatics screening to clinical validation. Aging (Albany NY). 2021;13(24):26161-79.
195. ROLAND M. Cytokeratin 20 in Human Carcinomas. Am J Pathol. 1992;140:427-47.
196. Heo C-K, Hwang H-M, Ruem A, Yu D-Y, Lee JY, Yoo JS, et al. Identification of a mimotope for circulating anti-cytokeratin 8/18 antibody and its usage for the diagnosis of breast cancer. Int J Oncol. 2013;42(1):65-74.

197. Wu R, Lin L, Beer DG, Ellenson LH, Lamb BJ, Rouillard J-M, et al. Amplification and overexpression of the L-MYC proto-oncogene in ovarian carcinomas. The American journal of pathology. 2003;162(5):1603-10.
198. Nau MM, Brooks BJ, Battey J, Sausville E, Gazdar AF, Kirsch IR, et al. L-myc, a new myc-related gene amplified and expressed in human small cell lung cancer. Nature. 1985;318(6041):69-73.
199. Nobori T, Miura K, Wu DJ, Lois A, Takabayashi K, Carson DA. Deletions of the cyclin-dependent kinase-4 inhibitor gene in multiple human cancers. Nature. 1994;368(6473):753-6.
200. Zhang C, Ye L, Guan S, Jin S, Wang W, Sun S, et al. Autoantibodies against p16 protein-derived peptides may be a potential biomarker for non-small cell lung cancer. Tumor Biology. 2014;35:2047-51.
201. Liu W, Peng B, Lu Y, Xu W, Qian W, Zhang J-Y. Autoantibodies to tumor-associated antigens as biomarkers in cancer immunodiagnosis. Autoimmunity Reviews. 2011;10(6):331-5.
202. Ye H, Sun C, Ren P, Dai L, Peng B, Wang K, et al. Mini-array of multiple tumor-associated antigens (TAAs) in the immunodiagnosis of breast cancer. Oncology letters. 2013;5(2):663-8.
203. Looi K, Megliorino R, Shi F-D, Peng X-X, Chen Y, Zhang J-Y. Humoral immune response to p16, a cyclin-dependent kinase inhibitor in human malignancies. Oncology reports. 2006;16(5):1105-10.
204. Yantiss RK, Woda BA, Fanger GR, Kalos M, Whalen GF, Tada H, et al. KOC (K Homology Domain Containing Protein Overexpressed in Cancer): A Novel Molecular Marker That Distinguishes Between Benign and Malignant Lesions of the Pancreas. The American Journal of Surgical Pathology. 2005;29(2):188-95.
205. Wang T, Fan L, Watanabe Y, McNeill P, Moulton G, Bangur C, et al. L523S, an RNA-binding protein as a potential therapeutic target for lung cancer. British journal of cancer. 2003;88(6):887-94.
206. Jiang Z, Chu PG, Woda BA, Rock KL, Liu Q, Hsieh C-C, et al. Analysis of RNA-binding protein IMP3 to predict metastasis and prognosis of renal-cell carcinoma: a retrospective study. The lancet oncology. 2006;7(7):556-64.
207. Li C, Zota V, Woda BA, Rock KL, Fraire AE, Jiang Z, et al. Expression of a novel oncofetal mRNA-binding protein IMP3 in endometrial carcinomas: diagnostic significance and clinicopathologic correlations. Modern Pathology. 2007;20(12):1263-8.
208. Pryor JG, Bourne PA, Yang Q, Spaulding BO, Scott GA, Xu H. IMP-3 is a novel progression marker in malignant melanoma. Modern Pathology. 2008;21(4):431-7.
209. Findeis-Hosey JJ, Yang Q, Spaulding BO, Wang HL, Xu H. IMP3 expression is correlated with histologic grade of lung adenocarcinoma. Human Pathology. 2010;41(4):477-84.
210. Zhang J-Y, Casiano CA, Peng X-X, Koziol JA, Chan EKL, Tan EM. Enhancement of antibody detection in cancer using panel of recombinant tumor-associated antigens. Cancer Epidemiol Biomarkers Prev. 2003;12(2):136-43.
211. Roudi R, Korourian A, Shariftabrizi A, Madjd Z. Differential expression of cancer stem cell markers ALDH1 and CD133 in various lung cancer subtypes. Cancer investigation. 2015;33(7):294-302.
212. Sarkar P, Basu K, Sarkar P, Chatterjee U, Mukhopadhyay M, Choudhuri MK, et al. Correlations of aldehyde dehydrogenase-1 (ALDH1) expression with traditional prognostic parameters and different molecular subtypes of breast carcinoma. Clujul Medical. 2018;91(2):181.
213. Rezaee M, Gheytanchi E, Madjd Z, Mehrazma M. Clinicopathological significance of tumor stem cell markers ALDH1 and CD133 in colorectal carcinoma. Iranian Journal of Pathology. 2021;16(1):40.

214. Resetkova E, Reis-Filho JS, Jain RK, Mehta R, Thorat MA, Nakshatri H, et al. Prognostic impact of ALDH1 in breast cancer: a story of stem cells and tumor microenvironment. Breast cancer research and treatment. 2010;123:97-108.

215. Chen J, Xia Q, Jiang B, Chang W, Yuan W, Ma Z, et al. Prognostic value of cancer stem cell marker ALDH1 expression in colorectal cancer: a systematic review and meta-analysis. PloS one. 2015;10(12):e0145164.

216. Moscat J, Diaz-Meco MT. p62 at the Crossroads of Autophagy, Apoptosis, and Cancer. Cell. 2009;137(6):1001-4.

217. Mathew R, Karp CM, Beaudoin B, Vuong N, Chen G, Chen H-Y, et al. Autophagy Suppresses Tumorigenesis through Elimination of p62. Cell. 2009;137(6):1062-75.

218. Smith HA, McNeel DG. The SSX Family of Cancer-Testis Antigens as Target Proteins for Tumor Therapy. Clinical and Developmental Immunology. 2010;2010:150591.

219. Cox AD, Fesik SW, Kimmelman AC, Luo J, Der CJ. Drugging the undruggable RAS: Mission possible? Nat Rev Drug Discov. 2014;13(11):828-51.

220. Waters AM, Der CJ. KRAS: the critical driver and therapeutic target for pancreatic cancer. Cold Spring Harbor perspectives in medicine. 2018;8(9):a031435.

221. Zhu G, Pei L, Xia H, Tang Q, Bi F. Role of oncogenic KRAS in the prognosis, diagnosis and treatment of colorectal cancer. Molecular Cancer. 2021;20(1):143.

222. Westcott PM, To MD. The genetics and biology of KRAS in lung cancer. Chin J Cancer. 2013;32(2):63-70.

223. Schabath MB, Cote ML. Cancer Progress and Priorities: Lung Cancer. Cancer Epidemiology, Biomarkers & Prevention. 2019;28(10):1563-79.

224. Chapman C, Murray A, Chakrabarti J, Thorpe A, Woolston C, Sahin U, et al. Autoantibodies in breast cancer: their use as an aid to early diagnosis. Ann Oncol. 2007;18(5):868-73.

225. Negm OH, Hamed MR, Schoen RE, Whelan RL, Steele RJ, Scholefield J, et al. Human Blood Autoantibodies in the Detection of Colorectal Cancer. PLoS One. 2016;11(7):e0156971.

226. Welberry C, Macdonald I, McElveen J, Parsy-Kowalska C, Allen J, Healey G, et al. Tumor-associated autoantibodies in combination with alpha-fetoprotein for detection of early stage hepatocellular carcinoma. PLOS ONE. 2020;15(5):e0232247.

227. Murray A, Chapman CJ, Healey G, Peek LJ, Parsons G, Baldwin D, et al. Technical validation of an autoantibody test for lung cancer. Ann Oncol. 2010;21(8):1687-93.

228. Shapiro A. Monte Carlo Sampling Methods. Handbooks in Operations Research and Management Science. 10: Elsevier; 2003. p. 353-425.

229. Bertsimas D, Tsitsiklis J. Simulated Annealing. Statistical Science. 1993;8(1):10-5, 6.

230. Kirkpatrick S, Gelatt CD, Vecchi MP. Optimization by Simulated Annealing. Science. 1983;220(4598):671-80.

231. Bhanot G. The Metropolis algorithm. Reports on Progress in Physics. 1988;51(3):429.

232. Xavier R, Natacha T, Alexandre H, Natalia T, Frédérique L, Jean-Charles S, et al. PanelomiX: A threshold-based algorithm to create panels of biomarkers. Translational Proteomics. 2013;1(1):57-64.

233. Claxton K, Sculpher M, Palmer S, Culyer AJ. Causes for concern: is NICE failing to uphold its responsibilities to all NHS patients? : Wiley Online Library; 2015. p. 1-7.

234. Price WN. Big data and black-box medical algorithms. Sci Transl Med. 2018;10(471).

235. Agencies HoM, Agency EM. HMA-EMA Joint Big Data Taskforce–Summary report. 2019.

236. Verbraak FD, Abramoff MD, Bausch GC, Klaver C, Nijpels G, Schlingemann RO, et al. Diagnostic accuracy of a device for the automated detection of diabetic retinopathy in a primary care setting. Diabetes care. 2019;42(4):651-6.

237. Bruno TC, Ebner PJ, Moore BL, Squalls OG, Waugh KA, Eruslanov EB, et al. Antigen-Presenting Intratumoral B Cells Affect CD4+ TIL Phenotypes in Non–Small Cell Lung Cancer PatientsTIL-Bs Present Antigen to CD4 TILs in NSCLC. Cancer immunology research. 2017;5(10):898-907.

238. Flahault A, Cadilhac M, Thomas G. Sample size calculation should be performed for design accuracy in diagnostic test studies. Journal of Clinical Epidemiology. 2005;58(8):859-62.

239. Sullivan FM, Mair FS, Anderson W, Armory P, Briggs A, Chew C, et al. Earlier diagnosis of lung cancer in a randomised trial of an autoantibody blood test followed by imaging. Eur Respir J. 2021;57(1).

240. Putcha G, Xu C, Shaukat M, Levin TR. Prevention of colorectal cancer through multiomics blood testing: The PREEMPT CRC study. J Clin Oncol. 2022;40.

241. Davies-Dear S. Lung Health Checks in Wessex and Yorkshire: Integrated Biomarker Studies. 2021.

242. Skates SJ, Xu FJ, Yu YH, Sjövall K, Einhorn N, Chang Y, et al. Toward an optimal algorithm for ovarian cancer screening with longitudinal tumor markers. Cancer. 1995;76(S10):2004-10.

243. Breiman L. Bagging predictors. Machine learning. 1996;24:123-40.

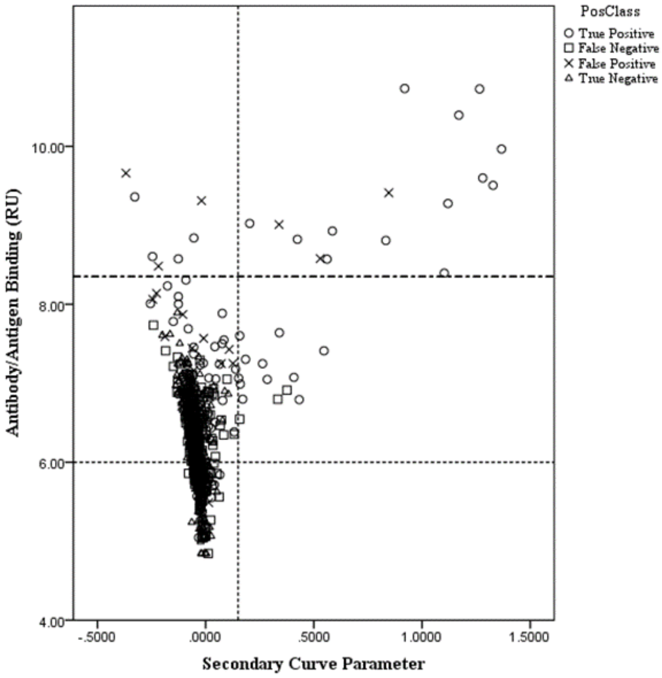
244. Begg CB. Biases in the assessment of diagnostic tests. Statistics in Medicine. 1987;6(4):411-23.

245. Sullivan F, Schembri S. Progress with an RCT of the Detection of Autoantibodies to Tumour Antigens in Lung Cancer Using the Early CDT-Lung Test in Scotland (ECLS). JOURNAL OF THORACIC ONCOLOGY; 2015: … INC 360 PARK AVE SOUTH, NEW …; 2015. p. S306-S.

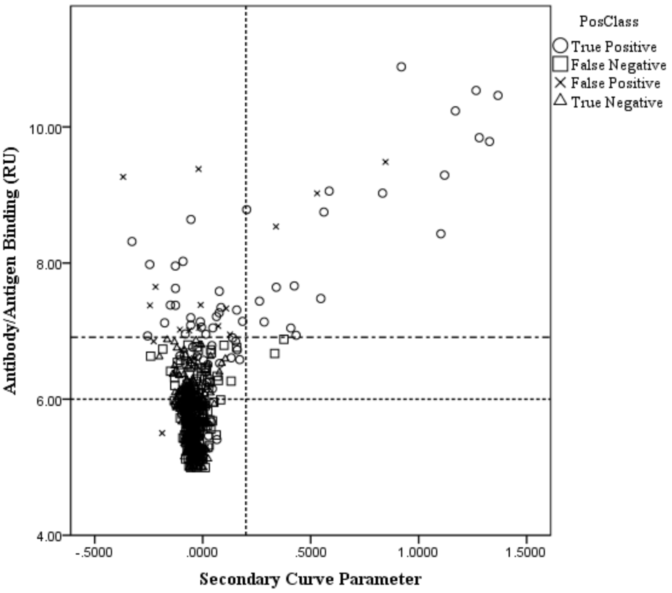
Chapter 12: Appendices

12.1 Appendix 2A: Application of Curve Characteristic Cut-off Thresholds to Improve Specificity

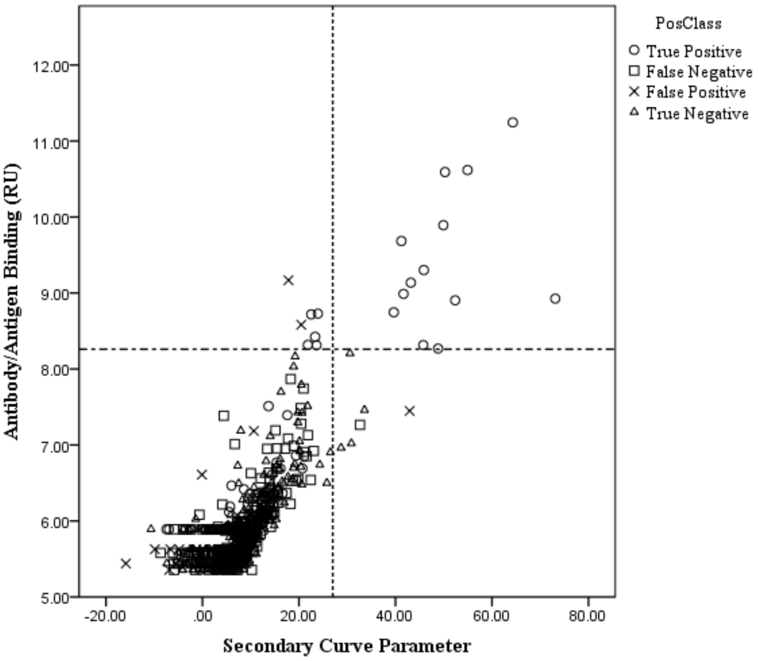
Training Set p53 160nM
Commercial threshold: RU≥8.35
Proposed Paired thresholds: RU≥6 & Secondary Parameter (Intercept)≥0.15



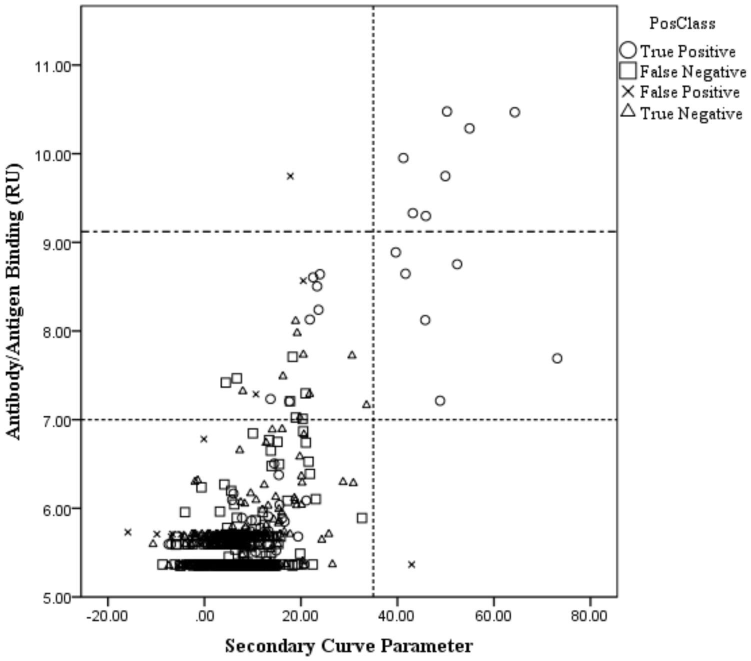
Training Set p53 50nM
Commercial threshold: RU≥6.91
Proposed Paired thresholds: RU≥6 & Secondary Parameter (Intercept)≥0.2



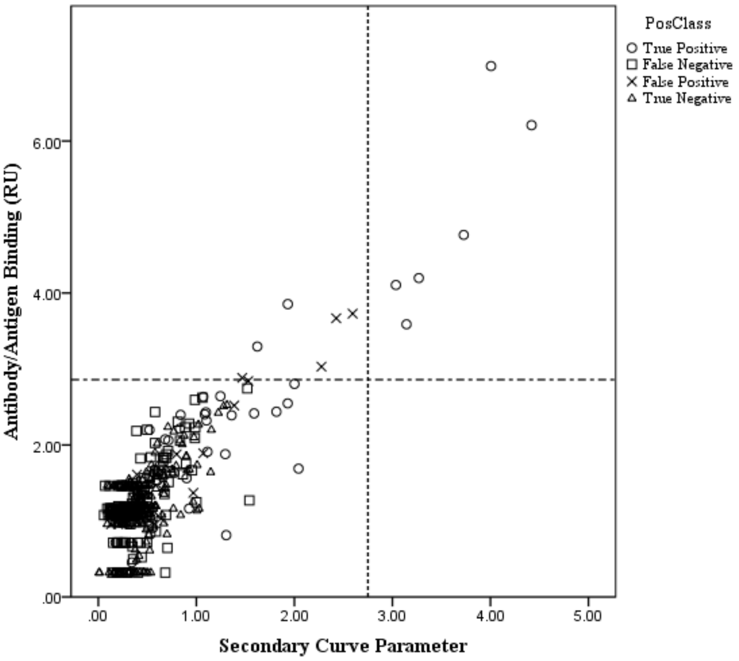
Training Set SOX2 160nM
Commercial threshold: RU≥8.26
Proposed Paired thresholds: RU≥8.26 & Secondary Parameter (AUC)≥27



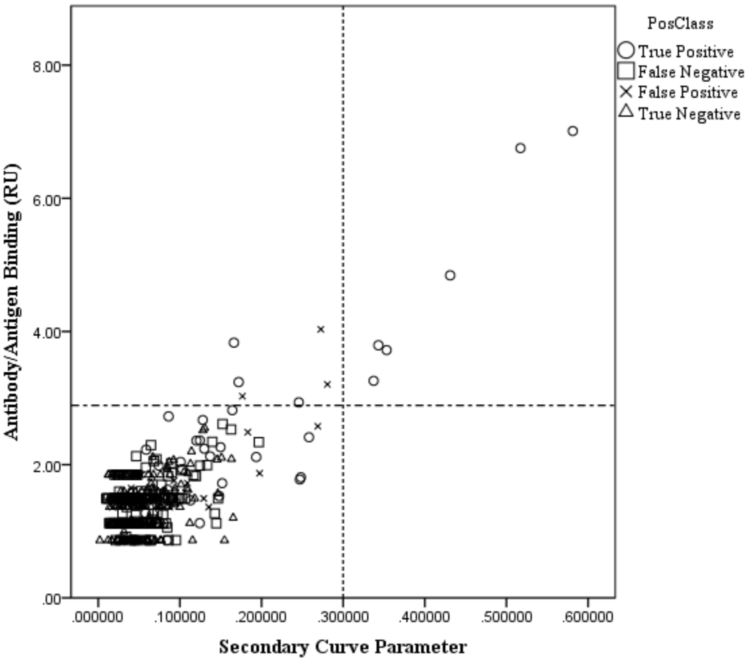
Training Set SOX2 50nM
Commercial threshold: RU≥9.12
Proposed Paired thresholds: RU≥7 & Secondary Parameter (AUC)≥35


PosClass
True Positive
False Negative
False Positive
True Negative
Antibody/Antigen Binding (RU)
11.00
10.00
9.00
8.00
7.00
6.00
5.00
-20.00
.00
20.00
40.00
60.00
80.00
Secondary Curve Parameter

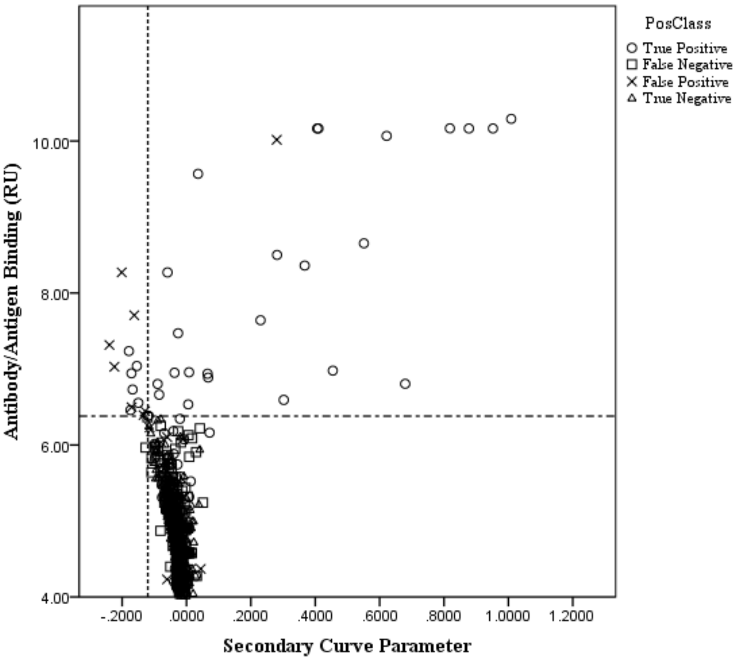
Training Set CAGE 160nM
Commercial threshold: RU≥2.86
Proposed Paired thresholds: RU≥2.86 & Secondary Parameter (AUC)≥2.8



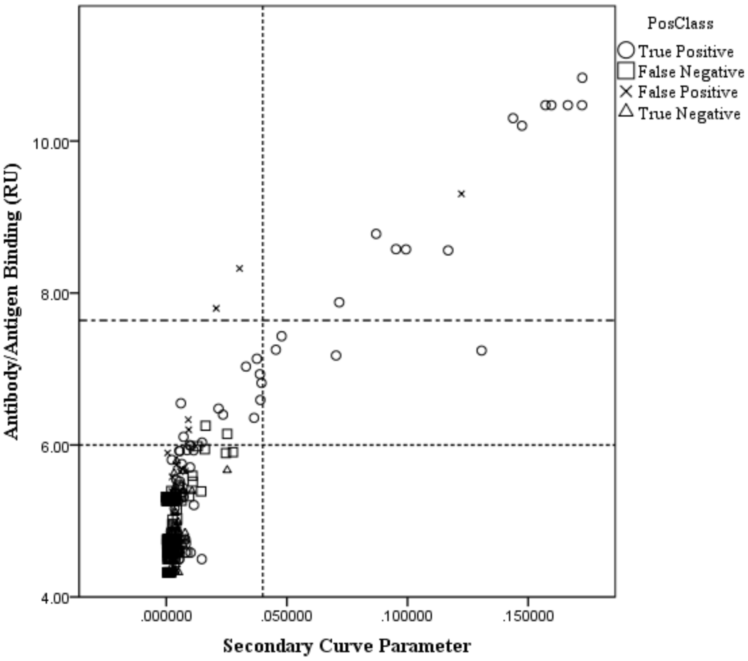
Training Set CAGE 50nM
Commercial threshold: RU≥2.89
Proposed Paired thresholds: RU≥2.89 & Secondary Parameter (SlopeMax)≥0.3



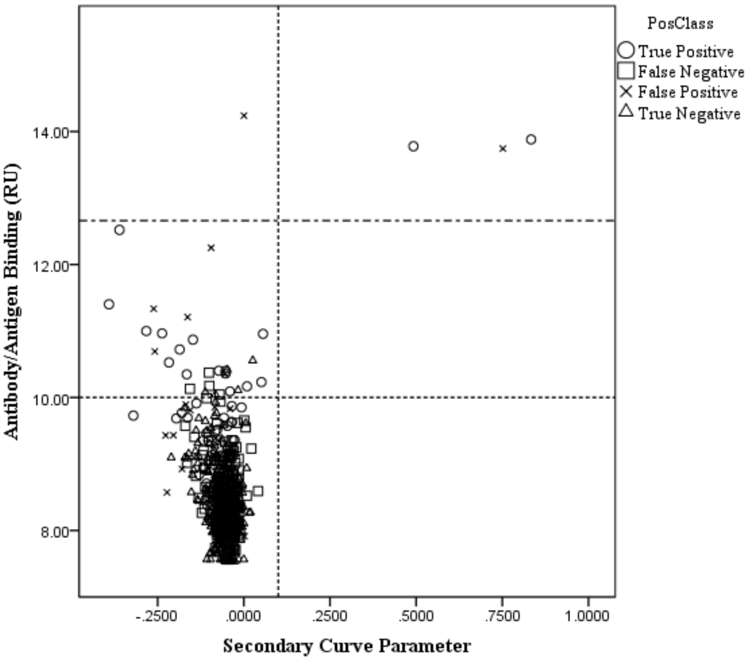
Training Set NY-ESO-1 160nM
Commercial threshold: RU≥6.38
Proposed Paired thresholds: RU≥6.38 & Secondary Parameter (Intercept)≥-0.12



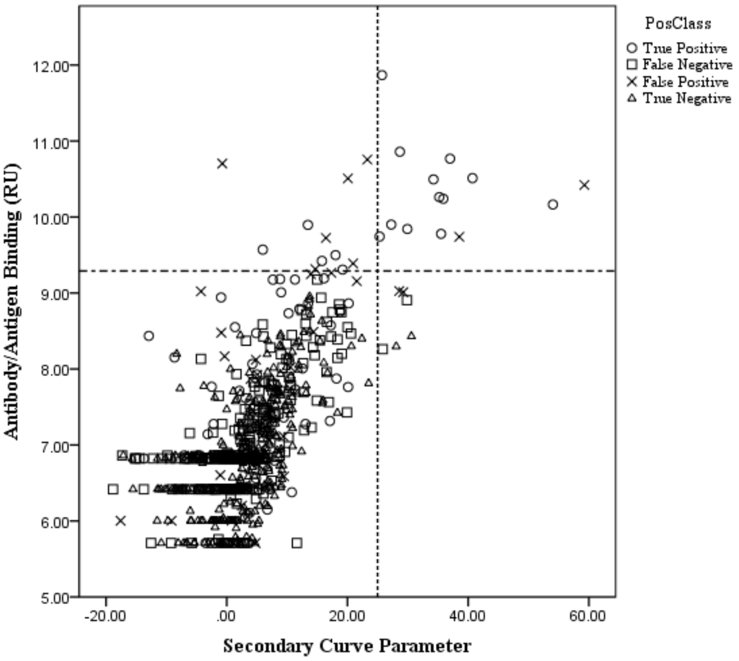
Training Set NY-ESO-1 50nM
Commercial threshold: RU≥7.64
Proposed Paired thresholds: RU≥6 & Secondary Parameter (SlopeMax)≥0.04



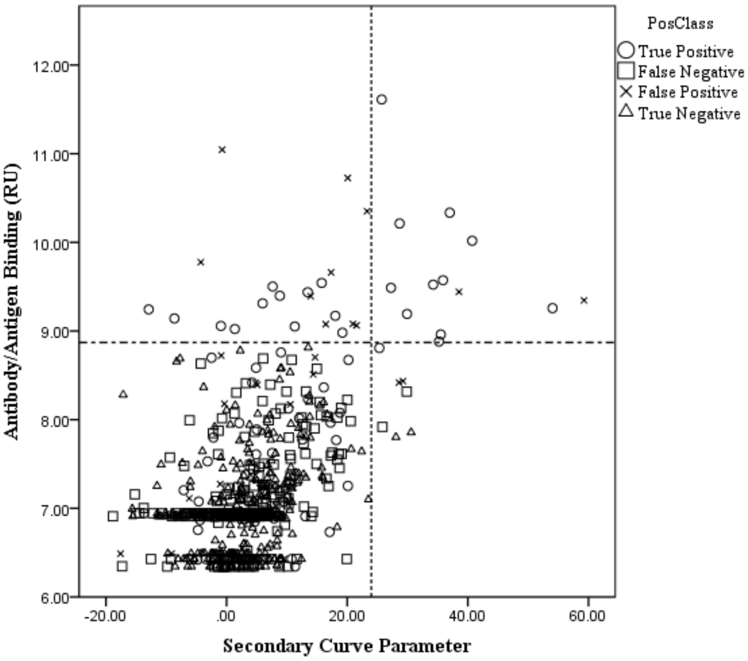
Training Set GBU 4-5 50nM
Commercial threshold: RU≥12.66
Proposed Paired thresholds: RU≥10 & Secondary Parameter (Intercept)≥0.1



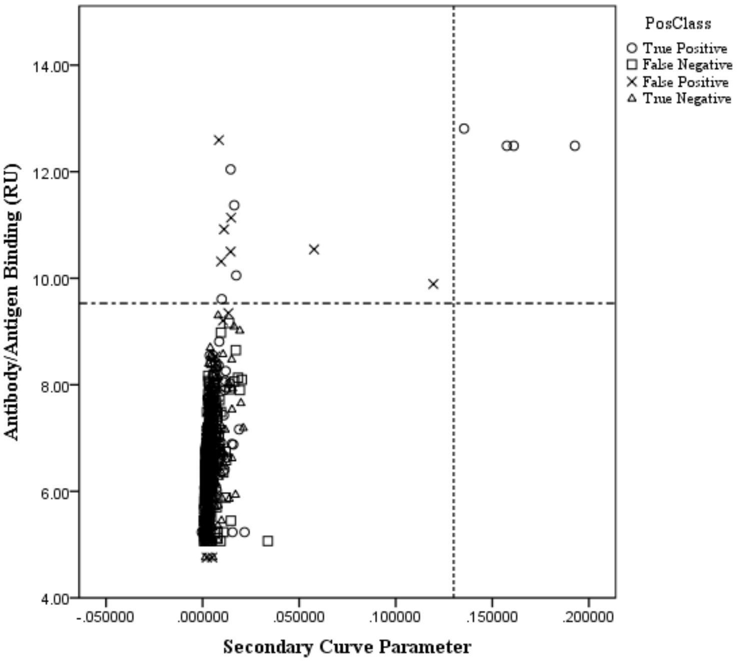
Training Set MAGE-A4 160nM
Commercial threshold: RU≥9.29
Proposed Paired thresholds: RU≥9.29 & Secondary Parameter (AUC)≥25



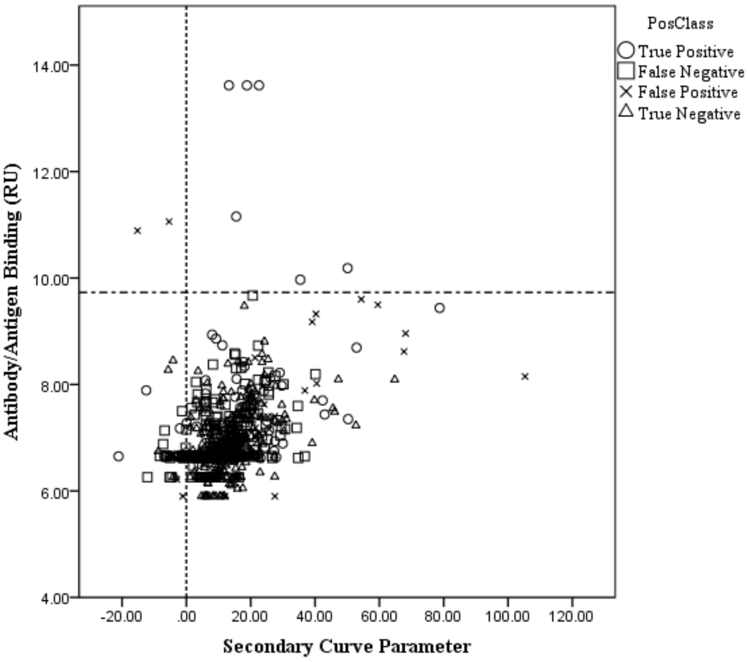
Training Set MAGE-A4 50nM
Commercial threshold: RU≥8.87
Proposed Paired thresholds: RU≥8.87 & Secondary Parameter (AUC)≥24



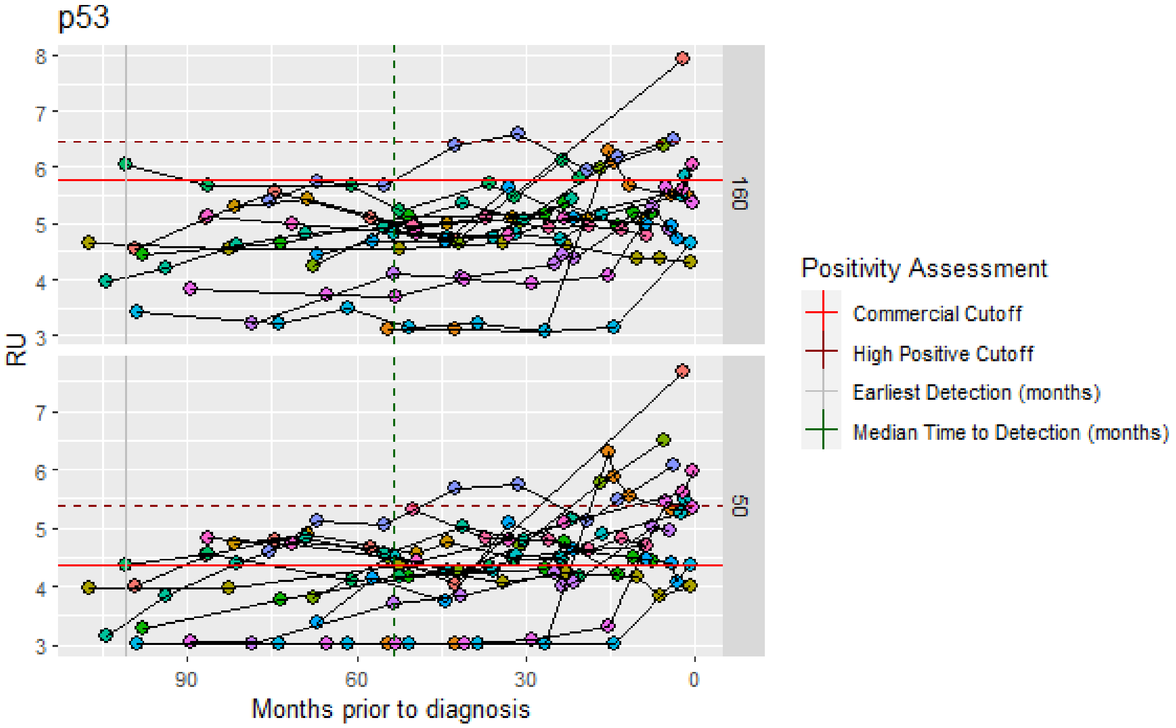
Training Set HuD 160nM
Commercial threshold: RU≥9.53
Proposed Paired thresholds: RU≥9.53 & Secondary Parameter (SlopeMax)≥0.13

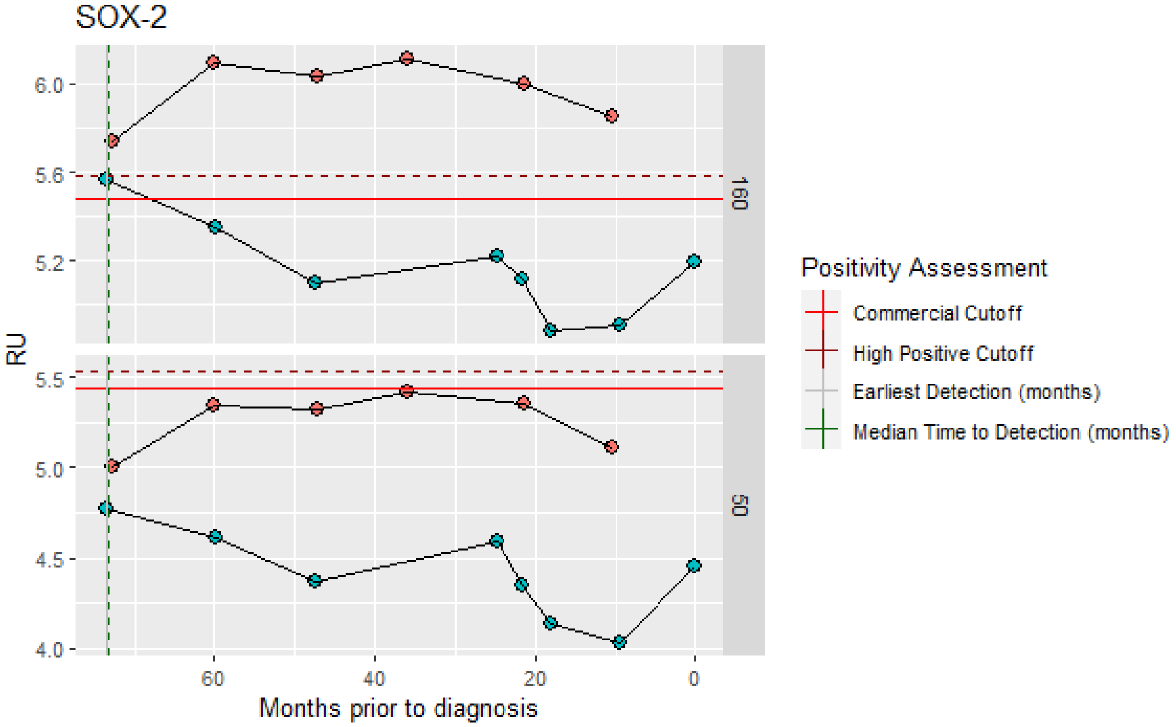


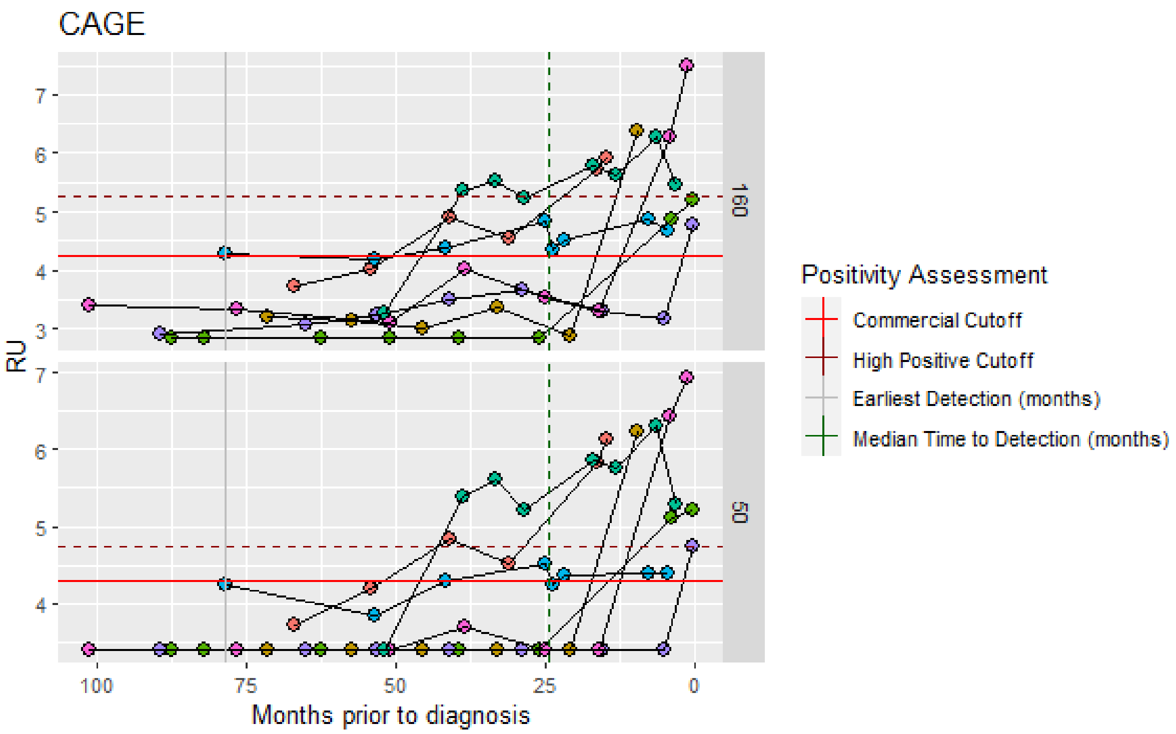
Training Set HuD 50nM
Commercial threshold: RU≥9.73
Proposed Paired thresholds: RU≥9.73 & Secondary Parameter (SlopeMax)≥0.0

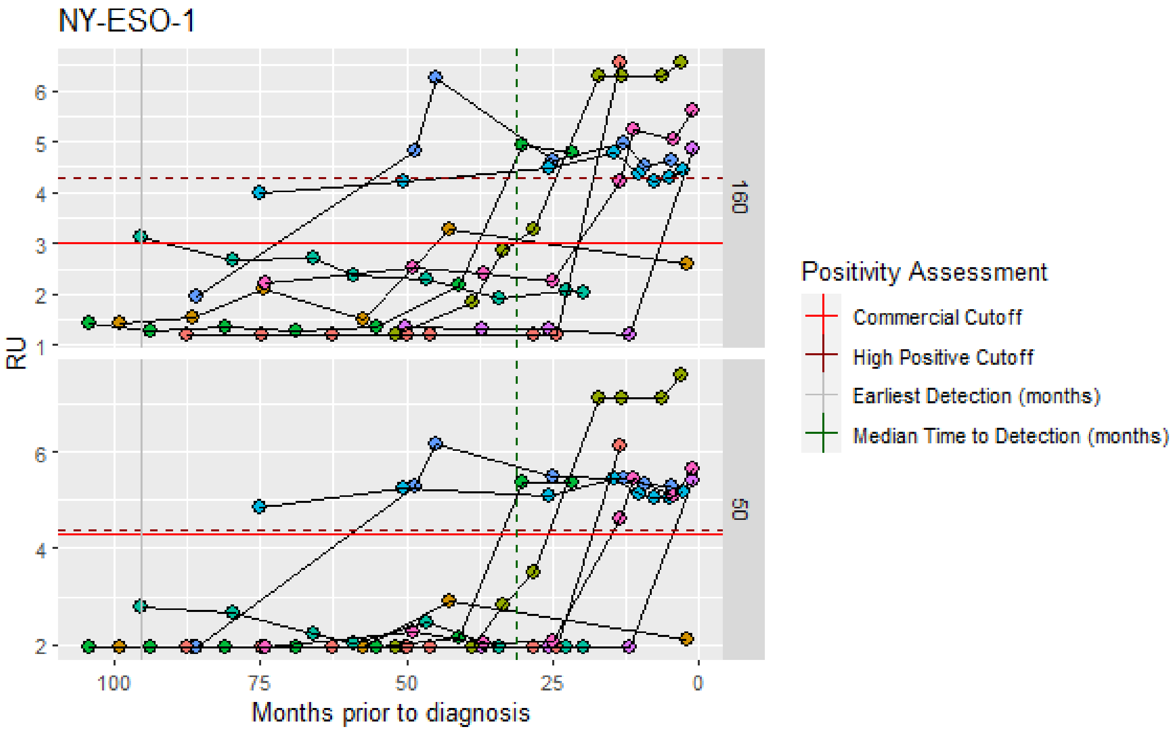


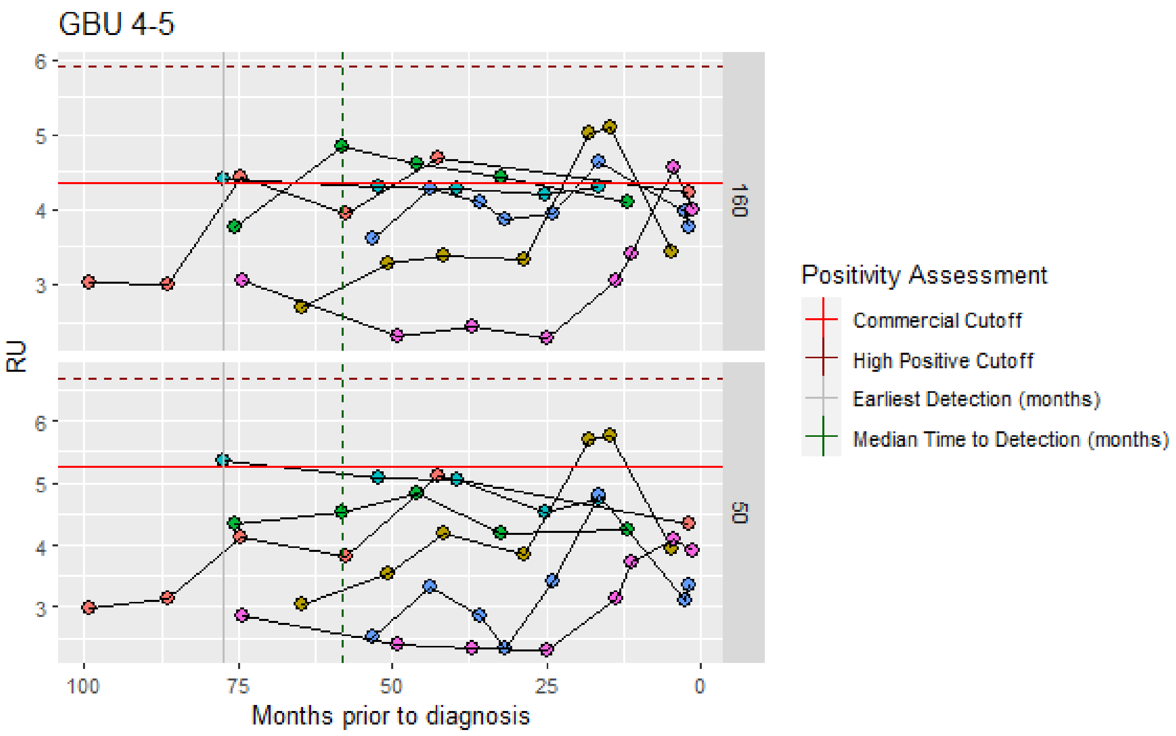
12.2 Appendix 3A: Longitudinal autoantibody profiles in positive cases



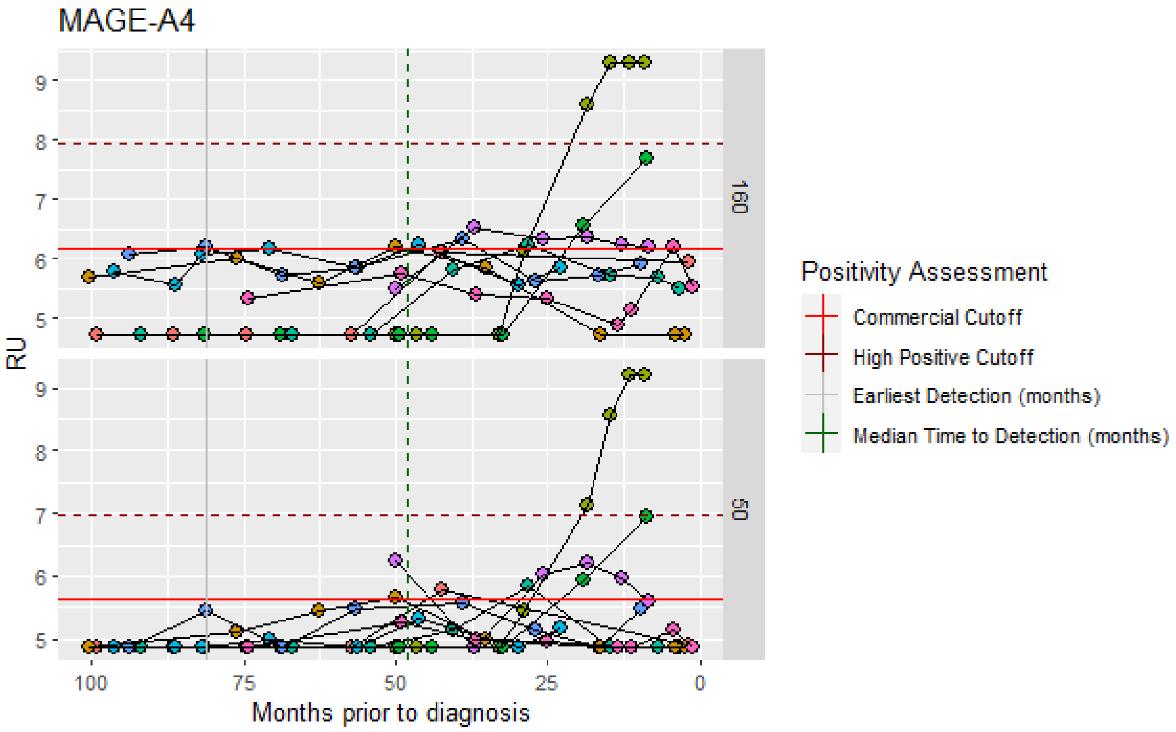



CAGE
RU
7
6
5
4
3
160
7
6
5
4
50
100
75
50
25
0
Months prior to diagnosis
Positivity Assessment
Commercial Cutoff
High Positive Cutoff
Earliest Detection (months)
Median Time to Detection (months)


NY-ESO-1
RU
6
5
4
3
2
1
160
6
4
2
50
100
75
50
25
0
Months prior to diagnosis
Positivity Assessment
Commercial Cutoff
High Positive Cutoff
Earliest Detection (months)
Median Time to Detection (months)



GBU 4-5



MAGE-A4

12.3 Appendix 4A: Parameters used for simulation of high risk population in the health economic model

Parameter	Description	Base-case value
Population parameters		
pop_size	Size of screening population	13,000,000
p_male	Proportion of males in screening population	0.482
pop_age_mean	Mean age of screening population	61.939
pop_age_sd	Standard deviation of age of screening population	9.000
pop_age_LL	Quantile for lower age limit of screening population	0.220
pop_age_UL	Quantile for upper age limit of screening population	0.978
Screening programme parameters		
p_respond	Probability of invited individual responding and returning questionnaire	0.307
p_join	Probability of joining screening if eligible	0.465
Disease natural history model parameters		
mu_AB	Lognormal probability distribution parameter (location) for pre-clinical incidence of lung cancer	4.7470
delta_mu_AB_F	Coefficient for women for lognormal probability distribution parameter (location) for pre-clinical incidence of lung cancer	0.0358
sigma_AB	Lognormal probability distribution parameter (shape) for pre-clinical incidence of lung cancer	0.3635
ln_lambda_pIA_pIB	Log rate of pre-clinical progression from stage Ia to Ib	0.0035
ln_lambda_pIB_pIIA	Log rate of pre-clinical progression from stage Ib to IIa	1.6451
ln_lambda_pIIA_pIIB	Log rate of pre-clinical progression from stage IIa to IIb	1.8006
ln_lambda_pIIB_pIIIA	Log rate of pre-clinical progression from stage IIb to IIIa	1.6258
ln_lambda_pIIIA_pIIIB	Log rate of pre-clinical progression from stage IIIa to IIIb	1.0797
ln_lambda_pIIIB_pIV	Log rate of pre-clinical progression from stage IIIb to IV	2.0803
ln_lambda_pIA_cIA	Log rate of clinical presentation at stage Ia	-2.4828
ln_lambda_pIB_cIB	Log rate of clinical presentation at stage Ib	-1.8726
ln_lambda_pIIA_cIIA	Log rate of clinical presentation at stage IIa	-1.6507
ln_lambda_pIIB_cIIB	Log rate of clinical presentation at stage IIb	-2.1362
ln_lambda_pIIIA_cIIIA	Log rate of clinical presentation at stage IIIa	-1.4088
ln_lambda_pIIIB_cIIIB	Log rate of clinical presentation at stage IIIb	-0.8811
ln_lambda_pIV_cIV	Log rate of clinical presentation at stage IV	-1.4027
Disease survival parameters		
lambda_lcs_Ia	Lambda constant for survival if diagnosed and treated from stage Ia	0.214

lambda_lcs_Ib	Lambda constant for survival if diagnosed and treated from stage Ib	0.274
lambda_lcs_IIa	Lambda constant for survival if diagnosed and treated from stage IIa	0.330
lambda_lcs_IIb	Lambda constant for survival if diagnosed and treated from stage IIb	0.475
lambda_lcs_IIIa	Lambda constant for survival if diagnosed and treated from stage IIIa	0.588
lambda_lcs_IIIb	Lambda constant for survival if diagnosed and treated from stage IIIb	0.909
lambda_lcs_IV	Lambda constant for survival if diagnosed and treated from stage IV	1.423
gamma_lcs_all_stages	Gamma constant for survival if diagnosed and treated at any stage	0.676
lambda_ocm_F	Gompertz distribution lambda parameter for other cause mortality in the female population	0.00019
gamma_ocm_F	Gompertz distribution gamma parameter for other cause mortality in the female population	0.1018
lambda_ocm_M	Gompertz distribution lambda parameter for other cause mortality in the male population	0.00059
gamma_ocm_M	Gompertz distribution gamma parameter for other cause mortality in the male population	0.0917
Risk prediction parameters		
risk_age	Risk prediction coefficient for age (years)	0.08985
risk_male	Risk prediction coefficient for male sex	0.30562
risk_smoker	Risk prediction coefficient for current/former smoker (vs never smoker)	1.45929
risk_lungcancer	Risk prediction coefficient for presence of lung cancer at baseline or within 3 years.	0.33488
risk_intercept	Risk prediction intercept	-11.39758
risk_SD	Risk prediction standard deviation (error term)	0.62920
Screening effectiveness		
sens_LDCT	Sensitivity of low dose CT screening for the detection of lung cancer	0.709
spec_LDCT	Specificity of low dose CT screening for the detection of lung cancer	0.624
sens_AABT	Sensitivity of autoantibody test screening for the detection of lung cancer	0.400
spec_AABT	Specificity of autoantibody test screening for the detection of lung cancer	0.900
mu_ind_scrn_delay	Mean time to screening after model entry	-2.823
sig_ind_scrn_delay	standard deviation of time to screening after model entry	0.820
Quality of life parameters		
u_base_male	Utility of male smoker in the UK general population (including with occult lung cancer)	0.7816
u_base_female	Utility of female smoker in the UK general population (including with occult lung cancer)	0.7531
u_dis_sII	Disutility of second stage cancer vs first stage	-0.04
u_dis_sIII	Disutility of third stage cancer vs first stage	-0.04

u_dis_sIV	Disutility of fourth stage cancer vs first stage	-0.05
u_dis_fp	Disutility associated with a false-positive screening result	-0.063
u_dis_scr_anxiety	Disutility associated with anxiety of a screening event	-0.010
t_dis_fp	Duration of disutility from false-positive screen	3.00
t_dis_scrn_anx	Duration of disutility from anxiety of a screening event	2.00
Costs		
c_invite	Cost of initial invite and questionnaire	£2.90
c_score	Cost of scoring questionnaire and risk stratification	£18.54
c_letter	Cost of follow-up letter and screening appointment	£1.74
c_gp_ref	Cost of GP consultations leading to lung cancer referral	£72.00
c_LDCT	Cost of low-dose CT scan	£98.80
c_AABT	Cost of autoantibody test	£60.00
c_scrn_nurse	Cost of nurse-led screening consultation	£6.25
c_false_pos	Cost of resourcing following a false-positive screen	£184.63
c_eol_lung	Cost of end-of-life care for lung cancer patient	£4589.04
c_rdtf_sIa_ini	Cost of initial diagnosis and treatment if diagnosed at stage Ia	£5558.14
c_rdtf_sIb_ini	Cost of initial diagnosis and treatment if diagnosed at stage Ib	£6411.63
c_rdtf_sIIa_ini	Cost of initial diagnosis and treatment if diagnosed at stage IIa	£7279.07
c_rdtf_sIIb_ini	Cost of initial diagnosis and treatment if diagnosed at stage IIb	£6558.14
c_rdtf_sIIIa_ini	Cost of initial diagnosis and treatment if diagnosed at stage IIIa	£6511.63
c_rdtf_sIIIb_ini	Cost of initial diagnosis and treatment if diagnosed at stage IIIb	£6046.51
c_rdtf_sIV_ini	Cost of initial diagnosis and treatment if diagnosed at stage IV	£5441.86
c_rdtf_sIa_re_ind_yr	Cost of index year diagnosis, treatment and follow-up if diagnosed at stage Ia	£5848.11
c_rdtf_sIb_re_ind_yr	Cost of index year diagnosis, treatment and follow-up if diagnosed at stage Ib	£5359.21
c_rdtf_sIIa_re_ind_yr	Cost of index year diagnosis, treatment and follow-up if diagnosed at stage IIa	£5637.60
c_rdtf_sIIb_re_ind_yr	Cost of index year diagnosis, treatment and follow-up if diagnosed at stage IIb	£6514.78
c_rdtf_sIIIa_re_ind_yr	Cost of index year diagnosis, treatment and follow-up if diagnosed at stage IIIa	£5415.46
c_rdtf_sIIIb_re_ind_yr	Cost of index year diagnosis, treatment and follow-up if diagnosed at stage IIIb	£4318.07
c_rdtf_sIV_re_ind_yr	Cost of index year diagnosis, treatment and follow-up if diagnosed at stage IV	£2787.31
c_rdtf_sIa_subyrs	Cost of treatment and follow-up in subsequent years if diagnosed at stage Ia	£1437.79
c_rdtf_sIb_subyrs	Cost of treatment and follow-up in subsequent years if diagnosed at stage Ib	£1483.75
c_rdtf_sIIa_subyrs	Cost of treatment and follow-up in subsequent years if diagnosed at stage IIa	£1628.19

c_rdtf_sIIb_subyrs	Cost of treatment and follow-up in subsequent years if diagnosed at stage IIb	£1647.88
c_rdtf_sIIIa_subyrs	Cost of treatment and follow-up in subsequent years if diagnosed at stage IIIa	£1503.45
c_rdtf_sIIIb_subyrs	Cost of treatment and follow-up in subsequent years if diagnosed at stage IIIb	£1306.49
c_rdtf_sIV_subyrs	Cost of treatment and follow-up in subsequent years if diagnosed at stage IV	£1037.31

12.4 Appendix 6A: Pearson/df ratio values for transformation strategies for all features

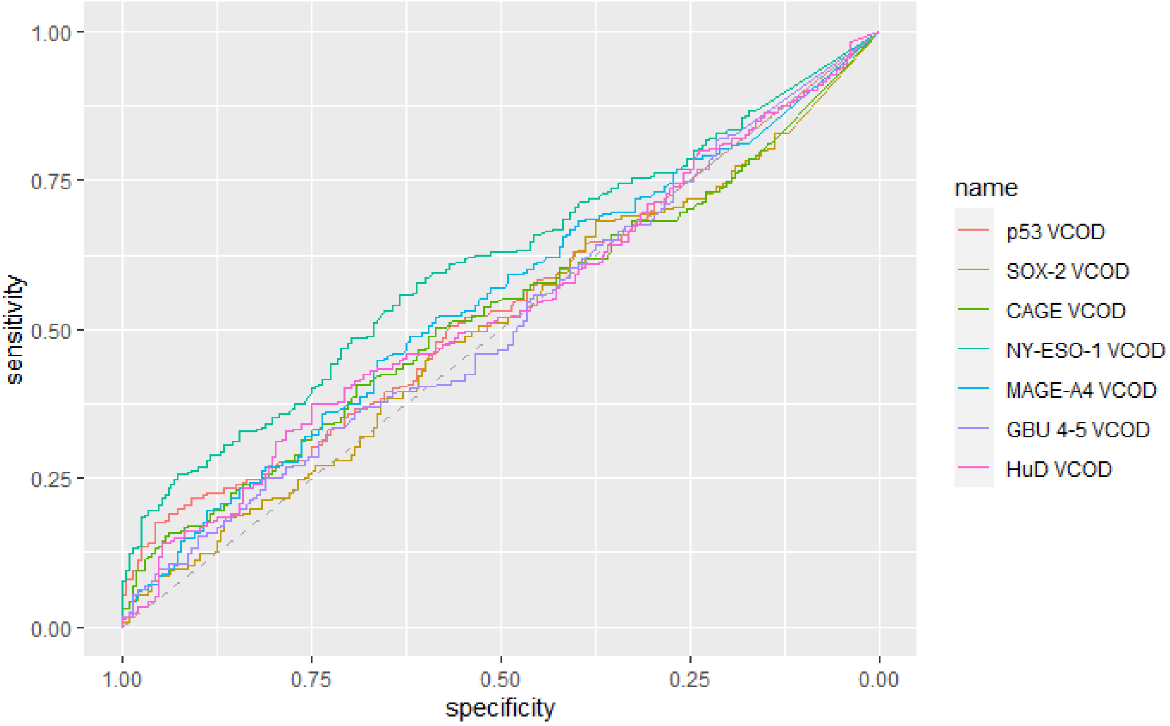
Feature	Antigen	Chosen Transform	log	sqrt	exp	arcsinh	box cox	yeo-johnson
OD	p53	boxcox	1.743	3.353	11.556	5.876	1.093	1.644
OD	SOX_2	boxcox	1.560	2.958	8.404	4.959	0.976	1.581
OD	CAGE	boxcox	1.411	2.530	6.422	3.958	1.114	1.533
OD	NY_ESO_1	boxcox	1.851	4.262	16.116	7.046	0.986	1.486
OD	GBU_4_5	boxcox	0.940	1.576	4.589	2.340	0.918	0.986
OD	MAGE_A4	boxcox	1.378	2.682	7.234	4.223	1.043	1.282
OD	HuD	boxcox	1.354	2.469	8.252	3.962	1.060	1.311
od_intercept	p53	boxcox	1.855	3.329	7.976	5.803	0.985	2.079
od_intercept	SOX_2	boxcox	1.936	3.425	7.934	5.811	1.259	2.513
od_intercept	CAGE	boxcox	1.888	3.235	7.220	5.677	1.185	2.543
od_intercept	NY_ESO_1	boxcox	2.380	4.408	11.338	7.113	1.390	2.883
od_intercept	GBU_4_5	boxcox	1.416	2.557	6.822	4.636	0.900	1.515
od_intercept	MAGE_A4	boxcox	1.723	3.134	7.293	5.298	1.132	2.073
od_intercept	HuD	boxcox	1.779	3.005	6.966	5.706	0.988	2.259
od_slope	p53	log_x	1.783	3.168		14.907		5.180
od_slope	SOX_2	log_x	1.534	2.363		13.683		4.511
od_slope	CAGE	log_x	1.760	2.604		18.625		4.704
od_slope	NY_ESO_1	log_x	2.612	4.001		18.739		6.770
od_slope	GBU_4_5	log_x	1.166	1.534		8.669		2.360
od_slope	MAGE_A4	log_x	2.689	2.842		25.088		3.817
od_slope	HuD	log_x	2.136	2.988		18.520		4.390
od_auc	p53	arcsinh_x	3.502	3.802	3.502	3.502		3.502
od_auc	SOX_2	arcsinh_x	4.036	4.059	4.036	4.036		4.036
od_auc	CAGE	arcsinh_x	3.689	4.626	3.689	3.689		3.689
od_auc	NY_ESO_1	arcsinh_x	3.472	4.347	3.472	3.472		3.472
od_auc	GBU_4_5	arcsinh_x	1.973	2.628	1.973	1.973		1.973
od_auc	MAGE_A4	arcsinh_x	2.984	3.589	2.984	2.984		2.984
od_auc	HuD	arcsinh_x	2.435	3.253	2.435	2.435		2.435
od_slopemax	p53	log_x	1.288	5.833		2.246		15.157
od_slopemax	SOX_2	log_x	1.259	4.371		2.439		13.678
od_slopemax	CAGE	log_x	1.117	3.526		2.038		11.149

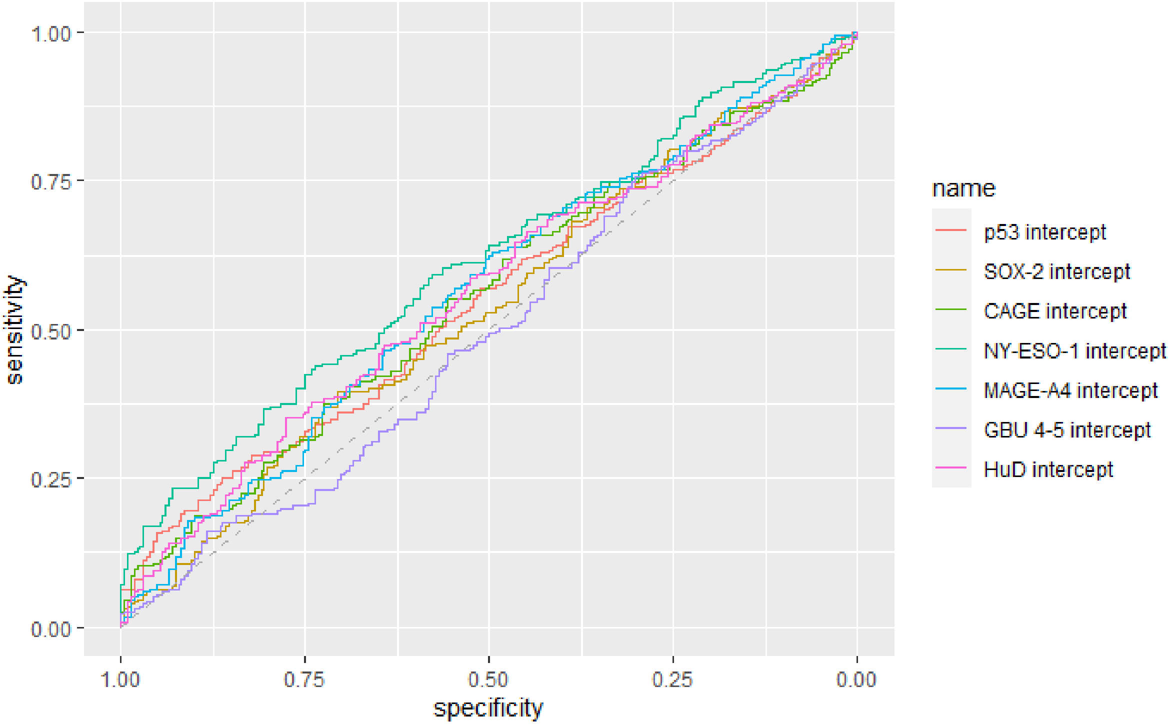
Feature	Antigen	Chosen Transform	log	sqrt	exp	arcsinh	box cox	yeo-johnson
od_slopemax	NY_ESO_1	log_x	1.577	8.129		4.343		19.017
od_slopemax	GBU_4_5	arcsinh_x	1.014	2.951		1.014	1.131	1.129
od_slopemax	MAGE_A4	log_x	1.115	3.211		5.309		7.566
od_slopemax	HuD	arcsinh_x	1.145	3.578		1.145	1.266	1.266
VCOD	p53	log_x	2.230	3.217	17.844	12.309		3.010
VCOD	SOX_2	sqrt_x	2.979	2.680	12.447	7.657		3.280
VCOD	CAGE	sqrt_x	2.144	1.975	9.733	6.884		2.335
VCOD	NY_ESO_1	log_x	2.783	4.348	26.387	15.368		2.789
VCOD	GBU_4_5	yeojohnson	1.716	1.490	5.733	3.280		1.172
VCOD	MAGE_A4	sqrt_x	2.086	2.048	9.156	6.165		2.054
VCOD	HuD	yeojohnson	2.032	2.513	11.321	6.869		1.762
vcod_intercept	p53	log_x	2.243	6.312	18.527	15.902		7.568
vcod_intercept	SOX_2	log_x	3.639	5.752	10.475	9.370		6.428
vcod_intercept	CAGE	log_x	1.973	3.548	8.721	8.383		5.636
vcod_intercept	NY_ESO_1	log_x	4.385	11.255	25.363	19.256		9.847
vcod_intercept	GBU_4_5	log_x	1.150	2.348	7.447	6.111		3.023
vcod_intercept	MAGE_A4	log_x	1.233	2.789	8.314	7.668		4.523
vcod_intercept	HuD	log_x	1.491	3.178	7.386	6.864		4.034
vcod_slope	p53	log_x	1.826	3.285		18.894		5.202
vcod_slope	SOX_2	log_x	1.469	2.053		20.382		4.234
vcod_slope	CAGE	log_x	1.486	2.542		18.525		4.873
vcod_slope	NY_ESO_1	log_x	1.975	3.567		21.348		6.363
vcod_slope	GBU_4_5	sqrt_x	1.965	1.301		7.327		1.766
vcod_slope	MAGE_A4	log_x	1.225	2.026		23.331		3.028
vcod_slope	HuD	log_x	1.596	2.204		21.075		3.827
vcod_auc	p53	arcsinh_x	2.933	3.148	2.933	2.933		2.933
vcod_auc	SOX_2	arcsinh_x	3.547	3.799	3.547	3.547		3.547
vcod_auc	CAGE	arcsinh_x	3.205	3.872	3.205	3.205		3.205
vcod_auc	NY_ESO_1	arcsinh_x	2.902	4.293	2.902	2.902		2.902
vcod_auc	GBU_4_5	arcsinh_x	1.648	2.200	1.648	1.648		1.648
vcod_auc	MAGE_A4	arcsinh_x	2.696	2.776	2.696	2.696		2.696
vcod_auc	HuD	arcsinh_x	2.491	2.795	2.491	2.491		2.491

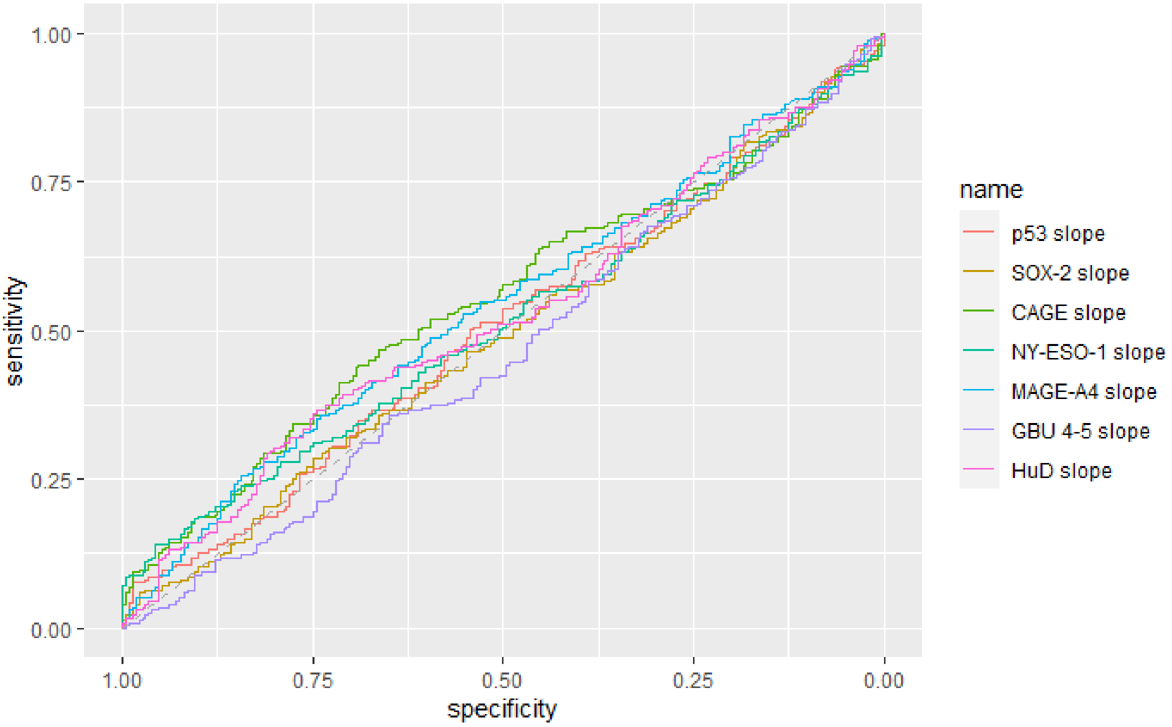
Feature	Antigen	Chosen Transform	log	sqrt	exp	arcsinh	box cox	yeo-johnson
vcod_slope max	p53	yeojohnson	4.000	5.567		2.725		1.801
vcod_slope max	SOX_2	yeojohnson	5.301	3.776		3.587		1.897
vcod_slope max	CAGE	yeojohnson	3.219	3.399		2.360		1.448
vcod_slope max	NY_ESO_1	yeojohnson	3.424	6.397		2.406		1.913
vcod_slope max	GBU_4_5	arcsinh_x	1.235	3.051		1.235	1.242	1.241
vcod_slope max	MAGE_A4	yeojohnson	5.990	2.662		4.037		1.361
vcod_slope max	HuD	yeojohnson	2.677	2.806		2.007		1.329
STVR	p53	yeojohnson	3.317	6.128	45.189	4.069	1.825	1.541
STVR	SOX_2	yeojohnson	3.251	4.956	29.704	3.789	2.708	2.330
STVR	CAGE	yeojohnson	2.707	4.565	32.740	3.292	1.784	1.598
STVR	NY_ESO_1	yeojohnson	4.928	9.928	48.046	5.988	2.162	1.839
STVR	GBU_4_5	boxcox	1.269	2.028	41.121	1.421	1.003	1.051
STVR	MAGE_A4	yeojohnson	1.850	2.898	31.204	2.272	1.262	1.233
STVR	HuD	yeojohnson	2.208	3.959	32.826	2.762	1.343	1.238
stvr_intercept	p53	yeojohnson	7.061	10.359	37.513	8.597	3.363	1.892
stvr_intercept	SOX_2	yeojohnson	4.766	6.370	14.854	5.732	3.377	2.925
stvr_intercept	CAGE	yeojohnson	3.461	4.325	11.005	4.001	2.362	1.731
stvr_intercept	NY_ESO_1	yeojohnson	9.698	15.309	43.729	12.247	4.168	1.816
stvr_intercept	GBU_4_5	yeojohnson	3.022	4.136	19.901	3.566	1.683	1.409
stvr_intercept	MAGE_A4	yeojohnson	2.847	3.652	10.482	3.381	1.967	1.409
stvr_intercept	HuD	yeojohnson	2.960	4.207	11.919	3.613	1.715	1.351
stvr_slope	p53	log_x	2.304	3.539		21.206		5.589
stvr_slope	SOX_2	log_x	1.478	2.186		21.302		4.119
stvr_slope	CAGE	log_x	1.706	2.935		19.453		5.940
stvr_slope	NY_ESO_1	log_x	1.938	3.317		23.148		6.555
stvr_slope	GBU_4_5	log_x	1.104	1.502		11.714		2.318
stvr_slope	MAGE_A4	log_x	1.235	1.895		24.437		2.861
stvr_slope	HuD	log_x	1.319	1.553		19.262		3.005
stvr_auc	p53	arcsinh_x	3.515	4.018	3.515	3.515		3.515
stvr_auc	SOX_2	sqrt_x	2.865	2.481	2.860	2.860		2.860
stvr_auc	CAGE	sqrt_x	3.255	2.807	3.255	3.255		3.255

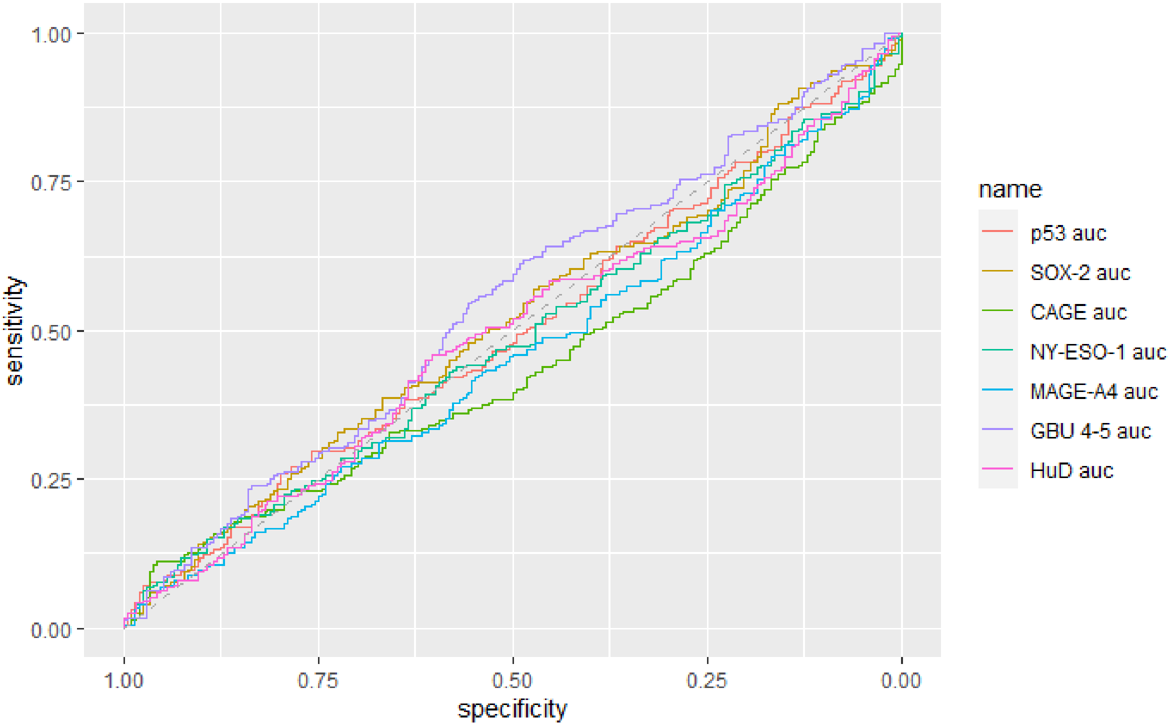
Feature	Antigen	Chosen Transform	log	sqrt	exp	arcsinh	box cox	yeo-johnson
stvr_auc	NY_ESO_1	arcsinh_x	3.789	4.688	3.788	3.788		3.788
stvr_auc	GBU_4_5	arcsinh_x	1.919	2.194	1.919	1.919		1.919
stvr_auc	MAGE_A4	arcsinh_x	2.148	2.481	2.148	2.148		2.148
stvr_auc	HuD	arcsinh_x	2.255	2.703	2.255	2.255		2.255
stvr_slopemax	p53	boxcox	1.333	4.763		1.333	1.291	1.291
stvr_slopemax	SOX_2	boxcox	1.275	2.873		1.275	1.259	1.259
stvr_slopemax	CAGE	log_x	1.165	2.626		3.971		6.882
stvr_slopemax	NY_ESO_1	arcsinh_x	1.439	5.507		1.439	1.513	1.518
stvr_slopemax	GBU_4_5	arcsinh_x	1.153	2.371		1.153	1.192	1.192
stvr_slopemax	MAGE_A4	log_x	0.982	2.082		2.481		8.977
stvr_slopemax	HuD	log_x	1.367	2.202		2.947		8.846

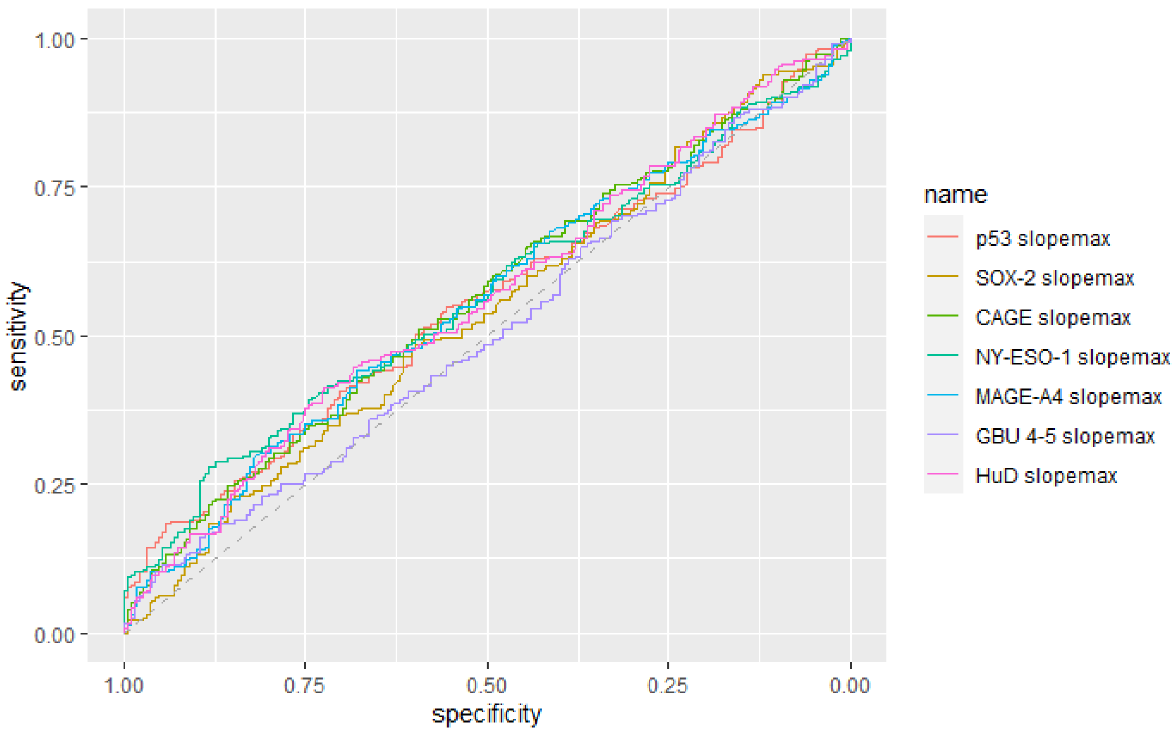
12.5 Appendix 7A: ROC Plots of VOL corrected magnitude and curve characteristic features.



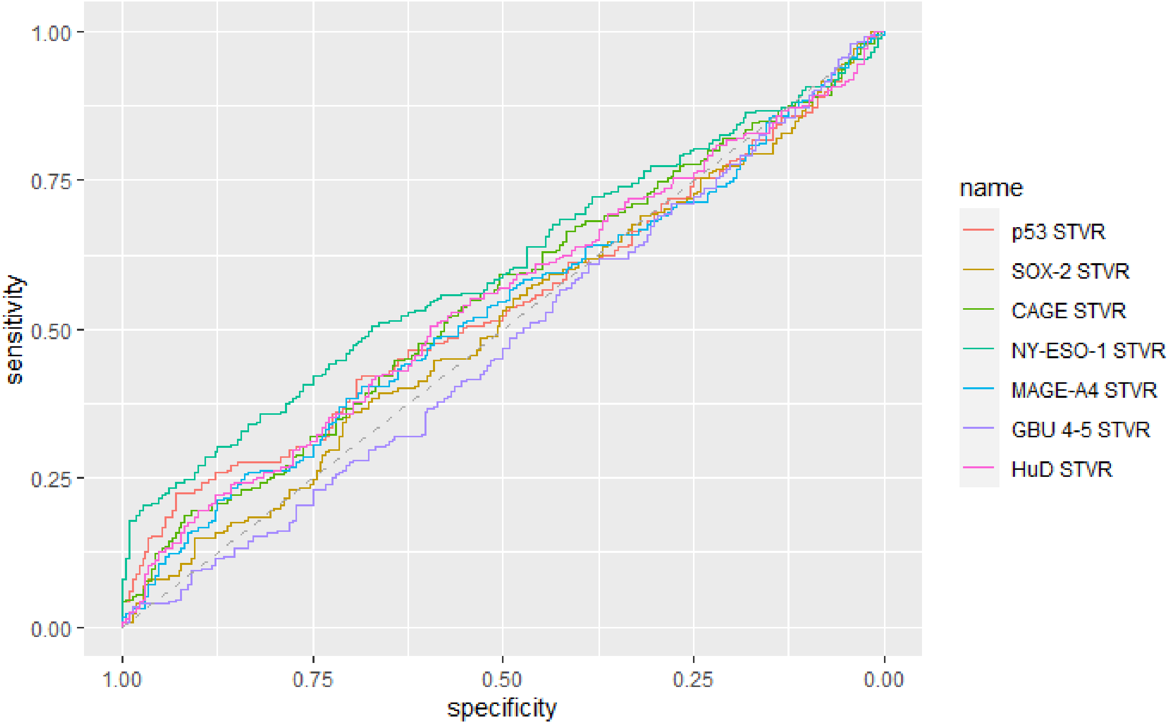


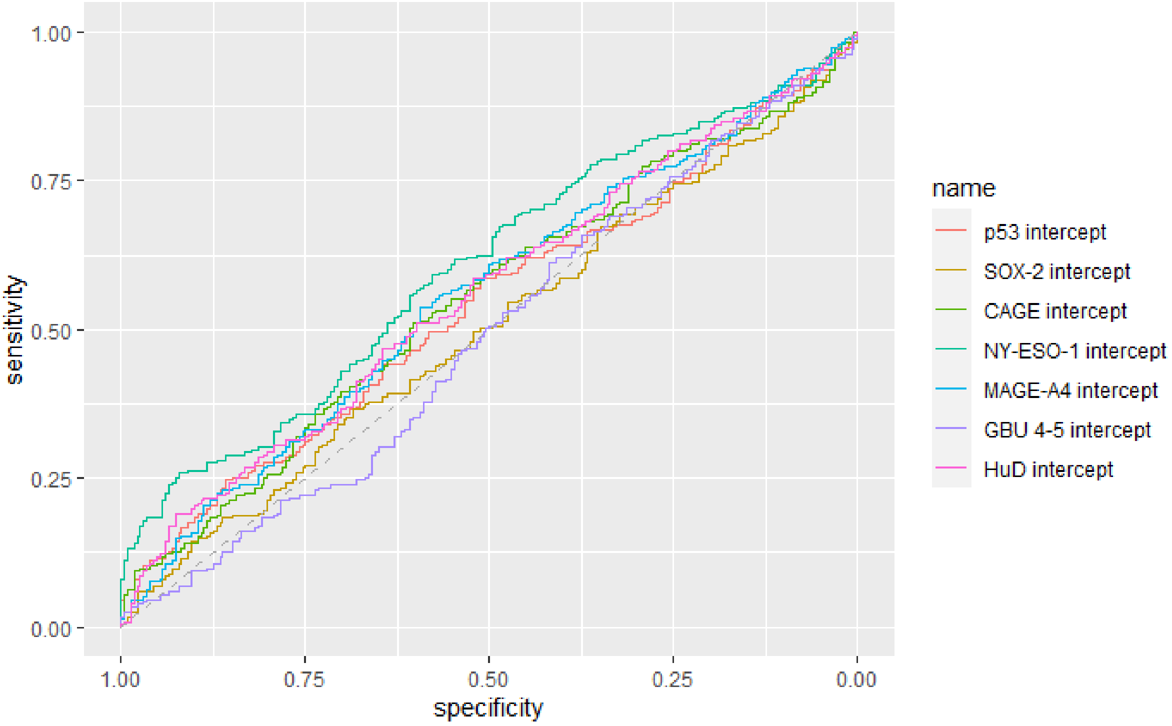

1.00
0.75
0.50
0.25
0.00
sensitivity
1.00
0.75
0.50
0.25
0.00
specificity
name
p53 slope
SOX-2 slope
CAGE slope
NY-ESO-1 slope
MAGE-A4 slope
GBU 4-5 slope
HuD slope

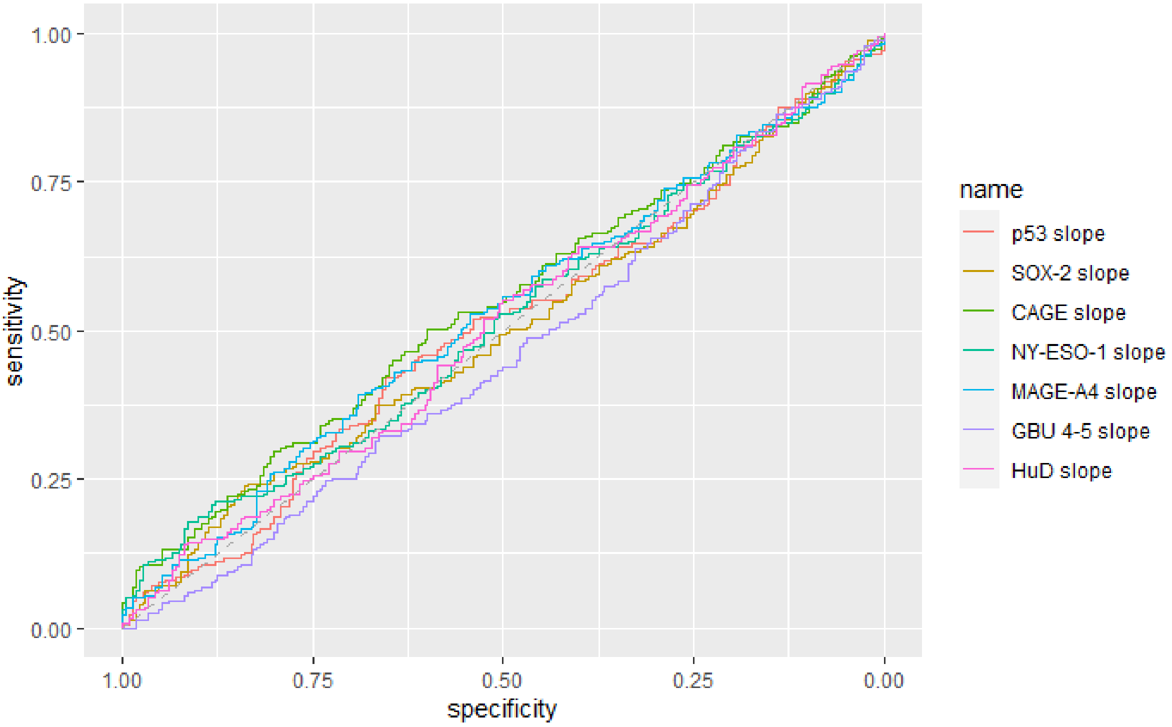

1.00
0.75
0.50
0.25
0.00
sensitivity
1.00
0.75
0.50
0.25
0.00
specificity
name
p53 auc
SOX-2 auc
CAGE auc
NY-ESO-1 auc
MAGE-A4 auc
GBU 4-5 auc
HuD auc

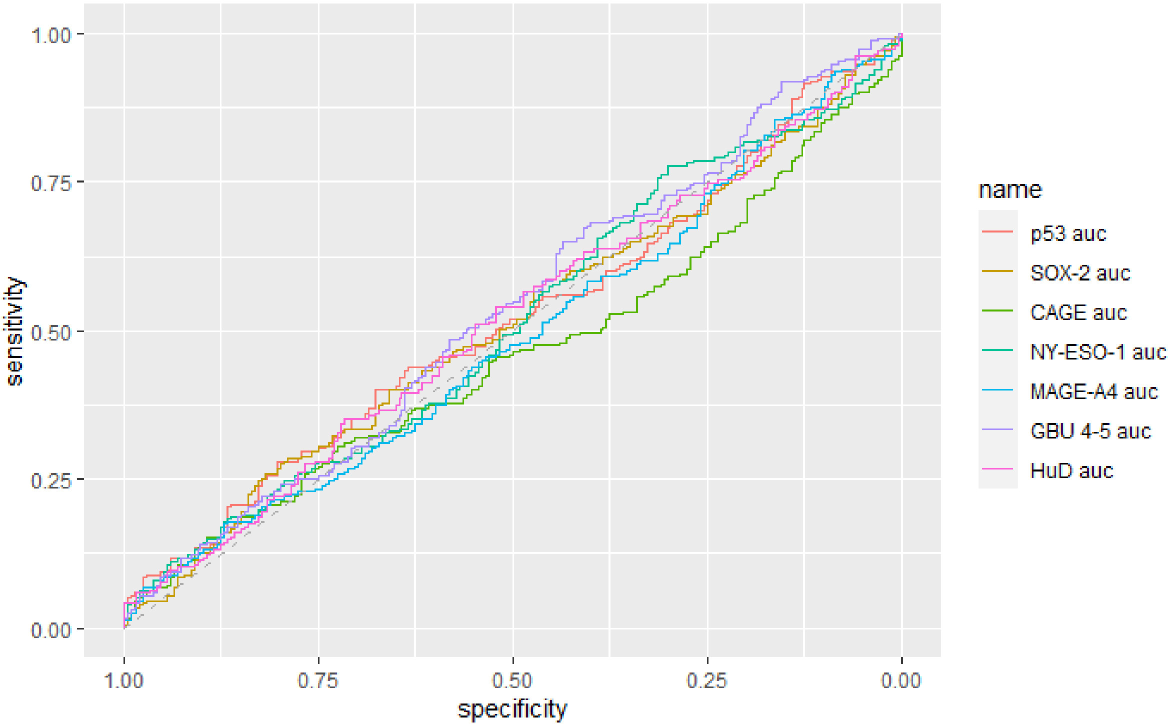

1.00
0.75
0.50
0.25
0.00
sensitivity
1.00
0.75
0.50
0.25
0.00
specificity
name
p53 slopemax
SOX-2 slopemax
CAGE slopemax
NY-ESO-1 slopemax
MAGE-A4 slopemax
GBU 4-5 slopemax
HuD slopemax

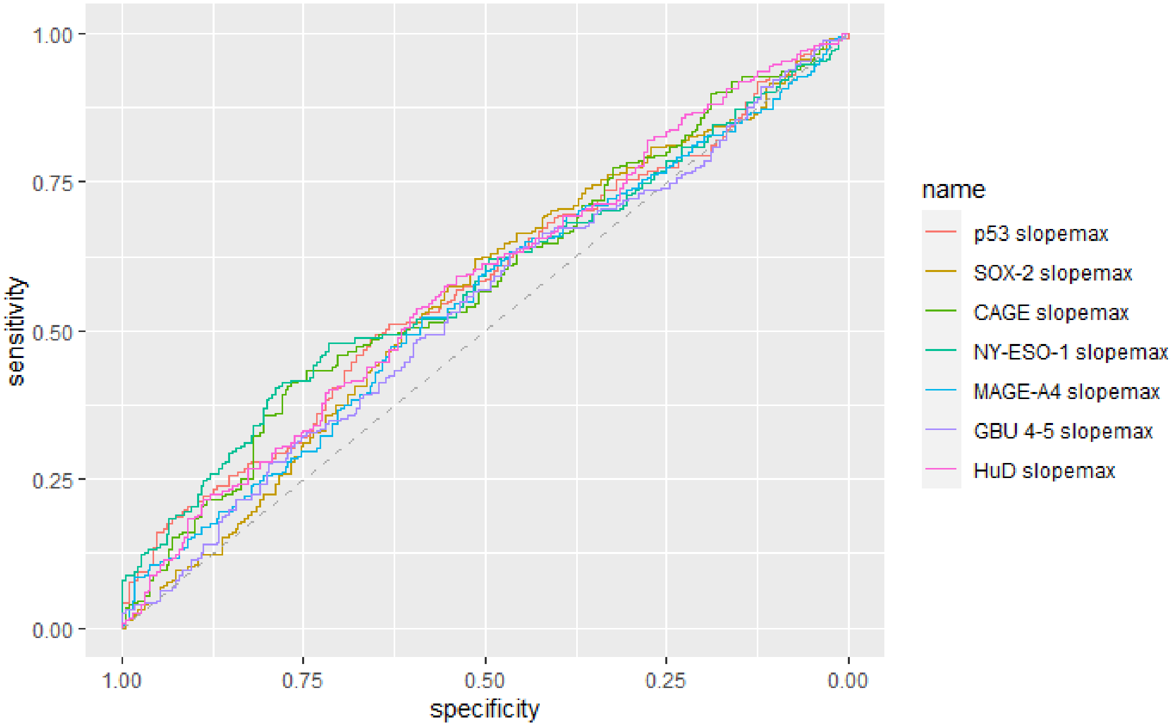
12.6 Appendix 7B: ROC Plots of Signal to Vol Ratio magnitude and curve characteristic features.






1.00
0.75
0.50
0.25
0.00
sensitivity
1.00
0.75
0.50
0.25
0.00
specificity
name
p53 slope
SOX-2 slope
CAGE slope
NY-ESO-1 slope
MAGE-A4 slope
GBU 4-5 slope
HuD slope


1.00
0.75
0.50
0.25
0.00
sensitivity
1.00
0.75
0.50
0.25
0.00
specificity
name
p53 auc
SOX-2 auc
CAGE auc
NY-ESO-1 auc
MAGE-A4 auc
GBU 4-5 auc
HuD auc


1.00
0.75
0.50
0.25
0.00
sensitivity
1.00
0.75
0.50
0.25
0.00
specificity
name
p53 slopemax
SOX-2 slopemax
CAGE slopemax
NY-ESO-1 slopemax
MAGE-A4 slopemax
GBU 4-5 slopemax
HuD slopemax

12.7 Conference Attendance and Publications

12.7.1 Conferences Attended:

Immuno-Oncology Summit Boston 2017

Public Health England: Cancer Data and Outcomes Conference 2017

CRUK Cambridge Centre: Early Detection International Summer School 2019

Cancer Research Nottingham Symposium 2020

Biodata World Congress 2020

12.7.2 Publications:

Due to commercial sensitivity of the data and results, unfortunately my ability to publish results was limited.

Manuscript in preparation for submission to PLOS One: Lung Cancer Associated Autoantibody Responses are Detectable Years Before Clinical Presentation. Allen, J., Healey, G., Macdonald I.

Manuscript in preparation: A Simulated Annealing Gradient Ascent Algorithm for Optimisation of High Specificity Multivariate Autoantibody Panels for Early Lung Cancer Detection. Allen, J. Healey G., Chapman, C., Grainge, M.