UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences

School for Information Technology

Master of Statistics

Master's thesis

Improvement of risk models to select individuals eligible for lung cancer screening: adding the metabolic phenotype

Ekiti Ekote

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics, specialization Biostatistics

SUPERVISOR :

Prof. dr. Ziv SHKEDY

Mevrouw Olajumoke Evangelina OWOKOTOMO

2018
2019

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

Faculty of Sciences

School for Information Technology

Master of Statistics

Master's thesis

Improvement of risk models to select individuals eligible for lung cancer screening: adding the metabolic phenotype

Ekiti Ekote

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics, specialization Biostatistics

SUPERVISOR :

Prof. dr. Ziv SHKEDY

Mevrouw Olajumoke Evangelina OWOKOTOMO

Acknowledgements

There are a few people I would like to thank for helping me during my master thesis journey. First, I would like to thank my supervisor Prof. Dr. Ziv Shkedy, for his mentorship, dedication, feedback and undiluted support throughout my thesis journey. He always answered present when I needed help. Secondly, I would like to sincerely thank my second supervisor Olajumoke Evangelina Owokotomo, for her invaluable insights and encouragement during this academically trying time in my life. She taught and guided me, and devoted an incredible amount of her time to understand my worries. Her R tips provided me with exceptional relief.

My unfeigned appreciation also goes to my entire family, especially my brothers, Ngipe Ekote Ramsay and Rene Ekoteson, my sister, Ekote Doris Mbulle, and my uncle, Nelson Ebwelleson for their prayers, encouragements and sponsorship.

I am profoundly thankful to my lecturers at Hasselt University and all CENSTAT members for the academy, scientific and moral mentorship they provided me all these years. Finally, I deeply thank all my friends and every other person who in one way or the other helped me move a step forward in this journey.

Abstract

Low-dose computer tomography is considered as the test offering the highest potential for clinically effective and cost-effective screening for lung cancer since several potential screening tests such as chest x-ray and sputum cytology, for lung cancer have failed to demonstrate effectiveness in randomised controlled trials.Classical risk models used for the selection of high-risk individuals qualified for lung cancer screening with low-dose computed tomography only take clinical risk parameters into account. The aim of this study is to examine whether this selection can be improved by adding parameters that reflect the plasma metabolic phenotype.

Predictive risk models were developed, tested and validated with cross validation using the penalized logistic regression model, the LASSO. One important aspect of the analysis was to test if the metabolic information is relevant (in addition to the clinical variables). The analysis was implemented using the R package "glmnet" which is available on CRAN. A user-defined R function was also used to implement the cross validation.

The results revealed that the selection of high-risk individuals eligible for lung cancer screening improves when metabolic information is considered. Therefore, in addition to clinical risk factors, metabolic information is relevant. However, not all metabolic features may be admissible.

Key Words: Lung cancer screening, The LASSO, High dimensional data, Classification, Cross validation

Contents

1 Introduction

1.1 Lung Cancer

Snowsill et al.(2018), defined lung cancer as: ”malignant growth of cells in the lung”. Lung cancer is generally common in elderly, present or former smokers. The likelihood of getting lung cancer is higher in males than in females, and males are at a higher risk of dead from lung cancer (Snowsill et al., 2018). Lung cancer has been the most prevalent type of cancer overall for several decades and also the most common cause of cancer related deaths in men and women worldwide (International Agency for Research on Cancer, 2018; Snowsill et al., 2018). In fact, according to Bach et al.(2012), lung cancer related deaths are as many as those from the next four most deadly cancers (breast, prostate, colon, and pancreas) combined. Most patients are diagnosed with advanced disease, resulting in a very low 5-year survival (Bach et al., 2012). The risk of lung cancer reduces when smoking is stopped. However, if there is a significant smoking history, formal smokers still remain at higher risk compared to those who have never smoked (Pinsky et al., 2015).

Even with recent advances in oncology and surgery, the likely development or course of diagnosed lung cancer cases is disheartening . This is because lung cancer is generally diagnosed when the cancer is in the later stages, and in elderly people, who often have concomitant diseases that tend to restrict therapeutic possibilities (Snowsill et al., 2018). Lung cancer therapy depends on the cancer location within the lung, the tumour size, whether or not it has extended and how far it has extended, and the overall health and fitness of the individual presenting. According to Snowsill et al.(2018), the main treatment options are chemotherapy, radiotherapy, surgery, chemoradiotherapy, control of symptoms and palliative care.

1.1.1 Lung Cancer Screening with Low-Dose Computed Tomography (LDCT)

According to the Commission of Chronic Illness,(1957), screening is defined as ”the presumptive identification of unrecognized disease or defect by the application of tests, examinations, or other procedures which can be applied rapidly”. Screening tests recognizes seemingly well individuals who presumably have a disease from those who perhaps do not. However, a screening test is not preconceived to be diagnostic (Commision of Chronic Illness, 1957). Therefore, people with positive or suspicious outcomes must be referred to their physicians for diagnosis and necessary treatment. Screening, otherwise known as early disease detection or case-finding is carried out mainly to discover and cure conditions which have already produced pathological change even though it have not so far reached a stage of urgent medical aid (Wilson and Jungner, 1968). Without screening, the usual clinical course for lung cancer will examine individuals who present symptoms such as persistent cough, haemoptysis or persistent breathlessness (Snowsill et al., 2018).

For some decades now, several potential screening tests such as chest x-ray and sputum cytology, for lung cancer have been examined. However, the effectiveness of neither of these tests have been demonstrated in randomised controlled trials. The development of computer tomography(CT) scanning with its ability to offer increasingly high quality images at lower radiation dosage has made it to be considered as the test offering the highest potential for clinically effective and cost-effective screening for lung cancer. Much research is however devoted to investigating whether or not this is the case (Manser et al., 2013). Computed tomography scanning technology uses computer-processed combinations of several X-ray images taken from different angles to generate cross-sectional (tomographic) images of particalar portions of a scanned object, enabling the user to see inside without cutting (Snowsill et al., 2018). According to Snowsill et al.(2018), an important issue is that LDCT screening for lung cancer is not a homogenous technology, so they recommend that careful attention be paid to the exact nature of the device being used, the procedure being used and precise criteria being employed to define an abnormality as potentially malignant, benign or indeterminate.

1.2 Additional Predictive Value

Nowadays, as a result of technical developments in the field and growing biological knowledge, novel omic measures, such as NMR (Nuclear Magnetic Resonance) proteomics are emerging as potentially powerful new biomolecular marker sets. Due to this, it is increasingly common for studies to collect a variety of omic measurements in the same set of individuals, using different measurement platforms and spanning different aspects of human biology (Rodriguez-Girondo et al., 2018).

In spite the fact that these molecular data have been used for disease outcome prediction or diagnosis purposes for more than a decade in biomedical research, the question of the added predictive value of such data given that classical clinical predictors are already available have not been given sufficient attention in the bioinformatics literature (Boulesteix and Hothorn, 2010; Boulesteix and Sauerbrei, 2011). According to Boulesteix and Hothorn, 2010; Boulesteix and Sauerbrei, (2011), a particular drawback from a statistical point of view is that these molecular predictors are often high-dimensional, which potentially leads to overfitting problems and overoptimistic conclusions on their additional predictive power. Moreover, getting meaningful summary measures and valid statistical procedures for testing the added predictive value are difficult tasks, even when considering the addition of a single additional biomarker in the classical regression context. As a result, investigating the added predictive ability of new biomarkers regarding classical, low dimensional, settings has been a topic of intense debate in the biostatistical literature in the last years (Rodriguez-Girondo et al., 2018).

Until date, not much attention has been given to the evaluation of the added predictive ability in high-dimensional settings, where the aforementioned issues are more pronounced and new ones surface, such as the simultaneous inclusion in a unique prediction model of predictors sets of very different nature (Rodriguez-Girondo et al., 2018). However, a few authors in the like of Boulesteix and Hothorn, (2010); Boulesteix and Sauerbrei, (2011); Goeman et al.(2004); Tibshirani and Efron, (2002); Hoefling and Tibshirani, (2008) have publications suggesting procedures to handle these problems.

1.3 Aims and Objectives

Classical risk models used for the selection of high-risk individuals eligible for lung cancer screening with low-dose computed tomography (LDCT) only take clinical risk parameters into account. The aim of this study is to examined whether this selection can be improved by adding parameters that reflect the plasma metabolic phenotype. In particular, the current study seeks to answer the following questions;

1. Given that clinical risk factors are already available, is the metabolic information relevant?

2. Is the selection of high-risk individuals eligible for lung cancer screening improved by adding metabolic information?

2 Data

The data analysed in this study was obtained from a case control study with 536 subjects. Individuals with no symptoms of lung cancer were subjected to lung cancer screening using low-dose computed tomography(LDCT). Those in whom lung cancer was detected were marked as cases (273) while the others, without lung cancer detection were included in the study as healthy controls (263). The main interest was to investigate whether an individual's lung cancer status could be predicted. For all subjects, conventional clinical risk factors for lung cancer were recorded at baseline. Meanwhile, the metabolic activity of individuals included in the study was also measured through blood samples collected at baseline. The data structure therefore, consist of a 536 x 1 vector of binary reponse(**Y**), a 536 x 14 matrix of clinical risk factors(**Z**) and a 536 x 102 matrix of metabolic features (**X**) as shown below.

$$\begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ Y_{536} \end{bmatrix} \qquad \begin{bmatrix} Z_{1,1} & Z_{2,1} & . & . & Z_{14,1} \\ Z_{1,2} & Z_{2,2} & . & . & Z_{14,2} \\ . & . & . & . & . \\ . & . & . & . & . \\ Z_{1,536} & Z_{2,536} & . & . & Z_{14,536} \end{bmatrix} \qquad \begin{bmatrix} X_{1,1} & X_{2,1} & . & . & X_{102,1} \\ X_{1,2} & X_{2,2} & . & . & X_{102,2} \\ . & . & . & . & . \\ . & . & . & . & . \\ X_{1,536} & X_{2,536} & . & . & X_{102,536} \end{bmatrix}$$

In this study, the combination of the binary response(**Y**) and the matrix of clinical risk factors (**Z**) will be referred to as "clinical data", the combination of the binary response(**Y**) and the matrix of metabolic features (**X**) will be referred to as "metabolic data", and the combination of the binary response(**Y**), the matrix of clinical risk factors (**Z**) and the matrix of metabolic features (**X**) will be referred to as "combined data". The clinical data consist of 14 variables (both continuous and categorical), while the metabolic data contains 102 continuous metabolic features.

3 Methods

This section is organized as follows; In the first place, a gentle introduction about logistic regression model is given, with a close attention on stepwise logistic regression for the outcome variable and clinical risk factors (see section 3.1), then we move a step forward to describe the so called univariate analysis for the metabolic features. The term univeriate analysis here refers to a logistic regression model with all clinical risk factors and a single metabolic variable at a time (see section 3.2). In section 3.3, we outline and elaborate on three different tests for additional predictive value for high-dimensional molecular data. Finally we spell out the penalized logistic regression technique (the LASSO), particularly, classification using this tool and also provide motivations for these statistical methods.

3.1 Logistic Regression

Logistic regression is the most popular model for binary data (Agresti, 2007). It is the conventional tool for building linear class prediction rules and evaluating the significance of each predictor (Boulesteix and Hothorn, 2010). For binary data, where the outcome Y falls into one of two categories, Yes or No, rather than modeling this response directly, logistic regression models the probability that Y belongs to a particular category (James et al., 2013). In this paper, Y belongs to the binary categories, lung cancer or healthy control, and we model the probability of having lung cancer.
The logistic regression model is given as follows

$$\text{g}(\pi) = \log \frac{P(Y_i = 1|Z_{1i}, ..., Z_{pi})}{1 - P(Y_i = 1|Z_{1i}, ..., Z_{pi})} = \beta_0 + \beta_1 Z_{1i} + ... + \beta_p Z_{pi} \qquad (1)$$

where $i = 1, ..., 536$ are the subjects, Y the binary response variable of interest and $Z_1, ..., Z_p$ $(p = 1, ..., 14)$ are the p predictors.

Logistic regression analysis studies the association between a binary response variable and a set of predictor variables. Therefore, to study the relationship between the binary (lung cancer or healthy control) dependent variable and the independent clinical risk factors in this study, we employ this technique. In particular we use this statistical analysis tool to select the best subset of predictor variables for the response.

3.1.1 Stepwise Selection

The goal of variable selection is to select features from the universe of variables based on a specific criterion. Typically, these methods are iterative and computationally intensive since they search for the best subset of predictors. Variable selection is preferred when the final model needs to preserve the original variables to understand the contributions of each variable to the model (James et al., 2013).

To this effect, in this study, we obtain the best subset of predictor variables related to the response using stepwise selection approach. In particular, backward step-wise selection, which is a computationally efficient alternative to the best subset approach (James et al., 2013) was employed. This procedure starts with a full model containing all predictors, and then removes predictors from the model, one-at-a-time, until all the predictors are removed from the model. Traditionally, for a logistic regression model, at each step the variable that gives the smallest additional improvement to the fit according to deviance (small deviance imply good fit) value is dropped from the model. Unlike best subset selection otherwise known as exhaustive search, which involves fitting 2^p models, backward stepwise selection involves fitting one full model, along with $p - k$ models in the k^{th} iteration, for $k = 0, ..., p-1$. This amounts to a total of $1 + \sum_{k=0}^{p-1}(p-k) = 1 + p(p+1)/2$ models (James et al., 2013). Backward step-wise selection requires that the number of samples n is larger than the number of variables, p (so that the full model can be fit). In contrast, forward stepwise can be used even when $n < p$, and so is the only viable subset method when p is very large. However, $n > p$ in this study, hence backward selection is an appropriate choice.

3.2 Univariate Analysis of Metabolic Features

3.2.1 Model Formulation

After selecting the best subset of clinical predictors associated with the response, we move a step forward and consider univariate models for the metabolic features. To this effect, we again make use of the logistic regression technique described in section 3.1, whereby, we look at the metabolic features one at a time in a logistic regression model already containing the clinical variables selected as described in section 3.1.1. In principle, one can argue that this is a multivariate method, since more than one variable is considered at a time. However, the motivation for adopting the term, is as a result of the fact that, for all metabolic features, the subset of clinical variables remain unchanged. This method is used mainly to explore the correlation between each metabolic feature and the response given the clinical factors.
The model is given as follows;

$$\mathrm{g}_J(\pi) = \log\frac{P(Y_i = 1|Z_{1i}, ..., Z_{ri}, X_{Ji})}{1 - P(Y_i = 1|Z_{1i}, ..., Z_{ri}, X_{Ji})} = \beta_0 + (\beta_1 Z_{1i} + ... + \beta_r Z_{ri}) + \beta_J X_{Ji} \qquad (2)$$

where $i = 1, ..., 536$ are the subjects, $J = 1, ..., 102$ are the metabolic features, Y the binary response variable of interest and $Z_1, ..., Z_r$ are the r selected best predictors.

3.2.2 Multiplicity Adjustment

In statistical hypothesis testing, the speculation that there is no difference between groups is referred to as the null hypothesis (Amaratunga et al., 2014). With the univariate model formulated in section 3.2.1, there are 102 null hypotheses being tested, the J^{th} null hypothesis, for $\mathrm{J} = 1, ..., 102$, being that the J^{th} metabolic feature has no additional predictive value in the model containing clinical risk factors. From model (2) in section 3.2.1 the null and alternative hypotheses are given as $\mathrm{H}_0 : \beta_J = 0$ and $\mathrm{H}_1 : \beta_J \neq 0$, $\mathrm{J} = 1, ..., 102$, respectively

When many hypotheses are tested, the probability that a type I error occurs increases sharply with the number of hypotheses (Lin et al., 2012). The usual methods to the multiplicity issue demands for controlling the family-wise error rate (FWER) (Benjamini and Hochberg, 1995). According to Benjamini and Hochberg, 1995, FWER is defined as ”the probability to reject erroneously at least one true null hypothesis”. In this paper, we employ an approach proposed by Benjamini and Hochberg (1995), which controls the false discovery rate (FDR), which they defined as ”the proportion of rejected null hypotheses which are erroneously rejected” (Benjamini and Hochberg, 1995). Techniques based on controlling the FDR have gained their popularity in the microarray and other high-dimensional data settings, because they result to higher power as compared to the methods that control the FWER (Lin et al., 2012).

3.3 A Global Test for Additional Predictive Value of Metabolic Features

In this study, we want to investigate whether the metabolic data has an additional predictive value on the binary response given the clinical factors. However, given that the metabolic data is high-dimensional, evaluation of its additional predictive value will require tools different from the standard statistical ones. Therefore, we will on the one hand, employ two different high-dimensional data global test approaches; one proposed by Boulesteix and Hothorn, (2010) (herein, Global Test I), and another by Goeman et al.(2004) (herein, Global Test II). On the other hand, we will make use of the Likelihood ratio test. This test is applicable since the dataset for this study has sufficiently a large number of observations(536 subjects).

3.3.1 Global Test I (Boulesteix and Hothorn, 2010)

Boulesteix and Hothorn, (2010), suggested a permutation-based testing procedure for a global test of additional predictive value for high-dimensional molecular data. Their technique is a two step approach which combines logistic regression and boosting regression. In the first step, a logistic regression model, as described in section 3.1 is fitted to the clinical covariates to obtain estimates for the logistic regression coefficients from which a linear predictor can be computed. In the second step, a boosting regression,which works by sequentially applying a stratification algorithm to reweighted versions of the training data and then taking a weighted majority vote of the sequence of classifiers thus produced (Friedman et al., 2000) is fitted to the metabolic features to obtain estimates for the regression coefficients. These estimates are then joined with estimated regression coefficients of the first stage to compute a linear predictor. Predictive probabilities are then calculated from the linear predictor and the corresponding average negative binomial log-likelihood, $\boldsymbol{l}$ is derived. Furthermore, a permutation test is considered to test the null hypothesis that the molecular data have no additional predictive power given the clinical covariates. To this effect, the model

$$\log\frac{P(Y=1)}{1-P(Y=1)} = \beta_0 + \sum_{i=1}^{p}\beta_i Z_i + \sum_{j=1}^{q}\beta_j^* X_j \qquad (3)$$

is fitted, where Z_i $(i = 1, ..., p)$ are the clinical covariates and X_j $(j = 1, ..., q)$ are the molecular features.

The null-hypothesis being tested is;

$$\mathrm{H}_0 : \beta_1^* = \beta_2^* = ... = \beta_q^* = 0$$

and it is tested using a permutation procedure by permuting $X_1, ..., X_q$ only. The two-step procedure is applied and the negative binomial log-likelihood is computed for this permuted dataset. This procedure is repeated a large number of times B, resulting to $l_1, ..., l_B$ negative binomial log-likelihoods. The permutation p-value is then obtained as

$$\text{p-value} = \frac{1}{B}\sum_{b=1}^{B}\mathbf{1}(l_b \leq l)$$

3.3.2 Global Test II (Goeman et al., 2004)

Goeman et al.(2004) suggested a global test to be used for the analysis of microarray and other high-dimensional data. According to these authors, the test can be used to determine whether the global expression pattern of a group of genes is significantly related to some clinical outcome of interest. Since the test gives one p-value for a group of genes, it is also suitable to test for the additional predictive value of a set of features just as it allows groups of genes of different sizes to be compared. Indeed, there is a close connection between finding differentially expressed genes and predicting the clinical outcome . This property is used to derive the test within the generalized linear model framework which includes logistic regression as a special case (Goeman et al., 2004). The model is formulated as follows;

$$E(\mathrm{Y}_i|\beta) = h^{-1}\left(\beta_0 + \sum_{j=1}^{J} X_{ij}\beta_j\right) \tag{4}$$

where β_0 is the intercept, J is the length and β are the regression coefficients, of the vector of genes (features) we want to test, h is the logit link function, and β_j is the regression coefficient for gene (feature) j. The hypothesis

$$\mathrm{H}_0 : \beta_1 = \beta_2 = ... = \beta_J = 0$$

that all the regression coefficients are zero is then tested

3.3.3 Likelihood Ratio Test (LRT)

The likelihood ratio test (LRT) is a statistical test used to compare two nested models. A relatively more general model, g is compared with a simpler model, s to determine if it fits a particular dataset significantly better. The LRT is only valid if the general model, differ from the simple one only by the addition of one or more parameters. That is, the two models are hierarchically nested. The general, g and the simpler, s models for this study are respectively, formulated as follows;

$$\mathrm{g}(\pi) = \log\frac{P(Y_i = 1|\mathbf{Z},\mathbf{X})}{1 - P(Y_i = 1|\mathbf{Z},\mathbf{X})} = \beta_0+(\beta_1 Z_{1i}+...+\beta_r Z_{ri})+(\beta_{r+1}X_{1i}+\beta_{r+2}X_{2i}+...+\beta_{r+J}X_{Ji}) \quad (5)$$

and

$$\mathrm{g}(\pi) = \log\frac{P(Y_i = 1|\mathbf{Z})}{1 - P(Y_i = 1|\mathbf{Z})} = \beta_0 + \beta_1 Z_{1i} + ... + \beta_r Z_{ri} \quad (6)$$

where $i = 1, ..., 536$ are the subjects, $J = 1, ..., 102$ are the metabolic features, Y the binary response variable of interest and $\mathbf{Z} = Z_1, ..., Z_r$ are the r selected best predictors.
The null hypothesis of interest is therefore formulated as follows;

$$\mathrm{H}_0 : \beta_{r+1} = \beta_{r+2} = ... = \beta_{r+J} = 0$$

The LRT can therefore be presented as a difference in the log-likelihoods as follows

$$\mathrm{LRT} = -2^* \left(\mathrm{lnL}_s - \mathrm{lnL}_g\right) \quad (7)$$

where L_s and L_g are the likelihood functions of the simpler and general models, respectively. This LRT statistic is approximately chi-square distributed with degrees of freedom equals to the difference in parameters between the two models.

3.4 The LASSO

Classical logistic regression and by extension, stepwise logistic regression for subset selection (described in section 3.1) uses least squares to fit a linear model that contains a subset of the predictors. This subset selection approach produces a model that is interpretable and has possibly lower prediction error than the full model. However, the process either retains or discards variables, the method often reveal high variance, and so doesn't reduce the prediction of the full model (Hastie et al., 2001). A possible alternative is to fit a model with all p predictors using a technique that constrains or regularizes the coefficient estimates, or equivalently, that shrinks the coefficient estimates towards zero (James et al., 2013). By shrinking the coefficient estimates, we can significantly reduce their variance. Ridge regression and LASSO (Least Absolute Shrinkage and Selection Operator) are the two best-known techniques for shrinking regression coefficients towards zero (james et al., 2013).

The conventional least squares fitting procedure estimates $\beta_0, \beta_1, ..., \beta_p$ uses values that minimize

$$\text{RSS} = \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2 \tag{8}$$

while for Ridge regression and LASSO, the coefficients are estimated by minimizing slightly different quantities. The ridge regression coefficient estimates, $\hat{\beta^R}$ are the values that minimize

$$\sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{9}$$

and the LASSO coefficient estimates $\hat{\beta^L}$ minimize

$$\sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| = \text{RSS} + \lambda \sum_{j=1}^{p} |\beta_j| \tag{10}$$

where $\lambda \geq 0$ is a tuning parameter, to be determined separately. It controls the amount of shrinkage.

For Ridge regression, the penalty $\lambda \sum_{j=1}^{p} \beta_j^2$ in equation (9) will shrink all of the coefficients towards zero, but it will not set any of them exactly to zero (unless $\lambda = \infty$). This may not be a problem for prediction accuracy, but it can lead to model interpretation challenges in settings where the number of variables p is quite large (James et al., 2013).

The LASSO is a suitable alternative technique that resolves this issue. In fact, unlike in Ridge regression, the so called l_1 penalty (LASSO penalty), $\lambda \sum_{j=1}^{p} |\beta_j|$ in equation (10) has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ

is sufficiently large. Hence, just like best subset selection,the LASSO performs variable selection. As a result, models generated from the LASSO are generally much easier to interpret than those produced by Ridge regression (James et al., 2013). Moreover, the LASSO can produce a very good prediction accuracy, because shrinking and removing the coefficients can reduce variance without a substantial increase of the bias, this is especially useful when you have a small number of observation and a large number of features.

3.4.1 Selecting the Tuning Parameter, λ

One very important aspect for the LASSO is selecting a good value of λ. In fact, according to James et al.(2013), "we know that bias increases and variance decreases when λ increases", hence, a trade-off between bias and variance has to be found. Cross-validation provides a relatively straightforward way to tackle this problem (James et al., 2013). To this effect, we choose a grid of λ values, and compute the cross-validation error for each value of λ. The tuning parameter value for which the cross-validation error is smallest is then selected. We then fit the model again using all of the available observations and the selected value of the tuning parameter.

3.5 Classification with the LASSO

James et al.(2013), defined classification as "the process of predicting qualitative responses". Predicting a qualitative response for an observation can be described as classifying that observation, since it involves designating the observation to a category, or class (James et al., 2013). In the classification setting, the data set will be divided into two sets, namely, the training and the test sets. We use the set of training observations to build a classifier, and we want the classifier to perform well on both the training data and on test observations that were not used to train the classifier. There are many possible classification techniques, or classifiers, that one might use to predict a qualitative response. The LASSO approach, or classifier described in section 3.4 will be employed here. One fascinating property of the LASSO is that along the solution path some coefficient estimates may become zero, thus performing gene (or variable) selection by choosing those genes (or variables) with non-zero coefficients (Amaratunga et al., 2014).

In this section, the objective is to build a classifier or model, based on two different sets of variables (clinical variables and metabolic features, herein, clinical data and metabolic data, respectively). As illustrated in Figure 1 below, we will build, on one hand, a model based on only the clinical data and on the other hand, a model based on both the clinical and the metabolic data. The clinical data classifier makes use of the best subset of clinical variables selected according to the backward stepwise selection technique described in section 3.1.1. These clinical factors are not subject to the LASSO penalty. i.e no further variable selection is done. Meanwhile, the classifier for the combined data makes use of the best subset of clinical variables and all the metabolic features. In this case, only the metabolic features are subject to the LASSO penalty and thus, the clinical factors are forced to have non-zero coefficients. With this approach, both classifiers will always take the clinical variables into account. Furthermore, both models will be trained on the same train set and tested on the same test set. Hence any observed difference between the measures of the two models or classifiers will be attributed to the metabolic features.

3.5.1 Model Evaluation

In order to evaluate the prediction accuracy and compare the two fitted models (one with the combined data and the other with only clinical data), we monitor their misclassification errors (MCE), sensitivity(SEN), specificity(SPE), positive predictive values(PPV) and negative predictive values(NPV). According to Amaratunga et al.(2014), sensitivity and specificity are defined as the proportions of veritable positives and negatives, respectively, which are accurately recognized as such. Meanwhile, the positive and negative predictive values are the proportions of positives and negatives, respectively, that are actual positives and actual negatives, respectively. And misclassification error is defined as the proportion of subjects for whom a wrong classification is made (Amaratunga et al. 2014).

		Predicted label	
		LC	**Control**
True label	**LC**	TP	FN
	Control	FP	TN

Table 1: *Confusion matrix indicating the number of true positive (TP), false positive(FP), true negative and false negative values*

From Table 1, we can then compute the above mentioned accuracy measures using the formulae; MCE=$\frac{FP+FN}{TP+TN+FN+FP}$, SEN=$\frac{TP}{TP+FN}$, SPE=$\frac{TN}{TN+FP}$, PPV=$\frac{TP}{TP+FP}$ and NPV=$\frac{TN}{TN+FN}$.

3.5.2 Cross Validation

In the validation set approach for estimating test error, a subset of the data is held-out as the validation set while the other portion is used in learning the method. The resulting estimate of the test error rate can be extremely fluctuating, conditional on exactly which observations are contained in the training set and which observations are contained in the validation set (James et al., 2013). Therefore, in order that the evaluation of the model performance is not biased because of how the data was split between the train and test sets, and to also allow for a kind of validation of the classifier, the 3-fold cross validation, which consist of splitting the data into 3 parts and setting one portion aside for testing (test set=$\frac{1}{3}$ of data) while the other two parts are used to train (train set=$\frac{2}{3}$ of data) the model was done 1000 times. In this cross validation, the original dataset is randomly divided into the train and the test sets 1000 times and two LASSO models (one with all variables and one with only clinical factors) are fitted to the data for each random split. The MCE, SEN, SPE, PPV and NPV for both models are then obtained for the 1000 iterations. The distributions and summary measures of these accuracy measures are then obtained and compared between the two models to evaluate which one has a better prediction accuracy (see Figure 1).

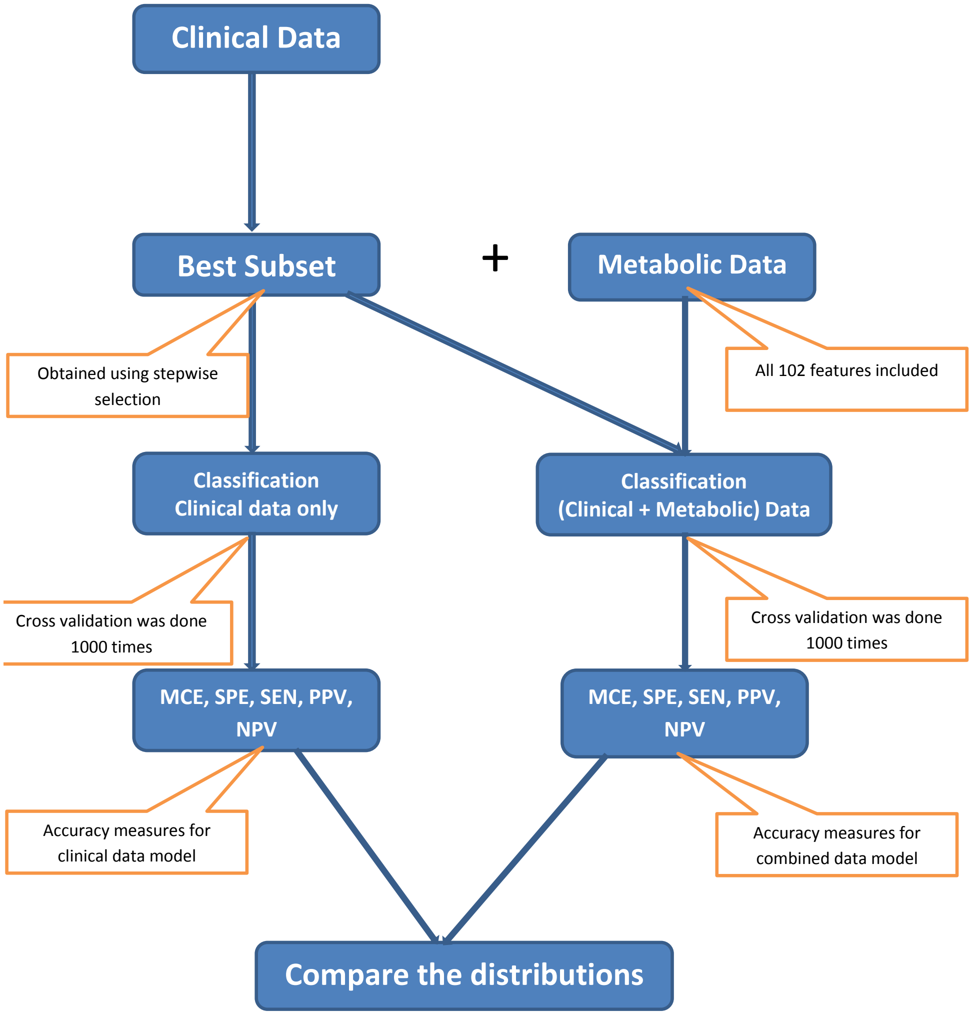
Clinical Data
Best Subset
+
Metabolic Data
Obtained using stepwise selection
All 102 features included
Classification Clinical data only
Classification (Clinical + Metabolic) Data
Cross validation was done 1000 times
Cross validation was done 1000 times
MCE, SPE, SEN, PPV, NPV
MCE, SPE, SEN, PPV, NPV
Accuracy measures for clinical data model
Accuracy measures for combined data model
Compare the distributions

Figure 1: *An illustration of the classification procedure for both classifiers using the 3-fold cross validation 1000 times*

4 Results and Interpretation

4.1 Logistic Regression

4.1.1 Stepwise Selection

Table 2 summarises the results of the fitted logistic regression model for the best subset of clinical variables. There were 9 variables selected according to the backward stepwise selection procedure described in section 3.1.1. After running a model with these variables, the significant ones were then retained as the best subset of clinical variables capable of predicting the outcome. In this study, we are interested in making predictions. However, the default (deviance) criterion for logistic regression model selection favours models with a low training error, whereas we wish to choose a model that has a low test error.Therefore the best model was selected based on *Akaike information criterion* (AIC), which corrects for the bias in using the training error as an estimate for the test error.

	Estimate	Std. Error	z value	Pr(>|z|)
Age	0.04	0.01	3.04	0.00
Packyears	0.02	0.01	2.72	0.01
BMI	-0.08	0.02	-3.15	0.00
COPD: Yes	1.70	0.26	6.50	0.00
Cardiac: Yes	1.47	0.42	3.47	0.00
Smoking: Never	-2.19	0.44	-4.93	0.00
Smoking: Stopped	-0.36	0.25	-1.45	0.15
Coagulation: Yes	-0.76	0.25	-3.00	0.00

Table 2: *Selected model for clinical data*

4.2 Univariate Analysis of Metabolic Features

Figure 2 displays a scatter plot of the 102 metabolic feature coefficient estimates against the logarithmic transformation of their p-values. The observations below the horizontal red line represent the coefficients with significant p-values at 5% significance level, while those above the red line are estimates with non-significant p-values. There were in total 53 significant coefficient estimates and 49 were non significant. These results reveals that metabolic features may have additional contribution in predicting the outcome. However, not all the features may contribute to this possible additional predictive value. Further analysis is therefore carried out to determine which metabolic features adds value to the prediction.

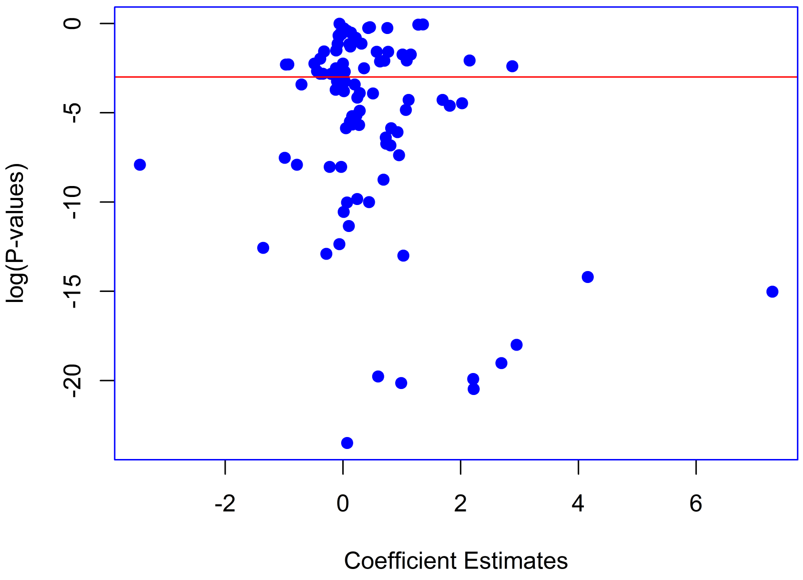

Figure 2: *Coefficient estimates for univariate analysis plotted against the logarithmic transformation of their p-values*

4.3 Global Test

Table 3 shows the results of the three global test for additional predictive value described in section 3.3 above. It can be seen that all three tests have p-values less than 0.05, which implies that, at 5% significance level, the tests reject the null hypothesis that the metabolic features have no additional predictive value given that we already have clinical factors. This is an indication that the prediction of subject outcome will be improved by taking subject's metabolic information into account in addition to already available clinical factors.

	Test	P-value
1	Global Test I	< 0.0001
2	Global Test II	< 0.0001
3	LRT	< 0.0001

Table 3: *Tests for additional predictive value*

4.4 The LASSO

4.4.1 Selection of Metabolic Features

In order to determine which metabolic features contributed to the evaluation of the predictive ability of the fitted model with both clinical and metabolic features, and whether the set of contributing metabolites vary across iterations, a count of the number of times a metabolite was included in the model was obtained by observing the number of its non-zero coefficients in the 1000 cross validations. Figure 3 (left panel) shows the distribution of the frequency of metabolic feature selection. It can be observed for example that, the first metabolic feature, represented by the first bar was selected about 440 times, meanwhile the second metabolic feature is seen to have been selected about 35 times. The difference in the bar lengths indicates variability in the metabolic future selection. In addition, all the bars are below the 1000 frequency mark, which is an indication that no metabolic feature was selected for every iteration. Further analysis revealed that one metabolic feature out of the 102 was never selected. However, some features have extremely high selection frequencies. A bar plot of the top 10 selected ones is displayed in Figure 3 (right panel).

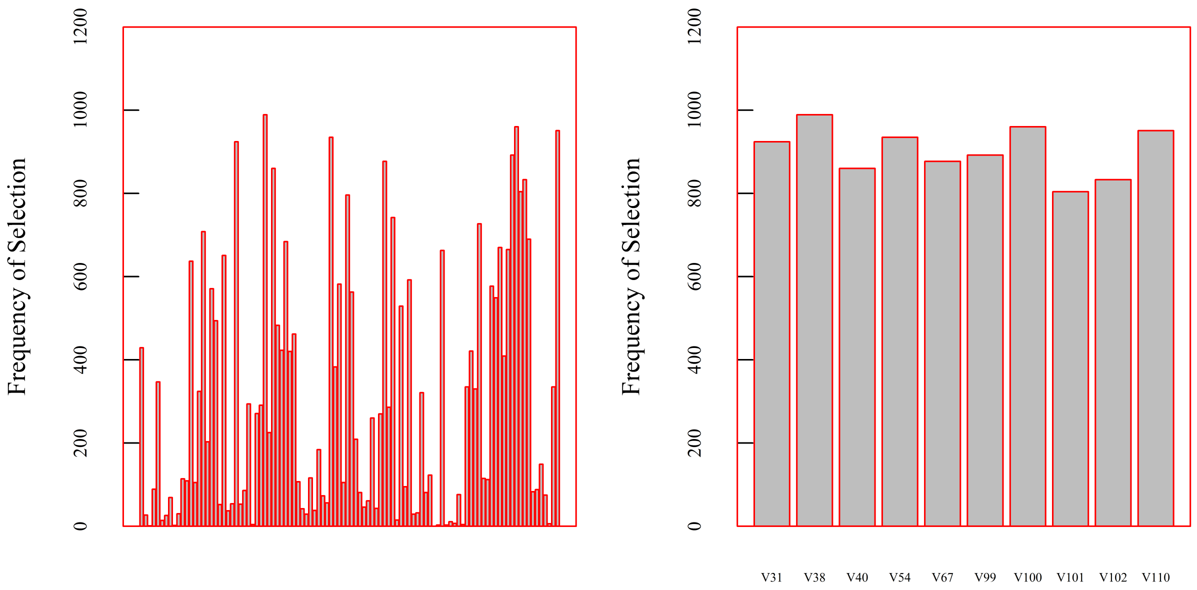

Figure 3: *Bar plots for the frequency of selection for metabolic features (left panel) and for the top 10 selected features (right panel)*

As can be seen on Figure 3 (right panel), the top 10 metabolic features were all selected more than 800 times. This implies that for more than 800 iterations, the features had non-zero coefficients. Further analysis showed V38 with 989 selections as the most frequently selected metabolic feature, closely followed by V100 with 960 selections. This results unveil the importance of metabolic information for the predictive ability of the model and hence for the classification of the individuals. However, not all metabolic features may be relevant.

4.4.2 Model Comparison

Misclassification Errors(MCE)

Figure 4 below is a display of box plots for the train and test misclassification errors (MCE) for the model with only clinical factors(left panel), and the model with both clinical factors and metabolic features(right panel). The blue boxes represent the train error while the red boxes represent the test error. By visual inspection, it can be seen that the the train MCE is lower than the test MCE in both cases. This is because for the train error, prediction was done with the same data that was used to train the model, leading to high prediction accuracy since the model had seen the data already. On the other hand, for the test MCE, prediction was done with the test data which was excluded while learning the classifier. This is a more realistic MCE since in principle prediction is done for new observations which were never part of the training set. Table 4 shows summary statistics of these MCE's and it can be seen that, in both cases the mean and median for the train MCE are lower than those for the test MCE while the maximum is higher for the test MCE.

From the red boxes in the Figure 4, it can be seen that the test MCE for the model with only clinical factors is higher than that of the model with both clinical factors and metabolic features. Moreover, from Table 4, it can be observed that the mean, median, minimum and maximum test MCE are higher for the model with clinical data compared to those for the model with the combined data. Therefore, we can reliably say that the MCE decreases when the combined data is used. This means the the classifier that takes metabolic information into account is seen to performs better than the classifier with only clinical risk factors.

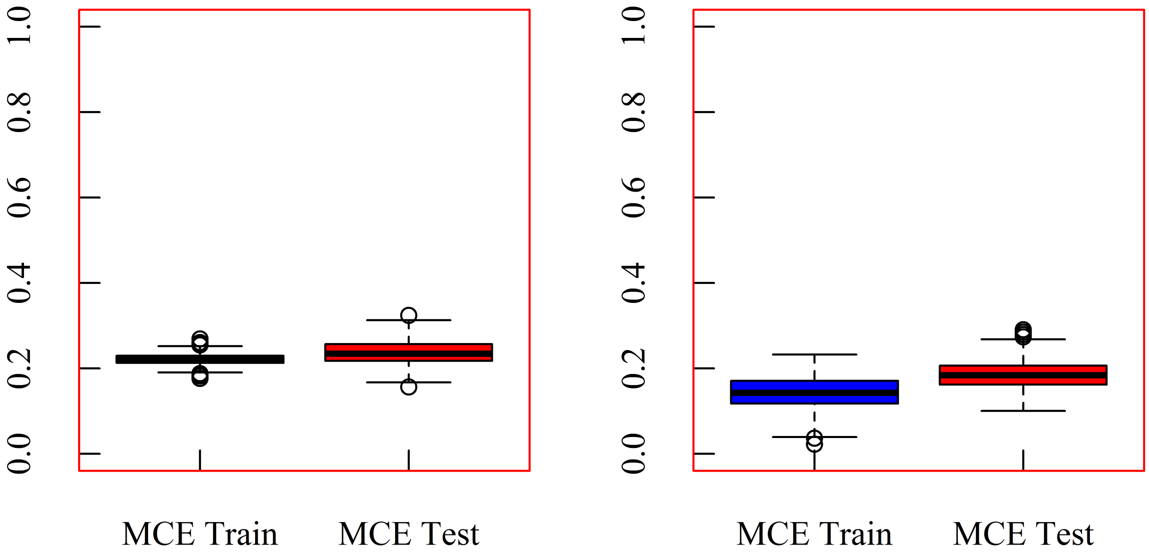

Figure 4: *MCE for clinical data model (left panel) and for the combined data model(right panel)*

Sensitivity(SEN), Specificity(SPE), Positive Predictive Value(PPV) and Negative Predictive Value(NPV)

From left to right and from top to bottom, the resulting box plots on Figure 5 represent the sensitivity, specificity, positive predictive values and negative predictive values, for both clinical data model (red) and for the combined data model(blue).
As can be seen on the plots for sensitivity (top left) and specificity (top right), the model with combined data (blue box) has higher values than the model with only clinical factors (red box). Furthermore, the results reported in Table 4 show that, the mean, median, minimum and maximum values for sensitivity and specificity are higher for the combined data model as compared to the model without the metabolic data. This implies that the true positive rate, as well as the true negative rates increases when metabolic data is used in combination with the classical clinical factors.

Similarly the plots for the negative and positive predictive values (bottom left and bottom right, respectively), show that the combined data (blue box) indicates higher values for both positive and negative predictive values as compared to the red box (clinical data model). This again reveals that the performance with the combined data outweighs that with only the clinical data. The results in Table 4 points in the same direction with higher values of the mean, median, minimum and maximum recorded for the model with the combined data.

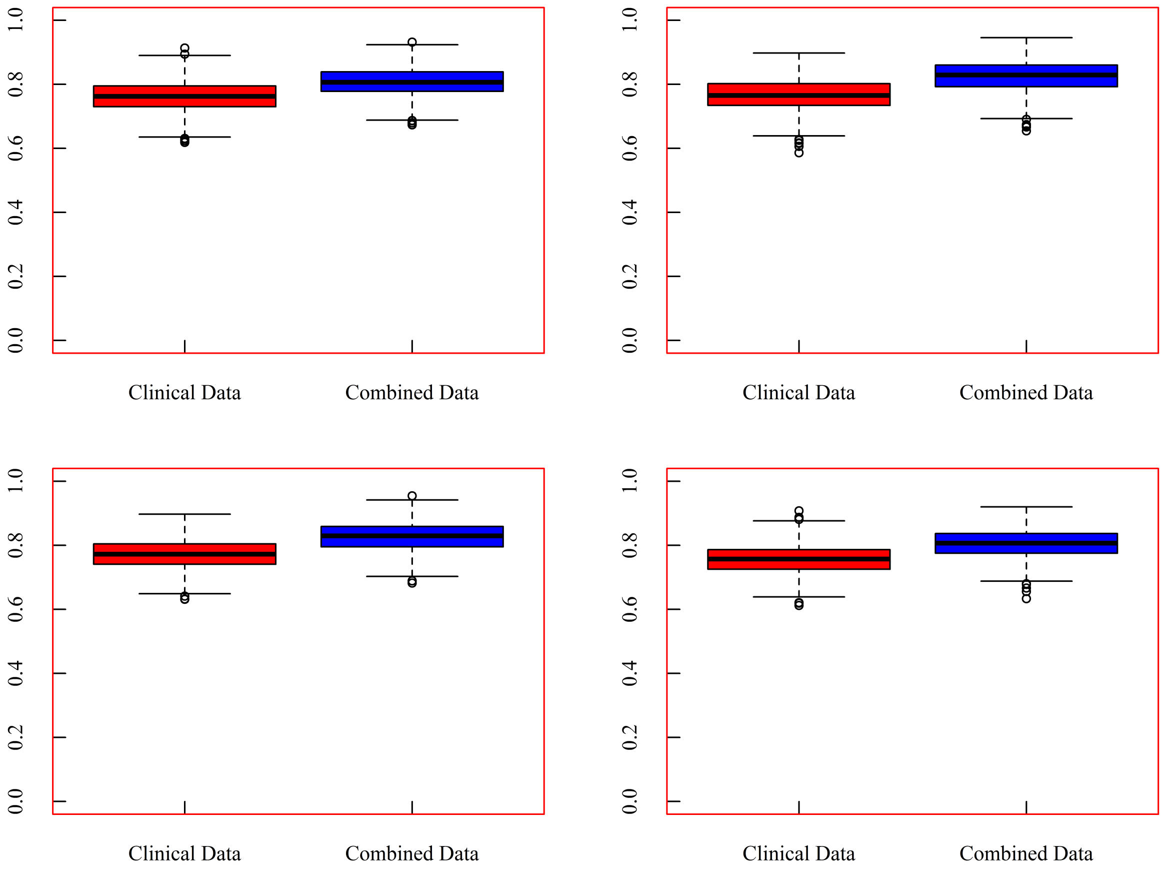

Figure 5: *From left to right and from top to bottom, the resulting box plots for sensitivity, specificity, positive predictive values and negative predictive values for both clinical data model (red) and for the combined data model(blue)*

Measure	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
			Clinical Data Measures			
MCE Train	0.1765	0.2129	0.2213	0.2215	0.2297	0.2689
MCE Test	0.1564	0.2179	0.2346	0.2372	0.2570	0.3240
SEN	0.6186	0.7303	0.7624	0.7613	0.7952	0.9136
SPE	0.5862	0.7340	0.7654	0.7665	0.8023	0.8977
PPV	0.6316	0.7412	0.7727	0.7723	0.8046	0.8974
NPV	0.6122	0.7253	0.7576	0.7563	0.7865	0.9079
			Combined Data Measures			
MCE Train	0.02241	0.11765	0.14286	0.14301	0.17087	0.23249
MCE Test	0.1006	0.1620	0.1844	0.1849	0.2067	0.2905
SEN	0.6735	0.7787	0.8066	0.8073	0.8388	0.9318
SPE	0.6548	0.7927	0.8295	0.8243	0.8602	0.9459
PPV	0.6824	0.7955	0.8298	0.8271	0.8590	0.9545
NPV	0.6333	0.7753	0.8068	0.8049	0.8370	0.9200

Table 4: *Summary measures for both models*

5 Discussion and Conclusion

In this study, three different global tests were employed to examine whether there is an additional predictive value by adding parameters that reflect the plasma metabolic phenotype to a classical risk model already containing clinical risk parameters. Excitingly, all the tests yielded significant results in favour of a possible additional predictive value of metabolic information. With this results known, it was possible to move a step further to investigate whether the selection of high-risk individuals eligible for lung cancer screening can be improved by adding metabolic information and also to see which metabolic features contribute to the additional predictive value. The LASSO was used to effect this. The model was considered or fitted under two different settings. First, the model was fitted without accounting for the metabolic information, that is, a classical risk model with only clinical risk factors was considered. In the second step, the model was again fitted with metabolic features appended to the previous model. The results of this model points in the same direction as those of the global tests. This spells out the need for incorporating metabolic information in risk models used in the selection of high-risk persons qualified for lung cancer screening.

A univariate analysis of the metabolic features was carried out and the association between some features and the binary clinical outcome was revealed. This univariate analysis was done while adjusting for the clinical risk factors. This is in agreement with the view point of Boulesteix and Sauerbrei, (2011), that univariate analysis without adjusting for clinical factors says nothing about added predicted value. Furthermore, these same authors are in support of the view that even when a model with clinical factors is compared with a model with molecular data, we gain no information about what can be achieved by combining both predictor types (Boulesteix and Sauerbrei, 2011). This explains why we employed the LASSO model under two setting, whereby the combined predictor types were incorporated in one model and then compared with the model with clinical predictors. Although this approach is believed to be a better way of evaluating additional predictive value, additionally, it was ensured that the best subset of clinical predictors is forced in the combined data model. This was to avoid the negligible difference often caused due to imbalance of clinical predictors in both models thereby causing a dilution of the additional predictive value of molecular data (Boulesteix and Sauerbrei, 2011).

Results of both models show that the misclassification errors(MCE) for the training sets are generally smaller than those of the test set. This was observed across the 1000 randomly generated training and test sets used in fitting both model. These results are in accordance to expectations as described by James et al. (2013). In addition, the misclassification errors were found to be lower in favour of the model with combined data. Meanwhile, sensitivity, specificity, and positive and negative predictive values where seen to generally have higher values for the combined data model compared to the model with only clinical factors. Therefore , revealing an improvement in the prediction by adding metabolic information. Boulesteix and Sauerbrei, (2011), emphasized the need to validate the added predictive value before concluding on the usefulness of the new model. In this light, they proposed that this could be done by using either independent validation data or cross validation in the absence of external data. In this study, the validation was done using the cross validation approach by dividing data between the training and testing sets randomly 1000 times since there was no external data available to validate the model.

Therefore, it is revealed from this study that, by incorporating metabolic information in classical risk models, the predictive ability improves and thus better decisions about selection of individuals eligible for lung cancer screening can be made.

References

[1] Agresti, A. (2007). An Introduction to Categorical Data Analysis, 2nd edn. *Hoboken, New Jersey [u.a.]: Wiley- Interscience.*

[2] Amaratunga, D., Cabrera, J. and Shkedy, Z. (2014). Exploration and Analysis of DNA Microarray and Other High-Dimensional Data. *New Jersey: Wiley & Sons*

[3] Bach, P., Mirkin, J., Oliver, T., Azzoli, C., Berry, D., Brawley, O., Byers, T., Colditz, G., Gould, M., Jett, J., Sabichi, A., Smith-Bindman, R., Wood, D., Qaseem, A., Detterbeck, F. (2012). Benefits and Harms of CT Screening for Lung Cancer. *JAMA: 307(22): 2418-29.*

[4] Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological), Vol. 57, No. 1, pp. 289-300*

[5] Boulesteix, AL. and Hothorn T. (2010). Testing the Additional Predictive Value of Hihg-dimensional Molecular Data. *BMC Bioinformatics,* ***11****:78*

[6] Boulesteix, AL. and Sauerbrei, W. (2011). Added predictive value of high-throughput molecular data to clinical data and its validation. *Briefings in Bioinformatics, Volume 12, Pages 215229, https://doi.org/10.1093/bib/bbq085*

[7] Commission on Chronic Illness (1957). Chronic illness in the United States: Volume I. Prevention of chronic illness. *Cambridge, Mass., Harvard University, Vol. I, Press, p. 45*

[8] Friedman, J., Hastie, T. and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting. *Annals of Statistics, 28:2000*

[9] Friedman, J., Hastie, T. and Tibshirani, R. (2009). Glmnet: Lasso and elastic-net regularized generalized linear models. *R Package Version. 1.*

[10] Friedman, J., Hastie, T. and Tibshirani, R. (2008). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software, Vol. 33(1), 1-22*

[11] Friedman, J., Hastie, T., Tibshirani, R., Simon, N., Narasimhan, B. and Qian, J.(2019).Lasso and Elastic-Net Regularized Generalized Linear Models. *R Package Version.2.0-18*

[12] Goeman, J.J., van de Geer, S.A., de Kort, F., and van Houwelingen, J.C. (2004). A global test for groups of genes: testing association with a clinical outcome.*Bioinformatics, 20(1):93-99*

[13] Golub, TR., Slonim, DK., Tamayo, P., et al (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science Vol 286.*

[14] Hastie, T., Tibshirani, R. and Friedman, J. (2001). The Elements of Statistical Learning: Data mining, Inference and Prediction. *New York: Spinger series in statistics.*

[15] Hoefling, H. and Tibshirani, R. (2008). A study of pre-validation. *The Annals of Applied Statistics, 2: 643-664.*

[16] International Agency for Research on Cancer (2018). GLOBOCAN 2018: estimated cancer incidence, mortality and prevalence worldwide in 2018 Lung cancer fact sheet.Accessed 16/05/2019. http://globocan.iarc.fr/Pages/fact_sheets_cancer.aspx?cancer=lung

[17] James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). An Introduction to Statistical Learning, with Applications in R. *Springer Texts in Statistics, DOI 10.1007/978-1-4614-7138-7 4.*

[18] Krishnamurthy, S. (2014). High Dimensionality in Large Datasets:Part I. *Wiley online library.* Accessed on 18/05/2019. https://doi.org/10.1002/wilm.10328

[19] Lin, D., Shkedy, Z., Yekutieli, D., Amaratunga, D., Bijnens, L.(2012). Modelling Dose-response Microarray Data in Early Drug Development Experiments Using R. *Springer-Verlag.*

[20] Manser, R., Lethaby, A., Irving, LB., Stone, C., Byrnes, G., Abramson, MJ. and Campbell, D. (2013). Screening for lung cancer. Cochrane Database Syst Rev ;6:CD001991. https://doi.org/10.1002/14651858.CD001991.pub3

[21] Molenberghs, G. and Verbeke, G. (2005). Models for Discrete Longitudinal Data. *New York: Springer.*

[22] Pinsky, P., Zhu, C. and Kramer, B. (2015). Lung cancer risk by years since quitting in 30+ pack year smokers.*J Med Screen;22:1517.* https://doi.org/10.1177/0969141315579119

[23] Rodriguez-Girondo, M., Salo, P., Burzykowski, T., Perola, M., Houwing-Duistermaat, J. and Mertens, B. (2018). Sequential double cross-validation for assessment of added predictive ability in high-dimensional omic applications. *The Annals of Applied Statistics, Vol 12, No 3, 1655-1678.*

[24] Simon, N., Friedman, J., Hastie, T., Tibshirani, R. (2011). Regularization Paths for Coxs Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software, Vol. 39(5) 1-13*

[25] Snowsill, T., Yang, H., Griffin, E., Long,L., Varley-Campbell, J., Coelho, H., Robinson, S. and Hyde, C. (2018).Low-dose computed tomography for lung cancer screening in high-risk populations: a systematic review and economic evaluation. *National Institute for Health Research Journals Library, vol. 22.*

[26] Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J. and Tibshirani, R.J. (2010). Strong Rules for Discarding Predictors in Lasso-type Problems.*Journal of the Royal Statistical Society: Series B (Statistical Methodology), 74(2), 245-266.*

[27] Tibshirani, R. and Efron, B. (2002). Pre-validation and inference in microarrays.Statistical Applications in Genetics and Molecular Biology, 1:Art. 1.

[28] Tukey, JW. (1977). Exploratory Data Analysis.*Reading(MA): Addison-Wesley.*

[29] Wilson, J. and Jungner, G. (1968) .Principles and Practice of Screening for Disease. *Public Health Papers 34,WHO, Geneva*

6 Appendix

6.1 Appendix A - Example: Application of the LASSO

This subsection is dedicated for a practical illustration of how the LASSO works. We use the dataset for this study to examplify this procedure as described in section 4.4. The results are displayed in appendix B (Figure 7) and appendix C (Table 5).

6.2 Appendix B - Figures

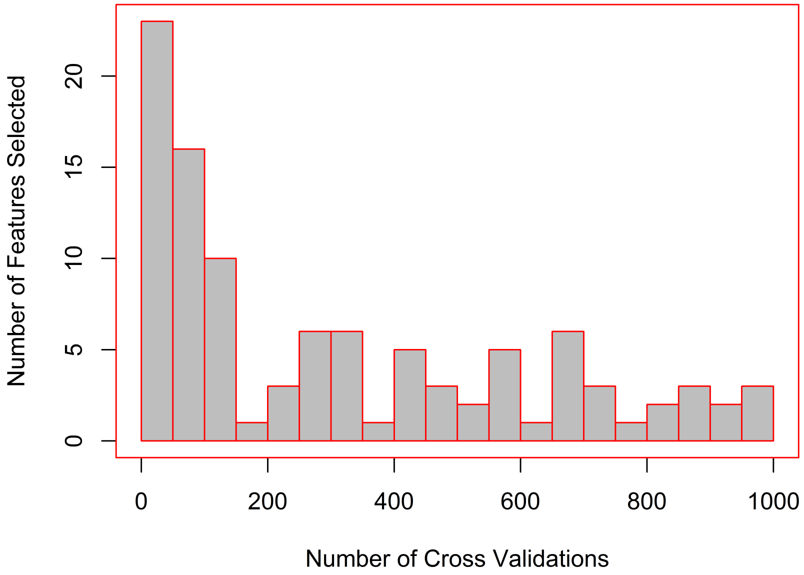

Figure 6: *Frequency of Selection for Metabolic Features*

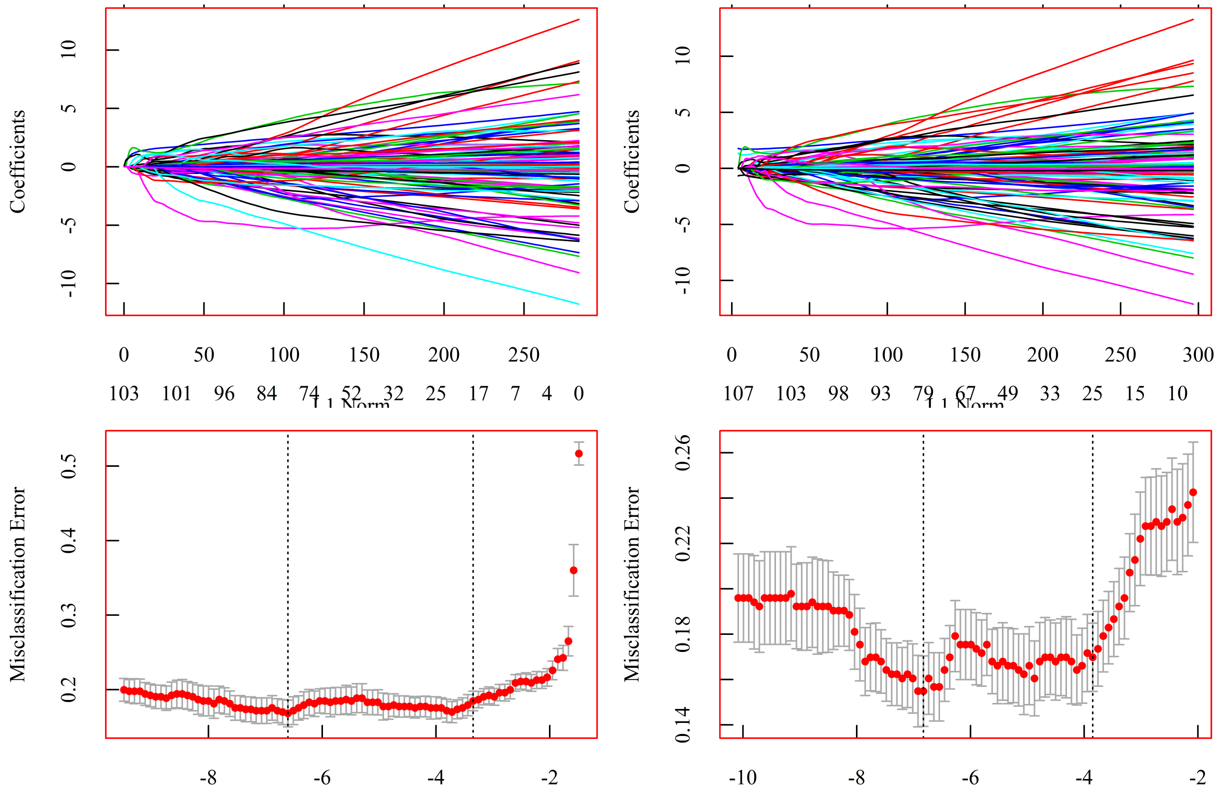

Figure 7: *From left to right and from top to bottom, the resulting plots of coefficients for different turning parameters, coefficients for different turning parameters(model with no penalty for clinical factors), selection of optimal turning parameter, selection of optimal turning parameter (model with no penalty for clinical factors)*

6.3 Appendix C - Tables

Table 4 below shows the results obtained by fitting a logistic regression model to the clinical data while Table 5 displays the results of classification with the lasso.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.04	1.00	0.04	0.97
Age	0.03	0.01	2.97	0.00
Gender:Male	0.42	0.45	0.94	0.35
Packyears	0.01	0.01	2.41	0.02
BMI	-0.08	0.03	-3.36	0.00
COPD:Yes	1.71	0.27	6.45	0.00
Diabetes:Yes	0.30	0.32	0.95	0.34
cholesterol:Yes	0.10	0.25	0.40	0.69
thyroid:Yes	-0.73	0.59	-1.23	0.22
Cardiac:Yes	1.51	0.43	3.52	0.00
Smoking:Never	-2.10	0.45	-4.67	0.00
Smoking:Stopped	-0.39	0.25	-1.56	0.12
PMW:No	-0.15	0.43	-0.35	0.73
PDMT:Yes	0.94	0.50	1.88	0.06
HBP:Yes	-0.47	0.27	-1.77	0.08
Coagulation:Yes	-0.84	0.27	-3.11	0.00

Table 5: *Parameter estimaes of the full clinical data model*

	Reference		
		0	1
Predicted	0	235	48
	1	28	225
	MCE= 0.14	SEN=0.89	SPC=0.82

Table 6: *Confusion Matrix for the example*

6.4 Appendix D - Implementation of Classification with the LASSO in R

The LASSO, as discussed in section 3.4 was implemented using the R package ”glmnet” (Friedman et al., 2009) which is available on CRAN. The package was initially developed for Lasso and Elastic-Net Regularized Generalized Linear Models and was later extended to incorporate cox’s proportional hazards models. The multiple-response Gaussian, and the grouped multinomial regression are the two most recent additions in the package (Friedman et al., 2019). The package provides and efficient approach and very easy to apply. It takes x,y data for regression models, and produces the regularization path over a grid of values for the tuning parameter lambda (Friedman et al., 2009; Friedman et al., 2019).

To implement the cross validation, whereby the data was randomly split into the training and test sets and model repeatedly fitted to obtain estimates, a user-defined R function was created to appropriately handle the analysis. The function and the analysis using this function that incorporate the the ”cv.glmnet” function of the ”glmnet” package is as follows;

```
   ###  FUNCTION ###
Fit.Lasso.kCV<-function(X,y,ind.train=1:nrow(X),ind.test=1:nrow(X),
       alpha.value=1,penality=NULL,indipend.data){
options(warn=-1)
if(is.null(penality)){
fit21 <- try(cv.glmnet(X[ind.train,],y[ind.train],
      alpha = alpha.value, standardize = F,family="binomial",
      type.measure="class",nfold=3),silent=T)
pred.fit21 <- try(predict(fit21,X[ind.test,],type="class"),
      silent=T)
pred.fit213 <-try(predict(fit21,X[ind.train,],type="class"),
      silent=T)
      Coefficients <-try(coef(fit21, s="lambda.min"),silent=T)
MCE.train<-try(1-sum(diag(table(y[ind.train],pred.fit213)))
            /length(y[ind.train]),silent=T)
MCE<-try(1-sum(diag(table(y[ind.test],pred.fit21)))
            /length(y[ind.test]),silent=T)

return(list(MCEi=MCE,MCE.train=MCE.train,pred.train=pred.fit213,
            pred.test=pred.fit21,train=ind.train,test=ind.test,
            Coefficients=Coefficients))
}
else {

fit21 <- try(cv.glmnet(X[ind.train,],y[ind.train],intercept=T,
            alpha = alpha.value,standardize = F,family="binomial",
            type.measure="class",penalty.factor =penality,nfold=3),silent=T)
pred.fit21 <- try(predict(fit21,X[ind.test,],type="class"),silent=T)
pred.fit213 <- try(predict(fit21,X[ind.train,],type="class"),silent=T)
            Coefficients <-try(coef(fit21, s="lambda.min"),silent=T)

MCE.train<-try(1-sum(diag(table(y[ind.train],pred.fit213)))
            /length(y[ind.train]),silent=T)
MCE<-try(1-sum(diag(table(y[ind.test],pred.fit21)))/length(y[ind.test]),
            silent=T)

return(list(MCEi=MCE,MCE.train=MCE.train,pred.train=pred.fit213,
            pred.test=pred.fit21,train=ind.train,test=ind.test,
            Coefficients=Coefficients))
}
}

### The LASSO Analysis ###
```

```
n.cv<-1000
MCE.test<-rep(NA,n.cv)
MCE.train<-rep(NA,n.cv)
Pred.train<-matrix(NA,357,n.cv)
Pred.test<-matrix(NA,179,n.cv)
Coeff<-matrix(NA,111,n.cv)
train<-matrix(NA,357,n.cv)
test<-matrix(NA,179,n.cv)

set.seed(110)
for (j in 1:n.cv){
ind.train <-as.vector(sort(sample(1:nrow(X),floor(2*nrow(X)/3),replace=F) ) )
ind.test <- as.vector(c(1:nrow(X))[-ind.train])
res.cv<-try(Fit.Lasso.kCV(X=X,y=y,ind.train=ind.train ,ind.test=ind.test,
      alpha.value=1, penality=p.fac2),silent=T)
MCE.test[j]<-try(res.cv$MCEi,silent=T)
MCE.train[j]<-try(res.cv$MCE.train,silent=T)
Pred.train[,j]<-try(as.numeric(res.cv$pred.train),silent=T)
Pred.test[,j]<-try(as.numeric(res.cv$pred.test),silent=T)
train[,j]<-ind.train
test[,j]<-ind.test
Coeff[,j]<-(try((res.cv$Coefficients)[1:111], silent=T))
print(j)
}
save(MCE.train,MCE.test,train,test,Pred.test,Pred.train,Coeff,
file="3foldCV_ALL_Data.Rdata")

npv.test<-ppv.test<-sen.test<-spe.test<-rep(NA,n.cv)

for (j in 1:n.cv){
```

```
tab.test<-table(y[test[,j]],Pred.test[,j])


npv.test[j]<-tab.test[1,1]/sum(tab.test[1,1],tab.test[2,1])
ppv.test[j]<-tab.test[2,2]/sum(tab.test[1,2],tab.test[2,2])
sen.test[j]<-tab.test[2,2]/sum(tab.test[2,1],tab.test[2,2])
spe.test[j]<-tab.test[1,1]/sum(tab.test[1,2],tab.test[1,1])
print(j)
}
save(MCE.train,MCE.test,train,test,npv.test,ppv.test,sen.test,spe.test,
     Coeff,file="3foldCV_ALL_Data_FINAL.Rdata")
```