UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences

School for Information Technology

Master of Statistics

Master's thesis

Association between 18F-FDG Positron Emission Tomography and metabolic profile of lung cancer patients

Mifflin-Rae Calvero

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

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

UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences

School for Information Technology

Master of Statistics

Master's thesis

Association between 18F-FDG Positron Emission Tomography and metabolic profile of lung cancer patients

Mifflin-Rae Calvero

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

SUPERVISOR :

Prof. dr. Ziv SHKEDY

Mevrouw Olajumoke Evangelina OWOKOTOMO

Contents

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Prof. dr. Ziv Shkedy for his guidance and his patience throughout the entire process of my thesis writing.

To my other supervisor, Olajumoke Evangelina Owokotomo, I would like to thank her for her support and her advice to make this thesis successful.

To all my teachers in Hasselt University these past two years, I would like to thank them for sharing their knowledge, their feedback, and their advice during lectures and exams.

To my classmates and my fellow ICP scholars, I would like to thank them for their friendship, their support, and the fun times we shared inside and outside classes.

To my friends in Belgium and in the Philippines, I would like to thank them for their encouragement and their support throughout my studies.

To my family, I would like to thank them for their patience, their encouragement, their love, and their prayers throughout my years of study and through the process of writing this thesis.

Finally, the Almighty God, for blessing me with His wisdom, motivation, and strength every day to finish this thesis. This accomplishment would not be possible without His divine grace.

Abstract

Lung cancer is the most commonly diagnosed type of cancer and the major cause of cancer death worldwide. Fluorine-18 fluorodeoxyglucose positron emission tomography (^{18}F-FDG PET) has been used widely in diagnosis and staging of cancer. Standardized uptake values (SUV) can be obtained from PET imaging which is the most common parameter to quantify metabolic activity of a tumor, along with metabolic tumor volume (MTV) and total lesion glycolysis (TLG). The use of human metabolic information to study cancer is recent and several studies attempted to identify biomarkers using metabolites to improve diagnosis, prognosis, to evaluate treatment response, and to gain deeper understanding of cancer. This study aims to determine the metabolites that are associated with the PET parameters.

Metabolic profiles of plasma from 222 lung cancer patients were obtained by hydrogen-1 nuclear magnetic resonance (^{1}H-NMR) spectroscopy. LASSO and Elastic Net methods were used to select and determine metabolites which are associated with the three PET parameters, integrating the clinical information from the patients. Cross-validation was used to evaluate the performance of the two methods. Multiple factor analysis (MFA) was also used to determine which metabolites are associated with the PET parameters and to explore their relationship.

The results showed that the metabolic profile of the patients is associated with the PET parameters. However, this cannot be used to predict the values of the PET parameters. Nevertheless, it has provided insights on the possibility of using the metabolic profile as biomarkers in lung cancer detection.

Keywords: *Lung cancer, ^{18}F-FDG PET, metabolites, LASSO, MFA*

1 Introduction

1.1 Cancer

Cancer incidence and mortality is growing worldwide. Bray et al. (2018) reported that it is expected to be the leading cause of death in every country of the world. Also in this report, it was presented that there will be 18.1 million new cases of cancer and 9.6 million deaths from cancer in 2018. The four most common cancers are breast, lung, prostate, and colorectal cancer (Bray et al., 2018; Cooper, 2000). In particular, lung cancer is the most commonly diagnosed cancer and the leading cause of cancer death, at 18.4% of the total cancer deaths worldwide (Bray et al., 2018).

Cancer occurs when there is an uncontrolled proliferation of any of the different kinds of cells in the body. The behavior and response to treatment of these cancer cells depend on which healthy normal cells and location in the body they affect (Cooper, 2000). This abnormal proliferation is called tumor, and it is either benign or malignant. However, only malignant tumors are considered as cancers. The tumor is benign if it does not invade nor spread to the surrounding normal tissue, while it is malignant if it does invade and spread throughout the body, which makes it dangerous. While benign tumors can be removed, the nature of malignant tumors makes them resistant and difficult to be treated locally (Cooper, 2000).

The cause of cancer can be attributed to many factors, including radiation, viruses, and chemicals. These factors are referred to as carcinogens, which damage the DNA, induce mutation in certain genes, or stimulate the proliferation of cells in the body (Cooper, 2000). For example, ultraviolet radiation from sun exposure is the most common cause of skin cancer and other factors include environmental pollutants such as arsenic (Fabbrocini et al., 2010), food contaminants such as aflatoxin causes liver cancer, and carcinogens in tobacco smoke, the most common cause of lung cancer (Cooper, 2000).

To understand the mechanisms of cancer in the body, imaging techniques have been developed. One of the imaging techniques used is the positron emission tomography (PET). PET imaging can be used in the diagnosis, prognosis, and staging of the disease, in measuring the response to therapy, and in identifying the site of disease, among other things (Bailey, Townsend, Valk, and Maisey, 2005). PET imaging commonly uses the tracer fluorine-18 fluorodeoxyglucose (^{18}F-FDG) in clinical oncology as this allows for quantification of metabolic activity of a tumor (Bailey et al., 2005; Moon, Hyun, and Choi, 2013). The standardized uptake value (SUV) is the most common parameter used in PET analysis. It is calculated as the ratio of the tissue concentration and the injected dose divided by the patient body weight over a region of interest (ROI), which is positioned centrally within a tumor (Adams, Turkington, Wilson, and Wong, 2010). SUV can be reported as the mean or the maximum SUV. The SUV$_{mean}$ is less sensitive to noise but is prone to observer variability, while SUV$_{max}$ is sensitive to noise, it is observer-independent, and therefore more commonly used (Adams et al., 2010; Soret, Bacharach, and Buvat, 2007). Other PET parameters, metabolic tumor volume (MTV) and total lesion glycolysis (TLG) have been developed to measure global changes in tumor metabolic activity (Larson et al., 1999; Zasadny, Kison, Francis, and Wahl, 1998). MTV refers to the volume of tumor with high metabolic activity (Moon et al., 2013) and TLG is the product of SUV and lesion volume (Larson et al., 1999).

The emergence of imaging technology for cancer has led to several studies on using these PET parameters for prognosis and diagnosis, for assessing the response to drug treatment or therapy, or for discovering biomarkers. One study by Chang et al. (2012) examined the correlation between PET parameters, Epstein-Barr virus (EBV) DNA (virus associated with nasopharyngeal carcinoma), and clinicopathological factors (e.g. age, sex, tumor stage), and found out that TLG is associated with EBV DNA and with tumor burden and clinical stage. Another study showed that MTV before treatment of esophageal cancer patients can predict the survival of patients after treatment (Shum et al., 2012). Another study by Cerfolio, Bryant, Winokur, Ohja, and Bartolucci (2004) showed that the percentage change of SUV$_{max}$ after treatment predicted the pathologic response of primary tumor in non-small cell lung cancer (NSCLC).

In combination with the imaging technology and cancer research is the development of cancer treatments and drugs. The most common treatments include surgery, chemotherapy, immunotherapy and radiation therapy, while new, developing ones include gene therapy and the use of nanotechnology to target cancer cells (Arruebo et al., 2011). The results of the cancer imaging and research can help the researchers and doctors understand the complexity and nature of cancer, which open the doors to developing targeted treatments and improving existing ones.

1.2 Metabolomics

The ”omics” sciences became popular when the technology in molecular biology, biochemistry, and analytical chemistry enabled the development of genomics, proteomics, transcriptomics, and metabolomics (Kiechle, Zhang, and Holland-Staley, 2004; Plaza, García-Galbis, and Martínez-Espinosa, 2017). This led to projects such as the ”Human Genome Project” where the human genome was sequenced (International Human Genome Sequencing Consortium, 2004; Venter, Smith, and Adams, 2015) and the still ongoing ”Human Proteome Project” where the human proteome is being mapped (Legrain et al., 2011), and the creation of databases to aid researchers in their studies. These projects help researchers gain a deeper understanding of the human biology and open up opportunities for clinical applications.

While genomics deals with the analysis of genes, proteomics deals with proteins, and transcriptomics deals with RNAs, metabolomics is the study of the small molecular weight molecules called metabolites in a biological specimen (Clish, 2015; Manzoni et al., 2018; Trivedi, Hollywood, and Goodacre, 2017). Moreover, the metabolome, which refers to the whole set of metabolites in an organism, is the final downstream product of gene transcription and therefore it is closest to the phenotype of the biological subject and can therefore be used to report on disease status and on the effect and the response of the subject to external stimuli, such as drug therapy, nutrition and exercise (Trivedi et al., 2017). The metabolome is considered to be more complex than the other ”omes” because it contains many diverse biological molecules (Horgan and Kenny, 2011).

The ”omics” technologies are used for detection and identification of genes, proteins, RNAs, and metabolites in a biological sample. For metabolomics, mass spectrometry and nuclear magnetic resonance (NMR) spectroscopy are the most commonly used for metabolic detection and identification. In mass spectrometry, also widely used in proteomics, metabolites are described by the mass-to-charge (m/z) values and the intensities of detected ions which together represent a mass spectrum. NMR spectroscopy also produces a spectrum which contains the chemical shift (ppm) and the intensities of molecules. NMR spectroscopy is highly selective, non-destructive, and sample preparation is easy, however, it is less sensitive than mass spectrometry, while mass spectrometry is both highly sensitive and selective and high-throughput (Lei, Huhman, and Sumner, 2011). Nevertheless, several studies have been conducted using these two methods.

Trivedi et al. (2017) listed some studies throughout the years aimed towards biomarker discovery using metabolomic approaches for diseases such as Alzheimer’s disease, cancer, cardiovascular diseases, diabetes, and multiple sclerosis and disorders such as autism, Down syndrome and schizophrenia. For example, a study on pancreatic cancer (Di Gangi et al., 2016) identified four metabolites with high discrimination between normal and pancreatic cancer patients. A more recent study on lung cancer (Moreno et al., 2018) found five metabolites for adenocarcinoma and two metabolites for squamous cell lung carcinoma that can discriminate between normal and lung tumor tissues. Other studies include detection of lung cancer using metabolic phenotyping of plasma (Louis et al., 2016a), and discrimination between lung and breast cancer also using metabolic phenotyping of plasma (Louis et al., 2016b).

The technology and research on cancer has given us a deeper understanding on how it works in the body and it has led us to several developments of different treatments, biomarkers, and methods of analysis. The use of metabolic profile in the analyses has also shed new light on cancer. Metabolomics is a relatively new field and the range of possibilities for further research and development is great as the science and technology continues to advance.

1.3 Aims and Objectives

The aim of this study is to determine the metabolites that are associated with the ^{18}F-FDG PET parameters: maximum standardized uptake value (SUV$_{max}$), total lesion glycolysis (TLG), and metabolic tumor volume (MTV), using penalized regression methods and multiple factor analysis.

2 Data Description

The data used in this study contains lung cancer patients (N = 222) included in the Limburg Positron Emission Tomography Center (Louis et al., 2016a) from March 2011 to June 2014. The patients were subjected to fluorine-18 fluorodeoxyglucose positron emission tomography (^{18}F-FDG PET) imaging and three PET parameters were obtained: (1) maximum standardized uptake value (SUV$_{max}$), (2) total lesion glycolysis (TLG), and (3) metabolic tumor volume (MTV).

In addition, to represent the metabolic profile of these patients, blood plasma samples were obtained and subjected to ^{1}H-NMR spectroscopy. The resulting ^{1}H-NMR spectra were preprocessed and divided into 110 regions, which were integrated and normalized, resulting to 110 normalized integration values, and these values represented the metabolic profile used in this study (Louis et al., 2015; Louis et al., 2016a).

Seven other variables, the clinical factors, were obtained from the patients, namely: age, gender, smoking habits, diabetes, glycemia, histology of lung cancer, and body mass index (BMI).

We denote the responses, the three PET parameters, as a 222 × 3 matrix $\boldsymbol{Y}$, the metabolic profile as a 222 × 110 matrix $\boldsymbol{X}$, and the clinical factors as a 222 × 7 matrix $\boldsymbol{Z}$. The data structure is shown below.

$$Y = \begin{bmatrix} Y_{1,1} & Y_{1,2} & Y_{1,3} \\ Y_{2,1} & Y_{2,2} & Y_{2,3} \\ Y_{3,1} & Y_{3,2} & Y_{3,3} \\ \vdots & & \\ Y_{222,1} & Y_{222,2} & Y_{222,3} \end{bmatrix}$$

$$\text{X} = \begin{bmatrix} X_{1,1} & X_{1,2} & X_{1,3} & \cdots & X_{1,110} \\ X_{2,1} & X_{2,2} & X_{2,3} & \cdots & X_{2,110} \\ X_{3,1} & X_{3,2} & X_{3,3} & \cdots & X_{3,110} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ X_{222,1} & X_{222,2} & X_{222,3} & \cdots & X_{222,110} \end{bmatrix} \text{Z} = \begin{bmatrix} Z_{1,1} & Z_{1,2} & Z_{1,3} & \cdots & Z_{1,7} \\ Z_{2,1} & Z_{2,2} & Z_{2,3} & \cdots & Z_{2,7} \\ Z_{3,1} & Z_{3,2} & Z_{3,3} & \cdots & Z_{3,7} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ Z_{222,1} & Z_{222,2} & Z_{222,3} & \cdots & Z_{222,7} \end{bmatrix}$$

The PET parameters and the characteristics of the patients are summarized in Table 1.

Table 1: *PET parameters and clinical characteristics of patients*

Variable	mean ± sd (range)
SUV$_{max}$	12.33 ± 7.42 (2.64 - 50.39)
TLG	750.31 ± 1280.70 (0.19 - 11594.91)
MTV	137.55 ± 244.63 (0.06 - 2131.26)
BMI (kg/m^2)	25.76 ± 4.72 (17.50 - 44.40)
Glycemia (mg/dl)	106 ± 21.19 (70.0 - 194.0)
Age (years)	68 ± 10 (43 - 88)
	n (%)
Sex	
Male	152 (68.5%)
Female	70 (31.5%)
Smoking status	
Yes	113 (50.9%)
No	5 (2.3%)
Former	104 (46.8%)
Histology	
ADENO	82 (36.9%)
SPINO	62 (27.9%)
SCLC	33 (14.9%)
NO	26 (11.7%)
NOS	9 (4.1%)
Other	10 (4.5%)
Diabetes	
Yes	40 (18%)
No	182 (82%)

3 Methods

The emergence of multi-omics data has resulted to the development of several statistical methods to analyze these type of data. When we want to fit a model, we might use a regression model with a subset of variables or predictors. In parallel with fitting a model is to evaluate its prediction accuracy or to provide an interpretation of its estimates. However, using ordinary least squares (OLS) regression has its drawbacks. One drawback is on its estimates, which often have low bias but large variance and affect the model's prediction accuracy. Another drawback is the model interpretation. We usually want to determine only a subset of predictors that are relevant and interpret the estimated coefficients. The method introduced by Tibshirani (1996) called LASSO, attempts to resolve these drawbacks by shrinking or setting to zero some coefficients, therefore, improving the prediction accuracy and performing variable selection at the same time. Another method is the Elastic Net, proposed by Zou and Hastie (2005), which also performs shrinkage of coefficients and variable selection simultaneously like LASSO, and in addition, selects groups of correlated variables. On the other hand, multiple factor analysis can be used when you have several different data sets coming from the same set of observations. This analysis investigates the relationship among these data sets (Abdi, Williams, and Valentin, 2013).

This section discusses the methods used to complete the objectives of this study, namely: the LASSO method (Section 3.2), the Elastic Net method (Section 3.3), the cross-validation (Section 3.4), and the Multiple Factor Analysis (Section 3.5). In addition, model-based hypothesis testing with multiplicity correction (Section 3.1), to determine which metabolites are associated with the PET parameters, is also used in the first part of the analysis.

3.1 Multiplicity Correction

Multiple testing refers to testing several hypotheses simultaneously. Similar to testing with only one hypothesis, multiple testing also comes with type I error, which refers to rejecting the null hypothesis when it is actually true. When we test for more than one hypothesis and the multiplicity of tests are not taken into account, the type I error increases (Romano, Shaikh, and Wolf, 2010; Sainani, 2009).

There are different correction methods to deal with multiplicity, such as the Bonferroni, Holm, and Benjamini and Hochberg (1995) multiplicity correction methods. The Bonferroni and Holm methods control the family-wise error rate (FWER), the probability of at least one type I error in a series of hypothesis tests, while the Benjamini-Hochberg method controls the false discovery rate (FDR), the expected proportion of errors among the rejected hypotheses (Benjamini and Hochberg, 1995). In this study, the Benjamini-Hochberg (BH) correction is used.

Let $\mathrm{H}_1, \mathrm{H}_2, \ldots, \mathrm{H}_m$ be the m hypotheses to be tested simultaneously, based on the corresponding p-values P_1, P_2, ..., P_m. Let $\mathrm{P}_{(1)} \leq \mathrm{P}_{(2)} \leq \ldots \leq \mathrm{P}_{(m)}$ be the ordered p-values and $\mathrm{H}_{(i)}$ be the null hypothesis corresponding to $\mathrm{P}_{(i)}$. The BH procedure is as follows:

1. Let k be the largest i such that $\mathrm{P}_{(i)} \leq \frac{i}{m}\alpha$.
2. Reject all $\mathrm{H}_{(i)}$ for $i = 1, 2, \ldots, k$.

for $m = 1, 2, \ldots, 110$. The procedure controls the FDR at α.

3.2 LASSO

The Least Absolute Shrinkage and Selection Operator (LASSO) is a penalized regression method (Tibshirani, 1996), where some coefficients are shrunk and others are set to zero. Thus, performing a variable selection by including only those variables with non-zero coefficients in the model.

Let $\boldsymbol{X}_i$ be the vector of the p predictor variables (metabolites) and $\boldsymbol{Y_i}$ are the responses, for $i = 1, 2, \ldots, 222$. The LASSO parameter estimate $(\hat{\alpha}, \hat{\beta})$ is defined as:

$$(\hat{\alpha}, \hat{\beta}) = \arg\min \left\{ \sum_{i=1}^{222} \left(Y_i - \alpha - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2 \right\} \quad \text{subject to} \sum_j |\beta_j| \leq t. \tag{1}$$

Equation 1 is equivalent to

$$\sum_{i=1}^{222} \left(Y_i - \alpha - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{2}$$

where λ is called the penalty or tuning parameter. This parameter controls the amount of shrinkage that is applied to the coefficients. If $\lambda = 0$, then then we have the OLS parameter estimates, while if $\lambda > 0$, then some coefficients will be set exactly to zero, thus performing variable selection.

Now, since we also want to include the clinical variables in the model, we want the LASSO method to retain these variables in the model. Let $\boldsymbol{X}_i$ be the vector of the $p - k$ predictor variables (metabolites), let $\boldsymbol{Z}_i$ be the vector of k clinical variables, and $\boldsymbol{Y_i}$ are the responses. Then equation (2) becomes

$$\sum_{i=1}^{222}\left(Y_i - \alpha - \sum_{j=1}^{p-k}\beta_j X_{ij} + \sum_{j=p-k+1}^{p}\beta_j Z_{ij}\right)^2 + \lambda\sum_{j=1}^{p-k}|\beta_j| \tag{3}$$

where the k predictor variables are not penalized.

3.3 Elastic Net

A method related to LASSO is Elastic Net (Zou and Hastie, 2005), which is a combination of ridge regression (not discussed here) and LASSO. Aside from performing variable selection and shrinkage of coefficients, Elastic Net also selects groups of correlated variables.

Let $\boldsymbol{X_i}$ be the vector of the p predictor variables (metabolites) and $\boldsymbol{Y_i}$ be the responses, for $i = 1, 2, \ldots, 222$. The Elastic Net estimates are the minimizers to:

$$\sum_{i=1}^{222}\left(Y_i - \alpha - \sum_{j=1}^{p}\beta_j X_{ij}\right)^2 + \lambda\left(\alpha\sum_{j=1}^{p}|\beta_j| + (1-\alpha)\sum_{j=1}^{p}\beta_j^2\right) \tag{4}$$

where λ is the penalty parameter and α is the mixing parameter. This is a mixture of the ridge regression and LASSO penalties since if $\alpha = 0$, then we have the ridge regression, while if $\alpha = 1$, then we have the LASSO, and a value for α between 0 and 1 is a blend of the two methods.

Since we want the clinical variables to be also included in the model, we do not want them to be penalized by the method. Let $\boldsymbol{X}_i$ be the vector of the $p - k$ predictor variables (metabolites), let $\boldsymbol{Z}_i$ be the vector of k clinical variables, and $\boldsymbol{Y_i}$ be the responses. Then equation (4) becomes

$$\sum_{i=1}^{222}\left(Y_i - \alpha - \sum_{j=1}^{p-k}\beta_j X_{ij} + \sum_{j=p-k+1}^{p}\beta_j Z_{ij}\right)^2 + \lambda\left(\alpha\sum_{j=1}^{p-k}|\beta_j| + (1-\alpha)\sum_{j=1}^{p-k}\beta_j^2\right) \tag{5}$$

where the k predictor variables are not penalized.

3.4 Cross-Validation

Cross-validation (CV) is a method used to evaluate the predictive performance of a model, usually by estimating the misclassification error rate for binary responses or the mean squared error (MSE) or the root mean squared error (RMSE) for continuous responses. It also addresses the problems of over-fitting and selection bias in the model (Cawley and Talbot, 2010) by fitting a model to the training set (known data) and assessing the model's predictive ability on the test set (unknown or unseen data). Ideally, we want our training set to be large enough and have a separate data for our test set. However, due to the constraints of a non-ideal world, such as limited data and limited resources to collect more data (Raschka, 2018), we usually only have one data set to work with and so this data set is partitioned into the training and the test set.

One of the most common types of cross-validation is the k-fold cross validation, where the data is split into k parts of approximately equal size: one part as the test set and the remaining $k-1$ parts as the training set. Another type of cross-validation is the leave-one-out cross-validation (LOOCV), which is a special case of the k-fold cross-validation, when $k = n$ (the number of observations). Here, the $n-1$ observations are used to fit a model and the remaining one observation is used to predict and evaluate the model's performance. In this study, the 3-fold cross-validation is used: the training set is 2/3 of the data and the test set is the remaining 1/3 of the data.

Furthermore, to obtain a reliable estimate of the model's performance, the cross-validation is repeated a large number of times. Krstajic, Buturovic, Leahy, and Thomas (2014) showed that repetition is important to have a more reliable model assessment than performing the cross-validation once. In a repeated k-fold CV, a different subset of the observations are assigned to the training set and the test set for every iteration. In this study, the 3-fold CV is repeated 1000 times. So, the performance of the model is evaluated 1000 times for 1000 random selections of lung cancer patients. The entire cross-validation procedure used in this study is summarized in Figure 1.

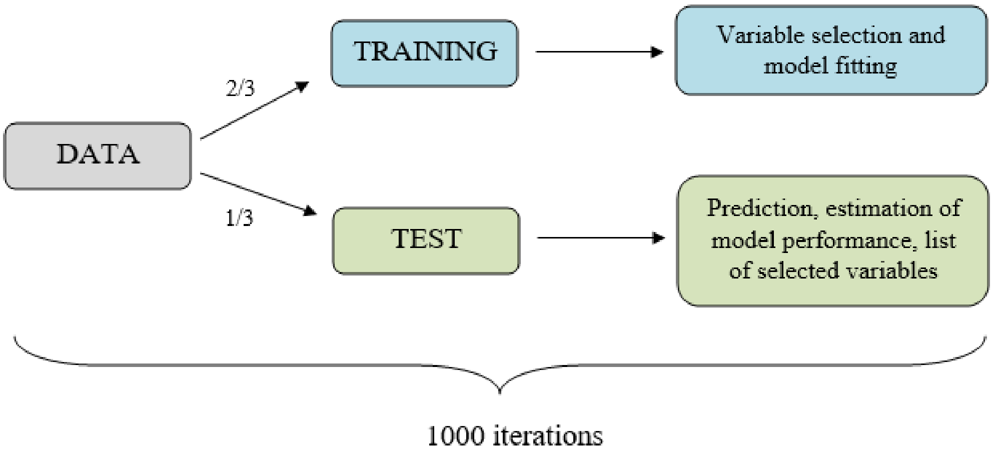

DATA
2/3
1/3
TRAINING
TEST
Variable selection and model fitting
Prediction, estimation of model performance, list of selected variables
1000 iterations

Figure 1: *Cross-validation procedure*

3.5 Multiple Factor Analysis

Multiple factor analysis (MFA) is a generalization of the principal component analysis (PCA), where several different data sets obtained from the same set of observations can be analyzed together (Abdi et al., 2013). The goal of MFA is to integrate these different data sets and examine the relationship between the observations, the variables, and the data sets (Abdi et al., 2013; Kasim, Shkedy, Kaiser, Hochreiter, and Talloen, 2016).

The first step in MFA is the normalization of each of the individual data sets to make them comparable. This step is needed so that the length of the first principal component (singular value) of each data set is equal to 1 and therefore no data set can dominate the common structure (Abdi et al., 2013; Kasim et al., 2016). Let $\boldsymbol{X_1}$ be the matrix of the combined vectors of the three responses (SUV$_{max}$, TLG, and MTV), with a dimension of 222×3, and $\boldsymbol{X_2}$ be the matrix of the metabolites, with a dimension of 222×110. To normalize, each matrix is divided by its first singular value, we can denote as ${}_1\varphi_1$, which is the square root of the first eigenvalue of the PCA. Then the two normalized matrices are given by:

$$\boldsymbol{Z_1} = {}_1\varphi_1^{-1} \times \boldsymbol{X_1} \tag{6}$$

$$\boldsymbol{Z_2} = {}_2\varphi_1^{-1} \times \boldsymbol{X_2} \tag{7}$$

The normalized matrices are then combined into one large matrix, $\boldsymbol{Z} = [\boldsymbol{Z_1}|\boldsymbol{Z_2}]$. The number of rows (samples) is still 222, which is the common dimension between the two matrices, and the number of columns (variables) is now 113. The next step is to perform a PCA on $\boldsymbol{Z}$. In MFA, just like in PCA, we can determine the importance of a principal component (dimension) by how much variability in the data is explained by this component, the factor scores of the observations on a component, and the factor loadings of the variables on a component that can be used to plot graphs to help us analyze the relationships.

4 Results and Interpretation

4.1 Preliminary model fitting

Prior to fitting LASSO and Elastic Net to the data, a linear model was fitted for each response with the clinical factors as the predictor variables. This was to keep only the clinical variables that have a significant effect on the responses and use them in the subsequent analyses. Table 2 shows the estimates with their standard errors and p-values.

Table 2: *Parameter estimates of the clinical variables for SUV_{max}, TLG, and MTV response variables*

	SUV_{max}		TLG		MTV	
Parameter	Estimate (s.e.)	p-value	Estimate (s.e.)	p-value	Estimate (s.e.)	p-value
Intercept	10.4770 (5.1144)	0.0418	835.5307 (874.0638)	0.3403	178.4777 (164.2581)	0.2785
BMI	0.0625 (0.1151)	0.5878	0.7627 (19.6745)	0.9691	-1.9195 (3.6973)	0.6042
Glycemia	-0.0305 (0.0250)	0.2235	-1.2934 (4.2697)	0.7623	0.2533 (0.8024)	0.7526
Sex (Male)	-0.7113 (1.1559)	0.5390	270.2701 (197.5358)	0.1728	61.0352 (37.1218)	0.1017
Smoking (No)	-5.3245 (4.5141)	0.2396	-28.0643 (771.4645)	0.9710	43.3544 (144.9772)	0.7652
Smoking (Yes)	-1.2923 (1.0775)	0.2318	90.3649 (184.1401)	0.6241	22.8343 (34.6044)	0.5101
Histo (NO)	-3.1464 (1.7811)	0.0788	-197.6910 (304.3884)	0.5168	-25.1193 (57.2021)	0.6610
Histo (NOS)	2.5098 (2.6223)	0.3397	959.4247 (448.1522)	0.0335*	179.6270 (84.2188)	0.0341*
Histo (SCLC)	0.0147 (1.5489)	0.9924	951.8338 (264.7056)	0.0004*	217.1066 (49.7447)	2.04e-05*
Histo (SPINO)	2.1687 (1.3306)	0.1047	174.4087 (227.4071)	0.382	16.4141 (42.7354)	0.7013
Age	0.0655 (0.0597)	0.2737	-5.7308 (10.1936)	0.5746	-1.6316 (1.9156)	0.3954

*p-value ≤ 0.05

No significant clinical factors are associated with SUV_{max}, while *Histology* factor levels *NOS* and *SCLC* are found to be significantly associated with both TLG and MTV. Therefore, the *Histology* factor is included in the models for TLG and MTV later.

4.2 Univariate tests

Univariate tests for the metabolites were also performed to get an idea which metabolites are relevant. This is a feature by feature analysis to determine significant metabolites by fitting a model for each PET parameter with one metabolite and clinical factors as the predictor variables. We have the model

$$Y_i = \alpha + \beta_j X_{ij} + \sum_{k=1}^{K} \delta_k Z_{ik} \tag{8}$$

where the same denotations as in Section 2 are used.

Here, the clinical factors included were the ones found significant in Section 4.1. There were 110 models simultaneously fitted for each response, because we have 110 metabolites, so the

Benjamini-Hochberg (BH) multiplicity correction was used. Using a significance level $\alpha = 0.05$, 5 metabolites were significant in explaining SUV$_{max}$, 15 significant metabolites in explaining TLG, and 21 significant metabolites in explaining MTV. Table 3 shows the significant metabolites for each response variable. The metabolites that are consistently significant in the three models are: VAR48, VAR10, and VAR49, while all significant metabolites in TLG are also in MTV. This is expected since TLG and MTV are functionally related.

Table 3: *Complete list of significant metabolites after multiplicity correction*

SUV$_{max}$		TLG		MTV	
Metabolite	p-value	Metabolite	p-value	Metabolite	p-value
VAR13	0.0017	VAR91	0.0015	VAR48	0.0019
VAR48	0.0155	VAR10	0.0015	VAR23	0.0019
VAR9	0.0155	VAR48	0.0016	VAR49	0.0019
VAR10	0.0155	VAR49	0.0016	VAR91	0.0027
VAR49	0.0281	VAR11	0.0020	VAR50	0.0034
		VAR50	0.0037	VAR10	0.0034
		VAR23	0.0037	VAR11	0.0034
		VAR65	0.0188	VAR45	0.0223
		VAR109	0.0219	VAR108	0.0227
		VAR108	0.0219	VAR46	0.0232
		VAR46	0.0252	VAR109	0.0232
		VAR106	0.0281	VAR65	0.0261
		VAR45	0.0281	VAR30	0.0261
		VAR107	0.0281	VAR33	0.0282
		VAR47	0.0310	VAR106	0.0282
				VAR47	0.0282
				VAR107	0.0282
				VAR8	0.0290
				VAR37	0.0290
				VAR36	0.0333
				VAR38	0.0377

The volcano plots of each response is shown in Figure 2. Volcano plots are scatter plots that can be used to visualize and to identify data points that are statistically significant. It plots the statistical significance on the y-axis and the measurement of a statistical signal (e.g. fold change) on the x-axis (Cui and Churchill, 2003). The gray points represent the significant metabolites without multiplicity correction, while the red points represent the metabolites that are still significant after multiplicity correction, and the black points are the non-significant metabolites (Figure 2).

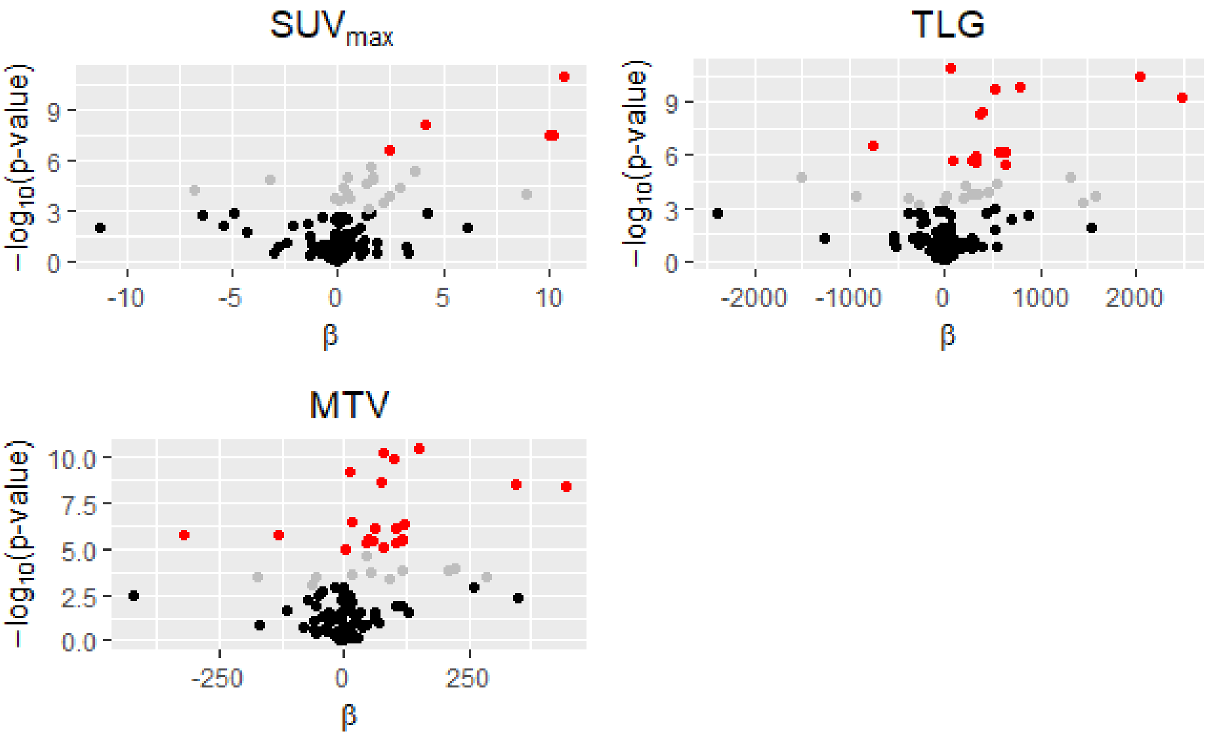


Figure 2: *Volcano plots of the three PET parameters*

4.3 Analysis using LASSO and Elastic Net

The LASSO and Elastic Net model fitting and validation were repeated 1000 times as indicated in Figure 1. For every iteration, the frequency of selection of variables, correlation of observed and predicted values, and MSE were recorded. The clinical factors included for the models for the three responses were mentioned in Section 4.1, and these variables were not penalized and were retained in the models throughout the cross-validation procedure.

Figure 3 shows the distribution of the correlation between the observed and predicted values of SUV$_{max}$. We can see that the values of the correlation are centered around 0.3, both for LASSO and Elastic Net. For a correlation, this value indicates a low positive correlation between the observed and predicted values. On the other hand, the distributions of the MSE of the two methods look similar and are centered around 50.

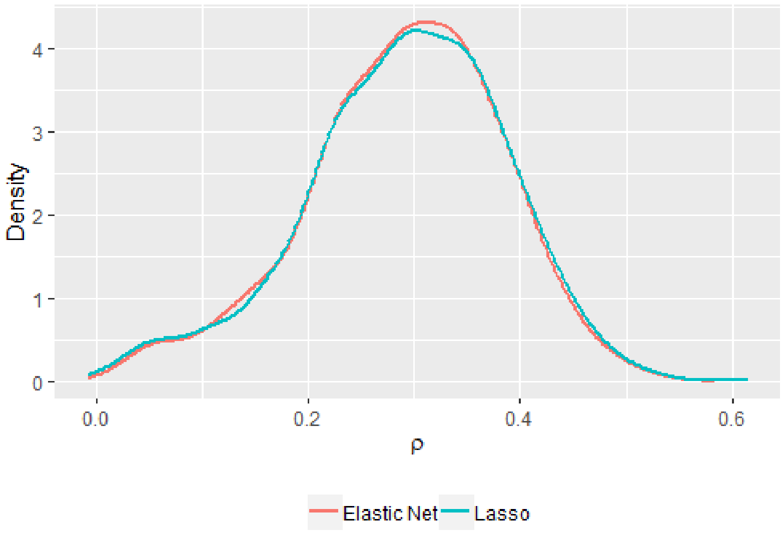


Figure 3: *Density of the correlation between observed and predicted values of* SUV_{max}

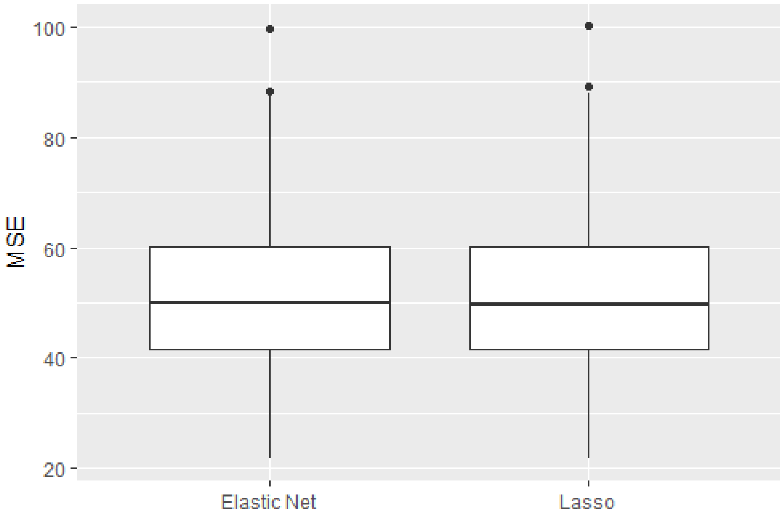


Figure 4: *Boxplots of MSE of* SUV_{max}

For TLG, the correlation between the observed and the predicted values are centered around 0.4 and its distribution looks similar for the two methods (Figure 5). Similarly, the distribution of the MSEs estimated from the two methods does not seem to differ from each other.

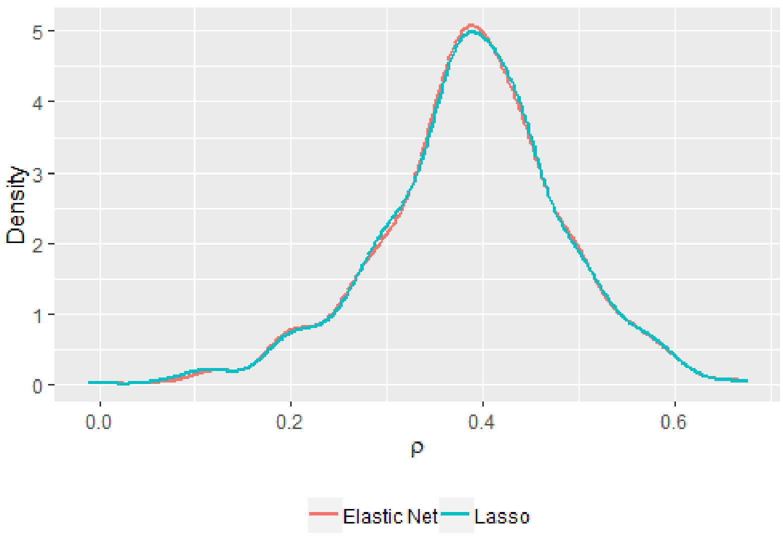


Figure 5: *Density of the correlation between observed and predicted values of TLG*

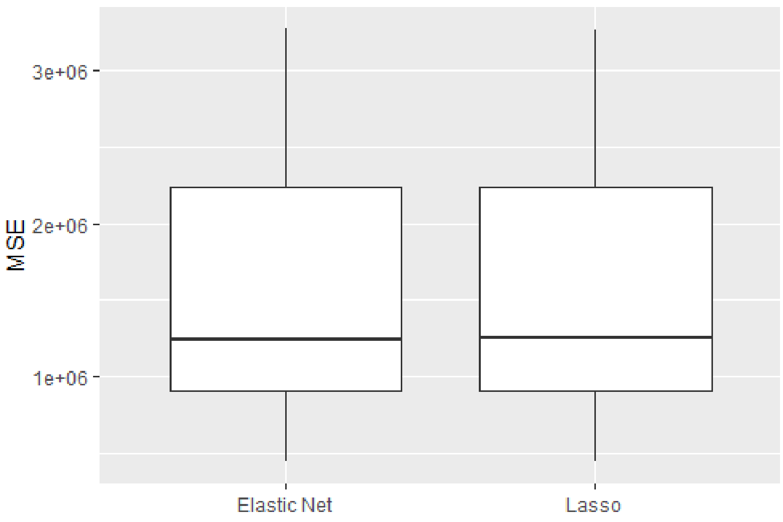


Figure 6: *Boxplots of MSE of TLG*

In Figure 7, similar with TLG, the distribution of the estimated correlation of the observed and the predicted values of MTV between the two methods are similar. The same can be said about the distribution of the MSE obtained from the two methods as shown in Figure 8.

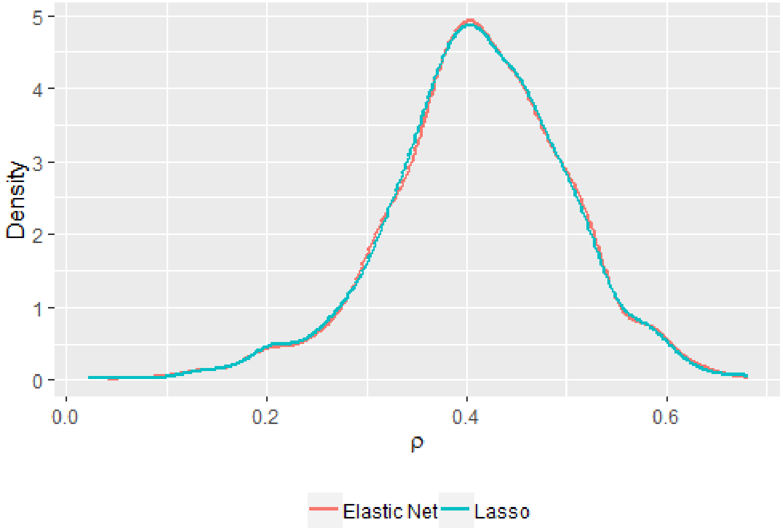


Figure 7: *Density of the correlation between observed and predicted values of MTV*

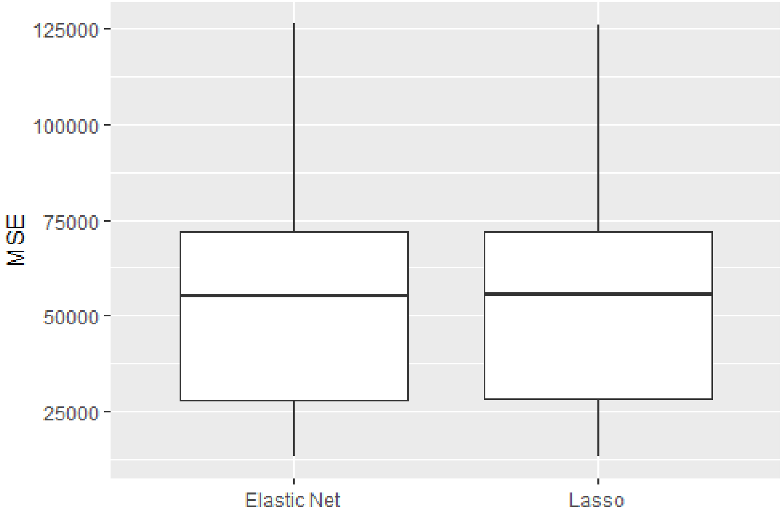


Figure 8: *Boxplots of MSE of MTV*

As mentioned in Sections 3.2 and 3.3, variable selection is performed by shrinking some coefficients of the variables to 0. To illustrate what LASSO and Elastic Net do, Figure 9 shows the shrinkage effect of the two methods. We can see that some of the coefficients from the univariate analysis are set to 0 in LASSO and Elastic Net.

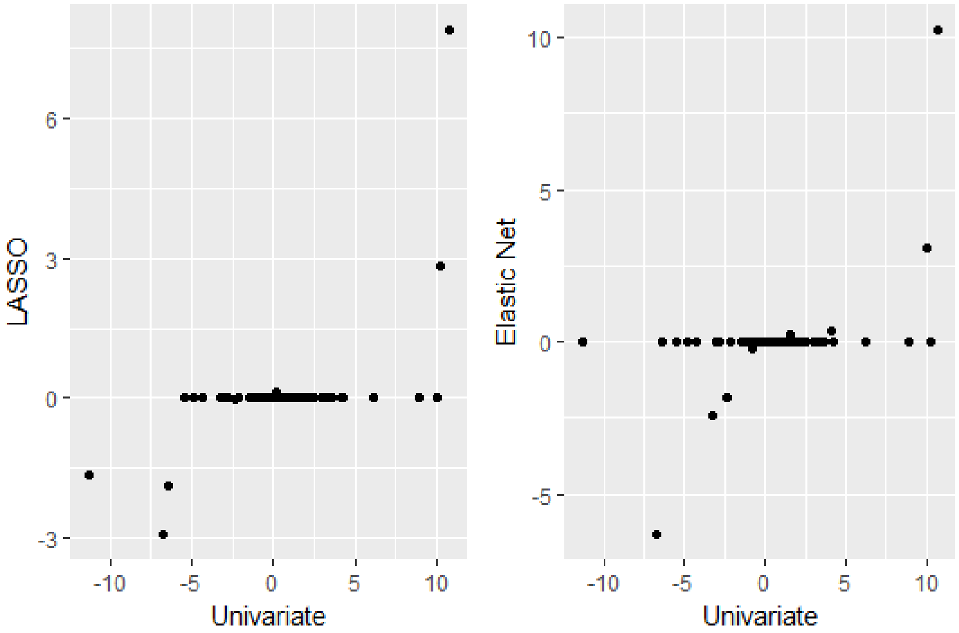


Figure 9: *Shrinkage effect: LASSO and Elastic Net versus univariate coefficients*

The frequency of selection of metabolites gives us an idea which metabolites are selected more than the others and indicates which metabolites are relevant. The top metabolites for each response for both methods are presented in Table 4. Both methods selected almost the same metabolites more frequently, and because of this, we observed similarities in the MSE and the correlation distributions between the two methods as shown in the previous plots. Furthermore, the Elastic Net selects more metabolites than LASSO since Elastic Net takes into account the correlation among the variables and tends to select these groups of (highly) correlated variables together in the model.

Some of these metabolites were also found to be significant from the univariate tests in Section 4.2. However, this does not mean that the other metabolites selected by LASSO and Elastic Net that were not significant in the univariate tests are not relevant. The significant metabolites that were also selected by LASSO or Elastic Net tell us that these metabolites are indeed important and are associated with the PET parameters.

Table 4: *Top metabolites selected more than 300 times*

	SUV$_{max}$	TLG	MTV
LASSO	VAR13, VAR10, VAR74, VAR106, VAR91, VAR9, VAR76, VAR75, VAR12, VAR48	VAR65, VAR91, VAR10, VAR23, VAR11, VAR48, VAR107, VAR2, VAR30, VAR28, VAR71, VAR15, VAR8, VAR22, VAR7, VAR80, VAR79	VAR23, VAR65, VAR91, VAR48, VAR30, VAR2, VAR11, VAR10, VAR8, VAR71, VAR15, VAR22, VAR28, VAR107, VAR32, VAR17, VAR33
Elastic Net	VAR13, VAR10, VAR74, VAR106, VAR9, VAR76, VAR91, VAR75, VAR12, VAR48, VAR49, VAR61	VAR65, VAR91, VAR10, VAR23, VAR11, VAR48, VAR107, VAR2, VAR30, VAR71, VAR28, VAR15, VAR8, VAR22, VAR7, VAR79, VAR80, VAR13, VAR33	VAR65, VAR23, VAR91, VAR48, VAR30, VAR11, VAR2, VAR10, VAR8, VAR71, VAR15, VAR107, VAR22, VAR28, VAR32, VAR33, VAR17, VAR64, VAR79

4.4 Analysis using MFA

MFA is applied to the combined data sets of the PET parameters and the metabolites. Table 5 shows the percentage of contribution of a data set to the component/factor. The larger the contribution, the more it contributes to the component (Kasim et al., 2016). As we can see, 54.88% of the variance of the first factor can be attributed to the PET data set and the rest can be attributed to the metabolites data set. For the second factor, 62.03% of its variance can be attributed to the metabolites data set.

Table 5: *Data set contribution to each factor*

Data set	Factor 1	Factor 2
PET	54.88	37.97
metabolites	45.12	62.03

The factor loadings for the first and second factors are illustrated in Figures 10 and 11. Those in red are the variables with relatively low and high loadings and are related to the first and second factors. We can see that the MTV and the TLG loadings are close to each other which indicates their almost equal contribution to the variance explained in the first factor. This is expected since the TLG is the product of SUV and MTV. The first factor can be called the PET factor. Moreover, the metabolites with relatively high positive and negative loadings are shown in Table 6. These metabolites are associated with the PET factor. In particular, VAR38 (0.68), VAR48 (0.68), VAR49 (0.69), VAR50 (0.70), and VAR100 (-0.65) have the highest loadings and are highly correlated with the PET factor. On the other hand, the second factor is mainly driven by the metabolites and can be called the metabolites factor. In particular, VAR86 (-0.65), VAR87 (-0.70), VAR88 (-0.67), and VAR100 (-0.72) have the highest loadings and are highly correlated with the metabolites factor.

Moreover, metabolites VAR38, VAR48, and VAR49 were also found to be significant from the univariate tests and were also selected more frequently than the other metabolites from LASSO and Elastic Net. This gives us another view on which metabolites are associated and how they are associated together with the PET parameters, particularly with MTV and TLG.

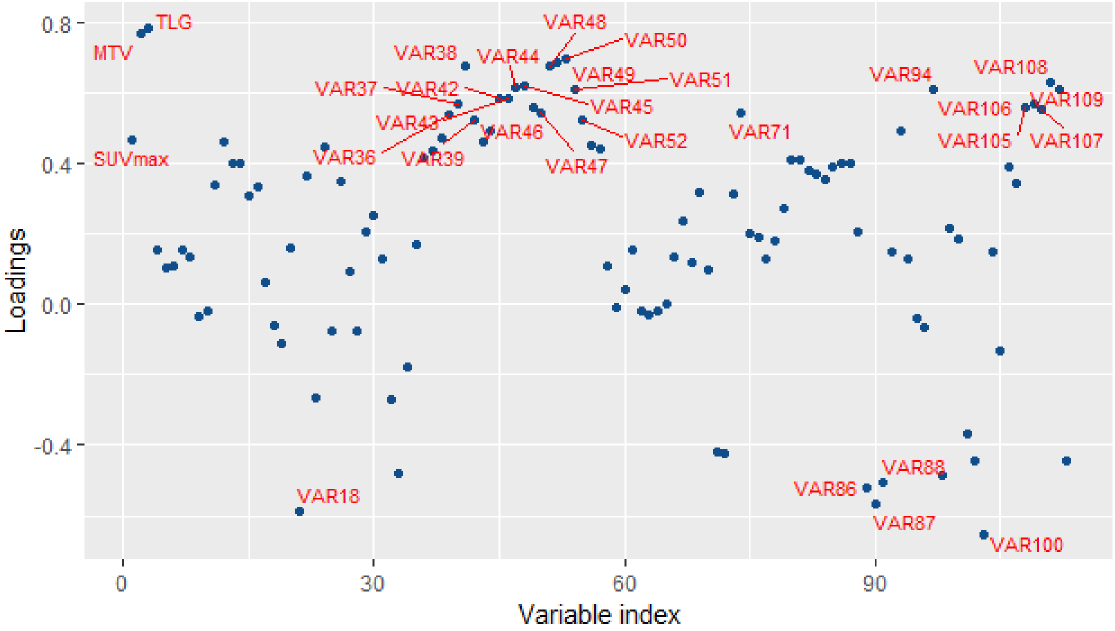


Figure 10: *Factor loadings of the variables for the first factor*

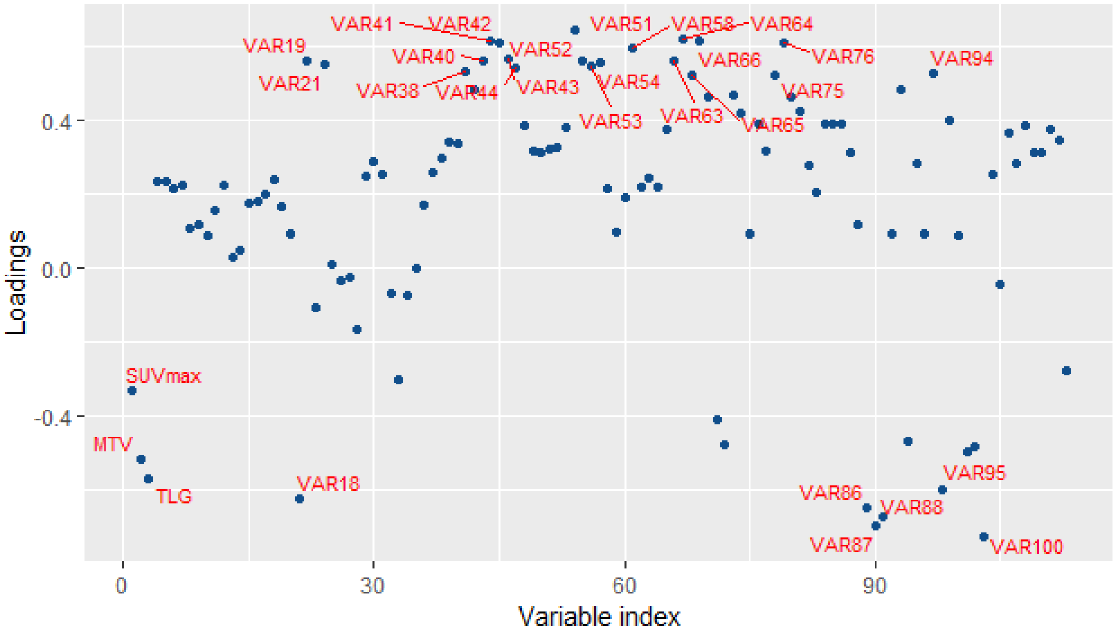


Figure 11: *Factor loadings of the variables for the second factor*

Table 6: *Factor loadings ($\geq |0.5|$) for the first and second factor*

Variable	Factor 1	Factor 2
SUV_{max}	0.4690	-0.3314
MTV	0.7704	-0.5173
TLG	0.7857	-0.5714
VAR18	-0.5865	-0.6206
VAR19		0.5581
VAR21		0.5472
VAR36	0.5369	
VAR37	0.5723	
VAR38	0.6750	0.5316
VAR39	0.5220	
VAR40		0.5591
VAR41		0.6103
VAR42	0.5854	0.6071
VAR43	0.5859	0.5662
VAR44	0.6144	0.5393
VAR45	0.6193	
VAR46	0.5572	
VAR47	0.5439	
VAR48	0.6776	
VAR49	0.6869	
VAR50	0.7002	
VAR51	0.6134	0.6424
VAR52	0.5245	0.5575
VAR53		0.5439
VAR54		0.5564
VAR58		0.5928

(*table continues*)

Variable	Factor 1	Factor 2
VAR63		0.5587
VAR64		0.6162
VAR65		0.5198
VAR66		0.6136
VAR71	0.5424	
VAR75		0.5215
VAR76		0.6063
VAR86	-0.5182	-0.6457
VAR87	-0.5671	-0.6969
VAR88	-0.5037	-0.6717
VAR94	0.6121	0.5249
VAR95		-0.5984
VAR100	-0.6532	-0.7244
VAR105	0.5578	
VAR106	0.5682	
VAR107	0.5564	
VAR108	0.6334	
VAR109	0.6086	

The factor scores for the first and second factor are illustrated in Figures 12 and 13. We can observe that patients 101 and 212 are the ones dictating the variance of the PET factor, as they have the highest factor scores. On the other hand, patient 212 showed the lowest factor score for the metabolites factor.

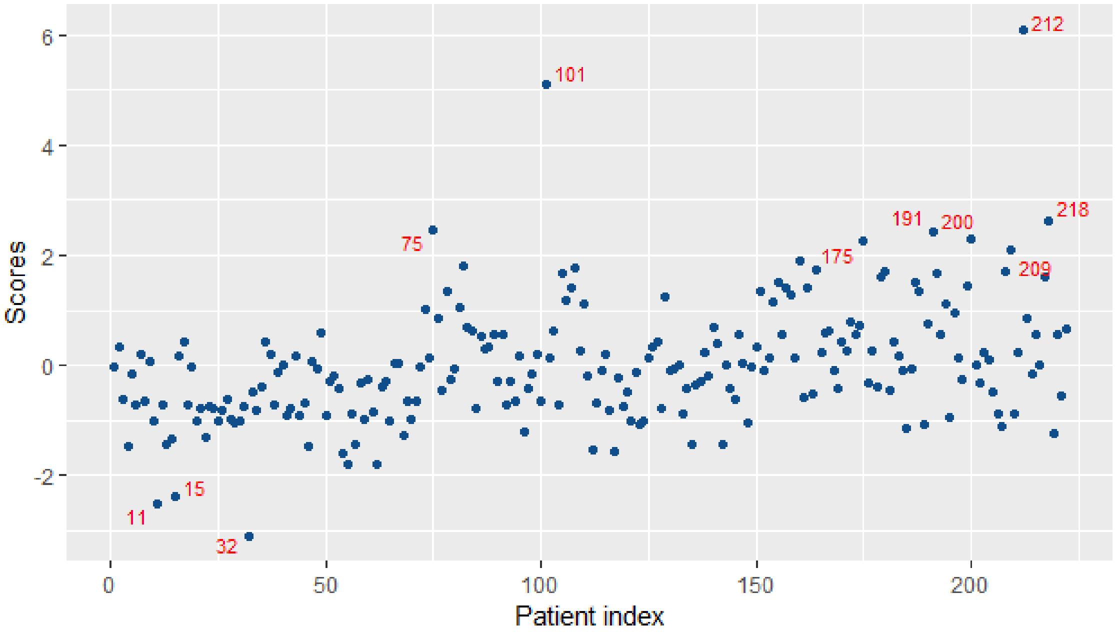


Figure 12: *Factor scores of the patients for the first factor*

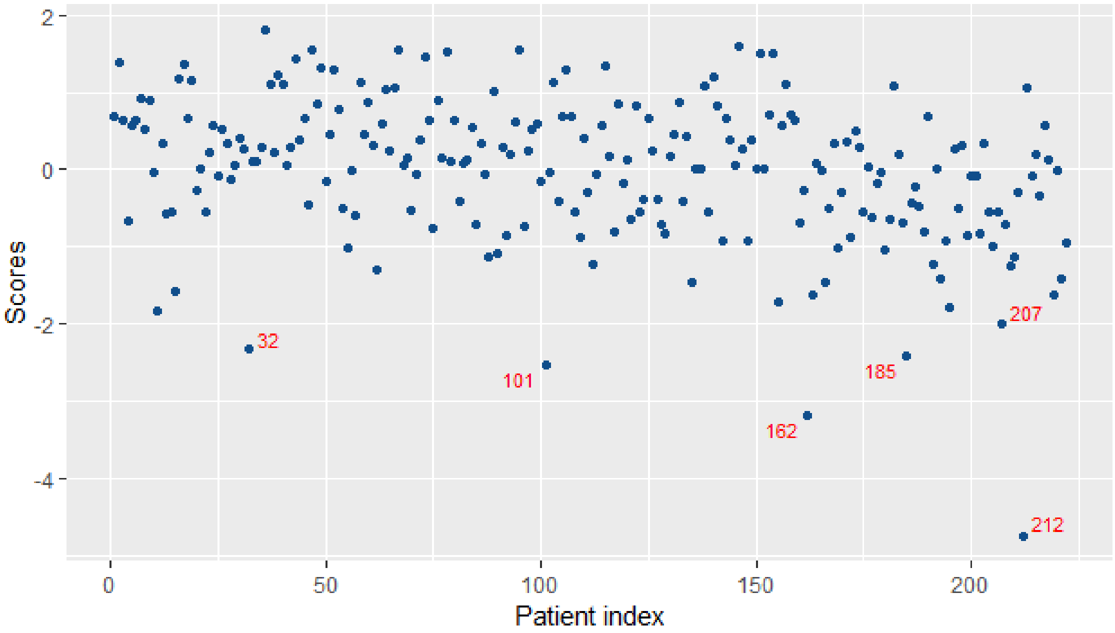


Figure 13: *Factor scores of the patients for the second factor*

5 Discussion and Conclusion

The use of PET with ^{18}F-FDG in the early diagnosis of cancer has become a standard component of diagnosis and staging in oncology. In PET, parameters such as SUV, TLG, and MTV are estimated to help assess cancer cells activities.

There are many factors that can be attributed to human diseases. The association of these factors with diseases can be explored as a way for an early disease detection and subsequent appropriate medical management. The association of various metabolites with some major diseases have been studied (Di Gangi et al., 2016; Louis et al., 2016a, 2016b; Moreno et al., 2018; Trivedi et al., 2017). This was the focus of this study, particularly in exploring metabolites and lung cancer association.

As there were several information available for analysis, the clinical information of patients in the study were first analyzed. Among the clinical information gathered, patients with increased level of glycemia were found to have lower SUV$_{max}$ compared with patients with lower glycemia levels though the association was not significant. This is consistent with previous works. Oh et al. (2014) in their study of colorectal cancer patients found similar association of decreasing SUV with increasing plasma glucose content in patients, relationship was also found insignificant. Gorenberg, Hallett, and O'Doherty (2002) in their study of lung cancer demonstrated however that there was no difference in log(SUV) between diabetic and non-diabetic patients.

Basically, blood sugar of patients is an important information before conducting PET scan. Glucose level in patients may compete with FDG transporters during tumor uptake in doing PET scan rendering uptake ineffective. This is the reason why SUV inversely correlates with blood sugar. Therefore, for an oncologic FDG-PET scan, fasting is required for patients to reduce blood sugar level (Lee et al., 2005). Fasting also reduces insulin levels and thus lowers the activity of glucose transporters in the surrounding tissues and muscles to further enhance FDG uptake in the tumors relative to its surroundings (Tenley, Corn, Yuan, and Lee, 2013).

A positive significant association of NOS with MTV and TLG was found in the study. These findings support the fact that these two PET parameters are associated with tumor volume and sites manifesting highest metabolic activities. NOS or "not otherwise specified" correlating with MTV and TLG reflects sensitivity of these PET parameters with accurate tumor prognosis and subsequent overall survival of patients from cancer. Studies have shown that TLG has high potentiality in esophageal cancer prognosis than MTV and SUV, indicating higher sensitivity and specificity for predicting overall survival among patients sampled (Hong et al., 2016). TLG and MTV were identified to be potential sensitive markers for tumor burden in patients with recurrent SCLC (small cell lung cancer) (Shi et al., 2015).

There was also a report of MTV and TLG correlating better with histopathological response in NSCLC (non-small cell lung cancer) compared to SUV$_{max}$ (Burger et al., 2016; Han et al., 2015; Huang et al., 2014). Similarly, changes in overall tumor response during chemotherapy had direct impact to MTV and TLG values (Larson et al., 1999).

MTV and TLG were also found associated with SCLC in our analysis. Shi et al. (2015) found that these two PET parameters are strongly correlated with SCLC through the neuron specific enolase (NSE). As neuroendocrine differentiation is considered to be an important feature of SCLC, NSE has been utilized as a marker for its diagnosis and therapeutic monitoring (Ono et al., 2012; Shi et al., 2015). In recurrent colorectal cancer patients, a correlation has been found between serum carcinoembryonic antigen (CEA) and metabolic tumor volume (MTV), as determined by FDG PET (Choi et al., 2005).

The main objective of the study is to correlate PET parameters with the metabolic profiles of the sampled lung cancer patients. Post analysis to determine the association of patients' clinical information with SUV$_{max}$, MTV, and TLG was carried out via univariate tests to further determine the metabolites in the patients' profile that are closely linked with these parameters. High numbers of metabolites were associated with MTV (21 metabolites), followed by TLG (15) and with SUV$_{max}$, the least (5).

LASSO and Elastic Net, on the other hand, fit a model with all the metabolites as the predictor variables, and perform variable selection by penalizing the predictor variables for being high-dimensional. The relevant clinical factors were also integrated in both methods and are not penalized so they were always selected. Both methods show similar results for the first few top metabolites selected, which tell us that these metabolites are indeed associated with SUV$_{max}$, TLG, and MTV. Moreover, more metabolites were selected by Elastic Net than LASSO since Elastic Net selects the group of metabolites that are correlated and do not select only one metabolite from the group like LASSO does (Tibshirani, 1996; Zou and Hastie, 2005). This tells us that these metabolites function together and interact with each other, affecting the metabolic tumor activity inside the body. On the other hand, the predictive performance of both methods are similar and is relatively low in terms of the correlation between the observed and predicted values. Therefore, these metabolites cannot be used to predict the values of the PET parameters.

Similar findings of non-correlation between lung cancer progression and blood plasma metabolites were reported by Lokhov, Trifonova, Maslov, and Archakov (2013) even though such metabolites were present in diseased patients. However in the approach used by Hori et al. (2011), it was demonstrated through partial least squares discriminant analysis that changes in metabolite pattern are useful for assessing the clinical characteristics of lung cancer.

MFA, on the other hand, revealed that VAR38, VAR48, VAR49, VAR50, VAR100 are correlated, based on their factor loadings, with the first factor, which is controlled by the PET parameters, specifically by TLG and MTV only. SUV$_{max}$, on the other hand, has low factor loading and was not strongly correlated with the first factor. This result further supported that TLG and MTV are more sensitive parameters to consider in lung cancer prognosis than SUV$_{max}$ that warrants further study. Though results through MFA only provide an exploratory view of the relationship between these variables, the outcome warrants further analysis of the metabolites and their function in lung cancer prognosis using, possibly, other multivariate analysis. It is believed that by doing so, a more conclusive result on their association with the PET parameters can be achieved.

A number of metabolites was determined associated with PET parameters following various analysis procedures. As by design, the analyses used yielded different outcomes, except for LASSO and Elastic Net since both have almost similar functionalities. The univariate test is shown to be the simplest procedure to determine variables that are linearly related to PET. The limitation however is by the model design itself as it eliminates both additive and interactive relationships between metabolites. On the other hand, the MFA procedure has provided additional view of how PET and metabolites interact in a multivariate dimension but lacks predictive ability.

The results that were obtained from the study have provided insights on the potentiality of using the metabolic profile in analyzing cancer activity which open up possibilities for further research on how metabolites can be evaluated as probable biomarkers in lung cancer detection.

6 References

1. Abdi, H., Williams L.J., & Valentin D. (2013). Multiple factor analysis: principal component analysis for multitable and multiblock data sets. *WIREs Computational Statistics, 5(2)*, 149-179.

2. Adams M.C., Turkington T.G., Wilson J.M., & Wong T.Z. (2010). A Systematic Review of the Factors Affecting Accuracy of SUV Measurements. *American Journal of Roentgenology, 195(2)*, 310-320.

3. Arruebo M., Vilaboa N., Sáez-Gutierrez B., Lambea J., Tres A., Valladares M., & González-Fernández Á. (2011). Assessment of the Evolution of Cancer Treatment Therapies. *Cancers, 3(3)*, 3279–3330.

4. Bailey, D.L., Townsend, D.W., Valk, P.E., & Maisey M.N. (Eds.). (2005). *Positron-Emission Tomography: Basic Sciences.* Secaucus, NJ: Springer-Verlag.

5. Benjamini, Y. & 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), 57(1)*, 289-300.

6. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians, 68*, 394-424.

7. Burger, I.A., Casanova R., Steiger S., Husmann, L., Stolzmann, P., Huellner, M.W., . . . Soltermann, A. (2016). FDG-PET/CT of non-small cell lung carcinoma under neoadjuvant chemotherapy: background based adaptive volume metrics outperform TLG and MTV in predicting histopathological response. *Journal of Nuclear Medicine, 57(6)*, 849–854.

8. Cawley, G.C. & Talbot, N.L.C. (2010). On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *Journal of Machine Learning Research, 11*, 2079-2107.

9. Cerfolio, R.J, Bryant, A.S, Winokur, T.S, Ohja, B., & Bartolucci, A.A. (2004). Repeat FDG-PET After Neoadjuvant Therapy is a Predictor of Pathologic Response in Patients With Non-Small Cell Lung Cancer. *The Annals of Thoracic Surgery, 78(6)*, 1903 - 1909.

10. Chang, K.P., Tsang, N.M., Liao, C.T., Hsu, C.L., Chung, M.J., Lo, C.W., . . . Yen, T.C. (2012). Prognostic Significance of ^{18}F-FDG PET Parameters and Plasma Epstein-Barr Virus DNA Load in Patients with Nasopharyngeal Carcinoma. *The Journal of Nuclear Medicine, 53(1)*, 21-28.

11. Choi, M.Y., Lee, K.M., Chung, J.K., Lee, D.S., Jeong, J.M., Park, J.G., . . . Lee, M.C. (2005). Correlation between serum CEA level and metabolic volume as determined by FDG PET in postoperative patients with recurrent colorectal cancer. *Annals of Nuclear Medicine, 19(2)*, 123–129.

12. Clish, C.B. (2015). Metabolomics: an emerging but powerful tool for precision medicine. *Cold Spring Harbor molecular case studies, 1(1)*, a000588.

13. Cooper G.M. (2000). The Development and Causes of Cancer. *The Cell: A Molecular Approach* (2nd ed.). Sunderland, MA: Sinauer Associates.

14. Cui, X. & Churchill, G.A. (2003). Statistical tests for differential expression in cDNA microarray experiments. *Genome Biology, 4(4)*, 210.

15. Di Gangi, I.M., Mazza, T., Fontana, A., Copetti, M., Fusilli, C., Ippolito, A., . . . Pazienza, V. (2016). Metabolomic profile in pancreatic cancer patients: a consensus-based approach to identify highly discriminating metabolites. *Oncotarget, 7(5)*, 5815–5829.

16. Fabbrocini, G., Triassi, M., Mauriello, M.C., Torre, G., Annunziata, M.C., De Vita, V., . . . Monfrecola, G. (2010). Epidemiology of Skin Cancer: Role of Some Environmental Factors. *Cancers, 2*, 1980-1989.

17. Gorenberg, M., Hallett, W.A, & O'Doherty M.J. (2002). Does diabetes affect ^{18}F-FDG standardised uptake values in lung cancer?. *European Journal of Nuclear Medicine and Molecular Imaging, 29(10)*, 1324-1327.

18. Han, E.J., Yang, Y.J., Park, J.C., Park, S.Y., Choi, W.H., & Kim, S.H. (2015). Prognostic value of early response assessment using ^{18}F-FDG PET/CT in chemotherapy-treated patients with non-small-cell lung cancer. *Nuclear Medicine Communications, 36(12)*, 1187–1194.

19. Hong, J.H., Kim, H.H., Han, E.J., Byun, J.H., Jang, H.S., Choi, E.K., . . . Yoo, IeR. (2016). Total Lesion Glycolysis Using ^{18}F-FDG PET/CT as a Prognostic Factor for Locally Advanced Esophageal Cancer. *Journal of Korean Medical Science, 31(1)*, 39–46.

20. Horgan, R.P. & Kenny, L.C. (2011). ‘Omic’ technologies: genomics, transcriptomics, proteomics and metabolomics. *The Obstetrician & Gynaecologist, 13*, 189–195.

21. Hori, S., Nishiumi, S., Kobayashi, K., Shinohara, M., Hatakeyama, Y., Kotani, Y., . . . Yoshida, M. (2011). A metabolomic approach to lung cancer. *Lung cancer, 74(2)*, 284-292.

22. Huang, W., Fan, M., Liu, B., Fu, Z., Zhou, T., Zhang, Z., . . . Li, B. (2014). Value of metabolic tumor volume on repeated ^{18}F-FDG PET/CT for early prediction of survival in locally advanced non-small cell lung cancer treated with concurrent chemoradiotherapy. *Journal of Nuclear Medicine, 55(10)*, 1584–1590.

23. International Human Genome Sequencing Consortium. (2004). Finishing the euchromatic sequence of the human genome. *Nature, 431*, 931-945.

24. Kasim, A., Shkedy, Z., Kaiser, S., Hochreiter, S., & Talloen, W. (Eds.). (2016). *Applied Biclustering Methods for Big and High-Dimensional Data Using R.* Boca Raton, FL: Chapman & Hall/CRC.

25. Kiechle, F.L., Zhang, X., & Holland-Staley, C.A. (2004). The -omics Era and Its Impact. *Archives of Pathology & Laboratory Medicine, 128*, 1337-1345.

26. Krstajic, D., Buturovic, L.J., Leahy, D.E., & Thomas, S. (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of cheminformatics, 6(1)*, 10.

27. Larson, S.M., Erdi, Y., Akhurst, T., Mazumdar, M., Macapinlac, H.A., Finn, R.D., . . . Ginsberg, R. (1999). Tumor Treatment Response Based on Visual and Quantitative Changes in Global Tumor Glycolysis Using PET-FDG Imaging: The Visual Response Score and the Change in Total Lesion Glycolysis. *Clinical Positron Imaging, 2(3)*, 159–171.

28. Lee, K.H., Ko, B.H., Paik, J.Y., Jung, K.H., Choe, Y.S., Choi, Y., & Kim, B.T. (2005). Effects of anesthetic agents and fasting duration on ^{18}F-FDG biodistribution and insulin levels in tumor-bearing mice. *Journal of Nuclear Medicine, 46(9)*, 1531-1536.

29. Legrain, P., Aebersold, R., Archakov, A., Bairoch, A., Bala, K., Beretta, L., . . . Omenn, G.S. (2011). The Human Proteome Project: Current State and Future Direction. *Molecular & Cellular Proteomics, 10(7)*, M111.009993.

30. Lei, Z., Huhman, D.V., & Sumner, L.W. (2011). Mass Spectrometry Strategies in Metabolomics. *The Journal of Biological Chemistry, 286*, 25435-25442.

31. Lokhov, P.G., Trifonova, O.P., Maslov, D.L., & Archakov, A.I. (2013). Blood plasma metabolites and the risk of developing lung cancer in Russia. *European Journal of Cancer Prevention, 22(4)*, 335-341.

32. Louis, E., Bervoets, L., Reekmans, G., De Jonge, E., Mesotten, L., Thomeer, M., & Adriaensens, P. (2015). Phenotyping human blood plasma by ^{1}H-NMR: a robust protocol based on metabolite spiking and its evaluation in breast cancer. *Metabolomics, 11(1)*, 225–236.

33. Louis, E., Adriaensens, P., Guedens, W., Bigirumurame, T., Baeten, K., Vanhove, K., . . . Thomeer, M. (2016a). Detection of Lung Cancer through Metabolic Changes Measured in Blood Plasma. *Journal of Thoracic Oncology, 11(4)*, 516-523.

34. Louis, E., Adriaensens, P., Guedens, W., Vanhove, K., Vandeurzen, K., Darquennes, K., . . . Mesotten, L. (2016b). Metabolic phenotyping of human blood plasma: a powerful tool to discriminate between cancer types?. *Annals of Oncology, 27(1)*, 178–184.

35. Manzoni, C., Kia, D.A., Vandrovcova, J., Hardy, J., Wood, N.W., Lewis, P.A., & Ferrari, R. (2018). Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Briefings in Bioinformatics, 19(2)*, 286–302.

36. Moon, S.H., Hyun, S.H., & Choi, J.Y. (2013). Prognostic significance of volume-based PET parameters in cancer patients. *Korean journal of radiology, 14(1)*, 1–12.

37. Moreno, P., Jiménez-Jiménez, C., Garrido-Rodríguez, M., Calderón-Santiago, M., Molina, S., Lara-Chica, M., . . . Calzado, M.A. (2018). Metabolomic profiling of human lung tumor tissues - nucleotide metabolism as a candidate for therapeutic interventions and biomarkers. *Molecular oncology, 12(10)*, 1778–1796.

38. Oh, D.Y., Kim, J.W., Koh, S.J., Kim, M., Park, J.H., Cho, S.Y., . . . Im, J.P. (2014). Does diabetes mellitus influence standardized uptake values of fluorodeoxyglucose positron emission tomography in colorectal cancer?. *Intestinal research, 12(2)*, 146–152.

39. Ono, A., Naito, T., Ito, I., Watanabe, R., Shukuya, T., Kenmotsu, H., . . . Yamamoto, N. (2012). Correlations between serial pro-gastrin-releasing peptide and neuron-specific enolase levels and the radiological response to treatment and survival of patients with small-cell lung cancer. *Lung Cancer, 76(3)*, 439–444.

40. Plaza, N.C., García-Galbis, M.R., & Martínez-Espinosa, R.M. (2017). Impact of the "Omics Sciences" in Medicine: New Era for Integrative Medicine. *Journal of Clinical Microbiology and Biochemical Technology, 3(1)*, 9-13.

41. Raschka, S. (2018). Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. arXiv:1811.12808v2 [cs.LG].

42. Romano, J.P., Shaikh, A.M., & Wolf, M. (2010). Multiple Testing. *The New Palgrave Dictionary of Economics.* London: Palgrave Macmillan.

43. Sainani, K.L. (2009). The Problem of Multiple Testing. *PM & R: The Journal of Injury, Function, and Rehabilitation, 1(12)*, 1098-1103.

44. Shi, P., Meng, X., Ni, M., Sun, X., Xing, L., & Yu, J. (2015). Association between serum tumor markers and metabolic tumor volume or total lesion glycolysis in patients with recurrent small cell lung cancer. *Oncology Letters, 10(5)*, 3123-3128.

45. Shum, W.Y., Ding, H.J., Liang, J.A., Yen, K.Y., Chen, S.W., & Kao, C.H. (2012). Use of Pretreatment Metabolic Tumor Volumes on PET-CT to Predict the Survival of Patients with Squamous Cell Carcinoma of Esophagus Treated by Curative Surgery. *Anticancer Research: International Journal of Cancer Research and Treatment, 32(9)*, 4163-4168.

46. Soret, M., Bacharach, S.L., & Buvat, I. (2007). Partial-Volume Effect in PET Tumor Imaging. *The Journal of Nuclear Medicine, 48(6)*, 932-945.

47. Tenley, N., Corn, D.J., Yuan, L., & Lee, Z. (2013). The effect of fasting on PET Imaging of Hepatocellular Carcinoma. *Journal of Cancer Therapy, 4(2)*, 561–567.

48. Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological), 58(1)*, 267-288.

49. Trivedi, D.K., Hollywood, K.A., & Goodacre, R. (2017). Metabolomics for the masses: The future of metabolomics in a personalized world. *New horizons in translational medicine, 3(6)*, 294–305.

50. Venter, J.C., Smith, H.O., & Adams, M.D. (2015). The Sequence of the Human Genome. *Clinical Chemistry, 61(9)*, 1207-1208.

51. Zasadny, K.R., Kison, P.V., Francis, I.R., & Wahl, R.L. (1998). FDG-PET Determination of Metabolically Active Tumor Volume and Comparison with CT. *Clinical Positron Imaging, 1(2)*, 123-129.

52. Zou, H. & Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology), 67(2)*, 301-320.

7 Appendix

7.1 Tables

Table 7: *Eigenvalues and percentage of variance of the first 5 components*

	eigenvalues	% of variance	cumulative % of variance
comp 1	1.2026	24.09375	24.09375
comp 2	0.8552	17.13426	41.22801
comp 3	0.5506	11.03133	52.25934
comp 4	0.3572	7.15547	59.41481
comp 5	0.2785	5.579983	64.99479

Table 8: *Factor loadings of all variables for the first and second factor*

Variable	Factor 1	Factor 2
SUV$_{max}$	0.4690	-0.3314
MTV	0.7704	-0.5173
TLG	0.7857	-0.5714
VAR1	0.1571	0.2318
VAR2	0.1055	0.2351
VAR3	0.1075	0.2160
VAR4	0.1572	0.2215
VAR5	0.1354	0.1049
VAR6	-0.0318	0.1162
VAR7	-0.0185	0.0856
VAR8	0.3395	0.1572
VAR9	0.4625	0.2233
VAR10	0.3986	0.0304
VAR11	0.4002	0.0498
VAR12	0.3103	0.1745
VAR13	0.3351	0.1802
VAR14	0.0609	0.1994
VAR15	-0.0617	0.2361
VAR16	-0.1094	0.1647
VAR17	0.1583	0.0913
VAR18	-0.5865	-0.6206
VAR19	0.3630	0.5581
VAR20	-0.2649	-0.1087
VAR21	0.4485	0.5472
VAR22	-0.0764	0.0086
VAR23	0.3490	-0.0367
VAR24	0.0918	-0.0264
VAR25	-0.0744	-0.1676
VAR26	0.2049	0.2502
VAR27	0.2521	0.2843
VAR28	0.1289	0.2539
VAR29	-0.2692	-0.0690
VAR30	-0.4800	-0.3038

(*table continues*)

Variable	Factor 1	Factor 2
VAR31	-0.1772	-0.0747
VAR32	0.1707	-0.0019
VAR33	0.4181	0.1716
VAR34	0.4350	0.2573
VAR35	0.4730	0.2964
VAR36	0.5369	0.3412
VAR37	0.5723	0.3362
VAR38	0.6750	0.5316
VAR39	0.5220	0.4814
VAR40	0.4613	0.5591
VAR41	0.4953	0.6103
VAR42	0.5854	0.6071
VAR43	0.5859	0.5662
VAR44	0.6144	0.5393
VAR45	0.6193	0.3846
VAR46	0.5572	0.3153
VAR47	0.5439	0.3106
VAR48	0.6776	0.3202
VAR49	0.6869	0.3240
VAR50	0.7002	0.3789
VAR51	0.6134	0.6424
VAR52	0.5245	0.5575
VAR53	0.4515	0.5439
VAR54	0.4404	0.5564
VAR55	0.1076	0.2157
VAR56	-0.0077	0.0988
VAR57	0.0436	0.1876
VAR58	0.1543	0.5928
VAR59	-0.0206	0.2185
VAR60	-0.0307	0.2411
VAR61	-0.0177	0.2173
VAR62	0.0008	0.3719
VAR63	0.1371	0.5587
VAR64	0.2397	0.6162
VAR65	0.1202	0.5198
VAR66	0.3169	0.6136
VAR67	0.1003	0.4636
VAR68	-0.4201	-0.4070
VAR69	-0.4234	-0.4776
VAR70	0.3163	0.4642
VAR71	0.5424	0.4188
VAR72	0.2024	0.0943
VAR73	0.1895	0.3892
VAR74	0.1280	0.3174
VAR75	0.1800	0.5215
VAR76	0.2736	0.6063
VAR77	0.4092	0.4620
VAR78	0.4134	0.4233
VAR79	0.3818	0.2756

(*table continues*)

Variable	Factor 1	Factor 2
VAR80	0.3723	0.2053
VAR81	0.3558	0.3912
VAR82	0.3898	0.3895
VAR83	0.4033	0.3880
VAR84	0.3996	0.3100
VAR85	0.2077	0.1178
VAR86	-0.5182	-0.6457
VAR87	-0.5671	-0.6969
VAR88	-0.5037	-0.6717
VAR89	0.1485	0.0907
VAR90	0.4936	0.4811
VAR91	0.1299	-0.4679
VAR92	-0.0402	0.2816
VAR93	-0.0671	0.0906
VAR94	0.6121	0.5249
VAR95	-0.4862	-0.5984
VAR96	0.2173	0.3973
VAR97	0.1842	0.0884
VAR98	-0.3691	-0.4978
VAR99	-0.4422	-0.4797
VAR100	-0.6532	-0.7244
VAR101	0.1526	0.2525
VAR102	-0.1336	-0.0429
VAR103	0.3894	0.3637
VAR104	0.3472	0.2807
VAR105	0.5578	0.3838
VAR106	0.5682	0.3105
VAR107	0.5564	0.3107
VAR108	0.6334	0.3746
VAR109	0.6086	0.3458
VAR110	-0.4416	-0.2758

Table 9: *Factor scores of all patients for the first and second factor*

Patient	Factor 1	Factor 2
1	-0.0284	0.6931
2	0.3176	1.393
3	-0.6216	0.6484
4	-1.4759	-0.6606
5	-0.1533	0.5784
6	-0.7109	0.6457
7	0.2051	0.9239
8	-0.6345	0.5117
9	0.0723	0.8945
10	-1.0109	-0.0457
11	-2.5291	-1.8360
12	-0.7277	0.3323
13	-1.4321	-0.5690
14	-1.3268	-0.5509

(*table continues*)

Patient	Factor 1	Factor 2
15	-2.3793	-1.5867
16	0.1684	1.1763
17	0.4448	1.3536
18	-0.7255	0.6515
19	-0.0274	1.1587
20	-1.0008	-0.2808
21	-0.7797	0.0049
22	-1.3152	-0.5592
23	-0.7435	0.2256
24	-0.7714	0.5584
25	-1.0217	-0.0847
26	-0.8166	0.5294
27	-0.6146	0.3405
28	-0.9856	-0.1339
29	-1.0440	0.0483
30	-0.9930	0.4060
31	-0.7478	0.2581
32	-3.0993	-2.3298
33	-0.4944	0.1091
34	-0.8193	0.1072
35	-0.3769	0.2842
36	0.4295	1.8078
37	0.2135	1.1103
38	-0.7228	0.2217
39	-0.1225	1.2229
40	-0.0064	1.0991
41	-0.9175	0.0459
42	-0.7751	0.2818
43	0.1770	1.4291
44	-0.9097	0.3929
45	-0.6719	0.6524
46	-1.4823	-0.4668
47	0.0729	1.5463
48	-0.0662	0.8551
49	0.5811	1.3232
50	-0.9117	-0.1593
51	-0.2866	0.4523
52	-0.1800	1.2865
53	-0.4082	0.7752
54	-1.5900	-0.5137
55	-1.7980	-1.0241
56	-0.8833	-0.0195
57	-1.4484	-0.5912
58	-0.3197	1.1329
59	-0.9819	0.4478
60	-0.2711	0.8724
61	-0.8605	0.3173
62	-1.8098	-1.2960
63	-0.3718	0.5839

(*table continues*)

Patient	Factor 1	Factor 2
64	-0.2985	1.0361
65	-0.9932	0.2346
66	0.0493	1.0572
67	0.0536	1.5557
68	-1.2676	0.0513
69	-0.6545	0.1595
70	-0.9749	-0.5196
71	-0.6440	-0.0631
72	-0.0286	0.3783
73	1.0159	1.4478
74	0.1283	0.6379
75	2.4481	-0.7548
76	0.8626	0.9039
77	-0.4498	0.1406
78	1.3568	1.5183
79	-0.2634	0.1034
80	-0.0545	0.6379
81	1.0659	-0.4049
82	1.8205	0.0719
83	0.6783	0.1171
84	0.6125	0.5555
85	-0.7714	-0.7229
86	0.5396	0.3350
87	0.3132	-0.0563
88	0.3467	-1.1366
89	0.5517	1.0072
90	-0.2745	-1.0878
91	0.5522	0.2894
92	-0.7104	-0.8586
93	-0.2839	0.2015
94	-0.6638	0.6119
95	0.1816	1.5463
96	-1.1966	-0.7457
97	-0.4075	0.2381
98	-0.1719	0.5131
99	0.1937	0.5917
100	-0.6508	-0.1591
101	5.0961	-2.5398
102	0.1379	-0.0465
103	0.6287	1.1303
104	-0.7248	-0.3996
105	1.6719	0.6756
106	1.1873	1.2923
107	1.4247	0.6760
108	1.7746	-0.5561
109	0.2612	-0.8877
110	1.1044	0.4065
111	-0.1863	-0.3042
112	-1.5387	-1.2221

(*table continues*)

Patient	Factor 1	Factor 2
113	-0.6891	-0.0564
114	-0.0996	0.5694
115	0.2148	1.3363
116	-0.8014	0.1718
117	-1.5676	-0.8095
118	-0.2317	0.8556
119	-0.7520	-0.1722
120	-0.4966	0.1254
121	-1.0109	-0.6524
122	-0.1178	0.8148
123	-1.0592	-0.5599
124	-1.0131	-0.3783
125	0.1317	0.6681
126	0.3274	0.2379
127	0.4428	-0.3885
128	-0.7907	-0.7236
129	1.2643	-0.8343
130	-0.0932	0.1842
131	-0.0493	0.4466
132	-0.0084	0.8822
133	-0.8778	-0.4128
134	-0.4034	0.4407
135	-1.4399	-1.4683
136	-0.3697	0.0038
137	-0.2790	0.0164
138	0.2411	1.0913
139	-0.2002	-0.5547
140	0.6963	1.1942
141	0.3959	0.8178
142	-1.4227	-0.9319
143	0.0032	0.6649
144	-0.4142	0.3751
145	-0.6257	0.0627
146	0.5564	1.6033
147	0.0321	0.2606
148	-1.0482	-0.9211
149	-0.0231	0.3769
150	0.3241	0.0195
151	1.3605	1.5077
152	-0.1060	0.0046
153	0.1473	0.7057
154	1.1431	1.5021
155	1.5062	-1.7153
156	0.5754	0.5729
157	1.4289	1.1039
158	1.2832	0.7166
159	0.1483	0.6357
160	1.9125	-0.6981
161	-0.5754	-0.2704

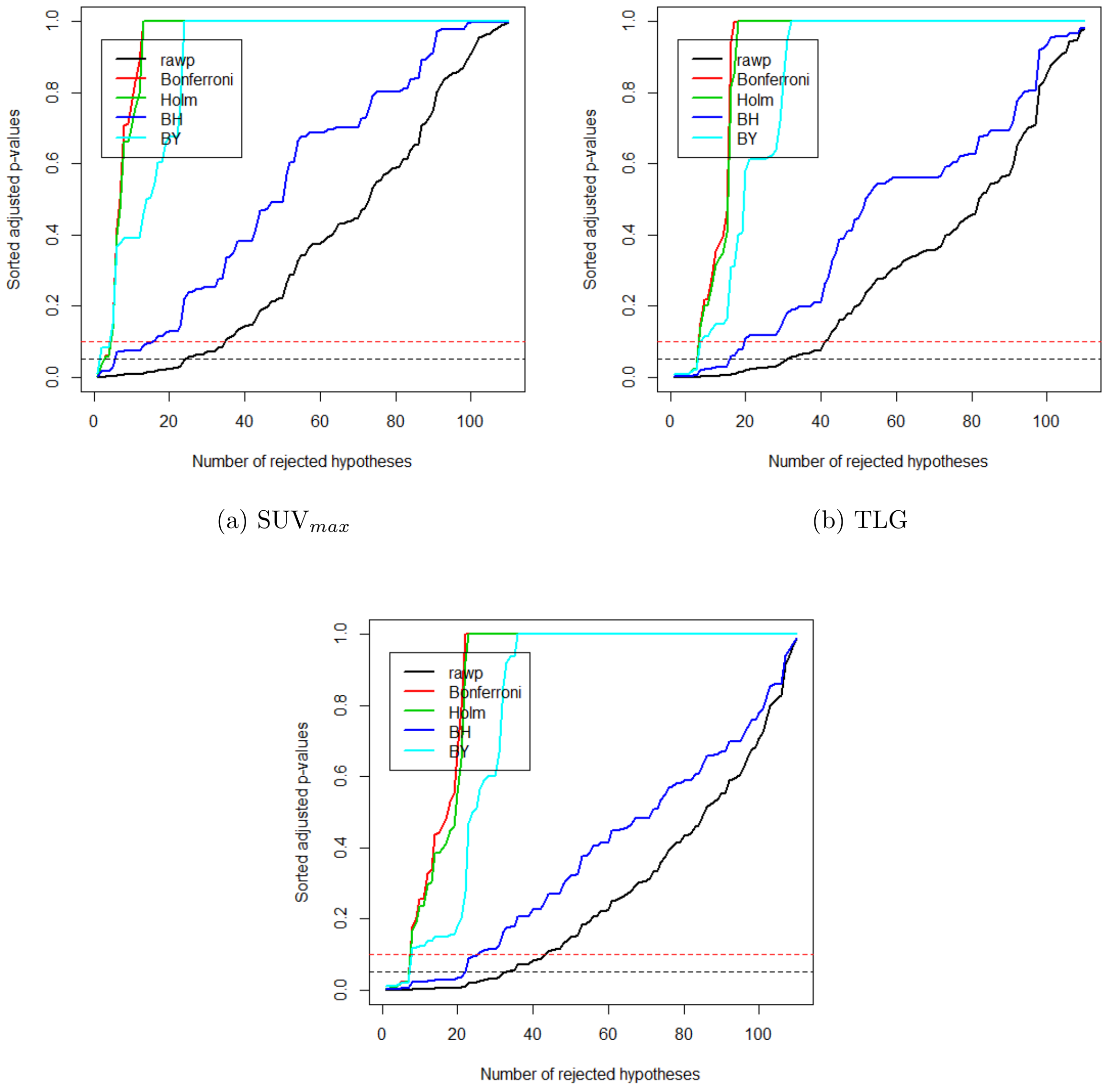
(*table continues*)

Patient	Factor 1	Factor 2
162	1.4072	-3.1843
163	-0.5207	-1.6144
164	1.7352	0.0729
165	0.2505	-0.0089
166	0.6104	-1.4572
167	0.6436	-0.4956
168	-0.0829	0.3261
169	-0.4143	-1.0125
170	0.4411	-0.2864
171	0.2579	0.3550
172	0.8021	-0.8829
173	0.5700	0.4913
174	0.7194	0.2963
175	2.2575	-0.5547
176	-0.3054	0.0373
177	0.2693	-0.6284
178	-0.3857	-0.1740
179	1.6032	-0.0374
180	1.7090	-1.0280
181	-0.4544	-0.6446
182	0.4439	1.0717
183	0.1642	0.1871
184	-0.0919	-0.6796
185	-1.1257	-2.4131
186	-0.0590	-0.4355
187	1.5064	-0.2295
188	1.3327	-0.4838
189	-1.0812	-0.8082
190	0.7594	0.6856
191	2.4415	-1.2337
192	1.6669	-0.0017
193	0.5755	-1.4170
194	1.1039	-0.9142
195	-0.9513	-1.7802
196	0.9516	0.2555
197	0.1368	-0.5139
198	-0.2400	0.3045
199	1.4475	-0.8607
200	2.3091	-0.0762
201	0.0072	-0.0883
202	-0.3066	-0.8309
203	0.2279	0.3415
204	0.0967	-0.5612
205	-0.4924	-1.0032
206	-0.8921	-0.5396
207	-1.1189	-2.0049
208	1.7008	-0.7214
209	2.0887	-1.2472
210	-0.8785	-1.1270

(*table continues*)

Patient	Factor 1	Factor 2
211	0.2277	-0.2849
212	6.0959	-4.7434
213	0.8628	1.0546
214	-0.1427	-0.0720
215	0.5697	0.1848
216	-0.0010	-0.3431
217	1.6102	0.5619
218	2.6108	0.1318
219	-1.2430	-1.6204
220	0.5500	-0.0126
221	-0.5467	-1.4045
222	0.6728	-0.9563

7.2 Figures

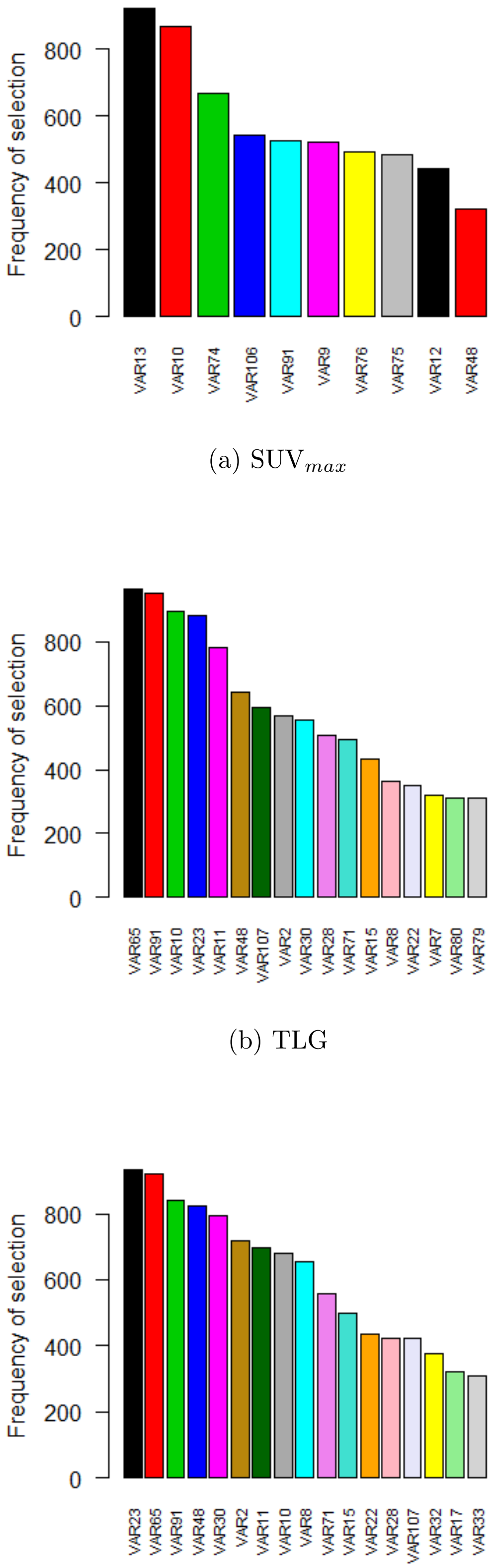


(a) SUV$_{max}$

(b) TLG

(c) MTV

Figure 14: *Number of rejected hypotheses after multiplicity correction*

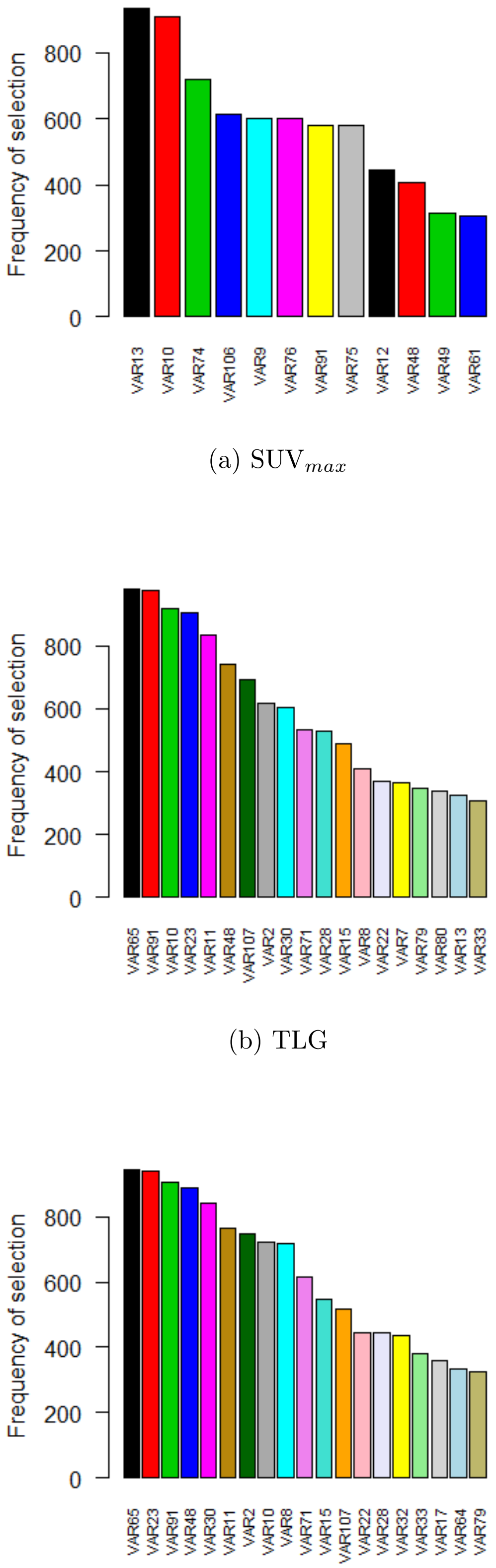


(a) SUV$_{max}$

(b) TLG

(c) MTV

Figure 15: *Frequency of selected metabolites in LASSO*



(a) SUV$_{max}$

(b) TLG

(c) MTV

Figure 16: *Frequency of selected metabolites in Elastic Net*