Improvement of risk models to select individuals eligible for lung cancer screening with low-dose computed tomography: adding the metabolic phenotype

Theophile Bigirumurame[1*], Louis Evelyne[2*],Vanhove Karolien[2,3], Mesotten Liesbet[2,4], Vandeurzen Kurt[5], Darquennes Karen[6], Vansteenkiste Johan[7], Dooms Christophe[7], Thomeer Michiel[2,8], Ziv Shkedy[1], Adriaensens Peter[9]

Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-BioStat),
Center for Statistics, Hasselt University, Belgium

Abstract

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 current study examined whether this selection can be improved by adding parameters that reflect the plasma metabolic phenotype and which are obtained by proton nuclear magnetic resonance (NMR) spectroscopy. Predictive risk models were developed in a training cohort and evaluated in two independent validation cohorts using random forests (RF) analysis and penalized logistic regression models.
All methods indicate that, and this for both validation cohorts, the addition of NMR metabolic phenotype data to a classical risk model significantly improves its discriminative power. RF analysis demonstrates an improvement in misclassification error (MCE) and positive predictive value (PPV) up to 11% and 14%, respectively. Complementing classical risk models which only take clinical risk factors into account with NMR metabolic phenotype data shows interesting potential to improve the selection of high-risk individuals eligible for lung cancer screening with LDCT, thereby reducing the false positive rate and corresponding financial burden.

Keywords: metabolic phenotype; ^{1}H-NMR spectroscopy; lung cancer; selection for screening; risk models

1 Introduction

Lung cancer represents the leading cause of cancer death worldwide, accounting for about 20% of all cancer deaths (Ferlay et al., 2015). This underscores the imperative need for screening programs which enable to reduce lung cancer mortality (Manser et al., 2004, Woolf et al., 2014). One of the main criteria for a screening test is the cost-effectiveness, meaning that the number of false positive results should be low to prevent unnecessary surgical interventions (Tammemagi and Lam, 2014, Wood et al., 2012). To maximize the benefit-risk balance, a high-risk target population needs to be selected (Field and Duffy, 2008, Field et al., 2013). Accurate selection of high-risk individuals in lung cancer screening programs necessitates robust methods for risk prediction (Field et al., 2013).
Current models to estimate lung cancer risk have tended to concentrate on clinical risk factors, including age, smoking behavior, previous history of cancer and family history of lung cancer (Cassidy et al., 2008, Spitz et al., 2007, Hoggart et al., 2012, Tammemagi et al., 2013, Bach et al., 2003). Since lung cancer predominantly occurs in elderly people and smoking is an important risk factor, the high-risk individuals in the two largest randomized controlled trials designed to evaluate the impact of low-dose computed tomography (LDCT) screening on lung cancer mortality were selected on the basis of age and smoking behavior (National Lung Screening Trial Research Team, 2011a, Zhaoa et al., 2011). More specifically, eligible participants for the North-American National Lung Screening Trial (NLST) were aged between 55 and 74 years and had a smoking history of at least 30 pack years. Former smokers were only included in the study if they had quit smoking within the past 15 years (National Lung Screening Trial Research Team, 2011a,b). Furthermore, the Dutch-Belgian lung cancer screening trial (Dutch acronym - NELSON) recruited subjects aged between 50 and 75 years, who smoked 15 or more cigarettes a day for more than 25 years ($\geq$18.75 pack years) or 10 or more cigarettes a day for more than 30 years ($\geq$15 pack years). Former smokers were only included in the study if they had quit smoking for less than 10 years (Zhaoa et al., 2011, van den Bergh et al., 2008). The major drawback of both LDCT screening studies is the low positive predictive value (PPV), ranging from 3.8% in the NLST study to 40.4% in the NELSON study. This indicates that more than half of the study participants were referred for further investigations, being not without cost and risk, on the basis of false positive results (National Lung Screening Trial Research Team, 2011b, 2013, Horeweg et al., 2014, Bach et al., 2012). Consequently, there is an increasing interest to improve the accuracy of risk models by adding lung cancer risk-related biomarkers. This in order to better select high-risk individuals eligible for lung cancer screening with LDCT and so to lower the false positive rate and corresponding financial burden.

A blood-based diagnostic biomarker signifies an attractive option to complement risk models used to select high-risk individuals eligible for lung cancer screening with LDCT since blood samples can be obtained in a non-invasive way and with minimal risk for the patient (Mamas et al., 2011, Tsay et al., 2014). Recently, our research group has demonstrated that the metabolic phenotype of blood plasma, determined by proton nuclear magnetic resonance (^{1}H-NMR) spectroscopy, not only enables to discriminate between cancer patients and controls (Louis et al., 2015b, 2016) but also between different cancer types like lung and breast cancer (Louis et al., 2015a). ^{1}H-NMR spectroscopy, one of the main analytical platforms used in metabolomics studies, is a very reproducible technique which permits a fast and non-invasive identification and specially quantification of complex mixtures of metabolites, as in plasma, with minimal sample preparation and relatively low cost on a per sample basis (Emwas et al., 2013, Lindon and Nicholson, 2008). Hence, ^{1}H-NMR-based metabolomics of blood plasma looks to provide attractive blood-based diagnostic biomarkers to add to risk models used for the selection of high-risk individuals eligible for lung cancer screening with LDCT.

The Liverpool Lung Project (LLP) risk model proposed by Cassidy et al. (2008) predicts individual risk to develop lung cancer based on clinical risk factors such as age, gender, smoking status, prior diagnosis of malignant tumor. Based on the individual risk produce by the LLP model patient can be classified as having lung cancer (LC) or healthy control (HC). In the present study, we investigate whether the addition of metabolic phenotype data to a risk model that only takes clinical risk factors into account has potential to improve the selection of high-risk individuals eligible for lung cancer screening with LDCT.

This paper is organized as follows: in Section 2, we introduce the study and the data used for the analysis presented in the paper. Section 3 presents the statistical methodology used for the development of the predictive model while Section 4 the proposed methods is applied to the data followed by a discussion in Section 5.

2 Data

Lung cancer patients (LC, n=357) were included in the Limburg Positron Emission Tomography Center (n=273) (Hasselt, Belgium) and at the Department of Respiratory Medicine of University Hospital Leuven (n=84) (Leuven, Belgium) from March 2011 to June 2014. The diagnosis of lung cancer was confirmed by a pathological biopsy or a clinician specialized in interpreting radiological and clinical lung cancer data. Clinical staging of the tumors was performed according to the $7th$ edition of the tumor, node, metastasis classification of malignant tumors (Goldstraw et al., 2007). Controls (HC, n=347) were patients with non-cancer diseases who were included at Ziekenhuis Oost-Limburg (Genk, Belgium) between March 2012 and June 2014. Exclusion criteria were: 1) not fasted for at least 6 h; 2) fasting blood glucose concentration $\geq$200 mg/dl; 3) medication intake on the morning of blood sampling and 4) treatment or history of cancer in the past 5 years. The study was conducted in accordance with the ethical rules of the Helsinki Declaration and Good Clinical Practice and was approved by the involved ethical committees. All study participants provided written informed consent. The current study is a secondary objective of the lung cancer-control study which is registered at clinical trials.gov (Trial registration ID: NCT02024113) and which is recently published in Journal of Thoracic Oncology (Louis et al., 2016). More details about the blood sampling, sample preparation and NMR analysis are found in Section 1.2 of the supplementary appendix for the paper.

273 of the patients treated in Limburg Positron Emission Tomography and 263 of the healthy controls recruited from Ziekenhuis Oost-Limburg (Genk, Belgium) formed the first study cohort while the 84 patients treated at the University Hospital Leuven together with a group of 84 healthy controls from Ziekenhuis Oost-Limburg (Genk, Belgium) form the second cohort study. An elaborate discussion of the usage of the two cohorts within a cross validation loop is given in Section 3.2.

The data described above can be presented in matrix notation as follows. Let $\mathbf{X}$ be a $n \times m$ data matrix, where n is the sample size and m the number of the molecular variables (metabolites in our case). The $\mathbf{X}$ matrix contains information about 102 integration regions on 536 samples. Let $\mathbf{Z}$ be a $n \times p$ clinical variables matrix. It is assumed that $p < n$, but no such restriction is put on m. Let $\mathbf{Y}$ $n \times 1$ be a vector in which the ith entry is an indicator variable which is equal to 1 if the patient has lung cancer (LC) and 0 otherwise. We further assume that:

$$Y_i = \begin{cases} 1 & : \text{Lung cancer}, \\ 0 & : \text{Otherwise.} \end{cases} \tag{1}$$

The three data sources $\mathbf{Y}$, $\mathbf{Z}$ and $\mathbf{X}$ can be represented as matrices given by:

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix}, \quad \mathbf{Z} = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & z_{22} & \cdots & z_{2p} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ z_{n1} & z_{n2} & \cdots & z_{np} \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}. \tag{2}$$

Note that only 102 out of the 110 NMR integration values were evaluated since 8 integration values had more than 10% of missing observations (i.e., $\mathbf{X}$ is a 536×102). The matrix $\mathbf{Z}$ is a 536×9 matrix consisting of nine clinical variables that were identified as potential risk factors associated with lung cancer: (1) gender, (2) body mass index (BMI), (3) smoking habits, (4) number of smoked packs per years, (5) previous mine-worker (PMW), (6) prior diagnosis of a malignant tumor (PDMT), (7) presence of chronic obstructive pulmonary disease (COPD), (8) intake of anti-arrhythmic medication and (9) intake of medication against high blood pressure.

3 Statistical methods

3.1 Classification based on the clinical risk factors

First, a stepwise logistic regression, in which disease status (lung cancer or control) was the response variable and the clinical risk factors were the predictors, was performed to identify clinical risk factors which had significant effect on disease status prediction (p-value ≤ 0.05). The following linear predictor based on the clinical risk factors was defined:

$$logit(\pi_i) = \alpha_0 + \sum_{l=1}^{p} \alpha_l Z_{il}. \tag{3}$$

Here $\pi_i = P(Y_i = 1)$. The logistic regression model comprising significant clinical risk factors was then considered as the baseline predictive model and was used for classification of the patients.

3.2 Cross validation

Classification of patients as LC or HC was done within a loop of 1000 iterations of 3-fold cross validation. In each iteration in the loop, the first cohort of 536 subjects (273 lung cancer patients treated in Limburg Positron Emission Tomography and 263 controls) was further split randomly into a training cohort (2/3 of the subjects) and a validation cohort (1/3 of the subjects, test cohort 1). The classifier was trained on the training cohort and was validated on the validation cohort. In addition, a fixed validation cohort (test cohort 2) of 168 subjects, 84 lung cancer patients treated in the University Hospital Leuven and 84 controls from Ziekenhuis Oost-Limburg (Genk, Belgium), was used for the validation.

Note that this procedure ensure that two independent validation sets are used: the first is a subset of 1/3 of the patients that selected within the cross validation loop (which are treated in the same hospital as the patients in the training set) while the second validation set consists of cancer patients that are not treated in the same hospital as the patients in the training set. Figure 1 depicts the cross validation procedure used for the development of the predictive models base on both penalized regression and random forest which will be discussed in the next Section.

3.3 Classification using clinical risk factors and metabolic data

First, the likelihood ratio test (Casella and Berger, 2001), the global test proposed by Goeman et al. (2005), and the global test proposed by Boulesteix and Hothorn (2010) were used to check whether the addition of these NMR parameters has potential to improve disease status prediction. Note that although these tests do not result in a predictive model, they do provide an initial indication about the association of the metabolic data with the diseases status given the clinical risk factors.

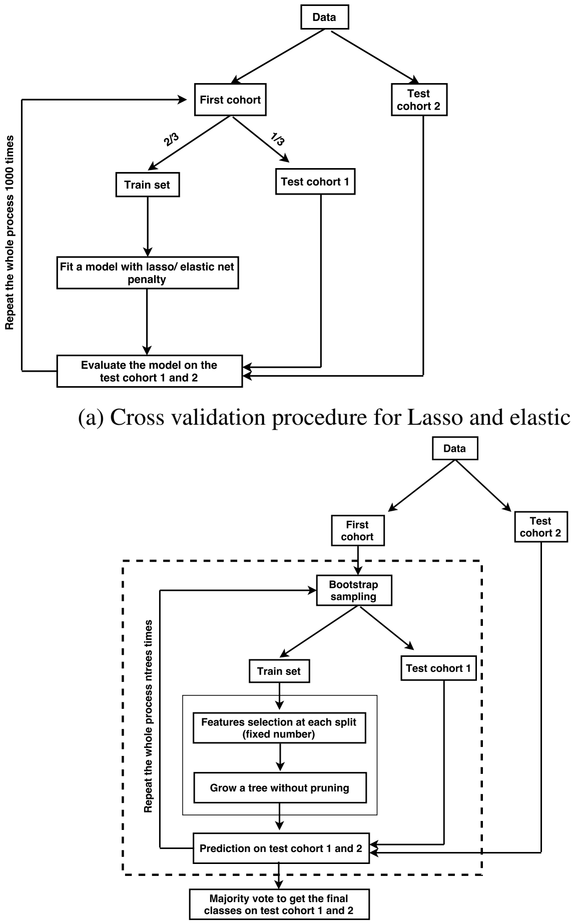
To study the added predictive value of the metabolic data given clinical risk factors, the following predictive models were developed: penalized logistic regression models using Lasso (Tibshirani, 1996, Friedman et al., 2001) and elastic net penalties (Friedman et al., 2001, Zou and Hastie, 2005), and random forest (RF, Breiman, 2001).

3.3.1 Predictive models based on Lasso and elastic net

Given significant clinical variables in equation (3) a model in which both data types are included is given by:

$$logit(\pi_i) = \alpha_0 + \sum_{l=1}^{p} \alpha_l Z_{il} + \sum_{j=1}^{m} \beta_j X_{ij}, \tag{4}$$

where, the matrices $\mathbf{Z}$ and $\mathbf{X}$ contain the clinical and metabolic data, respectively. The negative log-likelihood with penalty takes the following form:


Data
First cohort
Test cohort 2
2/3
1/3
Train set
Test cohort 1
Fit a model with lasso/ elastic net penalty
Evaluate the model on the test cohort 1 and 2
Repeat the whole process 1000 times
Data
First cohort
Test cohort 2
Bootstrap sampling
Train set
Test cohort 1
Features selection at each split (fixed number)
Grow a tree without pruning
Prediction on test cohort 1 and 2
Repeat the whole process ntrees times
Majority vote to get the final classes on test cohort 1 and 2

(a) Cross validation procedure for Lasso and elastic nett

(b) Sampling procedure for random forest

Figure 1: Cross validation schemes for Lasso and elastic net (panel a) and Random forest (panel b). Each cross validation loop consists of 1000 iterations. Test cohort 1: the test set which is randomly selected from the first cohort. Cancer patients in this cohort were treated in Limburg Positron Emission Tomograph center. Test cohort 2: a fixed test cohort in which cancer patients were treated in Leuven's University Hospital.

$$-\frac{1}{N}\sum_{i=1}^{N}\{y_i \log \Pr(Y=1|x_i,z_i)+(1-y_i)\log \Pr(\mathrm{Y}=0|\mathrm{x_i},\mathrm{z_i})\}+\lambda P_\alpha(\beta_j),$$

$$=-\frac{1}{N}\sum_{i=1}^{N}\{y_i(\beta_0+x_i'\beta+z_i'\gamma)-\log(1+e^{\beta_0+x_i'\beta+z_i'\gamma})\}+\lambda P_\alpha(\beta_j),$$

where, $P_\alpha(\beta_j)=\sum_{j=1}^{m}\left[\frac{1}{2}(1-\alpha)\beta_j^2+\alpha|\beta_j|\right]$ is the elastic net penalty (Zou and Hastie, 2005). The penalty term P_α is a compromise between the ridge-regression penalty (α = 0) and the Lasso penalty (α = 1). The penalty parameter λ introduces shrinkage in the coefficients β_j. The above model implies that only coefficients related to **X** matrix (metabolic data) are penalized (i.e., all the clinical risk factors identified in Section 3.1 are included in the predictive model). At each step of the cross validation loop, the predictive models were compared to the baseline model in equation (3) in terms of misclassification error (MCE), sensitivity, specificity, positive and negative predictive values in order to assess the added predictive value of the metabolic data in addition to the clinical risk factors.

3.3.2 Random forest

Random Forest (RF, Breiman, 2001) is an ensemble method for classification that builds many decision trees based on bootstrap samples from the original data and aggregates their predictions to improve the predictive accuracy and to control for over-fitting. The re-sampling scheme used in RF is different from the k-fold cross validation used for penalized regression models. Figure 1b depicts a schematic representation of RF re-sampling. Misclassification errors from random forest built using only the clinical risk data are compared to those from random forest built using both data types.

3.4 Permutation based inference

In order to assess the added value of the metabolic data to the classifier, a permutation procedure was used to evaluate the misclassification error. In total, 1000 permutations were performed. At each permutation, the columns of the metabolic matrix (**X**) were permuted while the columns of clinical risk factors (**Z**) and label variable (**Y**, the disease status) were kept fixed.

$$\mathbf{X}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} ---> \quad \mathbf{X}^*=\begin{bmatrix} x_{21}^* & x_{22}^* & \cdots & x_{2m}^* \\ x_{n1}^* & x_{n2}^* & \cdots & x_{nm}^* \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ x_{11}^* & x_{12}^* & \cdots & x_{1m}^* \end{bmatrix}. \tag{5}$$

Note that for each permuted data, a 3-fold cross validation procedure, consisting of a loop with 1000 iterations was preformed and the misclassification error (for both test cohorts) was calculated. The columns of $\mathbf{X}^*$ were used as predictors. This procedures implies that at the end of the permutations we are able to compare between the distribution of the misclassification error obtained for the permuted data and the non-permuted data. This allows us to test the null hypothesis that the metabolic data do not influence the predicative model in terms of reduction of the misclassification error. Note that under the alternative hypothesis, including the metabolic data in the predictive model is expected to reduced the misclassification error.

4 Application to the data

4.1 Identification of significant metabolic variables

All clinical risk factors were included in a multiple logistic regression model using a stepwise selection procedure. Table 1 presents the final model which includes age, smoking habits, number of smoking pack years, BMI, presence of COPD, intake of anti-arrhythmic medication and intake of anti-coagulants medication are significantly associated to lung cancer. This covariates set was used in each iteration in the cross validation loop as the baseline predictive model. In each iteration, the parameters were estimated using the train data set and classification was done on the two test cohorts describe in Section 3.2.

Table 1: Odds ratio estimates from a multiple logistic regression model comprising statistically significant clinical risk factors. Abbreviations: CI: confidence interval, OR: odds ratio.

	OR (95% CI)	**p-value**
(Intercept)	0.103 (0.012-0.835)	0.035
Age	1.035 (1.013-1.059)	0.002
Smoking pack years	1.016 (1.005-1.028)	0.0057
BMI	0.922 (0.879-0.965)	0.0006
COPD	5.466 (3.316-9.248)	<0.0001
Taking anti-arrhythmic medication	4.293 (1.932-10.103)	0.0005
Smoker	9.354 (4.031-23.156)	<0.0001
Ex-smoker	6.525 (3.052-14.996)	<0.0001
Taking anti-coagulants medication	0.406 (0.253-0.642)	0.0001

Subsequently, NMR metabolic phenotype data were included in this model to examine whether they have added value for disease status prediction. Both global tests (Goeman et al., 2005, Boulesteix and Hothorn, 2010) and the LRT indicate that NMR metabolic phenotype data is associated with the disease given the clinical risk factors (Table 2). The results obtained for a feature by feature analysis are shown in Section 1.4 in the supplementary appendix of the paper.

Table 2: P-values obtained from the three testing procedures.

Goeman et al.	**Boulesteix and Hothorn**	**LRT**
p-value	p-value	p-value
8.1×10^{-5}	0	2.2×10^{-16}

4.2 Development of predictive models for lung cancer using metabolica data: Lasso and elastic net

In a next step, predictive models with and without NMR metabolic phenotype data were developed by penalized logistic regression models using Lasso and elastic net penalties in order to evaluate the added predictive value of the NMR metabolic phenotype data. Table 3 presents the results obtained for all classification procedures.

Table 3: Prediction performance parameters resulting from the penalized logistic regression models using Lasso and elastic net penalties and the random forests analysis. Data are presented as the mean percentage over the cross validation loops (± standard deviation). Abbreviations: MCE: misclassification error, PPV: positive predictive value, NPV: negative predictive value, SEN: sensitivity, SPE: specificity.

		MCE (S.D)	**SENS (S.D)**	**SPEC (S.D)**	**NPV (S.D)**	**PPV (S.D)**
Logistic model	Test 1	24.6 (2.7)	76.6 (4.7)	74.4 (4.8)	75.4 (4.5)	75.7 (4.2)
(clinical)	Test 2	23.0 (1.7)	71.0 (3.5)	83.0 (2.1)	74.2 (2.2)	80.7 (1.8)
Lasso	Test 1	18.6 (3.4)	80.9 (4.5)	82.1 (5.0)	80.5 (4.6)	82.5 (4.6)
(Clinical + metabolic phenotype)	Test 2	22.6 (3.4)	70.9 (5.8)	83.9 (2.6)	74.4 (4.0)	81.5 (2.9)
Elastic net	Test 1	17.9 (3.4)	81.5 (4.5)	82.8 (5.0)	81.1 (4.5)	83.2 (4.6)
(Clinical + metabolic phenotype)	Test 2	21.8 (3.5)	72.5 (5.9)	83.8 (2.4)	75.5 (4.2)	81.7 (2.9)
Random Forests	Test 1	22.9 (0.9)	79.1 (1.5)	74.5 (0.9)	77.5 (1.3)	76.4 (0.7)
(Clinical)	Test 2	27.4 (2.1)	78.6 (2.8)	66.7 (2.1)	75.7 (2.8)	70.2 (1.8)
Random Forests	Test 1	16.4 (0.6)	80.2 (1.0)	86.7 (0.8)	80.9 (0.8)	86.3 (0.7)
(clinical + metabolic phenotype)	Test 2	16.1 (0.8)	84.5 (1.0)	84.5 (1.1)	84.3 (0.9)	84.3 (1.0)

For the Lasso method, the average misclassification error (MCE) of test cohort 1 dropped from 24.6% to 18.6% when the NMR metabolic phenotype data were included in the model. A smaller decrease is observed for test cohort 2, the average MCE dropped from 23.0% to 22.6%. Similar patterns were observed for the models based on the elastic net penalty, i.e. the average MCE of test cohort 1 declined from 24.6% to 17.9% and that of test cohort 2 dropped from 23.0% to 21.8% when the NMR metabolic phenotype data were included in the model. Density estimates for the distribution of the MCE are shown in Figure 2 (upper row). For test cohort 1, the sensitivity, specificity, PPV and NPV of the penalized Lasso and elastic net models increased when NMR metabolic phenotype data were added to the logistic regression model. However, for the test cohort 2 they remained quasi constant (Table 3). The density plots for PPV, NPV, sensitivity and specificity for the penalized Lasso and elastic net models are given in the supplementary (Figure S3-S6).

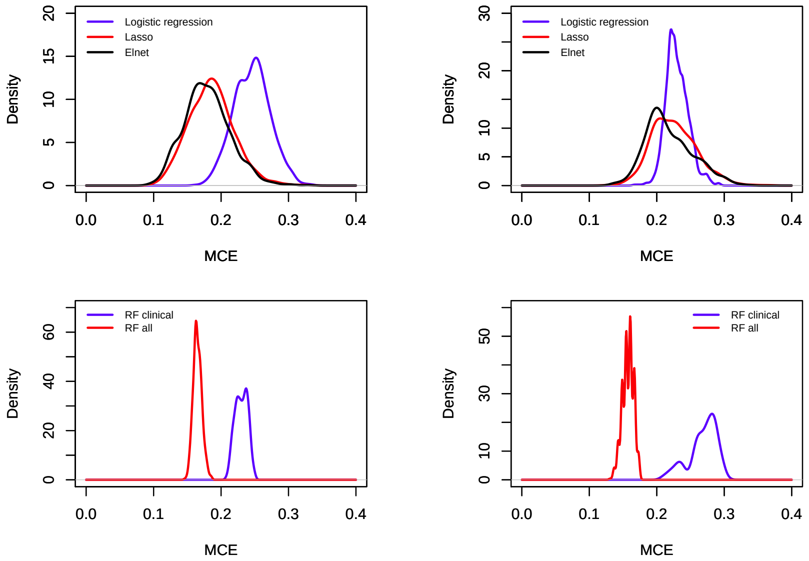


Figure 2: Density estimates for the distribution of the misclassification error obtained from the cross validation loop. Left panel: density plot for test cohort 1. Right panel: density plot for test cohort 2. Abbreviations: Elnet: Elastic Net, MCE: misclassification error, RF: random forest.

For the random forest analysis, the average MCE when only clinical risk factors were included was 22.9% for test cohort 1 and dropped to 16.4% when the NMR metabolic phenotype data were added to the model. Similarly, the average MCE of test cohort 2 declined from 27.4 to 16.1% when both clinical and metabolic phenotype data were included in the model. Density estimates for the distribution of the MCE are shown in Figure 2 (lower row).
Moreover, the PPV and the specificity increased for test cohort 1 when the metabolic phenotype data were included in the model, while the NPV and sensitivity remained stable. Furthermore, PPV, NPV, sensitivity and specificity increased significantly for test cohort 2 when the metabolic phenotype data were included in the model (Table 3). Figure 3 shows the variable importance for the projection plot which indicates that 24 out of 28 of the most discriminating variables constitute NMR integration values. These results further confirm that NMR metabolic phenotype information has added value for disease prediction. Besides the number of smoking pack years, the presence of COPD and the smoking habits, the plasma concentration of threonine (VAR49 and 50), glycerol (VAR45 and 46) and valine (VAR48) contribute the most to the discriminative power of the model. Thus, besides the known clinical risk factors for lung cancer (smoking and the presence of COPD), altered plasma levels of these metabolites seem to be predictive factors for lung cancer as well. Thus, both penalized logistic regression models using Lasso and elastic net penalties and RF analysis indicate that the addition of NMR metabolic phenotype data to classical risk models that only take clin-

ical risk factors into account reduces the MCE. These findings are comparable to those found for studies in which the addition of genetic risk markers and mutagen sensitivity data improved the performance of risk models including only clinical risk factors (Raji et al., 2010, Spitz et al., 2008). However, it has to be kept in mind that these results were achieved in a cohort consisting of controls and patients with a known diagnosis of lung cancer and that further independent validation is needed in large-scale prospective screening studies with asymptomatic, high risk-individuals who are eligible for LDCT screening.

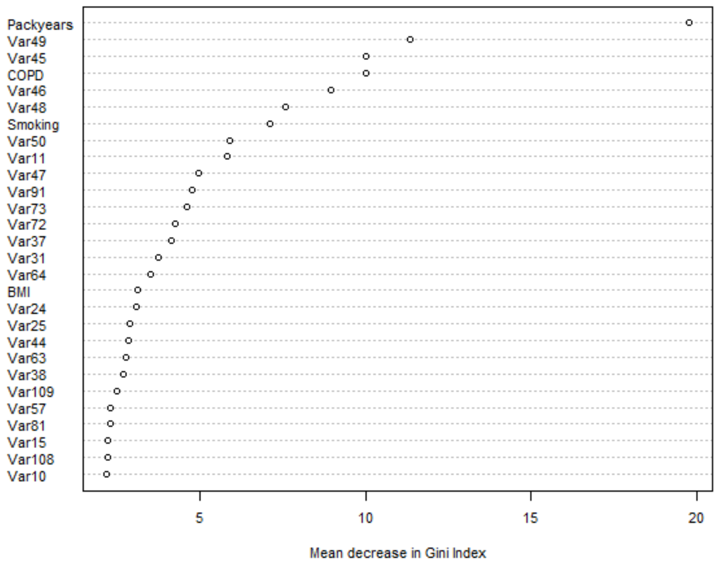


Figure 3: Variable importance plot obtained from a single random forests analysis including both clinical risk factors and ^{1}H-NMR metabolic phenotype data (top 28 important variables). Abbreviations: BMI: body mass index, COPD: chronic obstructive pulmonary disease.

4.3 Permutation based assessment

In this section, we compare the results, in terms of misclassification error, obtained for permutation procedure describe in Section 3.4 with the results obtained in the previous section. Our focus is placed on the distribution of the misclassification error, for each predictive model, with and without permutation.

This was done in order to assess if the obtained results without permutation were not due to chance. Note that the permuted data represents a scenario in which the metabolic data is not associated with the disease status and therefore do not expect to reduce the MCE. In total 1000 permutations were applied, for each permuted data a cross validation loop of 1000 iteration was used. Figure 4 (upper row) shows the distributions of the MCE in 1000 permutations (Lasso and elastic net) on the test cohort 1 (left panel) and the test cohort 2 (right panel). The lower row in Figure 4 shows the results obtained using RF method to the permuted data.

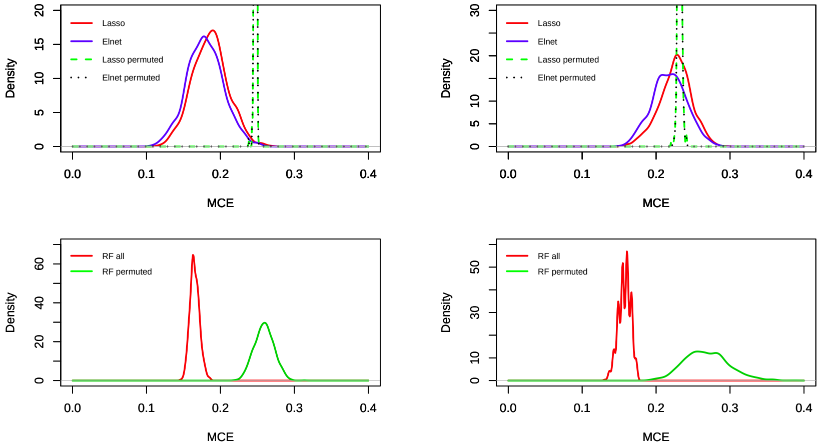


Figure 4: Density estimates for the distribution of the misclassification error. Left panel: density plot for test cohort 1. Right panel: density plot for test cohort 2. Upper panels: Lasso and elastic net. Lower panels: Random forest. Abbreviations: Lasso/Elnet/RF all: Lasso, elastic net, random forest built using both clinical risk factors and metabolic data, Lasso permuted/ Elnet permuted/ RF permuted: Lasso, elastic net, random forest built using clinical risk factors and permuted metabolic data.

Table 4: Average MCE values. The number in parentheses show the 2.5% and 97.5% quantiles of the estimated MCE in the 1000 cross validation iterations.

		Lasso	**Elastic net**	**RF**
Original data	Test 1	18.6 (16.9, 23.4)	17.9 (16.3, 23.0)	16.5 (16.0, 17.9)
	Test 2	22.6 (21.3, 26.8)	21.8 (20.2, 26.4)	15.7 (14.9, 17.3)
Permuted data	Test 1	24.7 (24.6, 24.9)	24.7 (24.6, 24.9)	25.9 (25.0, 28.5)
	Test 2	23.1 (23.0, 23.6)	23.1 (23.0, 23.6)	26.8 (25.0, 33.3)

We notice that, except for test cohort 2 for lasso and elastic net, the distribution of the misclassification error for the non permuted data is located to the left compared with the distribution of the misclassification error for the permuted data.

The permutation procedure described in this section allows us to test if the inclusion of the metabolic data in the predictive model leads to a significant reduction in the misclassification error. Under the null hypothesis the inclusion of the metabolic data in the predictive model do not reduced the MCE. Under the alternative, inclusion of the metabolic data in the predictive model is expected to reduce the misclassification error. Let $MCE(Y|Z,X)$ be the misclassification error obtained from a predictive model in which both clinical and metabolic data are included. We can approximate the distribution of $MCE(Y|Z,X)$ under the null hypothesis using the permutation procedure described 3.4. The permutation p value is calculated using

$$p\_value = \frac{1+\#\{MCE(Y|Z,X)_b \leq MCE(Y|Z,X)_{obs}\}}{B+1},$$

where, $MCE(Y|Z,X)_{obs}$ is the misclassification error estimated for non-permuted data (and reported in Table 3) while $MCE(Y|Z,X)_b$ is the misclassification error obtained from the bth permuted data, $b = 1,\ldots,1000$. Note that both $MCE(Y|Z,X)_b$ and $MCE(Y|Z,X)_{obs}$ are the average of 1000 cross validation iterations. For both test cohorts and for all methods, $MCE(Y|Z,X)_b > MCE(Y|Z,X)_{obs}$ which implies that the p-value is 0.001.

5 Discussion

Early detection and improving classification procedures for lung cancer are of primary interest. In the current paper we focused on the development of predictive models for lung cancer based on phenotype data and NMR metabolic data.
We have shown that, using both penalized regression models and random forest, the inclusion of the NMR metabolic data in addition to the phenotypic data in the predictive models results in a decline in misclassification error of the classifiers. We have shown that a classifier that was developed based on one population (Limburg Positron Emission Tomography Center , test cohort 1), can be successfully applied to another population (Department of Respiratory Medicine of University Hospital Leuven, test cohort 2) with the same magnitude of accuracy. Furthermore, we have shown that the reduction in misclassification error (i.e., the increment in accuracy) can be tested using a permutation based inference procedure. This implies that this type of risk models, which include both data sources, can improve accuracy in selection of high-risk individuals eligible for lung cancer screening with LDCT and so might pave the way to a reduction of false positive results and corresponding risk and financial burden.

Acknowledgements

This study is part of the Limburg Clinical Research Program (LCRP) UHasselt-ZOL-Jessa, supported by the foundation Limburg Sterk Merk, province of Limburg, Flemish government, Hasselt University, Ziekenhuis Oost-Limburg and Jessa Hospital. Samples are stored at the University Biobank Limburg (UBiLim). We thank the Research Foundation Flanders (FWO-Vlaanderen) for supporting via the MULTIMAR project and Gunter Reekmans for NMR assistance.

References

Amaratunga, D., J. Cabrera, and Z. Shkedy (2014): *Exploration and Analysis of DNA Microarray and Other High-Dimensional Data*, volume Second Edition, Wiley Series in Probability and Statistics.

Bach, P. B., M. W. Kattan, M. D. Thornquist, M. G. Kris, R. C. Tate, M. J. Barnett, L. J. Hsieh, and C. B. Begg (2003): "Variations in lung cancer risk among smokers," *Journal of the National Cancer Institute*, 95, 470–478.

Bach, P. B., J. N. Mirkin, T. K. Oliver, C. G. Azzoli, D. A. Berry, O. W. Brawley, T. Byers, G. A. Colditz, M. K. Gould, J. R. Jett, et al. (2012): "Benefits and harms of CT screening for lung cancer: a systematic review," *Jama*, 307, 2418–2429.

Benjamini, Y. and Y. Hochberg (1995): "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, 289–300.

Boulesteix, A.-L. and T. Hothorn (2010): "Testing the additional predictive value of high-dimensional molecular data," *BMC bioinformatics*, 11, 78.

Breiman, L. (2001): "Random forests," *Machine learning*, 45, 5–32.

Casella, G. and R. L. Berger (2001): *Statistical inference*, volume 2, Duxbury Pacific Grove, CA.

Cassidy, A., J. P. Myles, M. van Tongeren, R. Page, T. Liloglou, S. Duffy, and J. Field (2008): "The LLP risk model: an individual risk prediction model for lung cancer," *British journal of cancer*, 98, 270–276.

Emwas, A.-H. M., R. M. Salek, J. L. Griffin, and J. Merzaban (2013): "NMR-based metabolomics in human disease diagnosis: applications, limitations, and recommendations," *Metabolomics*, 9, 1048–1072.

Ferlay, J., I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, and F. Bray (2015): "Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012," *International Journal of Cancer*, 136, E359–E386.

Field, J. K. and S. Duffy (2008): "Lung cancer screening: the way forward," *British journal of cancer*, 99, 557–562.

Field, J. K., M. Oudkerk, J. H. Pedersen, and S. W. Duffy (2013): "Prospects for population screening and diagnosis of lung cancer," *The Lancet*, 382, 732–741.

Friedman, J., T. Hastie, and R. Tibshirani (2001): *The elements of statistical learning*, volume 2, Springer series in statistics Springer, Berlin.

Goeman, J. J., J. Oosting, A.-M. Cleton-Jansen, J. K. Anninga, and H. C. Van Houwelingen (2005): "Testing association of a pathway with survival using gene expression data," *Bioinformatics*, 21, 1950–1957.

Goldstraw, P., J. Crowley, K. Chansky, I. A. for the Study of Lung Cancer International Staging Committee, et al. (2007): “The iaslc lung cancer staging project: proposals for the revision of the tnm stage groupings in the forthcoming (seventh) edition of the tnm classification of malignant tumours,” *J Thorac Oncol*, 2, 706–714.

Hoggart, C., P. Brennan, A. Tjonneland, U. Vogel, K. Overvad, J. N. Østergaard, R. Kaaks, F. Canzian, H. Boeing, and A. Steffen (2012): “A risk model for lung cancer incidence,” *Cancer Prevention Research*, 5, 834–846.

Horeweg, N., E. T. Scholten, P. A. de Jong, C. M. van der Aalst, C. Weenink, J.-W. J. Lammers, K. Nackaerts, R. Vliegenthart, K. ten Haaf, U. A. Yousaf-Khan, et al. (2014): “Detection of lung cancer through low-dose CT screening (NELSON): a prespecified analysis of screening test performance and interval cancers,” *The Lancet Oncology*, 15, 1342–1350.

Lindon, J. C. and J. K. Nicholson (2008): “Spectroscopic and statistical techniques for information recovery in metabonomics and metabolomics,” *Annu. Rev. Anal. Chem.*, 1, 45–69.

Louis, E., P. Adriaensens, W. Guedens, T. Bigirumurame, K. Baeten, K. Vanhove, K. Vandeurzen, K. Darquennes, J. Vansteenkiste, C. Dooms, S. Ziv, M. Liesbet, and T. Michiel (2016): “Detection of lung cancer through metabolic changes measured in blood plasma,” *Journal of Thoracic Oncology*, 11, 516–523.

Louis, E., P. Adriaensens, W. Guedens, K. Vanhove, K. Vandeurzen, K. Darquennes, J. Vansteenkiste, C. Dooms, E. de Jonge, and M. Thomeer (2015a): “Metabolic phenotyping of human blood plasma: a powerful tool to discriminate between cancer types?” *Annals of Oncology*, mdv499.

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

Mamas, M., W. B. Dunn, L. Neyses, and R. Goodacre (2011): “The role of metabolites and metabolomics in clinically applicable biomarkers of disease,” *Archives of toxicology*, 85, 5–17.

Manser, R., L. B. Irving, C. Stone, G. Byrnes, M. Abramson, and D. Campbell (2004): “Screening for lung cancer,” *Cochrane Database Syst Rev*, 1, CD001991.

National Lung Screening Trial Research Team (2011a): “The national lung screening trial: Overview and study design,” *Radiology*, 258, 243–253.

National Lung Screening Trial Research Team (2011b): “Reduced lung-cancer mortality with low-dose computed tomographic screening,” *The New England journal of medicine*, 365, 395–409.

National Lung Screening Trial Research Team (2013): "Results of initial low-dose computed tomographic screening for lung cancer," *The New England journal of medicine*, 368, 1980–1991.

Raji, O. Y., O. F. Agbaje, S. W. Duffy, A. Cassidy, and J. K. Field (2010): "Incorporation of a genetic factor into an epidemiologic model for prediction of individual risk of lung cancer: the liverpool lung project," *Cancer Prevention Research*, 3, 664–669.

Spitz, M. R., C. J. Etzel, Q. Dong, C. I. Amos, Q. Wei, X. Wu, and W. K. Hong (2008): "An expanded risk prediction model for lung cancer," *Cancer Prevention Research*, 1, 250–254.

Spitz, M. R., W. K. Hong, C. I. Amos, X. Wu, M. B. Schabath, Q. Dong, S. Shete, and C. J. Etzel (2007): "A risk model for prediction of lung cancer," *Journal of the National Cancer Institute*, 99, 715–726.

Tammemagi, M. C., H. A. Katki, W. G. Hocking, T. R. Church, N. Caporaso, P. A. Kvale, A. K. Chaturvedi, G. A. Silvestri, T. L. Riley, and J. Commins (2013): "Selection criteria for lung-cancer screening," *New England Journal of Medicine*, 368, 728–736.

Tammemagi, M. C. and S. Lam (2014): "Screening for lung cancer using low dose computed tomography," *BMJ*, 348, g2253.

Tibshirani, R. (1996): "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Tsay, J.-C. J., C. DeCotiis, A. K. Greenberg, and W. N. Rom (2014): "Current readings: blood-based biomarkers for lung cancer," in *Seminars in thoracic and cardiovascular surgery*, Elsevier, 328–334.

van den Bergh, K. A., M.-L. Essink-Bot, E. M. Bunge, E. T. Scholten, M. Prokop, C. A. van Iersel, R. J. van Klaveren, and H. J. de Koning (2008): "Impact of computed tomography screening for lung cancer on participants in a randomized controlled trial (NELSON trial)," *Cancer*, 113, 396–404.

Wood, D. E., G. A. Eapen, D. S. Ettinger, L. Hou, D. Jackman, E. Kazerooni, D. Klippenstein, R. P. Lackner, L. Leard, A. N. Leung, et al. (2012): "Lung cancer screening," *Journal of the National Comprehensive Cancer Network*, 10, 240–265.

Woolf, S. H., R. P. Harris, and D. Campos-Outcalt (2014): "Low-dose computed tomography screening for lung cancer: How strong is the evidence?" *JAMA internal medicine*, 174, 2019–2022.

Zhaoa, Y. R., X. Xiea, H. J. de Koningb, W. P. Malic, R. Vliegentharta, and M. Oudkerka (2011): "NELSON lung cancer screening study," *Cancer Imaging*, 11, S79–S84.

Zou, H. and T. Hastie (2005): "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 301–320.

Improvement of risk models to select individuals eligible for lung cancer screening with low-dose computed tomography: adding the metabolic phenotype

Theophile Bigirumurame[1*], Louis Evelyne[2*], *et al.*

Supplementary Appendix

1.1 Introduction

This appendix include materials which were referred to in the main manuscript. In Section 1.2 we present the data collection protocol for the lung cancer study. In Section 1.3 we present patients' descriptive characteristics in the two cohorts, while in Section 1.4, we present the results obtained from the feature by feature analysis. Finally, in Section 1.5 we present the results of several validation scores mentioned in Section 3.3.1 in the paper.

1.2 Blood sampling, sample preparation and NMR analysis

Blood sampling and processing

Fasting venous blood was collected in 10 ml lithium-heparin tubes and stored at 4°C within 5 min. Within 8 h after collection, samples were centrifuged at 1600 g for 15 min and plasma aliquots of $500\mu l$ were transferred into cryovials and stored at -80°C. NMR sample preparation and analysis After thawing, plasma aliquots were centrifuged at 13000 g for 4 min at 4°C, followed by diluting $200\mu l$ of the supernatant with $600\mu l$ deuterium oxide containing $0.3\mu g/\mu l$ trimethylsilyl-2,2,3,3-tetradeuteropropionic acid (TSP) as a chemical shift reference. Samples were placed on ice until ^{1}H-NMR analysis. After mixing and transferring into 5 mm NMR tubes, the samples were acclimatized to 21°C during 7 min. The ^{1}H-NMR spectra were recorded on a 400 MHz (9.4 Tesla) Inova spectrometer (Agilent Technologies Inc.) at 21°C. Slightly T2-weighted spectra were acquired using the Carr-Purcell-Meiboom-Gill pulse sequence (total spin-echo time of 32 ms; interpulse delay of 0.1 ms), preceded by an initial preparation delay of 0.5 s and 3 s presaturation for water suppression. Other parameters were a spectral width of 6000 Hz, an acquisition time of 1.1 s, 13 K complex data points and 96 scans. Each free induction decay was zero-filled to 65 K points and multiplied by a 0.7 Hz exponential line-broadening function prior to Fourier transformation. Spectral processing ^{1}H-NMR spectra were phased, baseline corrected and referenced to TSP (δ=0.015 ppm) and segmented into 110 variable-sized integration regions, excluding water (4.7-5.2 ppm) and TSP (-0.3-0.3 ppm). The rational segmentation into 110 regions is based on spiking with known metabolites (Louis et al., 2015b). The 110 regions were integrated and normalized relative to the total integrated area (except water and TSP), resulting in 110 normalized integration values, being the variables for multivariate statistics.

1.3 Descriptive analysis for the clinical risk factors

The subject characteristics of the first cohort of 536 subjects (273 lung cancer patients and 263 controls) and the second fixed validation cohort (test cohort 2) of 168 subjects (84 lung cancer patients and 84 controls) are shown in Supplementary Table S 1 and 2.

Supplementary Table S 1: Subject characteristics of cohort 1. Data are presented as mean ± standard deviation, unless otherwise indicated. Univariate logistic regression models were used to calculate p-values for continuous variables, while a Chi-square test was used to compute p-values for categorical variables.

	Controls (n=263)	**Patients (n=273)**	**p-value**
Gender, n (%)			
Male	139 (53)	186 (68)	0.0003
Female	124 (47)	87 (32)	
Age, yrs	66 ± 11	68 ± 10	0.056
BMI, kg/m^2	28.0 ± 5.1	25.8 ± 4.5	<0.0001
Smoking habits, n (%)			
Smoker	60 (23)	131 (48)	<0.0001
Ex-smoker	111 (42)	131 (48)	
Non-smoker	92 (35)	11 (4)	
Smoking pack years	16 ± 23	33 ± 21	<0.0001
Previous mine-worker, n (%)			
Yes	15 (6)	24 (9)	0.001
No	124 (47)	162 (59)	
Not applicable	124 (47)	87 (32)	
Prior diagnosis of malignant tumor, n (%)			
Yes	9 (3)	24 (9)	0.016
No	254 (97)	249 (91)	
COPD, n (%)			
Yes	30 (11)	139 (51)	<0.0001
No	233 (89)	134 (49)	
Diabetes, n (%)			
Yes	47 (18)	47 (17)	0.932
No	216 (82)	226 (83)	
Taking lipid-lowering medication, n (%)			
Yes	149 (57)	142 (52)	0.322
No	114 (43)	131 (48)	
Taking malfunctioning thyroid medication, n (%)			
Yes	18 (7)	9 (3)	0.093
No	245 (93)	264 (97)	
Taking anti-arrhythmic medication, n (%)			
Yes	13 (5)	33 (12)	0.005
No	250 (95)	240 (88)	
Taking blood pressure-lowering medication, n (%)			
Yes	193 (73)	168 (62)	0.005
No	70 (27)	105 (38)	
Taking anti-coagulants medication, n (%)			
Yes	163 (62)	155 (57)	0.255
No	100 (38)	118 (43)	

Supplementary Table S 2: Subject characteristics of test cohort 2. Data are presented as mean standard deviation, unless otherwise indicated. Univariate logistic regression models were used to calculate p-values for continuous variables, while a Chi-square test was used to compute p-values for categorical variables. Abbreviations: BMI: body mass index, COPD: chronic obstructive pulmonary disease.

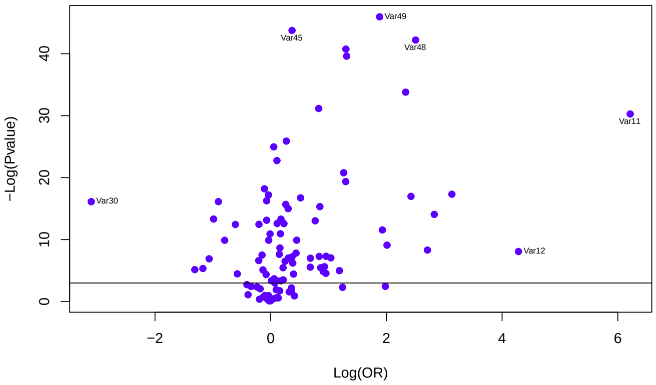
	Controls (n=84)	**Patients (n=84)**	**p-value**
Gender, **n (%)**			
Male	39 (46)	60 (71)	0.001
Female	45 (54)	24 (29)	
Age, yrs	70 ± 10	64 ± 9	0.0002
BMI, kg/m^2	28 ± 6	26 ± 5	0.007
Smoking habits, n (%)			
Smoker	9 (11)	44 (52)	<0.0001
Ex-smoker	36 (43)	37 (44)	
Non-smoker	39 (47)	3 (4)	
Smoking pack years	13 ± 17	41 ± 22	<0.0001
Previous mine-worker, n (%)			
Yes	4 (5)	2 (2)	0.002
No	35 (41)	58 (69)	
Not applicable	45 (54)	24 (29)	
Prior diagnosis of malignant tumor, n (%)			
Yes	3 (4)	12 (14)	0.0304
No	81 (96)	72 (86)	
COPD, n (%)			
Yes	7 (8)	31 (37)	<0.0001
No	77 (92)	53 (63)	
Diabetes, n (%)			
Yes	20 (24)	11 (13)	0.116
No	64 (76)	73 (87)	
Taking lipid-lowering medication, n (%)			
Yes	49 (58)	34 (41)	0.0307
No	35 (42)	50 (59)	
Taking malfunctioning thyroid medication, n (%)			
Yes	11 (13)	8 (10)	0.6261
No	73 (87)	76 (90)	
Taking anti-arrhythmic medication, n (%)			
Yes	9 (11)	4 (5)	0.2481
No	75 (89)	80 (95)	
Taking blood pressure-lowering medication, n (%)			
Yes	65 (77)	47 (56)	0.0054
No	19 (23)	37 (44)	
Taking anti-coagulants medication, n (%)			
Yes	60 (71)	32 (38)	<0.0001
No	24 (29)	52 (62)	

1.4 Feature by feature analysis

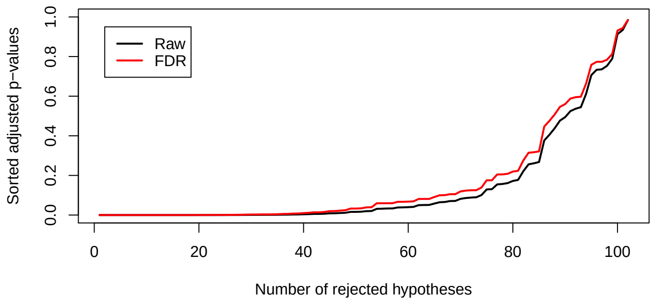
Multiple logistic regression models comprising significant clinical risk factors and one NMR variable (integration value) at a time were fitted to examine which integration values have a significant effect on disease status prediction. The linear predictor for the models is given by:

$$logit(\pi_i) = \alpha_0 + \sum_{l=1}^{p} \alpha_l Z_{il} + \beta_j X_{ij}. \qquad (6)$$

After adjusting for multiple testing, 53 out of the 102 integration values were found to have a significant effect on disease status prediction. The false discovery rate procedure was used to adjust for multiple testing (Benjamini and Hochberg, 1995, Amaratunga et al., 2014). Supplementary Figure S 2 shows the adjusted and unadjusted p-values obtained from the models in equation 6.



Supplementary Figure S 1: Volcano plot for the integration regions specific models.



Supplementary Figure S 2: Unadjusted and adjusted p-values obtained for the tests of the NMR integration values: p-values obtained from a model containing significant clinical risk factors and one NMR integration value at a time. Abbreviations: FDR: false discovery rate.

1.5 Validation scores

Several validation scores were used to evaluate the performance of the classifier in the two test cohorts. For a binary classification problem let Y_i and $\hat{Y}_i$ be the true and the predicted status (class) of a subject, respectively.

$$Y_i = \begin{cases} 1 & \text{observed class is cancer} \\ 0 & \text{observed class is control} \end{cases} \text{ and } \hat{Y}_i = \begin{cases} 1 & \text{predicted class is cancer} \\ 0 & \text{predicted class is control} \end{cases}$$

The observed and the predicted classes are used to form a confusion matrix from which the validation scores are computed. Table S 3 shows such a matrix.

Supplementary Table S 3: Confusion matrix.

		Predicted status	
		1	**0**
Observed Status	**1**	True Positive (TP)	False Negative (FN)
	0	False Positive (FP)	True Negative (TN)

The following statistics can be computed and used to assess the performance of a given classifier.

- The misclassification error is the total number of mistakes committed using the classification procedure, that is:

$$MCE = \frac{FN + FP}{(FN + FP + TP + FP)}.$$

- The specificity of a classifier measures the proportion of negative cases (controls) that are correctly classified,

$$SPE = \frac{TN}{(TN + FP)}.$$

- The sensitivity of a classifier measures the proportion of positive cases (cancer for instance) that are correctly classified,

$$SEN = \frac{TP}{(TP + FN)}.$$

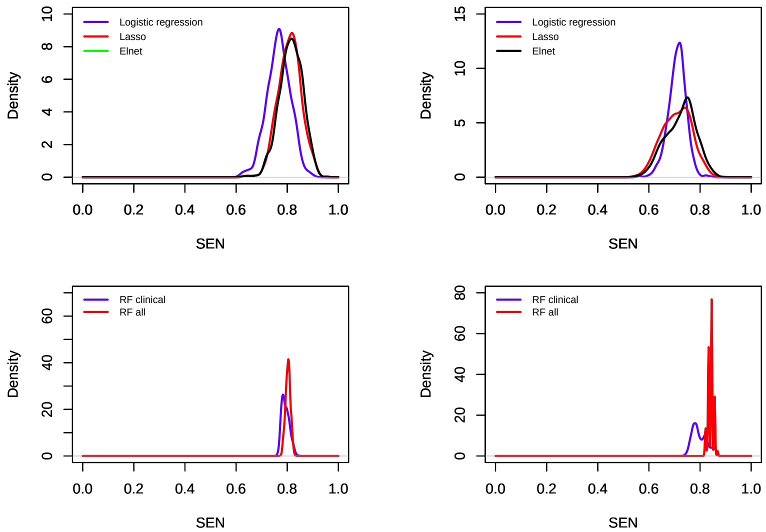
- The positive predictive value of a classifier measures the proportions of predicted positive cases (cancer subjects) that are true positive cases,

$$PPV = \frac{TP}{(TP + FP)}.$$

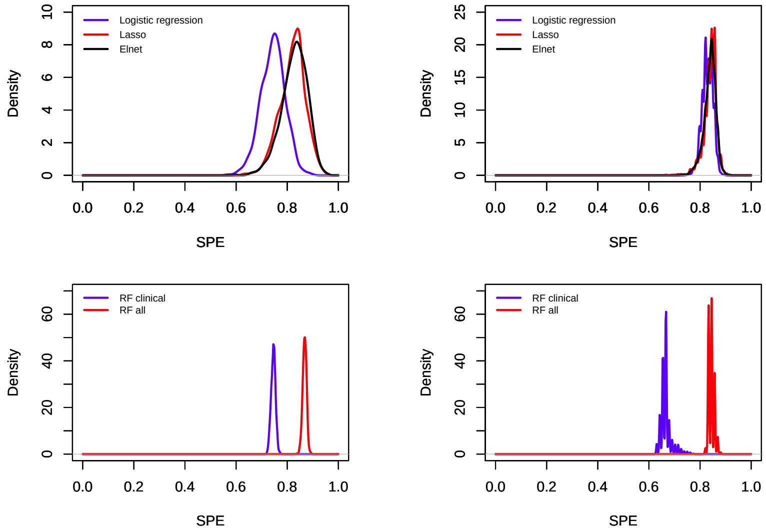
- The negative predictive value of a classifier measures the proportions of predicted negative cases (control subjects) that are true negative cases,

$$NPV = \frac{TN}{(TN + FN)}.$$

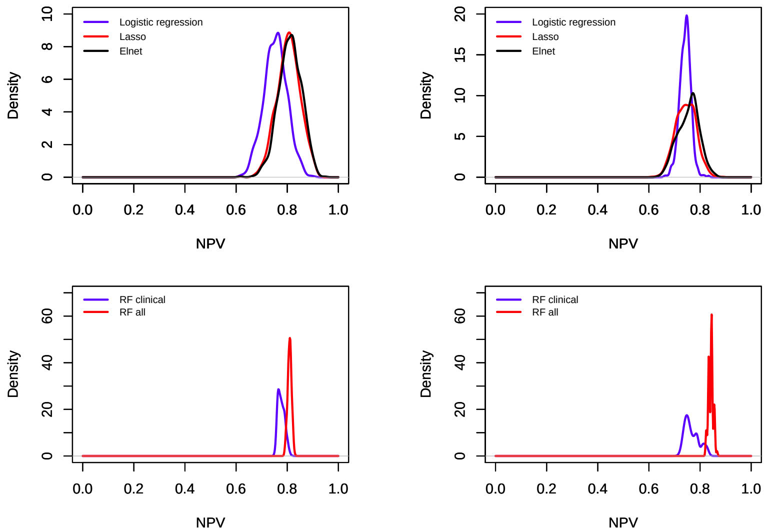
In this section, we present the density estimates for the validation scores mentioned in Section3.3.1 based on a 3 fold cross validation loop of 1000 steps. The density estimates are presented for both test cohort mentioned in Section 3.2.



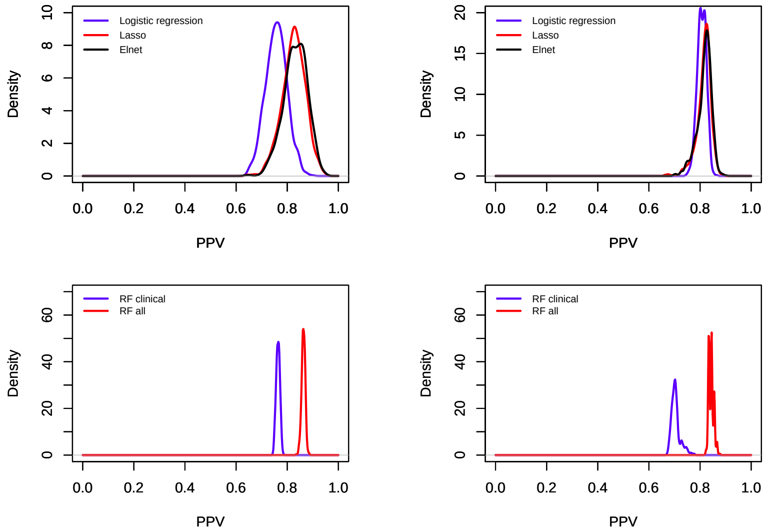
Supplementary Figure S 3: Density estimates for the distribution of the sensitivity. Left panel: density plot for test cohort 1. Right panel: density plot for test cohort 2. Abbreviations: Elnet: Elastic Net, SEN: sensitivity, RF: random forest.



Supplementary Figure S 4: Density estimates for the distribution of the specificity. Left panel: density plot for test cohort 1. Right panel: density plot for test cohort 2. Abbreviations: Elnet: Elastic Net, SPE: specificity, RF: random forest.



Supplementary Figure S 5: Density estimates for the distribution of the negative predictive value. Left panel: density plot for test cohort 1. Right panel: density plot for test cohort 2. Abbreviations: Elnet: Elastic Net, NPV: negative predictive value, RF: random forest.



Supplementary Figure S 6: Density estimates for the distribution of the positive predictive value. Left panel: density plot for test cohort 1. Right panel: density plot for test cohort 2. Abbreviations: Elnet: Elastic Net, PPV: positive predictive value, RF: random forest.