PROJECT ANALYSIS OF GENE EXPRESSION REPORT

METABOLIC SIGNATURE IN LUNG CANCER : READY FOR DIAGNOSTIC USE ?

MARC JOIRET

FOR THE ATTENTION OF ZIV SHKEDY*

April 28th, 2017

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

This "Analysis of Gene Expression" project has two parts (part A and part B). The aim of part A is to find out how many metabolic feature variables from blood samples are informative enough to build a classifier with the best performance as possible in detecting whether or not patients have lung cancer. Which specific variables are the best in discriminating between the two conditions is of interest. The metabolic signature classifier will be built with supervised learning methods trained and cross-validated on a so-called `Metabolite.Train` data set. The reliability (robustness) of this metabolic signature will eventually be assessed on completely independent data. The aim of part B is to show if the new metabolic signature can increase the predictive accuracy of existing models based on known standard risk factors for lung cancer prognostic. The existing predictor feature variables and data are provided in a so-called `Clinical.Train` data set.

* *Interuniversity Institute for Biostatistics and statistical Bioinformatics, Hasselt University, Belgium*

1 INTRODUCTION

1.1 Development of metabolic signature for lung cancer

WHAT IS A METABOLIC SIGNATURE ?

A metabolic signature is a particular profile of the concentration levels in a variety of metabolites that are analysed in blood sample from identified patients and which can be associated to a particular medical condition of the patient. Other groups have investigated distinct metabolic fingerprints between colon and ovarian cancer cell lines [1]. In this project, the metabolic profile is searched for from a data set of 200 patients for whom 110 blood sampled metabolites were recorded. Half of the patients (100) have lung cancer and half (100) are control cases. Clinical variables are also provided for each patient with classical standard features (age, gender, smoking status, pack-years, ...).

The project consists of two main tasks :

PART A : METABOLIC SIGNATURE FEATURES DEVELOPMENT AND CLASSIFIER CONSTRUCTION

The 110 metabolites are blind to the statistical analyst. Only a variable ID is provided (Var 1, Var2, ..., Var110) with the concentration levels of the 200 patients blood samples. The aim is to construct a metabolic signature profile from a subset of the 110 features variables. The number of informative variables is part of the question to solve. Not all the feature variables are necessarily useful for the construction of a good classifier. The classifier can be built from a variety of supervised learning machine techniques. The performance of the classifier will be measured with standard relevant metrics in biomedicine: misclassification rate, sensitivity, specificity, positive predictive values, negative predictive values, receiver operating curve (ROC) and its area under the curve (AUC). The task is to build the best classifier as possible to distinguish between future controls and lung cancer patients.

PART B : UTILITY EVALUATION OF THE NEWLY BUILT CLASSIFIER AS COMPARED TO EXISTING MODEL BASED ON STANDARD CLINICAL VARIABLES

Risk prediction models for lung cancer for a given individual, based on standard clinical variables are documented in the literature. In the Liverpool Lung Project (LLP)[2], significant classical variables were fitted into a multiple logistic regression model. Standard classical predictors are variables like gender, age, smoking duration, pack-years, prior diagnosis of malignant tumor, family history of lung cancer, exposure to asbestos. In this part B of the project, such a multiple logistic regression is mirrored from the LLP study and applied to the `Clinical.Train` received data set to build a reference standard model based on most significant known risk factors. The aim of part B is to test if the new prognostic metabolite signature developed in part A can provide increased predictive accuracy to the existing model in addition to the known risk factors.

1.2 Supervised learning special concerns

A special attention is dedicated to two most important issues in the supervised learning techniques used for this project, namely the splitting rule of the data for cross-validation and the variable selection rule. A pitfall would be to overfit the classifier on the available data, preventing successful generalization and successful application of the built classifier to future independent data. This is where cross-validation comes into play. We will select the feature variables (metabolic variables) and train the classifier several times on split subsets of the available samples and validate the classifier on the remain-

ing samples. To avoid any special sample effect, the splitting rule will be repeated a number of times and the results will be aggregated for the classifier performance evaluation. All samples are given the chance to be part of the training set or part of the validation set. The splitting rule addresses the samples only (patients in the data set). For each repetition, whatever the sample, all the 110 variables are systematically ranked for the significant difference they can show across the two conditions we want to classify : cancer or control. Variables that are more prone to discriminate between the two conditions will be detected (by a statistical Welch t test) and ranked (by increasing p-values). As the ranking result may depend on the sample selected, the sample splitting rule is repeated many times as already mentioned. In this way, all the samples from the available data set are given the chance to contribute to variable selection. How many variables will be selected depends on the marginal performance they bring in the classifier. We will explore what increase in classification performance is brought by incorporating an increased number of variables as part of the feature space. We will start with 2 variables and retain more and more variables until no further increase in classification performance occurs. It is considered possible ahead of the analysis that all the 110 variables could be retained but maybe this will not be necessary. The result of variable selection presents an interest of its own from a biomedicine perspective. Indeed, the selected (significant and best ranked) variables are possibly involved (not by chance) in special biochemical mechanisms the cancer condition has triggered or for which there is an indirect association, as shown in the article by Halama et al for ovarian and colon cancer cell lines. [1].

2 SCIENTIFIC QUESTION

2.1 Research Questions

Part A : Test and determine if a metabolite signature exists from which the most powerful classifier as possible can be built to distinguish between future lung cancer patients and controls.

Part B : Test whether a metabolic signature can lead to a gain in prediction accuracy when used in addition to standard known risk factors.

WHAT IS THE THE MEASURE OF CLASSIFIERS PERFORMANCE ?

The classifier performance is primarily assessed through the AUC of a ROC curve. Other biomedicine metrics will also be monitored : misclassification rate, sensitivity, specificity, positive predictive value and negative predictive value.

3 PART A

3.1 Handling of missing data

The missing data `NA` have been replaced by 0.

3.2 Unsupervised exploration of the feature space

DATA EXPLORATION ANALYSIS

Basically, in the `Metabolite.Train` file, the data are divided on the one hand, by the outcome we are interested in and for which we want to build a classifier : the binary outcome indicating whether or not the patient has a lung cancer (or is control); and on the other hand, by 110 metabolite continuous variables that we call the feature space.

To get some preliminary insight in the feature space, we conducted a principal component analysis (PCA) and presented the results graphically on figure 1 to check for patterns. Out of the 110 principal components (PC), the first 4 PC explain 58.6% of the variability. To capture more than 80% of the variability, at least 11 PC are required. Four pairwise graphs of the 4 first principal components are displayed on figure 1. No obvious pattern appears on these graphs. This result could be expected from a large number of variables which are basically noisy altogether. This result motivates the need of a variable selection procedure to search for the most informative variables where *informative* means the ability to discriminate between the two groups (cancer and control).

NORMALITY ASSUMPTION FOR VARIABLES IN THE FEATURE SPACE

We explore the normality of the variables from the feature space. We took, as illustrative example, Var45 from the data set `Metabolite.Train`. The histograms of the Var45 are displayed across the 2 groups on figure 2 with the QQ plot assessing the normality assumption. Normality appears to be a reasonable assumption for this metabolite.
This does not assure the normality of all the variables but we take this assumption as a first approximation useful for later Welsh `t` tests. Note that for `Var45`, the mean and standard deviation are higher for the Cancer group (lower left panel in figure 2) than for the control group (upper left panel in figure 2). The normality assumption is not necessarily required as we have shown that with non-parametric methods (Wilcoxon sum rank tests), the variable selection procedure lead to the same results.

3.3 Note on Software used

To conduct the variable selection, data splitting, classification with repeated cross-validation of the different learning algorithms we used the bioconductor package called CMA (Classification for microarrays) that was developed by Slawski, M. and Boulesteix, A.L.[3].

3.4 Rationale behind variable selection

HOW MANY METABOLITE VARIABLES SHOULD BE SELECTED AND WHICH ONES ?

From the previous section, it is suspected that all the metabolite variables are not informative to be used as features for building a classifier with good predictive accuracy. How many variables should

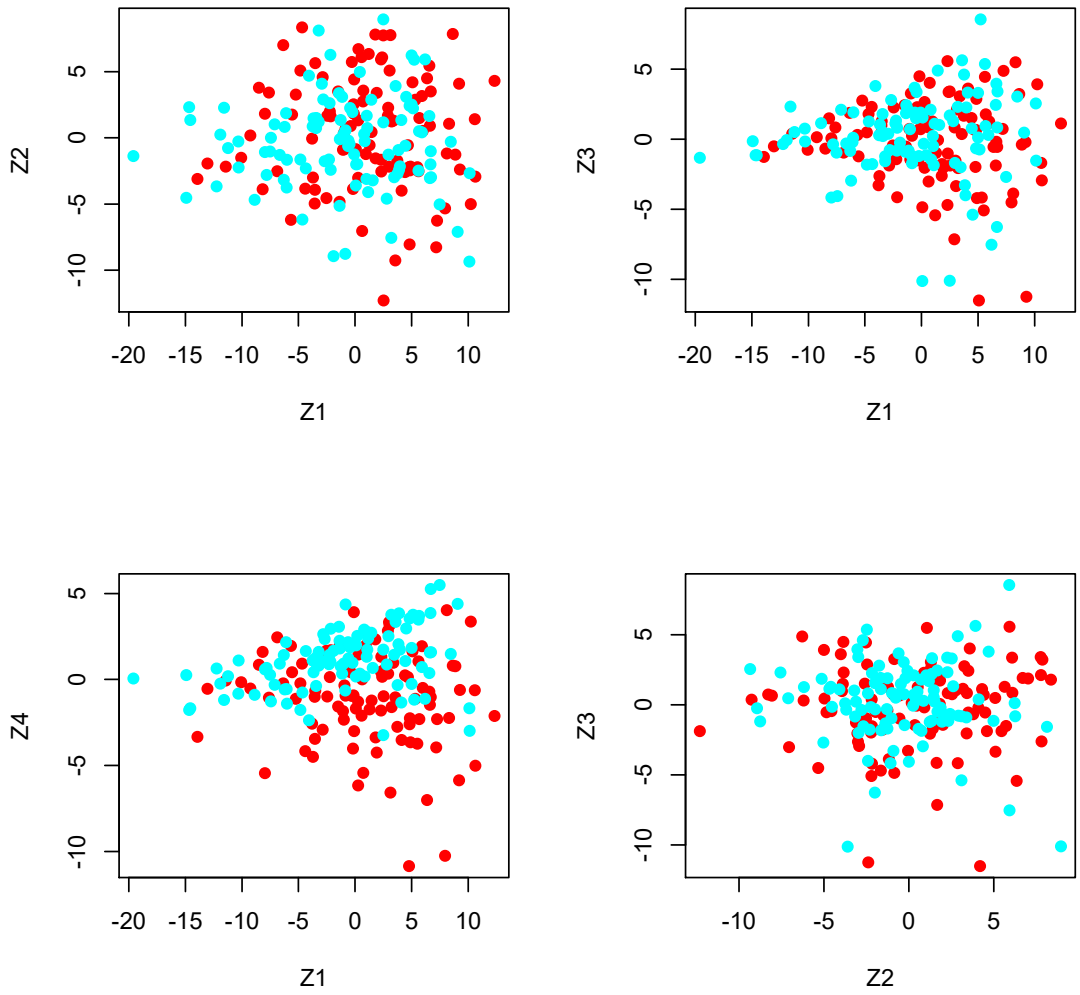

Figure 1: Scatter plot of PC by Group (Lung Cancer or Control), scaled data.

be selected and which ones ?

We will follow the following algorithm to determine how many variable we should use in the classifier to be build :

1. Rank the variables (features) for significant differences across the group (cancer or control). The ranking sorts the feature variables by increasing p-values of the Welsh t test testing whether the considered variable shows significant differences across the 2 groups. The exact samples used to conduct the Welch t test depends on the data splitting rule of the cross-validation (see next section). The reason why the modified Welsh test is used is that it accommodates for possible heteroscedasticity of the considered variable across the 2 groups[1].

2. Choose k as the number of variables to include as input features for a classifier. We start with k=2.

3. The first k ranked variables are used as input for a classifier.

1 The non-parametric Wilcoxon rank sum test gave similar results for the whole procedure (not reported here).

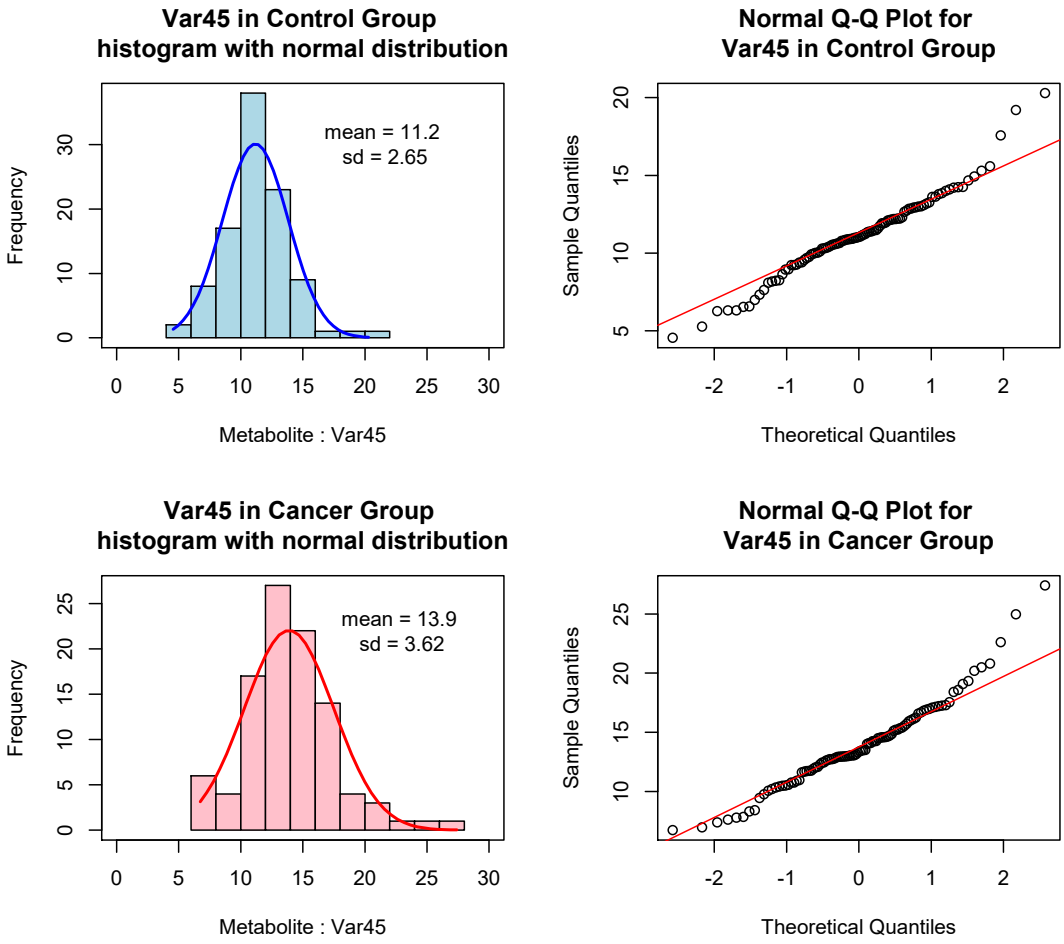

Figure 2: Assessing normality assumption for Var45 by group.

4. Build a classifier from the k ranked variables. The classification method is either Random Forest or SVM with linear kernel (linear boundaries and soft margin).

5. Check the performance of the obtained classifier. The performance is measured by misclassification rate, sensitivity, specificity, AUC in the ROC curve.

6. Increment k and return to step 1 until k=110 or until there is no more important increase in the classifier's performance.

Two different class prediction methods are used : Random Forest and Support Vector Classifier (linear boundaries and soft margin).
Random Forest and SVM as supervised machine learning statistical tools are detailed in [4].

The consistency of the two methods will be investigated as whether they provide the same number of variables.

3.5 Data splitting rule

As already touched upon in the introduction, we do not want to overfit the classifier. Overfitting would mean we can accurately predict the class of the training set but poorly predict the class of future independent data not yet observed. Besides, we want to capture as much information as possible from the training set. To prevent overfitting and yet capture as much information from the training set, cross-validation will be used to build the classifier and test it on data not used for learning the classifier. The number of variables selected in the variable selection procedure depends on the performance of the classifier. So the number of variables selected is nested in the cross-validation scheme.

5-FOLD CROSS VALIDATION, 10 REPEATS AND STRATIFIED SAMPLING

The cross validation is 5-fold. The samples (rows) in the available dataset `Metabolite.Train` will be split into 5 parts (5-fold CV) of equal size. 1 part is used as validation set, the 4 other parts are used as training set. The procedure is repeated 10 times (10 iterations). The sampling for the sample selection is stratified, i.e. the proportion of cancer and controls samples is the same in the training set and in the validation set. In practice, this means that, at each iteration, out of the 200 patients (=samples=rows), 40 will be isolated for the test, the other 160 patients will be used to train a classifier (supervised learning). The variable selection will rank the 110 variables of the feature space in increasing order of the p-values for the Welch t.test comparing the 80 cancer patients (half of 160 of the training set) to the 80 control patients (the other half of the training set) selected at this iteration. The first k ranked variables are used to learn and fit the classifier. The performance (sensitivity, ..., AUC) of the fitted classifier is evaluated on the 40 samples of the validating set. A confusion matrix is built. The whole procedure is repeated 10 times (10 iterations). All the 10 confusion matrices are summed and/or averaged and confidence intervals are calculated for the misclassification, sensitivity, specificity, AUC of ROC curves.

3.6 Variable selection

The variable selection is nested in the 5-fold cross-validation with 10 iterations. As explained in the section rationale behind the variable selection, the averaged performances of the classifier are stored for each set of k variables (k=2, k=3, k=4,..., k=110).

3.7 Variable selection results

The graphs of the measured performances of the classifier as a function of the number of selected variables is shown on figure 3 for the Random Forest Classifier and on figures 4 for the support vector classifier.

3.7.1 *Random Forest classifier*

It can be seen on the upper left panel of figure 3 that the misclassification rate of the Random Forest classifier sharply decreases when the number of variables increases from 2 to 16, then remains rather stable. Similarly, the sensitivity, specificity and AUC sharply increase when we move from 2 variables to 16, then remain rather stable.

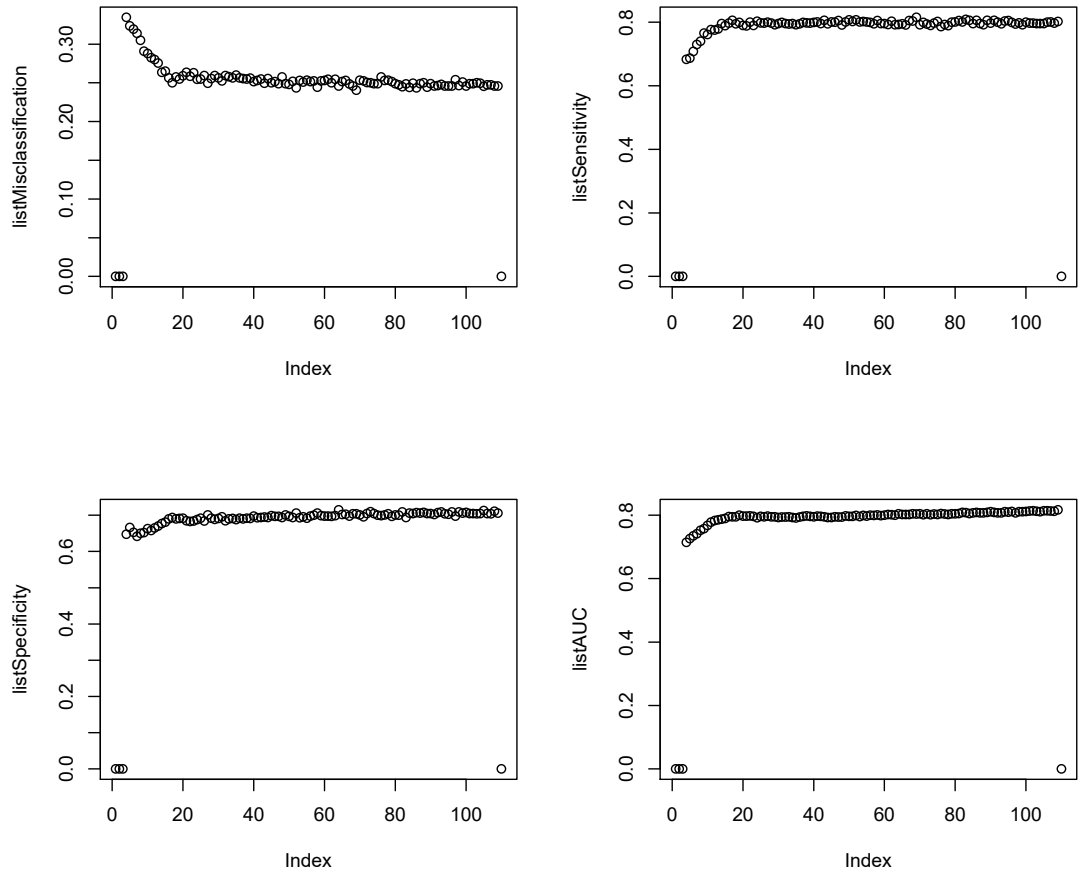

Figure 3: RF Classifier performance versus number of selected variables.

3.7.2 *Support Vector Classifier*

A support vector classifier is also used to check if we reach similar results as the ones from the previous section. Due to the more computer intensive method linked to the SVM cost tuning, the graphs are only constructed from k=2 to k=30. It can be seen however that the performances of the classifier are stabilized at k=16 variables as well.

3.7.3 *Shortlist of 16 most informative metabolite feature variables*

The variable selection variable procedures with RF and with SVM both motivate that 16 variables capture enough information to train a learning machine providing enough classification power. Both classifiers provide the same 16 best ranked shortlist of metabolite feature variables which is given hereafter :

```
var : 45  46  48  49  50  11  73  31  47  72  44  91  30  37  25  38
freq: 10  10  10  10  10   8   8   7   7   7   3   3   2   2   1   1
```

The second row of the previous shortlist shows how many times the considered variable was picked up as part of the ranked list in the iterated procedure of the cross-validation. The smaller the p-value,

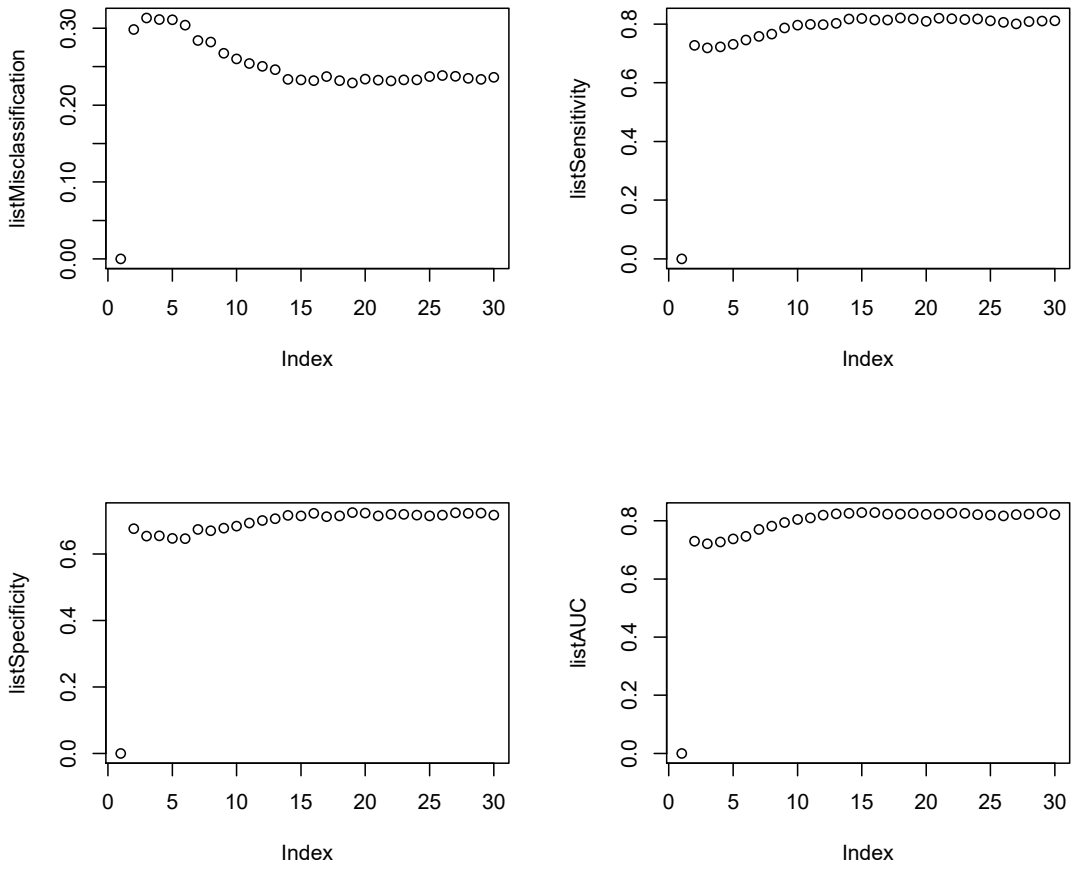

Figure 4: SVM Classifier performance versus number of selected variables.

the higher the variable is ranked and the more often it is picked up among the informative features for the classifier.

From now on, we will only use these 16 variables to build a classifier (any classifier).

3.8 Comparison of classifiers performance

The performances of the following classifiers are compared. All classifiers used the shortlist of the 16 metabolite variables that were identified previously.

- LDA Linear Discriminant Analysis
- FDA Fisher Discriminant Analysis (4 components)
- Random Forest Classifier
- SVM Support Vector Classifier with linear boundaries, soft margins

The 5-fold cross validation was carried out with the selected list of the 16 variables to train the classifier and produced the performance estimation tabulated in table 1.

Table 1: Classifiers compared performance

	missclassification	sensitivity	specificity	AUC
LDA	0.261	0.797	0.681	0.800
FDA	0.2610	0.797	0.681	0.800
rf	0.2625	0.788	0.687	0.794
svm	0.2445	0.807	0.704	0.819

The performances of the 4 classifiers are compared graphically on figures 5, 6, 7 and 8 showing the boxplots of the performance metrics (misclassification rate, sensitivity, specificity and AUC respectively). All the 4 classifiers show similar performance metrics.

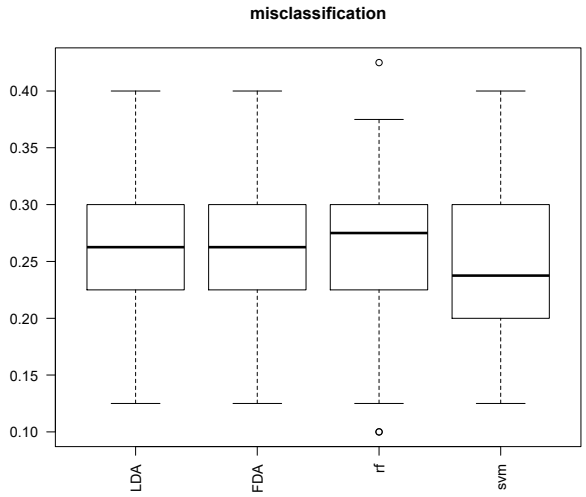

Figure 5: Comparison of the misclassification rates for the 4 classifiers.

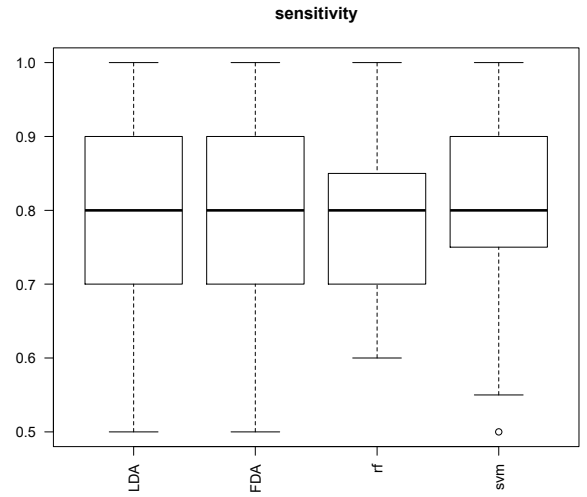

Figure 6: Comparison of the sensitivity for the 4 classifiers.

Our primarily fixed measure of performance is the AUC. Based on this AUC criterion, the support vector classifier (SVM with linear kernel, soft margin and linear boundaries) is selected as the best performing classification tool based on a metabolite signature. Note also that the misclassification is the smallest with the SVM classifier.

3.9 SVM Classifier performance

We resume hereafter the performances of the SVM classifier. The performance metrics were obtained from 50 learning sets and 50 validation sets (5 fold CV x 10 iterations from the 200 samples available in the `Metabolite.Train` dataset). It means that 1 validation provided 40 samples for classification and cross-checking. On the whole, with the 50 validation sets, a total of 2000 samples were used in classification and cross-checking.

3.9.1 *Confusion matrix*

The confusion matrix is tabulated in table 2.

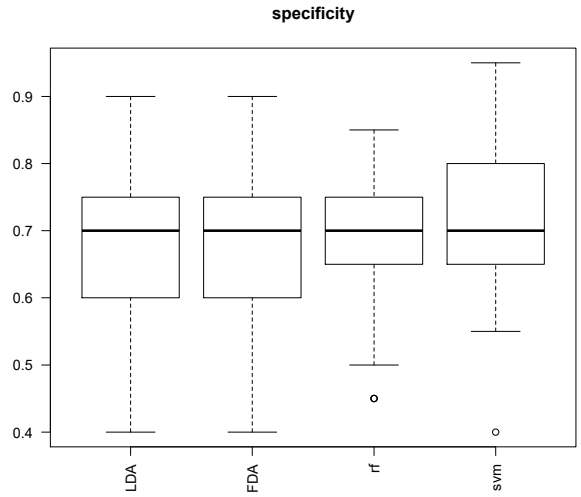

Figure 7: Comparison of the specificity for the 4 classifiers.

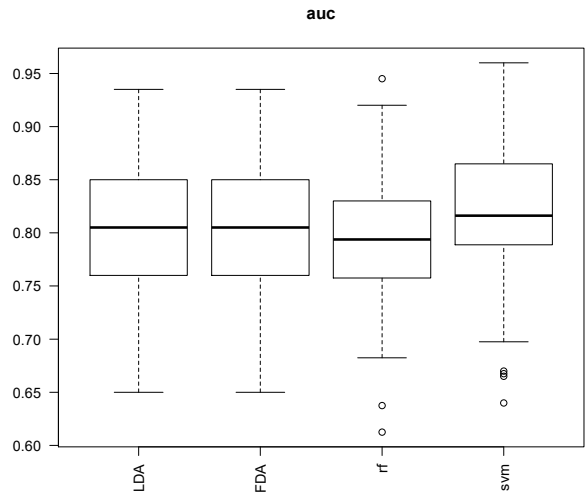

Figure 8: Comparison of the AUC ROC curve for the 4 classifiers.

Table 2: Confusion matrix of the metabolic signature based SVM classifier.

		Predicted	
		0	1
True	0	704	296
	1	193	807

The performances are :

1. Misclassification rate : 0.244
2. Sensitivity : 0.704
3. Specificity : 0.807
4. Positive Predictive value (PPV): 0.732
5. Negative Predictive value (NPV) : 0.785

3.9.2 *ROC Curve*

The ROC curve of the SVM classifier is displayed on figure 9.

The LLP risk model based on classical risk factors implementing logistic regression has an AUC of 0.71. The LLP study ROC curve is presented in Figure 1 of the article by Cassidy et al. [2].

Although the LLP study [2] addresses a next 5 years prognostic and the SVM classifier we developed addresses a current diagnostic purpose, it is encouraging that the AUC value of 0.81 for the SVM classifier is higher than the one based on classical clinical variables.

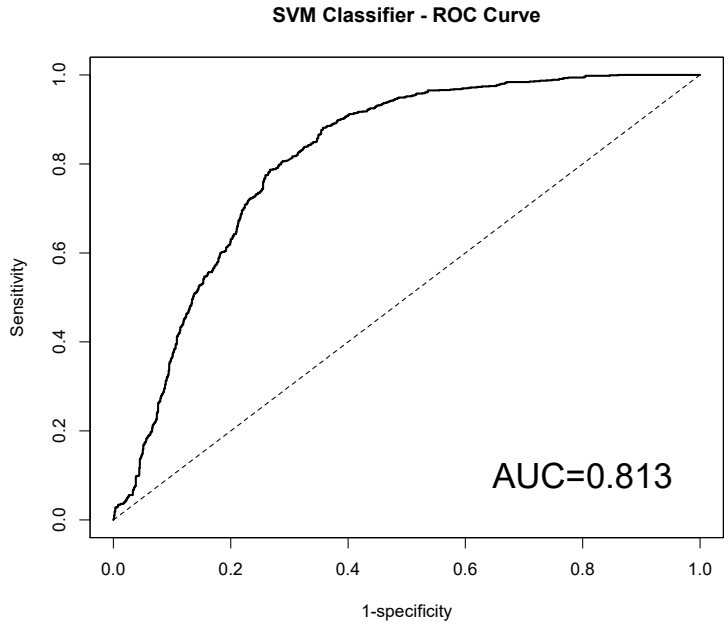

Figure 9: Receiver operating curve for the SVM classifier.

3.10 Test the SVM classifier on independent data

To get confidence on the robustness of the metabolic signature classifier, it should be tested on at least one completely independent dataset. The attached R script code is made available for further testing.

We assumed the second file that was provided, named `Metabolite.test` and containing more than 330 samples from different patients, was an independent dataset. We chose randomly 180 samples from this dataset and used our SVM classifier as is (without new fitting) to classify these 180 samples based on the very same 16 metabolic variables we previously selected. The cross-check of our predictions provide the following ROC curve (red) compared to the ROC curve computed from the previous training set, see figure 10. The red ROC curve shows that the classifier performs very good on independent data.

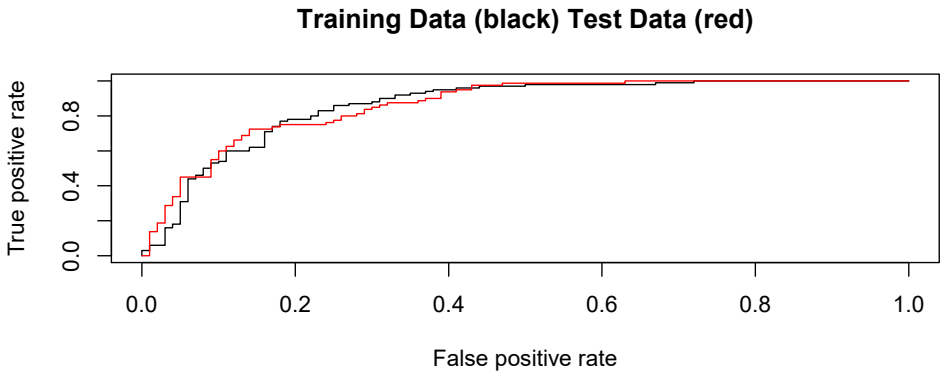

Figure 10: ROC curve of training set and independent test set.

4 PART B

Risk prediction models for lung cancer for a given individual, based on standard clinical variables are documented in the literature. In the Liverpool Lung Project (LLP)[2], significant classical variables were fitted into a multiple logistic regression model. Standard classical predictors are variables like gender, age, smoking duration, pack-years, prior diagnosis of malignant tumor, family history of lung cancer, exposure to asbestos. In this part B of the project, such a multiple logistic regression is mirrored from the LLP study and applied to the `Clinical.Train` received data set to build a reference standard model based on most significant known risk factors. The aim of part B is to test if the new prognostic metabolite signature developed in part A can provide increased predictive accuracy to the existing model in addition to the known risk factors.

WHAT SHOULD BE INCORPORATED IN THE EXISTING MODEL BASED ON STANDARD RISK FACTORS ?

The idea is to incorporate in the classical logistic regression model one single extra predictor, i.e. the SVM metabolic signature as a categorical variable (positive signature or negative signature for cancer). The positive or negative signature is simply the algebraic sign of the classifier according to whether it is on one or the other side of the SVM hyperplane cutoff. This hyperplane was built from the previously selected 16 metabolites variables.
In the next section we start by building the classical logistic regression model based on the standard clinical variables.

4.1 Existing multiple logistic regression

Existing Multiple Logistic Regression risk model :

We use as standard clinical variables for the risk factors : age of the patient (continuous), smoking status (ordered categorical : Never, Smoking stopped more than 6 month ago, Smoking Active) and pack-years (continuous).

The model estimates the log odds of having a lung cancer for the different classical clinical risk factors.

$$Y_i \sim \text{Bernoulli}(\pi_i) \tag{1}$$

$$\text{logit}(\pi_i) = \beta_0 + \sum_{j=1}^{J} \beta_{ij} \cdot X_{ij} \tag{2}$$

The results for the fitted logistic regression classical model from the given dataset `Clinical.train` are :

```
glm(formula = Group ~ Age + Smoke + Packyears, family = binomial,
    data = Clinical.Train)
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.94578  -0.84008  -0.07517   0.88853   2.58904
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -4.613441   1.354885  -3.405 0.000662 ***
Age                0.036043   0.017773   2.028 0.042561 *
SmokeStopped > 6m  1.203034   0.594161   2.025 0.042892 *
SmokeActive        1.838525   0.633110   2.904 0.003685 **
Packyears          0.039398   0.009634   4.089 4.33e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 277.26  on 199  degrees of freedom
Residual deviance: 208.60  on 195  degrees of freedom
AIC: 218.6
Number of Fisher Scoring iterations: 4
```

All the identified risk factors are significant.

4.2 Comparing the existing logistic regression with the SVM classifier alone

A ROC curve from this classical standard risk factors model is displayed (in black) on figure 11 and is compared to the the ROC curve (in red) from the SVM classifier using the metabolite signature only. For a false positive discovery rate controlled at values less than 0.2, the sensitivity of the classical model is better than the SVM classifier based on the metabolic signature only. For higher tolerated values of the false positive rate (or at lower tolerated values of the specificity), the sensitivity of the metabolic signature classifier is better.

4.3 Enriched multiple logistic regression with metabolite signature

Enriched Multiple Logistic Regression risk model with metabolite signature :

The full enriched model writes

$$\begin{aligned} \text{logit}(\pi_i) &= \beta_0 + \sum_{j=1}^{J} \beta_{ij} \cdot X_{ij} + \beta_Z \cdot Z_i \\ Z_i &= 0 \text{ if (+) signature (cancer) or } 1 \text{ if (-) (control)} \end{aligned}$$

where Z_i stands for the metabolic signature of the SVM classifier based on 16 metabolic variables in the dataset for patient i.

We have incorporated the metabolic signature as an extra categorical predictor in the lung cancer risk model.
Does the metabolic signature make a difference ?
Is $\beta_Z \neq 0$ significantly ?

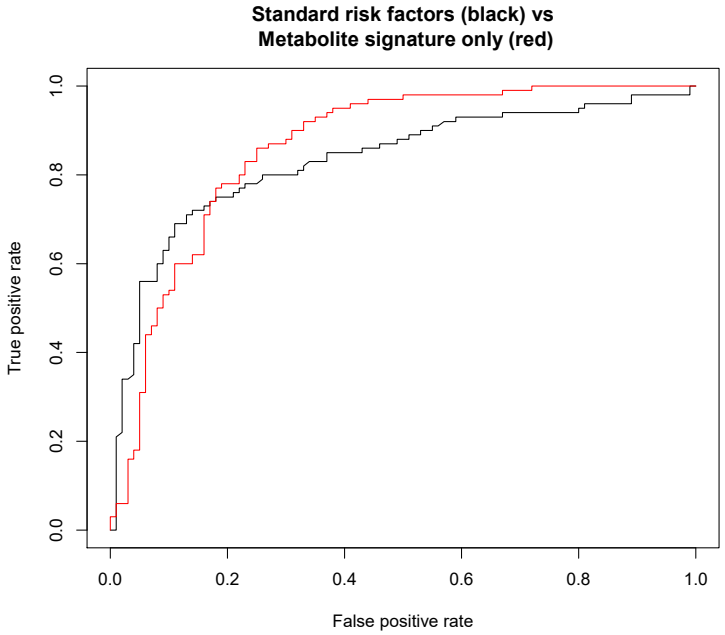

Figure 11: ROC curve for standard versus metabolic signature separately.

The enriched model is fitted on the given Clinical.train dataset :

```
glm(formula = Group ~ Age + Smoke + Packyears + metabo.signature,
    family = binomial, data = Clinical.Train)
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.91540  -0.60542  -0.05047   0.55882   2.95099
Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)             -3.04949    1.66816  -1.828 0.067542 .
Age                      0.03698    0.02166   1.707 0.087803 .
SmokeStopped > 6m        1.31806    0.69387   1.900 0.057488 .
SmokeActive              1.63590    0.74455   2.197 0.028009 *
Packyears                0.03809    0.01086   3.508 0.000452 ***
metabo.signatureControl -2.62293    0.42719  -6.140 8.25e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 277.26  on 199  degrees of freedom
Residual deviance: 161.41  on 194  degrees of freedom
AIC: 173.41
```

We see that β_Z is indeed different from zero with the lowest p-value of all the predictors. Here, β_Z is negative as it is compared to the signature for the cancer reference ($Z = 0$ for cancer) in the implemented model. The interpretation is that the log odds for having lung cancer is reduced by 2.62293 when a patient shows a negative metabolic signature as compared to a positive signature. Stated otherwise, it means that the odd ratio for lung cancer is $0.0726 = \exp(-2.62293)$ when you have a negative metabolic signature compared to a positive metabolic signature. The chance that a patient has a lung cancer are 13.77 times lower if the patient has a negative metabolic signature when compared to a patient with a

positive metabolic signature, all the other clinical variable kept the same.
Incorporation of the metabolic signature in the full model results in the lost of significance for the age predictor.

The Chi-square version of ANOVA comparing the classic reduced model with the full enriched model shows that the enriched model fits better with the metabolic signature and adds significantly beyond the classical standard clinical variable predictors.

```
> anova(fit.classic.reduced, fit.full, test="Chisq")
Analysis of Deviance Table
Model 1: Group ~ Age + Smoke + Packyears
Model 2: Group ~ Age + Smoke + Packyears + metabo.signature
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       195     208.60
2       194     161.41  1   47.197 6.419e-12 ***
---
```

HAS THE PREDICTIVE ACCURACY OF THE MODEL INCREASED ?

The misclassification rate has decreased from 0.215 to 0.18 with the enriched model.

Most importantly, the AUC of the ROC curve has increased as can be seen from the red curve of the full model displayed on figure 12 as compared to the black curve for the reduced model.

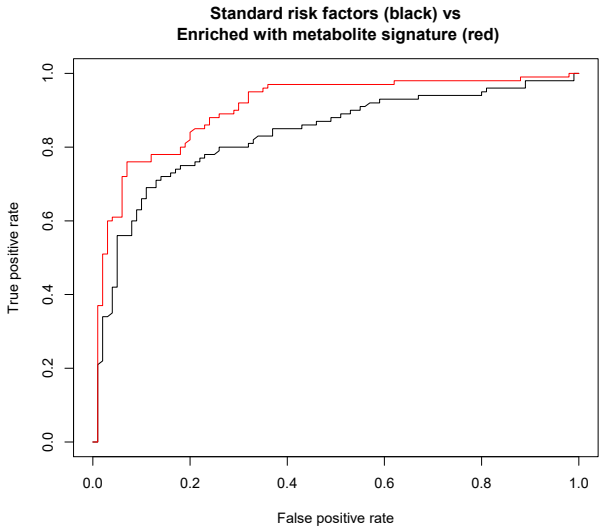

Figure 12: ROC curve for standard versus enriched cancer risk model with metabolic signature.

5 DISCUSSION AND CONCLUSION

A metabolic signature for lung cancer diagnostic has been devised successfully from a sufficiently minimal number of informative metabolites in the patient blood samples.

Sixteen metabolic variables appear to capture enough information for building a binary classifier which shows similar predictive power as the existing lung cancer prognostic model based on standard clinical predictors like gender, age, smoking status and pack-years.

The feature space variables were provided blindly to our analysis, i.e. the exact chemical names of the metabolites were not provided in the dataset. Instead, we only had anonymous metabolite names ID (Var1, ..., Var110). The identification of these 16 metabolites should be investigated further and should be checked for their biochemical role, their direct or indirect association to lung cancer in particular. The named list of these 16 metabolites contributing to the lung cancer signature is :

Metabolite Variable ID's: `45 46 48 49 50 11 73 31 47 72 44 91 30 37 25 38`.

The concentration levels of these metabolites are informative enough to classify a patient between two groups : lung cancer or control groups.

The built classifier is of the support vector classifier type with linear boundaries and soft margins (SVM with linear kernel and a cost parameter = 0.05). The classifier is built only on the 16 indicated metabolite variables.

As a stand alone metabolite based classifier, it shows good classification metrics on averaged test sets results of cross-validation: misclassification rate = 0.24, sensitivity = 0.70, specificity= 0.81, AUC = 0.813, very similar or better than existing risk models based on standard clinical variables. The ROC curve obtained from completely independent data is good (see figure 10).

When incorporated to existing risk model in addition to known risk factors, the enriched model shows a better accuracy (0.82 instead of 0.785) and a higher positive predictive value and most importantly a higher AUC in the receiver operating (ROC) curve (see figure 12).

If the 16 metabolites levels in patients blood samples are easily and cheaply obtained, their direct use as input for the built svm classifier would be beneficial in the diagnostic toolbox. The gain in predictive power is positive. The misclassification has been shown to decrease by more than 3 points (from 0.215 to 0.18).

In conclusion the two research questions addressed are answered :

PART A : a metabolite signature exists for lung cancer from which a classifier can be built to distinguish future lung cancer patients and controls with high predicting accuracy.

PART B : This identified metabolic signature can lead to a gain in prediction accuracy when used in addition to standard known risk factors.

REFERENCES

[1] Halama, A, et al. (2015) *Metabolic signatures differentiate ovarian from colon cancer cell lines*, Journal of Translational Medecine, 13:223.

[2] Cassidy, A. et al. (2008) *The LLP risk model: an individual risk prediction model for lung cancer*. British Journal of Cancer, 98:270-276.

[3] Slawski, M., Boulesteix, A.L. (2016) *CMA package vignette*, Bioconductor Package vignette.

[4] James,G., Witten, D., Hastie, T., Tibshirani, R. (2013) *An Introduction to Statistical Learning with applications in R*, Springer.