Supplement for Manuscript

Profiling Serum Biomarkers in Patients with COPD: Associations with Clinical Parameters

Victor Pinto-Plata [1], John Toso[2], Kwan Lee[2], Daniel Park[2], John Bilello[2],
Hana Mullerova[2], Mary M. De Souza[2], Rupert Vessey[2], Bartolome Celli[1].

[1] Pulmonary, Critical Care and Sleep Division. Caritas St Elizabeth’s Medical Center.
Tufts University. Boston, MA.
[2] Discovery Medicine, High Throughput Biology and Biomedical Data Sciences,
GlaxoSmithKline R&D.

Correspondence: Bartolome R. Celli, M.D.
Caritas St. Elizabeth’s Medical Center
736 Cambridge Street. Boston. MA.02135

Tel: (617) -789-2554
Fax: (617) 562-7756
Email: bcelli@copdnet.org

Statistical Appendix

Overview of Statistical Methods Used for Biomarker Selection

There is no standard or agreed upon statistical methods used for ranking and selection of biomarkers related to a disease or a drug. Many different methods are used and they can potentially yield different rankings and selections. There are two different types of methods in general – one, an univariate method and the other is a multivariate method.

Univariate Method

Univariate methods consider one biomarker at a time without considering association with others. The most frequently used simple t-test belongs to this category. For a disease marker selection, for example, t-test compares two group means between the normal and diseased samples. This is equivalent to the use of Pearson’s correlation coefficient between the biomarker expression (“x”) and an indicator variable (“y”) coded as 0’s for normal and 1’s for diseased samples. The same result can be obtained by considering the simple linear regression between the “x” and “y” and test the regression coefficients for significance (i.e. whether the regression coefficient is significantly different from zero). The statistic used in this model is another t statistic formed by the ratio of the estimated regression coefficient to its standard error. The actual ranking of biomarkers can be done by corresponding p-values and some cut point can be used for the final selection. The technique used in this study also adjust the p-values for the multiplicity of the testing using the concept of false discovery rate (FDR). The idea of FDR is to control the average proportion of false positives in the selected list of biomarkers. In this simple linear regression setting, the regression coefficient is proportional to the Pearson’s correlation mentioned above and essentially tests the same thing – the association between a biomarker and the disease ignoring the association with other biomarkers. In other words, Univariate analysis is based on the marginal correlation between a marker and the clinical endpoint.

Multivariate

Multivariate methods in general consider all the biomarkers in the experiment together in a single model and include many different regression and classification method. These tools are called supervised learners and also called a wrapper based approach in the machine learning community. The types of regression (or classification) models can be either linear or nonlinear. Ordinary least squares (OLS) and logistic regressions are frequently used linear models. Decision trees and neural networks are examples of nonlinear models. Let's consider as an example the selection of disease markers using a multiple linear regression model. The regression coefficient in this case essentially measures the partial correlation between a biomarker and the disease adjusted for all the other biomarkers. This is the reason why multivariate models are preferred since the partial correlation takes into account the association with other markers. We know that all the biomarkers are related and their associations approximate the biological network in the disease pathways. It is well known that partial correlation is generally better measure of direct association between the two markers than the marginal correlation in the association network modeling. Generally nonzero marginal correlation can mean either direct association or effects of other indirect variables.

Ordinary least squares or logistic regression, however, can have a serious problem especially when the data is of high dimensional and as a result the biomarker expressions are highly correlated. This is a well known multi-collinearity problem for linear regression and OLS' regression coefficients can be very misleading since their standard errors are so large that sometimes they even have wrong signs in their estimates. The same is true for logistic regression for classification and we can not rely on their coefficients and p-values for the ranking and selection of the biomarkers. These models are unstable under multi-collinearity and are of high variance structure.

A shrinkage method of estimation such as principal component regression (PCR), partial least squares (PLS) and ridge regression (RR) can bypass this multi-collinearity problem by regularization of the estimation process (1). They may introduce a small bias but can reduce the variance of the estimated coefficients appreciably and hence are more stable. We have used partial least squares (PLS) regression and its discriminant analysis (PLS-DA) to deal with high dimensional biomarker selection and found them very competitive with other methods of shrinkage estimation. The PLS-DA is simply PLS applied to a categorical response variable. For a binary response, it is typically coded as 0 or 1 but other scaling of the response does not alter the ranking of the regression coefficients and hence interpretation of the result remains the same. The software package SIMCA (2) implemented PLS and PLS-DA in a very user friendly manner with an excellent graphical user interface. We found the package very useful for high dimensional data analysis in general. There is a recent study comparing different shrinkage methods and currently active research is being done to improve the accuracy and flexibility of ridge regression to high dimensional biomarker selection (3).

Nonlinear models such as decision trees and neural networks can improve the accuracy of their predictions by adopting nonlinearity but are of high variance structure and can be unstable as well. Decision tree algorithms are unstable at times since variable selection is done in a stepwise manner and is of discrete nature (greedy algorithm).

This can be true for any stepwise variable (biomarker) selection algorithms.

Neural networks use many parameters in the estimation process (in many cases overparametrized) and a trade off can be again instability of the model. One interesting computer intensive method called Random Forest (4) is based on the bootstrap aggregation of the many (> 500 for example) decision tree models and is a very promising tool for high dimensional biomarker selection. Our limited experience showed that the PLS coefficients gave similar rankings of the biomarkers to the Random Forest in many cases of high dimensional data.

Model Validation.

Validation of a prediction model can be done externally on a separate test data or internally using a cross validation. Typically cross validation is applied to come up with a best performing model e.g. to minimize a performance measure such as predicted residual sums of squares for a regression model. Once a cross validated performance is obtained, the statistical significance of the performance measure is obtained by a permutation test. The permutation test in this case is to randomly permute the labels of the response part of the data to assess the significance of the actual performance measure against those obtained from random permutations of labels. If none of the models from the 100 different random permutations of the labels of the response showed better performance than the model from original data then we can conclude the model is significant at P less than 0.01. Our approach of model validation was based on combining the ideas of the cross validation and the permutation test.

Data analysis strategy used in the stusy.

The ranking and selection of biomarkers is not a pure statistical exercise but should be a collaborative effort between statisticians and scientists. We could obtain a ranking of biomarkers by a univariate statistical test and select a few in the top of a list or use a cut point based on p-values. In many cases, people also adjust the p-values for the multiplicity of the testing but recently the concept of false discovery rate (FDR) became popular for the decision making, which professor Efron calls one of the genuinely useful new ideas (5). The idea of FDR is to control the average proportion of false positives in the selected list of biomarkers. However selecting biomarkers solely based on a univariate ranking may ignore the associations among the biomarkers and may end up with markers that have all similar functions. In order to select diverse set of markers for COPD our approach of selecting a panel of predictive biomarkers for COPD was to cluster the biomarkers into a few clusters/ groups (say 30) first and then evaluate the predictiveness of each cluster for COPD. Then we would select a few representative biomarkers from each group of the predictive clusters. The particular clustering tool we have used was Variable Clustering (VARCLUS) procedure in SAS (6). The VARCLUS procedure attempts to divide a set of variables into non-overlapping clusters in such a way that each cluster can be interpreted as essentially unidimensional. Underlying computation of VARCLUS is very similar to a factor analysis and roughly a factor is equivalent to a cluster in VARCLUS. The predictiveness of each cluster was then determined by computing partial correlation of each cluster centroid with COPD given all the other markers using partial least squares discriminant analysis (PLS-DA). Each of the regression coefficients from the PLS-DA is essentially equivalent to the partial correlation between the cluster centroid and the response. In this case the response is coded as a binary indicator variable and as long as the indicator variable has two distinct values such as 0 for control or 1 for COPD patient it does not matter what the scale is. Hence a regression coefficient essentially measures the partial correlation between an average biomarker in a cluster with COPD that is adjusted for all other cluster averages.

Description of the analytes included in the micro-arrays.

The total number of analytes included on arrays 1-5. Note data from CRP on array 5 was not useable due to CRP levels well above the upper detection limit of the assay which resulted in a ‘Hook effect”.

Array 1 analytes

	Analyte	Name
1	ANG	Angiogenin
2	BLC (BCA-1)	B-lymphocyte chemoattractant
3	EGF	Epidermal growth factor
4	ENA-78	Epithelial cell-derived neutrophil-activating peptide
5	Eot	Eotaxin
6	Eot-2	Eotaxin-2
7	Fas	Fas (CD95)
8	FGF-7	Fibroblast growth factor-7
9	FGF-9	Fibroblast growth factor-9
10	GDNF	Glial cell line derived neurotrophic factor
11	GM-CSF	Granulocyte macrophage colony stimulating factor
12	IL-1ra	Interleukin 1 receptor antagonist
13	IL-2 sRα	Interleukin 2 soluble receptor alpha
14	IL-3	Interleukin 3
15	IL-4	Interleukin 4
16	IL-5	Interleukin 5
17	IL-6	Interleukin 6
18	IL-7	Interleukin 7
19	IL-8	Interleukin 8
20	IL-13	Interleukin 13
21	IL-15	Interleukin 15
22	MCP-2	Monocyte chemotactic protein 2
23	MCP-3	Monocyte chemotactic protein 3
24	MIP-1α	Macrophage inflammatory protein 1 alpha
25	MPIF	Myeloid progenitor inhibitory factor 1
26	OSM	Oncostatin M
27	PlGF	Placental growth factor

Array 2 analytes

	Analyte	Name
1	AR	Amphiregulin
2	BDNF	Brain-derived neurotrophic factor
3	Flt-3 Lig	fms-like tyrosine kinase-3 ligand
4	GCP-2	Granulocyte chemotactic protein 2
5	HCC4 (NCC4)	Hemofiltrate CC chemokine 4
6	I-309	I-309
7	IL-1α	Interleukin 1 alpha
8	IL-1β	Interleukin 1 beta
9	IL-2	Interleukin 2
10	IL-17	Interleukin 17
11	MCP-1	Monocyte chemotactic protein 1
12	M-CSF	Macrophage colony stimulating factor
13	MIG	Monokine induced by interferon gamma
14	MIP-1β	Macrophage inflammatory protein 1 beta
15	MIP-1δ	Macrophage inflammatory protein 1 delta
16	NT-3	Neurotrophin 3
17	NT-4	Neurotrophin 4
18	PARC	Pulmonary and activation-regulated chemokine
19	RANTES	Regulated upon activation, normal T expressed and presumably secreted
20	SCF	Stem cell factor
21	sgp130	Soluble glycoprotein 130
22	TARC	Thymus and activation regulated chemokine
23	TNF-RI	Tumor necrosis factor receptor I
24	TNF-α	Tumor necrosis factor alpha
25	TNF-β	Tumor necrosis factor beta
26	VEGF	Vascular endothelial growth factor

Array 3 analytes

	Analyte	Name
1	BTC	Betacellulin
2	DR6	Death receptor 6
3	Fas Lig	Fas ligand
4	FGF acid (FGF-1)	Fibroblast growth factor acidic
5	Fractalkine	Fractalkine
6	GRO-β	Growth related oncogene beta
7	HCC-1	Hemofiltrate CC chemokine 1
8	HGF	Hepatocyte growth factor
9	HVEM	Herpes virus entry mediator
10	ICAM-3 (CD50)	Intercellular adhesion molecule 3
11	IGFBP-2	Insulin-like growth factor binding protein 2
12	IL-2 Rγ	Interleukin 2 receptor gamma
13	IL-5 Rα (CD125)	Interleukin 5 receptor alpha
14	IL-9	Interleukin 9
15	Leptin/OB	Leptin
16	L-Selectin (CD62L)	Leukocyte selectin
17	MCP-4	Monocyte chemotactic protein 4
18	MIP-3β	Macrophage inflammatory protein 3 beta
19	MMP-7 (total)	Matrix metalloproteinase 7
20	MMP-9	Matrix metalloproteinase 9
21	PECAM-1 (CD31)	Platelet endothelial cell adhesion molecule-1
22	RANK	Receptor activator of NF-kappa-B
23	SCF R	Stem cell factor receptor
24	TIMP-1	Tissue inhibitors of metalloproteinases 1
25	TRAIL R4	TNF-related apoptosis-inducing ligand receptor 4
26	VEGF-R2 (Flk-1/KDR)	Vascular endothelial growth factor receptor 2
27	ST2	Interleukin 1 receptor 4

Array 4 analytes

	Analyte	Name
1	ALCAM	Activated leukocyte cell adhesion molecule
2	β-NGF	beta-nerve growth factor
3	CD27	CD27
4	CTACK	Cutaneous T-cell attracting chemokine
5	CD30	CD30
6	Eot-3	Eotaxin-3
7	FGF-2	Fibroblast growth factor-2 (FGF-basic)
8	FGF-4	Fibroblast growth factor-4
9	Follistatin	Follistatin
10	GRO-γ	Growth related oncogene gamma
11	ICAM-1	Intercellular adhesion molecule 1
12	IFN-γ	Interferon gamma
13	IFN-ω	Interferon omega
14	IGF-1R	Insulin-like growth factor I receptor
15	IGFBP-1	Insulin-like growth factor binding protein 1
16	IGFBP-3	Insulin-like growth factor binding protein 3
17	IGFBP-4	Insulin-like growth factor binding protein 4
18	IGF-II	Insulin-like growth factor II
19	IL-1 sR1	Interleukin 1 soluble receptor I
20	IL-1 sRII	Interleukin 1 soluble receptor II
21	IL-10 Rβ	Interleukin 10 receptor beta
22	IL-16	Interleukin 16
23	IL-2 Rβ	Interleukin 2 receptor beta
24	I-TAC	Interferon gamma-inducible T cell alpha chemoattractant
25	Lptn	Lymphotactin
26	LT βR	lymphotoxin-beta receptor
27	M-CSF R	Macrophage colony stimulating factor receptor
28	MIP-3α	Macrophage inflammatory protein 3 alpha
29	MMP-10	Matrix metalloproteinase 10
30	PDGF Rα	Platelet-derived growth factor receptor alpha
31	PF4	Stromal cell-derived factor beta
32	sVAP-1	Soluble Vascular Adhesion Protein-1
33	TGF-α	Transforming growth factor alpha
34	TIMP-2	Tissue inhibitors of metalloproteinases 2
35	TRAIL R1	TNF-related apoptosis-inducing ligand receptor 1
36	VE-cadherin	Vascular Endothelial Cadherin
37	VEGF-D	Vascular endothelial growth factor-D

Array 5 analytes

	Analyte	Name
1	4-1BB (CD137)	4-1BB
2	ACE-2	Angiotensin I converting enzyme-2
3	AFP	Alpha fetoprotein
4	AgRP	Agouti-related protein
5	CD141	Thrombomodulin/CD141
6	CD40	CD40
7	CNTF Rα	Ciliary neurotrophic factor receptor alpha
8	CRP	C-reactive protein
9	D-Dimer	D-Dimer
10	E-Selectin	E-selectin
11	HCG	Human chorionic gonadotrophin
12	IGFBP-6	Insulin-like Growth Factor Binding Protein 6
13	IL-12 (p40)	Interleukin 12 p40
14	IL-18	Interleukin 18
15	LIF Rα (gp190)	Leukemia inhibitory factor souble receptor alpha
16	MIF	Macrophage migration inhibitory factor
17	MMP-8 (total)	Matrix Metalloproteinase-8
18	NAP-2	Neutrophil Activating Peptide 2
19	Neutrophil elastase	Neutrophil elastase
20	PAI-II	Plasminogen activator inhibitor-II
21	Prolactin	Prolactin
22	Protein C	Human Protein C
23	Protein S	Human Protein S
24	P-Selectin	P-Selectin
25	TSH	Thyroid stimulating hormone

References

1. Hatie, T., Tibshirani, R. & Friedman, J. (2001), The Elements of Statistical Learning; Data mining, Inference and Prediction, Springer Verlag, New York.

2. SIMCA-P, version 10.5, Jan. 2004, Copyright ©, Umetrics 1993 – 2003, http://www.umetrics.com/

3. Hui Zou and Trevor Hastie. Regularization and Variable Selection via the Elastic Net (pdf). JRSSB (2005) 67(2) 301-320

4 Breiman, L. (1996), "Bagging predictors", Machine Learning 24, 123-140

5. Bradley Efron (2005), Bayesians, Frequentists, and Scientists, Journal of American Statistical Association, vol. 100, no.469, 1-5.

6. SAS/STAT User's Guide, Version 8, Chapter 68. The VARCLUS Procedure, pp 3593-3620.