CHRONIC OBSTRUCTIVE PULMONARY DISEASE

Profiling serum biomarkers in patients with COPD: associations with clinical parameters

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

Thorax 2007;**62**:595–601. doi: 10.1136/thx.2006.064428

See end of article for authors' affiliations

Correspondence to: Dr Bartolome R Celli, Caritas St Elizabeth's Medical Center, 736 Cambridge Street, Boston, Massachusetts 02135, USA; bcelli@copdnet.org

Received 4 May 2006
Accepted 10 January 2007
Published Online First 13 March 2007

Background: Chronic obstructive pulmonary disease (COPD) is an inflammatory lung disease associated with significant systemic consequences. Recognition of the systemic manifestations has stimulated interest in identifying circulating biomarkers in these patients. A systematic analysis was undertaken of multiple protein analytes in the serum of well characterised patients with COPD and matched controls using novel protein microarray platform (PMP) technology.

Methods: Forty-eight patients (65% men) with COPD (forced expiratory volume in 1 s <55%) and 48 matched controls were studied. Anthropometric parameters, pulmonary function tests, 6-minute walk distance, the BODE index and the number of exacerbations were measured and the association of these outcomes with the baseline levels of 143 serum biomarkers measured by PMP was explored.

Results: Thirty biomarker clusters were identified and ranked by computing the predictive value of each cluster for COPD (partial least squares discriminant analysis). From the 19 best predictive clusters, 2–3 biomarkers were selected based on their pathophysiological profile (chemoattractants, inflammation, tissue destruction and repair) and the statistical significance of their relationship with clinically important end points was tested. The selected panel of 24 biomarkers correlated ($p<0.01$) with forced expiratory volume in 1 s, carbon monoxide transfer factor, 6-minute walk distance, BODE index and exacerbation frequency.

Conclusion: PMP technology can be useful in identifying potential biomarkers in patients with COPD. Panels of selected serum markers are associated with important clinical predictors of outcome in these patients.

Chronic obstructive pulmonary disease (COPD) is projected to be the third leading cause of death in the world by the year 2020.[1 2] Despite the well-documented role of cigarette smoking in the genesis of COPD, it is unclear what steps are involved in its pathogenesis.[3] Most, if not all, patients with COPD develop a combination of lung emphysema with its characteristic pattern of alveolar destruction and abnormal repair as well as small airway inflammation that persists even years after smoking cessation.[4]

The current pathogenetic theories for the development of COPD include an imbalance between the protease and antiprotease system, dysregulation of oxidant-antioxidant activity and chronic airway inflammation, processes that lead to the progressive destruction and abnormal repair of the lung connective tissue matrix.[5] Recent studies have suggested that increased apoptosis of the alveolar wall accounts in part for the loss of lung tissue that characterises emphysema.[6 7] Transgenic and null mutant mouse studies have identified a number of genes and pathways that, when altered, result in the morphological changes of emphysema.[8–10]

Although COPD primarily affects the lungs, it is associated with important systemic consequences which include malnutrition with a low body mass index (BMI)[11] and impaired peripheral muscle function.[12] These clinically relevant expressions of the disease have been associated with detectable systemic changes including evidence of increased oxidative stress, activation of circulating inflammatory cells and increased levels of proinflammatory cytokines.[13 14] The multidimensional expression of COPD can be expressed by a clinical score including BMI, degree of obstruction (O), perception of dyspnoea (D) and exercise capacity (E) by the 6-minute walk distance known as the BODE index.[15] This index predicts mortality better than the forced expiratory volume in 1 s (FEV_1).

We reasoned that the pathobiological processes that occur in the lungs and possibly in systemic tissues such as the peripheral muscles of patients with COPD could be associated with systemic biomarker levels detectable in the systemic circulation. Despite the many studies aimed at identifying the pathogenesis of COPD, to our knowledge only one study[16] has explored the potential value of high-density microarray technology to systematically define the serum protein expression profile in patients with COPD. Using a novel protein microarray platform (PMP) technology, we compared the serum proteomic profile of 143 serum biomarkers in patients with COPD with that of age and sex-matched controls. We also explored the relationship between a selected subset of 24 biomarkers with clinically important outcome variables in COPD including lung function, the BODE index and its components and the frequency of exacerbations.

METHODS

Patient recruitment

This is a matched case-control study of 48 patients with severe COPD (FEV_1 <55% predicted), 8 of whom were current smokers. We then matched 8 control smokers and 40 subjects who had smoked <5 pack years and had stopped at least

Abbreviations: AR, amphiregulin; BAL, bronchoalveolar lavage; BDNF, brain-derived neurotropic factor; BMI, body mass index; BODE index, body mass index (B), degree of obstruction (O), perception of dyspnoea (D) and exercise capacity (E); COPD, chronic obstructive pulmonary disease; FDR_p, false discovery adjusted p value; FEV_1, forced expiratory volume in 1 s; IFNγ, interferon γ; IL, interleukin; MMP, metalloproteinase; 6MWD, 6-minute walk distance; βNGF, nerve growth factor β; PLS-DA, partial least squares discriminant analysis; PMP, protein microarray platform; RCA, rolling cell amplification; TGFα, tissue growth factor α; TIMP-1, tissue inhibitor of metalloproteinase 1; T_{LCO}, carbon monoxide transfer factor; TNFα, tumour necrosis factor α; VEGF, vascular endothelial growth factor

Table 1 Cases and controls stratified by exacerbation frequency and smoking status

	Never/ex-smoker	Active smoker	Total
COPD			
Group 1 (no exacerbations)	12	4	16
Group 2 (⩽2 exacerbations/year)	12	4	16
Group 3 (>2 exacerbations/year)	15	0	15
Controls	40	8	48

20 years previously or who had always been non-smokers. All controls had a ratio of FEV_1 to forced vital capacity (FVC) of >0.7 and FEV_1 >70% predicted. Participants were >35 years of age and patients with COPD had to be clinically stable and without exacerbations for at least 3 months. Subjects with a history of asthma or atopy, conditions precluding performance of the tests, and a systemic infection or an inflammatory process that could be associated with abnormal biomarker profile were excluded. All patients were followed for 1 year and were stratified according to smoking history into ex-smokers (never smoked or ex-smokers for >15 years and <20 pack-years) and active smokers. The controls were frequency matched according to sex, age and smoking history (table 1).

The pulmonary function tests were measured according to ATS standards[17] and the BODE index was calculated as previously reported.[15] Exacerbations were defined as episodes of increased dyspnoea, sputum or cough lasting >24 h and requiring treatment with antibiotics and/or corticosteroids.[18] After follow-up for 1 year, patients were stratified into no exacerbations (n = 12), <2 exacerbations (n = 12) and ⩾2 exacerbations (n = 15).

Specimen collection

Blood samples were drawn, centrifuged and the serum frozen at –80°C. Rolling cell amplification (RCA) immunoassay was performed by Molecular Staging Inc (MSI, New Haven, Connecticut, USA) using a protein microarray platform that measured levels of 143 analytes (see table S1 available online at http://thorax.bmj.com/supplemental) on five separate arrays.[19 20] After incubating and washing the serum samples on microarrays, the captured proteins were detected by specific biotinylated second antibodies and a universal anti-biotin antibody was bound to the secondary antibodies. The anti-biotin antibody contained an oligonucleotide DNA primer used for amplification. During the process, a circular DNA hybridises to the oligonucleotide DNA primer in the presence of DNA polymerase and fluorescent nucleotides to generate a signal. Following RCA, the slides were scanned (L200 scan, TECAN, Durham, North Carolina, USA) using a proprietary software. The fluorescence intensity of microarray spots was analysed and the resulting mean intensity values were measured. Dose-response curves for the biomarkers were determined with increasing intensity indicating increasing analyte concentration.

Data analysis

A more complete discussion of the analysis used in this study is available in the online supplement at http://thorax.bmj.com/supplemental. In summary, two independent statistical approaches were used: (1) we tested the distribution of biomarkers for an association with COPD by univariate analysis adjusting for multiple comparisons using false discovery rate analysis;[21] and (2) we used a variable clustering (VARCLUS) tool which divides the biomarkers into non-overlapping unidimensional groups or clusters,[22] a process similar to factor analysis. Each cluster's predictive value was determined by computing the partial regression coefficient of individual cluster centroids with COPD using partial least squares discriminant analysis (PLS-DA). After the initial analysis, we selected a group of 24 biomarkers from those clusters that showed a significant association with the diagnosis of COPD (clinical history and presence of airflow limitation). The biomarkers were chosen to reflect a variety of pathobiological mechanisms relevant to the disease process. The resultant panel of biomarkers was then tested for strength of association with variables known to predict outcome in COPD, including transfer factor for carbon monoxide (T_{LCO}),[23] 6-minute walk distance (6MWD), the BODE index and exacerbation frequency.

Table 2 Basic characteristics of patients with COPD and controls*

Variable	COPD (n = 47) M = 28/F = 19	Controls (n = 48) M = 29/F = 19	T test for independent samples (p value)
Age	64 (8)	64 (7)	NS
BMI (kg/m^2)	26.6 (5.2)	27.6 (4.8)	NS
Smoking history (pack-years)	70 (40)	23 (3)	<0.001
Pre-bronchodilator FEV_1 (l)	0.87 (0.33)	2.46 (0.61)	<0.001
Pre-bronchodilator FEV_1 (% predicted)	32 (10)	90 (15)	<0.001
Post-bronchodilator FEV_1 (l)	0.94 (0.36)	NA	
Post-bronchodilator FEV_1 (% predicted)	35 (11)	NA	
TLC (l)	7.00 (1.57)	5.33 (1.48)	<0.001
TLC (% predicted)	126 (2)	95 (19)	<0.001
RV (l)	4.46 (1.36)	2.07 (0.81)	<0.001
RV (% predicted)	211 (60)	96 (33)	<0.001
T_{LCO} (ml/min/mmHg)	9.9 (3.8)	19.5 (6)	<0.001
T_{LCO} (% predicted)	48 (5)	92 (24)	<0.001
MRC dyspnoea	2.5 (0.8)	0.08 (0.3)*	<0.001
6MWD (m)	365 (109)	540 (95)	<0.001
SGRQ composite score	53 (18)	4.8 (2.6)*	<0.001
BODE index	4.7 (1.8)	2.2 (1.8)	<0.001

NA, not available; NS, not significant; BMI, body mass index; FEV_1, forced expiratory volume in 1 s; TLC, total lung capacity; RV, residual volume; T_{LCO}, carbon monoxide transfer factor; MRC, Medical Research Council; 6MWD, 6-minute walk distance; SGRQ, St George's Respiratory Disease Questionnaire; BODE, body mass index (B), obstruction (O), dyspnoea (D) and exercise endurance (E).
Data presented as mean (SD).
*Log transformation was performed on variables with non-normal distribution.

Table 3 Statistical comparison of biomarkers in patients with COPD and controls

		Controls		COPD patients				
Sequence	Protein ID	Mean	SD	Mean	SD	T ratio	FDR_p	p Value
1	MMP-9	4129.05	1524.12	6298.61	2437.93	5.07	0.00028	0.00000
2	Eotaxin-2	3821.17	2786.26	6897.35	4303.97	4.37	0.00078	0.00003
3	MMP-7	860.72	340.62	1431.51	1021.53	4.37	0.00078	0.00003
4	MMP-10	1111.17	445.91	1813.39	1424.64	4.54	0.00078	0.00002
5	TARC	178.84	124.23	386.29	541.28	4.44	0.00078	0.00002
6	TNF-α	83.24	19.97	112.99	48.21	4.40	0.00078	0.00003
7	BDNF	293.34	347.45	992.07	1348.16	4.17	0.00106	0.00007
8	GCP-2	100.41	26.19	131.26	45.91	4.16	0.00106	0.00007
9	MIF	702.08	165.77	1062.19	931.81	4.15	0.00106	0.00007
10	NE	3951.64	925.28	4917.56	1284.20	4.18	0.00106	0.00006
11	IL-1-sR1	173.18	89.25	264.36	157.92	4.04	0.00144	0.00011
12	MMP-8	7479.60	5112.69	14511.88	9305.13	4.00	0.00151	0.00013
13	TIMP-1	2885.88	1477.78	4677.85	2424.25	3.91	0.00197	0.00018
14	AR	117.76	30.93	146.85	47.46	3.77	0.00253	0.00029
15	IL-10rβ	689.21	200.70	874.60	287.25	3.70	0.00253	0.00037
16	IL-12p40	130.59	41.13	162.29	54.10	3.69	0.00253	0.00037
17	IL-1rα	176.72	53.05	221.43	61.74	3.80	0.00253	0.00026
18	IL-2rβ	275.40	92.33	366.01	155.99	3.73	0.00253	0.00033
19	IL-8	209.52	80.41	278.40	101.25	3.79	0.00253	0.00027
20	I-TAC	86.84	51.75	138.41	95.26	3.71	0.00253	0.00036
21	VEGF	127.21	40.98	168.05	65.40	3.70	0.00253	0.00036
22	VEGF-D	263.56	142.27	361.02	186.09	3.56	0.00377	0.00058
23	CD30	158.85	69.99	210.88	102.48	3.48	0.00457	0.00077
24	Eotaxin	235.07	80.48	282.00	65.23	3.49	0.00457	0.00075
25	MCP-1	633.02	303.92	936.95	584.67	3.44	0.00496	0.00087
26	IL-1α	91.98	31.06	113.74	38.44	3.41	0.00508	0.00096
27	Prolactin	925.25	244.88	1274.07	838.37	3.42	0.00508	0.00093
28	TNF-R1	525.09	455.95	800.09	616.36	3.35	0.00597	0.00117
29	IL-1β	98.21	34.65	124.21	53.06	3.31	0.00662	0.00134
30	BLC	192.76	89.61	294.84	240.72	3.29	0.00675	0.00142
31	FGF-4	347.42	95.24	434.21	169.61	3.26	0.00719	0.00156
32	MPIF-1	624.64	331.03	825.75	398.73	3.17	0.00932	0.00209
33	IGFBP-4	6118.33	1188.73	7136.63	2703.80	3.15	0.00962	0.00222
34	CTACK	3480.02	1367.85	4745.23	3090.78	3.12	0.01005	0.00239
35	HCC-1	1201.77	352.24	1549.81	870.31	3.07	0.01130	0.00276
36	Eotaxin-3	229.35	114.73	609.19	2006.73	3.06	0.01138	0.00291
37	IL-17	343.13	69.30	390.90	84.96	3.05	0.01138	0.00302
38	TGF-α	210.31	107.82	595.09	2168.58	3.05	0.01138	0.00299
39	MIP-1δ	2094.72	836.45	2878.11	1762.62	2.99	0.01326	0.00362
40	EGF	158.28	64.86	216.59	164.16	2.93	0.01458	0.00428
41	M-CSF	78.34	24.86	96.89	35.82	2.93	0.01458	0.00422
42	Protein C	8112.46	1788.08	9087.41	1603.96	2.94	0.01458	0.00412
43	HCC-4	8572.51	2501.97	10071.78	2581.08	2.90	0.01498	0.00461

The protein markers were filtered by false discovery adjusted p-value (FDR_p) of 0.015. The FDR adjusted p value (sometimes called just FDR) is a practical way of dealing with multiple testing issues and can be interpreted as estimated proportion of false positives in the list. The T ratio was computed from common log scale. The names of all the analytes are given in the online supplement available at http://thorax.bmj.com/supplemental.

RESULTS

Study population

The characteristics of the patients and the controls are summarised in table 2. As expected, the patients had higher smoking exposure (pack-years), significant airflow limitation, higher lung volumes, worse BODE scores and health-related quality of life than controls. Patients and controls were of similar age, sex and BMI.

Biomarkers that distinguish between patients with COPD and controls

In the univariate analysis, 43 biomarkers were identified that differed between patients and controls. To adjust for multiple analysis, these were filtered by false discovery rate adjusted p value (FDR_p) of <0.015 (table 3).

The second approach (variable cluster analysis) resulted in 30 different clusters, 19 of which correlated significantly with the diagnosis of COPD. We selected biomarkers from among these 19 clusters to reflect a variety of pathophysiological mechanisms considered relevant to COPD. In order to enrich the exploratory value of the panel, two biomarkers—prolactin and plasminogen activator inhibitor type 2 (PAI-II)—were included despite lack of an obvious disease association. The selected panel biomarkers are shown in table 4 and their full description is given in the online supplement available at http://thorax.bmj.com/supplemental.

Associations of the biomarker panel with FEV_1, TLCO, 6MWD, BODE index, BMI and exacerbation rate in patients with COPD

In the patients with COPD, the selected biomarkers tested in the panel correlated significantly with FEV_1 (fig 1). The findings were replicated for the TLCO (fig 2), the BODE index (fig 3) and the exacerbation rate (fig 4).

We also observed a correlation with the 6MWD while there was no correlation with BMI (not shown). The same selected biomarkers are shown for each analysis. Most of the markers were associated with all of the physiological indicators of disease, but the strength of the association differed from outcome to outcome as did the rank order of each biomarker.

DISCUSSION

This study had two important findings: (1) that PMP technology can be useful in identifying potential biomarkers in patients with COPD; and (2) that a pattern of systemic

Table 4 Biomarkers selected for analysis

Pathobiological function	
Chemoattractants	I-TAC, eotaxin-2, MPIF-1, MCP-1, MIP-1β, IL-8, PARC
Inflammation	IL-15, IL-1ra, IL-17, TNFα, TNF R1, IFNγ, IL-12 p40, IL-2Rγ
Destruction and repair	TGFα, VEGF, AR, BDNF, βNGF, MMP-9, TIMP-1
Novel markers	PAI-II, prolactin

The biomarkers were selected from clusters statistically associated with the diagnosis of COPD and thought to have known or potential significance in the pathobiology of COPD.
AR, amphiregulin; BDNF, brain-derived neurotrophic factor; βNGF, β-nerve growth factor; IFNγ, interferon γ; IL, interleukin; IL-1ra, interleukin 1 receptor antagonist; IL-2Rγ, interleukin 2 receptor gamma; I-TAC, interferon γ-inducible T cell α chemoattractant; MCP-1, monocyte chemotactic protein 1; MIP-1β, macrophage inflammatory protein 1β; MMP-9, matrix metalloproteinase 9; MPIF-1, myeloid progenitor inhibitory factor 1; PAI-II, plasminogen activator inhibitor II; TGFα, transforming growth factor α; TIMP-1, tissue inhibitors of metalloproteinases 1; TNFα, tumour necrosis factor α; TNF R1, tumour necrosis factor receptor I; VEGF, vascular endothelial growth factor.

biomarkers identified in these patients can be associated with different clinical variables known to predict disease outcome including degree of airflow limitation, lung transfer factor, functional capacity, the BODE index and exacerbation frequency.

Several groups have shown an increase in a number of circulating inflammatory biomarkers in COPD,[24–26] suggesting that it might be possible to characterise patients with COPD using systemic biomarkers. To address this question we used a novel technology that simultaneously evaluated analytes covering diverse potential processes including inflammation, chemoattraction, cell activation, tissue destruction and repair. Based on a collaborative effort of statistical results and scientific plausibility, a subset of 24 biomarkers was identified and selected for subsequent testing against a variety of clinically important parameters. Many studies have been published on the association between a specific marker and COPD disease status, with both positive and negative results being reported.[24–31] The disagreement in results can be attributed to different factors, including the heterogeneity of COPD phenotypes, low sample size, or the use of different methodologies and assays. The development of a panel of biomarkers addressing preconceived multiple pathophysiological pathways may provide a more specific tool to serve as an intermediate end point reflecting the natural history of the disease.

One obvious limitation of this preliminary dataset is that the biomarkers identified are limited by the pool of analytes that were available for the primary assessment. Clearly, use of an ''open'' proteomic platform would give information about a much broader range of proteins and might provide additional

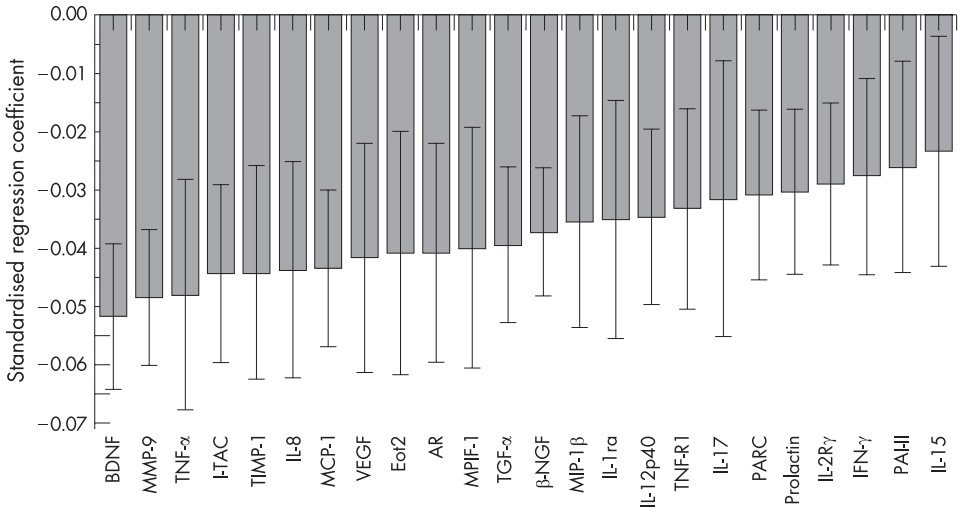

Figure 1 Correlation of the selected biomarker panel with forced expiratory volume in 1 s (FEV_1) in patients with COPD. The size of the bar in the graph indicates the magnitude of the regression coefficients and the 95% confidence interval is also indicated for each bar. If the confidence interval includes zero, the associated biomarker is ''not significant''. The overall regression model was significant by a permutation test ($p<0.01$). The standardised coefficients for this and for figs 2, 3 and 4 are for scaled and centred markers and scaled response. These coefficients can be used to interpret the influence of the markers on the clinical response. The standardised regression coefficient for each marker measures the effect of the marker on the clinical response adjusted for all other markers in the regression (partial correlation). The coefficients can also be compared across the clinical responses since they are scaled. Definitions of the biomarkers are given in the footnote to table 4.

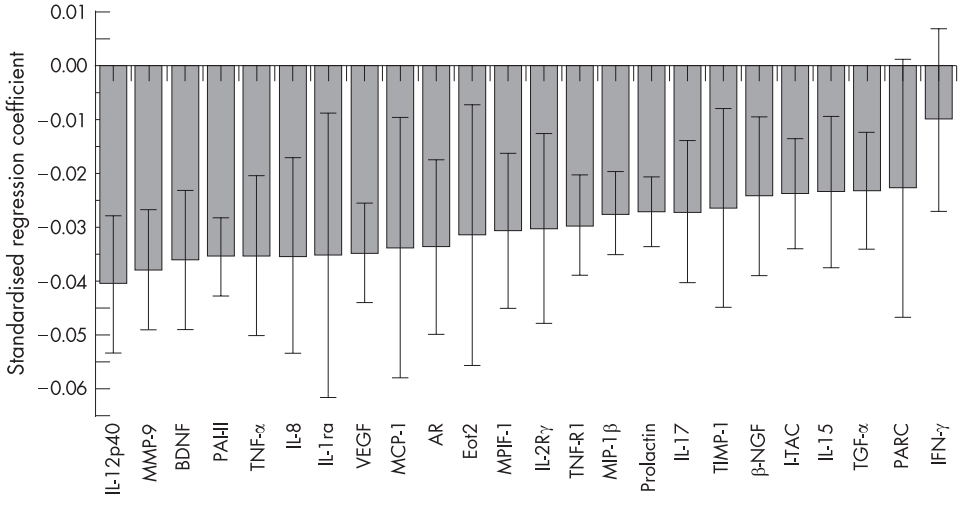

Figure 2 Correlation of the selected biomarker panel with carbon monoxide transfer factor (T_{LCO}) in patients with COPD. The size of the bar in the graph indicates the magnitude of the regression coefficients and the 95% confidence interval is also indicated for each bar. If the confidence interval includes zero, the associated biomarker is ''not significant''. The overall regression model was significant by a permutation test ($p<0.01$). Definitions of the biomarkers are given in the footnote to table 4.

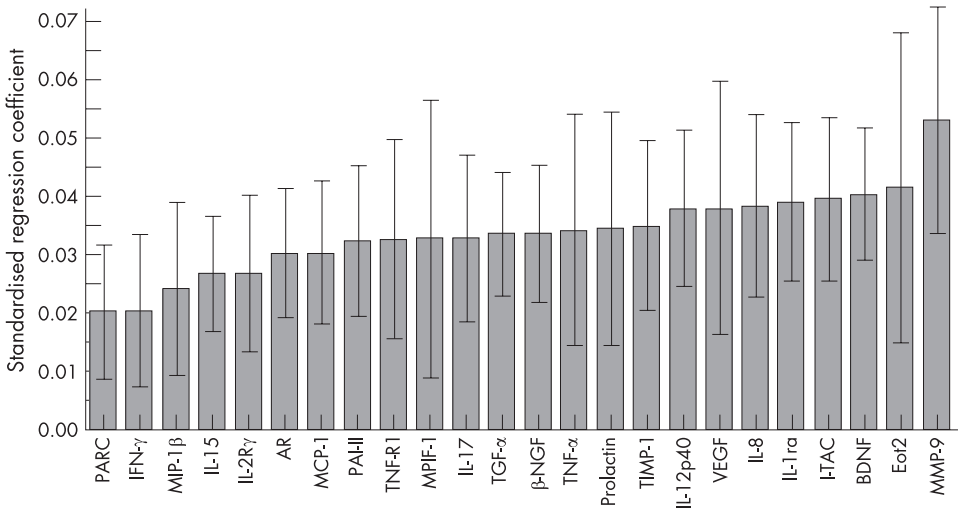

Figure 3 Correlation of the selected biomarker panel with the BODE index in patients with COPD. The size of the bar in the graph indicates the magnitude of the regression coefficients and the 95% confidence interval is also indicated for each bar. If the confidence interval includes zero, the associated biomarker is "not significant". The overall regression model was significant by a permutation test ($p<0.01$). Definitions of the biomarkers are given in the footnote to table 4.

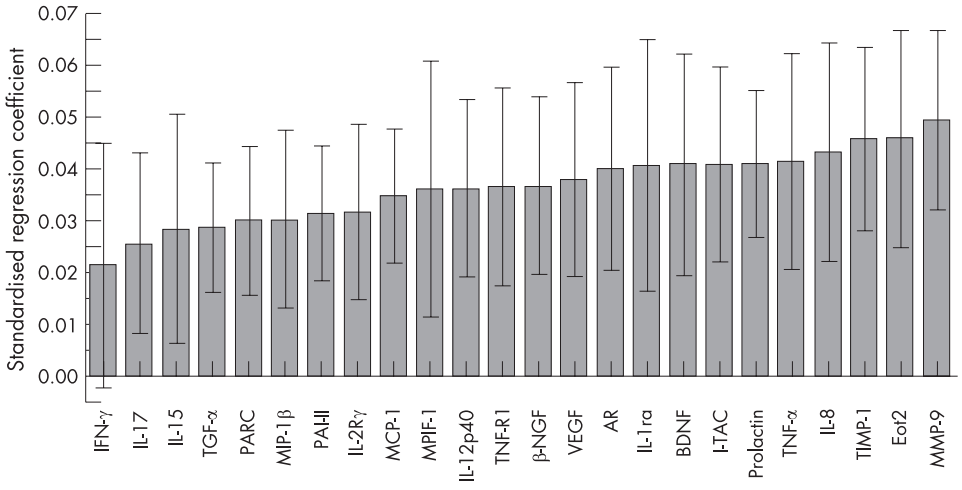

Figure 4 Correlation of the selected biomarker panel with the exacerbation rate in patients with COPD. The size of the bar in the graph indicates the magnitude of the regression coefficients and the 95% confidence interval is also indicated for each bar. If the confidence interval includes zero, the associated biomarker is "not significant". The overall regression model was significant by a permutation test ($p<0.01$). Definitions of the biomarkers are given in the footnote to table 4.

insights into biomarker selection and disease processes. However, a recent report[16] using a panel developed from the one reported here shows that the panel as developed is valid and capable of reflecting changes induced by exacerbations. Recognising the fact that the discussion is valid only for the analytes explored, our findings may help to shed light on the underlying pathogenetic processes involved in this disease.

It has been proposed that various proteases break down lung connective tissue components to cause emphysema,[5 6] leading to aberrant remodelling and/or degradation of the extracellular matrix. In our study, several proteins (table 3) related to the protease-antiprotease mechanism were clearly different between patients with COPD and controls. Thus, metalloproteinases 7, 8, 9 and 10 (MMP-7, MMP-8, MMP-9 and MMP-10) were among the proteins with large differences between groups. Of these, MMP-9 showed the strongest association with FEV_1 and TLCO, which is interesting because MMP-9 has been implicated in the experimental genesis of emphysema.[32 33] The tissue inhibitor of metalloproteinase 1 (TIMP-1), a collagenase inhibitor, was also different between patients and controls, providing evidence that the final expression of the disease may rest upon the appropriate balance of the system.[33]

Differences were also found in enzymes other than the metalloproteinases that are related to tissue destruction, as well as proteins related to repair, that deserve some comments. While the fold increase of neutrophil elastase in COPD was not as great as that found for the metalloproteinases, the difference was still statistically significant. Previous studies of experimental emphysema produced by pancreatic or neutrophil elastase showed that increased levels of elastase enzymes lead to the degradation of connective tissue components and, thus, enlargement of distal airspaces.[34] While both elastin and collagen are rapidly re-synthesised in these animal models and mRNA levels for both are increased, the connective tissue remodelling process is ineffective and lung mechanical properties remain abnormal.[35]

The differences in tissue growth factor alpha (TGFα), amphiregulin (AR), brain-derived neurotropic factor (BDNF) and nerve growth factor β (βNGF) and their association with low FEV_1 and TLCO (figs 1 and 2) suggest that connective tissue remodelling continues even in severe advanced COPD in humans, but the process fails effectively to restore the mechanical properties of the diseased lung. The role of TGFα is something of a mystery. Mice genetically manipulated to overexpress TGFα develop emphysema postnatally,[36] yet an in vitro model of alveolar re-epithelialisation showed that TGFα induced faster wound repair.[37] The presence of significant associations between BDNF and lung function and the BODE index (fig 3) is particularly interesting. Recent evidence indicates that BDNF decreases conversion from oxygen to hydrogen peroxide in experimental cell cultures,[38] suggesting a role in the modulation of oxidative stress, and makes this an interesting marker to study. Furthermore, similar to results seen with AR, exogenous BDNF can protect cells from serum deprivation-induced cell death.[39]

It has been suggested that angiogenesis and apoptosis of the alveolar wall may have a role in emphysema. While little is

known about the role of the EGF family member AR in the aetiology of COPD, one study has found that AR can inhibit apoptosis of non-small cell lung cancer cell line.[40] Blockade of vascular endothelial growth factor R2 (VEGF-R2) receptor in rats induces apoptosis of the alveolar cell wall and results in an emphysema-like pathology.[41 42] Several studies have found decreased expression of VEGF in induced sputum or bronchoalveolar lavage (BAL) fluid from patients with obstructive lung disease in comparison with normal subjects.[43 44] These studies have also shown a direct association between the reduction in VEGF and FEV_1. While our study showed an increase in VEGF serum content that was inversely associated with FEV_1, this difference could be due to differential expression of VEGF in lung tissue and serum. Studies of VEGF expression in human lung tissue by immunohistochemistry have shown increased VEGF in pulmonary and airway smooth muscle in subjects with COPD that correlated with decreased FEV_1.[45] Furthermore, patients with cystic fibrosis show an inverse relationship in the level of VEGF in serum and BAL fluid compartments. These patients had a higher level of VEGF in serum and a lower level of VEGF in BAL fluid compared with controls.[46] The role of apoptosis and its relationship to inflammation and repair seem supported by our findings.

Current thinking places inflammation at the centre of the pathogenetic mechanisms of COPD. The inflammation is characterised by increased numbers of alveolar macrophages, neutrophils and T lymphocytes, together with the release of multiple inflammatory mediators that result in a high level of oxidative stress. Multiple proteins related to inflammation were detected in the serum of patients with COPD (table 4). These included interleukin (IL)-12, IL-15, IL-17, IL-1 receptor antagonist (IL-1ra), tumour necrosis factor α (TNFα), tumour necrosis factor receptor 1 (TNF R1), interferon γ (IFNγ), IL-12p40 and IL-2Rγ. There is experimental evidence for the participation of all of these proteins in the inflammation that characterises COPD, and raises the possibility that the systemic manifestations of COPD may be intimately related to this process. Indeed, the association between inflammatory markers and exacerbation rate (fig 4) suggests that this manifestation of the disease could be modulated by amplification of the inflammatory cascade. In this regard, eotaxin-2—which had one of the strongest associations with the exacerbation rate in our patients—is a strong chemotactic cytokine for eosinophils,[47] cells that have been found to be increased in airway biopsy tissue from patients with COPD exacerbations.[48] Indeed, although the inflammatory pathways of COPD appear to be more related to lymphocytes expressing a T helper 1 (Th1) bias,[49] a high level of Th2 chemokines have been reported in experimental models of emphysema induced by cigarette smoking.[10]

There were several novel proteins that differed between patients with COPD and controls. We selected two of them—plasminogen activator inhibitor type 2 (PAI-II) and prolactin—because of their presence in one of the eight clusters with the strongest association with COPD. PAI-II belongs to the serpine class of protease inhibitors and is involved in the thrombogenic cascade. Known to be produced by activated monocytes in the peripheral blood,[50] this protein (together with PAI-I) may have a role in tissue remodelling in airways disease.[51] These data warrant further investigation to explore the possible role of serpines in COPD.[52] Prolactin upregulation presents an enigma. Prolactin receptor has recently been reported to be upregulated in the lungs of mice exposed to lipopolysaccharide,[53] and prolactin can activate the inflammatory natural killer (NF)-κβ cascade in pulmonary fibroblasts.[54] It is therefore plausible that prolactin may play a role in the inflammatory environment in COPD.

There are a number of important limitations to our study. Not all of the possible proteins that participate in the complex mechanism of COPD were tested. Absent were some with a known relationship to COPD such as C-reactive protein and fibrinogen, and some of potential importance such as MMP-12. The reason for their omission was not any preconceived mechanistic bias. Our study was designed as a proof of principle rather than a totally comprehensive evaluation of all of the markers that could potentially be explored. Many complex diseases have components related to inflammation, tissue remodelling, apoptosis and chemoattraction of specific cell types. This observation suggests that a panel of analytes might provide insight into the pathobiology of the disease under study in the absence of, or in conjunction with, novel ''disease-specific'' biomarkers. We also acknowledge that not all phenotypic expressions of COPD were analysed; for example, it would have been interesting to have related the biomarkers to changes in the CT scan of patients with emphysema, but unfortunately the technique needed to quantitatively express CT changes was not available. However, the T_{LCO} does relate to the phenotypic expression of emphysema. We believe that this study represents a proof of concept and opens a window for hypothesis testing and perhaps the discovery of yet to be described pathway interactions and targets.

For the correlation analyses we attempted to address the issue of many proteins representing the same pathophysiological mechanism by empirically grouping them according to their statistical strength and their presumed pathobiological role. We acknowledge the latter to be empirical, but it is based on the data currently available and aimed at simplifying the prospective testing. Furthermore, the inclusion of too many proteins may be intellectually desirable but may cause important cross-correlative noise that may actually cloud the interpretation of the results. We also acknowledge that the patients included in the study do not represent the large population of patients with COPD since all of them had severe disease. However, the patients included represent those likely to be seen by clinicians and to benefit from new therapeutic strategies. On the other hand, this study is unique in that patients and controls were phenotypically well characterised and matched by age, sex and—very importantly—by smoking habits to minimise the hypothetical influence of these confounding factors. Indeed, the inclusion or exclusion of smokers in each of the groups did not affect the results. In addition, the evaluation of important associations of the panel markers with clinical markers of COPD such as the BODE index and its individual components offers a more comprehensive picture of the value of the technique. The association with exacerbation frequency is particularly interesting because exacerbations constitute an extremely important outcome and one where elucidation of the factors that may help prevent their occurrence would prove extremely useful. Finally, we also acknowledge that the stability of biomarker levels in serum samples is not well characterised and that we did not repeat the tests at different times. However, the recent report by Hurat and colleagues[16] using the panel derived from this study independently validated our findings.

In summary, using a serum PMP, we have identified a biomarker profile whose expression levels can distinguish patients with COPD from smokers and non-smokers without COPD. We have also found an association between the level of selected biomarkers and lung function, the degree of airflow limitation and T_{LCO}, a marker of lung tissue destruction. Furthermore, we documented an association between the expression of the serum biomarkers and the integrated local and systemic manifestations of the disease as represented by the functional capacity and the BODE index. The expression of biomarkers was also associated

with the exacerbation rate, crucial events in the natural course of the disease. The ease of sampling of peripheral blood and the continuing improvement and availability of multiplexed immunoassay technology should provide us with a new tool for research in this deadly disease.

Further information is given in the online supplement available at http://thorax.bmj.com/supplemental.

Authors' affiliations

Victor Pinto-Plata, Bartolome Celli, Pulmonary, Critical Care and Sleep Division, Caritas St Elizabeth's Medical Center, Tufts University, Boston, Massachusetts, USA

John Toso, Kwan Lee, Daniel Park, John Bilello, Hana Mullerova, Mary M D e Souza, Rupert Vessey, Discovery Medicine, High Throughput Biology and Biomedical Data Sciences, GlaxoSmithKline R&D, USA

This work was supported by an unrestricted grant from GlaxoSmithKline and the Thoracic and Overholt Foundation. No funds were received from the tobacco industry.

Competing interests: JT, KL, DP, JB, HM, MMDS and RV are all full time employees of GlaxoSmithKline.

REFERENCES

1 **Ezzati M**, Lopez AD. Estimates of global mortality attributable to smoking in 2000. *Lancet* 2003;**362**:847–52.
2 **Mannino DM**, Homa DM, Akinbami LJ, *et al.* Chronic obstructive pulmonary disease surveillance – United States, 1971–2000. *MMWR* 2002;**51**:1–16.
3 **Celli BR**, MacNee W. Standards for the diagnosis and treatment of COPD. *Eur Respir J* 2004;**23**:932–46.
4 **Hogg JC**, Chu F, Utokaparach S, *et al.* The nature of small-airway obstruction in chronic obstructive pulmonary disease. *N Engl J Med* 2004;**350**:2645–53.
5 **Barnes PJ**, Shapiro SD, Pauwels RA. Chronic obstructive pulmonary disease: molecular and cellular mechanisms. *Eur Respir J* 2003;**22**:672–88.
6 **Tuder RM**, Zhen CYL. Cho L, *et al.* Oxidative stress and apoptosis interact and cause emphysema due to vascular endothelial growth factor receptor blockade. *Am J Respir Cell Mol Biol* 2003;**29**:88–97.
7 **Aoshiba K**, Yokohori N, Nagai A. Alveolar wall apoptosis causes lung destruction and emphysematous changes. *Am J Respir Cell Mol Biol* 2003;**28**:555–62.
8 **Fehrenbach H**. Animal models of chronic obstructive pulmonary disease: some critical remarks. *Pathobiology* 2002;**70**:277–83.
9 **Mahadeva R**, Shapiro SD. Chronic obstructive pulmonary disease: experimental animal models of pulmonary emphysema. *Thorax* 2002;**57**:908–14.
10 **Elias J**. The relationship between asthma and COPD. Lessons from transgenic mice. *Chest* 2004;**126**:111–6S.
11 **Schols AM**, Slangen J, Volovics L, *et al.* Weight loss is a reversible factor in the prognosis of chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 1998;**157**:1791–7.
12 **Maltais F**, Simard AA, Simard C, *et al.* Oxidative capacity of the skeletal muscle and lactic acid kinetics during exercise in normal subjects and in patients with COPD. *Am J Respir Crit Care Med* 1996;**153**:288–93.
13 **Agustí AG**, Noguera A, Sauleda J, *et al.* Systemic effects of chronic obstructive pulmonary disease. *Eur Respir J* 2003;**21**:347–60.
14 **Rahman I**, Morrison D, Donaldson K, *et al.* Systemic oxidative stress in asthma, COPD, and smokers. *Am J Respir Crit Care Med* 1996;**154**:1055–60.
15 **Celli BR**, Cote CG, Marin JM, *et al.* The body mass index, airflow obstruction, dyspnea and exercise capacity index in chronic obstructive pulmonary disease. *N Engl J Med* 2004;**350**:1005–12.
16 **Hurst JR**, Donaldson GC, Perea WR, *et al.* Utility of plasma biomarkers at exacerbation of chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 2006;**174**:867–74.
17 **American Thoracic Society**. Lung function testing: selection of reference values and interpretative strategies. *Am Rev Respir Dis* 1991;**144**:1202–18.
18 **Rodriguez-Roisin R**. Toward a consensus definition for COPD exacerbations. *Chest* 2000;**117**:398–401S.
19 **Schweitzer B**, Wiltshire S, Lambert J, *et al.* Immunoassays with rolling circle DNA amplification: a versatile platform for ultrasensitive antigen detection. *Proc Natl Acad Sci USA* 2000;**97**:10113–9.
20 **Perlee L**, Christiansen J, Dondero R, *et al.* Development and standardization of multiplexed antibody microarrays for use in quantitative proteomics. *Proteome Sci* 2004;**2**:9.
21 **Bradley E**. Bayesians, frequentists and scientists. *J Am Stat Assoc* 2005;**100**:1–5.
22 **SAS/STAT**. *User's Guide. Version 8. Chapter 68. The VARCLUS procedure*, 3593–620.
23 **Nakano Y**, Muro S, Sakai H, *et al.* Computed tomographic measurements of airway dimensions and emphysema in smokers. Correlation with lung function. *Am J Respir Crit Care Med* 2000;**162**:1102–8.
24 **Gan WQ**, Man SF, Senthilselvan A, *et al.* Association between chronic obstructive pulmonary disease and systemic inflammation: a systematic review and a meta-analysis. *Thorax* 2004;**59**:574–80.
25 **Vernooy JH**, Kucukaycan M, Jacobs JA, *et al.* Local and systemic inflammation in patients with chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 2002;**166**:1218–24.
26 **Schols AMWJ**, Buurman WA, Staal-van den Brekel AJ, *et al.* Evidence for a relation between metabolic derangement and increased levels of inflammatory mediators in a subgroup of patients with chronic obstructive pulmonary disease. *Thorax* 1996;**51**:819–24.
27 **Aaron SD**, Angel JB, Lunau M, *et al.* Granulocyte inflammatory markers and airway infection during acute exacerbation of chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 2001;**163**:349–55.
28 **Noguera A**, Busquets X, Sauleda J, *et al.* Expression of adhesion molecules and G proteins in circulating neutrophils in chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 1998;**158**:1664–8.
29 **Malo O**, Sauleda J, Busquets X, *et al.* Systemic inflammation during exacerbations of chronic obstructive pulmonary disease. *Arch Bronconeumol* 2002;**38**:172–6.
30 **Broekhuizen R**, Wouters EF, Creutzberg EC, *et al.* Raised CRP levels mark metabolic and functional impairment in advanced COPD. *Thorax* 2006;**61**:17–22.
31 **Pinto-Plata V**, Mullerova H, Toso J, *et al.* C-reactive protein in patients with COPD, control smokers and non-smokers. *Thorax* 2006;**61**:23–8.
32 **Culpitt SV**, Rogers DF, Traves SL, *et al.* Sputum matrix metalloproteases: comparison between chronic obstructive pulmonary disease and asthma. *Respir Med* 2005;**99**:703–10.
33 **Higashimoto Y**, Yamagaga Y, Iwata T, *et al.* Increased serum concentrations of tissue inhibitor of metalloproteinase-1 in COPD patients. *Eur Respir J* 2005;**25**:885–90.
34 **Churg A**, Wright JL. Proteases and emphysema. *Curr Opin Pulm Med* 2005;**11**:153–9.
35 **Shapiro SD**, Goldstein NM, Houghton AM, *et al.* Neutrophil elastase contributes to cigarette smoke-induced emphysema in mice. *Am J Pathol* 2003;**163**:2329–35.
36 **Hardie WD**, Piljan-Gentle A, Dunlavy MR, *et al.* Dose-dependent lung remodeling in transgenic mice expressing transforming growth factor-alpha. *Am J Physiol Lung Cell Mol Physiol* 2001;**281**:1088–94.
37 **Kheradmand F**, Folkesson HG, Shum L, *et al.* Transforming growth factor-alpha enhances alveolar epithelial cell repair in a new in vitro model. *Am J Physiol* 1994;**267**:728–38.
38 **Yamagata T**, Satoh T, Ishikawa Y, *et al.* Brain-derived neurotropic factor prevents superoxide anion-induced death of PC12h cells stably expressing TrkB receptor via modulation of reactive oxygen species. *Neurosci Res* 1999;**35**:9–17.
39 **Zheng WH**, Quirion R. Comparative signaling pathways of insulin-like growth factor-1 and brain-derived neurotrophic factor in hippocampal neurons and the role of the PI3 kinase pathway in cell survival. *J Neurochem* 2004;**89**:844–52.
40 **Hurbin A**, Dubrez L, Coll JL, *et al.* Inhibition of apoptosis by amphiregulin via an insulin-like growth factor-1 receptor-dependent pathway in non-small cell lung cancer cell lines. *J Biol Chem* 2002;**277**:49127–33.
41 **Voelkel NF**, Cool CD. Pulmonary vascular involvement in chronic obstructive pulmonary disease. *Eur Respir J* 2003;**46**(Suppl):28–32s.
42 **Kasahara Y**, Tuder RM, Taraseviciene-Stewart L, *et al.* Inhibition of VEGF receptors causes lung cell apoptosis and emphysema. *J Clin Invest* 2000;**106**:1311–9.
43 **Kanazawa H**, Hirata K, Yoshikawa J. Imbalance between vascular endothelial growth factor and endostatin in emphysema. *Eur Respir J* 2003;**22**:609–12.
44 **Kanazawa H**, Asai K, Hirata K, *et al.* Possible effects of vascular endothelial growth factor in the pathogenesis of chronic obstructive pulmonary disease. *Am J Med* 2003;**114**:354–8.
45 **Kranenburg AR**, de Boer WI, Alagappan VK, *et al.* Enhanced bronchial expression of vascular endothelial growth factor and receptors (Flk-1 and Flt-1) in patients with chronic obstructive pulmonary disease. *Thorax* 2005;**60**:106–13.
46 **Meyer KC**, Cardoni A, Xiang ZZ. Vascular endothelial growth factor in bronchoalveolar lavage from normal subjects and patients with diffuse parenchymal lung disease. *Lab Clin Med* 2000;**135**:332–8.
47 **Tachimoto H**, Kikuchi M, Hudson SA, *et al.* Eotaxin-2 alters eosinophil integrin function via mitogen-activated protein kinases. *Am J Respir Cell Mol Biol* 2002;**26**:645–9.
48 **Zhu J**, Qui YS, Majumdar S, *et al.* Exacerbations of bronchitis: bronchial eosinophilia and gene expression for interleukin-4, interleukin-5, and eosinophil chemoattractants. *Am J Respir Crit Care Med* 2001;**164**:109–16.
49 **Grumelli S**, Corry DB, Song LZ, *et al.* An immune basis for lung parenchymal destruction in chronic obstructive pulmonary disease and emphysema. *PLoS Med* 2004;**1**:e8.
50 **Ritchie H**, Booth NA. The distribution of the secreted and intracellular forms of plasminogen activator inhibitor 2 (PAI-2) in human peripheral blood monocytes is modulated by serum. *Thromb Haemost* 1998;**79**:813–7.
51 **Kucharewicz I**, Kowal K, Buczko W, *et al.* The plasmin system in airway remodeling. *Thromb Res* 2003;**112**:1–7.
52 **DeMeo DL**, Mariani TJ, Lange C, *et al.* The SERPINE2 gene is associated with chronic obstructive pulmonary disease. *Am J Hum Genet* 2006;**78**:253–64.
53 **Corbacho AM**, Valacchi G, Kubala L, *et al.* Tissue-specific gene expression of prolactin receptor in the acute-phase response induced by lipopolysaccharides. *Am J Physiol Endocrinol Metab* 2004;**287**:750–7.
54 **Macotela Y**, Mendoza C, Corbacho AM, *et al.* 16K prolactin induces NF-kappaB activation in pulmonary fibroblasts. *J Endocrinol* 2002;**175**:13–8.

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.