RESEARCH ARTICLE

Identification of predictive cytokine biomarkers of scleroderma via local causal neighborhood methods [version 1; peer review: 1 approved with reservations, 1 not approved]

Ali Shojaee Bakhtiari ¹, Galina S. Bogatkevich², Alexander V. Alekseyenko ^{1,3}

¹Biomedical Informatics Center, Department for Public Health Sciences, Medical University of South Carolina, Charleston, South Carolina, 29425, USA

²Division of Rheumatology and Immunology, Department of Medicine, Medical University of South Carolina, Charleston, South Carolina, 29425, USA

³Department of Oral Health Sciences, Medical University of South Carolina, Charleston, South Carolina., 29425, USA

V1 First published: 24 Oct 2017, 6:1875
<https://doi.org/10.12688/f1000research.12563.1>
 Latest published: 24 Oct 2017, 6:1875
<https://doi.org/10.12688/f1000research.12563.1>

Abstract

Background: Scleroderma is an autoimmune disease with established relationship between immune cytokines and prognosis. Therefore, it is necessary to identify and investigate the causal relationship between cytokines and scleroderma diagnosis and to use this information to identify predictive biomarkers of scleroderma status.

Methods: Forty scleroderma positive patients and twenty-four healthy controls have been included in this study. Twenty-nine cytokines implicated in scleroderma have been measured in the bronchoalveolar lavage fluid of these patients, and eight have been found to be univariately associated with scleroderma status.

Results: Using local causal neighborhood learning methods, we have found two cytokines, Osteoprotegerin (OPG), also known as osteoclast genesis inhibitory factor, or tumor necrosis factor receptor superfamily member 11B and macrophage inflammatory protein-1 delta, to be causally related to the scleroderma status. Logistic regression predictor based on these cytokines achieves 73% AUC for the task of identifying the scleroderma status.

Conclusions: Our results demonstrate the feasibility of developing predictive local causal neighborhood biomarkers of scleroderma status based on bronchoalveolar lavage fluid.

Keywords

Local causal neighborhood, Predictive modeling, Scleroderma, Cytokine biomarkers.

Open Peer Review

Approval Status

	1	2
version 1		
24 Oct 2017	view	view
1. Mark M Wurfel , University of Washington, Seattle, USA 2. Kathryn Torok , University of Pittsburgh, Pittsburgh, USA		
Any reports and responses or comments on the article can be found at the end of the article.		

Corresponding author: Ali Shojaee Bakhtiari (ali.shojaee.bakhtiari@gmail.com)

Author roles: **Shojaee Bakhtiari A:** Conceptualization, Data Curation, Formal Analysis, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **S. Bogatkevich G:** Data Curation, Funding Acquisition, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Alekseyenko AV:** Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Methodology, Resources, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: The authors declare no competing interests.

Grant information: The project described was supported by the NIH National Center for Advancing Translational Sciences (NCATS) through Grant Number UL1 TR001450. AVA is funded by NIH/NCI R01 CA164964, NIH/NIDCR R34 DE025085, and NIH/NIAMS R21 AR067459. AVA and ASB are funded by MUSC College of Medicine Enhancing Team Science (COMETS) Pilot. GSB is funded by NIH/NIAMS P60 AR062755.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2017 Shojaee Bakhtiari A *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data associated with the article are available under the terms of the [Creative Commons Zero "No rights reserved" data waiver](#) (CC0 1.0 Public domain dedication).

How to cite this article: Shojaee Bakhtiari A, S. Bogatkevich G and Alekseyenko AV. **Identification of predictive cytokine biomarkers of scleroderma via local causal neighborhood methods [version 1; peer review: 1 approved with reservations, 1 not approved]** F1000Research 2017, 6:1875 <https://doi.org/10.12688/f1000research.12563.1>

First published: 24 Oct 2017, 6:1875 <https://doi.org/10.12688/f1000research.12563.1>

Introduction

Scleroderma, also known as systemic sclerosis (SSc), is an autoimmune disease of the connective tissues. The disease is characterized by accumulation of collagen and other connective tissue macromolecules in skin as well as internal organs. The disease is not considered heritable, although it is argued that genetic predisposition plays an important role in its development¹.

The mechanisms that cause interstitial lung disease in scleroderma (SSc-ILD) remain poorly understood. It is well documented that fibrosis is associated with increased expression of the profibrotic cytokines and chemokines. Several cytokines such as transforming growth factor (TGF)², connective tissue growth factor (CTGF)³, interleukins, tumor necrosis factor (TNF)⁴, oncostatin M (OSM)⁵ and others have been reported to be increased in bronchoalveolar lavage fluid (BALF) or serum from scleroderma patients when compared to healthy controls. It has been postulated that production of profibrotic cytokines and chemokines is likely to be key components in the pathology that leads to progressive pulmonary fibrosis in scleroderma⁶.

It has been suggested that T-helper (Th) cells and their associated cytokines may play a role in the pathophysiology of scleroderma⁷. Different subsets of Th cells play prominent roles in the progression of scleroderma⁷. Although heavily implicated in disease pathogenesis, immunoproteomic biomarkers do not currently provide an effective way to predict SSc-ILD⁸.

In this paper, we use computational local causal neighborhood (LCN) discovery techniques^{9,10} to identify the cytokines that have direct causal relationship to scleroderma status. Rather than reconstructing a detailed causal graph, LCN methods identify variables (cytokines) that are causally 'close' to a target variable (SSc-ILD). Based on theoretical results and empirical studies, this task can be achieved even from observational data under very common assumptions about the distribution of the data^{9,10}. The causal nature of the identified variables is complemented by provably advantageous properties with respect to the predictive value of these variables. Specifically, the set of biomarkers in the LCN of a target node make all other biomarkers conditionally independent. This implies that LCN yield the most parsimonious yet maximally predictive diagnostic biomarkers. This has been demonstrated in a number of empirical studies with real biological data^{11,12}. Therefore, we seek to develop a biomarker of SSc-ILD by discovering its LCN in cytokine data measured from the bronchoalveolar lavage fluid (BALF).

Materials and methods

Study cohort

Forty patients with SSc-ILD (20 African American: 5 males, 15 females, mean age 43.7 ± 11.0 ; 20 Caucasian: 10 males, 10 females, mean age 53.8 ± 11.6) and 24 healthy subjects (12 African American: 5 males, 7 females, mean age 32.0 ± 9.5 ; 12 Caucasian: 4 males, 8 females, mean age 29.6 ± 9.9), all nonsmokers, were examined.

BALF Preparation. Bronchoalveolar lavage was performed as a part of standard care after informed consent was obtained under

a protocol approved by the Institutional Review Board for Human Research of Medical University of South Carolina. Recovered BALF was centrifuged at 500 g for 10 minutes at 4°C. Pellet was removed and subjected to total and differential cell counts. Supernatant was dialyzed against sterile distilled H₂O overnight at 4°C, lyophilized, and stored at -80°C until assayed.

Cytokine Array. Lyophilized BALF powder was recovered in 5 mM Tris, pH 7.4 to a protein concentration of 1 mg/ml, and analyzed using Human Cytokine Antibody Array V from RayBiotech, Inc. (Norcross, GA) according to the manufacturer's instructions. Briefly, 500 µg of BALF samples were incubated with array support at room temperature for 2 h followed by the incubation with cocktail of biotin conjugated antibodies at 4°C overnight. The arrays were then incubated with horseradish peroxidase-conjugated streptavidin at room temperature for 2 h, and developed by using enhanced chemiluminescence-type solution. The images were scanned and analyzed with the NIH Image software. The net optical density (OD) was obtained by subtracting a background measurement of the same size negative area from the OD measurement of the signal area. Positive controls (six per membrane) were used to normalize the results from different membranes being compared. The variation between two identical cytokine spots ranged from 0 to 10% in duplicated experiments.

The final dataset consisted of 28 measured cytokines. [Figure 1](#) shows the description of the cytokine array map.

Data handling and preparation

The raw cytokine measurements have been log transformed. We have imputed the missing cytokine values from their 10 nearest neighbors identified by K-nearest neighbors (KNN) algorithm using Bioconductor package *impute*, version 3.4¹³. Adjustment for sex and race have been performed by fitting linear regression model to individual cytokine values and extracting the residuals for further analysis. All analyses have been performed in R, version 3.2.5¹⁴.

Univariate analyses

We have performed Welch t-test¹⁵ to find cytokines univariately associated with scleroderma status. We have adjusted for multiple comparisons using False Discovery Rate correction (FDR)¹⁶ with significance threshold set at 0.05. [Table 2](#).

Local causal neighborhood biomarker selection

We define LCN of a variable as a set of variables in its vicinity. LCNs can be constructed in such a way as to maximize their utility for biomarker development. For instance, Markov blanket (MB) of a variable is an LCN defined as the most compact set of variables (cytokines) rendering other variables independent of a target variable (SS)¹⁷. Intuitively, MB is the optimal solution of variable selection problem for predictive purposes. In practice, this means that the selected biomarkers contain all essential predictive information about the target node and are causally close to it (causal parents, children or spouses). Inference of MB requires inference of causal directionality, which is infeasible in most biomedical datasets due to their limited sample size and complexity. A closely related LCN that preserves much of the predictive utility of MB

	A	B	C	D	E	F	G	H	I	J	K
1	Pos	Pos	Pos	Pos	Neg	Neg	ENA-78	GCSF	GM-CSF	GRO	GRO- α
2	I-309	IL-1 α	IL-1 β	IL-2	IL-3	IL-4	IL-5	IL-6	IL-7	IL-8	IL-10
3	IL-12	IL-13	IL-15	IFN- γ	MCP-1	MCP-2	MCP-3	MCSF	MDC	MIG	MIP-1 β
4	MIP-18	RANTES	SCF	SDF-1	TARC	TGF- β 1	TNF- α	TNF- β	EGF	IGF-I	Ang
5	OSM	TPO	VEGF	PDGF-B	Leptin	BDNF	BLC	Ck β 8-1	Eotaxin	Eotaxin-2	Eotaxin-3
6	FGF-4	FGF-6	FGF-7	FGF-9	Flt-3 ligand	Fractalkine	GCP-2	GDNF	HGF	IGFBP-1	IGFBP-2
7	IGFBP-3	IGFBP-4	IL-16	IP-10	LIF	Light	MCP-4	MIF	MIP-3a	NAP-2	NT-3
8	NT-4	Ost	PARC	PIGF	TGF- β 2	TGF- β 3	TIMP-1	TIMP-2	Neg	Pos	Pos

Figure 1. RayBiotech human cytokine array map. Pos, positive; Neg, negative; ENA, epithelial cell-derived neutrophil-activating peptide; GCSF, granulocyte-colony stimulating factor; GM-CSF, granulocyte/macrophage-colony stimulating factor; GRO, growth-regulated oncogene; IL, interleukin; IFN, interferon; MCP, monocyte chemotactic protein; MCSF, macrophage colony stimulating factor; MDC, macrophage-derived chemokine; MIG, monokine induced by gamma interferon; MIP, macrophage inflammatory protein; RANTES, regulated upon activation in normal T cells, expressed, and secreted; SCF, stem cell factor; SDF, stromal cell-derived factor; TARC, thymus and activation regulated chemokine; TNF, tumor necrosis factor; EGF, epidermal growth factor; IGF, insulin-like growth factor; Ang, angiogenin; OSM, oncostatin; TPO, thrombopoietin; VEGF, vascular endothelial growth factor; PDGF, platelet-derived growth factor; BDNF, brain-derived neurotrophic factor; BLC, B-lymphocyte chemoattractant; IGFBP, insulin-like growth factor binding protein; IP-10, interferon-inducible protein 10; LIF, leukemia inhibitory factor; LIGHT, lymphotoxins, inducible expression, competes with HSV glycoprotein for HVEM, a receptor expressed on T-lymphocytes; MIF, macrophage migration inhibitory factor; NAP, neutrophil activating peptide; NT, neurotrophin; PARC, pulmonary and activation regulatory chemokine; PIGF, placenta growth factor; TIMP, tissue inhibitors of metalloproteinases.

is the parent-child set (PC-set), which only consists of direct causal parents and children of the target node. As opposed to MB the PC-set does not include common confounders of the effects of the target node, and thus can be slightly less predictive. Nonetheless, PC-sets are much easier to discover computationally. Overall, the major advantage LCNs for biomarker discovery is that the size of the biomarker is typically small and has a causal interpretation.

In order to infer LCN of SS, we have used the HITON-PC algorithm¹⁸ from causal explorer toolbox¹⁹ in MATLAB Release 2016b²⁰. This algorithm performs a series of conditional independence tests (Fisher's Z test with 0.05 significance threshold) to infer the PC-set of the target variable (SS).

Development of a predictive model based on the local causal neighborhood

We use logistic regression model for predictive analysis. To ensure accurate and unbiased estimates of predictive power of the local causal neighborhood, we have performed 1,000 cross validation (CV) runs of the predictive analysis. First, we have randomly split the sample into training (70% of the data) and testing sets preserving case-control balance in each of the splits. Next, we have identified the PC-set of SSc-ILD using HITON-PC algorithm. The cytokines in the PC-set served as the predictors in building the logistic regression model of SSc-ILD from the training set data. Finally, we estimated the performance of the logistic regression model on the data in the test set. The performance estimates from each of the 1,000 CV runs have been used to determine the overall predictive power of the PC-set.

Our main metric of predictive performance is area under receiver operating characteristic (ROC) curve (AUC). Curve specifies the relationship between the true positive rate (TPR) and the false positive rate (FPR) of the model under all possible model parameter thresholds. The value of AUC from 0.5 (no predictive signal) to 1 (perfect prediction). We have extracted the following additional performance metrics from the ROC: (i) *mean classification success rate*, defined as the average rate of correctly identifying the scleroderma status on the test set, over the entire simulation runs; (ii) *mean optimal sensitivity*, the average value of the optimal sensitivity of the predictive model over the entire simulation runs. (ii) *mean optimal specificity*, the average of the optimal specificity of the predictive model over the entire simulation runs.

Results

Cytokine array

Cytokine arrays were utilized to explore and compare the expression levels of multiple cytokines in BALF samples. The array revealed elevated expression of 28 cytokines in BALF from scleroderma patients when compared to controls, which were selected for further analysis. [Figure 2](#).

Scleroderma status is univariately associated with 8 out of 28 cytokines measured in BALF

From the 28 cytokines measured, 8 are significantly associated with scleroderma status at p-value of 0.05 ([Table 1](#)). After correction for multiple comparisons using FDR, 3 cytokines, MIP-1delta, VEGF and Osteoprotegerin, remain significantly associated. The same cytokines are significant after adjustment for subject sex and race.

Figure 2. A comparison of RayBiotech human cytokine array map for different. Select cytokine expression locations are shown by arrows. In this instance Caucasian and African American controls are brought against Caucasian and African American with positive scleroderma status.

Table 1. Demographic information of the cohort.

	Cases (N=40)		Controls (N=24)	
Sex	Male	Female	Male	Female
	15	25	9	15
Race	African American	White	African American	White
	20	20	12	12

Table 2. Results of univariate screening of cytokines for association with SS.

	Mean		Unadjusted		Adjusted	
	Healthy(sd)	Scleroderma(sd)	P-value*	FDR	P-value*	FDR
MCP.1	2.002(1.85)	3.010(1.48)	0.029	0.110	0.028	0.107
MIP.1delta	2.887(1.34)	3.833(0.74)	0.003	0.030	0.002	0.022
SCF	1.184(1.14)	1.900(0.85)	0.011	0.075	0.010	0.067
ANG	3.934(0.79)	4.390(0.50)	0.016	0.080	0.015	0.077
TPO	2.286(1.03)	2.863(1.07)	0.038	0.128	0.035	0.120
VEGF	3.223(0.46)	3.627(0.47)	0.001	0.018	0.000	0.006
Osteoprot	2.439(0.91)	3.186(0.57)	0.001	0.018	0.000	0.006
PARC	4.130(0.87)	4.588(0.24)	0.018	0.080	0.017	0.077

* Welch t-test, p-values significant at 0.05 are shown in bold.

Local causal neighborhood of SSc-ILD consists of OPG and MIP1delta

HITON-PC algorithm identifies two cytokines, OPG and MIP-1delta, as the members of the PC-set LCN of SS in both adjusted and unadjusted data. [Figure 3](#) shows the inferred LCN for the scleroderma status.

In order to assess the stability of the PC-Set estimate, we performed a re-sampling analysis as described in the methods section. The histogram of the appearance of the different cytokines in the causal neighborhood of scleroderma status is shown in [Figure 4](#). Osteoprot and MIP-1delta appeared most frequently in the PC-set LCN of SS. In the 1000 simulation runs, either one or both cytokines appeared in the PC-set of scleroderma status 847 times. The scatterplot of MIP1-delta against Osteoprot adjusted values ([Figure 5](#)) shows apparent concentration of scleroderma cases in the top right quadrant, indicating association of higher expression values of these cytokines with disease.

LCN cytokines predict SS with 73% AUC

We first fit the logistic regression model using all cytokines as predictors. The resulted model attains cross-validated predictivity indistinguishable from random predictor on the testing data and almost perfect predictivity in the training data. This indicates high levels of noise in the weakly predictive data, which requires feature selection before attempting to build predictive model in order to obtain accurate models and estimates of model predictivity. To estimate the predictive utility of PC-set LCN in an unbiased way, we have performed 1,000 iterations of repeated cross-validated predictive model building using logistic regression. In data adjusted for sex and race, the mean classification success rate is 68.46% (st. dev. 6.66%). The mean optimal sensitivity of the model is 65.49% (st. dev. 11.38%). The mean optimal specificity of the model is 73.44% ($\pm 11.88\%$ st. dev.). We estimate the AUC of the cross-validated model to be 73% ([Figure 6](#)), indicating non-negligible power of LCN biomarker for SSc prediction.

Figure 3. Local causal neighborhood of the scleroderma status. HITON-PC algorithm identified two cytokines, Osteoprot and MIP-1delta, to be in the PC-set of the scleroderma status (SS). 6 other cytokines are found to be univariately associated with SS. These additional cytokines as well as every other cytokine not in PC-set, however, are conditionally independent of the SS, given the PC-set. The causal connectivity of the remaining cytokines with SS is unknown.

Figure 4. Osteoprot and MIP-1delta LCN biomarkers are robust to repeated cross-validation. We performed 1,000 iterations of repeated cross validation to determine stability of the two cytokine biomarkers. The number of times each cytokine (adjusted for sex and race) appears in the PC-set LCN of SS is depicted on this plot. Osteoprot and MIP-1delta appear in the LCN of SS with high frequency, indicating high likelihood of their causal proximity to scleroderma status. These results are similar if no adjustments for sex and race are made (data not shown).

Figure 5. Scatterplot of Osteoprot vs. MIP-1delta. The scatterplot shows the distribution of the two cytokines against each other in data adjusted by sex and race. Red and black points indicate cases and controls, respectively. The concentration of the red points in the upper right quadrant indicates potential presence of predictive signal towards disease status in these cytokines.

Figure 6. The receiver operating characteristic curve of the logistic regression model for SS prediction with sex and race adjusted cytokines. The ROC curve shows the estimated values of true positive rate against the false positive rate for the predictive model. The estimated value for the area under curve is 73%. The significant difference of the AUC value from 0.5 confirms the strength of the identified biomarkers in predicting the SS.

Discussion

ILD is the leading cause of morbidity and mortality in scleroderma patients. The mechanisms leading to SSc lung disease remain unknown. However, a variety of cytokines and growth factors have been reported to be increased in BALF from SSc patients. It is known fact that, Osteoprotegerin acts as a competitive receptor for the membrane-bound receptor activator of nuclear factor-kappaB²¹. It is also known that, osteoprotegerin-deficient mice exhibit osteoporosis at early stage of life span with the activation of RANK ligand (RANKL)²². Macrophage inflammatory protein (MIP) has been previously shown to be associated with scleroderma diagnosis²³. OPG in primary myelofibrosis (PMF) expressed significantly higher (up to 71-fold) when compared with prefibrotic cellular PMF and control cases²⁴.

In this paper, we present a model for inferring the cytokines predictive of SSc-ILD using local causal neighborhood framework. The model identified osteoprotegerin (OPG) and macrophage inflammatory protein-1 delta (MIP-1delta), to be closely related to the scleroderma status. We tested the hypothesis that the identified cytokines can be used as predictive biomarkers.

We compared the biomarkers selected by LCN with those resulting from LASSO regression²⁵. In our data, the regression model selects the same features as LCN. This suggests that LCN approach can result in equivalent inference to LASSO. However, the major advantage of an LCN feature selection approach is in the theoretical causal guarantees it provides. LCN identifies features based on their causal proximity to the target variable, something that LASSO regression feature selection cannot match.

In this study, clinical disease characterization and demographic information, other than sex and race, have not been incorporated in the predictive model of SS. In practice, we maintain that combining molecular biomarkers with clinical and demographic data should improve the quality of the predictor. Such a biomarker, however, will require a larger cohort of patients to be developed.

Understanding the role of BALF's cytokines in pathogenesis of SSc-ILD may identify new targets for the development of diagnostic biomarkers predicting the biological behavior of the disease.

Data and code availability

ZENODO: *SLE_cytokines - Identification of predictive cytokine biomarkers of scleroderma via local causal neighborhood methods data*

Raw and imputed cytokine array expressions, relevant code provided in Matlab and R format, and the results of applying LASSO regression on the BALF cytokine arrays are provided in the repository. <https://zenodo.org/record/1001044>; DOI: [10.5281/zenodo.1001044](https://doi.org/10.5281/zenodo.1001044)²⁶

Ethics and consent

All procedures were performed as a part of patient standard care after informed consent was obtained under a protocol approved by the Institutional Review Board for Human Research of Medical University of South Carolina.

Competing interests

The authors declare no competing interests.

Grant information

The project described was supported by the NIH National Center for Advancing Translational Sciences (NCATS) through Grant Number UL1 TR001450. AVA is funded by NIH/NCI R01 CA164964, NIH/NIDCR R34 DE025085, and NIH/NIAMS R21 AR067459. AVA and ASB are funded by MUSC College of Medicine Enhancing Team Science (COMETS) Pilot. GSB is funded by NIH/NIAMS P60 AR062755.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

- Assassi S, Radstake TR, Mayes MD, *et al.*: **Genetics of scleroderma: implications for personalized medicine?** *BMC Med.* 2013; **11**(1): 9. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Roberts AB, Sporn MB, Assoian RK, *et al.*: **Transforming growth factor type beta: rapid induction of fibrosis and angiogenesis in vivo and stimulation of collagen formation in vitro.** *Proc Natl Acad Sci U S A.* 1986; **83**(12): 4167–71. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Moussad EE, Brigstock DR: **Connective tissue growth factor: what's in a name?** *Mol Genet Metab.* 2000; **71**(1–2): 276–92. [PubMed Abstract](#) | [Publisher Full Text](#)
- Leithäuser F, Dhein J, Mechtersheimer G, *et al.*: **Constitutive and induced expression of APO-1, a new member of the nerve growth factor/tumor necrosis factor receptor superfamily, in normal and neoplastic cells.** *Lab Invest.* 1993; **69**(4): 415–29. [PubMed Abstract](#)
- Gearing DP, Bruce AG: **Oncostatin M binds the high-affinity leukemia inhibitory factor receptor.** *New Biol.* 1992; **4**(1): 61–5. [PubMed Abstract](#)
- Atamas SP, White B: **Cytokine regulation of pulmonary fibrosis in scleroderma.** *Cytokine Growth Factor Rev.* 2003; **14**(6): 537–50. [PubMed Abstract](#) | [Publisher Full Text](#)
- Kurzinski K, Torok KS: **Cytokine profiles in localized scleroderma and relationship to clinical features.** *Cytokine.* 2011; **55**(2): 157–64. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Castro SV, Jimenez SA: **Biomarkers in systemic sclerosis.** *Biomark Med.* 2010; **4**(1): 133–147. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Aliferis CF, Statnikov A, Tsamardinos I, *et al.*: **Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part I: Algorithms and Empirical Evaluation.** *J Mach Learn Res.* 2010; **11**: 171–234. [Reference Source](#)
- Aliferis CF, Statnikov A, Tsamardinos I, *et al.*: **Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part II: Analysis and Extensions.** *J Mach Learn Res.* 2010; **11**: 235–284. [Reference Source](#)
- Statnikov A, Alekseyenko AV, Li Z, *et al.*: **Microbiomic signatures of psoriasis: feasibility and methodology comparison.** *Sci Rep.* 2013; **3**: 2620. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Statnikov A, Henaff M, Narenda V, *et al.*: **A comprehensive evaluation of multicategory classification methods for microbiomic data.** *Microbiome.* 2013; **1**(1): 11. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Hastie T, Tibshirani R, Narasimhan B, *et al.*: **Impute: Imputation for microarray data.** R package version 1.48.0. 2016. [Reference Source](#)
- R Core Team: **R: A Language and Environment for Statistical Computing.** 2015. [Reference Source](#)
- Welch BL: **The generalisation of student's problems when several different population variances are involved.** *Biometrika.* 1947; **34**(1–2): 28–35. [PubMed Abstract](#) | [Publisher Full Text](#)
- Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.** *J R Stat Soc Series B (Methodol).* 1995; **57**(1): 289–300. [Reference Source](#)
- Pearl J: **Probabilistic reasoning in intelligent systems: networks of plausible inference.** Morgan Kaufmann Publishers Inc. 1988; 552. [Reference Source](#)
- Aliferis CF, Tsamardinos I, Statnikov A: **HITON: a novel Markov Blanket algorithm for optimal variable selection.** *AMIA Annu Symp Proc.* 2003; **2003**: 21–25. [PubMed Abstract](#) | [Free Full Text](#)
- Aliferis CF, Statnikov AR, Tsamardinos I, *et al.*: **Causal explorer: A causal probabilistic network learning toolkit for biomedical discovery.** In *International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS'03)*. 2003. [Reference Source](#)
- MATLAB version 2016b, Natick, Massachusetts: The MathWorks Inc.
- Coetzee M, Kruger MC: **Osteoprotegerin-receptor activator of nuclear factor-kappaB ligand ratio: a new approach to osteoporosis treatment?** *South Med J.* 2004; **97**(5): 506–11. [PubMed Abstract](#)
- Bucay N, Sarosi I, Dunstan CR, *et al.*: **osteoprotegerin-deficient mice develop early onset osteoporosis and arterial calcification.** *Genes Dev.* 1998; **12**(9): 1260–8. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Scala E, Pallotta S, Frezzolini A, *et al.*: **Cytokine and chemokine levels in systemic sclerosis: relationship with cutaneous and internal organ involvement.** *Clin Exp Immunol.* 2004; **138**(3): 540–6. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Kreipe H, Büsche G, Bock O, *et al.*: **Myelofibrosis: molecular and cell biological aspects.** *Fibrogenesis Tissue Repair.* 2012; **5**(Suppl 1): S21. [PubMed Abstract](#) | [Free Full Text](#)
- Tibshirani R: **Regression Shrinkage and Selection via the Lasso.** *J R Stat Soc Series B (Methodol).* 1996; **58**(1): 267–288. [Reference Source](#)
- Bakhtiari AS, *et al.*: **ashojee/SLE_cytokines: Identification of predictive cytokine biomarkers of scleroderma via local causal neighborhood methods data.** *Zenodo.* 2017. [Publisher Full Text](#)

Open Peer Review

Current Peer Review Status:

Version 1

Reviewer Report 11 December 2017

<https://doi.org/10.5256/f1000research.13606.r28122>

© 2017 Torok K. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Kathryn Torok

Children's Hospital of Pittsburgh of UPMC, University of Pittsburgh, Pittsburgh, PA, USA

The authors present an interesting paper regarding the study of cytokines derived from BALF which offers the possibility to study local disease mechanisms and potential biomarkers associated with disease propagation. Most of the paper and references were dedicated to the local casual neighborhood (LCN) biostatistical approach. My main comments to develop this paper further would be to 1) display the directionality and interaction of cytokines on a correlation matrix or cluster graphic to allow for more insight regarding the relationship of these cytokines/insight into pathways, and 2) expand the clinical information about the systemic sclerosis subjects to assist the reader in understanding the generalizability to the patient population they are treating.

A few constructive comments to address these concerns:

1. Regarding citations, this institution has published before regarding certain cellular subtypes and cytokines determined in SSc BALF. It would be a strength to include these and also potentially tie into the discussion if similar or different cytokines were determined in these current analyses. Particularly, MIP-1 alpha (not MIP-1 delta as in current paper), was found to associate with degree of alveolitis.

Cytokine concentrations in bronchoalveolar lavage fluid of patients with systemic sclerosis. Arthritis Rheum. 1997 Apr;40(4):743-51.
Bolster MB1, Ludwicka A, Sutherland SE, Strange C, Silver RM.

2. In a similar framework, these prior publications have included more SSc patient clinical data, which would be helpful to understand the subset of patients analysed and the generalizability to clinical practice. I would recommend including featured such as disease subtype (dcSSc, lcSSc), disease duration, degree of lung fibrosis or CT changes if quantified, average FVC, DLCO etc if available.
2a. In the discussion it was mentioned these were not displayed due to lack of numbers for

subanalyses, which is understandable, but the general summary of these characteristics would be helpful if added to Table 1.

3. Presenting the general cell types derived from the BALF would be helpful between SSc and controls. This may give more insight in context with cytokines found more predominately in SSc BALF.
4. Figure 4 displays the frequency of cytokines generated with repeated analyses, although MIP-1delta and OPG are the most frequent, PARC and VEGF are not close behind. A Figure showing the Causal modeling PC algorithm for cytokines in proximity and direction to the target variable (SSc disease status) would be very instructional here. One would think PARC and VEGF might be close enough to include in the ROC curve combining these 4 cytokines. Revising Figure 3 to display directionality from the casual modeling, which is the strength of utilizing this statistical approach, would be advisable to help understand the narrow focus of just the 2 cytokines.

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Partly

If applicable, is the statistical analysis and its interpretation appropriate?

Partly

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Partly

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 01 December 2017

<https://doi.org/10.5256/f1000research.13606.r27316>

© 2017 Wurfel M. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Mark M Wurfel

Department of Medicine, University of Washington, Seattle, WA, USA

This is an interesting report that addresses an important clinical problem. However, the methods and presentation potentially obscure what the key findings from these data might be. The following should be addressed.

1. There is very little data presented describing the study subjects. They authors should present more demographic information (age) and disease characteristics (time from diagnosis, organs involved, lung function, current therapy).
2. The rationale for using lyophilized BALF for the analyses is not presented nor are the potential challenges in interpretation addressed. For instance, the absolute numbers will not be useful for future diagnostics as it is highly unlikely that alveolar fluid would be lyophilized/concentrated before measurement. It would be helpful to know what the concentration of albumin or other background proteins might be to appreciate the relative abundance of the mediators measured.
3. The text refers to Table 1 as having results of the cytokine analyses but this appears in Table 2. Table 2 has "unadjusted" and "adjusted" analyses without any statement of how the "adjusted" analyses were performed. Given its exploratory nature I would like to see the entirety of the dataset presented here as it has more utility to the community than a multivariate model that is not validated and is likely to be highly over-fit.
4. The rationale for using LCN for these analyses is unclear, particularly given that LASSO appears to perform similarly and is more straightforward to interpret. Notably, there does not appear to be any comparison of LCN and LASSO presented even though it is mentioned in the discussion.
5. Figure 4 and Figure 5 are not very useful. I would like to see more details about the outputs from the LASSO model (i.e. what were the driving mediators and model performance measures like BIC or other measures of fit/overfit). It would also be helpful to readers to see some form of correlation matrix or cluster graphic to show how the different mediators track or don't track with each other. This will allow for more insights on pathways that are differentially up/down regulated.

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and is the work technically sound?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?

No

If applicable, is the statistical analysis and its interpretation appropriate?

Partly

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Partly

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research