Wednesday, August 30, 2017

ON LINE DATA SUPPLEMENT

European Respiratory Journal

REVISED VERSION

MULTI-LEVEL DIFFERENTIAL NETWORK ANALYSIS OF COPD EXACERBATIONS

Guillaume Noell[1,2*], Borja G Cosío[1,3*], Rosa Faner[1,2*],

Eduard Monsó[1,4], German Peces-Barba[5], Alfredo de Diego[6], Cristobal Esteban[7],

Joaquim Gea[1,8], Robert Rodriguez-Roisin[1,2,9], Marian Garcia-Nuñez[1,4],

Francisco Pozo-Rodriguez[1,10], Susana G. Kalko[2] and Alvar Agusti[1,2,9,11]

* Co-primary authors

On-line supplement text: 1,135 words; References: 8; Tables: 0; Figures: 0.

METHODS

Measurements

The following variables were determined both during ECOPD and at convalescence, except specifically indicated. Measurements were performed in each participating hospital following routine standardized procedures except for some specific determinations (as indicated) that were measured centrally.

Clinical data

These included demographics, previous medical history and co-morbidities, smoking history and status (current vs. former), body mass index (BMI) and degree of breathlessness using the modified Medical Research Council (mMRC) scale and a visual analog-scale. We also recorded standard vital signs, including respiratory and heart rates.

General Biochemistry

Measurements included hemoglobin level, erythrocyte sedimentation rate (ESR) and liver and kidney function tests, as determined routinely in participating hospitals.

Lung function

Lung function measurements included forced spirometry, static lung volumes, carbon monoxide diffusing capacity (DLCO), arterial blood gases and 6-minutes walking distance (6MWD), according to international recommendations [1]. Reference values were those of a Mediterranean population [2, 3].

Cardiovascular function

To assess cardiovascular function, we: *(1)* obtained an standard 12-lead electrocardiogram (EKG); *(2)* performed echocardiography (at ECOPD only); and, *(3)* quantified (in a central laboratory) the plasma levels of N-terminal prohormone of brain natriuretic peptide (BNP), fibrinogen, creatine phosphor-kinase (CPK) and troponin in peripheral venous blood samples using standard methodology.

Sputum microbiology

Spontaneous sputum was obtained and cultured for common bacterial species during ECOPD using standard methodology. The presence of respiratory viruses in sputum was assessed using the clinical arrays Pneumovir kit (CLART Pneumovir, Genomica, Spain). All PCR determinations were centralized at the Institut de Recerca en Ciències de la Salut Germans Trias i Pujol, Badalona, Spain. Bacterial and viral total nucleic acid (DNA/RNA) was extracted using the QIAamp Viral RNA Mini kit (Qiagen) from a starting volume of 200 µl with a 50 µl eluate.

Lung Imaging

These included chest x-ray at each visit and helical chest CT-scan at ECOPD to determine the presence/absence (qualitative assessment by two experienced radiologists at each center) of emphysema, bronchiectasis, alveolar infiltrates and/or pulmonary embolism by an experienced radiologist.

Inflammatory markers

The concentration of IL-8, IL-6, IL-1β, TNF-α, both in plasma and sputum supernatant were quantified at the Fundació Investigació Hospital General de Valencia, Spain, using a Luminex® assay and a modified flow cytometer (Luminex X100) and fluorescent-dyed microspheres (Fluorokine MAP human base kit A, R&D). TNF-SR, TGF-β, SAA, and Procalcitonin were determined (also centrally in Valencia) using a Zeptosens® assay following manufacturer instructions, like TAS which was measured using a colorimetric (Randox Laboratories Ltd, Crumlin, UK) following manufacturer recommendations.

Data analysis

Multi-level correlation network construction

Pairs of variables were included in the network according to the following criteria: *(1)* for two *continuous variables,* a Spearman correlation coefficient (Rho) >0.3 and a nominal p-value < 0.01; *(2)* for two *categorical variables*, Fisher p value <0.01; and, *(3)* for a *continuous and a categorical* variable, Mann-Whitney p value <0.01. Hubs in each network were defined as variables with a high Kleinberg score (from 0.8 to 1) [4].

Differential Network Analysis

The randomization procedure swaps iteratively and randomly the clinical condition of patient's data (ECOPD or convalescence) as part of a Monte Carlo sampling, then recalculates from the permuted data the absolute Rho differences between ECOPD and convalescence (test-statistic) and compares it with the original difference to get an estimation of the test-statistic associated p-value. We observed that 10,000 iterations were sufficient to accept or reject the null hypothesis (Rho ECOPD not statistically different from Rho convalescence) when filtering correlations with a FDR p-value threshold of 0.05 and minimum absolute correlation difference of 0.3: $|\text{Rho ECOPD} – \text{Rho convalescence}| > 0.3$. Edges that connect categorical variables were found rarely and, therefore, not included in the Monte Carlo randomization test.

ECOPD biomarkers

To identify the potential ECOPD biomarkers, we: *(1)* acknowledge that variable values determined at convalescence describe their distribution in clinically stable COPD, which may not necessarily be within the normal range observed in a healthy subject precisely because of the presence of COPD; *(2)* determined, for each variable, the proportion of ECOPD values which were above the 95^{th} or below the 5^{th} percentile of that same variable at convalescence (i.e., measured during clinical stability). These ECOPD *outlier* values were considered “abnormal” for a clinically stable COPD patient and, therefore, representative of an episode of ECOPD; *(3)* built a correlation network (as described above) that include only those ECOPD variables with a significant (FDR p value<0.5 by a Monte Carlo permutation test [5]) proportion of *outliers* above the 95^{th} or below the 5^{th} percentile. The former ($>95^{th}$) were graphically depicted as triangles pointing upwards, whereas the latter ($<5^{th}$) were identified by triangles pointing downwards, symbol size being proportional in each case to the proportion of outliers in each variable. Nodes in the network were linked by edges, whose thickness was proportional to the percentage of patients with shared outlier values of these two particular variables albeit, to facilitate graphic display, we decided to include in the network only those with ≥25% shared *outlier* values; and, *(4)* computed the receiver operating characteristic (ROC) profiles, and the corresponding area under the curve (AUC) values, of all potential diagnostic biomarkers (continuous variables) identified above using the R package ROCR [6], albeit we only considered those with an AUC>0.8, which we later used to build a generalized linear model to predict the status of the patient.

RESULTS

Microbiologic data during ECOPD

Identified bacteria included *Haemophilus influenzae* (38%) and *parainfluenzae* (30%), *Streptecocus pneumoniae* (24%), *Moraxella catharralis* (12%) and *Pseudomona aeruginosa* (6%). Two or more bacteria were present in 36% of samples, being *H. influenzae* or *parainfluenzae* present in all combinations. The most frequently virus found was *Rhinovirus* (7%). At least one bacteria and one virus were present in 21.3% of patients. Only 8.7% of virus traces showed no co-presence of bacteria, and only 6% of patients did not show any microorganisms both by sputum culture or PCR.

Description of modules during ECOPD and at convalescence

During ECOPD, MM1 was basically a biochemistry module since it contained the hemoglobin, urea, creatinine, erythrocyte sedimentation rate, fibrinogen and CRP nodes. Systemic inflammatory markers were mostly distributed across two clusters loosely defined: MM3 and MM4 with low modularity scores (0.093 and 0.041, respectively). MM2 was mainly composed of pulmonary gas exchange (PaO2, PaCO2, pH) and lung volumes variables, in addition to dyspnea, respiratory rate and presence of CT bronchiectasis. Finally, MM5 was basically composed by sputum inflammatory markers and circulating lymphocytes and neutrophils. The best defined modules according to the algorithm used were MM2 (score=0.321) and MM5 (score 0.312).

At convalescence, the network become denser (as described above) and six modules were identified by the fast-greedy community algorithm [7, 8]. MM6 contained biochemistry variables, similarly to MM1. MM7 was the only well-defined cluster at convalescence (score=0.279) and included most lung function variables. Inflammatory biomarkers were spread between MM8, MM9, MM10 and sputum variables were divided between MM9 and MM11 (Figure 2).

REFERENCES

1. Clausen JL. Pulmonary function testing. Guidelines and controversies. Equipment, Methods and Normal Values. 1 ed. Grune & Stratton, Inc., Orlando, 1984.

2. Roca J, Sanchis J, Agustí-Vidal A, Segarra J, Navajas D, Rodriguez-Roisín R, *et al.* Spirometric reference values for a mediterranean population. *Bull Eur Physiopathol Respir* 1986: 22: 217-224.

3. Roca J, Rodriguez-Roisín R, Cobo E, Burgos F, Perez J, Clausen JL. Single-breath carbon monoxide diffusing capacity (DL_{CO}) prediction equations for a mediterranean population. *Am Rev Respir Dis* 1990: 141: 1026-1032.

4. Kleinberg J. Authoritative sources in a hyperlinked environment. *Proc 9th ACM-SIAM Symposium on Discrete Algorithms* 1998.

5. Fishman GS. Monte Carlo: Concepts, Algorithms and Applications. Springer, New York, 1995.

6. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. *Bioinformatics* 2005: 21(20): 3940-3941.

7. Newman ME, Girvan M. Finding and evaluating community structure in networks. *Phys Rev E Stat Nonlin Soft Matter Phys* 2004: 69(2 Pt 2): 026113.

8. Su G, Kuchinsky A, Morris JH, States DJ, Meng F. GLay: community structure analysis of biological networks. *Bioinformatics* 2010: 26(24): 3135-3137.