ON LINE SUPPLEMENT

Surveillance for lower airway pathogens in mechanically ventilated patients by metabolomic analysis of exhaled breath: a case-control study

Stephen J Fowler*[1, 2], Maria Basanta-Sanchez*[1], Yun Xu[3], Royston Goodacre[3], Paul M Dark[1, 4]

*Both authors contributed equally to this work

[1]University of Manchester, Manchester Academic Health Science Centre, and NIHR Respiratory and Allergy Clinical Research Facility, University Hospital of South Manchester, Manchester, UK

[2] Lancashire Teaching Hospitals NHS Foundation Trust, Preston, UK

[3] School of Chemistry & Manchester Institute of Biotechnology, University of Manchester, Manchester, UK

[4] Salford Royal Hospitals NHS Foundation Trust, Salford, UK

Correspondence: Stephen J Fowler,

Education and Research Centre,

University Hospital of South Manchester,

Southmoor Road,

Manchester M23 9LT,

UK

Tel: +44 161 291 5864 or +44 1772 523237

Fax: +44 161 291 5730

Email: stephen.fowler@manchester.ac.uk

ANOVA-mean centre (ANOVA-MC); the model

ANOVA-MC (1) is an extension to principal component analysis (PCA) (2). The aim of ANOVA-MC is to fit the data obtained from a single factor, one-way ANOVA experiment design with the problem of having high between-subject variability and to allow discovery of the underlying pattern which relates to the experimental question(s). The PCA model decomposes the observed data matrix (e.g. the GC-MS data matrix obtained from breath VOC analysis) ***X*** into the product of a scores matrix ***T*** and a loadings matrix ***P***, with the unfitted error put into a residue matrix ***E*** as shown in the eq (1):

$$\boldsymbol{X} = \boldsymbol{T} \times \boldsymbol{P}^T + \boldsymbol{E} \qquad \text{eq(1)}$$

The pattern in the samples is presented in ***T*** while the contribution of the variables to such pattern is presented in ***P***.

In order to prevent T to be dominated by the high between-subject variations, ANOVA-MC adds a pre-processing step and uses the pre-processed data matrix $\boldsymbol{X}_{anova\_mc}$ instead of ***X*** subject to PCA, given in eq(2).

$$\begin{aligned}\mathbf{X}_{anova\_mc} &= \mathbf{X}_f + \boldsymbol{\varepsilon}_{anova\_mc} \\ &= \begin{bmatrix} \mathbf{1}_1 \cdot \mathbf{m}_1^T \\ \mathbf{1}_2 \cdot \mathbf{m}_2^T \\ \vdots \\ \mathbf{1}_c \cdot \mathbf{m}_c^T \end{bmatrix} + \boldsymbol{\varepsilon}_{anova\_mc}\end{aligned} \qquad \text{eq(6)}$$

In which $\boldsymbol{1}_i$ (i=1,2,…,c) is a column vector of 1s and the length of each vector equals the number of samples of each class; $\boldsymbol{m}_i^T$ (i =1,2,…,c) is a row vector which is the mean vector of all the samples of class i. The mean vectors are all calculated from mean centred ***X***. The residual matrix $\boldsymbol{\varepsilon}_{anova\_mc}$ is is obtained by firstly mean centring the original data matrix, then calculating the mean of each subject and subtracting them from the corresponding rows (samples) as shown in eq(7).

$$\boldsymbol{\varepsilon}_{anova\_mc} = \boldsymbol{X} - \boldsymbol{1} \cdot \boldsymbol{m}^T - \begin{bmatrix} \boldsymbol{1}_1 \cdot \boldsymbol{a}_1^T \\ \boldsymbol{1}_2 \cdot \boldsymbol{a}_2^T \\ \vdots \\ \boldsymbol{1}_s \cdot \boldsymbol{a}_s^T \end{bmatrix} \qquad \text{eq(7)}$$

In which $\boldsymbol{1}_j$ (j =1,2,…,s) is a column vector of ones and the length of the vector equals the number of samples of the test subject j; $\boldsymbol{a}_j^T$ is a row vector which is the mean vector of all the samples collected from the test subject j.

If there were no significant dynamic effect between the repeated measurements of the same subject, $\boldsymbol{\varepsilon}_{anova\_mc}$ is essentially the variation caused by experiment itself, e.g. sampling error, instrument measurement error etc. Analysing it together with the $\boldsymbol{X}_f$ is equivalent to superimposing the between-group difference onto the unavoidable variance introduced by the experiment and assessing the significance level of the between groups variance. $\boldsymbol{X}_f$ could either be added to the residual matrix $\boldsymbol{\varepsilon}_{anova\_mc}$ back and then have PCA performed on it, or subjected to decomposition directly to obtain the loadings first, then $\boldsymbol{\varepsilon}_{anova\_mc}$ added back and projected into the subspace via the loadings. In this study, we employed the ANOVA-MC by using the approach adding $\boldsymbol{\varepsilon}_{anova\_mc}$ back to $\boldsymbol{X}_f$.

Validation procedure for ANOVA-MC

In ANOVA-MC, the labelling information about which samples belong to which group has been used when performing the localised mean centring. Therefore, the mean matrix of interest itself ($\boldsymbol{X}_f$) has become a latent factor which would cluster samples into expected groups according to their class labels, thus there is no question about whether the samples from different groups could be separated from each other in the PCA model applied to such testing matrix. The question arises as to whether such separation is statistically significant when comparing it to the background variations, i.e. the residue matrix and, more importantly, the chance that such separation shown in the PCA

model had been a false discovery. We employed a validation procedure based on random re-sampling and permutation. We firstly assumeds that there were *c* known classes and the results of ANOVA-MC had showed a clear separation between these classes which matched the class labels well, and that *k* PCs are required to separate all the known classes (ideally *k* should be no greater than *c* - 1). The aim of the validation is to assess the reproducibility of the separation between the groups. This is achieved by randomly re-sampling the data sets *R* times to generate *R* different subsets of the data. In this study we employed bootstrap re-sampling strategy (3) and performed 1,000 iterations. Each subset of the data was then analysed by the method to be validated. The *K*-means clustering analysis (4) was then performed on the final PCA results using a sufficient number of PCs which were able to separate the classes. The number of clusters was set to *c* and the initial cluster centroid positions set to be the mean of each class, calculated from the subset of samples using the known group labels. This way the clusters identified by the *K*-means clustering should have a 1-to-1 correspondence with the expected classes. A pattern with the known classes well separated from each other would be expected to see a high consistency between the known class labels and the labels identified by *K*-means clustering. Such high consistency should also be reproducible for the models which were built on different subsets of samples. By contrast, if the observed separation was caused by chance or there was no genuine separation the results of the *K*-means clustering would be rather unpredictable. If there was no true underlying difference between the expected classes, the PCA scores obtained would be expected to be a homogeneous mixture and the clusters identified by the *K*-means clustering merely arbitrary collections of samples, depending on the relative distance between them, and there should be little to no agreement between the expected group labels and those assigned by the *K*-means clustering. Thus for each subset of the data obtained by the random re-sampling, the same analysis as described above was repeated a second time by using the same data but with the class labels randomly permuted, i.e. each sample was randomly assigned a class membership. The consistency between the known class labels and labels identified by *K*-means clustering were calculated both for the model using the original labels, and the one using the

permuted labels. If the separation between the known classes were genuine, the label consistency of the models using the original labels (the observed consistency) should be always higher than those using the permuted labels (the null consistency). An empirical *p*-value can be derived by counting the number of cases when the null consistency value had been higher than the observed consistency value and divide it by *R*. In addition, a confusion matrix can be calculated by comparing these two types of labels. In the confusion matrix, each row contains the percentage of the samples in one particular cluster coming from each of the known class while each column contains the percentage of the samples in one particular class allocated into each of the clusters identified by the *K*-means clustering. Such a confusion matrix gives a more detailed information of the distribution of the classes, e.g. which class(es) were better separated from others and which classes may have certain amount of overlap between them, similar to the confusion matrix provided by supervised classification models.

Reference

(1) Xu Y, Fowler SJ, Bayat A, et al. Chemometrics models for overcoming high between subject variability: applications in clinical metabolic profiling studies. Metabolomics. 2014;10:375-85.

(2) Brereton RG, Chemometrics: Data analysis for the laboratory and chemical plant, Wiley, Chichester, 2003.

(3) Efron B, Tibshirani R, *An introduction to the Bootstrap*, Chapman & Hall, New York, 1993.

(4) Hartigan, JA, Wong MA, A *K*-means Clustering Algorithm, Journal of the Royal Statistical Society, Series C (Applied Statistics), 1979:28:100-08.

Subject Number	**Sample day 1**	**2**	**3**	**4**	**5**	**KEY:**	
ICU001	**GBS**					AF	*Aspergillus fumigatus*
ICU002						C	*Citrobacter* species
ICU003						ECl	*Enterobacter cloacae*
ICU004		**C; SL**				ECo	*Escherichia coli*

ICU005	**HI**				
ICU008	**HI**				
ICU009			**KO**		
ICU010	**HI; SA**		**HI; SA**		
ICU011		**HI; Y**			
ICU012			**SA**		
ICU013			**AF; Y**		
ICU014					
ICU016	**SA**	**SA**			
ICU017	**ECl; PA**	**ECl; PA**			
ICU018					
ICU019					
ICU020					
ICU021	**SA**	**SA**			
ICU022					
ICU023		**EF**			
ICU024		**HI;SP**			
ICU026					
ICU027	**SA**				
ICU028	**SA; HP**	**SA; SP; PP**	**SA; HP**	**SA**	
ICU030		**SA; Y**			
ICU031	**SA**	**HI**	**KP**		
ICU032					
ICU033					
ICU034					
ICU035					
ICU036	**HI**				
ICU037	**HI; SP**				
ICU038					
ICU039					
ICU040					
ICU041					
ICU043		**HI**			
ICU045					
ICU046					
ICU047					
ICU048		**SA**		**SA**	**Ps**
ICU049		**EC**			
ICU052					
ICU053	**SA; SP**	**SA**		**ECl**	
ICU054	**GCS**				
ICU055	**HI**	**SA; ECo**		**ECo**	**ECo**

EF	*Enterococcus faecium*
GBS	Group B *Streptococcus*
GCS	Group C *Streptococcus*
HI	*Haemophilus influenzae*
HP	*Haemophilus parainfluenzae*
KO	*Klebsiella oxytocia*
KP	*Klebsiella pneumoniae*
PA	*Pseudomonas aeruginosa*
PP	*Pasteurella pneumoniae*
Ps	*Pseudomonas species*
SA	*Staphylococcus aureus*
SL	*Serratia liquifaciens*
SP	*Streptococcus pneumoniae*
Y	Yeast (unidentified)

Supplemental Table: All time points are shown where a breath sample was collected for each patient. Also shown are the concomitant positive microbiological cultures, with the organism(s) isolated identified by abbreviations, shown in full in the key.