Online supplement 3.

Statistics of exhaled VOCs.

It is well known that the distribution of some exhaled biomarkers (VOCs, but also NO) is not normally distributed and therefore specific attention is needed to appropriate analysis and presentation of data. It is required that the study provides information on the distribution of data and provides estimates of data dispersion. This may include the presentation of interquartile intervals or 95% interval (2.5 to 97.5 percentile). By scaling the influence of VOCs with large variation in multi-variate modelling can be reduced. The intensity of data points can be adjusted for (1) the sum, (2) the mean or (3) the (square root of the) standard deviation [1]. Alternatively, log-transformation may be attempted to approximate a Gaussian distribution and reduce variation.

Processed data can subsequently be used for statistical analysis. Useful guidelines for data interpretation and presentation are provided by the Metabolomics Standards Initiative [2]. Initial data description relates total number of features, average features per sample, and how many shared features are common across samples. Volatiles present in a minority of samples may be excluded but this may risk omission of important but rare peaks (*e.g.* related to specific pathogens or drugs).

Exploratory (unsupervised; hypothesis-free) data analysis typically involves principal components analysis (PCA) or other clustering methods, which can reveal novel relationships within the data and generate research questions. Success may be limited because dominant influences frequently arise from non-disease factors such as gender, environment or co-morbidities [3,4].

For hypothesis-driven (supervised) research univariate analysis can be used, when adjusted for multiple testing [5]. Multivariate techniques are then applied on the predictor matrix or on a projection (principal component analysis: PCA). Common methods include discriminant analysis, support vector machines, neural networks, decision trees and Bayesian approaches [6,7].

References

1. Smolinska A, Hauschild AC, Fijten RR, Dallinga JW, Baumbach J, van Schooten FJ. Current breathomics-a review on data pre-processing techniques and machine learning in metabolomics breath analysis. J Breath Res 2014: 8(2): 027105.

2. Goodacre R, Broadhurst D, Smilde AK, Kristal BS, Baker JD, Beger R, Bessant C, Connor S, Capuani G, Craig A, Ebbels T, Kell DB, Manetti C, Newton J, Paternostro G, Somorjai R, Sjöström M, Trygg J, Wulfert F. Proposed minimum reporting standards for data analysis in metabolomics. Metabolomics 2007: 3(3): 231-241.

3. Fens N, Douma RA, Sterk PJ, Kamphuisen PW. Breathomics as a diagnostic tool for pulmonary embolism. J Thromb Haemost 2010: 8(12): 2831-2833.

4. Kischkel S, Miekisch W, Sawacki A, Straker EM, Trefz P, Amann A, Schubert JK. Breath biomarkers for lung cancer detection and assessment of smoking related effects—confounding variables, influence of normalization and statistical algorithms. Clin Chim Acta 2010: 411(21): 1637-1644.

5. Broadhurst D, Kell D. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. Metabolomics 2006: 2(4): 171-196.

6. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. Data Mining, Inference, and Prediction. 2nd edition ed. Springer New York, 2009.

7. Gromski PS, Correa E, Vaughan AA, Wedge DC, Turner ML, Goodacre R. A comparison of different chemometrics approaches for the robust classification of electronic nose data. Anal Bioanal Chem 2014:406(29):7581-90.