Published in final edited form as:
J Breath Res. 2014 March ; 8(1): 012001. doi:10.1088/1752-7155/8/1/012001.

On the importance of statistics in breath analysis - Hope or curse?

Sandrah P. Eckel[1], **Jan Baumbach**[2], and **Anne-Christin Hauschild**[3]

[1]Division of Biostatistics, Department of Preventive Medicine, University of Southern California, Los Angeles, CA, USA

[2]Computational Biology group, Institute for Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark

[3]Computational Systems Biology group, Max Planck Institute for Informatics, Saarbrucken, Germany

Abstract

As we saw at the 2013 Breath Analysis Summit, breath analysis is a rapidly evolving field. Increasingly sophisticated technology is producing huge amounts of complex data. A major barrier now faced by the breath research community is the analysis of these data. Emerging breath data require sophisticated, modern statistical methods to allow for a careful and robust deduction of real-world conclusions.

Keywords

breath analysis; exhaled nitric oxide; machine learning; robustness; statistics

The scientific program at the 2013 Breath Analysis Summit provided stimulating insights into the wealth of information that can be gleaned from air exhaled by humans. Since exhaled breath can be sampled continuously and non-invasively, there is great potential for breath analysis to lead to the development of biomarkers with widespread clinical and public health applications. Beyond breathalyzers in law enforcement, exhaled breath monitoring has become routine in clinical practice for monitoring patients undergoing anesthesia. The fractional concentration of exhaled nitric oxide (FeNO)—a marker of aspects of airway inflammation—has been studied extensively in research settings and considered for clinical applications in asthma. The breadth of developmental applications discussed at this year's summit was remarkable, ranging from diagnosis of diseases (e.g., lung cancer, tuberculosis) to locating survivors trapped in rubble following natural disasters.

As an emerging field, breath analysis is rapidly evolving. Increasingly sophisticated technology is producing huge amounts of increasingly complex data. Major data barriers now faced by the breath research community include standardizing sampling protocols,

Contact information: Sandrah P. Eckel, Department of Preventive Medicine, University of Southern California, 2001 N. Soto Street, MC-9234, Los Angeles, CA 90089, eckel@usc.edu, Phone: (323)-442-2030, Fax: (323)-442-2349.

optimizing study design, and appropriately analyzing the resultant data. In this perspective, we focus on methods for data analysis. Emerging breath data require sophisticated, modern statistical methods to produce valid and robust real-world conclusions. Statistical methodology was memorably featured in the panel discussion on the Future of Breath Analysis when it was said that "statistics will be the death of this subject". It is our opinion that statistical approaches tailored to modern breath data will be a new beginning—rather than an end—to breath analysis.

In our opinion, one historical source of criticism for proposed breath markers has been a lack of proven translations to real-world settings. Improved statistical analyses of breath data can increase the chances for successful translation of research findings to clinical and public health practice. Many statistical methods relevant for modern breath data already exist and are being regularly applied in other scientific disciplines. For example, the research questions in breath metabolomics are similar to those in other omics research areas (genomics, transcriptomics, proteomics, etc.). These other fields have at least partially overcome these criticisms by moving from demonstrating *separability* to providing evidence for *predictability*. There is strong potential to make strides in breath research by applying these existing methods. In other cases, opportunities may exist for the development of new statistical methods addressing unique features of breath data. Below, we briefly describe two areas of breath research (single compound and multiple compounds analysis) in which statistical methods offer great promise.

For single compound analysis, we focus on FeNO as an example. Society guidelines exist for standardized assessment of FeNO at the 50 ml/s exhalation flow rate [1]. FeNO is flow-dependent, which provides the possibility to partition FeNO into airway and alveolar sources (quantified by “NO parameters”) using data on FeNO measured at multiple flow rates and a statistical model that estimates the NO parameters from a deterministic model of NO exhalation [2]. NO parameters have shown promise as a tool for studying several diseases [3], but no guidelines exist for standardized assessment. Several important questions need to be addressed before NO parameter guidelines can be developed.

Three of these open questions regarding NO parameters are inherently statistical and should be addressed through interdisciplinary collaboration using rigorous statistical approaches. First, increasingly complex deterministic models have been proposed [4–7], but there is no agreement on which models adequately reproduce observed FeNO data in standard research and clinical settings. This question can be partially addressed by comparing the goodness of fit of the statistical models used to estimate NO parameters from each deterministic model to observed FeNO data across a range of flow rates. Second, for any given deterministic model, there is no agreement on the optimal statistical method to estimate NO parameters from that model. For example, the numerous statistical methods proposed to estimate NO parameters in the deterministic two-compartment model should be compared in terms of model fit as well as bias and efficiency in estimation. Third, there is no agreement on the set of flow rates at which FeNO should be assessed. Within a range of feasible flow rates, a set of optimal flow rates—for a given statistical method and deterministic model—can be derived according to standard criteria of statistical study design (e.g., to minimize the variance of parameter estimates). Controversy in the field of multiple flow FeNO analysis related to

these issues has left some researchers to simply use high flow FeNO as a proxy for alveolar NO [8] despite the limitations of this approach [9]. This is a symptom of the barrier to progress resulting from lack of statistical contributions to this area. Careful statistical consideration of important open questions in this field has the potential to move the field forward. This work could draw on the relevant statistical methods developed for similarly structured data in fields such as population pharmacodynamics [10]. Statistical methods work on single compound analysis would be complementary to work on developing refined mathematical models to represent physiological processes (e.g., recent work developing and refining models of the exhalation kinetics of isoprene [11] and acetone [12]) and on developing clinical and public health applications.

In multiple compound analysis, the goal is typically to identify the smallest subset of the universe of exhaled compounds that can provide accurate (e.g. high sensitivity and/or specificity) and robust (generalizable) predictions regarding clinical outcomes, such as diagnosing or managing disease and drug response. Methodological challenges of these analyses have been identified previously [13–14] and include: addressing sources of extraneous variability (e.g. from physiological processes and sample storage), accounting for important clinical and demographic characteristics that might be associated with both exhaled compounds and disease (i.e., confounders), filtering out exogenous compounds related to disease, controlling the inflated Type I error or “voodoo” correlations [15] inherent to studying a large number of compounds in a small number of subjects, and generalizing results from small, non-random study samples to larger populations. In our opinion, the most important but most ignored issue is avoiding model overfitting. Overfitting is a major concern when the number of study participants (N) is small (e.g., $N<50$ as in [16]) and the number of compounds (p) is much larger than N ($p \gg N$ as in [16–19]). Due to the small number of participants in many pilot studies, prediction models are often trained and tested on the whole dataset, leading to overfitted prediction models that are not generalizable to the target population. In other words, just because we find a set of compounds whose joint distribution may distinguish between two groups in a given dataset (*separability)* this by no means allows us to draw the conclusion that we have found a robust set of biomarkers for a certain disease (*predictability*). Permutation-based tests can be used to compare the observed prediction accuracy to the distribution of the same statistic calculated under the null hypothesis of no association between disease and compounds (i.e., using versions of the dataset where the disease status labels have been randomly permuted), resulting in a non-parametric p-value for the significance of the prediction accuracy in the given dataset (*separability*) [20–21]. When only a single dataset is available, cross-validation is a simple and widely used method that produces measures of predictive accuracy closer to what would be expected in an independent validation sample [17, 22–24] (*predictability* rather than *separability*). However, the gold standard for evaluating *predictability* is to apply the model to validation dataset(s) completely distinct from the training dataset. Finally, note that if the study sample is biased (not a random sample of the target population), the results of any analysis may not be generalizable to the target population. This is a general problem we face with any scientific endeavor where early developmental work is conducted in convenience samples and our knowledge of variability in the overall population is still limited.

Common statistical approaches to multiple compound analysis include: basic hypothesis tests (e.g., t-tests or their non-parameteric analog Wilcoxon-Mann-Whitney (WMW) -U-tests [16, 25–27]), basic correlation analyses (e.g., correlating amines in exhaled air with uremic breath [27] or propophol concentration in blood and breath [28]), or principal component analysis (PCA [29]). PCA is primarily used to reduce high-dimensional multiple compound data to a small number of uncorrelated principal components that explain a majority of the variation in the data which are then related to environmental factors or disease (e.g., as in [18, 30]). PCA summarizes high-dimensional data, but it is not designed to: (a) search for patterns of compounds in exhaled breath related to disease or (b) select the smallest set of compounds necessary for high quality and robust predictions. Alternative statistical methods are available that better address the research questions in multiple compound analysis. For example, as an alternative to the usual approach of relating one or more principal components to disease, linear discriminant analysis (LDA) is used to build linear combinations of compounds that discriminate between class labels of interest (e.g. disease vs. healthy), so it is more suitable for linear feature extraction and dimensionality reduction [31]. Note that the aforementioned methods investigate linear relationships between compounds and disease status, but they may overlook the nonlinear relations that are more likely in complex biological systems [32]. Nonlinear relations have been investigated recently in breath studies using other, well-established statistical learning methods, [17, 33–37]. While some of these methods (e.g., support vector machines (SVM) [38] or neural networks) may be criticized as producing “black-box” predictive models, other methods exist that are more intuitive. The simplest is the decision tree [39]. In a decision tree, the data is split at each branch according to the compound that best separates the set of samples until each leaf node contains only one class. More sophisticated tree-based methods (e.g., random forest [40] and boosting [24]), in simple terms, combine the results of a set of decision trees to reduce the variance of a single tree and thereby dramatically improve performance. They also provide a measure of variable importance for each compound, which can be used for dimension reduction and to identify a set of compounds that are potential biomarkers.

Acknowledgments

Funding sources:

SPE is grateful for funding from the National Institute of Environmental Health Sciences (grant 1K22ES022987) and the James H. Zumberge Research and Innovation Fund. ACH is grateful for funding from the International Max Planck Research School (IMPRS) and the Cluster of Excellence for Multimodal Computing and Interaction (MMCI) from the German Research Foundation. JB is grateful for financial support from MMCI and the Danish SDU2020 initiative.

References

1. ATS/ERS. ATS/ERS recommendations for standardized procedures for the online and offline measurement of exhaled lower respiratory nitric oxide and nasal nitric oxide, 2005. Am J Respir Crit Care Med. 2005; 171(8):912–930. [PubMed: 15817806]
2. George SC, et al. Modeling pulmonary nitric oxide exchange. J Appl Physiol. 2004; 96(3):831–839. [PubMed: 14766761]
3. Hogman M. Extended NO analysis in health and disease. J Breath Res. 2012; 6(4):047103. [PubMed: 22677778]

4. Tsoukias NM, George SC. A two-compartment model of pulmonary nitric oxide exchange dynamics. J Appl Physiol. 1998; 85(2):653–666. [PubMed: 9688744]
5. Condorelli P, et al. A simple technique to characterize proximal and peripheral nitric oxide exchange using constant flow exhalations and an axial diffusion model. J Appl Physiol. 2007; 102(1):417–425. [PubMed: 16888048]
6. Kerckx Y, Michils A, Van Muylem A. Airway contribution to alveolar nitric oxide in healthy subjects and stable asthma patients. J Appl Physiol. 2008; 104(4):918–924. [PubMed: 18218917]
7. Shelley DA, Puckett JL, George SC. Quantifying proximal and distal sources of NO in asthma using a multicompartment model. J Appl Physiol. 2010; 108(4):821–829. [PubMed: 20093668]
8. Barregard L, et al. Experimental exposure to wood smoke: effects on airway inflammation and oxidative stress. Occup Environ Med. 2008; 65(5):319–324. [PubMed: 17704195]
9. Eckel SP, Salam MT. Single high flow exhaled nitric oxide is an imperfect proxy for distal nitric oxide. Occup Environ Med. 2013; 70(7):519–520. [PubMed: 23645622]
10. Davidian M, Giltinan DM. Nonlinear models for repeated measurement data: an overview and update. J Agric Biol Envir S. 2003; 8(4):387–419.
11. King J, et al. Physiological modeling of isoprene dynamics in exhaled breath. J Theor Biol. 2010; 267(4):626–637. [PubMed: 20869370]
12. King J, et al. A mathematical model for breath gas analysis of volatile organic compounds with special emphasis on acetone. J Math Biol. 2011; 63(5):959–999. [PubMed: 21234569]
13. Miekisch W, Herbig J, Schubert JK. Data interpretation in breath biomarker research: pitfalls and directions. J Breath Res. 2012; 6(3):036007. [PubMed: 22854185]
14. Pleil JD, Stiegel MA, Sobus JR. Breath biomarkers in environmental health science: exploring patterns in the human exposome. J Breath Res. 2011; 5(4):046005. [PubMed: 21904020]
15. Vul E, et al. Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. Perspectives on psychological science. 2009; 4(3):274–290.
16. Bessa V, et al. Detection of volatile organic compounds (VOCs) in exhaled breath of patients with chronic obstructive pulmonary disease (COPD) by ion mobility spectrometry. Int. J. Ion Mobility Spectrom. 2011; 14:7–13. (Copyright (C) 2011 American Chemical Society (ACS). All Rights Reserved.).
17. Finthammer M, et al. Probabilistic Relational Learning for Medical Diagnosis Based on Ion Mobility Spectrometry. Int. J. Ion Mobility Spectrom. 2010; 13(2):83–92.
18. Westhoff M, et al. Statistical and bioinformatical methods to differentiate chronic obstructive pulmonary disease (COPD) including lung cancer from healthy control by breath analysis using ion mobility spectrometr. Int. J. Ion Mobility Spectrom. 2011; 11(4):139–149.
19. Hauschild A-C, Baumbach JI, Baumbach JI. Integrated Statistical Learning of Metabolic Ion Mobility - Spectrometry Profiles for Pulmonary Disease Identification. Journal Integrative Biology. 2012
20. Golland, P., et al. Learning Theory. Berlin Heidelberg: Springer; 2005. Permutation Tests for Classication; p. 501-515.
21. Ojala M, Garriga GC. Permutation tests for studying classifier performance. The Journal of Machine Learning Research. 2010; 99:1833–1863.
22. Stone M. Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society Series B. 1974; 36:111–147.
23. Phillips M, et al. Volatile organic compounds in the breath of patients with schizophrenia. J Clin Pathol. 1995; 1995:466–469. [PubMed: 7629295]
24. T. Hastie, RT.; Friedman, JH., editors. The Elements of Statistical Learning. Second Edition ed.. Springer;
25. Fay MP, Proschan MA. Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. Statist. Surv. 2010; 4:1–39.
26. Gordon SM, et al. Volatile organic compounds in exhaled air from patients with lung cancer. Clinical Chemistry. 1985; 31(8):1278–1282. [PubMed: 4017231]
27. Simenhoff ML, et al. Biochemical Profile of Uremic Breath. N Engl J Med. 1977; 297:132–135. [PubMed: 865584]

28. Kreuder A-E, et al. Characterization of propofol in human breath of patients undergoing anesthesia. Int. J. Ion Mobility Spectrom. 2011; 14(4):167–175.
29. Hotelling H. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology. 1933; 24:417–441. 498–520.
30. Gordon SM. Identification of exposure markers in smokers' breath. Journal of Chromatography A. 1990; 511:291–302.
31. Fisher RA. The use of multiple measurements in taxonomic problems. Annals of Eugenics. 1936; 7(2):179–188.
32. Isabelle Guyon AE. An introduction to variable and feature selection. The Journal of Machine Learning Research. 2003; 3:1157–1182.
33. Hu, Y., et al. EPIDEMIOLOGY. 2002. Using data mining techniques to identify volatile organic compounds associated with asthma attack. LIPPINCOTT WILLIAMS & WILKINS 530 WALNUT ST, PHILADELPHIA, PA 19106-3621 USA
34. Polikar R, et al. Artificial intelligence methods for selection of an optimized sensor array for identification of volatile organic compounds. Sensors and Actuators B: Chemical. 2001; 80(3): 243–254.
35. Mazzone PJ, et al. Diagnosis of lung cancer by the analysis of exhaled breath with a colorimetric sensor array. Thorax. 2007; 62:565–568. [PubMed: 17327260]
36. Van Berkel JJBN, et al. Development of accurate classification method based on the analysis of volatile organic compounds from human exhaled air. Journal of Chromatography B. 2008; 861(1): 101–107.
37. Baumbach J, et al. IMS2 -- An integrated medical software system for early lung cancer detection using ion mobility spectrometry data of human breath. Journal of Integrative Bioinformatics. 2007; 4(3):75, 1–12.
38. Vapnik, V. The Nature of Statistical Learning Theory. New York: Springer; 1996.
39. Breiman, L., et al. Classification and Regression Trees. New York: Chapman & Hall/CRC; 1984.
40. Breiman L. Random forests. Machine Learning. 2001; 45(1):5–32.
41. Osborne JW, Costello AB. Sample size and subject to item ratio in principal components analysis. Practical Assessment, Research & Evaluation. 2004; 9(11)
42. Hastie T, Tibshirani R. Expression arrays and the p≫ n problem. 2003

In closing, the breath analysis is undergoing a period of change: from breath biomarker discovery to biomarker validation, from *separation* ("Is there a difference?") to *predictions* ("Can we exploit it?"). This change is difficult but it offers an opportunity. Interdisciplinary collaboration between breath researchers and statisticians can benefit both fields, potentially offering a new beginning rather than the end of breath analysis.

Table 1

Summary of statistical methods referenced for multiple compound analysis, including: whether the method emphasizes separability (S) or predictability (P)* and typical number of samples (N) relative to the number of compounds (p) in practice.

Method	Application & Best practice	S/P	N/p
Hypothesis tests e.g. t-tests or Wilcoxon-Mann-Whitney-U-tests	Test whether a single compound is related to disease. When performing many tests, correct for multiple-comparisons (e.g. using Bonferroni method) to control overall Type-I (false positive) error.	S	N = test dependent p=1
Correlation e.g. Pearson or Spearman	Statistic that quantifies the strength of the linear association between two compounds.	-	p=2
Principal component analysis (PCA)	Method for identifying uncorrelated linear combinations of compounds that explain the highest variability in the set of compounds, and which may help indicate patterns. Often used for dimension reduction in large datasets, but may lead to loss of information relevant for later prediction of disease.	-	N/p > 10 suggested but often p > N in practice [41]
Linear discriminant analysis (LDA)	Method somewhat analogous to PCA, but that searches for linear combinations of compounds that best discriminate disease labels. Commonly used for dimension reduction. Occasionally used for linear prediction.	S (P)	N>p classically but extensions allow for N<p [42]
Decision tree	Intuitive model for disease prediction based on binary splits on compounds. Model simplicity comes at the cost of reduced predictive accuracy. Algorithms typically include internal cross-validation or training/validation datasets for model selection. Suffers from high variance: Small data set changes may lead to large changes in the model.	(P)	p>N**
Tree-based prediction models e.g. Random forest or Boosting	“Black-box” models for disease prediction based on combination of decision trees or other “weak learners”. Greatly improved predictive accuracy at cost of model simplicity, but provides interpretable feature importance measures. Algorithms typically include internal cross-validation or training/validation datasets for model selection.	P	p>N**
Other prediction models e.g. SVM or Neural networks	“Black-box” models for disease prediction based on other methods. Occasionally, SVM versions perform significantly better than tree-based methods, strongly depending on the nature of the data. Some provide feature importance. However, interpretation is often difficult.	P	p>N**

* Predictability relies on cross-validation or distinct training/validation datasets.

** Generally N>p but with special-purpose and highly application-specific tools for feature reduction, regularization and tree pruning, this can be relaxed to p>N.