

## Editorial

# Predicting atopic asthma by using eNose breath profiles with machine learning

Wei Wu, PhD Pittsburgh, Pa

**Key words:** Atopy, asthma, eNose, breath profile, classification, machine learning

Cluster analyses have been used to identify asthma phenotypes. Among these studies, atopy has been repeatedly identified as a key feature distinguishing different phenotypes,<sup>1-3</sup> confirming its importance in asthma phenotyping. The criterion standard for detecting atopy is by using the IgE assay or the skin prick test<sup>4</sup>; however, although the former can be limited by the availability of antigen-specific antibodies and is time-consuming, the latter, despite being fast and simple, is invasive and can have adverse effects on patients. Because atopy is a treatable trait, there is a clear need to find an alternative means to detect atopy rapidly and non-invasively at the point of care for asthma management.

Exhaled breath contains numerous disease-relevant volatile organic compounds (VOCs). Electronic noses (eNoses) can quickly detect VOC signals by capturing complex mixtures of the VOCs with multiple cross-reactive sensors. Even though these sensors cannot recognize individual VOCs, the captured signals can be potentially used as quick and noninvasive biomarkers for prediction of disease or treatable traits at the point of care.

But can eNose breath profiles obtained from patients with asthma be used to classify atopy? In this issue, Abdel-Aziz et al<sup>5</sup> investigated this question using 4 independent cohorts of adult and pediatric participants with asthma or wheezing: BreathCloud, U-BIOPRED (the Unbiased Biomarkers for the Prediction of Respiratory Disease Outcomes) adult and pediatric, and PACMAN2 (Pharmacogenetics of Asthma medication in Children: Medication with ANti-inflammatory effects 2). Exhaled breath from participants was measured by 1 of the 2 types of eNoses, with real-time eNoses (SpiroNoses) used in BreathCloud and the offline eNoses used in the other cohorts.

One of the strengths of this study is that to predict atopy (a response variable), the eNose profiles were analyzed by 3 supervised machine learning (ML) algorithms to get a robust estimate of classification accuracy. Most ML algorithms need to be understood well, and parameters in an algorithm should be carefully tuned so that common problems such as overfitting can

be avoided. Overfitting happens when a classifier fits the training data too well; consequently, the “trained” classifier predicts the response variable in the training data with high accuracy (called *training accuracy*), but it does not generalize well and predicts that in the validation/test data with lower accuracy (called *validation/test accuracy*). To prevent overfitting, the commonly used ML techniques include regularization and parameter tuning with internal cross-validation.<sup>6</sup> In this study,<sup>5</sup> both techniques were used to try to avoid overfitting during classification (summarized in Fig 1). However, these techniques are not foolproof in practice, and biological/clinical expert opinions are sometimes also needed to guide selection of the optimal parameters in an algorithm.

The algorithms used,<sup>5</sup> adaptive Lasso, sparse partial least squares discriminant analysis (sPLS-DA), and gradient boosting machine (GBM), were developed previously by making different assumptions about the data and using different prediction rules to distinguish different classes<sup>6</sup> (eg, atopic vs nonatopic). These algorithms were described in the Methods section in Abdel-Aziz et al.<sup>5</sup> Briefly, adaptive Lasso—an approach derived from the well-known Lasso regression<sup>6</sup>—performs linear regression with a weighted L1 regularization, which allows the algorithm to produce a sparse solution to avoid overfitting.<sup>6</sup> sPLS-DA accounts for correlation among predictor variables, and also uses regularization to avoid overfitting. However, GBM is an ensemble approach that combines outputs of a number of weak base-classifiers to produce a stronger ensemble prediction.<sup>6</sup> Ensemble methods are often found to achieve higher classification accuracy than single classifiers, but because they contain more parameters, various techniques need to be used to avoid overfitting.

Abdel-Aziz et al<sup>5</sup> first used the eNose profiles in each individual cohort to train the classifiers and predict atopy with 10-fold cross-validation. All 3 classifiers achieved high classification accuracy, with the area under a receiver operating characteristic curve (AUC) of more than 0.85 for all 4 individual cohorts. To get a more independent estimate of the prediction accuracy, another strategy was used. Three cohorts (U-BIOPRED adults and pediatric, and PACMAN2) that used offline eNoses were first pooled so that the sample size can be increased. BreathCloud used real-time eNoses, so it could not be pooled with the other cohorts. Then in either BreathCloud or the pooled cohorts, participants were randomly split into a training set with 75% of the patients and a validation set with 25% of the patients. The training and validation sets were then used by the classifiers to obtain training and validation accuracy, respectively.

For BreathCloud, all 3 classifiers achieved high accuracy (AUCs > 0.9), specificity, and sensitivity, on both the training and validation sets. This is an encouraging result, suggesting that real-time eNoses can indeed be potentially used as point-of-care tools for atopy prediction. For the pooled cohorts, classification accuracies by GBM and sPLS-DA were noticeably much higher in

From the Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh.

W.W. was supported by grants from the National Institutes of Health (grant nos. P30DA035778 and P01 AI106684).

Disclosure of potential conflicts of interest: The author declares no relevant conflict of interest.

Received for publication June 11, 2020; revised August 5, 2020; accepted for publication August 7, 2020.

Corresponding author: Wei Wu, PhD, Computational Biology Department, School of Computer Science, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213. E-mail: [weiwu2@cs.cmu.edu](mailto:weiwu2@cs.cmu.edu).

J Allergy Clin Immunol 2020;■■■■:■■■■-■■■■.  
0091-6749/\$36.00

© 2020 American Academy of Allergy, Asthma & Immunology  
<https://doi.org/10.1016/j.jaci.2020.08.009>

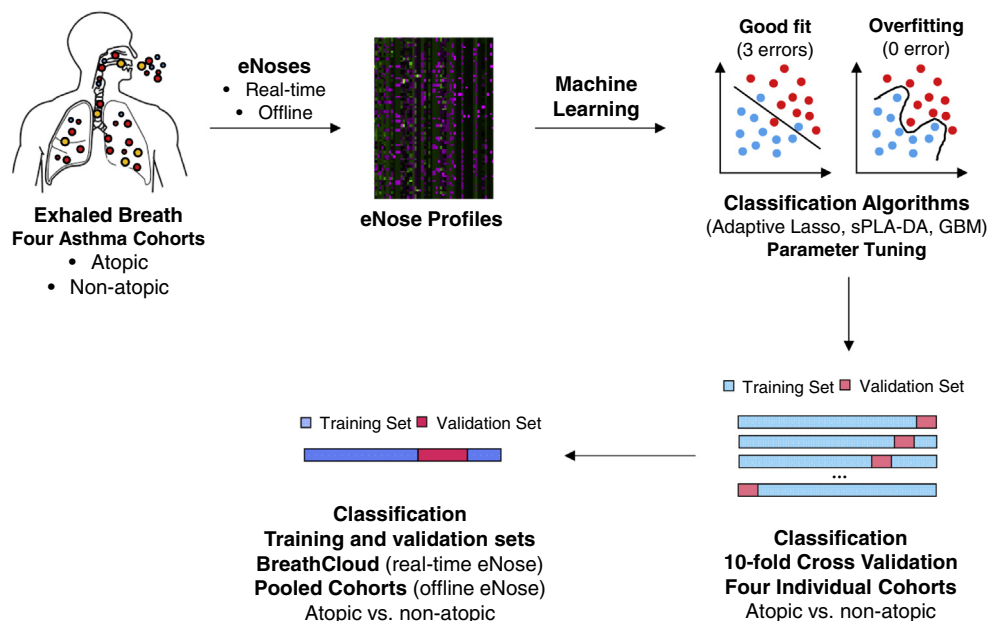


**FIG 1.** Schematic illustration of the supervised learning strategy used by Abdel-Aziz et al<sup>5</sup> for classification of atopy in patients with asthma by eNoses.

the training set than in the validation set (eg, GBM achieved the AUC of 0.95 on the training set, but the AUC of 0.762 on the validation set), indicating that these classifiers suffered from overfitting even though various strategies were used to avoid such problem.<sup>5</sup> Adaptive Lasso performed similarly on both the training and validation sets, indicating that it did not have the overfitting problem. Nevertheless, for the validation accuracy, all 3 classifiers showed similar modest-good performance, with the AUCs in the range of 0.718 to 0.762. However, fractional exhaled nitric oxide could not distinguish well patients with or without atopy.

The main reasons why validation accuracy on the pooled cohorts was lower than that on BreathCloud include the following—the former had a more heterogeneous population (with both pediatric and adult participants and recruited from >10 European countries), had a smaller sample size, used offline eNoses (vs real-time ones in BreathCloud), and had batch effects in the eNose breath profiles.

Finally, a Bayesian Network approach was used to investigate confounders associated with atopy in BreathCloud. The network results (shown in Fig E3 in Abdel-Aziz et al<sup>5</sup>) suggested that signals from the eNose sensors were directly connected to atopy, oral corticosteroids, and blood eosinophils, but not connected to confounders such as coffee, confirming similar findings from previous studies.<sup>7</sup>

Multiple future directions can be pursued. First, the results in this study should ideally be validated in an independent cohort. Considering the limitations of eNoses,<sup>8</sup> a replication study should also be conducted to investigate the repeatability of the classification results in the same cohorts. Second, if the goal is to identify eNose biomarkers for atopy and improve classification accuracy, several strategies can be considered; for example, adult and pediatric participants can be classified separately. Also, participants can be stratified on the basis of their sensitization patterns to allergens, and then classifiers can be used to predict the patterns. For biomarker discovery, an optimal subset of the predictive variables

from eNose profiles can be identified using feature selection, and this subset of the variables can be served as potential biomarkers for atopy prediction. Finally, to understand pathophysiologic links between exhaled breath and atopy, individual VOCs in the predictive eNose profiles can be characterized using gas chromatography mass spectrometry, and metabolomics of exhaled breath condensate can be identified by nuclear magnetic resonance spectroscopy.<sup>8</sup> These individual VOCs and the metabolomics can then be correlated with atopy-relevant genes, proteins, and lipids should they become available from the ongoing omics studies in U-BIOPRED.<sup>9</sup>

Despite the encouraging eNose results, it should be noted that IgE assays and skin prick tests remain primary approaches for the diagnosis of atopy in clinical practice. Furthermore, even for patients classified as atopic by the eNose-based approach, if immunologic therapies are needed for treating them, information on specific allergens still needs to be determined for treatment such as subcutaneous immunotherapy<sup>10</sup> or total IgE level needs to be decided for dosing should omalizumab be prescribed for an anti-IgE therapy.

## REFERENCES

1. Haldar P, Pavord ID, Shaw DE, Berry MA, Thomas M, Brightling CE, et al. Cluster analysis and clinical asthma phenotypes. *Am J Respir Crit Care Med* 2008;178:218-24.
2. Moore WC, Meyers DA, Wenzel SE, Teague WG, Li H, Li X, et al. Identification of asthma phenotypes using cluster analysis in the Severe Asthma Research Program. *Am J Respir Crit Care Med* 2010;181:315-23.
3. Wu W, Bleecker E, Moore W, Busse WW, Castro M, Chung KF, et al. Unsupervised phenotyping of Severe Asthma Research Program participants using expanded lung data. *J Allergy Clin Immunol* 2014;133:1280-8.
4. Ali FR. Does this patient have atopic asthma? *Clin Med (Lond)* 2011;11:376-80.
5. Abdel-Aziz MI, Brinkman P, Vijverberg SJH, Neerincx AH, de Vries R, Dagelet YWF, et al. eNose breathprints as a surrogate biomarker for classifying asthma patients by atopy. *J Allergy Clin Immunol* 2020;146 XXX-XXX.
6. Hastie T, Tibshirani R, Friedman JH. *The elements of statistical learning: data mining, inference, and prediction*. Springer; 2009.

7. de Vries R, Dagelet YWF, Spoor P, Snoey E, Jak PMC, Brinkman P, et al. Clinical and inflammatory phenotyping by breathomics in chronic airway diseases irrespective of the diagnostic label. *Eur Respir J* 2018;51.
8. Santini G, Mores N, Penas A, Capuano R, Mondino C, Trové A, et al. Electronic nose and exhaled breath NMR-based metabolomics applications in airways disease. *Curr Top Med Chem* 2016;16:1610-30.
9. Shaw DE, Sousa AR, Fowler SJ, Fleming LJ, Roberts G, Corfield J, et al. Clinical and inflammatory characteristics of the European U-BIOPRED adult severe asthma cohort. *Eur Respir J* 2015;46:1308-21.
10. Cox L, Nelson H, Lockey R, Calabria C, Chacko T, Finegold I, et al. Allergen immunotherapy: a practice parameter third update. *J Allergy Clin Immunol* 2011;127:S1-55.