Proof of concept for identifying cystic fibrosis from perspiration samples

Zhenpeng Zhoua, Daniel Alvarzb, Carlos Milla, and Richard N. Zare*1

*Department of Chemistry, Stanford University, Stanford, CA 94305; and Center for Excellence in Pulmonary Biology, Stanford University School of Medicine, Stanford, CA 94305

Contributed by Richard N. Zare, October 11, 2019 (sent for review June 4, 2019; reviewed by Nathalie Y. R. Agar, Alan K. Jarmsusch, and Bineet Sharma)

The gold standard for cystic fibrosis (CF) diagnosis is the determination of chloride concentration in sweat. Current testing methodology takes up to 3 h to complete and has recognized shortcomings on its diagnostic accuracy. We present an alternative method for the identification of CF by combining desorption electrospray ionization mass spectrometry and a machine-learning algorithm based on gradient boosted decision trees to analyze perspiration samples. This process takes as little as 2 min, and we determined its accuracy to be 98 ± 2% by cross-validation on analyzing 277 perspiration samples. With the introduction of statistical bootstrap, our method can provide a confidence estimate of our prediction, which helps diagnosis decision-making. We also identified important peaks by the feature selection algorithm and assigned the chemical structure of the metabolites by high-resolution and/or tandem mass spectrometry. We inspected the correlation between mild and severe CFTR gene mutation types and lipid profiles, suggesting a possible way to realize personalized medicine with this noninvasive, fast, and accurate method.

desorption electrospray ionization | mass spectrometry | machine learning | cystic fibrosis

Cystic fibrosis (CF) is the most common life-threatening autosomal recessive disease in the United States and is the result of mutations in the CF transmembrane conductance regulator (CFTR) gene (1, 2). Although there is no cure, CFTR modulator drugs and intensive therapy regimens can improve the outlook for people with CF, if diagnosed early (3). The data accumulated from several longitudinal studies (4–6) demonstrated the benefits of early diagnosis prompted by newborn screening (NBS), which is now universally available in the United States. However, NBS only identifies newborns at risk for CF, and the benefits can only be fully realized when the appropriate confirmatory diagnostic testing is in place. Despite its initial description more than 60 y ago, the sweat chloride test by cholinergic pilocarpine iontophoresis remains to this date as the clinical standard for CF diagnosis (7).

The sweat chloride test (8) includes first stimulation of a high rate of sweat secretion in a small area of skin by pilocarpine iontophoresis, collection of sweat, and determination of chloride concentration in the sweat collected (9). The testing routine, particularly when working with newborns, entails great technical skill, sophisticated equipment, and cumbersome sweat collection devices. It is also recognized that newborns are an especially challenging population to test as the failure rate for sample collection will lead to the highest rate of success in obtaining a valid sample. To accomplish this, we selected a standard microscope

Significance

Early diagnosis and characterization of the severity of CFTR mutations carried in cystic fibrosis (CF) impacts life expectancy and quality of life for patients. We demonstrate a testing platform that combines analysis of perspiration samples by desorption electrospray ionization mass spectrometry and a machine-learning method of gradient boosted decision trees, with an accuracy for the correct identification of CF cases of 98 ± 2%. Our sampling method is minimally invasive; it only requires swiping a standard microscope slide across the patient’s forehead, with no sample processing. The whole collection and testing process takes less than 2 min, which suggests a faster alternative with comparable accuracy to the conventional sweat chloride test, which takes up to 3 h.

Author contributions: Z.Z. and R.N.Z. designed research; Z.Z. and D.A. performed research; Z.Z., D.A., C.M., and R.N.Z. analyzed data; and Z.Z., C.M., and R.N.Z. wrote the paper.


The authors declare no competing interest.

Published under the PNAS license.

Data deposition: Raw data can be accessed through the Open Science Framework at https://osf.io/j59h2/?view_only=cd123d071271499095550a65d4b0213, DOI:10.17605/OSF.IO/J59H2.

1To whom correspondence may be addressed. E-mail: zare@stanford.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1909630116/-/DCSupplemental.

First published November 18, 2019.
glass slide as the collection device and simply swiped it across the forehead or nose of a subject to collect the sample. The swiping process involves applying gentle pressure to the microscope slide as it is moved across the forehead of the patient, a task that takes about 5 s to perform. Then, the samples collected in this manner are the products of perspiration present in the skin surface rather than stimulated sweat. This introduces an important feature of our method as it obviates the need for active sweat gland stimulation and eliminates the known influence of the secretory rate achieved on the concentrations of analytes in sweat. Next, we applied DESI directly to the glass slide without any extra sample preparation steps. We employed a machine-learning model of gradient boosted decision trees (GBDT) (34) to recognize the pattern of the metabolites in the sample and searched for the distinction of CF versus healthy controls. The reason why we choose GBDT method is that it not only exhibits high classification performance but also provides an explainable model. The tree model makes predictions based on a series of tests of the intensities of peaks. Although molecules from exogenous sources such as lotions or creams are picked up by the mass spectrometer, the machine learning algorithm tends to pick the features that exist in all subjects of a class (i.e., CF or healthy control). Therefore, exogenous sourced molecules are usually not selected because their large variance across different subjects causes them to be rejected by the machine learning algorithm.

In addition, statistical bootstrap (35) was used to provide an estimate of the uncertainty in the prediction, which provides further information for healthcare providers to decide whether additional diagnostic testing (e.g., further gene sequencing, CFTR functional assays (7)) is necessary. In addition, we selected important mass peaks by the feature selection algorithm to gain insight as to which metabolites and lipids contribute to the prediction of disease state. The concept of important features selected by the algorithm are different from the traditional definition of biomarkers, but they are related in the following 2 ways: 1) if an important feature has been identified as a known biomarker, it will add credibility to the model and 2) if an important feature is previously unknown, it may suggest a biomarker candidate for future study. Finally, given the large number of CFTR mutations with variable spectrum of dysfunction associated (36), we examined the correlation between gene mutations and patterns in lipid composition.

Results

Mass Spectrometric Analysis of Perspiration Samples. DESI mass spectrometry was applied on 277 perspiration samples at negative ion mode. On the surface of each sample, 30 spots are randomly selected to be analyzed by DESI; their mass spectra are then averaged. The spots are chosen randomly from the glass area seen to contain the perspiration sample. In DESI-MS imaging of tissues, each pixel or spot is different because they are composed of different cells. However, each spot is taken from the perspiration sample obtained from the skin surface and is believed to represent a homogeneous distribution of perspiration molecules. Therefore, we believe it is better to analyze the average rather than individual spots. It is not necessary to remove the background because the algorithm tends to choose the feature that is different between the CF and healthy control samples. The background peaks that are contained in both samples will not be used during classification. Of the 277 samples, 57 were collected from CF patients, while 220 came from healthy controls. The CF and healthy controls have similar age brackets (SI Appendix, Fig. S1).

Fig. 1 shows the average mass spectra for CF patients and healthy controls at negative ion mode. SI Appendix, Fig. S2, shows the same spectra in an overlaid manner. We introduce the terms \( m/z \) (the mass-to-charge ratio of an ion analyzed by the mass spectrometer) and peak (molecules with a specific \( m/z \) appear as a peak in the mass spectrum). Most of the ions detected in the region \( m/z = 200 \) to 350 are fatty acids, while the ions in the region \( m/z = 450 \) to 600 are mostly fatty acid dimers or diacylglycerols. A visual examination reveals that there is little difference between CF and healthy control samples in the \( m/z = 200 \) to 350 region.

Predictive Diagnosis with Machine Learning. Approximately 1,000 distinct peaks were extracted from the whole set of samples. GBDT (34) was applied to the samples to classify them between CF and healthy states. GBDT is a tree-based algorithm that uses a sequence of rules to evaluate the chance of the sample being in a specific category. Given the limited size of samples, 6-fold rotational cross-validation was employed to evaluate the performance of the model. The dataset was randomly split into 6 parts; in each of the 6 rounds, 1 part was chosen to be the test set, and all other parts were used as the training set. The goal of cross-validation is to test the ability of the model to predict unseen data. With the accuracy of 6-fold cross-validation (SI Appendix, Table S1), the accuracy of the model was determined to be 98 ± 2%. With the reported sweat chloride test performance metrics for the diagnosis of CF of an accuracy of 98%, specificity of 92.8%, and sensitivity of 100% (24), our method for analyzing perspiration reaches a comparable performance but with the added advantage of it being a highly feasible and faster alternative.

Table 1 presents the values of all of the metrics. The definitions of recall (sensitivity), precision, and specificity are

\[
\text{Recall} = \frac{tp}{tp + fn},
\]

\[
\text{Precision} = \frac{tp}{tp + fp},
\]

\[
\text{Specificity} = \frac{tn}{tn + fp},
\]

where \( m \) denotes true negative, \( tp \) denotes true positive, \( fn \) denotes false negative, and \( fp \) denotes false positive. These metrics measure the performance of the model under different conditions. AUC is the area under the receiver operating characteristics curve (SI Appendix, Fig. S3), which illustrates the diagnostic ability of the model as its discrimination threshold is varied.

Uncertainty of Diagnosis. Considering the needs in clinical practice, healthcare providers are often faced with a necessity to make decisions based on the level of confidence that a given test result has as a predictor. Further diagnostic tests may have to be
the cross-validation process.

diction accuracy of the confident samples is 100% throughout

average 76% of the samples marked as confident, and the pre-
cross-validation process as previously described, we have on av-
prediction is lower than the threshold. Over a similar 6-fold
predict will be marked as
agnostic practice. We established a threshold of 0.1, which means
the prediction will be marked as

Further, we performed an experiment to simulate real di-
agnostic practice. We established a threshold of 0.1, which means
the prediction will be marked as “confident” only if the SD of the
prediction is lower than the threshold. Over a similar 6-fold
cross-validation process as previously described, we have on av-
average 76% of the samples marked as confident, and the pre-
diction accuracy of the confident samples is 100% throughout the
cross-validation process.

Feature Selection and Identification. The model described above
is able to determine whether a sample comes from a CF pa-
tient or a healthy control without explicitly knowing what
biomarker or set of biomarkers are contributing to the dis-
tinction. However, it is still desired to identify the important
chemical species in the classification procedure. An impor-
tant feature, which is identified as a known biomarker, will help
rationalize the decision-making process. Moreover, an important
feature that is previously unknown will lead to possible metabolite
discovery with relevance to the disease state that ultimately could
serve as a biomarker of the degree of function or dysfunction
present at the individual level.

The GBDT algorithm is capable of evaluating the impor-
tance of a feature by measuring the number of times when
using this feature as a branching point for the tree. The more
the feature is used, the higher the importance. Fig. 3 shows a
bar plot of the peak importance calculated by the GBDT
model. Over the 1,222 peaks extracted from the samples, 280
peaks (SI Appendix, Table S2) are selected to be important
features to distinguish between healthy controls and CF
patients. The important features are mostly lipids because the
perspiration samples we collected consist of mainly lipids.
The intensity of a peak in a mass spectrum correlates with
the concentration of 1 or more chemical species in the sample.
The chemical species that are represented by the important
peaks were tentatively identified using high-resolution and/or
tandem mass spectrometry (SI Appendix, Figs. S6–S33). SI Ap-
pendix, Table S3, lists several important peaks and their identifi-
cations. Many of those identified molecules are found to be
biologically relevant. For example, m/z = 305.2480 (error 0.0
ppm) and its tandem mass spectrum matched those of Mead
acid FA(20:3). Here 20 stands for 20 carbons in the fatty acid
molecule, while 3 means there are 3 carbon–carbon double
bonds. The proposed chemical structure of FA(20:3) is shown in
SI Appendix, Fig. S5A. This molecule was found to have a higher
mean concentration in CF samples, which agrees with the result of increased Mead acid level in blood and tissues of
CF patients (37). Another peak selected to be important by
GBDT is m/z = 255.2328, which is identified to be palmitic acid
FA(16:0) based on high-accuracy mass measurements (mass
error 1.57 ppm) and tandem mass spectrometry. The proposed
chemical structure of FA(20:3) is shown in SI Appendix, Fig.
S5B. Previous studies show that palmitate has a higher con-
centration in the plasma of CF patients (38), and CF patients
absorb more of this fatty acid than healthy controls (39). Those
studies confirm our result, which shows that the concentration
of FA(16:0) is higher in CF patients. Although chemical iden-
tification is not necessary, the feature selection adds credibility
to the model.

Correlation Between Gene Mutations and Lipid Profiles. A large
number of mutations are known to be associated with CF (40).
This contributes to the known variability seen from patient to
patient. Different CFTR mutations usually vary in the severity of
the defect produced and have different cellular mechanisms (36).
The genetic mutation of each CF patient that contributed a
sample is presented in SI Appendix, Table S4. Following the cate-
gorization scheme proposed by McKone et al. (41), the mutations
were classified into 2 categories, mild and severe, for practical
considerations. For each peak, a test was performed on the null
hypothesis of the distributions of that specific peak intensity are
the same among mild and strong mutations. SI Appendix, Table
S5, shows the peaks with P values less than 0.05. Also, statistical
feature selection is performed with least absolute shrinkage and
selection operator (Lasso) (42), whose test accuracy is 77 ± 9% on
a 6-fold cross-validation. The peaks with nonzero coefficients

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>98 ± 2%</td>
</tr>
<tr>
<td>Recall</td>
<td>96 ± 7%</td>
</tr>
<tr>
<td>Precision</td>
<td>94 ± 7%</td>
</tr>
<tr>
<td>Specificity</td>
<td>98 ± 2%</td>
</tr>
</tbody>
</table>

Table 1. The average performance of the prediction model

Fig. 2. (A) Graphical illustration of bootstrap methods. (B) The empirical distribution of the SD of correct and wrong predictions. The y axis is the number of
samples for which the prediction is correct or wrong.
are recognized as important features, which are shown in SI Appendix, Table S6. Interestingly, the peak at m/z = 609.5109 is believed to be significantly different by hypothesis testing, as well as an important feature by Lasso. This peak is tentatively identified as triacylglycerol TG(34:0) by a high-accuracy mass measurement (mass error 1.48 ppm). A previous study on lipid profiles in CF demonstrated that CF patients in all age groups had higher triacylglycerol (43). As a further step, our results show that TG levels are different between mild and severe CFTR mutations, and there is a correlation between CFTR gene mutations and lipid profiles.

Discussion

We present here a method for simple, fast, and accurate CF testing, which uses DESI mass spectrometry to analyze the chemical composition of perspiration and machine learning to recognize underlying patterns that distinguish between healthy controls and CF patients. Compared with other methods including electrospray ionization mass spectrometry and capillary electrophoresis–mass spectrometry, DESI enables the analysis of a sample surface without sample processing, which reduces the time needed, as well as the variation induced in sample preparation steps. The advantages introduced by our simple sampling procedures constitute an unquestionable improvement over current gold standard CF diagnostic methodology. In addition, we employed statistical bootstrap to estimate the confidence level of the prediction given by the algorithm, which provides further information for healthcare providers to make decisions in practice. Although metabolite identification is not necessary for diagnosis, we have pinpointed the important chemical features in the prediction process and tentatively identified some of them by high-accuracy and/or tandem mass spectrometry. The discovery of biologically related molecules adds credibility to the model. In the final part, we correlated the CFTR gene mutations with the change of lipid profiles in perspiration, which suggests an alternative to gene sequencing for building a patient model for personalized medicine. Future improvement to this work can be achieved by enrolling more patients to increase the sample size of a genetically diverse CF population, as well as further identification of the important metabolites present in CF patients. In addition, we believe our method opens the possibility of using the peaks identified to monitor the change in CFTR function introduced by novel CFTR modulators being introduced to the clinic.

Materials and Methods

Sample Collection. All experiments involving human subjects were approved by Stanford University's Institutional Review Board, which required adults to give an oral consent form and that all data collected would be anonymized. Perspiration samples are collected by gently swiping a standard microscope glass slide across the forehead of each participant. It is possible that some dead skin is also collected in this procedure, but it is not readily dissolved by the droplet spray, so we refer to the samples analyzed as perspiration. Samples were stored at room temperature until batch processing. The interval between collection and mass spectrometric analysis ranges from 1 to 8 wk. The lipid profile changes during storage, but the formulation of our algorithm tends to use features that do not vary much with time for classification, as both CF and healthy control samples were stored for comparable time periods. The rationale behind our claim is if a feature varies strongly with time, then the difference between CF and normal samples with respect to this feature will disappear, and it will be difficult to use this feature to separate CF and normal samples.

Mass Spectrometry Analysis. The DESI method was employed for sweat sample analysis. A laboratory-built DESI source with an x–y stage was set up in front of an LTQ-Orbitrap XL mass spectrometer (Thermo Scientific). The mass spectrum was collected under negative ion mode with m/z of 50 to 1,000. The DESI source used methanol–water (9:1 vol/vol) as the solvent with a flow rate of 4 μL/min. The nitrogen gas pressure was set to 120 psi. In all experiments, the Orbitrap ion analyzer was calibrated and operated under a resolution of 60,000 to ensure accurate measurements. On the surface of each sample, 30 spots were randomly selected to be analyzed by DESI.

Data Analysis. The Thermo proprietary raw files were converted to mzML file formats and then read in Python with the mzmml package (44). The 30 spectra of each sample were averaged. A handwritten peak finding algorithm was employed to convert the continuous spectrum to sparse peak profiles. A total of 1,222 peaks are chosen such that each peak appears in at least 60 samples. Each sample was then vectorized by the peak values with a resolution of 0.1 m/z, which means that each sample is converted to a 1,222-dimensional vector based on the intensities of the specific peaks. Six-fold rotational cross-validation was employed to evaluate the performance of a model. The dataset was randomly split into 6 parts; in each of the 6 rounds, 1 part was chosen to be the test set, and all other parts are used as the training set. The mechanism of the GBDT model is that it only uses the important features to make a prediction. Therefore, the metrics we reported are based only on the selected features.

Differential classification algorithms of logistic regression with 11 regularization (Lasso), support vector machines, random forests, and gradient boosted decision trees (GBDT) are tested based on the performance of cross-validation. The number of trees in the GBDT model is 100. The algorithms are implemented in open-source packages of scikit-learn (45) and lightgbm (34).

Chemical Identification. The important features selected by the GBDT algorithm are correlated with specific peaks in the mass spectrum. The chemicals represented by the peaks are identified by searching the database of Lipid Maps (https://www.lipidmaps.org), METLIN (https://metlin.scripps.edu), and MassBank (www.massbank.jp), according to their high-resolution and high-accuracy m/z values. Further, we performed collision-induced dissociation (CID) tandem mass spectrometry to obtain the fragmentation profile of a specific peak. The raw data can be found at https://osf.io/4j5yk/?view_only=c0212307a2714d909559550a65db0213. The CID spectra corresponding to the chemical identification are provided as SI Appendix, Figs. S6–S33. The fragmentation profiles are compared with the standard from the database, if available. Given the complicated biological matrix (the perspiration sample), not all fragmented ions can be matched with the standard samples. Therefore, we claim tentative identification of the chemicals. In addition, the exact position and stereochemistry of the unsaturated bonds are not able to be determined with tandem mass spectrometry. However, given the source of the samples is human beings, we can claim the most probable metabolites are the human-sourced ones.

ACKNOWLEDGMENTS. Z.Z. is grateful for the support of a Stanford Graduate Fellowship. This work was supported by National Science Foundation under the Data-Driven Discovery Science in Chemistry (D3SC) for Early Concept Grants for Exploratory Research (EAGER Grant CHE–1734082) and the Ross Mosier CF Laboratories at Stanford.