

Analytical and Bioanalytical Chemistry

Electronic Supplementary Material

Rapid breath analysis for acute respiratory distress syndrome diagnostics using a portable two-dimensional gas chromatography device

Menglian Zhou, Ruchi Sharma, Hongbo Zhu, Ziqi Li, Jiliang Li, Shiyu Wang, Erin Bisco, Justin Massey, Amanda Pennington, Michael Sjoding, Robert P. Dickson, Pauline Park, Robert Hyzy, Lena Napolitano, Christopher E. Gillies, Kevin R. Ward, Xudong Fan

Section S1: Portable 2D GC Description and Operation

S1.1. Materials

DB-1ms Agilent J&W, nonpolar column (length 10 m, i.d. 250 μm , film thickness 0.25 μm) was purchased from Agilent Technologies (P/N: 122-0162, Agilent Technologies). SUPELCOWAX[®] 10 polar column (length 3 m, i.d. 250 μm , film thickness 0.25 μm) was purchased from Sigma Aldrich (P/N: 24077, Sigma-Aldrich). Copper tube (length 10 cm, i.d. 1 mm, o.d. 1.5 mm) was purchased from Swagelok and glass wool was purchased from Sigma Aldrich. Teflon tape was purchased from Grainger (Ann Arbor, MI). Shrink tube was purchased from Digi-Key Electronics. Other materials are the same as those described in Ref. [1].

S1.2. Design, fabrication, and characterization of components

Various microfabricated components were used in the present portable 1 x 2 channel 2D GC device, including thermal desorption tube, micro-fabricated thermal injector (μTI), micro-Deans switch (μDS), and micro-photoionization detector (μPID). All of these components were fabricated and characterized in-house. The details of μTI , μDS , and μPID can be found in Ref. [1]. The detail of the thermal desorption tube is given as follows.

The thermal desorption tube was made of a 5 cm long copper tube having an inner diameter of 1 mm. Both Carbopack[™] X and B granules, 10 mg each, were loaded into the hollow cylindrical copper tube using a diaphragm pump. Glass wool was used to separate the Carbopack[™] X and B, as well as to seal the copper tube from both ends. Swagelok fittings were used to connect a stainless steel tube of i.d. 250 μm at both the ends of the copper tube. For temperature ramping, the nickel wire was wrapped around the entire length of the copper tube. The nickel wire was insulated from the copper tube using a Kapton tape. A type K thermocouple was attached to the copper tube using a Kapton tape to monitor the temperature in real time. Finally, the thermal desorption tube was preconditioned at 300 $^{\circ}\text{C}$ for 12 h under helium flow.

S1.3. Device assembly

The portable 1 x 2 channel 2D GC device is similar to the 1 x 4 channel 2D GC device described in Ref. [1]. As illustrated in Fig. 3 in the main text, the 2D GC consisted of a sampling module, a 1st-dimensional separation module, and a 2nd-dimensional separation module. The sampling module consisted of a sampling tube, a thermal desorption tube loaded with Carbopack[™]

X and B, valves, and a pump. The 1st-dimensional module consisted of a μ TI loaded with CarbopackTM X and B, a 10 m long Agilent J&W DB-1ms, and a μ PID. The 2nd-dimensional module had two identical channels. Each channel consisted of a μ TI, a 3 m long SUPELCOWAX[®] 10 column, and a μ PID. The eluent from the 1^D column was transferred to one of the 2^D columns via a μ DS. All the modules and components were connected via tubings, universal connectors, and Y-connectors. The entire device was housed in a customized plastic case (see Fig. 2 in the main text) and had a total weight less than 5 kg, including the weight of the He gas cartridge (231 g). LabVIEWTM based codes were developed in-house for user interface, and device control and automation.

The portable GC can be operated in 1D GC alone (in which case the 2nd-dimensional module was disabled or detached) or comprehensive 2D GC. Operation as 1D GC is straightforward. Operation as comprehensive 2D GC is described in Section S1.4 below.

During the measurement, the device was secured on a rolling cart (see Fig. 2) and placed outside the ventilated patient's room. A 2 m long polytetrafluoroethylene (PTFE) tubing (0.64 cm i.d.) was used to connect the output of the ventilator to the GC device, through the 7.6 mm port of 22M-22F straight connector (shown in Fig. 2). The straight connector was discarded after a single use and the PTFE tube was cleaned (first rinsed with 70% 2-propanol, then flushed with deionized water and finally dried with pressured air, to eliminate pathogenic microorganism and remove residual VOCs) after each sampling to avoid patient-to-patient transmission as well as cross contamination between patient breath samples.

S1.4. Operation of the portable 1x2-channel 2D GC

The operation procedures and parameters of the portable GC in the comprehensive 2D GC mode are described as follows.

- (1) Sampling: The exhaled breath of the patient was drawn by the diaphragm pump through the 2-port valve and adsorbed by the thermal desorption tube at a flow rate of 70 mL/min for 5 min.
- (2) Desorption and injection: The 2-port valve was closed and the helium gas was flowed through the 3-port valve at a flow rate of 2 mL/min. Meanwhile, the thermal desorption tube was heated to 300 °C for 5 min to transfer the analytes onto the μ TI 1. After 5 min, μ TI 1 was heated to 270 °C in 0.6 s and then kept at 250 °C for 30 s for complete thermal desorption and injection of the analytes into the 1^D column.

(3) Separation: The analytes underwent separation through the 10 m long ¹D column and then detected by μ PID 1. During the separation, the column was kept at 25 °C for 2 min, then first ramped at a rate of 10 °C min⁻¹ to 80 °C, next ramped at a rate of 20 °C min⁻¹ to 120 °C, and kept at 120 °C for 4 min. The flow rate was 2 mL/min for the ¹D column.

We used a modulation period of 10 s to inject the eluent from the ¹D column into the ²D columns. The first 10 s long slice of the eluent from the ¹D column was routed to and trapped by μ TI 2A, which were kept at room temperature (25 °C). The μ TI 2A was then heated to 270 °C in 0.6 s and then kept at 250 °C for 5 s to inject the trapped analytes to Column 2A. In the meantime, the second 10 s long slice of the eluent from the ¹D column was routed to and trapped by μ TI 2B, which was subsequently injected into Column 2B. The same operation was repeated between Columns 2A and 2B alternatively throughout the analysis. The analyte underwent ²D separation through one of ²D columns (kept isothermally at 75 °C during entire operation) and then detected by μ PID 2. During the entire operation, the helium flow was 3 mL/min for each of ²D columns.

(4) Cleaning: After analysis, the outlet of the μ TI 1 was disconnected from the inlet of the ¹D column so that it was open to the ambient air. Then the thermal desorption tube was heated to 300 °C for 5 min followed by heating μ TI 1 to 270 °C in 0.6 s and then kept at 250 °C for 30 s at a helium flow rate of 25 mL/min. This process was repeated twice in order to completely remove the residual analytes (if any) trapped in the thermal desorption tube and μ TI 1. Note that cleaning of μ TI 2A and 2B was not needed.

The total assay time was 33 minutes, which included 5 minutes of sample collection, 5 minutes of desorption/transfer, 13 minutes of separation, and 10 minutes of cleaning.

It should be noted that multiple μ PIDs were used to measure the analytes eluted from ¹D column and ²D columns. The responsivity of those μ PIDs may be different due to variations in aging and amplification, etc. During the experiment, μ PID 2A and 2B were calibrated against μ PID 1 using toluene (50 ppb), as discussed in detail in Ref. [1]. This calibration was carried out approximately every 300 hours of operation.

Operation of the portable GC as 1D GC is similar to the steps in (1)-(4) above, except that the inlet of the μ DS is detached from the outlet of the μ PID 1 or the 2nd-dimensional module is powered off.

Section S2: Two-dimensional gas chromatogram construction

Once the analysis is completed, two-dimensional gas chromatogram was generated from all three channels' (1D, 2A, 2B) PID signals. There were a total of 97 peaks found in about 800 seconds of 2D separation (~800 seconds of 1st-dimensional separation and 20 seconds of 2nd-dimensional separation). Each of the peak may represent only one analyte (no co-elution) or multiple analytes (co-elution). Note that the 2D chromatogram of a particular patient may contain only a subset of the 97 peaks. Also note that the volume (analyte mass) of each peak is normalized to the total peak volume of the entire 2D chromatogram, which is one of the most commonly used normalization techniques [2-7].

First, the signal from each channel was preprocessed for baseline correction and peak detection. Second, all the 2D peaks was traced back to their correspondence 1D peak based on the 1D period they were sampled from, so that both 1D and 2D retention time and peak shapes could be found. Third, the 1D chromatogram was aligned to the reference chromatogram to fix the 1D retention time drift. Finally, the two-dimensional gas chromatogram for each peak was generated by multiplying its 1D peak shape to its 2D peak shape. Adding all individual peak's two-dimensional gas chromatogram yielded the completed two-dimensional gas chromatogram. All algorithms were implemented in the Matlab™ programming environment with a user-friendly graphical interface.

- (1) Baseline correction: The baseline of gas chromatograms drift slowly due to column bleeding at high temperature, flow fluctuation and detector performance. This baseline drift can negatively affect the analytical results quantitatively hence should be corrected before performing further data analysis [8]. Here we use adaptive iterative reweighted Penalized Least Squares (airPLS) algorithms, developed by Zhang et al. [9], which iteratively changing weights of sum squares errors (SSE) between the original signals and fitted baseline until the termination criteria is met. This method requires no user intervention and has been applied to chromatograms, NMR and Raman spectra.
- (2) Peak detection: After the baseline correction, peak detection is applied to both 1D and 2D chromatograms. The peak apexes are found via the method developed by Morris et al. [10]. In this method, the signal is first denoised via wavelet regression using the undecimated discrete wavelet transform (UDWT), then searched for all local maxima and the associated peak endpoints. At last, the peaks that do not meet the peak height and FWHM criteria are eliminated

[10-14]. Once the peak apexes are found (including single and co-eluted peaks), the peak shapes are fitted by the exponentially modified Gaussian (EMG) model. This peak fitting method has been described previously¹.

- (3) Peak assignment: Within each modulation, the ¹D peak will be injected into either a 2A or 2B subsystem. Each ²D peak is assigned to one or multiple ¹D peak IDs, depending on the total peaks within each modulation. For each individual peak, multiplying its ¹D peak shape by its normalized ²D peak shape yields a two-dimensional chromatogram for this individual peak.
- (4) Peak alignment: Gas chromatograms may contain distortions of the retention time due to column aging, changes in temperature, or other unknown deviations in instrumental conditions. Fluctuations in retention times across various measurements obscure statistical analysis and thus discovery of relevant patterns in the data [15]. Since retention time shifts are observed in our ¹D chromatogram, we applied the correlation optimized wrapping (COW) algorithm [16] for peak alignment. This method is a piecewise or segmented data preprocessing method (operating on one sample record at a time) aimed at aligning a sample data vector towards a reference vector by allowing limited changes in segment lengths on the sample vector. The outcome of this method contains the correlation between the reference chromatogram retention time and the new chromatogram retention time. With this correlation, a shift time could be found for each peak based on its original retention time and the single peak two-dimensional gas chromatogram could shift on the ¹D based on this shift time.
- (5) Summation of individual two-dimensional chromatograms: Adding all individual peaks' two-dimensional chromatograms together after applying the shift time yields the complete two-dimensional chromatogram.

Section S3: Algorithm description

To distinguish ARDS and non-ARDS patients based on their breath chromatograms, linear discriminant analysis (LDA) was applied to find a linear function that could be used to separate these two groups. However, LDA can only be applied if the number of samples (patient chromatograms) is much larger than the number of features [2] (i.e., the number of VOCs, which was 97 in our study). To overcome this limitation and to decrease the computational complexity of the pattern classifier, principal component analysis (PCA) was applied prior to LDA to reduce the dimensionality of the feature space.

Since PCA is an unsupervised dimension reduction method, it only performs a linear mapping of the data to a lower-dimensional space in such a way that the variance of the data point is maximized. If we directly apply PCA to the overall VOC dataset, the VOCs relevant to ARDS may get overlooked and the interference VOCs that have bigger variance among patients will be kept. Therefore, to produce the best classification result with PCA-LDA, it is critical to find the features (VOCs) that are relevant to ARDS and discard all other interference features. The physical interpretation is that not all of those peaks in the 2D chromatogram may be relevant to ARDS. For example, some of the peaks may be from indoor air background, normal metabolic activities, or other conditions that a patient has. Those irrelevant peaks interfere in the correct classification of ARDS and non-ARDS groups. It is therefore critical to determine which subset of the peaks is most responsible for the differences observed between ARDS and non-ARDS groups.

S3.1. Generation of possible peak subsets

We first assume that there are a total of m different peaks found in all patients' 2D chromatograms with different quantities. For a particular patient not all m peaks are present. The quantities of those missing peaks are assigned to be 0. All m peaks and their quantities form the entire dataset can be expressed as:

$$\begin{pmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{p1} & \cdots & x_{pm} \end{pmatrix}, \quad (1)$$

where x_{ij} is the quantity of the j^{th} peak of the i^{th} patient. In total, there are m peaks and p patients. We further assume that there are N peaks relevant to ARDS and non-ARDS classification. Consequently, there are C_m^N possible peak combinations (subsets) that can be selected from the dataset in Eq. (1). One such subset can be written as:

$$\begin{pmatrix} x_{1k_1} & \cdots & x_{1k_N} \\ \vdots & \ddots & \vdots \\ x_{pk_1} & \cdots & x_{pk_N} \end{pmatrix}, \quad (2)$$

where (k_1, k_2, \dots, k_N) is the subset formed by N peaks.

S3.2. Criteria of peak subset selection

PCA was first used for data reduction of the N peaks VOC subsets and then LDA was applied to the primary two principal component scores for classification. The total accuracy (true positive plus true negative rate) of classification was used as the criteria for peak subset relevancy to ARDS. For each possible peak subset, PCA was first applied to the p -by- N dataset to produce p -by- N principal component scores. Then, the primary two principal component scores (p -by-2) and the classifier (1 as ARDS and 0 as non-ARDS) for each patient were used to train an LDA model and yield a linear boundary between the ARDS and non-ARDS groups. The total accuracy (number of patients falling in the correct side of the boundary divided by the total patients) was calculated and used as criteria of the relevancy of this VOC subset. Equations (3) and (4) illustrate the methods and processes described above.

$$\begin{pmatrix} x_{1k_1} & \cdots & x_{1k_N} \\ \vdots & \ddots & \vdots \\ x_{pk_1} & \cdots & x_{pk_N} \end{pmatrix} \xrightarrow{PCA} \begin{pmatrix} s_{11} & \cdots & s_{1N} \\ \vdots & \ddots & \vdots \\ s_{p1} & \cdots & s_{pN} \end{pmatrix}, \quad (3)$$

where s_{ij} is the j^{th} principal component score of the i^{th} patient.

$$\begin{pmatrix} s_{11} & s_{12} \\ \vdots & \vdots \\ s_{p1} & s_{p2} \end{pmatrix} \text{ with } \begin{pmatrix} c_1 \\ \vdots \\ c_p \end{pmatrix} \xrightarrow{LDA} \text{linear boundary} \xrightarrow{\text{yields}} \text{Classification accuracy}, \quad (4)$$

where c_i is the classifier (1 for ARDS and 0 for non-ARDS) of the i^{th} patient. Finally, the peak combinations (i.e., subsets) with highest accuracy were selected. For each of these selected peak combinations (subsets), the patients' coordinates on the PCA plot were decided by their principal component scores. The mean distance of the 20% patients closest to the boundary line, normalized by the mean distance, was calculated. The peak subset with highest boundary distance was chosen as the optimal peak subset.

S3.3. Iterative peak subset selection

Since human breath is a complex mixture, the total peak number m is large and the total number of possible combinations of peaks (i.e., peak subsets), $\sum_{N=1}^m C_m^N$, is enormous. As a result, it

requires a great amount of computation time to evaluate all the peak subsets. To expedite the selection process, in this work we started with peak subsets formed by a small number of peaks (e.g., n peaks, which resulted in C_m^n subsets to be evaluated). Once the most relevant peak subset was determined, more peaks were added to this selected subset, aiming to achieve higher accuracy.

Assuming that there are n' more peaks that are relevant to ARDS (n' is another small number of VOCs in order to save the computation time), $C_{m-n}^{n'}$ possible peak combinations (subsets) can be found and added to the previously optimized VOC subset to form a new peak subset, i.e.,

$$\begin{pmatrix} x_{1k_1} & \cdots & x_{1k_n} & x_{1l_1} & \cdots & x_{1l_{n'}} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_{pk_1} & \cdots & x_{pk_n} & x_{pl_1} & \cdots & x_{pl_{n'}} \end{pmatrix}, \quad (5)$$

where $(l_1, l_2, \dots, l_{n'})$ is the peak subset with n' peaks.

With the new peak subset, PCA and LDA were applied to calculate the accuracy and the boundary distance. If the classification accuracy increased or the boundary distance increased, those n' peaks were kept and more peaks out of $m-n-n'$ peaks would be added iteratively in the same manner described above. If accuracy no longer increased or the boundary distance no longer increased, then the iteration process was ended, and the final optimal peak subset was determined. The overall iteration processes are illustrated in Fig. S2.

S3.4. Training and testing with ARDS and non-ARDS patients

The entire patient data set was divided into two sets: training set (p patients) and testing set (q patients). The training set was used to select the optimal peak subset for best classification and determine the linear boundary, whereas the testing set was used to validate the determined optimal peak subset and the boundary on the PCA plot.

Assuming the final optimal peak subset containing N peaks. With the N peaks subset, the PCA analysis yields an N -by- N PCA coefficient and a linear boundary line between ARDS and non-ARDS groups.

$$\begin{pmatrix} x_{t_1k_1} & \cdots & x_{t_1k_N} \\ \vdots & \ddots & \vdots \\ x_{t_pk_1} & \cdots & x_{t_pk_N} \end{pmatrix} \xrightarrow{PCA} \begin{pmatrix} Coeff_{11} & \cdots & Coeff_{1N} \\ \vdots & \ddots & \vdots \\ Coeff_{N1} & \cdots & Coeff_{NN} \end{pmatrix} \text{ and } \begin{pmatrix} s_{t_11} & \cdots & s_{t_1N} \\ \vdots & \ddots & \vdots \\ s_{t_p1} & \cdots & s_{t_pN} \end{pmatrix} \quad (6)$$

where (t_1, t_2, \dots, t_p) is the training set patients; in total there are p patients in the training set.

$$\begin{pmatrix} s_{t_11} & s_{t_12} \\ \vdots & \vdots \\ s_{t_p1} & s_{t_p2} \end{pmatrix} \text{with} \begin{pmatrix} c_{t_1} \\ \vdots \\ c_{t_p} \end{pmatrix} \xrightarrow{LDA} \text{linear boundary} \quad (7)$$

With the N-by-N PCA coefficient acquired from the training set, the PC scores of the testing set can be calculated by multiplying the PCA coefficient to the N peak subset of all testing patients. With the linear boundary line acquired from the training set, the final classification accuracy can be calculated.

$$\begin{pmatrix} x_{v_1k_1} & \cdots & x_{v_1k_N} \\ \vdots & \ddots & \vdots \\ x_{v_qk_1} & \cdots & x_{v_qk_N} \end{pmatrix} \begin{pmatrix} Coeff_{11} & \cdots & Coeff_{1N} \\ \vdots & \ddots & \vdots \\ Coeff_{N1} & \cdots & Coeff_{NN} \end{pmatrix} \xrightarrow{yields} \begin{pmatrix} s_{v_11} & \cdots & s_{v_1N} \\ \vdots & \ddots & \vdots \\ s_{v_q1} & \cdots & s_{v_qN} \end{pmatrix}, \quad (8)$$

where (v_1, v_2, \dots, v_q) is the testing set patients; in total there are q patients in testing set;

$$\begin{pmatrix} s_{v_11} & s_{v_12} \\ \vdots & \vdots \\ s_{v_q1} & s_{v_q2} \end{pmatrix} \text{with linear boundary} \xrightarrow{yields} \text{final classification accuracy} \quad (9)$$

S3.5. Selection of the optimal subset of peaks relevant to ARDS

In our study, a total of $m=97$ peaks were found in 2D chromatograms. We first assumed that there are $n=4$ peaks relevant to classification of ARDS and non-ARDS. We found that the 4-peak subset of Peak #(2, 44, 72, 79) provides the best classification with the total accuracy of 88.4% (see the corresponding peaks on the 2D GC chromatogram in Fig. S3 and the PCA-LDA results in Fig. S4(a)). Then $n'=5$ peaks were added and we found that the 9-peak subset of [(2, 44, 72, 79) + (34, 38, 62, 66, 81)] provides the best classification with the total accuracy of 93.0% (see the corresponding peaks on a 2D GC chromatogram in Fig. S3 and the PCA-LDA results in Fig. S4(b)). Then another $n'=5$ peaks were added and we found that the 14-peak subset of [(2, 44, 72, 79) + (34, 38, 62, 66, 81) + (54, 61, 63, 71, 75)] provides the best classification with the total accuracy of 93.0% (see the corresponding peaks on a 2D GC chromatogram in Fig. S3 and the PCA-LDA results in Fig. S4(c)). Since the classification accuracy and the boundary distance does not improve from the 9-peak subset to the 14-peak subset (i.e., the ARDS and non-ARDS groups are not clustered/separated better), the 9-peak subset was selected as the final optimal peak subset, which consists of Peak #(2, 44, 72, 79, 34, 38, 62, 66, 81) in the 2D chromatogram.

These 9 peaks were tentatively identified by coupling our portable GC with Thermo Scientific Single Quadrupole Mass Spectrometer (ISQ™ Series) and analyzing with Chromeleon™ 7 Software. Their names, CAS numbers, and formulas are presented in Table S4.

S3.6. Receiver Operating Characteristic (ROC) curve analysis

With the LDA model acquired from the training set, the posterior probability can be calculated for any given \mathbf{S}_p :

$$P(ARDS|\mathbf{S}_p) = \frac{P(\mathbf{S}_p|ARDS)P(ARDS)}{P(\mathbf{S}_p|ARDS)P(ARDS)+P(\mathbf{S}_p|non-ARDS)P(non-ARDS)}, \quad (10)$$

where \mathbf{S}_p is the principal component scores (S_{p1} S_{p2}) for any given patient p. $P(ARDS)$ and $P(non-ARDS)$ are the prior probability (fraction of ARDS and non-ARDS patients within the training set), respectively. $P(\mathbf{S}_p|ARDS)$ and $P(\mathbf{S}_p|non-ARDS)$ are the ARDS and non-ARDS multivariate Gaussian distribution density functions, with $\boldsymbol{\mu}_{ARDS}$, $\boldsymbol{\mu}_{non-ARDS}$ being the means and $\boldsymbol{\Sigma}$ being the shared covariance matrix across ARDS and non-ARDS.

$$P(\mathbf{S}_p|ARDS) = \frac{1}{(2\pi|\boldsymbol{\Sigma}|)^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{S}_p-\boldsymbol{\mu}_{ARDS})^T \boldsymbol{\Sigma}^{-1}(\mathbf{S}_p-\boldsymbol{\mu}_{ARDS})}, \quad (11)$$

$$P(\mathbf{S}_p|non-ARDS) = \frac{1}{(2\pi|\boldsymbol{\Sigma}|)^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{S}_p-\boldsymbol{\mu}_{non-ARDS})^T \boldsymbol{\Sigma}^{-1}(\mathbf{S}_p-\boldsymbol{\mu}_{non-ARDS})}. \quad (12)$$

With the ARDS/non-ARDS labels and the posterior probability of the patients in the training set, testing set, and all patients, their ROC curves and the corresponding AUC (area under curve) were computed and shown in Fig. S7.

Section S4: Peak capacity of portable 2D GC

Table S1 lists the peak capacity estimated from three exemplary peaks. For GC \times GC, the peak capacity is defined as $n_{GC \times GC} = n_1 \times n_2$, where n_1 and n_2 are the peak capacity for 1D and 2D , respectively [1]. The conventional method for calculation of peak capacity using 4σ - bottom-to-bottom-width $w_{4\sigma}$ is given by: $n_{4\sigma} = (t_R^1/w_{4\sigma}^1)(CP_M/w_{4\sigma}^2)$, where C is the number of channels in 2D and P_M is the modulation period. The peak capacity value for three selected peaks is listed in the table below as $n_{4\sigma}$.

Table S1 Peak capacity for the portable 2D GC calculated for three exemplary peaks

| Peak # | t_R^1 | $w_{4\sigma}^1$ | t_R^2 | $w_{4\sigma}^2$ | $n_{4\sigma}$ |
|--------|---------|-----------------|---------|-----------------|---------------|
| 16 | 154.0 s | 4.8 s | 5.0 s | 1.5 s | 428 |
| 35 | 315.7 s | 6.6 s | 6.0 s | 2.5 s | 383 |
| 58 | 544.3 s | 7.0 s | 10.4 s | 6.8 s | 229 |

t_R^1 : 1D retention time

t_R^2 : 2D -dimensional retention time

$w_{4\sigma}^1$: 1D peak width (4σ - bottom-to-bottom)

$w_{4\sigma}^2$: 2D peak width (4σ - bottom-to-bottom)

C: Number of 2D channels (in our case: C = 2)

P_M : modulation time (in our case: $P_M = 10$ s)

Section S5: Patient medical history during time series testing dates

Patient #2 was a healthy subject tested for 4 days.

Patient #3 was a healthy subject tested for 4 days.

Patient #7 was sampled for 4 days and had ARDS by the Berlin Criteria since the 1st testing day.

No signs of recovery at least 4 days after the last testing day.

Patient #11 was a potential and undetermined ARDS patient on the 1st test day and then upgraded to ARDS on the next day. By the Berlin Criteria he/has ARDS for all 3 days.

Patient #12 was suspected for pneumonia on the 1st testing day. This patient was tested for 3 days and no ARDS was developed during this period based on the Berlin Criteria.

Patient #27 was a potential and undetermined ARDS patient on the 1st test day and then upgraded to ARDS on the next day. Based on the Berlin Criteria he/has ARDS for all 3 days.

Patient #30 had pneumonia and ARDS based on the Berlin Criteria since 1st test day and was tested for 3 days. No signs of recovery and was shifted to comfort care after the last testing day.

Patient #31 had acute respiratory failure on the 1st testing day but no ARDS based on the Berlin Criteria during the 2 testing days.

Patient #34 had hypoxemic respiratory failure on the 1st testing day but no ARDS based on the Berlin Criteria during the 3 testing days.

Patient #35 had no ARDS based on the Berlin Criteria during the 2 testing days.

Patient #36 had ARDS since the 1st sampling day. On the 3rd day the patient was still listed as ARDS patients based on the Berlin Criteria and then got extubated and discharged from ICU on the 5th day.

Patient #38 was a healthy subject tested for 3 days.

Patient #39 had ARDS based on the Berlin Criteria during the 2 testing days. No signs of recovery.

Patient #40 had pneumonia on the 1st testing day but no ARDS based on the Berlin Criteria during the 2 testing days.

Patient #42 had ARDS based on the Berlin Criteria during the 2 testing days. No signs of recovery.

Patient #45 had ARDS based on the Berlin Criteria during the 5 testing days. No signs of recovery and was shifted to comfort care on the last testing day.

Patient #46 had no ARDS based on the Berlin Criteria during the 3 testing days.

Patient #47 was sampled for 4 days and got extubated and discharged on the 6th day. Based on the Berlin Criteria this patient had ARDS for all first 4 days.

References

1. Lee J, Zhou M, Zhu H, Nidetz R, Kurabayashi K, Fan X. Fully Automated Portable Comprehensive 2-Dimensional Gas Chromatography Device. *Anal Chem*. 2016;88:10266-74.
2. Smolinska A, Hauschild AC, Fijten RRR, 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;8:027105.
3. Wang C, Li M, Jiang H, Tong H, Feng Y, Wang Y, et al. Comparative Analysis of VOCs in Exhaled Breath of Amyotrophic Lateral Sclerosis and Cervical Spondylotic Myelopathy Patients. *Scientific Reports*. 2016;6:26120.
4. Pereira J, Porto-Figueira P, Cavaco C, Taunk K, Rapole S, Dhakne R, et al. Breath Analysis as a Potential and Non-Invasive Frontier in Disease Diagnosis: An Overview. *Metabolites*. 2015;5(1):3-55.
5. Smolinska A, Klaassen EMM, Dallinga JW, van de Kant KDG, Jobsis Q, Moonen EJC, et al. Profiling of Volatile Organic Compounds in Exhaled Breath As a Strategy to Find Early Predictive Signatures of Asthma in Children. *PLOS ONE*. 2014;9(4):e95668.
6. Wang C, Feng Y, Wang M, Pi X, Tong H, Wang Y, et al. Volatile Organic Metabolites Identify Patients with Mesangial Proliferative Glomerulonephritis, IgA Nephropathy and Normal Controls. *Scientific Reports*. 2015;5:14744.
7. Lau H-C, Yu J-B, Lee H-W, Huh J-S, Lim J-O. Investigation of Exhaled Breath Samples from Patients with Alzheimer's Disease Using Gas Chromatography-Mass Spectrometry and an Exhaled Breath Sensor System. *Sensors (Basel, Switzerland)*. 2017;17(8):1783.
8. Andrade L, Manolakos ES. Signal background estimation and baseline correction algorithms for accurate DNA sequencing. *J VLSI Sig Proc Syst*. 2003;35:229-43.
9. Zhang ZM, Chen S, Liang YZ. Baseline correction using adaptive iteratively reweighted penalized least squares. *Analyst*. 2010;135:1138-46.
10. Morris JS, Coombes KR, Koomen J, Baggerly KA, Kobayashi R. Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics*. 2005;21:1764-75.
11. Yasui Y, Pepe M, Thompson ML, Adam BL, Wright Jr. GL, Qu Y, et al. A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. *Biostatistics*. 2003;4:449-63.
12. Donoho DL, Johnstone IM. Adapting to unknown smoothness via wavelet shrinkage. *J Am Stat Assoc*. 1995;90:1200-24.
13. Strang G, Nguyen T. *Wavelets and Filter Banks*. Wellesley, MA: Wellesley-Cambridge Press; 1996.
14. Coombes KR, Tsavachidis S, Morris JS, Baggerly KA, Hung MC, Kuerer HM. Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. *Proteomics*. 2005;5:4107-17.
15. Liang YZ, Xie PS, F. Chau. Chromatographic fingerprinting and related chemometric techniques for quality control of traditional Chinese medicines. *J Sep Sci*. 2010;33:410-21.
16. Tomasi G, van den Berg F, Andersson C. Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *J Chemom*. 2004;18:231-41.

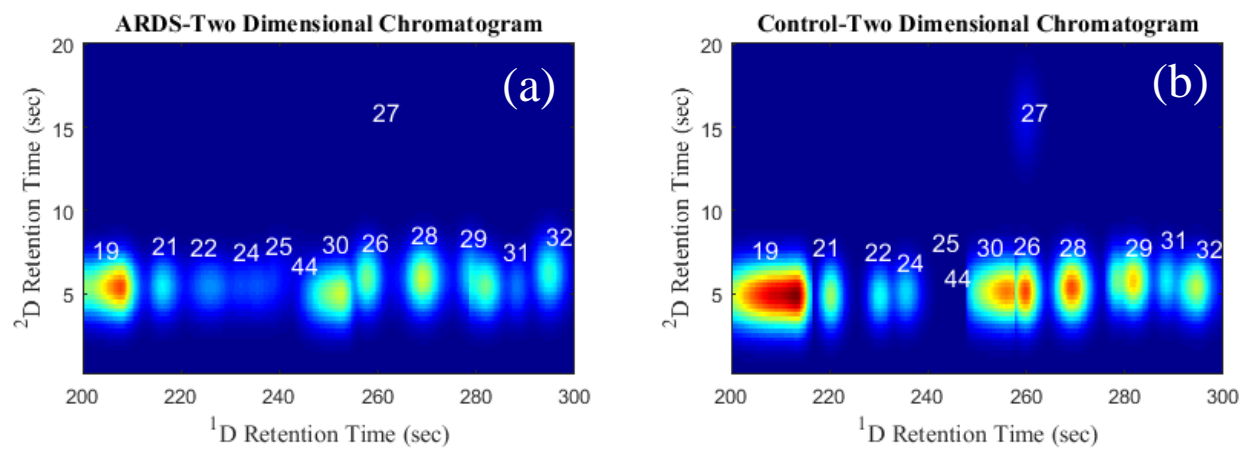


Fig. S1 (a) Zoomed-in portion of Fig. 4(b). (b) Zoomed-in portion of Fig. 4(d)

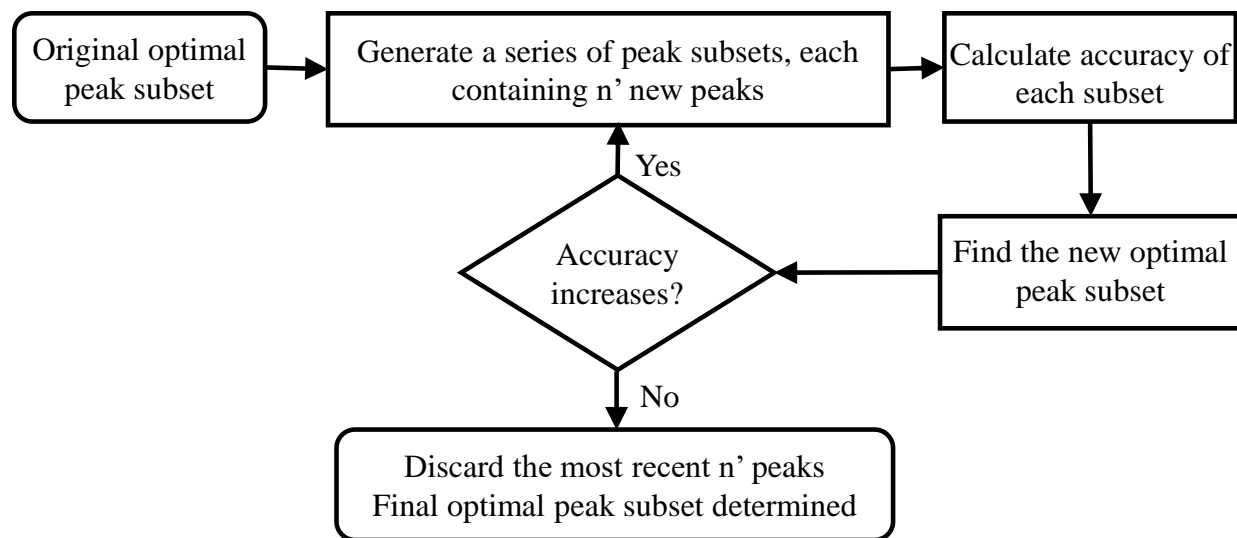


Fig. S2 Iterative peak subset selection procedure

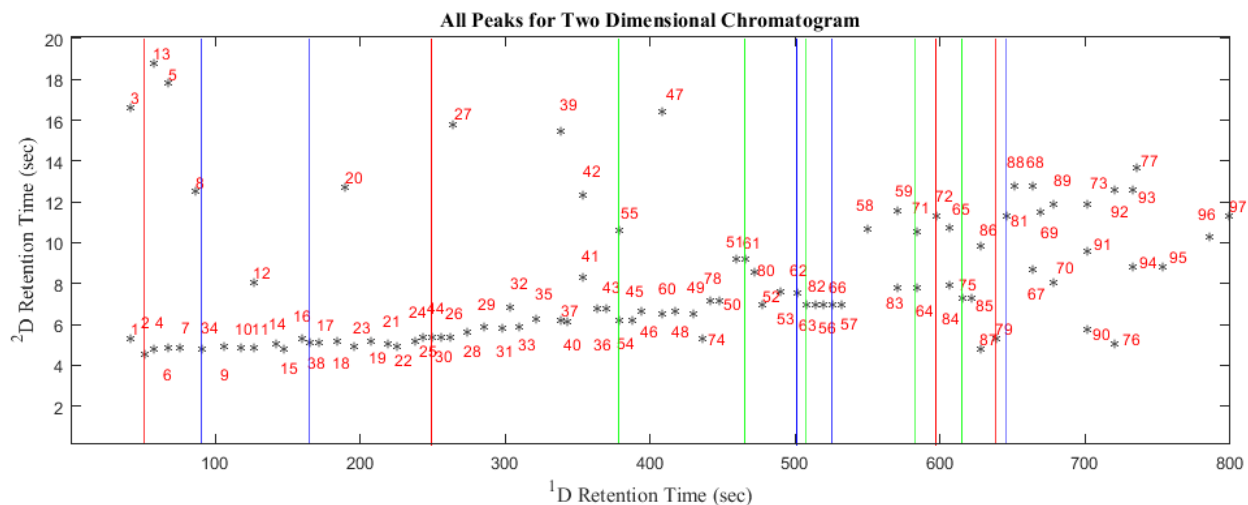


Fig. S3 Selection of the optimal subset of peaks relevant to ARDS. Red lines mark the ¹D retention time of the 4 peaks selected in the first iteration. Blue lines mark the ¹D retention time of the additional 5 peaks selected in the second iteration. Green lines mark the ¹D retention time of the additional 5 peaks selected in the third iteration. Peak #34 in the 9-peak subset nearly co-elutes with Peak #8. Peak #54 and #71 in the 14-peak subset co-elutes with Peak #55 and Peak #64, respectively

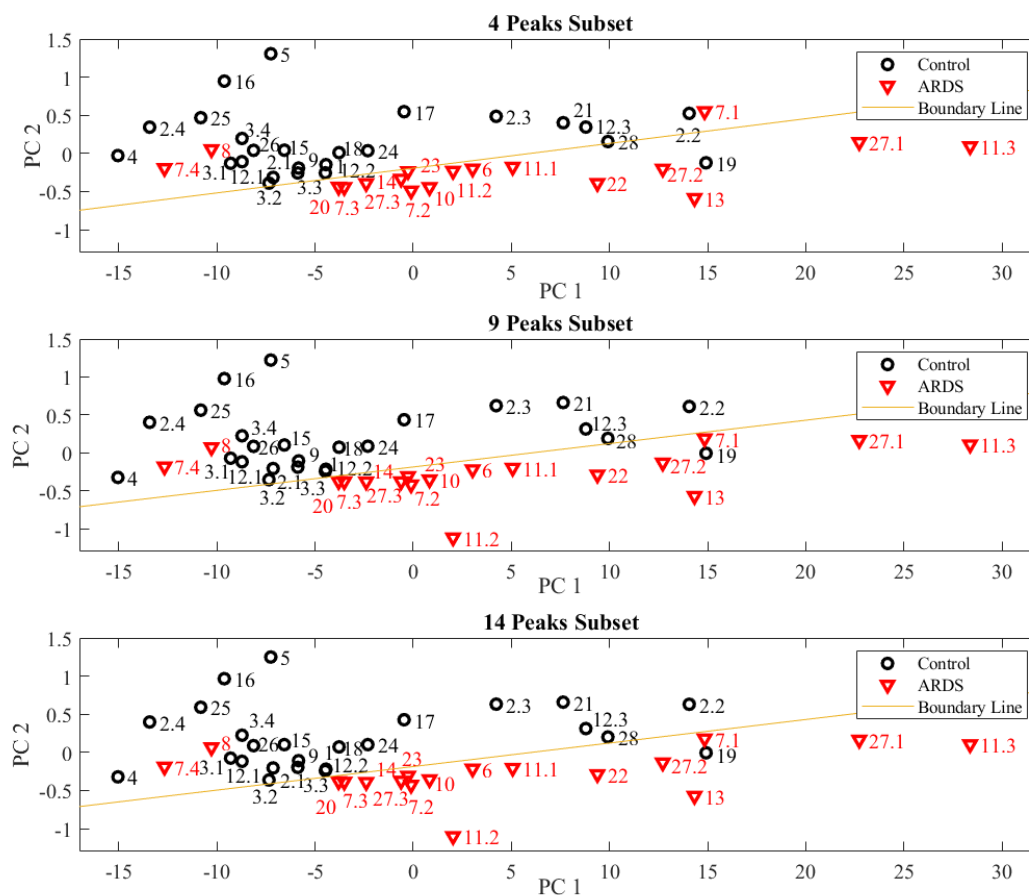


Fig. S4 PCA plots using the subset containing 4 peaks, 9 peaks, and 14 peaks for the training set of patients. The red and black symbols denote respectively the ARDS and non-ARDS patients adjudicated by physicians using the Berlin criteria. The patient numbers are given by the symbol. For example, “11.1” and “11.3” denote Patient #11, Day 1 and Day 3 results, respectively. The bottom/top zone below/above the boundary line represents respectively the ARDS/non-ARDS region using the breath analysis method

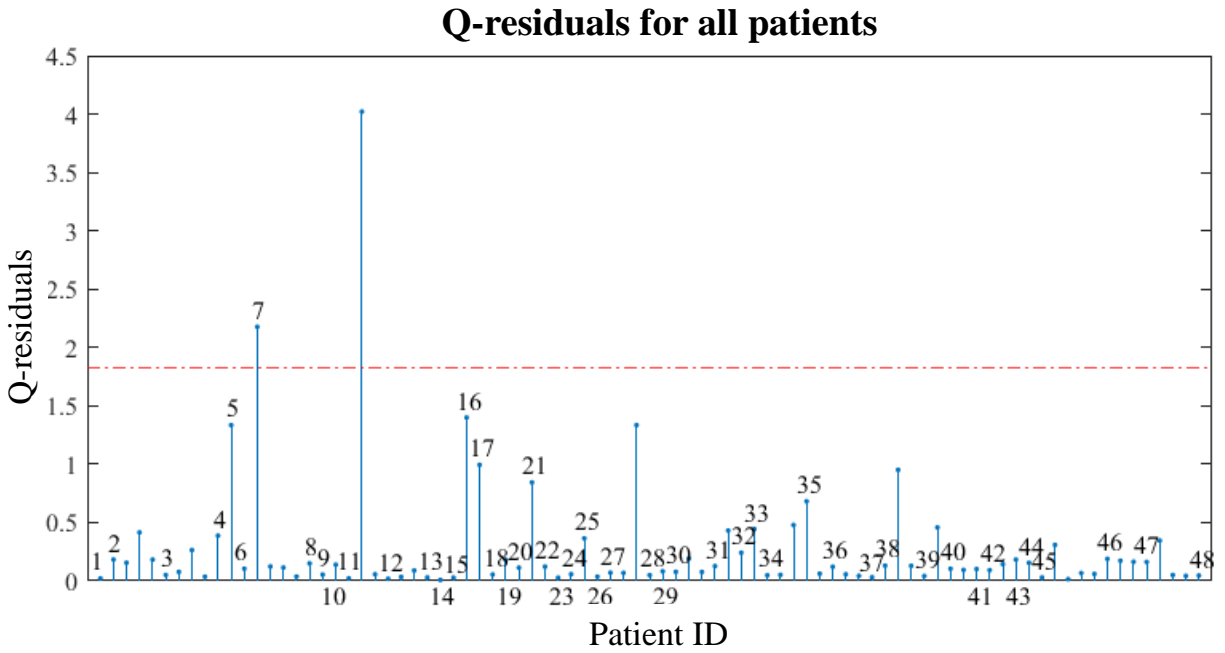


Fig. S5 Q-residuals of the PCA model (Figure 7) for all recruited patients. For the patients with time series tests, only the 1st test day is marked with the patient ID. The red dashed curve shows the 99% confidence level

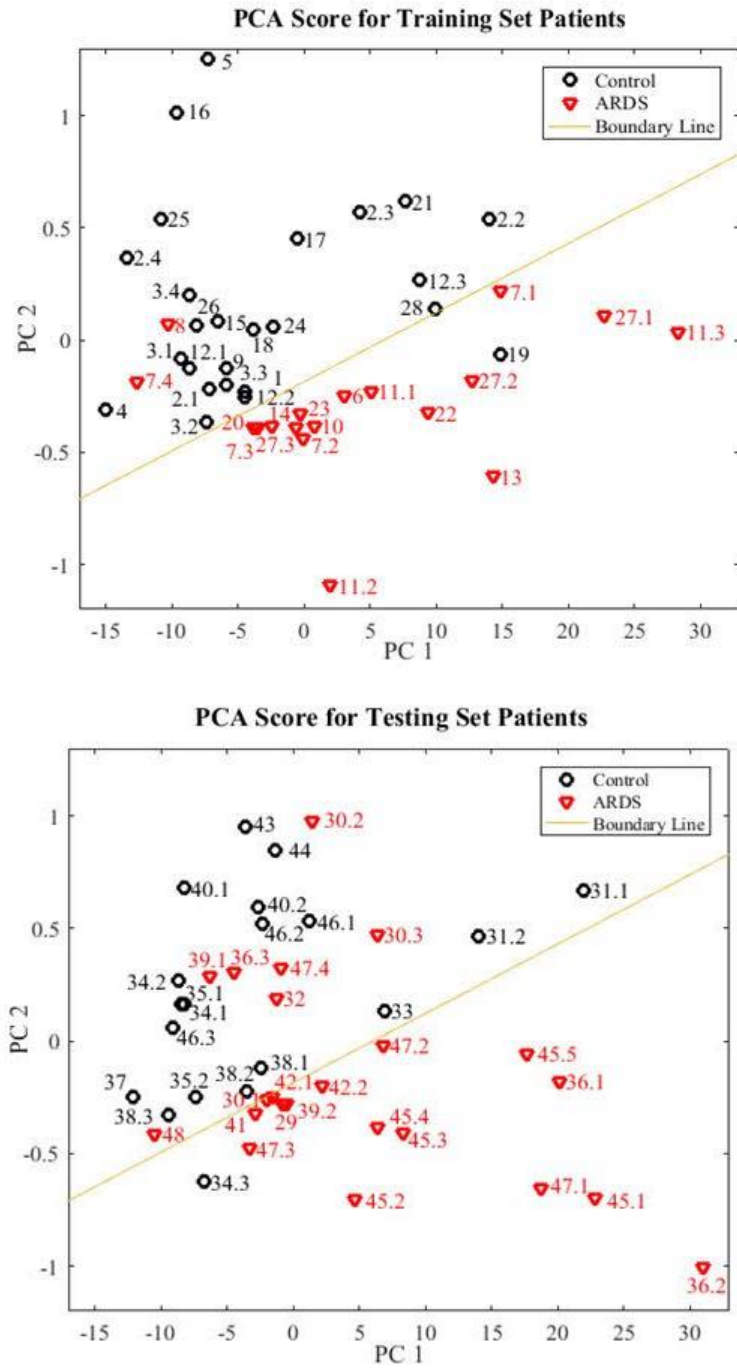


Fig. S6 PCA plot for the training and testing set of patients. The corresponding statistics is given in Table S2

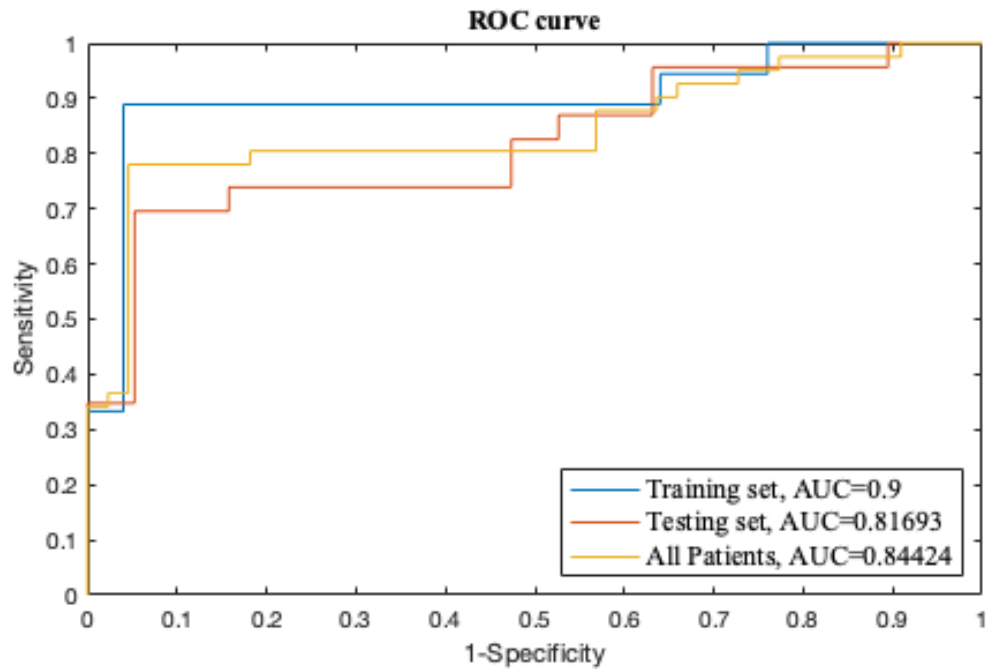


Fig. S7 Receiver operating characteristic (ROC) curves for the training set, testing set, and all patients

PCA Score for Patients with Time Series Tests

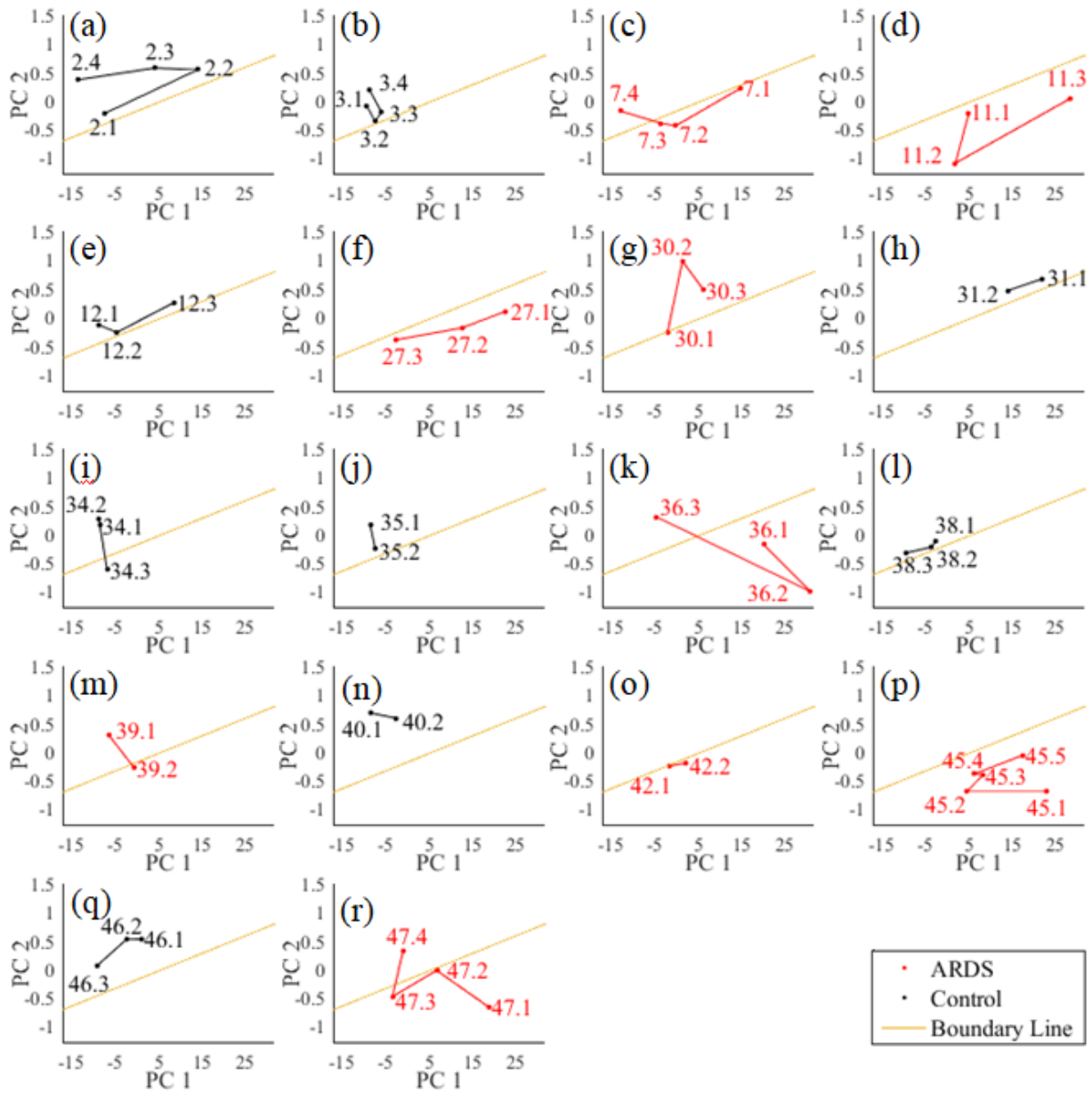


Fig. S8 Individual trajectories of all 18 patients with time series tests on the PCA plot. Refer to Section S5 for the patient medical history

Table S2 Statistics for the training and testing set of patients

| Training Set | ARDS | Non-ARDS | Total |
|----------------------------------|--------|----------|-------|
| Positive (our results) | 16 | 1 | 17 |
| Negative (our results) | 2 | 24 | 26 |
| Column total | 18 | 25 | 43 |
| | | | |
| Specificity | 96.0% | | |
| Sensitivity | 88.9 % | | |
| Positive predictive value | 94.1% | | |
| Negative predictive value | 92.3% | | |
| Total accuracy | 93.0% | | |
| | | | |
| Testing Set | ARDS | Non-ARDS | Total |
| Positive (our results) | 16 | 1 | 17 |
| Negative (our results) | 7 | 18 | 25 |
| Column total | 23 | 19 | 42 |
| | | | |
| Specificity | 94.7 % | | |
| Sensitivity | 69.6 % | | |
| Positive predictive value | 94.1% | | |
| Negative predictive value | 72.0% | | |
| Total accuracy | 80.9% | | |

Table S3 Statistics for the 4-fold cross-validation

| Cross Validation - 4 fold | Model 1 | Model 2 | Model 3 | Model 4 |
|----------------------------------|---------|---------|---------|---------|
| Specificity | 93.2% | 93.2% | 93.2% | 93.2% |
| Sensitivity | 78.0% | 75.6% | 78.0% | 75.6% |
| Positive predictive value | 91.4% | 91.2% | 91.4% | 91.2% |
| Negative predictive value | 82% | 80.4% | 82% | 80.4% |
| Total accuracy | 85.9% | 84.7% | 85.9% | 84.7% |

Table S4 Tentative chemical identification for the 9-peak subset

| Peak ID | Chemical Name | CAS Number | Formula |
|---------|---------------------------|------------|-----------------------------------|
| 2 | Pentane, 2-methyl- | 107-83-5 | C ₆ H ₁₄ |
| 44 | Heptane, 3-methyl- | 589-81-1 | C ₈ H ₁₈ |
| 72 | Heptane, 2,3,5-trimethyl- | 20278-85-7 | C ₁₀ H ₂₂ |
| 79 | 2,2,7,7-Tetramethyloctane | 1071-31-4 | C ₁₂ H ₂₆ |
| 34 | Pentane, 2,4-dimethyl- | 108-08-7 | C ₇ H ₁₆ |
| 38 | Cyclohexane, methyl- | 108-87-2 | C ₇ H ₁₄ |
| 62 | α -Pinene | 80-56-8 | C ₁₀ H ₁₆ |
| 66 | 3-Octene, 2,2-dimethyl- | 86869-76-3 | C ₁₀ H ₂₀ |
| 81 | 1-Decanol, 2-ethyl- | 21078-65-9 | C ₁₂ H ₂₆ O |