SUPPORTING INFORMATION

Covid-19 Automated Diagnosis and Risk Assessment through Metabolomics and Machine Learning

Jeany Delafiori1,2, Luiz Cláudio Navarro2,2, Rinaldo Focaccia Siciliano3,4, Gisely Cardoso de Melo5,6, Estela Natasha Brandt Busanello7, José Carlos Nicolau4, Geovana Manzan Sales1, Arthur Noim de Oliveira1, Fernando Fonseca Almeida Val5,6, Diogo Noim de Oliveira1, Adriana Eguti7, Luiz Augusto dos Santos8, Talia Falcão Dalçóquio9, Adriane Justi Bertolin6, Rebeca Linhares Abreu-Netto5,6, Rocio Salsoso4, Djane Baía-da-Silva5,6, Fabiana G Marcondes-Braga4, Vanderson Souza Sampaio5,6, Carla Cristina Judice10, Fabio Trindade Maranhão Costa10, Nelson Duran11, Mauricio Wesley Perroud7, Ester Cerdeira Sabino12, Marcus Vinicius Guimarães Lacerda5,13, Leonardo Oliveira Reis14, Wagner José Fávaro11, Wuelton Marcelo Monteiro5,6, Anderson Rezende Rocha2,10 and Rodrigo Ramos Catharino1,2

1 Innovare Biomarkers Laboratory, School of Pharmaceutical Sciences, University of Campinas, Campinas, Brazil - Rua Cinco de Junho, 350 - 13083-970 - Cidade Universitária Zeferino Vaz, Campinas/SP - Brazil
2 RECOD Laboratory, Computing Institute, University of Campinas, Campinas, Brazil - Rua Saturnino de Brito, 573 - 13083-852 - Cidade Universitária Zeferino Vaz, Campinas/SP - Brazil
3 Clinical Division of Infectious and Parasitic Diseases, University of São Paulo Medical School, Brazil - Av. Dr. Arnaldo, 455 - 01246-903 - Cerqueira César, São Paulo/SP, Brazil
4 Instituto do Coracao (InCor), Hospital das Clinicas HCFMUSP, Faculdade de Medicina, Universidade de São Paulo, São Paulo, Brazil - Av. Dr. Enéas de Carvalho Aguiar, 44 - 05403-900 – Cerqueira César, São Paulo/SP, Brazil
5 Tropical Medicine Foundation Dr. Heitor Vieira Dourado, Manaus, Brazil - Av. Pedro Teixeira, s/n - 69040-000 - Dom Pedro, Manaus/AM - Brazil
6 Amazonas State University, Manaus, Brazil - Av. Pedro Teixeira, nº 25 - 69040-000 - Dom Pedro, Manaus/AM - Brazil
7 Sumaré State Hospital, Sumaré, Brazil - Av. da Amizade, 2400 - 13175-490 - Jardim Bela Vista, Sumaré/SP, Brazil
8 Paulinia Municipal Hospital, Paulinia, Brasil - Rua Miguel Vicente Cury, 100 - 13140-000 - Nova Paulinia, Paulinia/SP, Brazil
9 Health Surveillance Foundation of Amazonas State, Manaus, Brazil - Av. Torquato Tapajós, 4010 - 69093-018 - Colônia Santo Antônio, Manaus/AM, Brazil
10 Laboratory of Tropical Diseases, Institute of Biology, University of Campinas, Campinas, Brazil - Rua Bertrand Russel, s/n - 13083-970 - Cidade Universitária Zeferino Vaz, Campinas/SP, Brazil
11 Laboratory of Urogenital Carcinogenesis and Immunotherapy, University of Campinas, Campinas, Brazil - Av. Bertrand Russel, s/n - 13083-865 – Cidade Universitária Zeferino Vaz, Campina/SP, Brazil
12 Institute of Tropical Medicine, University of São Paulo, São Paulo, Brazil - Avenida Dr. Enéas Carvalho de Aguiar, 470 - 05403-000 – Cerqueira César, São Paulo/SP, Brazil
13 Leónidas and Maria Deane Institute, FIOCRUZ, Manaus, Brazil - Rua Terezina, 476 - 69057-070 - Adrianópolis, Manaus/AM, Brazil
14 UroScience Laboratory, University of Campinas, Campinas, Brazil - Rua Tessália Vieira de Camargo, 126 - 13083-887 - Cidade Universitária Zeferino Vaz, Campinas/SP, Brazil

*corresponding author e-mail: Rodrigo Ramos Catharino, rrc@unicamp.br; Anderson Rezende Rocha, anderson.rocha@ic.unicamp.br

Table of Contents

Mass spectra data pre-processing ................................................................. S2
Algorithms selection and tuning ............................................................... S4
Markers importance ranking ................................................................. S6
Table S1. Performance metrics for tree-based algorithms ....................... S8
Table S2. Performance metrics for PLS and SVM algorithms ............... S8
Table S3. Proposed biomarkers to COVID-19 diagnosis model (M1) ........ S9
Table S4. Proposed biomarkers to risk assessment model (M2) ............. S11
Table S5. Proposed biomarkers to low-risk discrimination model (M3) ... S14
Models markers combination – Table S6. Confusion matrix ................. S16
Mass spectra data pre-processing

The pre-processing of mass spectrometry data acquired has the following steps:

1. The scan points included in each .RAW file (Thermo Fisher) are extracted using the Python interface developed by François Allain wrapping Thermo DLLs provided in https://pypi.org/project/pymsfilereader/. We configured the spectrometer to acquire 50 full scans for each replicate, and we acquired 10 replicates per patient sample. Figure 1 below explains the scan points. Scan points are stored in a JSON file with our lab-internally specified format, which will be complemented with the further steps of identification of peaks and final features.

![Scan points of collected raw data.](image)

2. The scan points are then processed looking for spectral peaks, fitting a Gaussian in each peak. Peaks which do not have at least 50% (configurable) of the scan points are considered noise (Figure 2). For the peaks extracted measurements we store: maximum intensity, gaussian peak, FWHM (full wide at half maximum), area. Those data points are also stored in a new expanded JSON file (which includes the previous scans and allows us to visualize all the processing steps made for each peak, if necessary).
3. Using the FWHM collected from all the spectrum, we compute the average peak resolution of the spectrometer, describing it as a linear function of the \( m/z \).

4. After collecting and describing the peaks of each spectrum, we compare the 10 replicates of the patient, discarding peaks that are not consistent across the replicates. Peaks are considered the same if they are superposed in their ranges of FWHM. For comparison, the replicates are normalized with maximum intensity equal to one (for each replicate). Two thresholds are defined for the RMSE error: the feature RMSE threshold and the vector RMSE threshold.

- If a normalized peak has median zero upon the replicates of the same patient, it is considered noise and deleted of all replicates.

- If a peak has the standard deviation of the intensity (measured by the gaussian maximum) greater than feature RMSE threshold then it is considered noise and deleted of all replicates.

- After the two steps above, if the RMSE error of a replicate, computed against the linear regression of all peak values of all replicates of the same patient, is above the vector RMSE threshold, then the replicate is considered to be acquired with bad quality and the entire replicate is discarded.
The previous steps guarantee that the data acquired from the samples are consistent and robust to random noise. Other types of noise and more consistent interferences still can be present and will be eliminated by the machine-learning process.

After identifying the peaks consistently within the replicates of the same patients, we build the bag of peaks (a concept like the bag of words in a text processing). All peaks from all spectra are collected and aligned using the spectrometer resolution at the \( m/z \). Peaks which are superposed at the resolution range are considered the same and represented by their \( m/z \) weighted average (using intensity as the weight). The \( m/z \) representatives in the bag of peaks are considered to be the final features. The feature vector matrix is composed \( F = [f_{i,j}] \) where \( i \) is the index of the \( m/z \), \( j \) is the index of the vector (replicate) and \( f_{i,j} \) is the intensity of the peak represented by the \( m/z(i) \) in the vector(j).

A final pre-processing annotation is performed over the feature vector matrix, discarding features whose number of non-zero intensities are below 10% of the number of vectors. Mass spectra pre-processing data is compiled and available via Zenodo link: [https://doi.org/10.5281/zenodo.4329381](https://doi.org/10.5281/zenodo.4329381).

**Algorithms selection and tuning**

The most important principles of our method are:

1. The ability to compute the feature importance on the predictive model generated by the method; and
2. the optimization process of recursively using the feature importance, discarding less important ones until achieving maximum predictive results, which identifies the features that carry most information about the phenomena under analysis.

With those principles in mind, we devised methods to implement them. For that, we adopted initially decision trees to estimate the variable importance. PLS is also a method which allows us to compute the variable importance in projection (VIP). Then we rely upon methods using combinations of trees such as Random Forests, Extreme Random Forests, Gradient Boosting Decision Forests (XGBoost), ADA Boosting Decision Forests and PLS. All of them seek to implement principle #1 above and point out the most important features for decision-making allowing us to recursively evaluate a set of features, eliminate the less important ones, and iterate (principle #2 above).

In our experiments during the method's design, ADA boosting slightly outperforms its counterparts on selecting the most important features, immediately followed by XGBoost. Extreme Random Forest and Random Forest come next, and PLS prediction performance depends on the linearity of the relations between variables, but it is the best for visualization purposes of the separation between classes in the projection space. Therefore, we adopted ADA boosting for the feature selection, PLS for visualization and XGBoost for
deployment (final classification) if it outperforms ADA in the final results. We note that adopting the two methods and choosing the one with higher confidence in a classification is straightforward as each method gives a classification along with its confidence in doing so.

SVM (Support Vector Machines) are used only at the end of the process to evaluate the prediction performance of the final variables with a completely different algorithm just to assess the richness of the identified features. We hypothesized that if the found features are robust and discriminatory an independent classifier would also offer good classification results, as it is the case here. As an observation, we opted NOT to consider SVM as one candidate for feature selection because this method is not straightforward to be adapted for this intent as it relies upon a transformed geometrical space. The main optimization algorithm can be expressed as a recursive application of the following steps:

1. Compute the best parameters for the current vector length using either random or grid search
   a. Train and validate the algorithm in 10 rounds with the fitting dataset split into training patients (all replicates together) and validate patients. Compute the mean and standard deviations of the confusion matrix metrics considering the metrics Accuracy, Sensitivity, Specificity, Precision, F1-Score, and MCC for a broader view of the performance.
   b. Compute the feature importance by averaging the feature importance in each of the 10 trained models.
   c. Discard features that represent less than 5% of the total summed importance.

   These steps are recursively performed until no features remain in the vector. The length associated with the maximum MCC (we adopt MCC for reference as it is symmetrical on positive and negative class metrics) is chosen and features selected by this method are the candidates of putative biomarkers.

   ![Graph of Most Discriminant Features](image)

   **FIGURE S4. Recursive fitting.**

Candidates of putative biomarkers are then analyzed by the values distribution and also the m/z are searched in the metabolomics databases verifying if they have a relation with metabolisms which can be applied to the phenomena under analysis. As an example, if there is an m/z related to a known medicine commonly given to patients, it could appear as an important prediction feature, and we must eliminate it from the experiment (depending on the case, restarting the ML method from the beginning eliminating this feature from the feature vectors). This is a selection process based on the prediction power of the molecules present on the samples (represented by the m/z relative intensity measured by the spectrometer).
After selecting putative biomarkers, the process is repeated for the paired features, which computes the relation between the putative biomarkers. The rationale here is to increase robustness to the analyzed biomarkers with respect to the acquisition equipment (e.g., different possible associated acquisition noise). In this way, each biomarker acts as an internal standard for the other biomarkers, making the spectrum independent from equipment and the relative scale, as the prediction model will work over features that represent the relation between biomarkers, mainly on biomarkers which are up-regulated against the down-regulated (in this case, the variation is amplified by the opposite direction of the single biomarkers variation).

Markers importance ranking

ΔJ is a metric we present to meet the requirements for measuring the potential of a single m/z to differentiate the classes of interest (e.g., control v. covid-19) as the probability to present higher intensity values above the median of each class. This measure compares the integral of the probability function (which is the cumulative distribution function CDF) using as a reference point the median of each class. In the ΔJ definition, there is an important condition that above the median, the difference of CDFs keeps the signal, guaranteeing that above the median of the selected class, the cumulative probability is always higher. We can consider ΔJ as a more restrictive measure than Kolmogorov–Smirnov two samples test (which we apply before computing ΔJ).

Machine-learning algorithms perform a multivariate statistical analysis to determine the prediction model at the end, which considers the features selected. On the other hand, for biomarkers, we want to see molecules that present a higher or lower value in the conditions, then ΔJ is a way of measuring the univariate independent contribution of the variable to the model (Figure 5).

We proposed an early version for ΔJ use for Paracoccidioidomycosis diagnosis (https://doi.org/10.1128/mSystems.00258-20).
FIGURE S5. Feature distribution using $\Delta J$. 

Feature m/z 553.3 ranked #0 values distribution

Feature m/z 806.5 ranked #9 values distribution

Cumulative distributions comparison
Table S1. Performance metrics using pairwise features on 10 validation tests for tree-based algorithms.

| Algorithm | GDB   | M1  | M2  | M3  | ADA   | M1  | M2  | M3  | RF    | M1  | M2  | M3  | XRF   | M1  | M2  | M3  |
|-----------|-------|-----|-----|-----|-------|-----|-----|-----|-------|-----|-----|-----|-------|-----|-----|-----|
| Vector length | 39   | 32  | 29  | 39  | 32   | 29  | 39  | 32  | 29    | 39  | 32  | 29  | 39    | 32  | 29  | 29  |
| # of Estimators | 260 (3) | 260 (3) | 260 (3) | 260 (3) | 260 (3) | 132 (3) | 20 (3) | 132 (3) | 132 (3) | 36 (3) | 68 (3) |
| TN         | 90 (3) | 37 (2) | 40 (2) | 89 (4) | 38 (2) | 40 (2) | 87 (3) | 35 (3) | 37 (2) | 87 (3) | 35 (2) | 36 (2) |
| FP         | 5 (2)  | 6 (2)  | 2 (1)  | 6 (3)  | 5 (2)  | 2 (1)  | 8 (3)  | 8 (3)  | 5 (2)  | 7 (3)  | 8 (2)  | 6 (3)  |
| FN         | 4 (2)  | 5 (2)  | 3 (1)  | 5 (2)  | 4 (2)  | 3 (1)  | 9 (3)  | 7 (2)  | 4 (1)  | 9 (2)  | 7 (2)  | 3 (1)  |
| TP         | 101 (4) | 33 (2) | 31 (2) | 101 (3) | 33 (3) | 31 (2) | 97 (5) | 31 (3) | 30 (2) | 96 (4) | 31 (3) | 31 (2) |
| Accuracy (%) | 95.6 (1.1) | 87.1 (2.7) | 93.0 (2.3) | 94.6 (1.7) | 88.7 (3.2) | 93.4 (1.8) | 91.9 (1.7) | 81.9 (3.0) | 88.5 (3.7) | 91.7 (1.3) | 82.1 (3.5) | 88.0 (3.3) |
| Sensitivity (%) | 95.9 (1.8) | 87.4 (4.8) | 91.6 (4.0) | 95.4 (1.6) | 88.1 (4.6) | 91.8 (3.1) | 91.7 (3.0) | 82.1 (6.0) | 88.5 (3.3) | 91.0 (1.9) | 82.1 (6.2) | 91.1 (4.2) |
| Specificity (%) | 95.2 (2.1) | 86.8 (4.4) | 94.5 (2.3) | 93.8 (3.5) | 89.3 (4.7) | 95.0 (2.4) | 92.1 (2.9) | 81.6 (6.1) | 88.4 (5.1) | 92.4 (2.7) | 82.0 (5.2) | 85.0 (5.6) |
| Precision (%) | 95.3 (1.9) | 87.1 (3.7) | 94.3 (2.2) | 94.1 (3.1) | 89.3 (4.2) | 94.9 (2.3) | 92.1 (2.6) | 82.1 (4.5) | 88.6 (4.7) | 92.4 (2.4) | 82.2 (4.1) | 86.1 (4.4) |
| F1 Score (%) | 95.6 (1.1) | 87.1 (2.7) | 92.9 (2.5) | 94.7 (1.7) | 88.6 (3.2) | 93.3 (1.9) | 91.9 (1.7) | 81.9 (3.1) | 88.5 (3.5) | 91.7 (1.3) | 82.0 (3.8) | 88.4 (3.0) |
| MCC        | 0.91 (0.02) | 0.74 (0.05) | 0.86 (0.04) | 0.89 (0.03) | 0.78 (0.06) | 0.87 (0.04) | 0.84 (0.03) | 0.77 (0.06) | 0.83 (0.07) | 0.84 (0.03) | 0.84 (0.07) |

Numbers correspond to individual’s classification average and standard deviations in parenthesis. Abbreviations: ADA – ADA Boosting; GDB – Gradient Tree Boosting; RF – Random Forest; XRF – eXtreme Random Forest; FN – False negative; TN – True negative; TP – True positive; MCC – Mathew’s Correlation Coefficient.

Table S2. Performance metrics using pairwise features on 10 validation tests for PLS and SVM algorithms.

| Algorithm | PLS | SVM |
|-----------|-----|-----|
| Vector length | 39  | 32  |
| TN         | 82 (6) | 36 (2) |
| FP         | 13 (4) | 6 (2) |
| FN         | 6 (2)  | 8 (2)  |
| TP         | 99 (3)  | 30 (3)  |
| Accuracy (%) | 90.3 (2.1) | 82.5 (2.4) |
| Sensitivity (%) | 94.1 (2.2) | 79.6 (5.5) |
| Specificity (%) | 86.6 (4.5) | 85.3 (4.4) |
| Precision (%) | 87.7 (3.6) | 84.7 (3.6) |
| F1 Score (%) | 90.7 (1.9) | 81.9 (2.8) |
| MCC        | 0.81 (0.04) | 0.65 (0.05) |

Numbers correspond to individual’s classification average and standard deviations in parenthesis. Abbreviations: PLS – Partial Least Squares; SVM – Support Vector Machine.
Table S3. Proposed biomarkers to m/z discriminant features elected by Machine Learning algorithm group first by model contribution to COVID-19 diagnosis (M1), followed by molecule class.

| Class | Molecule | Exact mass m/z | Correlation | DeltaJ | Molecular Formula | Adduct | Error (ppm) | MSMS | Metlin ID |
|-------|----------|----------------|-------------|--------|-------------------|--------|-------------|------|-----------|
| **COVID-19 POSITIVE** | | | | | | | | | |
| Purine | Deoxyguanosine and/or Adenosine | 268.1050 | Marker | 22.0 | C_{16}H_{12}N_{4}O_{4} | [M+H]^+ | 3.7 | 236-250-236-240-222-226-150-95 | 3395 and/or 86 |
| Glycerophospholipid | PG(20:5)§ | 553.2561 | Marker | 33.1 | C_{20}H_{40}O_{3}P | [M+Na]^+ | 4.2 | 299-301-419-521-535-495 | 80018 |
| | PE(38:4)§ | 806.5068 | Marker | 30.2 | C_{22}H_{36}N_{2}O_{3}P | [M+K]^+ | -3.6 | 588-241-747-788-623-537 | 60752 |
| | PC(38:8)§ | 784.5252 | Marker | 26.2 | C_{26}H_{50}N_{3}O_{4} | [M+H+H2O]^+ | -3.7 | 725-752-724-740-774 | 59948 |
| Glycerolipid | DG(34:1)§ | 577.5189 | Marker | 17.1 | C_{21}H_{36}O_{3} | [M+H+H2O]^+ | -1.2 | 559-545-531-447-405-265-195 | 4260 |
| | DG(34:2)§ | 575.5031 | 577.5189 | 14.9 | C_{22}H_{36}O_{3} | [M+H+H2O]^+ | -1.4 | 557-543-529-319-263 | 58796 |
| | DG(36:3)§ | 601.5189 | 577.5189 | 19.7 | C_{24}H_{50}O_{3} | [M+H+H2O]^+ | -1.2 | 527-569-583-555-265 | 58855 |
| | TG(50:1)§ | 855.7415 | 879.7408 | 18.4 | C_{30}H_{100}O_{6} | [M+Na]^+ | 0.4 | 599-573-577-551-823 | 4705 |
| | TG(52:2)§ | 891.7568 | 879.7408 | 26.2 | C_{32}H_{102}O_{6} | [M+Na]^+ | -0.1 | 599-625-577-863-603 | 4798 |
| | TG(52:2)§ | 897.7310 | Marker | 25.3 | C_{32}H_{102}O_{6} | [M+K]^+ | 0.2 | 641-615-321-613-599-627-865 | 4793 |
| | TG(52:3)§ | 879.7408 | Marker | 25.2 | C_{32}H_{100}O_{6} | [M+Na]^+ | -0.5 | 623-597-599-577-575-847-861 | 61738 |
| | TG(52:4)§ | 893.6987 | 879.7408 | 20.2 | C_{32}H_{98}O_{6} | [M+K]^+ | -0.9 | 637-613-611-317-623-597-875 | 4839 |
| | TG(52:4)§ | 877.7246 | 879.7408 | 20.7 | C_{32}H_{98}O_{6} | [M+Na]^+ | -1.1 | 597-621-575-599-623-845 | 4839 |
| | TG(54:4)§ | 905.7562 | 879.7408 | 26.6 | C_{34}H_{102}O_{6} | [M+Na]^+ | -0.8 | 623-625-601-603-599-887-873 | 4994 |
| | TG(54:4)§ | 921.7297 | 879.7408 | 24.4 | C_{34}H_{102}O_{6} | [M+K]^+ | -1.2 | 639-641-665-319-367-625-889 | 101990 |
| Unknown | Peptide 1§ | 581.3655 | Marker | 22.9 | C_{12}H_{44}N_{2}O_{3} | [M+H]^+ | 4.3 | 525-549-489-535-563-435 | 221206 |
| | Unknown 1 | 469.3824 | Marker | 9.7 | - | - | - | 437-451-329-413-206-290 | - |
| | Unknown 2 | 822.4814 | Marker | 22.2 | - | - | - | 604-386-790-763-766-804-639 | - |
| | Unknown 3 | 974.6389 | Marker | 17.5 | - | - | - | 941-955-675-717-703-689-661 | - |
| | Unknown 4 | 1121.8588 | Marker | 22.9 | - | - | - | 1102-864-836-989 | - |
| **COVID-19 NEGATIVE** | | | | | | | | | |
| Unsaturated Fatty Acid | Tridecadienoic acid | 228.1955 | Marker | -14.4 | C_{13}H_{22}O_{2} | [M+NH]_{4}^{+} | -1.3 | 211-193-175-109-95 | 73920 |
| | Eicosatetraenoic acid | 345.2183 | Marker | -15.3 | C_{20}H_{32}O_{2} | [M+K]^+ | -2.0 | 257-299-327-273-287-215 | 259 |
| Sphingolipids | Sphingosine(14:2)§ | 242.2110 | Marker | -17.5 | C_{14}H_{22}NO_{2} | [M+H]^+ | -2.1 | 210-224-225-186-184-175-142 | 53907 |
| Glycerophospholipid | LysoPC(16:0)§ | 496.3392 | Marker | -14.1 | C_{22}H_{50}NO_{3}P | [M+H]^+ | -1.2 | 478-184-419-464 | 61692 |
| Lipid                      | m/z     | Marker  | Retention Time | Exact Mass (Da) | Mass Difference (Da) | Purity (Da) |
|----------------------------|---------|---------|----------------|-----------------|----------------------|-------------|
| LysoPC(18:2) §             | 520.3394| 496.3392| -17.7          | C₂₆H₅₀NO₇P      | [M+H]⁺               | 0.8         |
|                            |         |         |                |                 |                      | 502-184-461-474-488 | 61696      |
| LysoPC(18:1) §             | 522.3550| 496.3392| -15.7          | C₂₆H₅₂NO₇P      | [M+H]⁺               | 0.8         |
|                            |         |         |                |                 |                      | 504-184-490-476   | 61695      |
| PAF C-16 and/or LysoPC(18:0) § | 524.3706| 496.3392| -11.5          | C₂₆H₅₄NO₇P      | [M+H]⁺               | 0.8         |
|                            |         |         |                |                 |                      | 506-299-270-184-492-451-478 | 34488 |
|                            |         |         |                |                 |                      | 61694      |
| PC(O-34:3) §               | 742.5735| Marker  | -18.5          | C₂₆H₅₀NO₇P      | [M+H]⁺               | 1.3         |
|                            |         |         |                |                 |                      | 683-724-696-710-502-559-486-460 | 40070 |
| PC(O-34:2) §               | 744.5892| 742.5735| -13.9          | C₂₆H₅₂NO₇P      | [M+H]⁺               | 1.3         |
|                            |         |         |                |                 |                      | 726-712-685-698-659-655-641 | 62003 |
| PC(O-36:3) §               | 770.6043| Marker  | -17.0          | C₂₆H₅₄NO₇P      | [M+H]⁺               | 1.9         |
|                            |         |         |                |                 |                      | 752-738-711-724   | 43415      |
| PS(O-38:4) §               | 798.5640| Marker  | -35.1          | C₂₆H₅₀NO₇P      | [M+H]⁺               | 0.4         |
|                            |         |         |                |                 |                      | 739-780-766-515-461-752 | 78718 |
| Cholesterol                | 369.3511| Marker  | -11.4          | C₂₁H₄₆O          | [M+H-H₂O]⁺           | -2.7        |
|                            |         |         |                |                 |                      | 343-359-215-273-287-203-189-351-161-175-147-135-229 | 163 |
| Dihydroxy-cholenoic acid § | 408.3087| Marker  | -12.4          | C₂₃H₄₈O₄        | [M+NH₄]⁺             | -5.1        |
|                            |         |         |                |                 |                      | 319-345-375    | 84534      |
| CE(16:0) §                 | 647.5738| Marker  | -18.3          | C₂₃H₅₆O₂        | [M+Na]⁺              | 0.2         |
|                            |         |         |                |                 |                      | 591-385-601-615-535-279 | 41701 |
| Glycine-Phenylalanine §    | 261.0644| Marker  | -21.2          | C₁₁H₁₄N₂O₃      | [M+K]⁺               | 3.1         |
|                            |         |         |                |                 |                      | 173-215-189-243-233 | 85897 |
| Unknown 5                  | 205.0382| Marker  | -11.6          | -                | -                    | -           |
|                            |         |         |                |                 |                      | 173-187-188-159-191 | -          |
| Unknown 6                  | 723.6610| Marker  | -20.6          | -                | -                    | -           |
|                            |         |         |                |                 |                      | 664-691-677-705-587-540-467-355 | - |
| Unknown 7                  | 986.9247| Marker  | -10.8          | -                | -                    | -           |
|                            |         |         |                |                 |                      | 954-968-688-704-730-674 | - |
Table S4. Proposed biomarkers to m/z discriminant features elected by Machine Learning algorithm group first by model contribution to risk assessment (M2), followed by molecule class.

| Class         | Molecule                                      | Exact Mass m/z | Correlation | ΔJ  | Molecular Formula | Adduct         | Error (ppm) | MSMS                  | Metlin ID |
|---------------|-----------------------------------------------|----------------|-------------|-----|-------------------|----------------|-------------|-----------------------|-----------|
| SEVERE        |                                               |                |             |     |                   |                |             |                       |           |
| Purine        | Deoxyguanosine and/or Adenosine\(^6\)          | 268.1054       | Marker      | 27.5| C\(_{10}\)H\(_{15}\)N\(_3\)O\(_4\) | [M+H]\(^+\)  | -5.2        | 236-250-236-240-222-226-150-95 | 3395 and/or 86 |
| N-acyl amino acids | N-stearoyl valine                          | 406.3293       | Marker      | 11.7| C\(_{23}\)H\(_{49}\)NO\(_3\) | [M+Na]\(^+\) | -0.2       | 360-388-362-316-336-266 | 75504    |
| Sterol Lipid  | Dihydroxy-cholanoyl-glycine\(^4\)            | 450.3192       | 406.3293    | 11.9| C\(_{23}\)H\(_{49}\)NO\(_3\) | [M+H]\(^+\)  | 4.9         | 359-404-417-432-377-363 | 43193    |
| Unknown       | Unknown 8                                     | 296.1927       | Marker      | 3.5 | -                 | -              | -           | 264-234-282-278-252-250-224 | -         |
| Unknown       | Unknown 9                                     | 398.3743       | Marker      | 36.1| -                 | -              | -           | 319-352-366-380-310-338-326-368-298 | -         |
| Unknown       | Unknown 10                                    | 541.4215       | Marker      | 6.5 | -                 | -              | -           | 480-439-495-478-509-523-399 | -         |
| Unknown       | Unknown 11                                    | 617.5984       | Marker      | 15.8| -                 | -              | -           | 571-599-585-557-525-337 | -         |
| Unknown       | Unknown 12                                    | 1057.4970      | Marker      | 6.1 | -                 | -              | -           | 1039-1025-989-799 | -         |
| MILD          |                                               |                |             |     |                   |                |             |                       |           |
| N-acyl ethanolamines | Eicosaenoylcolanoylamine                  | 336.3263       | Marker      | -11.3| C\(_{22}\)H\(_{47}\)NO\(_2\) | [M+H-H\(_2\)O]\(^+\) | 0.9         | 263-281-290-308-304-318 | 3722     |
| N-acyl amino acids | Docosanoylcolanoylamine                  | 366.3731       | 336.3263    | -9.4| C\(_{22}\)H\(_{47}\)NO\(_2\) | [M+H-H\(_2\)O]\(^+\) | 1.4         | 334-295-320-336-348 | 3727     |
| Glycerophospholipid | N-Linoleoyl glycerine                   | 360.2498       | Marker      | -12.3| C\(_{20}\)H\(_{43}\)NO\(_3\) | [M+Na]\(^+\)  | 3.1         | 314-316-288-330-328-342-184 | 43426    |
| LysoPC(18:2)\(^{#\#}\) | LysoPC(18:2)\(^{#\#}\)                     | 520.3401       | Marker      | -39.1| C\(_{20}\)H\(_{43}\)NO\(_3\) | [M+H]\(^+\)  | -0.6        | 502-184-461-474-488 | 61696    |
| LysoPC(18:2)\(^{#\#}\) | LysoPC(18:2)\(^{#\#}\)                     | 542.3222       | Marker      | -38.5| C\(_{20}\)H\(_{43}\)NO\(_3\) | [M+Na]\(^+\)  | -0.9        | 496-483-524-215 | 61696    |
| PC(34:2)\(^{#\#}\) | PC(34:2)\(^{#\#}\)                         | 758.5692       | Marker      | -11.0| C\(_{22}\)H\(_{48}\)NO\(_4\) | [M+H]\(^+\)  | 0.3         | 698-740-419-502-375 | 39430    |
| PC(34:2)\(^{#\#}\) | PC(34:2)\(^{#\#}\)                         | 780.5515       | Marker      | -17.8| C\(_{24}\)H\(_{50}\)NO\(_4\) | [M+Na]\(^+\)  | -1.0        | 721-762-748-734-575 | 39430    |
| Arachidonoyl PAF C-16 and/or PC(36:2)\(^{#\#}\) | Arachidonoyl PAF C-16 and/or PC(36:2)\(^{#\#}\) | 768.5901       | Marker      | -32.3| C\(_{24}\)H\(_{49}\)NO\(_3\)P | [M+H]\(^+\)  | 0.0         | 736-754-750-740-722-709-724 | 43414    |
| PC(36:2)\(^{#\#}\) | PC(36:2)\(^{#\#}\)                         | 786.6008       | 758.5692    | -16.5| C\(_{24}\)H\(_{50}\)NO\(_3\)P | [M+H]\(^+\)  | 1.0         | 740-768-740-603-478-401 | 102731   |
| PC(36:3)\(^{#\#}\) | PC(36:3)\(^{#\#}\)                         | 784.5867       | Marker      | -22.3| C\(_{24}\)H\(_{50}\)NO\(_3\)P | [M+H]\(^+\)  | -2.0        | 725-766-693-572-601 | 39634    |
| PC(36:4)\(^{#\#}\) | PC(36:4)\(^{#\#}\)                         | 782.5692       | 758.5692    | -19.5| C\(_{24}\)H\(_{50}\)NO\(_3\)P | [M+H]\(^+\)  | 0.3         | 723-599-750-764-736 | 39671    |
| Eicosapentaenoyl PAF C-16 and/or PC(36:3) | 766.5743 | Marker | -35.9 | C_{46}H_{89}NO_5P | C_{46}H_{89}NO_5P | [M+H]^+ | [M+H-H_2O]^+ | 0.3 | 1.0 | 748-720-734-649-721-706 | 43430 | 39637 |
| PC(38:3) | 812.6163 | 784.5867 | -29.4 | C_{46}H_{89}NO_5P | [M+H]^+ | 0.1 | 753-794-766-530-502 | 59780 |
| Docosahexaenoil PAF C-16 and/or PC(38:4) | 792.5914 | 766.5743 | -34.5 | C_{46}H_{89}NO_5P | C_{46}H_{89}NO_5P | [M+H]^+ | [M+H-H_2O]^+ | 1.5 | -0.9 | 746-774-760-733-699-783 | 62936 | 39544 |
| PC(38:4) | 810.6010 | 758.5692 | -24.3 | C_{46}H_{89}NO_5P | [M+H]^+ | -0.4 | 751-792-778-627 | 59436 |
| PC(38:4) | 832.5831 | 808.5847 | -28.2 | C_{46}H_{89}NO_5P | [M+Na]^+ | -0.5 | 773-814-800-786-534 | 39545 |
| PC(38:5) | 808.5847 | Marker | -28.1 | C_{46}H_{89}NO_5P | [M+H]^+ | 0.5 | 749-625-790-776-762 | 59879 |
| PC(38:6) | 806.5693 | 758.5692 | -21.9 | C_{46}H_{89}NO_5P | [M+H]^+ | 0.1 | 747-788-623-774-760 | 39685 |
| PC(40:4) and/or PC(O-40:6) | 820.6215 | 766.5743 | -28.5 | C_{46}H_{89}NO_5P | C_{46}H_{89}NO_5P | [M+H-H_2O]^+ | 0.6 | 788-761-802-773-637-550-335 | 75947 | 40136 |
| PC(O-34:1) | 746.6058 | 703.5750 | -15.8 | C_{42}H_{85}NO_5P | [M+H]^+ | 0.0 | 728-494-508-714 | 40013 |
| PC(O-38:4) | 796.6213 | 703.5750 | -29.3 | C_{46}H_{89}NO_5P | [M+H]^+ | 0.3 | 737-778-764-613 | 102733 |
| PC(O-38:5) | 794.6060 | 703.5750 | -32.4 | C_{46}H_{89}NO_5P | [M+H]^+ | -0.3 | 776-762-748-735-611 | 40018 |
| Glycerolipid | | | | | | | | | | | | |
| TG(46:5) | 786.6563 | 703.5750 | -28.8 | C_{46}H_{89}O_5 | [M+NH_3]^+ | 5.5 | 529-543-559-571-503 | 99210 |
| TG(48:4) | 816.7036 | Marker | -32.4 | C_{51}H_{99}O_5 | [M+NH_3]^+ | 4.9 | 559-757-797-479-573-587 | 99318 |
| Sterol Lipid | | | | | | | | | | | | |
| Cholesterol | 369.3517 | Marker | -31.1 | C_{27}H_{46}O | [M+H-H_2O]^+ | 1.1 | 343-359-215-273-287-203-189-351-161-175-147-135-229 | 163 |
| SM(34:1) | 703.5750 | Marker | -15.2 | C_{46}H_{89}N_2O_5P | [M+H]^+ | -0.1 | 685-642-656-671-657-447 | 83743 |
| SM(35:1) | 717.5904 | 703.5750 | -19.4 | C_{46}H_{89}N_2O_5P | [M+H]^+ | 0.1 | 699-670-685-419-461 | 83745 |
| SM(36:2) | 729.5905 | 703.5750 | -21.0 | C_{46}H_{89}N_2O_5P | [M+H]^+ | 0.0 | 711-699-697-683-592-392 | 53977 |
| SM(38:2) | 757.6215 | 703.5750 | -28.4 | C_{46}H_{89}N_2O_5P | [M+H]^+ | 0.4 | 501-529-725-739-698-711-420 | 83760 |
| Sphingolipid | | | | | | | | | | | | |
| SM(40:1) | 809.6519 | 703.5750 | -20.8 | C_{46}H_{89}N_2O_5P | [M+Na]^+ | -1.5 | 750-791-777-553-737-425 | 41589 |
| SM(40:2) | 785.6526 | 758.5692 | -30.4 | C_{46}H_{89}N_2O_5P | [M+H]^+ | 0.6 | 529-767-739-515 | 83771 |
| SM(41:1) | 801.6841 | 703.5750 | -31.9 | C_{46}H_{89}N_2O_5P | [M+H]^+ | 0.4 | 545-531-503-375-754 | 62432 |
| SM(42:2) | 813.6842 | 703.5750 | -18.7 | C_{46}H_{89}N_2O_5P | [M+H]^+ | 0.2 | 557-754-767-795 | 83781 |
| SM(42:3) | 811.6682 | 758.5692 | -19.1 | C_{47}H_{91}N_2O_5P | [M+H]^+ | 0.7 | 555-529-720-541-751-779-793 | 83780 |
| Peptide 2 | 409.1752 | Marker | -11.9 | C_{15}H_{30}N_2O_5S | [M+H]^+ | -0.2 | 319-391-377-363-339-333-321 | 235779 |
| Unknown | | | | | | | | | | | | |
| Unknown 13 | 266.1720 | Marker | -9.0 | - | - | - | 194-245-234-220-204-212-248 | - |
| Unknown 14 | 537.3950 | Marker | -4.2 | - | - | - | 298-284-299-282-520-491-477 | - |
| Unknown 15 | 547.5190 | Marker | -12.5 | - | - | 501-503-529-485-487-479-398-515 |
| Unknown 16 | 788.6089 | Marker | -16.8 | - | - | 536-616-770 [536]*284-282 |
| Unknown 17 | 986.9265 | Marker | -33.6 | - | - | 953-697-703-729-967-673-657 |
| Unknown 18 | 1008.9080 | Marker | -30.6 | - | - | 990-976-961-940-891-751-698 |

≠ (Carbon number : double bond); § Isomers with the same exact m/z and similar fragmentation profile to be distinguished by FI-MS; * MS3.

Abbreviations: LysoPC – Lysophosphatidylcholine; PAF – Platelet Activating Factor; PC – Phosphatidylcholine; SM – Sphingomyelin; TG – Triacylglycerol.

Table S5. Proposed biomarkers to m/z discriminant features elected by Machine Learning algorithm group first by model contribution to low-risk discrimination (M3), followed by molecule class.
| Class         | Molecule       | Exact Mass m/z | Correlation | ΔJ  | Molecular formula | Adduct     | Error (ppm) | MSMS            | Metlin ID |
|--------------|----------------|----------------|-------------|-----|-------------------|------------|-------------|------------------|-----------|
| MILD         | PG(20:5) †‡    | 531.2744       | Marker      | 15.3| C_{26}H_{45}O_5P | [M+H]^+    | -5.1        | 470-489-499-459-349-307-234 | 80018    |
| Glycerophospholipid | PG(20:5) †‡    | 553.2561       | Marker      | 25.3| C_{26}H_{45}O_5P | [M+Na]^+   | -4.3        | 299-301-419-521-535-495   | 80018    |
|               | PC(38:8) †‡    | 784.5252       | Marker      | 26.5| C_{46}H_{76}NO_3P| [M+H-H_2O]^+ | 3.7         | 725-752-724-740-774     | 59948    |
|               | PE(38:4) ‡‡§   | 806.5068       | Marker      | 23.3| C_{43}H_{78}O_3P | [M+K]^+    | -3.6        | 588-241-747-788-623-537   | 60752    |
| Glycerolipid | DG(34:0) †‡    | 579.534        | Marker      | 22.6| C_{33}H_{52}O_5 | [M+H-H_2O]^+ | 2.1         | 533-547-451-563-519-493-475-195 | 4274    |
|               | DG(34:1) †‡    | 577.5189       | Marker      | 20.4| C_{33}H_{52}O_5 | [M+H-H_2O]^+ | -1.2        | 559-545-531-447-405-265-195 | 4259    |
|               | DG(34:2) †‡    | 575.5031       | Marker      | 21.6| C_{33}H_{52}O_5 | [M+H-H_2O]^+ | -1.4        | 557-543-529-393-319-263   | 58796    |
|               | DG(36:2) †‡§   | 603.534        | Marker      | 16.0| C_{36}H_{72}O_5 | [M+H-H_2O]^+ | 2.0         | 543-557-529-571-585-235-265-247 | 64623    |
|               | DG(36:3) †‡    | 601.5189       | Marker      | 24.9| C_{39}H_{90}O_5 | [M+H-H_2O]^+ | -1.2        | 527-569-583-555-265      | 4362    |
|               | TG(52:4) ‡‡     | 877.7246       | Marker      | 19.7| C_{35}H_{90}O_5 | [M+Na]^+   | 1.1         | 597-621-575-595-599-623-845 | 4839    |
| Unknown      | Peptide 3 ‡†    | 505.1811       | Marker      | 25.2| C_{28}H_{30}N_6O_7| [M+K]^+    | -0.6        | 371-473-459-283-487-469-441-437 | 152631  |
|               | Unknown 19      | 214.1414       | Marker      | 18.3| -                   | -          | -           | 319-353-366-310-309-380-326-338-368-298 | -       |
|               | Unknown 20      | 429.3727       | Marker      | 24.5| -                   | -          | -           | 341-319-165-359-295-401-411-209 | -       |
|               | Unknown 21      | 443.3344       | Marker      | 9.9 | -                   | -          | -           | 411-425-397-371-339-309-367-355-433-387 | -       |
|               | Unknown 22      | 515.4134       | Marker      | 14.9| -                   | -          | -           | 469-483-497-381 [469]*207-335-291-261-333 | -       |
|               | Unknown 23      | 696.379        | Marker      | 18.6| -                   | -          | -           | 327-328-649-678-312-384-593-634-664 | -       |
|               | Unknown 24      | 851.3977       | Marker      | 25.2| -                   | -          | -           | 577-551-571-595-623-832-819 | -       |
|               | Unknown 25      | 1073.4943      | Marker      | 19.2| -                   | -          | -           | 1055-1041-883-677-1027 | -       |
|               | Unknown 26      | 1169.0771      | Marker      | 25.4| -                   | -          | -           | 1151-1123-1108-913-575 | -       |
|               | Unknown 27      | 1279.5905      | Marker      | 20.2| -                   | -          | -           | 783-759-496-478-520-1261-886 | -       |
| COVID-19 NEGATIVE | Unsaturated Fatty Acid | Eicosatrienoic acid 345.2183 | Marker | -14.3| C_{20}H_{34}O_2 | [M+K]^+ | -2.0        | 256-257-255-299-327-313-273-287-215 | 259    |
| Sterol Lipid          | Keto-DHEA § | Marker | -19.0 | C₁₉H₂₆O₃ | [M+Na]⁺ | -1.5 | 279-283-293-281-307-253-271-255-297 | 263488 |
|----------------------|-------------|--------|-------|-----------|---------|------|---------------------------------|--------|
| Dehydrocholesterol  §| 367.3361    | Marker | -26.7 | C₂₇H₄₄O  | [M+H-H₂O]⁺ | 1.1  | 352-241-227-213-295-285-255-227-219-173 | 3902   |
| Dihydroxycholenoic   acid § | 429.2404    | Marker | -23.5 | C₂₄H₃₈O₄ | [M+K]⁺  | -0.5 | 341-319-165-209-411-401-359-205-271-163 | 84538  |
| Peptide 4 §          | 425.1490    | Marker | -11.6 | C₁₈H₂₆N₄O₇S | [M+H-H₂O]⁺  | 1.2  | 365-393-407-379-337-369          | 239104 |
| Unknown              |             |        |       |           |         |      |                                  |        |
| Unknown 28           | 398.3743    | Marker | -10.0 |          |         |      | 319-352-366-380-310-309-338-326-298 |        |
| Unknown 29           | 221.1751    | Marker | -12.3 |          |         |      | 175-177-189-203-193              |        |
| Unknown 30           | 984.9091    | Marker | -19.2 |          |         |      | 672-686-702-727-966-952-924-664  |        |

≠ (Carbon number : double bond); § Isomers with the same exact m/z and similar fragmentation profile to be distinguished by FI-MS; * MS3.

Abbreviations: DG - Diacylglycerol; DHEA – Dehydroepiandrosterone; PC – Phosphatidylcholine; PE – Phosphatidylethanolamine; PG – Phosphatidylglycerol; TG – Triacylglycerol.
Models markers combination

We combined the markers found by the M1 model (COVID-19 diagnosis) with the M2 (risk assessment), and M3 (low-risk discrimination) in a single list of biomarkers and performed classification with the separated blind test. The results show some potential for this analysis but with numbers slightly below the full performance obtained with custom-tailored binary detectors. However, we deem the results promising for future work in multiclass research front.

Table S6. Confusion matrix combining markers found by M1 with the M2 and M3 models.

| Real Class | Control | Mild | Severe | Inconclusive | Sensitivity |
|------------|---------|------|--------|--------------|-------------|
| Control    | 93.9%   | 2.0% | 0%     | 4.1%         | 93.9%       |
| Mild       | 4.9%    | 63.4%| 22.0%  | 9.8%         | 63.4%       |
| Severe     | 0.0%    | 12.5%| 87.5%  | 0.0%         | 87.5%       |
| Precision  | 95.1%   | 81.3%| 79.9%  |              | 81.1%       |