Journal of Breath Research

Instrumental drift removal in GC-MS data for breath analysis: the short-term and long-term temporal validation of putative biomarkers for COPD

To cite this article before publication: Raquel Rodríguez-Pérez *et al* 2018 *J. Breath Res.* in press https://doi.org/10.1088/1752-7163/aaa492

Instrumental Drift Removal in GC-MS Data for Breath Analysis: The Short-term and Long-term Temporal Validation of Putative Biomarkers for COPD

Raquel Rodríguez-Pérez[1], Roldán Cortés[2], Ana Guamán[1,3], Antonio Pardo[3], Yolanda Torralba[4], Federico Gómez[4], Josep Roca[4], Joan Albert Barberà[4], Marta Cascante[2], Santiago Marco[1,3]

[1] Signal and Information Processing for Sensing Systems, Institute for Bioengineering of Catalonia (IBEC), The Barcelona Institute of Science and Technology, Barcelona, Spain

[2] Department of Biochemistry and Molecular Biology, IBUB, University of Barcelona, Barcelona, Spain

[3] Department of Electronics and Biomedical Engineering, University of Barcelona, Barcelona, Spain

[4] Department of Pulmonary Medicine, Hospital Clínic, CIBERES, IDIBAPS, Barcelona, Spain

Abstract

Breath analysis holds the promise of a non-invasive technique for the diagnosis of diverse respiratory conditions including COPD and lung cancer. Breath contains small metabolites that may be putative biomarkers of these conditions. However, the discovery of reliable biomarkers is a considerable challenge in the presence of both clinical and instrumental confounding factors. Among the latter, instrumental time drifts are highly relevant, as since question the short and long-term validity of predictive models. In this work we present a methodology to counter instrumental drifts using information from interleaved blanks for a case study of GC-MS data from breath samples. The proposed method includes feature filtering, and additive, multiplicative and multivariate drift corrections, the latter being based on Component Correction. Biomarker discovery was based on Genetic Algorithms in a filter configuration using Fisher´s ratio computed in the Partial Least Squares – Discriminant Analysis subspace as a figure of merit. Using our protocol, we have been able to find nine peaks that provide a statistically significant Area under the ROC Curve (AUC) of 0.75 for COPD discrimination. The method developed has been successfully validated using blind samples in short-term temporal validation. However, in the attempt to use this model for patient screening six months later was not successful. This negative result highlights the importance of increasing validation rigour when reporting biomarker discovery results.

Introduction

Volatile biomarker discovery is increasingly gaining attention in the improvement of screening, diagnosis, and prognosis in medicine.[1] *Volatilome* is the volatile fraction of metabolome,[2] which contains all those volatile organic compounds (VOCs) generated within an organism. VOCs are organic analytes with a substantial vapour pressure (typically less than 300 Da) and their analysis allows monitoring human chemistry and health related conditions.[3] VOCs reflect metabolic processes that occur in the body and which may change with disease. Thus, finding specific VOC fingerprints that are characteristic of a pathology is an important field of research.[4,5,6] The analysis of exhaled breath is a specifically attractive and promising means of access to the volatilome, as breath can be obtained non-invasively and contains potentially informative VOCs. VOCs can reach the alveoli, cross the alveolar interface and then be exhaled by the subject.[7,8]

The discovery of new disease markers is based on measuring the global metabolic profile of a sample without bias, which is known as *untargeted metabolomics* or *metabolic fingerprinting*.[9] Advanced analytical instrumentation, mainly mass spectrometry (MS) and nuclear magnetic resonance (NMR), provides the researchers with possibility of examining hundreds or thousands of metabolites in parallel. Gas chromatography – mass spectrometry (GC-MS) in particular, is one of the most popular and powerful analytical techniques for breath analysis.[8]

The comparison of metabolic fingerprints among different experimental groups; namely condition and control, can lead to the identification of metabolic patterns that have changed due to a disease, and which may be used as a diagnostic tool. Chemometric or pattern recognition methods are required to extract information from exhaled breath data and discover relevant VOCs.[10] However, many data processing techniques are not designed to deal with large amounts of irrelevant or correlated features, which is the case with current analytical technologies that are applied to metabolomics. For instance, in a review on the statistical analysis of metabolomics data, Vinaixa[11] observes a number of metabolomics studies (based on Liquid Chromatography-Mass Spectrometry, LC-MS) where the number of features ranges between 4000-10000, while the number of subjects is always below 30. In other words, from a data-analysis perspective, metabolomics data is characterized by high dimensionality and small sample counts. Consequently, the 'Curse of Dimensionality'[12] has to be taken into account, and data processing methods must be scrutinized in their ability to deal with small sample-to-dimensionality ratios.[13,14] The inherent difficulty in the analysis of this type of data has long-been recognized and methods to deal with it have been proposed since the 60s.[15] The application of conventional statistical testing is plagued with theoretical difficulties,[16] and in many cases machine learning approaches are preferred. However, the scarcity of examples poses problems with respect to the complexity of those predictive models that may be built. Complexity control can be attained through regularization or through dimensionality- reduction techniques.[13] Some authors advocate the use of penalized likelihood estimation.[17] One example is the use of the Least Absolute Shrinkage and Selection Operator (LASSO) based on L1 penalty in the quest for sparse solutions,[18] while others propose projection methods, such as Principal Component Analysis (PCA)[19] or Partial Least Squares-Discriminant Analysis (PLS-DA).[20] However, the use of feature selection methods based on iterative searches and optimization procedures, using either wrappers or filters as objective functions is perhaps the most popular approach.[21] The performance of these methods in high-dimensionality, small-sample conditions has been analysed previously.[22] Among them, Genetic Algorithms (GA) have been used extensively.[23,24,25] The combination of GA and PLS-DA in metabolomics has been previously reported.[26]

The application of dimensionality-reduction techniques based on feature selection methods is therefore the natural choice.[17]

Breath analysis is particularly challenging due to the lack of standards for sampling and storage that are employed.[27,28,29] As such, researchers prefer to analyse breath samples soon after collection. If patient recruitment takes a long time, then analysis may extend over several months. In these conditions, instrumental changes, such as chromatographic column aging, temperature variations and the effects of contamination can shift intensity value measurements over time.[30,31,32,33] As computational methods in metabolomics rely on a quantitative comparison between metabolite abundances in diverse groups, instrumental drift may become an important source of errors. It is therefore extremely important to block this potentially confounding factor at the design stage, if possible.[34] Another inherent difficulty involved in breath analysis is the large inter and intra-individual variability of exhaled breath, even among healthy subjects,[35] and the change of VOCs patterns according to food consumption, smoking, gender, age, and so forth.[36,37] There are definitely several sources of unwanted variance that may act as confounding factors, and which can be divided into two types: instrumental (e.g. different location, operator, instrument or sampling conditions), and clinical (e.g. gender, age, smoking status, comorbidities or treatments).[38,29] If some factors cannot be controlled with experimental design,[39] they must be taken into consideration during data analysis.[40]

Normalization methods adjust data for biases caused by non-biological conditions or unwanted variations.[36,41,42] Most adopted normalization methods rely on scaling factors using all the data set (e.g. total sum or norm). However, these methods have been shown to adjust data incorrectly, as an increased abundance in some metabolites leads to a decrease in other metabolites.[43]

Alternatively, internal standards (IS) and quality control (QC) samples are often used for the posterior removal of certain systematic errors that maybe platform specific.[40] The internal standards method is based on the addition of known metabolites (in a determined amount) to the samples in order to improve quantitative analysis and normalize each sample according to their variation. The simplest approach is based on a single internal standard and assumes that the variation in sensitivity observed for this metabolite is constant across all the analytes. Thus, a multiplicative correction factor may be applied. However, it is not clear that the underlying hypothesis can be sustained, and therefore the use of multiple internal standards has been proposed, but then the selection of the proper normalization method remains an open problem. Application examples of normalization methods based on IS have been reported on literature.[43,40] For instance M. Sysi-Aho et al. proposes that the correction factors are a linear function of the variation observed for the IS, and he optimizes the coefficients of the model to maximize the likelihood of the observed data.[32] However, in untargeted studies, the metabolites that are to be detected are not known *a priori,* and therefore the addition of numerous IS for detection and to normalize for analytical variation is not practical, indeed it would prejudice the integrity of the samples.[31] As such, QC are often preferred in untargeted metabolomics in order to avoid changing the physical sample and using IS that may coelute with metabolites of interest.[44] QC samples may be either commercial or pooled (i.e., a mixture of equal aliquots from a representative set of the study samples). The latter are of special interest, as their measurement contains all those metabolites under investigation. Even though pooled QC samples may be relatively easy to prepare for some biofluids, such as urine or plasma,[45,46] exhaled breath presents inherent complications in terms of sampling and storage, which prevents its use, especially in long-term studies. Different signal correction methods rely on QC samples,[44,47] which can be employed to estimate sample- or feature-based correction factors.[31] According to previous works,[42,48] one of the best normalization methods that employs QC samples is Probabilistic Quotient Normalization (PQN).[49]

Since machine drifts are frequently observed in chromatography and MS, the assessment of reproducibility and repeatability is of utmost importance.[50,51] Over the last decade, a number of computational algorithms have been applied in order to correct instrumental drifts, or batches in data. For instance, ComBat technique was proposed in the genomics field and was later also applied in metabolomics, for adjusting data for batch effects.[52,30] Many methods are based on projection filters. Examples of these methods are Component Correction (CC) and Common Principal Component Analysis (CPCA). They consist of the estimation of the drift subspace with a reference class or the entire data set, respectively, and the subtraction of this subspace from the original data.[30,53] The dimensionality of the drift subspace should be carefully determined using calibration data, in order to avoid valuable information removal. For example, Fernández-Albert et al. used Dunn and Silhouette indexes to compare different drift correction methods and the dimensionalities of the drift subspace.[30] Another example in this family is Orthogonal Signal Correction (OSC), which removes the data variance that is uncorrelated to the class information.[54] Similarly, orthogonal projections to latent structures, such as Orthogonal Partial Least Squares algorithms (O-PLS or O2-PLS), have also been applied for drift compensation.[55,56,57] The existing literature mostly uses these methods for predictive models, but not specifically in biomarker discovery, where the impact of instrumental drift on computational biomarker discovery is not yet totally understood, and remains an under-studied topic.

Unfortunately, those studies of biomarker discovery in metabolomics that have been undertaken are usually plagued by statistical bias, and involve small sample sizes, which lead to very poor reproducibility. This is aggravated by the use of weak validation methodologies that mostly consist of internal validation. We believe that only external validation (blind samples) can provide further reliability to biomarker discoveries. Moreover, an additional temporal validation, undertaken months after the model development process has been closed, should be a recommended practise. However, in most published literature this long-term validation is missing. If we expect that discovered biomarkers can be adapted and be put into clinical use, then their use has to be reliable enough to maintain predictive power in a variety of instrumental conditions, or even instrument vendors. Reporting long-term validation results would allow the discovery of the limitations of current research and speed up further investigation into the most promising biomarker discovery studies.

In this work, methods for the instrumental drift in GC-MS data from exhaled breath are proposed and investigated through their application in a practical case. The analysed dataset consists of breath samples obtained from patients with lung cancer (LC), chronic obstructive pulmonary disease (COPD), and control subjects, in order to discover the markers of these diseases. Since this dataset was acquired over a two-year period, the correction of instrumental drift in the entire dataset was one of the data analysis objectives. The influence of time varying effects requires special attention to the application of strict validation procedures, which was an essential part of our workflow. We not only used external validation, but also an additional temporal validation in order to assess the predictive power of the biomarkers, six months after the end of the model development phase, which is a validation level that is not usually reported in similar studies. From a machine-learning standpoint, biomarker discovery may be addressed using feature-selection techniques.

In this quest for putative biomarkers, additional issues, such as potential clinical confounders have also been investigated. Finally, biomarker selection and classification was applied to the binary problem case of COPD vs. control.

Experimental Methods

Sampling System

Exhaled breath was collected using the tidal breath sampler (TBS) shown in **Figure 1**, which was and developed by the Department of Biochemistry and Molecular Biology of the University of Barcelona.[58] It consists of a glass tube with a specific shape that allows the collection of tidal exhaled air. The subject breathes into a unidirectional Rudolph valve, which allows medical air (22% O_2, 78% N_2) contained in a Douglas bag to pass through his mouth and which then transports the exhaled breath to the glass tube. An air pump (FLEX Air Pump 1001) extracts the air from the first section of the glass tube and drives it to a sorbent trap that is filled by a fibre comprising 200 mg Tenax and 200 mg Unicarb. Between the glass tube and the fibre is a filter made of silica gel, to avoid humidity, which distorts mass spectra. In order to avoid breath condensation on the walls of the glass, the tube is heated using a long, warmed filament, which is wrapped along its length. See Supplementary Material for sampling protocol information.

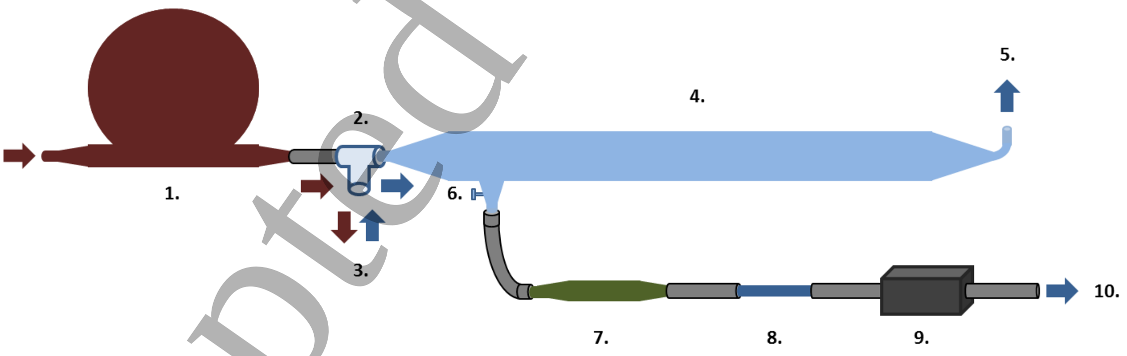


Figure 1. Tidal Breath Sampler. A diagram of the sampling system used to collect VOCs from exhaled breath is shown above. 1: Douglas bag with medical air. 2: Unidirectional Rudolph valve. 3: Mouthpiece for subject´s breath. 4: Glass tube at high temperature. 5: Outlet for the vast majority of air. 6: Outlet for the final part of the

subject´s exhalation (containing the VOCs from gas exchange). 7: Silica gel filter for humidity control. 8: Sorbent trap. 9: Air pump. 10: Air exiting the pump.

GC-MS Analysis

Samples were injected into the gas chromatograph system using a Unity thermal desorption unit for sorbent tubes (Markes). This unit applied a flow of hot carrier gas through the fibres contained in the sorbent traps, desorbing the VOCs in its surface and carrying them to the GC system (FocusGC, fThermo Scientific). The chromatographic column used was a 60 m DB-624 capillary column with an internal diameter of 0.32 mm and a stationary phase thickness of 1.8 µm. The temperature ramp used to optimize the separation and sensitivity of the chromatographic process was: 40 ºC (5 minutes) – 10 ºC (1 minute) – 180 ºC (1 minute) – 15 ºC (1 minute) – 230 ºC (10 minutes). Once eluded from the column at different retention times, the separated compounds were injected into the MS (DSQII MS, from ThermoScientific), where they were ionized via electron impact and subsequently fragmented. In order to reuse the sorbent traps, once desorbed, they were then cleaned to eliminate any possible remaining VOCs, with a flow of N_2 at 320 ºC for 2 hours, before being closed and stored in vacuum.

Dataset and Data Processing Algorithms

Dataset Description

The dataset contained four classes or medical groups: control, COPD, LC, and both diseases (COPDLC). Standard diagnostic criteria were used for the inclusion of each study group. The main exclusion criteria were: (i) Lack of clinical stability; (ii) Abnormal non-obstructive forced spirometry results; (iii) Previous history of cancer; (iv) Previous history of active inflammatory disease; (v) Treatment with steroids or immunomodulators; and, (vi) Data suggesting infectious disease. The protocol was submitted and approved by the Human Studies Ethical Committee at Hospital Clinic and all patients signed informed consent forms before any procedure was initiated.

The sample collection procedure lasted almost two years (from 20/05/11 to 6/5/13) and was divided in two campaigns. The first campaign lasted from 20/05/11 to 7/3/12 and was divided into calibration (initial 80%) and short-term external validation set (last 20% of each class). The second campaign lasted from 10/7/12 to 6/5/13. It is important to note that the collection and analysis protocol was identical in both campaigns. The purpose of the second campaign was to further validate the results obtained from the study carried out in the first campaign. This second campaign served not only to provide blind samples, but also to test the stability of the predictive model. It is important to note that this stability test is neglected in most studies and validation is internal to the same measurement campaign. All model development and optimization, including drift counteraction strategies, was undertaken within the calibration set. The number of samples included in the study is summarized in **Table 1**.

Data subsets	Control	COPD	COPDLC	LC
Calibration	15	22	15	10
Short-term validation	6	7	4	2
Long-term validation	6	5	5	-
Total	27	34	24	12

Table 1. Data Set Overview. The number of samples of each class is reported by data subsets (calibration, short-term, and long-term validation) and also the total number. COPD: Chronic Obstructive Pulmonary Disease, LC: Lung Cancer, COPDLC: COPD and LC.

The metadata of the patients was also available, including sex, age, weight, height, smoking status, functional respiratory results (forced expiratory volume and forced vital capacity) and haematology tests, current treatments or drugs that the patients were taking, comorbidities (e.g. hypertension, asthma, diabetes), and diagnostics of the studied disease.

Apart from the patient samples, two types of spectra blanks were available: (i) 7 fibre blanks (GC-MS analysis of new sorbent traps that were not previously used for any sample), and (ii) 49 sample blanks (GC-MS analysis of sorbent traps in each day of measurement). Fibre blanks are used to assess the existence of specific peaks due to the presence of fibre, whereas sample blanks serve to assess the potential contamination of the system. Even if no sample is injected, a signal can be obtained, due to the VOCs trapped in the column, while sample blanks provide quantification.

Data Pre-processing

Data pre-processing was performed with *MZmine2*[59] software. Peak detection was based on the *Noise Amplitude* algorithm. In this algorithm, intensity range is specified by the user and the algorithm automatically locates the where the noisy baseline is concentrated, and establishes the baseline level at this level. A peak is recognised when the chromatographic signal is registered above the noise level during a certain time span. A smoothed second derivative, Savitzky-Golay digital filter is then used to detect the borders of the peaks. After detection, peaks were aligned using the RANSAC aligner. Here the retention time tolerance was 1 minute, with an m/z tolerance of 0.4. The number of iterations was automatically determined by the software. MZmine also provides tools for isotopic peak grouping and gap filling in the event of missing values.[60] After MZmine pre-processing, 3049 peaks with an associated intensity were obtained. Each peak was identified by its mass to charge ratio (m/z) and retention time (RT). The area under the peaks was computed in order to obtain the data matrix. Since there were peaks with the same RT (differences less than the GC resolution, 0.1 min) and large correlation (Pearson correlation r > 0.9) between them, they were considered to be part of the same original compounds. By clustering these peaks and removing near zero variance features, the dimensionality of the data was reduced from 3049 to 1749 features. In total, 1297 features were removed by peak clustering, while 3 were removed, as they had near zero variance. Finally, 1793 features (peak areas) were retained.

Rank Products

Rank Products (RP) is a feature-selection algorithm proposed by Breitling that is used to find the differentially expressed genes.[61] It is an intuitive method that was designed to rank genes according to the up or down-regulation magnitude in all microarray replicates. If one variable is highly ranked in many replicates, greater confidence can be given to the consistence of the results. The Rank Product method has been widely applied in other domains, such as proteomics and metabolomics.[62,63] RP provides an estimator that may be computed for each feature. It is important to compare the obtained RP value with the sampling distribution under the null hypothesis that the differential expression values are identically distributed. In order to do this, we have used the permutation sampling procedure originally proposed by Breitling, using 100 iterations. In order to calculate the RP estimator, the *RankProd* library in R was used.[64] For each feature *f* in *k* replicates, each containing n_i features, the rank product is calculated as:

$$\mathrm{RP}_{\mathrm{f}}^{\mathrm{up}} = \prod_{\mathrm{i}=1}^{\mathrm{k}}\left(\frac{\mathrm{r}_{\mathrm{i,f}}}{\mathrm{n}_{\mathrm{i}}}\right) \qquad \text{(Eq.1)}$$

Where $r_{i,f}$ is the ranking position of the feature *f* in the replicate *i*.

Probabilistic Quotient Normalization

PQN method estimates a gain for every sample by using a reference spectrum.[49,42] This reference spectrum is normally a Quality Control (QC) sample made up of a pooled mix of equal aliquots from a representative set of the study samples. The median of the ratio between the sample and reference spectrum is computed and serves as a quotient to compensate for variations.

$$x_i^{PQN} = \left[\frac{x_{i,1}}{\mathrm{median}\left(\frac{x_{i,1}}{x_{ref,1}}\right)}, \dots, \frac{x_{i,n}}{\mathrm{median}\left(\frac{x_{i,n}}{x_{ref,n}}\right)}\right] \qquad \text{(Eq.2)}$$

for sample *i* with features $x_{i,j}$ - with j ranging from 1 to *n*. x_i^{PQN} is the corrected feature vector for sample *i* and x_{ref} is the reference spectrum with features $x_{ref,j.}$ Note that the median is calculated across samples for each feature. The challenges associated with normalization techniques in metabolomics are discussed by Filzmoser.[42]

Component Correction

CC assumes that the variance introduced by the drift is explained with one or more principal components (PCs) of a given reference class.[30] PCs that define the drift subspace are subtracted from the original data.[54]

$$X_c = X - (X \cdot P)P^T \qquad \text{(eq. 3)}$$

where X_c is the corrected data, X is the original data matrix, and P are the loadings (or principal components) spanning the drift subspace.

Genetic Algorithms

Genetic Algorithms (GA) are a feature-selection technique based on the survival of the fittest individual.[65] An individual is defined as a subset of selected features (binary vector with 1s and 0s, which are taken to mean selected or not selected, respectively). GA initializes a population of many individuals and evaluates their performance according to a predefined criterion (fitness function). The fittest individuals are selected and used to generate a new population through crossover; either by mating and/or by mutation The chosen figure of merit has been the Generalized Fisher Ratio,[66] computed in latent variable space using internal validation samples. This goal function is more sensitive than the classification rate (CR), and it is known that filters are less computationally expensive than wrappers.

The general framework was obtained from *GA* library.[67] Readers interested in an overview of feature selection techniques for omics readers are referred to Gromski et al.[68] and Saeys et al. [21] and the examples therein.

Partial Least Squares – Discriminant Analysis

Partial Least Squares – Discriminant Analysis (PLS-DA) can be understood as a PLS regression between a data matrix, *X*, and a categorical one, *Y*.[69,70] *X* contains the set of predictors, whereas *Y* consists of labels or responses, in this case a binary vector. PLS regression is based on an iterative process to define a new subspace of latent variables (LV).[71] In order to define it, PLS considers a compromise between maximum variance modelling *X* and maximum correlation with *Y*. Different algorithms are used to compute PLS, in this work, the kernel PLS algorithm implemented in *pls* R package has been applied.[72] The use of PLS-DA for metabolomics has been covered extensively in the literature.[73,74]

Random Forests

Random Forests (RF)[75] are machine-learning techniques often used for omics data analysis for the purposes of classification.[76] RF is an ensemble classifier that consists of multiple decision trees, each of which is built using a subset of features and data samples. They also possess an additional advantage in that they deliver an automatic ranking of variable importance that can be used for biomarker discovery.[77] Here the *RandomForest* R package was used.

Data Analysis Protocol

Instrumental drift correction

Data was inspected by PCA in order to determine the effect of time. Blanks were then used for in order to filter out noisy features. To this end, two strategies were adopted: (i) Non-parametric Wilcoxon test for fibre blanks, and (ii) RP for sample blanks. The one-sided Wilcoxon test[78] was used to check which features had a median distribution of intensity values that were higher in sample than in fibre blanks. Metabolites that were equally or less abundant in sample than in fibre blanks were considered as contamination, as they were likely to be non-informative, and as a result, they were discarded. The testing of this hypothesis was performed for all classes taken together, versus fibre blanks, due to the availability of only 7 fibre blanks. The RP method, however, was applied in order to compare the metabolic profiles of each class versus the sample blanks, for which a larger sample of 49 blanks was available. Therefore, four contrasts were defined in order to assess which features possessed more intensity in any of the classes than in sample blanks. The consideration of all samples (control, COPD, COPD-LC, and LC) in the same group in order to undertake the contrast versus sample blanks may have resulted in peaks that were highly abundant only in one class being removed. Thus, in order to avoid removing the relevant variables, all those variables that were more present in any patient group than in blanks were retained.

Contamination and memory effects in the conserved ion fragments were assumed to be additive noise that could be corrected. The intensity measured in the previous blank was then subtracted from the intensity of an ion fragment from a sample.[79,80] The possibility of gain variations in time was explored beyond the additional correction of potential instrument contamination. A normalization strategy inspired in PQN was introduced, which was aimed at removing the so-called *size-effect*.[42] The major difference between PQN and the applied method is that the gain variations were estimated with sample blanks. The multiplicative factor of a sample is given by the median of the ratio between the reference blank and a blank acquired previous to sampling. The reference blank chosen was one of 13/12/11, as intensity fluctuations within that time period were limited. The expression that encompasses both the removal of sample blanks and multiplicative correction is the following.

$$x_{iC} = \text{median}\left(\frac{x_{Blank\,ref}}{x_{Blank\,i}}\right)(x_i - x_{Blank\,i}) \qquad \text{(Eq. 4)}$$

Where x_i is a sample, x_{iC} is the corrected sample, $x_{Blank\ i}$ is the previous blank to the sample, and $x_{Blank\ ref}$ is the blank of reference in the middle of the study (13/11/11).

Moreover, CC was applied using sample blanks as the reference class to estimate drift subspace. Two figures of merit were considered at this stage, in order to optimize the number of PCs to be eliminated: the Pearson Correlation Coefficient of data with time,[81] and the Hotelling T^2 Statistic, which is used to estimate the distance between classes.[82]

Potential Clinical Confounders

Then hypothesis testing was used to check if controlled clinical variables were equally distributed between the defined groups, or if they acted as confounding factors. Tests were carried out to compare classes two-by-two. The Chi-squared test (or Fisher´s Test, if at least one expected frequency was lower than 5) was applied to the factorial variables; sex, smoking status, presence of hypertension, and presence of diabetes.[83] For the numerical variables; age and body mass index (BMI), the non-parametric Wilcoxon Test was used.

Biomarker Discovery

Pre-processing included outlier detection using the Mahalanobis distance, centring and auto-scaling to unit variance. Biomarker candidates were selected by GA, Fisher´s ratio computed in PLS-DA latent variable space being the figure of merit to maximize.[84] Some measures were adopted following the recommendations of R. Leardi and A. Lupiáñez González to ensure a proper choice of selected features that do not lead to model overfitting.[85] One of them was to limit the number of GA iterations according to the fitness saturation of the best individual. Moreover, in order to avoid the exploration of only a reduced part of the search domain, different GA runs with random initialization were performed, and feature selection frequencies were computed. The entropy of this feature selection frequency distribution was introduced as a stopping criterion for GA.[86] The saturation of entropy was related to a stable feature selection that does not change with more runs, i.e., more executions do not lead to an increase in information. The frequency distribution of a random selection (simulation of random selection 1,000 times) was computed, and the final subset consisted of those features selected in more runs than would be expected by chance. A PLS-DA model was built with this feature subset, and the number of latent variables (LV) was optimized using a cross-validation in the calibration set. External validation of the model was carried out using short-term and long-term validation data, and permutation tests of 10,000 iterations were employed to check if the results were statistically significant as compared to a random classification.

At this point it is important to note that the data processing workflow implemented that includes feature selection using GA provides only fragment candidates. Mass spectra matching is needed to obtain analytes from fragment candidates. The NIST Mass Spectral Search Program 2.3 was used for this purpose.

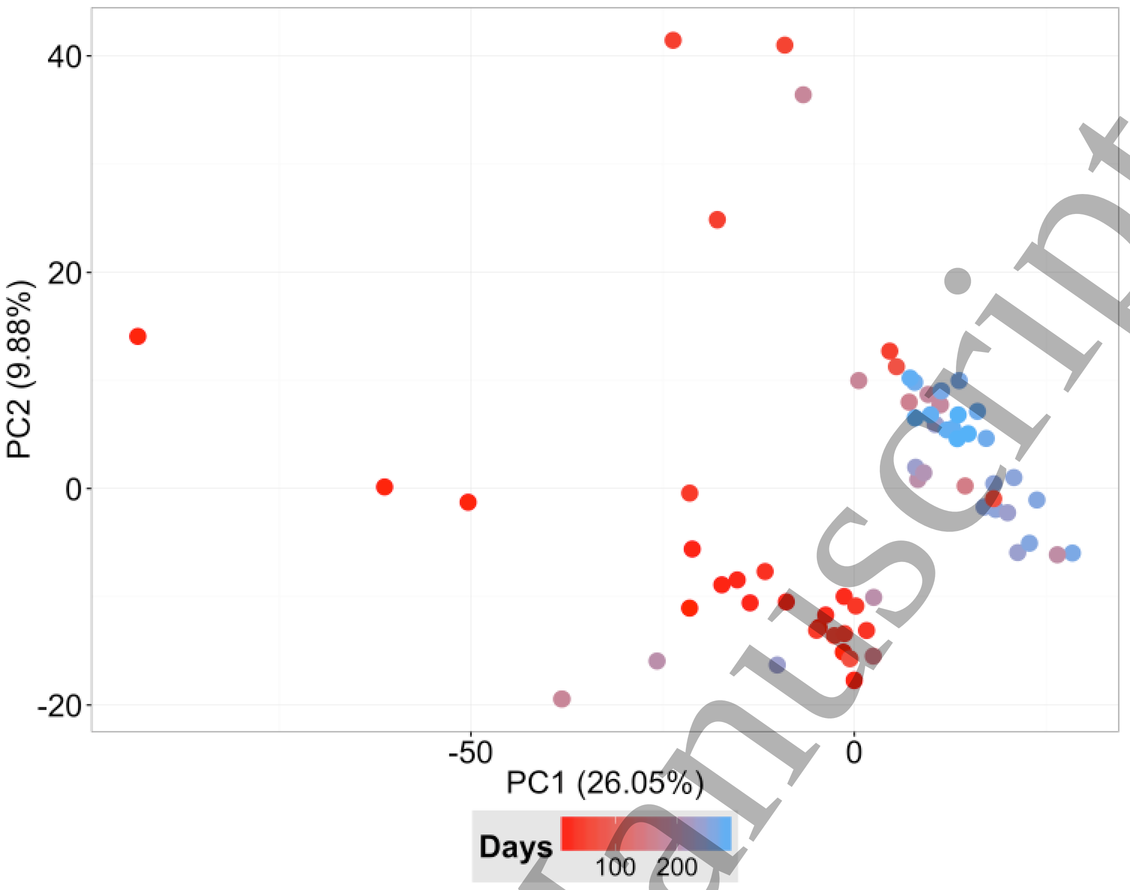
For reasons of comparison, the use of feature selection based on PLS-DA and RF models was also explored and compared to the proposed methodology of GA and PLS-DA. On the one hand, the Variable Importance in Projection (VIP) was applied to the global PLS-DA model (without GA).[87] Moreover, two variable importance measures were used for the RF model. The first was the total decrease in node impurities from splitting on the variable, averaged over all trees (for classification, the Gini index). The second measure of importance was based on the error rate in classification for out-of-bag data, when variable predictors are permuted and averaged over all trees.

Results and Discussion

Instrumental Drift Correction

Data Inspection: Unsupervised Exploration of Time Effects in Uncorrected Data

For a better understanding of high dimensional data, it is convenient to visually inspect the projection to subspace that capture maximum variance using PCA. **Figure 2** shows PCA score plots for calibration and short-term validation subsets. **Figure 2a** is coloured according to date of acquisition and reveals a clear relationship between the maximum variance directions in the data set and time. The two observed clusters are not due to the medical conditions under study, as it can be seen from **Figure 2b**, but consist of unwanted variance that should be removed, since it seems to be causing a separation between samples collected at various times. Therefore, through unsupervised data exploration, an instrumental signal drift explaining a considerable variance of the dataset was detected. However, it was not possible to specify a particular date in which the instrument changed within the calibration subset.


40
20
0
-20
PC2 (9.88%)
-50
0
PC1 (26.05%)
Days
100
200

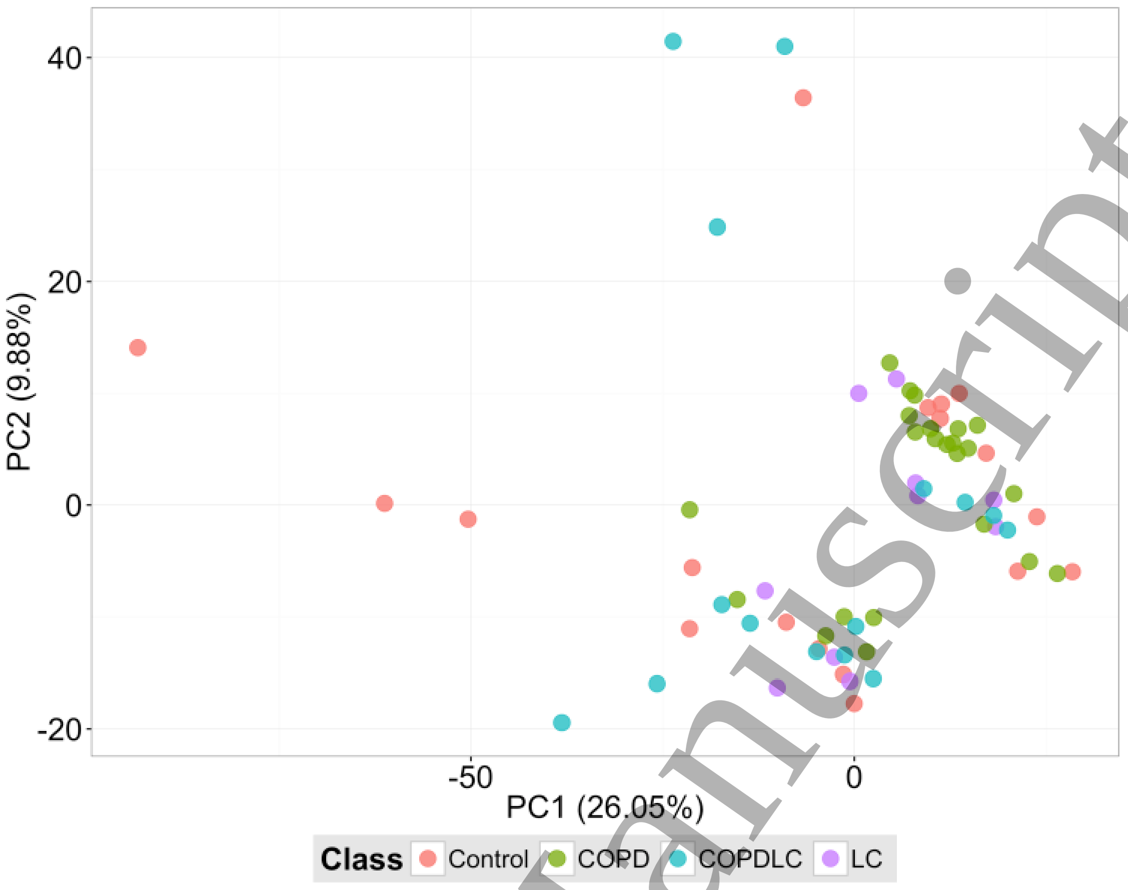


Figure 2. PCA score plot of calibration data. Score plots of PC1 and PC2 are shown **(a)** colouring samples (using a two-colour gradient) according to the date of acquisition (i.e., number of days since the first data acquisition of the study), and **(b)** colouring samples according to the class under study.

Feature filtering

In this section, we set up procedures to select features that are informative about the breath composition of the subjects, while rejecting sample contamination due to the absorbent fibre used or memory effects in the instrument. In order to do so, we took advantage of the two types of blanks were available in this study, fibre and sample blanks. They provide a measurement of baseline noise due to peaks caused respectively by the fibre or memory in the instrument in each day of acquisition. These data were employed to reduce the dimensionality of the dataset by removing noisy features. We would like to emphasize that this feature filtering was exclusively done using calibration data.

First, a univariate Wilcoxon test indicated that 1693 of 1752 features had significantly higher intensity profile in patients' samples than in the analysis of fibre blanks (new sorbent traps without samples). For the rest, the null hypothesis was accepted, and 59 features were removed from the data. In this step, features that fulfilled the null hypothesis were filtered out. With the intention of removing only those features that clearly belong to the blanks, no multiple testing correction was implemented. Multiple correction testing is stricter in order to identify samples that do not conform to the null hypothesis but, in this case, the aim was to identify features that clearly do conform to the null hypothesis.

Second, an alternative use of the RP method, which is often utilized as a univariate feature selection for biomarker discovery in genomics, was proposed. Instead of comparing classes under study among them to find putative biomarkers, binary contrasts of each class vs. sample blanks were computed. Features can have larger value distributions in one or more classes with respect to sample blanks, and RP results indicated that 79 features had higher intensity in one study class, 66 in two, 84 in three, and 243 in the four classes with respect to blanks. With this RP strategy, dimensionality was further reduced, from 1693 to 472 features. **Figure 3** shows that eliminated features mainly correspond to compounds eluting at longer RTs, which might be related to compounds of lower volatility, higher molecular weight, or more affinity with the chromatographic column. Since these RTs were shown to contribute more in the PCA loadings of the initial data, they largely hide the effect of lower intensity features. As a consequence, feature filtering did not only reduce computational cost, but probably improved the power in identifying real biomarkers.

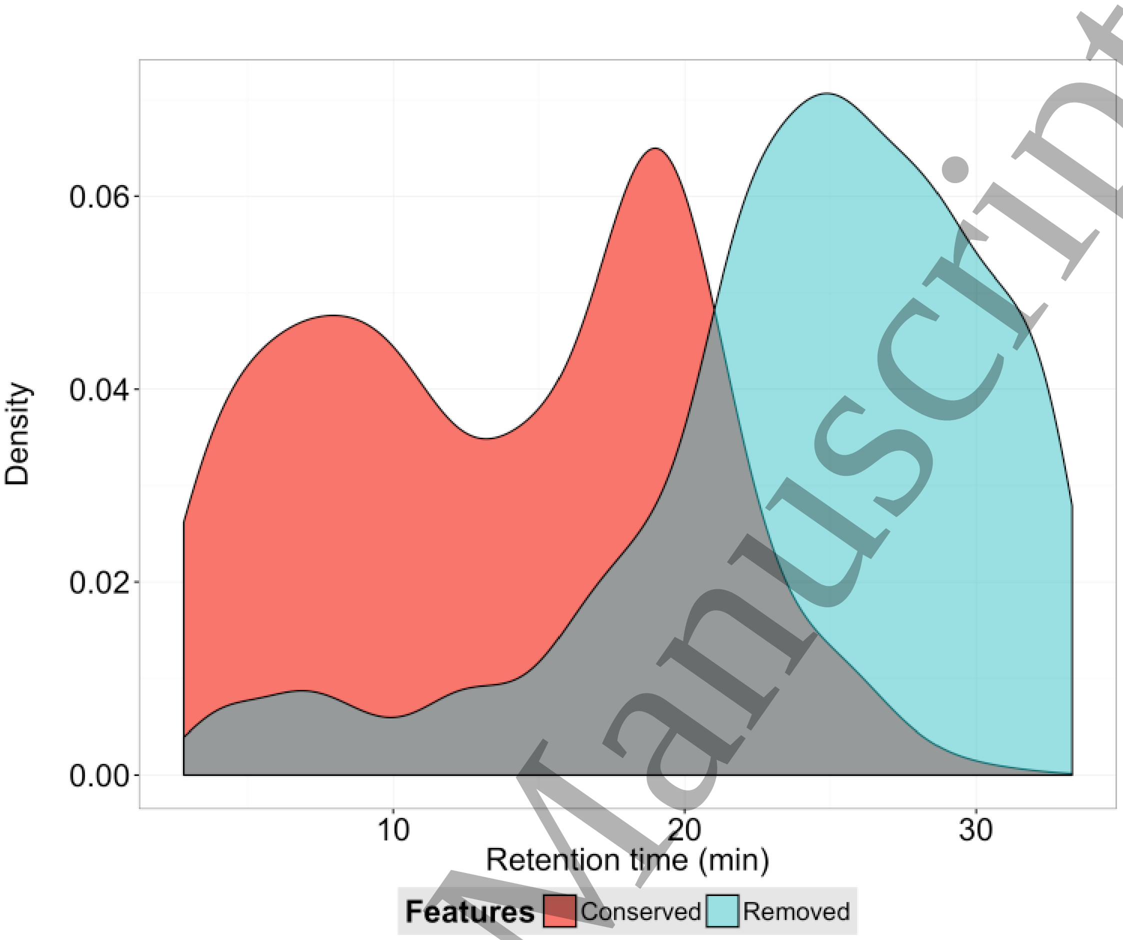


Figure 3. Data filtering based on sample blanks data. Density estimations are shown for the retention times of the conserved (red) and removed (blue) features using sample blanks data.

The score plot in **Figure 4** indicates that despite this filtering, the data still showed a temporal separation of samples. In this figure, long-term validation samples were also added to show how they differ from the other data points. When the long-term validation subset is considered, the first PC reflects the separation between them and the rest of samples, and the second PC is the one which reveals the two previously observed clusters.

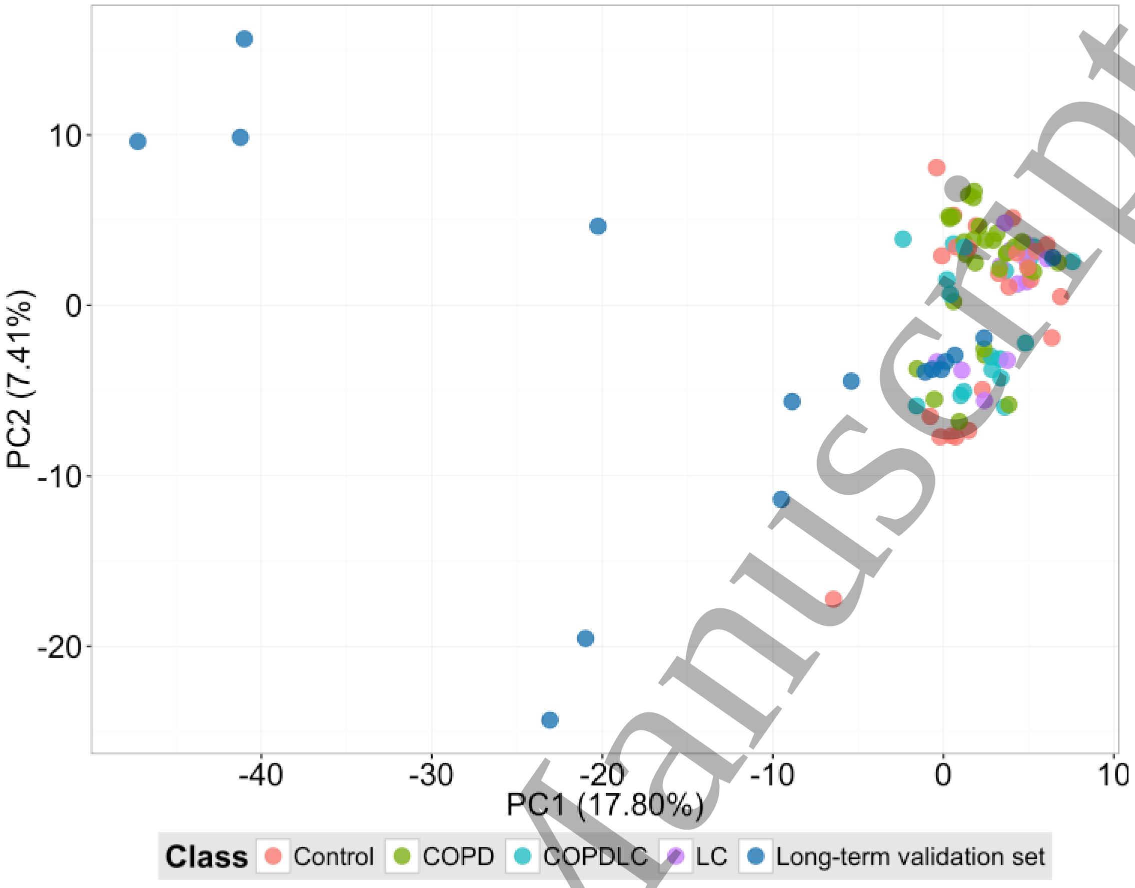


Figure 4. PCA score plot of all the data of the study after filtering. Score plots of PC1 and PC2 for all the samples (calibration, short-term, and long-term blinds) after data filtering using blanks data (i.e., data set with 472 features).

Figure 5 represents the first PC scores with respect to the date of acquisition and also shows that long-term data are remarkably different from the samples collected during the first year of the study. Acquisition of the last data subset was over a long period of time, which results in different batches even within long-term validation samples.

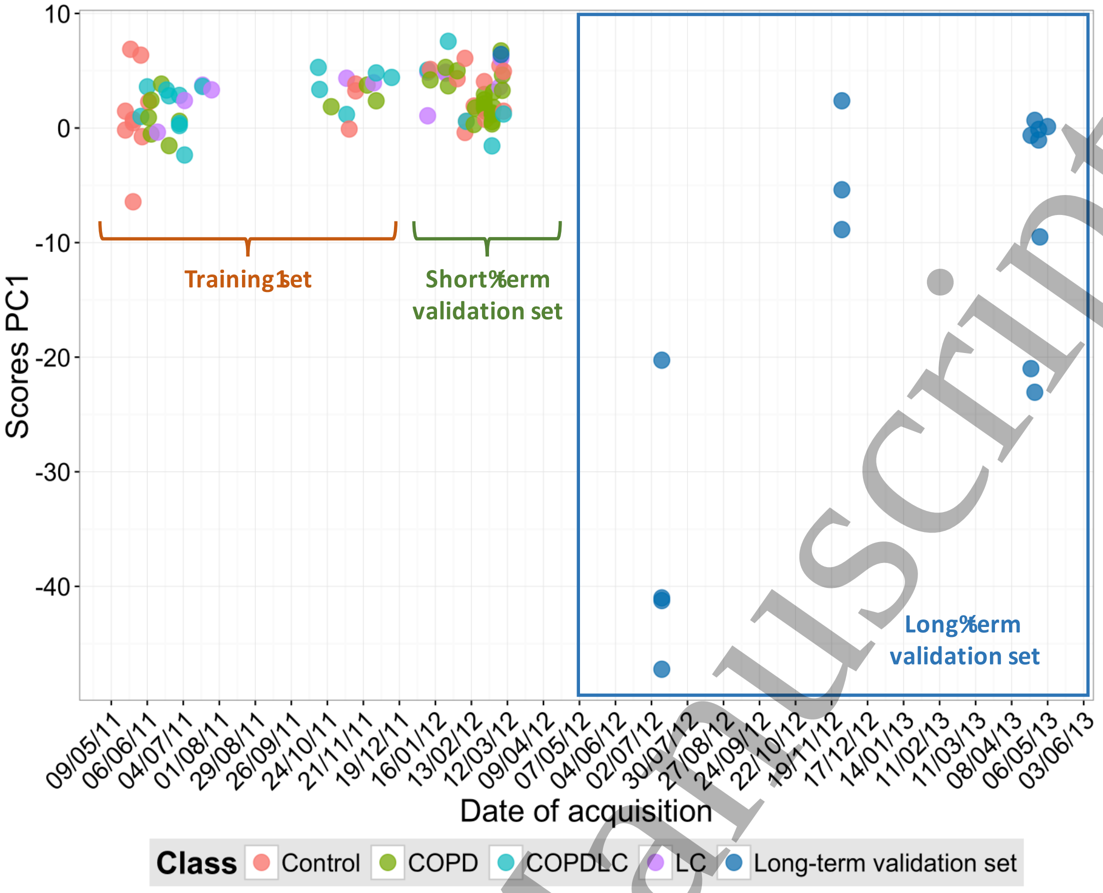


Figure 5. Change of PC1 scores over time. Score plots of PC1 are reported for all the samples with respect to the date of acquisition for filtered data (i.e., 472 features).

Even though in this study there were no QC samples, inspection of the changes in sample blanks allowed us to learn about time effects. Our initial observation was that sample blanks shared a lot of common fragments. On this basis, additive and multiplicative corrections were proposed.

Additive and multiplicative corrections

We have already mentioned that ion fragments with the same amplitude in the real samples and in the blanks were discarded. Beyond this additive correction, sample blanks showed contamination and memory effects in the conserved features. Therefore, each sample had the intensity values of the previous blank subtracted from its intensities. Even after additive correction, large variations in the intensity of the chromatograms were still observed and the possibility of gain variations over time was explored. **Figure 6** shows boxplots of the intensity ratios between a reference blank in the middle of the study and the rest. While ratios among fragments showed some scatter, there was a clear central value that differed across days. This effect leads us to the hypothesis of instabilities of sensitivity over time. In order to compensate for this, a gain correction was estimated through the median of the ratios across ion fragments. **Figure 6** reports a clear pattern for a change in the gain over time. There was a sharp variation in the gain on a particular date (10/07/12), which separates the group of the short-term and long-term validation data.

In order to compensate for the observed time effect, the multiplicative correction method described by Equation 4 was used—this assumes that there is a uniform gain variation across features for each sample. Additionally, an underlying hypothesis for this correction is that the estimated gain variation with sample blanks is the same for all the samples measured in this particular day. This normalization approach to correct the multiplicative effect is inspired by PQN, but using data from blanks as reference.

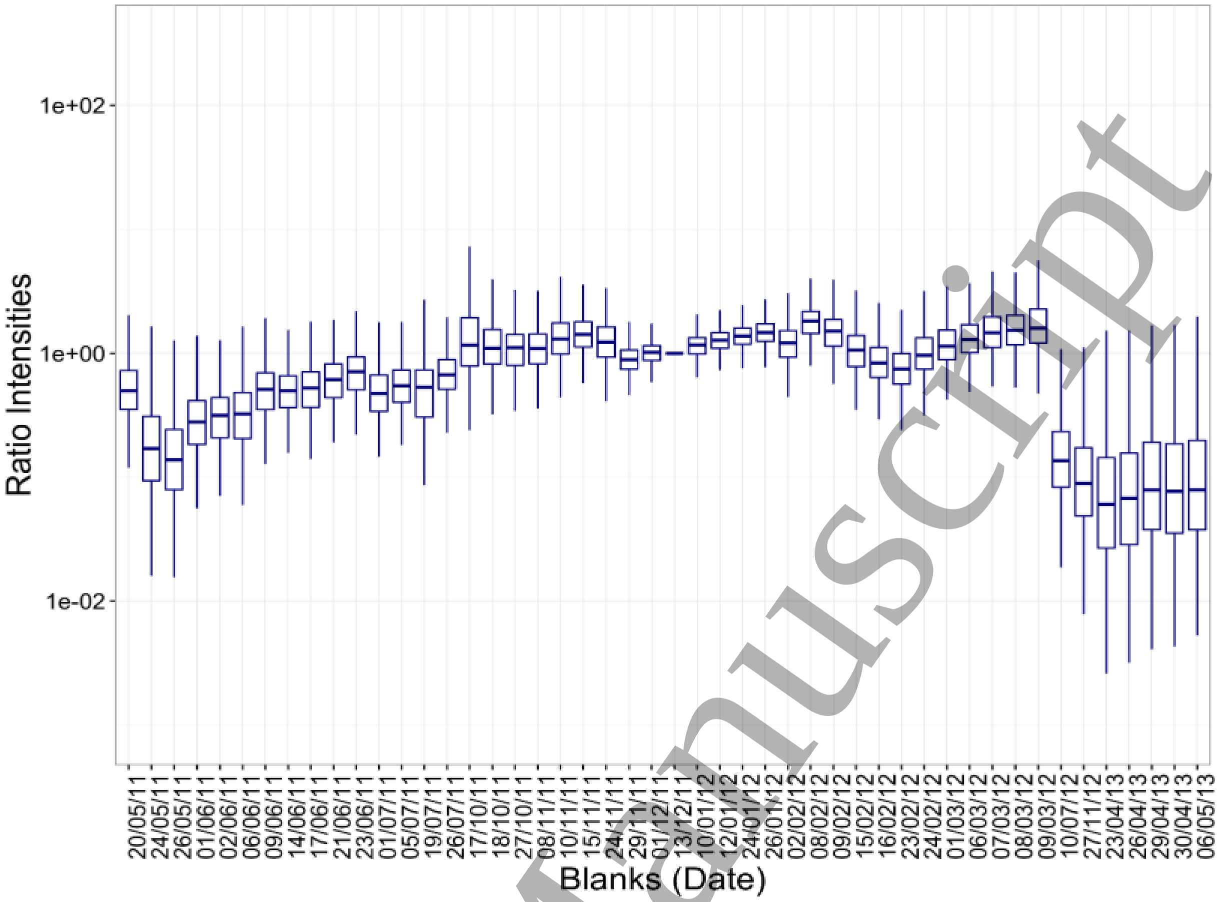


Figure 6. Intensity ratios of sample blanks. The ratios between a reference blank (13/12/11) and the rest of blanks are shown for each date of acquisition. Note that the time scale in x-axis is not uniform.

The application of the former normalization method minimized the effect of gain variations, and the long-term validation data was especially effective in diminishing their variability with respect to the rest of samples. However, temporal variation might also be correlated among features and, consequently, some privileged drift directions may exit. To account for such multivariate correction of instrumental drift, CC was implemented.

Component correction

For CC, sample blanks were chosen as the reference class. Variance in the reference class was modelled with PCA. The technique assumes that the drift subspace was shared between the blanks and the samples. This drift subspace is spanned by a number of principal components estimated from the PCA of the blanks. Since the removed component(s) might be explaining not merely the drift, but also biological differences under study, it is important to optimize the number of removed components and making an appropriate trade-off is essential. Correlation of data with time decreased with respect to the original data when removing the first PCs of sample blanks, so CC appeared to diminish the temporal dependence. Moreover, the use of the two-sample Hotelling T^2 statistic was proposed to estimate distance between classes. Results showed that removing the first PC of the blanks in the original data matrix maximizes the Hotelling T^2 statistic. Therefore, one PC was shown to minimize correlation with time and maximize distance between groups in binary contrasts (as measured by the Hotelling statistic).

Figure 7 shows the PCA score plot for data corrected by the methodology we have outlined. The two “date of acquisition” data clusters vanished after the correction was applied. In addition, long-term validation samples are closer to the other data points. Visually there has been a great improvement, but the correlation of the data with time was also computed to compare the methods, in particular the four cases explained above i.e. feature filtering (FF), feature filtering plus additive and multiplicative correction (AM), feature filtering plus CC, and the combination of all the methods (FF+AM+CC).

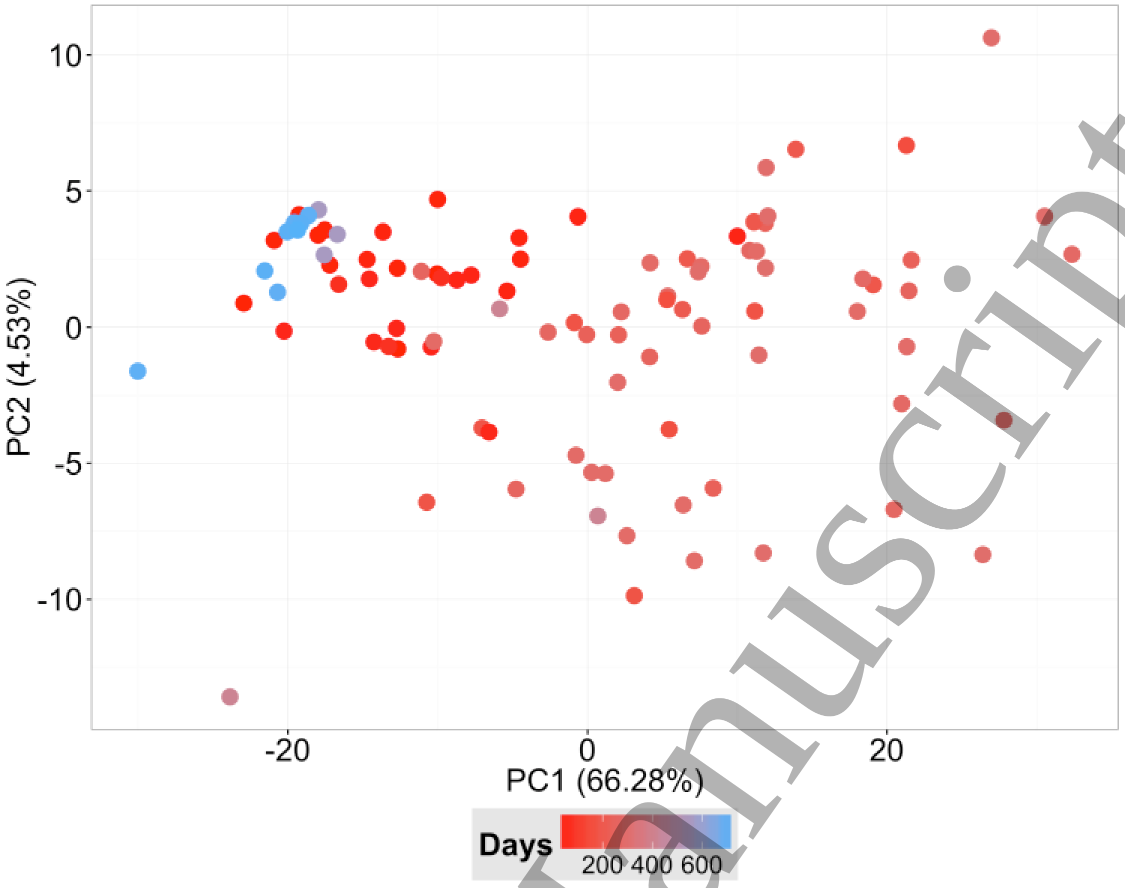


Figure 7. PCA score plot of all data after multiplicative and Component Correction. Score plots of PC1 and PC 2 are reported after multiplicative and component correction based on sample blanks data.

Figure 8 shows that correlation with time substantially diminished when applying the three strategies: feature filtering, additive and multiplicative correction, and CC. Therefore, blanks were successfully used to diminish the significant instrumental drift present in the GC-MS data.

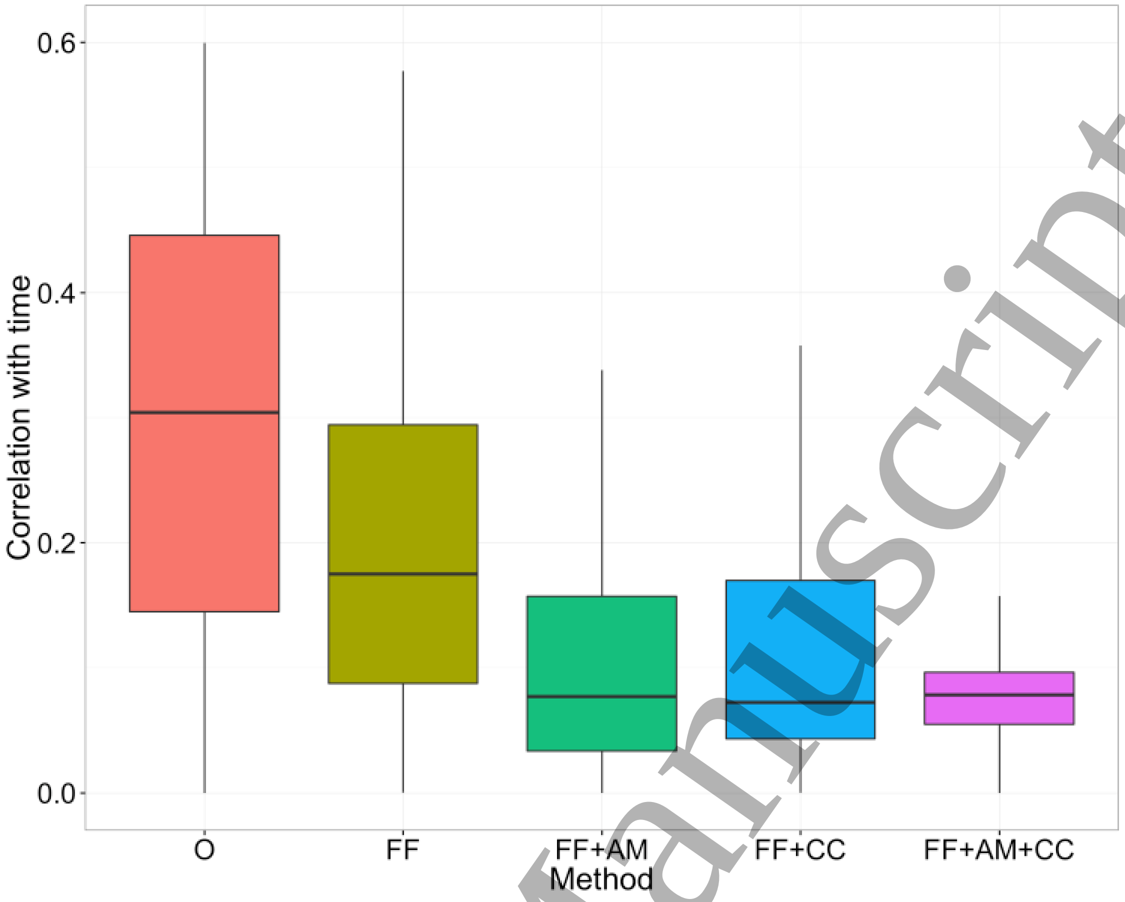


Figure 8. Comparison of applied corrections. The correlation with time (r coefficient) is shown for original data (O), data after feature filtering (FF), FF plus additive and multiplicative corrections (AM), FF plus component correction (CC), and the combination of the three strategies, FF+AM+CC.

Potential clinical confounders

The clinical variables, sex, age, BMI, smoking status, and the two more frequent comorbidities in the dataset, hypertension and diabetes mellitus II, were tested for different distributions between study groups. These factors were equally distributed between the study groups COPD and control (with 95% confidence). However, there were potential confounding factors for the cases of COPDLC (hypertension, age, and smoking status) and LC (hypertension).

Biomarker discovery

LC and COPD are complex diseases; indeed, both are a family of condition subtypes and COPD can be an indirect warning of LC. Hence, identification of VOCs signatures which allow proper classification of all condition subtypes is a real challenge, especially with insufficient data. Additionally, the consideration of clinical confounders in the final model complicates biomarker discovery even further. Due to these difficulties, the present study has corrected instrumental drift for the complete data set, but has focused on the case of COPD for the biomarker discovery stage. COPD intensity profiles were compared with those of the control group. This is the condition-control binary classification problem but with more samples available (15 for control, after eliminating 3 outliers, and 23 for COPD in the calibration set), and clinical confounders were not found.

Biomarker selection

Our biomarker selection strategy aimed at finding a small subset of features that may discriminate the two classes, in contrast to other strategies that identify volatile fingerprints with several tens or even hundreds of compounds. The discrimination of disease by a small subset of variables opens the possibility of further development of specific sensor kits for point of care, or clinical applications after biomarker identification.

The search technique based on GA looked for the features that maximize the distance between the classes in the PLS-DA subspace. The selection reached the entropy saturation criterion after 35 GA runs. **Figure 9** reports the frequency of selection of the features after these 35 runs. The hypothesis of this approach is that combinations of features that exhibit random correlations tend to be selected less often than real discriminant features after some trials. **Figure 9** shows that nine features were systematically chosen by the algorithm. Hence, they were assumed to contain class discriminant information and constituted the subset of putative biomarkers.

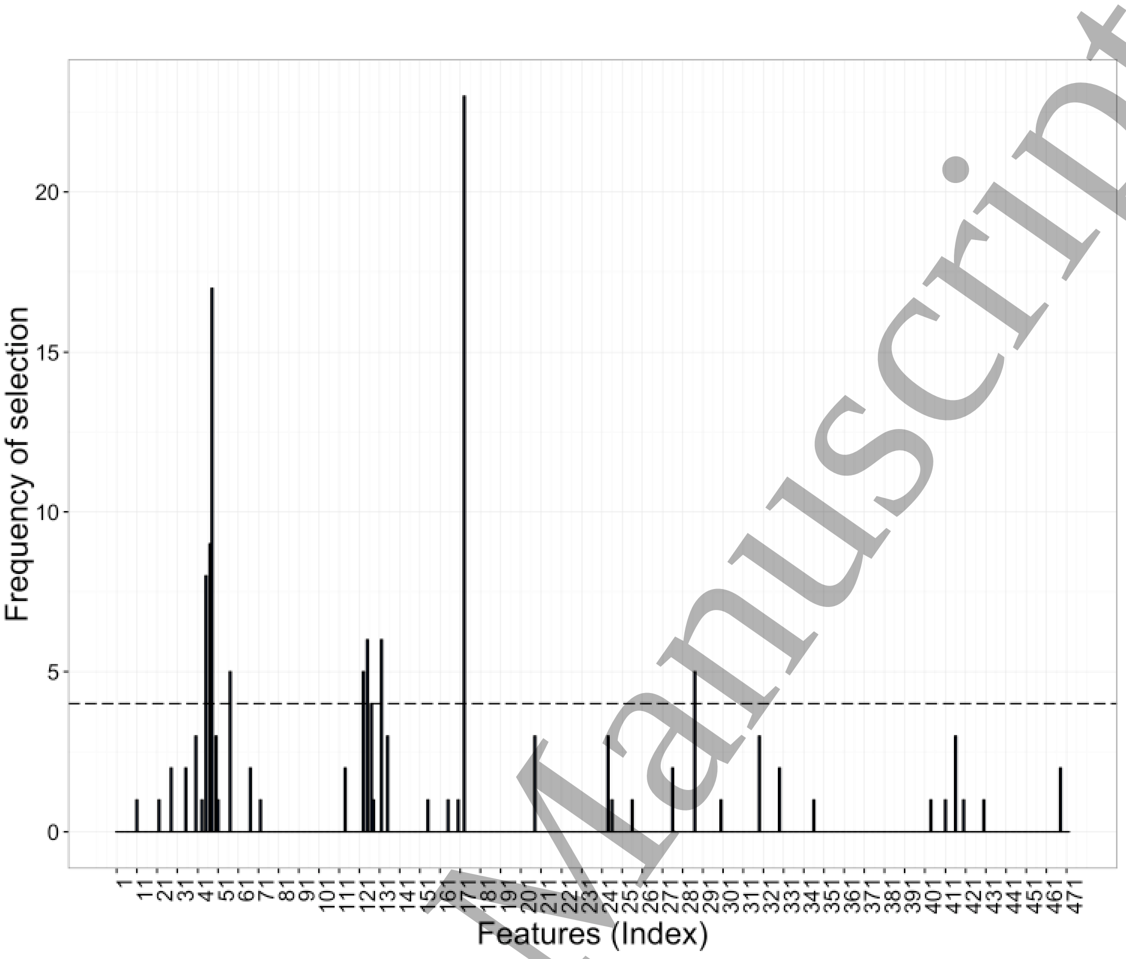


Figure 9. GA selection. Frequency of selection for every feature after 35 runs of the GA. Features above the threshold were ultimately selected, while features below the threshold were selected as often as would be expected at a random.

Metabolite identification

As we have already mentioned, the methodology implemented provides only fragments, or groups of fragments, that are characterized by an RT and a, m/z. To identify the metabolites, the correlation threshold for fragment grouping, which initially was set to 0.9, was decreased. By decreasing this threshold to 0.6 for fragments appearing at the same RT (with a window of 12 seconds), more complete mass spectra were obtained, which allowed the identification of three putative biomarkers. A fourth set of fragments did not match any candidate compound from the NIST library.

(i) **Isoprene**. This compound appears both in healthy and COPD subjects. Isoprene is a major metabolite in the breath. It has been mentioned as a product of lipid peroxidation and previous studies have linked it to cholesterol metabolism.[88] This compound has been proposed as biomarker for COPD in previous studies[5] and has been detected in breath associated with other respiratory diseases,[89] including lung cancer.[90] The utility of isoprene as biomarker has been reviewed by Salerno-Kennedy and Cashman.[91]

(ii) **Succinic Acid (Succinate).** Succinate in living organism has many roles as a metabolic intermediate and as a signal of metabolic state at the cellular level.[92] It is generated in the mitochondria through the Krebs cycle. Succinate has been proposed as a metabolic signal for inflammation,[93] and has been found at an increased level in inflammatory bowel disease and colitis.[93]

(iii) **Pentamethylene sulphide**. This compound appears in diverse patents as related to medication of COPD as a chymase inhibitor[94,95] and also in the formulation of anti-inflammatory agents.[96] Consequently, we suspect that this is not a biogenic compound, and may be related to the patients' medication.

The ranking of fragments provided by VIP in PLS-DA global model, and the feature importance provided by RF, were also compared. We note that the models were much less sparse than the ones obtained after our GA feature selection. Particularly for VIP in PLS-DA, there were many fragments providing similar values. There was not a good matching between the fragments selected by GA, and the ones ranked higher in VIP. On the other hand, RF gave a very high (top five positions) ranking to fragments related to Isoprene, but the rest of fragments selected by GA were medium-ranked by RF.

Predictive modelling: Assessment of accuracy in short-term external validation

Calibration data were used to assess model complexity assessment, both feature selection and optimization of the number of LV. Internal cross-validation determined that LV=2 was the optimum hyperparameter for a PLS-DA classifier with the selected features. A final PLS-DA model was built with all the calibration data, in order to estimate its predictive performance when classifying new samples as control or COPD.

Figure 10 shows the PLS-DA score plot for control and COPD classification, including calibration and short-term validation subsets. Control and COPD groups have distributions that are slightly separated. The first LV gives two density distributions with a certain overlap and the inclusion of the second LV in the model provides a small increment in the distance between classes. For short-term validation data, the model correctly classified 77% of the samples (CR=77%) and had an Area Under the ROC Curve (AUC) of 0.75. Sensitivity was 100%, i.e., all patients with the disease had a positive result, whereas the specificity was lower, 50%. Obviously, the trade-off between sensitivity and specificity can be adjusted by proper selection of the classifier output threshold. Moreover, permutation tests for CR and AUC distributions gave p-values of 0.015 and 0.029, respectively, which are statistically significant.

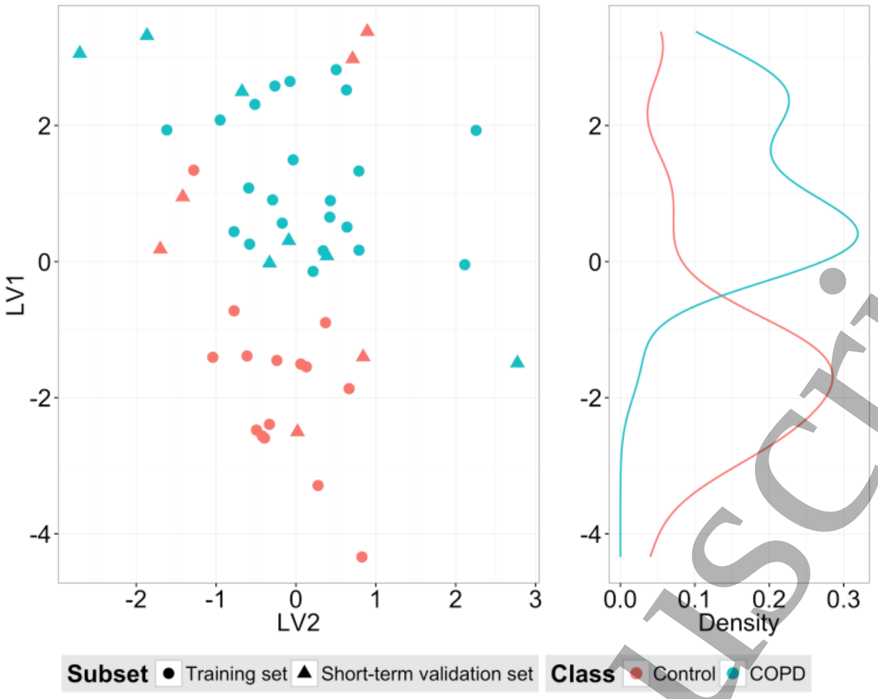


Figure 10. PLS-DA scores. LV1 and LV2 scores for calibration and short-term validation data subsets of control and COPD groups are reported.

Results were compared to other standard classifiers in metabolomics, namely: (i) PLS-DA on the corrected data without any feature selection using GA and (ii) RF. For PLS-DA, internal validation provided a model with five latent variables with a classification in the short-term external validation which resulted in an accuracy of CR=0.73±0.13 and an AUC=0.74. Maximum accuracy resulted in a sensitivity of 100%, but a poor specificity of 33%. For RF, the same level of accuracy CR=0.69±0.13 was reached, but with a poorer AUC of 0.60. These results were achieved with an optimum number of variables (*mtry* in the RandomForest Package) of 25.

Model	CR	Sensitivity	Specificity	AUC

PLS-DA with GA	77±12%	100%	50%	0.75
Random Forest	69±13%	100%	33%	0.60
PLS-DA	73±13%	100%	40%	0.70

Table 2: Predictive model performance in short-term external validation.

Table 2 shows that all models show a moderate prediction capability, despite the drift in the dataset. However, there is clear evidence that, after feature selection, the PLS-DA model achieves a higher accuracy, and has the added advantage of providing parsimonious models involving a reduced number of biomarkers.

Predictive modelling: Assessment of accuracy in long-term external validation

One of the objectives of the present investigation is to improve the validation of the predictive models, and the biomarkers involved in the prediction of the subject conditions. Unfortunately, most published works disregard this step, and present the results with internal validation or external validation data extracted from the same measurement regime. When these external validation data (blind samples), are time interleaved with calibration samples, the time effects do not play a role since, the model 'knows' about the future changes in the dataset. However, in the quest for the maximum reliability of the chosen biomarkers, the predictive models should, in the future, be able to deliver accurate classification of the calibration. In other words, the model should be resilient to instrumental or operator changes.

Results for the long-term validation of the PLS-DA model with the selected biomarkers were equivalent to a random classification (CR=55% with a p-value of 0.726). Therefore, the built model had a certain degree of predictive ability in the near future, but this did not hold for samples acquired during the second year of the study. Instrumental signal drift caused considerable changes on intensity profiles, chiefly during long-term validation samples acquisition, making these samples fall outside the applicability domain of the model. This loss of predictive power was also observed for PLS-DA without feature selection (CR=45%), and for the RF (CR=36%). These results show that, on many occasions, instrumental drift called the final application of the selected biomarkers into question. More emphasis on long-term validation of results and further research on instrumental drift correction are required.

Conclusions

We do not yet completely understand how instrumental drift affects biomarker discovery studies. The presence of, inter alia, column aging, temperature variations, or maintenance operations causes signal drift and hinders the development of models with long-term predictive power. This raises a dilemma for biomarker discovery studies: analysing more samples leads to the risk of there being more instrumental drift in the signals, but a small sample size may not be statistically robust. This is particularly relevant for breath analysis, since standardized procedures for the storage of breath samples over a long timescale, or for the use of pooled QC samples, simply do not exist.

In the present study, a GC-MS dataset was obtained from exhaled breath samples and a posterior analysis conducted. We have shown how the instrumental drift problem can be counteracted in one specific biomarker discovery study. For this particular study, we have proposed data processing techniques to minimize instrumental drift using information from blanks. In summary, the methodology consisted of using such data for feature filtering and removal, a multiplicative correction to compensate for the estimated instrumental gain variations in each spectrum, and CC to account for a drift subspace correlated among features. This combination of techniques was successfully applied and considerably diminished the correlation of the data with time. It is important to be aware that a complete removal of instrumental drift might also eliminate discriminant information. Therefore, signal drift compensation is a trade-off between removing an unwanted source of variance, and losing the variability of interest. As far as we know, RP have not been previously applied in this context of feature filtering, since it is generally used for condition vs. control instead of condition/control vs. blanks. We also proposed two figures of merit for the choice of the number of PCs to be extracted in CC: correlation with time (minimization) and the Hotelling T^2 statistic (maximization).

In addition, the combination of a feature selection strategy based on GA, and a PLS-DA model, allowed us to classify short-term validation samples in COPD or control classes with an accuracy of 77% and an AUC of 0.75, both these magnitudes being statistically significant under a permutations test. These results were slightly better than competing algorithms such as PLS-DA without GA, and RF.

While the methodology implemented is sufficient to preserve model accuracy in external short-term validation, the model could not extrapolate and satisfactorily predict long-term test samples which, due to instrumental variations, fall outside its domain of applicability. In other words, the methods implemented were not capable of predicting the instrumental variations over a longer time horizon.

In conclusion, we note that instrumental drift entails substantial difficulties in building models that are nonlocal in time. The predictive models obtained are unduly sensitive to instrumental conditions and in consequence are not robust enough for use in the field. We encourage other researchers to devise additional validation levels to improve research reproducibility.

Acknowledgements

This work was partially funded by the Spanish MINECO program, under grants TEC2011-26143 (SMART-IMS), TEC2014-59229-R (SIGVOL), and BES-2015-071698 (SEVERO-OCHOA). The Signal and Information Processing for Sensor Systems group is a consolidated Grup de Recerca de la Generalitat de Catalunya and has support from the Departament d'Universitats, Recerca i Societat de la Informació de la Generalitat de Catalunya (expedient 2014-SGR-1445). This work has received support from the Comissionat per a Universitats i Recerca del DIUE de la Generalitat de Catalunya and the European Social Fund (ESF). Additional financial support has been provided by the Institut de Bioenginyeria de Catalunya (IBEC). IBEC is a member of the CERCA Programme / Generalitat de Catalunya.

References

1 A. J. Atkinson, W. A. Colburn, V. G. DeGruttola, D. L. DeMets, G. J. Downing and D. F. Hoth, 2001, **69**, 89–95.

2 A. Amann, B. D. L. Costello, W. Miekisch, J. Schubert, B. Buszewski, J. Pleil, N. Ratcliffe and T. Risby, *J. Breath Res.*, 2014, **8**, 34001.

3 Y. Y. Broza, L. Zuri and H. Haick, *Sci. Rep.*, 2014, **4**, 1–6.

4 R. F. Machado, D. Laskowski, O. Deffenderfer, T. Burch, S. Zheng, P. J. Mazzone, T. Mekhail, C. Jennings, J. K. Stoller, J. Pyle, J. Duncan, R. A. Dweik and S. C. Erzurum, *Am. J. Respir. Crit. Care Med.*, 2005, **171**, 1286–1291.

5 J. J. B. N. Van Berkel, J. W. Dallinga, G. M. Möller, R. W. L. Godschalk, E. J. Moonen, E. F. M. Wouters and F. J. Van Schooten, *Respir. Med.*, 2010, **104**, 557–563.

6 B. Schmekel, F. Winquist and A. Vikström, *Anal. Chim. Acta*, 2014, **840**, 82–86.

7 W. Miekisch, J. K. Schubert and G. F. E. Noeldge-Schomburg, *Clin. Chim. Acta.*, 2004, **347**, 25–39.

8 C. Lourenço and C. Turner, *Metabolites*, 2014, **4**, 465–498.

9 K. Dettmer, P. A. Aronov and B. D. Hammock, 2007, 51–78.

10 A. Smolinska, L. Blanchet, L. M. C. Buydens and S. S. Wijmenga, *Anal. Chim. Acta*, 2012, **750**, 82–97.

11 M. Vinaixa, S. Samino, I. Saez, J. Duran, J. J. Guinovart and O. Yanes, *Metabolites*, 2012, **2**, 775–795.

12 R. Bellman, *Adaptive control processes : a guided tour*, .

13 C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer-Verlag, 2006.

14 D. Ghosh and L. M. Poisson, *Genomics*, 2009, **93**, 13–16.

15 D. Donoho, *AMS Math Challenges Lect.*, 2000, 1–33.

16 P. K. Sen, *Austrian J. Stat.*, 2016, **35**, 197–214.

17 J. Fan and J. Lv, *Stat. Sin.*, 2010, **20**, 101–148.

18 P. Buhlmann and S. Van De Geer, *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer Science and Business Media, 2011.

19 N. C. Chung and J. D. Storey, *Bioinformatics*, 2015, **31**, 545–554.

20 A.-L. Boulesteix and K. Strimmer, *Brief. Bioinform.*, 2006, **8**, 32–44.

21 Y. Saeys, I. Inza and P. Larrañaga, *Bioinformatics*, 2007, **23**, 2507–17.

22 A. Jain and D. Zongker, *IEEE Trans. Pattern Anal. Mach. Intell.*, 1997, **19**, 153–158.

23 W. Siedlecki and J. Sklansky, *Pattern Recognit. Lett.*, 1989, **10**, 335–347.

24 E. Correa and R. Goodacre, *BMC Bioinformatics*, 2011, **12**, 33.

25 R. and Leardi, *Compr. Chemom.*, 2009, 631–653.

26 Z. Ramadan, D. Jacobs, M. Grigorov and S. Kochhar, *Talanta*, 2006, **68**, 1683–1691.

27 W. Miekisch, S. Kischkel, A. Sawacki, T. Liebau, M. Mieth and J. K. Schubert, *J. Breath Res.*, 2008, **2**, 26007.

28 a. Amann, W. Miekisch, J. Pleil, T. Risby and J. Schubert, *Eur. Respir. Monogr.*, 2010, 96–114.

29 W. Miekisch, J. Herbig and J. K. Schubert, *J. Breath Res.*, 2012, **6**, 36007.

30 A. Perera, *Bioinformatics*, 2014, **30**, 2899–2905.

31 M. A. Kamleh, T. M. D. Ebbels, K. Spagou, P. Masson and E. J. Want, *Anal. Chem.*, 2012, **84**, 2670–2677.

32 M. Sysi-Aho, M. Katajamaa, L. Yetukuri and M. Oresic, *BMC Bioinformatics*,

2007, **8**, 93.

33 C. Deport, J. Ratel, J. L. Berdagué and E. Engel, *J. Chromatogr. A*, 2006, **1116**, 248–258.

34 M. M. W. B. Hendriks, F. A. van Eeuwijk, R. H. Jellema, J. A. Westerhuis, T. H. Reijmers, H. C. J. Hoefsloot and A. K. Smilde, *TrAC - Trends Anal. Chem.*, 2011, **30**, 1685–1698.

35 W. Cao and Y. Duan, *Clin. Chem.*, 2006, **52**, 800–811.

36 T. H. Risby, *J. Breath Res.*, 2008, **2**, 30302.

37 C. O. Phillips, Y. Syed, N. M. Parthalain, R. Zwiggelaar, T. C. Claypole and K. E. Lewis, *J. Breath Res.*, 2012, **6**, 36003.

38 S. Marco, *Anal. Bioanal. Chem.*, 2014, **406**, 3941–3956.

39 M. A. Pourhoseingholi, A. R. Baghestani and M. Vahedi, *Gastroenterol. Hepatol. from Bed to Bench*, 2012, **5**, 79–83.

40 A. M. De Livera, M. Sysi-Aho, L. Jacob, J. A. Gagnon-Bartsch, S. Castillo, J. A. Simpson and T. P. Speed, *Anal. Chem.*, , DOI:10.1021/ac502439y.

41 a Smolinska, a C. Hauschild, R. R. Fijten, J. W. Dallinga, J. Baumbach and F. J. van Schooten, *J. Breath Res.*, 2014, **8**, 27105.

42 P. Filzmoser and B. Walczak, *J. Chromatogr. A*, 2014, **1362**, 194–205.

43 A. M. De Livera, D. A. Dias, D. De Souza, T. Rupasinghe, D. L. Tull, U. Roessner, M. J. Mcconville and T. P. Speed, *Anal. Chem.*, 2012, **84**, 10768–10776.

44 R. Wehrens, J. A. Hageman, F. van Eeuwijk, R. Kooke, P. J. Flood, E. Wijnker, J. J. B. Keurentjes, A. Lommen, H. D. L. M. van Eekelen, R. D. Hall, R. Mumm and R. C. H. de Vos, *Metabolomics*, , DOI:10.1007/s11306-016-1015-8.

45 C. Christin, H. C. J. Hoefsloot, a. K. Smilde, B. Hoekman, F. Suits, R. Bischoff

and P. Horvatovich, *Mol. Cell. Proteomics*, 2013, **12**, 263–276.

46 W. Zou, J. She and V. V Tolstikov, *Metabolites*, 2013, **3**, 787-819, 33 .

47 J. Kuligowski, A. Sanchez-Illana, D. Sanjuan-Herraez, M. Vento and G. Quintas, *Analyst*, 2015, **140**, 7810–7817.

48 R. Di Guida, J. Engel, J. W. Allwood, R. J. M. Weber, M. R. Jones, U. Sommer, M. R. Viant and W. B. Dunn, *Metabolomics*, , DOI:10.1007/s11306-016-1030-9.

49 F. Dieterle, A. Ross, G. Schlotterbeck and H. Senn, *Anal. Chem.*, 2006, **78**, 4281–4290.

50 Lammerhofer M. and Weckwerth W, *Metabolomics in Practice*, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 2013.

51 K. A. Veselkov, L. K. Vingara, P. Masson, S. L. Robinette, E. Want, J. V. Li, R. H. Barton, C. Boursier-Neyret, B. Walther, T. M. Ebbels, I. Pelczer, E. Holmes, J. C. Lindon and J. K. Nicholson, *Anal. Chem.*, 2011, **83**, 5864–5872.

52 W. E. Johnson, C. Li and A. Rabinovic, *Biostatistics*, 2007, **8**, 118–127.

53 a. Ziyatdinov, S. Marco, a. Chaudry, K. Persaud, P. Caminal and a. Perera, *Sensors Actuators B Chem.*, 2010, **146**, 460–465.

54 M. Padilla, a. Perera, I. Montoliu, a. Chaudry, K. Persaud and S. Marco, *Chemom. Intell. Lab. Syst.*, 2010, **100**, 28–35.

55 M. Bylesjö, D. Eriksson, A. Sjödin, S. Jansson, T. Moritz and J. Trygg, *BMC Bioinformatics*, 2007, **8**, 207.

56 J. Trygg and S. Wold, *J. Chemom.*, 2002, **16**, 119–128.

57 J. Trygg, *J. Chemom.*, 2002, **16**, 283–293.

58 M. Phillips, K. Gleeson, J. M. Hughes, J. Greenberg, R. N. Cataneo, L. Baker and W. P. McVay, *Lancet (London, England)*, 1999, **353**, 1930–3.

59 T. Pluskal, S. Castillo, A. Villar-Briones and M. Oresic, *BMC Bioinformatics*,

2010, **11**, 395.

60 M. Katajamaa, J. Miettinen and M. Oresic, *Bioinformatics*, 2006, **22**, 634–636.

61 R. Breitling, P. Armengaud, A. Amtmann and P. Herzyk, *FEBS Lett.*, 2004, **573**, 83–92.

62 S. Smit, M. J. van Breemen, H. C. J. Hoefsloot, A. K. Smilde, J. M. F. G. Aerts and C. G. de Koster, *Anal. Chim. Acta*, 2007, **592**, 210–217.

63 A. Fukushima, M. Kusano, H. Redestig, M. Arita and K. Saito, *BMC Syst. Biol.*, 2011, **5**, 1.

64 F. Hong, R. Breitling, C. W. McEntee, B. S. Wittner, J. L. Nemhauser and J. Chory, *Bioinformatics*, 2006, **22**, 2825–2827.

65 D. E. (David E. Goldberg, *Genetic algorithms in search, optimization, and machine learning*, Addison-Wesley Longman Publishing Co., Inc., 1989.

66 Q. Gu, Z. Li and J. Han, .

67 L. Scrucca, *J. Stat. Softw.*, 2013, **53**, 1–37.

68 P. S. Gromski, Y. Xu, E. Correa, D. I. Ellis, M. L. Turner and R. Goodacre, *Anal. Chim. Acta*, , DOI:10.1016/j.aca.2014.03.039.

69 M. Barker and W. Rayens, *J. Chemom.*, 2003, **17**, 166–173.

70 J. a. Westerhuis, H. C. J. Hoefsloot, S. Smit, D. J. Vis, A. K. Smilde, E. J. J. Velzen, J. P. M. Duijnhoven and F. a. Dorsten, *Metabolomics*, 2008, **4**, 81–89.

71 S. Wold, M. Sjöström and L. Eriksson, *Chemom. Intell. Lab. Syst.*, 2001, **58**, 109–130.

72 B.-H. Mevik and R. Wehrens, *J. Stat. Softw.*, 2007, **18**, 1–23.

73 P. S. Gromski, H. Muhamadali, D. I. Ellis, Y. Xu, E. Correa, M. L. Turner and R. Goodacre, *Anal. Chim. Acta*, 2015, **879**, 10–23.

74 E. Szyma??ska, E. Saccenti, A. K. Smilde and J. A. Westerhuis, *Metabolomics*,

2012, **8**, 3–16.

75 L. Breiman, *Mach. Learn.*, 2001, **45**, 5–32.

76 W. G. Touw, J. R. Bayjanov, L. Overmars, L. Backus, J. Boekhorst, M. Wels and A. F. T. Sacha van Hijum, *Brief. Bioinform.*, 2013, **14**, 315–326.

77 C. Strobl, A.-L. Boulesteix, A. Zeileis and T. Hothorn, *BMC Bioinformatics*, 2007, **8**, 25.

78 H. Oja, *Multivariate Nonparametric Methods with R: An approach based on spatial signs and ranks*, Springer, 2010.

79 N. W. Lutz, J. V. Sweedler and R. A. Wevers, *Methodologies for Metabolomics: Experimental Strategies and Techniques*, Cambridge University Press, New York, USA, 2013.

80 M. Berg, M. Vanaerschot, A. Jankevics, B. Cuypers, R. Breitling and J.-C. Dujardin, *Comput. Struct. Biotechnol. J.*, 2013, **4**, e201301002.

81 J. Wang, in *Encyclopedia of Systems Biology*, Springer, New York, USA, 2013, p. 1671.

82 H. Hotelling, *Ann. Math. Stat. 2*, 1931, **3**, 360–378.

83 M. L. McHugh, *Biochem. medica*, 2013, **23**, 143–9.

84 K. M. Pierce, J. C. Hoggard, J. L. Hope, P. M. Rainey, A. N. Hoofnagle, R. M. Jack, B. W. Wright and R. E. Synovec, *Anal. Chem.*, 2006, **78**, 5068–5075.

85 R. Leardi and A. Lupiáñez González, *Chemom. Intell. Lab. Syst.*, 1998, **41**, 195–207.

86 C. M. Bishop, *Pattern recognition and machine learning*, Springer, 2006.

87 T. Mehmood, K. H. Liland, L. Snipen and S. Sæbø, *Chemom. Intell. Lab. Syst.*, 2012, **118**, 62–69.

88 J. Dummer, M. Storer, M. Swanney, M. McEwan, A. Scott-Thomas, S. Bhandari,

S. Chambers, R. Dweik and M. Epton, *TrAC - Trends Anal. Chem.*, 2011, **30**, 960–967.

89 P. J. P.J. Mazzone, *Eur. Respir. Monogr.*, 2010, 130–139.

90 a Bajtarevic, C. Ager, M. Pienz, M. Klieber, K. Schwarz, M. Ligor, T. Ligor, W. Filipiak, H. Denz, M. Fiegl, W. Hilbe, W. Weiss, P. Lukas, H. Jamnig, M. Hackl, a Haidenberger, B. Buszewski, W. Miekisch, J. Schubert and a Amann, *BMC Cancer*, 2009, **9**, 348.

91 R. Salerno-Kennedy and K. D. Cashman, *Wien. Klin. Wochenschr.*, 2005, **117**, 180–186.

92 L. Tretter, A. Patocs and C. Chinopoulos, *Biochim. Biophys. Acta - Bioenerg.*, 2016, **1857**, 1086–1101.

93 E. Mills and L. A. J. O’Neill, *Trends Cell Biol.*, 2014, **24**, 313–320.

94 US 8501749 B2, 2013.

95 US8193214 B2, 2012.

96 US6319921 B1, 2000.