Standardization of complex biologically-derived spectrochemical datasets

Camilo L.M. Morais[a,*,1], Maria Paraskevaidi[a,*,1], Li Cui[d], Nigel J Fullwood[b], Martin Isabelle[c], Kássio M.G. Lima[e], Pierre L. Martin-Hirsch[f], Hari Sreedhar[h], Júlio Trevisan[g], Michael J Walsh[h], Dayi Zhang[i], Yong-Guan Zhu[d], Francis L. Martin[a,1]

[a]*School of Pharmacy and Biomedical Sciences, University of Central Lancashire, Preston PR1 2HE, UK;* [b]*Division of Biomedical and Life Sciences, Faculty of Health and Medicine, University of Lancaster, Lancaster LA1 4YQ, UK;* [c]*Spectroscopy Products Division Renishaw plc, New Mills, Wotton-under-Edge, Gloucestershire GL12 8JR, UK;* [d]*Key Lab of Urban Environment and Health, Institute of Urban Environment, Chinese Academy of Sciences, Xiamen 361021, China;* [e]*Institute of Chemistry, Biological Chemistry and Chemometrics, Federal University of Rio Grande do Norte, Natal 59072-970, Brazil;* [f]*Department of Obstetrics and Gynaecology, Lancashire Teaching Hospitals NHS Foundation, Preston PR2 9HT, UK;* [g]*Institute of Astronomy, Geophysics and Atmospheric Sciences, University of São Paulo, Cidade Universitária, R. do Matão, 1226 - Butantã, São Paulo - SP, 05508-090, Brazil;* [h]*Department of Pathology, University of Illinois at Chicago, Chicago, Illinois, USA;* [i]*School of Environment, Tsinghua University, Beijing 100084, China*

[1]To whom correspondence should be addressed: Email: cdlmedeiros-de-morai@uclan.ac.uk; Email: mparaskevaidi@uclan.ac.uk; Email: flmartin@uclan.ac.uk; Tel: +44 (0) 1772 89 6482

*Contributed equally

Abstract

The use of spectroscopy techniques, such as Fourier-transform infrared (FTIR) spectroscopy, has been a successful method to study the interaction of light with biological materials and facilitate novel cell biology analysis. Disease screening and diagnosis, microbiological studies, forensic and environmental investigations make the use of spectrochemical analysis very attractive due to its low cost, minimal sample preparation, non-destructive nature and substantially accurate results. However, there is now an urgent need for repetition and validation of these methods in large-scale studies and across different research groups, which would bring the method closer to clinical and/or industrial, implementation. In order for this to succeed, it is important to eliminate the chance of random spectral alterations caused by inter-individual, inter-instrument and/or inter-laboratory variations. Thus, it is evident that spectral standardization is crucial for the widespread adoption of these spectrochemical technologies. By using calibration transfer procedures, different sources of variations can be normalized into a single model using computational-based methods; therefore, measurements performed under different conditions can eventually generate the same result, eliminating the need for a full recalibration. In this paper, we have constructed a protocol for model standardization using different transfer technologies described for FTIR spectrochemical applications. This is a critical step towards the construction of a practical spectrochemical analysis model for daily routine analysis, where uncertain and random variations are present.

Introduction

Vibrational spectroscopy has shown great promise as an analytical tool for the investigation of numerous sample types with wide applications in diverse sectors, such as biomedicine, pharmaceutics or environmental sciences. Fourier-transform infrared (FTIR) spectroscopy is one of the preferred techniques for identification of biomolecules through the study of their characteristic vibrational movements. Using chemometric approaches, the system is trained to recognize unique spectral features within a sample, so that when unknown samples are introduced an accurate classification is feasible. Alterations in these measurement parameters could interfere with the spectral signature and produce random variations. Therefore, a crucial step is spectral correction, or standardization, which would provide comparable results and allow system transferability. The idea is that non-biological variations, such as those arising from different users, locations or instruments, will no longer affect the classification result; therefore any collected data could be imported into a central database and handled for further exploration or diagnostic purposes. Several groups and companies worldwide are developing spectrochemical approaches for diagnosis, discrimination and monitoring of diseases, as well as for other uses. Combination of multiple datasets would facilitate the conduction of large-scale studies, which are still lacking in the field of biospectroscopy.

Sensor-based technologies

Sensor-based technologies are an integral part of daily life ranging from locating sensor-based technology, such as global positioning system (GPS)[1], to image biosensors, such as X-rays[2-5] and γ-rays[6-8], which are used extensively for medical applications. Other powerful approaches that make use of sensor-based technologies toward medical disease examination and diagnostics include circular dichroism (CD) spectroscopy[9-12], ultraviolet

(UV) or visible spectroscopy[13,14], fluorescence[15-19], nuclear magnetic resonance (NMR) spectroscopy[20-24] and ultrasound (US) [2,25-28].

Over the last two decades, optical biosensors employing vibrational spectroscopy, particularly IR spectroscopy, have seen tremendous progress in biomedical and biological research. A number of studies using the above-mentioned method~~s~~ have focused on cancer investigation with malignancies such as brain[29-32], breast[33-35], oesophagus[36,37], skin[38-42], colorectal[43-45], lung[46-48], ovarian[49-53], endometrial[50,54,55], cervical[56-59] and prostate[60-63] cancer being some of them. Non-cancerous diseases have also been examined, namely neurodegenerative disorders[64-67], HIV/AIDS[68], diabetes[69-71], rheumatoid arthritis[72,73], cardiovascular diseases[74,75], malaria[76-78], alkaptonuria[79], cystic fibrosis[80], thalassemia[81], prenatal disorders[82,83], macular degeneration[84,85], atherosclerosis[75,86] and osteoarthritis[87-89].

Limitations

Spectrochemical approaches are advantageous when compared with traditional molecular methods as they provide a holistic status of the sample under interrogation, thus generating typical spectral regions widely known as “fingerprint regions”. These methods have also been shown to be rapid, inexpensive and non-destructive while they also improve diagnostic performance and eliminate subjective diagnosis (*e.g*., histopathological diagnosis), where inter- and intra-observer variability are present[90]. However, similarly to any other analytical method, vibrational spectroscopy also comes with some limitations. For instance, prior to FTIR studies, optimization of instrumental settings, sample preparation and operation mode also needs to be conducted in order to improve the spectral quality and molecular sensitivity[91-93]. Overall, the above-mentioned barriers can be overcome after careful consideration of the experimental design.

A considerable limitation that is yet under-investigated in the field of spectrochemical techniques is associated with the difficulties entailed in data conformation and system standardization. Currently, there are multiple pilot studies showing promising results but an approach towards standardization for biological applications is lacking. Random variation between studies can originate from differences in instrumentation, operators, and environmental conditions, such as room temperature and humidity.

The main objective of this article is to present a protocol for model standardization, which can be applied in FTIR spectrochemical techniques to rule out the chance of random spectral alterations. Inter-individual, inter-instrument, inter-sample and/or inter-laboratory variations can be a source of unwanted, non-biological alterations, thus leading to incorrect conclusions. However, for a method to become reliable and clinically translatable, it is important that measurements performed under different conditions generate comparable results. The aim of the spectral standardization model presented here is to expedite multi-centre studies with large numbers of samples; this would bring these spectrochemical techniques closer to clinical implementation and facilitate life-changing decisions. We describe a protocol that has four main components: (i) sample preparation, (ii) spectral acquisition, (iii) data pre-processing and (iv) model standardization. The current protocol has an in-depth insight obtained from cross-laboratory collaborations with leading experts in the field. This article offers a step-by-step procedure, which can be implemented by a non-specialist in spectrochemical studies. For further information about instrumental and software options, spectral acquisition steps and data analysis for a range of different analytical systems the reader is directed towards additional protocols[91,94-101].

Applications

Spectrochemical approaches, in combination with computational analysis, have been proven to be effective for biomedical research through facilitating the diagnosis, classification, prognosis, treatment stratification and modulation or monitoring of a disease and treatment. However, these techniques are widely applicable to other fields as well, namely food industry[102-105], toxicology[106-109], microbiology[110-115], forensics[116-120], pharmacy[108,121,122], environmental and plant science[123-125], as well as defence and security[126-128]. Applications of standardization algorithms vary according to the spectral technique and sample matrix studied, where mostly are based on Raman and Fourier-transform near-infrared (FT-NIR) spectroscopy. Table 1 summarizes some standardization applications.

Table 1. Examples of applications involving standardization techniques.

Sample matrix	Spectroscopic technique	Aim	Ref.
Tissue	Raman	Standardization of various perturbations on Raman spectra for diagnosis of breast cancer based on snap frozen tissues	129
	Raman	Standardization of spectra acquired in 3 different sites for analysing oesophageal samples based on snap frozen tissues	130
Cells	Raman	Standardization of spectra acquired with 4 different instruments for classification of three different cultured spore species	131
Biofluids	FT-NIR	Standardization of spectra acquired with 3 different instruments for measuring haematocrit in the blood of grazing cattle	132
	LC-MS	Standardization of spectra acquired with 2 different instruments for mapping rendition times and matching metabolite features of subjects diagnosed with small cell lung cancer based on blood serum and plasma samples analysis	133
Pharmaceutical materials	Raman	Standardization of spectra acquired with 5 different instruments for analysing various pharmaceutical excipients, active pharmaceutical ingredients (APIs) and common contaminants	134
	FT-NIR	Standardization of spectra acquired with 2 different instruments for simultaneous determination of rifampicin and isoniazid in pharmaceutical formulations	135
	FT-NIR	Standardization of spectra acquired with 2 different instruments for predicting content of 654 pharmaceutical tablets	136
Food	FT-NIR	Standardization of spectra acquired with 3 different instruments for predicting parameters in corn samples	136
			137
	FT-NIR	Standardization of spectra acquired with 2 different instruments for predicting vitamin C in navel orange	138
	FT-NIR	Standardization of spectra recorded in 4 different labs for determining moisture, proteins and oil content in soy seeds	139
	FT-NIR	Standardization of spectra acquired by a benchtop and portable instrument for determining total soluble solid contents in single grape berry	140
	UV-Vis	Standardization of visible spectra acquired with 3 different instruments for measuring pH of Sala mango	141
Plant	FT-NIR	Standardization of spectra acquired with 2 different instruments for predicting baicalin contents in radix scutellariae samples	137
	FT-NIR	Standardization of spectra acquired by 2 different instruments and in three physical states (powder, filament and intact leaf) for determining total sugars, reducing sugars and nicotine in tobacco leaf samples	142
	NMR	Standardization of spectra acquired with 3 different instruments for authenticity control of sunflower lecithin	143
Cosmetic	CD spectroscopy	Standardization of spectra acquired between standard and real-world samples for determining Pb^{2+} in cosmetic samples	144
Inorganic substances	FT-IR	Standardization of interferogram spectra acquired with 2 instruments for classifying acetone and SF_6 samples	145
Fuel	FT-IR	Standardization of spectra acquired with 2 different instruments for predicting density of crude oil samples	146

Model transferability

Transferability models have been previously developed, however this is still an under-investigated field, especially for biomedical applications. An inclusive standardization protocol that could be implemented in a range of different spectrochemical approaches is of great need. Differences are present even between identical instruments; for instance, changes in signal intensity caused by replacement, alignment or ageing of optical and spectrometer components, natural variations in optics and detectors construction, changes in measurement conditions (temperature and humidity), changes in physical constitution of the sample (particle size and surface texture) and operator discrepancies could all lead to wavenumber shifts and artefacts in the spectra. In all of these cases, prediction errors can become very large, especially when the whole spectrum is used in the model. Standardization techniques aim to generate a uniform spectral response under differing conditions, ensuring the interchangeability of results obtained in different situations, without having to perform a full calibration for each situation.

Previous standardization methods include the use of simple slope and bias correction[147,148], direct standardization (DS)[149-153], piecewise direct standardization (PDS)[147,154-156], piecewise linear discriminant analysis (PLDA)[145], guided model reoptimization (GMR)[156], back-propagation neural network (BNN)[145], generalized least squares weighting (GLSW)[157], model updating (MU)[158,159], orthogonal signal correction (OSC)[160,161], orthogonal projections to latent structures (OPLS)[146], wavelet hybrid direct standardization (WHDS)[155], maximum likelihood PCA (MLPCA)[162], Shenk and Westerhaus method (SW)[163,164], positive matrix factorization (PMF)[165,166], artificial neural networks (ANN) drift correction[167], transfer *via* extreme learning machine auto-encoder method (TEAM)[168], calibration transfer based on the maximum margin criterion (CTMMC)[169], calibration transfer based on canonical correlation analysis (CTCCA)[170] and calibration methods, such as wavenumber offset correction, instrument response correction and baseline correction[130].

Direct standardization. DS is one of the most used methods for data standardization. It was initially proposed to correct relatively large spectral differences between data collected by two instruments[147]. In DS, the entire spectrum from a new secondary response (*e.g.*, a different instrument) is transformed to resemble the spectrum from the primary source (*e.g.*, original instrument)[149]. This is performed based on a linear relationship between the data acquired under different circumstances[158]:

$$\mathbf{S}_1 = \mathbf{S}_2\mathbf{F} \qquad (01)$$

where $\mathbf{S}_1$ represents the data acquired for the primary response; $\mathbf{S}_2$ represents the data acquired for the secondary response; and **F** is the transformation matrix that maintains the relationship between $\mathbf{S}_1$ and $\mathbf{S}_2$.

The transformation matrix **F** is estimated in a least-squares sense by[171]:

$$\mathbf{F} = \mathbf{S}_2^{+}\mathbf{S}_1 \qquad (02)$$

where $\mathbf{S}_2^{+}$ is the pseudo-inverse of $\mathbf{S}_2$, calculated by:

$$\mathbf{S}_2^{+} = (\mathbf{S}_2^{\mathrm{T}}\mathbf{S}_2)^{-1}\mathbf{S}_2^{\mathrm{T}} \qquad (03)$$

in which T stands for the matrix transpose operation.

Then, when samples are measured under the secondary system, the signals generated **X** are transformed to resemble the primary system response by[158]:

$$\widehat{\mathbf{X}}^{\mathrm{T}} = \mathbf{X}^{\mathrm{T}}\mathbf{F} \qquad (04)$$

where $\widehat{\mathbf{X}}$ is the standardized response for **X**.

Problems related to different background information between instruments can affect the standardization procedure. To correct for this, the standardization process is usually adapted with the background correction method[171], in which the transformation matrix described in Eq. 02 is calculated with a background correction factor ($\mathbf{F}_\mathrm{b}$) and an additive background correction vector $\mathbf{b}_\mathrm{s}$ as follows:

$$\mathbf{S}_1 = \mathbf{S}_2\mathbf{F}_\mathrm{b} + \mathbf{1}\mathbf{b}_\mathrm{s}^\mathrm{T} \tag{05}$$

where $\mathbf{1}$ is an all-ones vector and $\mathbf{b}_\mathrm{s}$ is obtained by:

$$\mathbf{b}_\mathrm{s} = \mathbf{s}_{1\mathrm{m}} - \mathbf{F}_\mathrm{b}^\mathrm{T}\mathbf{s}_{2\mathrm{m}} \tag{06}$$

in which $\mathbf{s}_{1\mathrm{m}}$ is the mean vector of $\mathbf{S}_1$ and $\mathbf{s}_{2\mathrm{m}}$ is the mean vector of $\mathbf{S}_2$.

One of the key steps for DS is the selection of the number of samples to transfer (called “transfer samples”). These are samples from the primary system ($\mathbf{S}_1$) that will be used to transform the signal obtained using the secondary system ($\mathbf{S}_2$). Usually, the procedure for selecting transfer samples is based on sample selection techniques, such as Kennard-Stone (KS) algorithm[172] or leverage[147]. Subsequently, the number of transfer samples is evaluated using a validation set through an arbitrary cost function. For quantification applications, a common cost function is the root-mean-square error of prediction, while for classification one can use the misclassification rate.

A disadvantage of DS is that each transformed variable is calculated using the whole spectrum, which carries a high risk of overfitting. The estimation of $\mathbf{F}$ in Eq. (02) is a ill-conditioned problem, because the number of variables may be much larger than the number of standard samples.

Piecewise direct standardization. PDS is another standardization procedure commonly employed for system transferability. It is based on DS, however it uses windows (*e.g.*, wavenumber portions) to make the standardization process more suitable for smaller regions of the data. When compared to DS, PDS is calculated by using the transformation matrix **F** with most of its off-diagonal elements set to zero[147]. With this, PDS fits minor spectral modifications not covered by DS. PDS is the technique of preference for correcting smaller spectral variations, such as small wavelengths shift, intensity variations, and bands enlargement and reduction[147]. In addition, an advantage of PDS compared to DS is that the local rank of each window will be smaller than the rank of the whole data matrix, which means that the number of standard samples can be smaller, and indeed good results have been obtained with very few samples.

One disadvantage of PDS is the need of an additional optimization process, because in addition to the number of transfer samples, PDS also needs a window size optimization, which might lead to a risk of overfitting. Herein, the window size optimization is made using a cost function expressed as the misclassification rate calculated for each window size tested, being evaluated using a validation set where the window with smaller misclassification is selected for final model construction.

Experimental Design

A specified number of steps are required for a study using vibrational spectroscopy, starting from careful experimental design, protocol optimisation and development of experimental procedure document, sample collection and preparation, spectral collection, pre-processing of the derived information and lastly the use of chemometrics for exploratory, classification and standardization purposes. FTIR spectroscopy is described in more detail in this study, however, the standardization protocol described here can be adapted to a range of techniques, including attenuated total reflection (ATR-FTIR), transmission and transflection FTIR, near-IR (NIR), UV-visible, NMR spectroscopy and MS. Nevertheless, intrinsic features of each technique should be taken into consideration before standardization and the protocol may change depending on the application of interest.

A number of biological samples can be analyzed with the above-mentioned analytical methods such as tissues, cytological materials or biological fluids. Sample type and preparation may differ depending on the technique that is employed each time. For instance, IR spectroscopy is limited by water interference at the fingerprint region that can mask the signal of the analyte close to the water peak. This could be addressed with an extra step of sample drying, in contrast to Raman spectroscopy, for example, where water does not generate signal in this region.

Typical steps for sample preparation, acquisition of spectra and data pre-processing are briefly presented here. However, the main focus of this protocol is placed on the calibration transfer and standardization procedures. Readers are directed to additional literature for more detailed information regarding sample format and preparation, suitability of substrates, instrumentation settings or available software packages (Table 2) and manufacturers[91,94-96,101,173-176].

Table 2. Software packages for data standardization.

Software	Website	Description	Availability
PLS_Toolbox	http://www.eigenvector.com/	MATLAB toolbox for chemometric analysis. Contains standardization routines using DS, PDS, double window PDS, spectral subspace transformation, GLSW, OSC, and alignment of matrices.	Commercial
Unscrambler® X	http://www.camo.com/	Software for multivariate data analysis and design of experiments. Contains standardization routines using interpolation, bias and slope correction, and PDS.	Commercial
OPUS	https://www.bruker.com/	Spectral acquisition software with data processing features. Contains a standardization routine using PDS.	Commercial
Pirouette®	https://infometrix.com/	Chemometrics modelling software. Contains standardization routines using DS and PDS.	Commercial

Experimental design: sampling

Sample preparation. Biological samples have been studied extensively with spectrochemical techniques for disease research. Tissue specimens can be analysed fresh, snap-frozen or formalin-fixed, paraffin-embedded (FFPE). Fresh or snap-frozen histology sections are preferable as they are devoid of contaminants whereas FFPE treatment contributes to characteristic peaks, hindering the biological information. FFPE tissues can be deparaffinized either by chemical methods (*e.g.*, incubation in xylene, hexane or Histo-Clear solutions)[91], which can alter tissue structures and be inefficient for the complete wax removal[177], or by applying chemometrics (*e.g.*, digital dewaxing)[178,179], which keeps the tissue intact but might introduce artefacts due to over- or under-estimation of the wax contribution[177].

Fixatives, such as ethanol, methanol or formalin, are often used for the preservation of cytological material, also generating strong peaks and interfering with the spectra; thus, a washing step is crucial before spectroscopic interrogation. Fixation in tissue or cells for preservation purposes generates protein cross-linking which can cause changes in the spectra, especially on the Amide I peak[180]. Alternatively, cells can be studied live after washing from residual medium.

Preparation and pre-treatment of biological fluids depend on the sample type. Some of the biofluids that have been previously used in spectroscopic studies include blood (whole blood, plasma or serum), urine, sputum, saliva, tears, cerebrospinal fluid (CSF), synovial fluid, ascitic fluid or amniotic fluid[181-183]. A centrifugation step should precede in cases where the cells present in these fluids are not the focus of the study; the supernatant could then be kept for further analysis. In blood-based studies, the user should also consider the anticoagulant of preference (*e.g.*, EDTA, citrate or heparin) as it could generate unwanted spectral peaks[184-186]. Careful planning of experiments as well as consistence throughout a study are of great importance for the generation of robust results. Samples should be very stable, since the spectral differences between the data collected under different situations (*e.g.*, different instruments or temperature) should be directly related to the difference between the systems and not a change caused by chemical or physical degradation of the samples. Optimal sample thickness, suitability of substrates and sample formats can differ from one analytical technique to another and thus the user should decide and tailor these according to the study's objective. Another consideration is the number of freeze-thaw cycles and long-term storage as these could compromise the integrity of the samples[184,187]. Preferably, FFPE tissue samples should be analysed after thorough dewaxing and freeze-thaw cycles or long-term storage avoided since these could result in many confounding factors for analysis.

Spectral acquisition. Depending on the study's objective, FTIR spectral information can be collected using either point spectra or imaging. FTIR spectra can be collected in different operational modes, namely ATR-FTIR, transmission or transflection. Instrument parameters such as resolution, aperture size, interferometer mirror velocity and co-additions have to be optimised before acquisition of spectra to achieve high SNR[91,94]. Metal surfaces can also be used to increase the IR signal in a technique known as surface-enhanced IR absorption (SEIRA)[188,189]. As water interference can mask biological information in IR spectra, the user can purge the spectrometer with dry air or nitrogen gas to reduce the instrument internal humidity, or use computational analysis to remove the water signature. In addition, samples should be dried until all water content evaporates; however, drying of a sample is not without consequences, since chemical changes may occur such as loss of volatile compounds. A background sample is collected regularly to account for any changes in the atmospheric or instrument conditions.

For analysing homogenous samples (*e.g.*, biofluids), measurements can be performed by acquiring spectra on different regions of the centre of a drop and across its borders. In transmission measurements, the sample can be measured raw or diluted. Usually, 10 spectra are collected per sample. A higher number of spectral replicas can be performed to decrease the standard-deviation (SD) between measurements, since the SD is proportion to $1/\sqrt{n}$, where n is the number of replicas. For heterogeneously distributed samples (*e.g.*, tissues), spectra should be acquired covering the sample surface as much uniformly as possible, to ensure that all sources of information is contemplated in the spectral data. Samples replicas are also recommended at least as triplicates. For precision estimation, at least six replicates at three levels should be performed. The minimum number of samples for analysis can be estimated using a power test at an 80% power[190]. Further details regarding sampling methodologies for analysing biological materials using FTIR spectroscopy can be found in our previous protocols[91,94].

Experimental design: data quality evaluation

Before processing, the data can be assessed to identify presence of anomalous behaviours or biased patterns. This can be made initially by visual inspection (e.g., identification of very anomalous spectra) followed by Hotelling T^2 *versus* Q residuals charts using only the mean-centred spectra. PCA residuals[191] can explored to identify biased patterns, in which heteroscedastic distributions are signs of biased experimental measurements; while homoscedastic distributions are associated with good sampling. Also, mistakes performed during experimental data acquisition can be evaluated by R^2 values. Negative R^2 indicates that the sample variance is smaller than the model residuals variance, which should not happen. SNR can be estimated by dividing the power (P) of signal by the power of noise, that is $\mathrm{SNR} = P_{signal}/P_{noise} = \left(A_{signal}/A_{noise}\right)^2$, where A is the amplitude; or by the inverse of the coefficient of variation, when only non-negative variables are measured. Collinearity can be evaluated by calculation of condition number, which is naturally high for spectral data (high collinearity).

Experimental design: pre-processing

Data pre-processing is employed for maximizing the SNR. This process is fundamental for correcting physical interfering, such as light scattering, different sample thickness, different optical paths and instrumental noise. Therefore, the pre-processing step has fundamental importance to highlight the signal of interest and reduce interfering.

For standardization applications, the pre-processing step is also important for reducing differences between the different systems that are used. Before any additional pre-processing, the biofingerprint region should be truncated (*e.g.*, 900-1800 cm^{-1}) before analysis. This region contains the main absorptions from biochemical compounds and it suffers minor effects of environmental variability, such as air humidity (free νO-H = 3650–3600 cm^{-1}, hydrogen-bonded νO-H = 3400 – 3300 cm^{-1}) and air CO_2 ($\nu_s CO_2$ = 2350 cm^{-1})[192]. Table **3** summarizes the main pre-processing techniques for correcting noise in biologically-derived datasets.

Table 3. Main pre-processing used for biologically-derived datasets.

Pre-processing	Interfering	Technique	Advantage	Disadvantage	Optimization
Savitzky-Golay smoothing[193]	Instrumental noise	ATR-FTIR, FTIR, NIR, Raman, NMR, UV-Vis	Corrects spectral noise without changing the shape of data significantly	The polynomial order and window size for polynomial fit affects the result	The polynomial function should have an order similar to the spectral data (*e.g.*, 2nd order polynomial function for IR data) and the window size should be an odd number and not too small (keeping the noise) or too large (changing the spectral shape)
Multiplicative scatter correction (MSC)[194]	Light scattering (Mie scattering), different pressure over the sample when using ATR or probe, different lengths of optical path	ATR-FTIR, FTIR, NIR, Raman, NMR, UV-Vis	Corrects light scattering maintaining the same spectral shape and signal scale	Need of a reference spectrum representative of all measurements	The reference spectrum is regularly set as the average spectrum across all training samples
Standard normal variate (SNV)[195]	Light scattering (Mie scattering), different pressure over the sample when using ATR or probe, different lengths of optical path	ATR-FTIR, FTIR, NIR, Raman, NMR, UV-Vis~~,~~ ~~fluorescence EEM~~	Corrects light scattering maintaining the same spectral shape	Creates negative signals since the data are centralized to zero (y-scale)	--
Spectral differentiation[193]	Light scattering (Mie scattering), different pressure over the sample when using ATR or probe, different lengths of optical path, background absorption interfering	ATR-FTIR, FTIR, NIR, Raman, NMR, UV-Vis	Corrects light scattering and baseline problems; highlights smaller spectral differences	Changes the signal scale, shifts the data and increases noise	The order of the derivative function should be used carefully to avoid increased noise (usually 1st or 2nd order differentiation is preferred). The differentiation can be coupled to Savitzky-Golay smoothing
Baseline correction[196]	Background absorption interfering	ATR-FTIR, FTIR, NIR, Raman, NMR, UV-Vis, MS	Corrects the baseline maintaining the same spectral shape	--	There are many methods for baseline correction (*e.g.*, rubber band, automatic weighted least squares, Whittaker filter). The method chosen should be maintained consistent for all systems used
Normalization[90]	Different sample thickness and concentration	ATR-FTIR, FTIR, Raman	Avoids influence of non-desired signals among the	The normalization might hide signal differences	--

		samples	between samples at important bands, such as Amide I and Amide II; and also may introduce non-linearities

Figure 1 depicts the effect of a pre-processing approach employed for a blood plasma dataset acquired under different experimental conditions (*i.e.*, different systems and operators). In this Figure, the reduction of the spectral differences between the systems is evident after data pre-processing (Savitzky-Golay smoothing, MSC, baseline correction and normalization).

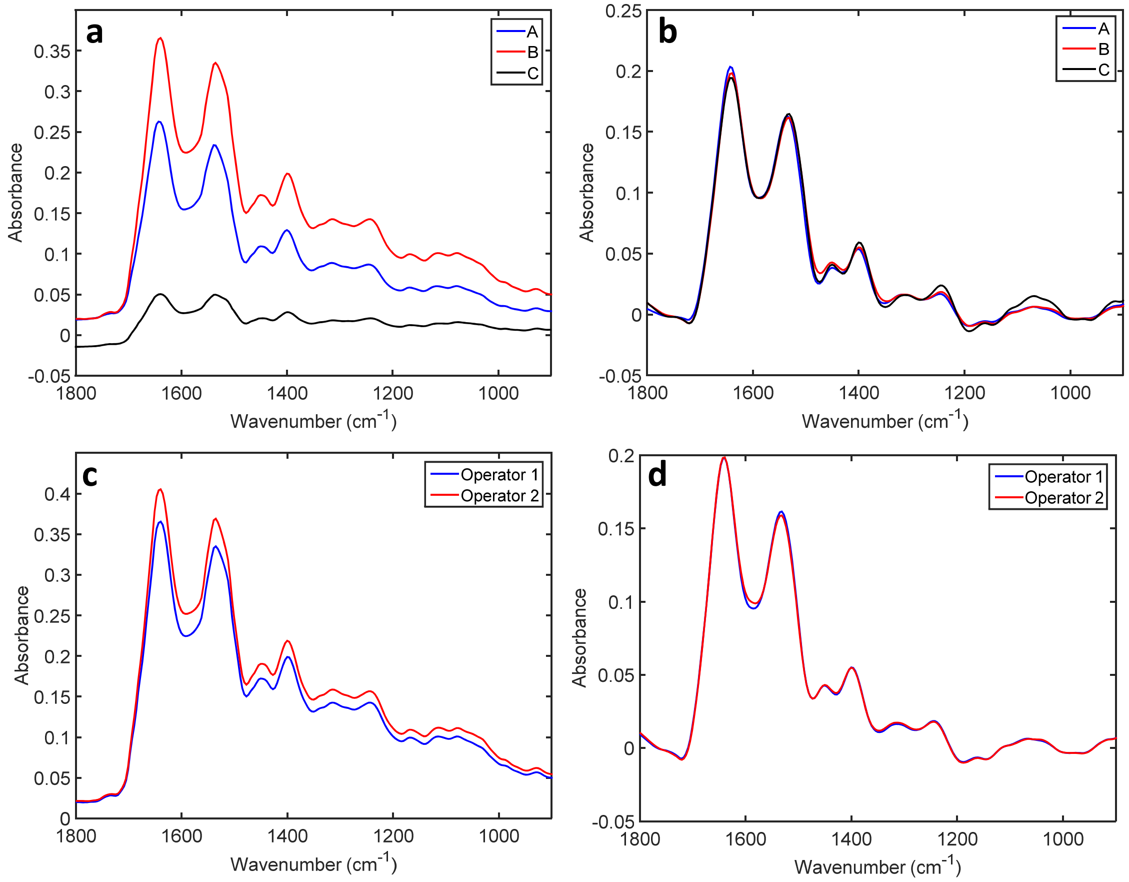


Figure 1. Average (a) raw and (b) pre-processed IR spectra for healthy control samples across three different systems (A, B and C). Average (c) raw and (d) pre-processed IR spectra for healthy control samples across two different operators (Operator 1 and 2).

After the pre-processing techniques displayed in Table 3, scaling should be employed as most classification methods require all the variables (*e.g.*, wavenumbers) in the dataset to be at the same scale in order to work properly.

For spectral data, mean-centring (also referred as “standardization” by Hastie et al.[197]) is a very reasonable approach, after which all variables in the dataset will have zero mean. When data contain values represented by different scales (*e.g.*, after data fusion using both IR and Raman spectra), block-scaling should be used, where each block of data would have the same sum-of-squares (normally after mean-centring).

Another important aspect of pre-processing is the order in which each step is applied. Pre-processing should be employed in a logical order so that the next pre-processing step is not affected by the previous one. For example, pure spectral differentiation cannot be employed before smoothing, since the spectral differentiation will increase the original noise. Therefore, smoothing should be applied before differentiation. Albeit, Savitzky-Golay routine incorporates smoothing and spectral differentiation so, in practical terms, these can be performed together. To summarise, the suggested order of pre-processing is as follows:

1. Spectral Truncation
2. Smoothing
3. Light scattering correction
4. Baseline correction
5. Normalization
6. Scaling

When using different instruments but same type of sample, the pre-processing steps should be the same for the data acquired under different circumstances.

Experimental design: data analysis

Sample splitting. Sample splitting is fundamental for constructing a predictive chemometric model. The splitting procedure can be performed manually or by computer-based methodologies. Manual splitting can generated biased results, therefore computational-based split is more recommended. In this case, some strategies includes random selection, leverage[147] or the KS algorithm[172]. KS works based on Euclidian distance calculation by firstly assigning the sample with the maximum distance to all other samples to the calibration set, and then by selecting the samples which are as far away as possible from the selected samples to this set, until the designed number of selected samples is reached. This ensures that the calibration model will contain samples that uniformly cover the complete sample space, where no or minimal extrapolation of the remaining samples are necessary; avoiding problems of manual or random selection, such as non-reproducibility and non-representative selection. Usually, the dataset is split with 70% of the samples assigned for training, 15% for validation and 15% for test. In this case, the test set is dependent on the initial group of samples measured, and it is not a regular independent test set where a new set of similar samples are measured.

Exploratory analysis. Exploratory analysis is an important tool to provide an initial assessment of the data. Using exploratory analysis, the analyst can see the clustering patterns and then draw conclusions related to the nature of samples, outliers and experimental errors. One of the most common techniques for exploratory analysis is principal component analysis (PCA), in which the original data are decomposed into a few principal components (PCs) responsible for most of the variance within the original dataset. The PCs are orthogonal to each other and are generated in a decreasing order of explained variance, so that the first PC represents most of the original data variance, followed by the second PC and so on[198]. Mathematically the decomposition takes the form:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \tag{07}$$

where $\mathbf{X}$ represents the pre-processed data (*e.g.*, pre-processed samples' spectra); $\mathbf{T}$ are the scores; $\mathbf{P}$ are the loadings; and $\mathbf{E}$ are the residuals.

The PCA scores represent the variance in the sample direction and they are used to assess similarities/dissimilarities among the samples, thus detecting clustering patterns. The PCA loadings represent the variance in the variable (*e.g.*, wavenumber) direction and they are used to detect which variables show the highest importance for the pattern observed on the scores. The PCA loadings are commonly employed as a tool for searching spectral markers that distinguish different biological classes[199]. The PCA residuals represent the difference between the decomposed and original data and can be used to identify experimental errors. Ideally, the PCA residuals should be random and close to zero, representing a heteroscedastic distribution. Otherwise, they can indicate experimental bias according to a homoscedastic distribution.

For standardization applications, PCA is a fast, intuitive and reliable tool to observe if there are differences between the spectra acquired by different systems. Ideally, if the same sample is measured under different conditions (different laboratories, instrument manufacturers or user operators) their PCA scores should be random and completely superposed. If a discrimination pattern is observed on the PCA scores, then it is indicative that the data need standardization. Figure 2 illustrates a PCA scores plot from the same samples (blood plasma of healthy controls) measured using three IR instruments before (Fig. 2a) and after (Fig. 2b) DS. Even though the samples in Fig. 2a are pre-processed, three different clusters are still evident. After DS the samples measured using different systems are normalized into a single cluster.

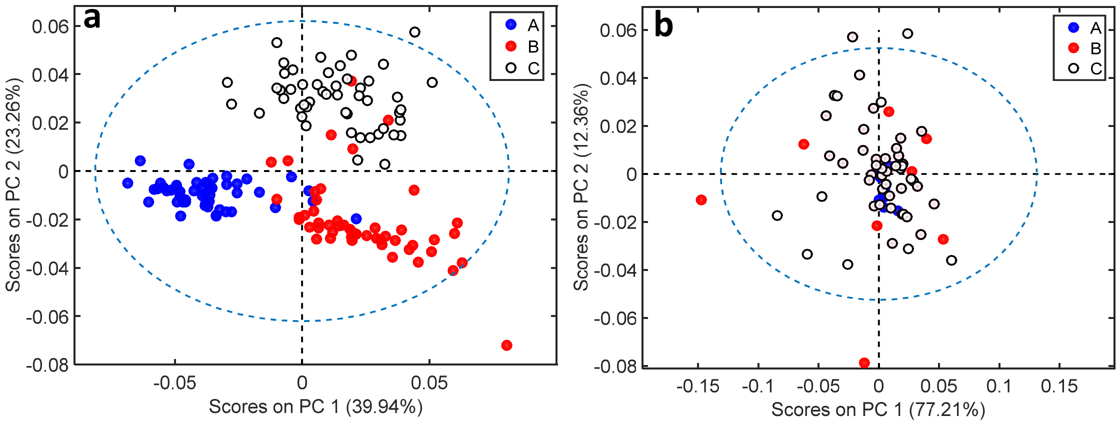


Figure 2. (a) PCA scores for healthy control samples across three different instruments (A, B and C) after pre-processing but before DS; (b) PCA scores for healthy control samples across three different instruments (A, B and C) after DS. The dotted blue circle shows 95 % confidence ellipse (two-sided).

Outlier detection. Outlier detection is important to prevent samples, which differ from the original dataset, from affecting the results using predictive models. Outliers can be attributed to experimental errors, such as inconsistent sample preparation or spectral acquisition, or to larger experimental noise, such as Johnson noise, shot noise, flicker noise and environmental noise. These samples can have large leverage for classification, masking the real signal from the samples of interest; therefore, it is advised that they be removed from the dataset used to train the predictive model.

To detect outliers, techniques such as Jack-knife[200], Z-score[201] or *K*-modes clustering[202] can be utilised among others[203]. One of the most popular and visually intuitive technique for detecting outliers is the Hotelling T^2 vs Q residual test[204]. In this test, a chart is created using the Hotelling T^2 values in x-axis and the Q residuals in the y-axis, generating a scatter plot. The Hotelling T^2 represents the sum of the normalized squared scores, which is the distance from the multivariate mean to the projection of the sample onto the PCs[205]. The Q residuals represent the sum of squares of each sample in the error matrix, thus measuring the residues between a sample and its projection onto the PCs[205]. All samples far from the origin of this graph are considered outliers and should be removed one at a time, as the PCA is highly influenced by the samples that are included in the model. Samples with high values in both Hotelling T^2 and Q residuals are the worst outliers; while samples with high values in only one of these axis are the second worst outliers. Supplementary Method 1 illustrates an example for outlier detection. Squared confidence limits can be draw based on this graph; however, this can hinder outlier detection. For example, in squared confidence limits at a 95% level, certain amount of data-points (5%) are set outside these limits.

Classification. Classification techniques are employed for sample discrimination. Using chemometric analysis, one can distinguish classes of samples based on their spectral features and then make further predictions based on these. The prediction capability of a classification model should be evaluated with external samples (unknown samples) through the calculation of figures of merit, including accuracy (proportion of samples correctly classified considering true positives and true negatives), sensitivity (proportion of positives that are correctly identified) and specificity (proportion of negatives that are correctly identified)[206].

There are many types of classification techniques for spectral data. Table 4 summarizes the main classification techniques employed for biospectroscopy applications, along with their advantages and disadvantages.

Table 4. Classification techniques.

Classification Technique	Advantage	Disadvantage
Linear discriminant analysis (LDA)[207]	Simplicity, fast calculation	Needs data reduction, does not account for classes having different variance structures, greatly affected by classes having different sizes
Quadratic discriminant analysis (QDA)[207]	Fast calculation, accounts for classes having different variance structures, not much affected by classes having different sizes	Needs data reduction, higher risk of overfitting
Partial least squares discriminant analysis (PLS-DA)[208]	Fast calculation, high accuracy	Greatly affected by classes having different sizes, needs optimization of the number of latent variables (LVs)
K-Nearest Neighbours (KNN)[209]	Simplicity, non-parametric, suitable for large datasets	Time consuming, needs optimization of the distance calculation method and *k* value, highly sensitive to the “curse of dimensionality”[197]
Support vector machines (SVM)[210]	Non-linear classification nature, high accuracy	High complexity, high risk of overfitting, needs optimization of kernel function and SVM parameters, time consuming
Artificial neural networks (ANN)[211]	Non-linear classification nature, ability to work with incomplete knowledge, high accuracy	High computational cost, needs optimization of the number of neurons and layers, no interpretability (“black box” model)
Random forests[212]	Non-linear classification nature, high accuracy, relatively low computational cost	High risk of overfitting, needs optimization of the number of trees, no interpretability (“black box” model)
Deep learning approaches[213]	Non-linear classification nature, native feature extraction (e.g., in convolutional neural networks (CNN)), local spatial coherence (CNN), high accuracy	High computational cost, needs hyperparameter optimization, needs large datasets, time consuming, no interpretability (“black box” model)

When employing classification techniques, one must follow a parsimony order[214], where the simplest algorithms should be used first, reducing the need for more complex algorithms which would require more optimization steps. An order for using these classification algorithms is: LDA>PLS-DA>QDA>KNN>SVM>ANN>Random forests>Deep learning approaches, from the simplest to the most complex.

Classification algorithms can be coupled to feature extraction and feature selection techniques in order to reduce data collinearity/redundancy, thus reducing the risk of overfitting in the classifier training, and speeding up such training, as there are less variables involved. An additional benefit of such a feature extraction/selection step is to provide spectral markers identification as a “side-effect” (depending on the feature extraction/selection method applied). For feature extraction, the most popular technique is PCA. In this case, a PCA is firstly applied to the data, and then the PCA scores are used as the input variables (instead of the wavenumbers data points) for the classification techniques mentioned above[215]. PLS-DA is also a feature extraction technique[208], and normally it performs better than a PCA followed by LDA, as the scores from a PCA does not necessarily describe the difference between the samples, but rather the variance in the data. In PLS-DA, a partial least squares (PLS) model is applied to the data in an interactive process reducing the original variables to a few number of LVs, where a LDA is used for classifying the groups[216]. Other discriminant classifiers, in particular QDA, also could be used in this classification step to circumvent problems observed with LDA. For feature selection, there are many techniques commonly employed in biological datasets, including genetic algorithm (GA)[217] and successive projections algorithm (SPA)[218]. The variables (*e.g.*, wavenumbers) selected by these techniques are used as input variables for the classification models described in Table 2. An important advantage of GA is its relatively low-computational cost compared to SPA and reduction of data collinearity. Furthermore, GA-based techniques are intuitive and simple to understand in the algorithmic sense but they also have a non-deterministic nature and require optimization of many parameters. SPA’s advantage relies on its deterministic nature, minor parameter optimization and reduction of data collinearity, however, it is very time consuming. For hyperspectral imaging, feature selection also can be performed by Minimum Redundancy Maximum Relevance (mRMR) algorithm[219], where the selection process is based on maximizing the relevance of extracted features and simultaneously minimize redundancy between them.

Standardization. Data standardization should be employed when a primary classification model is built and new data comes to be predicted from a secondary system (different laboratory or instrument manufacturers), or when there is a change in instrument components (*e.g.*, laser, gratings, etc.) or when the data of the chemometric model are acquired under different circumstances (different analysts, days, instrumental settings, etc.). As previously mentioned, the most common and reliable methods for data standardization are the DS and PDS algorithms. These methods can be found in a few software packages (described in Table 3).

Figure 3 summarises the standardization protocol using DS applied to spectra acquired under different conditions. The first step consists of applying KS algorithm for selecting the number of transfer samples from the primary system as well as the number of training samples for the secondary systems, which is ideally 70% of the dataset. Thereafter, the DS transform generation algorithm is employed to estimate the transform matrix. The validation set of the secondary system is then used with the classification model of the primary system to evaluate the optimum number of transfer samples. This optimization step is repeated depending on the number of transfer samples from the primary system. After this number is defined, the validation set of the secondary system is finally standardized and the final classification model is subsequently applied. This procedure is realized with a certain number of samples measured in all instruments being standardized. This procedure should be realized in as similar manner as possible to reduce spectral differences. After the model is standardized and proper validated, new external samples can be measured in any of the instruments and predicted by the standardized classification model.

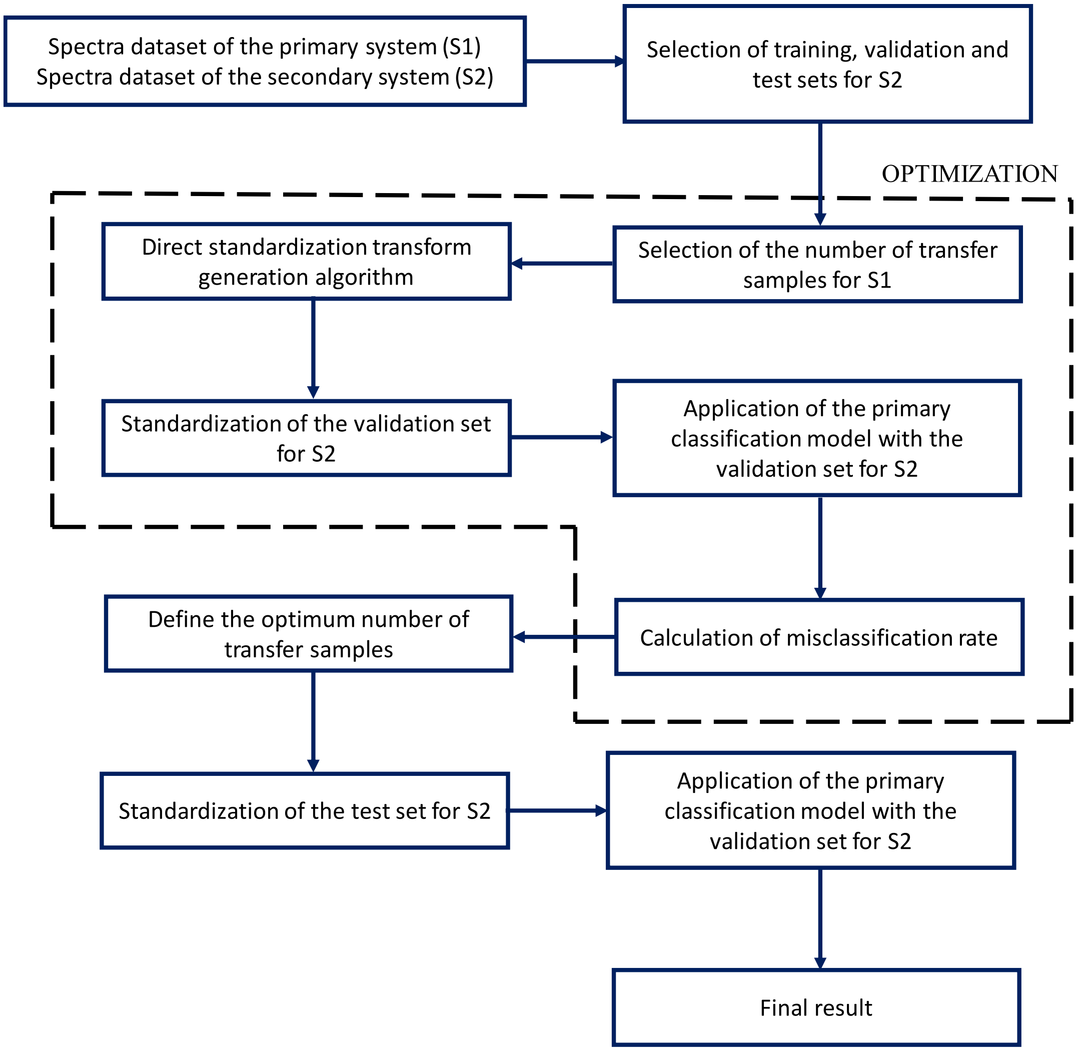

Spectra dataset of the primary system (S1)
Spectra dataset of the secondary system (S2)
Selection of training, validation and test sets for S2
OPTIMIZATION
Direct standardization transform generation algorithm
Selection of the number of transfer samples for S1
Standardization of the validation set for S2
Application of the primary classification model with the validation set for S2
Define the optimum number of transfer samples
Calculation of misclassification rate
Standardization of the test set for S2
Application of the primary classification model with the validation set for S2
Final result

Figure 3. Flowchart for standardization using Direct Standardization (DS).

For PDS, an extra step is added after defining the number of transfer samples to estimate the optimum window size. The dashed region in Fig. 3 is repeated according to the window size.

For multi-laboratory studies the flowchart depicted in Fig. 4 illustrates how the standardization protocol should be employed.

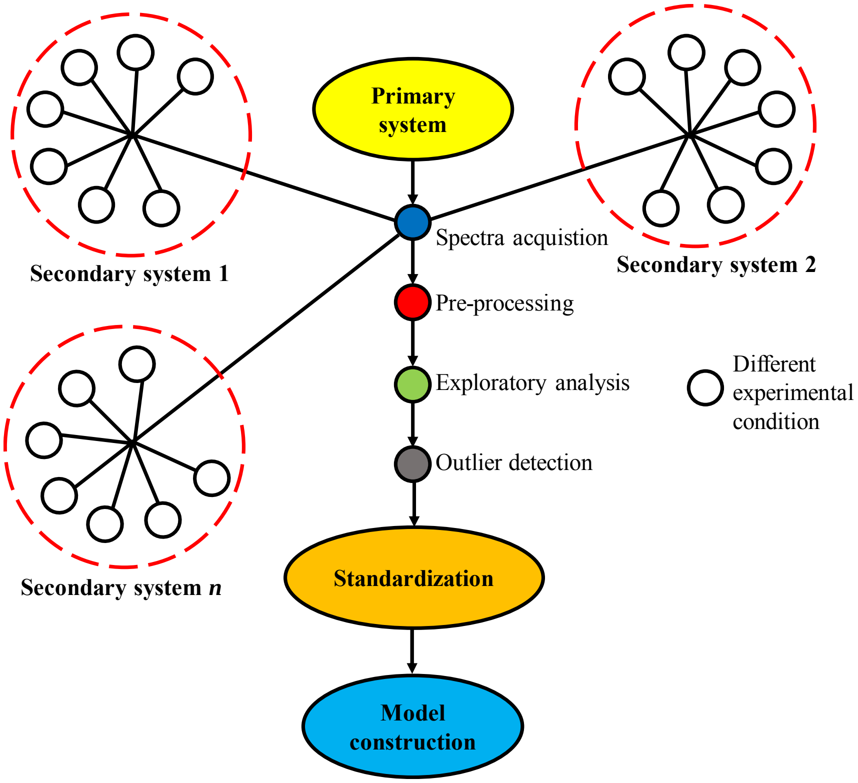


Figure 4. Flowchart for a standardization protocol using different experimental conditions.

In Fig. 4, spectra acquired under different experimental conditions are used for a global standardization model. A primary system should be designated and then all spectra from secondary systems are equally pre-processed, followed by an exploratory analysis to assess samples' similarities/dissimilarities, outlier detection, standardization by the method depicted in Figure 3; the final model construction follows last. With this, all sources of variations present in different systems can be included into a general chemometric model.

MATERIALS

REAGENTS

- Biological samples (tissue, cells, biofluids).

▲ **CRITICAL** Human samples should be collected with appropriate local institutional review board for ethical approval and adhere to the Declaration of Helsinki principles. Similarly, for studies involving animals, all experiments should be performed in accordance with relevant guidelines and regulations. Ethical approval has to be obtained before any sample collection.

- Optimal cutting temperature (OCT) compound (Agar Scientific, cat. no. AGR1180)
- Liquid nitrogen (BOC, CAS no. 7727-37-9) **! CAUTION** Asphyxiation hazard; make sure room is well ventilated. Causes burns; wear face shield, gloves and protective clothing.
- Paraplast Plus paraffin wax (Thermo Fisher Scientific, cat. no. SKU502004)
- Isopentane (Fisher Scientific, cat. no. P/1030/08) **! CAUTION** Extremely flammable, irritant, aspiration hazard and toxic; use in a fume hood.
- Distilled water
- PBS (10×; MP Biomedicals, cat. no. 0919610)
- Virkon (Antec, DuPont, cat. no. A00960632)
- Trypsin–EDTA (0.05%, Sigma-Aldrich, Thermo Fisher Scientific cat. no. 25300054)

Anticoagulants

- EDTA (Thermo Fisher Scientific, BD Vacutainer, cat. no. 02-687-107)
- Sodium citrate (Thermo Fisher Scientific, BD Vacutainer)
- Lithium/sodium heparin (Thermo Fisher Scientific, BD Vacutainer)

Fixative and preservative agents

- Formalin, 10% (vol/vol; Sigma-Aldrich, cat. no. HT501128) **! CAUTION** Potential carcinogen, irritant and allergenic; use in a fume hood.

- Ethanol (Fisher Scientific, cat. no. E/0600DF/17)
- Methanol (Fisher Scientific, cat. no. A456-212) **! CAUTION** Toxic vapours; use in a fume hood.
- Acetone (Fisher Scientific, cat. no. A19-1) **! CAUTION** Acetone vapors may cause dizziness; use in a fume hood.
- ThinPrep (PreservCyt Solution, Cytyc Corp)
- SurePath (Becton Dickinson Diagnostics)

Dewaxing agents

- Xylene (Sigma-Aldrich, cat. no. 534056) **! CAUTION** Potential carcinogen, irritant and allergenic; use in a fume hood.
- Histo-Clear (Fisher Scientific, cat. no. HIS-010-010S) **! CAUTION** It is an irritant.
- Hexane (Fisher Scientific, cat. no. 10764371) **! CAUTION** Extremely flammable liquid, can cause skin irritation; use protective equipment as required; use in a fume hood.

EQUIPMENT

- Microtome (Thermo Fisher Scientific, cat. no. 902100A; or cat. no. 956651)
- Wax dispenser (Electrothermal, cat. no. MH8523B)
- Sectioning bath (Electrothermal, cat. no. MH8517)
- Centrifuge (Thermo Fisher Scientific, cat. no. 75002410)
- Desiccator (Thermo Fisher Scientific, cat. no. 5311-0250)
- Desiccant (Sigma-Aldrich, cat. no. 13767)
- Laser power meter (Coherent, cat. no. 1098293)
- Spectrometer
- Computer system

Substrates

▲ CRITICAL Substrate should be carefully chosen depending on the spectrochemical approach that will be used.

- Low-E slides (Kevley Technologies, CFR)
- BaF_2 slides (Photox Optical Systems)
- CaF_2 slides (Crystran, cat. no. CAFP10-10-1)
- Silicon multi-well plate (Bruker Optics)
- Glass slides (Fisher Scientific, cat. no. 12657956)
- Quartz slides (UQG Optics, cat. no. FQM-2521)
- Aluminum-coated slides (EMF, cat. no. AL134)
- Mirrored stainless steel (Renishaw, cat. no. A-9859-1825-01)

REAGENT SETUP

Tissue For FFPE tissue, the excised specimen is immersed in fixative (*e.g.*, formalin), dehydrated in ethanol, cleared in xylene and embedded in paraffin wax. Specimens can then be stored indefinitely at room temperature. For snap-frozen tissue, the specimen is immersed in OCT, followed by cooling of isopentane with liquid N_2.

▲ CRITICAL Snap-frozen tissue should be thawed before analysis. Spectroscopic analysis should be performed directly after excision in case of fresh tissue to avoid sample degradation.

Cells Cells can be treated with a suitable fixative or preservative solution or studied alive.

▲ CRITICAL In case cells are fixed or stored in a preservative solution, a number of washing steps using centrifugation should be followed prior to spectroscopic analysis to remove unwanted signature. If cells are studied alive, optimum living conditions (*e.g.*, growth medium, temperature and pH) should be maintained; washing of live cells from medium is also necessary.

Biofluids Biofluids can be collected in designated, sterile tubes using standard operating procedures to achieve uniformity of performance. Preparation of biofluids depends on the sample type and the experiment's objective. If cellular material is not directly studied, it should be removed from the biofluid before storage. Biofluids can be analysed right after their collection or stored at a -80°C freezer.

▲ **CRITICAL** If biofluids have been stored in a freezer, it is essential that they are fully thawed before acquiring aliquots for spectroscopic analysis.

▲ **CRITICAL** Users are advised to store biofluids in smaller, single-use aliquots at -80°C to avoid repeated freeze-thaw cycles.

EQUIPMENT SETUP

The user can choose from a range of different instrumental setups and spectral acquisition modes. General information about FTIR systems is provided below. For more details about equipment setup see refs.[91,94,95].

The FTIR spectrometer can be left on for long periods of time. Before spectral acquisition, the user should check the interferogram signal for amplitude and position and keep a record of the measurements.

▲ **CRITICAL** For detectors that require a prior cooling step using liquid nitrogen (*e.g.*, mercury cadmium telluride (MCT) detectors), the signal should be allowed to stabilize for approximately 10 min before data collection.

▲ **CRITICAL** In case that the interferogram signal deviates from the last measurement, re-alignment or part replacement may be required.

Software: Software for spectral acquisition is typically provided by the manufacturer. Software packages for spectral analysis and data standardization are provided in Table 3.

PROCEDURE

Sample preparation

1| Prepare the biological samples for spectrochemical analysis using the following steps: option A for FFPE tissue samples, option B for snap-frozen or fresh tissue samples, option C for cells and option D for biofluids.

▲ **CRITICAL** Sample preparation is briefly presented in this protocol. More details about sample preparation can be found in ref.[91,94,95].

(A) Tissue (FFPE) ● TIMING 1-1.5 h

(i) Acquisition of FFPE tissue blocks.

(ii) Whole tissue block has to be sectioned using a microtome to obtain tissue sections at desired thickness (2-10 µm).

▲ **CRITICAL** Cooling of the tissue on an ice block allows easier sectioning.

(iii) Tissue ribbons are floated in a warm H_2O bath and then deposited onto the substrate of choice.

(iv) The tissue slide is then allowed to dry either at room temperature (30 min) or in a 60°C oven (10 min).

▲ **CRITICAL** The tissue slide may be dried in the oven for longer periods of time, depending on the type of tissue, to ensure optimal melting of the wax initially.

(v) Dewaxing is then performed by three sequential immersions in a dewaxing reagent such as fresh xylene, Histo-Clear solution or hexane (at least 5 min).

▲ CRITICAL Thorough dewaxing is important for eliminating all spectral peaks attributed to paraffin.

(vi) Tissue slide is immersed in acetone or ethanol (5 min) to remove the xylene and then left to air-dry.

■ PAUSE POINT Slides can be stored in a desiccator at room temperature for at least 1 year.

(B) Tissue (Snap-frozen or fresh) ● TIMING 2 h + drying time (3 h for FTIR only)

▲ CRITICAL Snap-frozen tissue can be stored at -80°C for several months.

▲ CRITICAL For fresh tissue, proceed to step 1B(iii).

(i) Acquire snap-frozen tissue from freezer and place onto a cryostat (30 min) to allow the tissue to reach the cryostat's temperature (-20°C).

(ii) Tissue block can be sectioned using the cryostat to obtain tissue sections at desired thickness (8-10 µm).

(iii) The tissue sections are deposited onto an appropriate substrate before spectra are collected.

▲ CRITICAL For FTIR studies the tissue sections need to dry for at least 3 h to remove the H_2O interference with the IR spectra.

▲ CRITICAL Exposure to light should be minimised to prevent sample degradation due to oxidation.

(C) Cells (fixed or live) ● TIMING 30 min + desiccation time (3 h for FTIR only)

▲ CRITICAL If cells are studied live proceed to step 1C(ii)

(i) Fixed cells need to be washed from the fixative or preservative solution to remove any spectral interference in the fingerprint region. Three sequential washes with distilled H_2O or PBS have been shown to remove unwanted peaks.

(ii) Live cells in suspension have to be detached from the growth substrate using trypsin and then washed from the medium and trypsin with PBS (×3 times).

▲ CRITICAL All reagents should be warmed to 37°C to reduce the shock to cells and maintain morphology.

(iii) After the final wash, the remaining cell pellet is resuspended in distilled H_2O and mounted on a substrate of choice.

▲ CRITICAL The final suspension of cells should be evenly deposited on the slide either by cytospinning or by micro-pipetting.

▲ CRITICAL For FTIR studies the sample needs to dry for at least 3 h.

(D) Biofluids (frozen or fresh) ● TIMING 5 min + thawing (20 min) + drying (1-1.5 h)

▲ CRITICAL If biofluids are analysed fresh, immediately after collection, continue to step 1D(ii).

(i) Acquire biofluids from the -80°C freezer and allow them to fully thaw.

(ii) Mix or gently vortex the sample before obtaining the desired volume for analysis.

▲ CRITICAL Only a small amount of the biofluid is typically required for spectroscopic studies (1-100 μL). However, this depends and should be tailored according to the study and experimental design.

(iii) Deposit the biological fluid onto an appropriate substrate.

▲ CRITICAL For ATR-FTIR spectroscopic studies, an alternative option is to deposit the sample directly on the ATR crystal instead of a substrate if the instrumentation setting allows (*i.e.*, if crystal is facing upwards). However, if the sample is sufficiently thick (>2-3 μm) to avoid substrate interference, then the use of a holding substrate is advantageous as it allows measurements from multiple locations as well as longer storage.

▲ CRITICAL For FTIR studies the sample needs to dry adequately before spectroscopic analysis (50 μl dry within approximately 1 h at room temperature). Drying can be sped up by using a gentle stream of air.

Spectral acquisition

2| Spectrochemical information can be collected as follows for FTIR spectroscopy.

▲ CRITICAL Spectral acquisition is briefly presented in this protocol. More details can be found in ref.[91,94,95].

FTIR spectroscopy ● TIMING 2 - 5 min per spectrum

(i) Settings should be optimised before a new study to increase the SNR (see ‘Experimental: spectral acquisition’).

▲ CRITICAL Some of the parameters that need to be adjusted include the resolution, spectral region of interest, co-additions, aperture size, interferometer mirror velocity, and interferogram zero-filling.

(ii) Depending on the sampling mode that has been chosen (ATR-FTIR, transmission or transflection), sample is deposited onto the appropriate holding substrate.

(iii) Load the sample and visualise the region of interest; information can then be acquired either as point map or as image maps.

▲ **CRITICAL** Typically, 5-25 point spectra are collected per sample while for image maps the step size should be the same or smaller than the selected aperture size divided by two. Sampling can be performed with 6 replicates in 3 levels.

▲ **CRITICAL** A background spectrum should be acquired before every sample to account for atmospheric changes.

▲ **CRITICAL** To improve reproducibility and decrease differences between the data collected by different operators, the spectral resolution should be set constant, since it can cause major differences between data collected across different experimental setups.

▲ **CRITICAL** The pressure applied on the sample in the ATR mode affects the signal intensity (*i.e.*, absorbance) between data collected by different instruments and operators. Thus, the pressure applied on the sample should be as closest as possible across different experimental setups to reduce differences between the spectra collected.

■ **PAUSE POINT** Save the acquired data in a database until further analysis.

Data quality evaluation ● **TIMING 15 min – 4 h (depending on the size of the dataset)**

▲ **CRITICAL** Before pre-processing, the raw data can be evaluated using some quality tests to identify anomalous spectra or biased patterns. This can be made by visual inspection of the collected spectra followed by Hotelling T^2 *versus* Q residuals charts using only the mean-centred data, and analysis of PCA residuals.

Data pre-processing ● **TIMING 15 min – 4 h (depending on the size of the dataset)**

▲ **CRITICAL** Steps 1-6 below can vary depending on the nature of the dataset. Table 1 provides more details about these pre-processing steps. In case of an ATR-FTIR dataset acquired under different experimental conditions, the pre-processing method should follow this order:

1. **Cutting at biofingerprint region (900-1800 cm^{-1}).** The spectra should be truncated to the biofingerprint region to reduce atmospheric interference.
2. **Savitzky-Golay smoothing for removing spectral-noise.** Window size varies according to the size of the spectra dataset (*e.g.*, wavenumber). The window size should be an odd number and the analyst should vary it from 3 to 21 and observe how the spectra change (in shape) and how the noise is reduced. The smallest window that removes the noise considerably whilst maintaining the original spectral shape should be used. Using a spectral resolution of 4 cm^{-1}, the biofingerprint region (900-1800 cm^{-1}) usually contains 235 wavenumbers. In that case, a window size of 5 points should be used. The polynomial order for Savitzky-Golay fitting should be 2nd order for IR spectroscopy due to the band shape.
3. **Light scattering correction using either multiplicative scatter correction (MSC), SNV or 2nd derivative.** The user should prioritize MSC or SNV as these methods maintain the spectral scale and original spectral shape. In case of unsatisfactory results, 2nd derivative should be then employed.
4. **Baseline correction using automatic weighted least squares or rubber band baseline correction.** If spectral differentiation is applied as light scattering correction method, baseline correction is not necessary.
5. **Normalization** to the amide I peak, amide II peak or vector normalization (2-Norm, length = 1) should be applied to correct different scales across spectra (*e.g.*, due to different sample thicknesses when using FTIR in transmission mode).
6. **Scaling (*i.e.*, for each variable, mean-centring followed by division by the variable standard deviation).** In case of data fusion, block-scaling should be used.

Data analysis

(A) Exploratory analysis. ● TIMING **1h – 4 d (depending on the data size)**

Exploratory analysis should be primarily conducted using PCA. The PCA scores plot (PC1 vs PC2) should be used for identification of the need of a standardization procedure.

(B) Outlier detection. ● TIMING **1h – 1 d (depending on the data size)**

Apply PCA to the dataset and then estimate the Q residuals and Hotelling T^2 values. Use the chart of Q residuals *versus* Hotelling T^2 to identify outliers. The outliers (*e.g.*, cosmic rays, artefacts, low signal spectra and substrate only (non-tissue) spectra) should be removed from the data set before proceeding to the next steps.

(C) Sample split. ● TIMING **1 – 4 h (depending on the data size)**

Sample split should be performed before construction of standardization of multivariate classification models. The samples can be split into training (70%) and test (30%) sets, using a cross-validated model; or split into training (70%), validation (15%) and test (15%) sets without using cross-validation. To maintain consistency and account for a well-balanced training model, KS algorithm should be employed.

(D) Standardization. ● TIMING **1h – 4 d (depending on the data size)**

▲ CRITICAL Standardization methods should be employed in the following order: DS > PDS. The data from the secondary response should be separated into training (70%), validation (15%) and test (15%) sets using KS algorithm. The number of transfer samples should be firstly optimized using the validation set from the secondary response. Then, when employing PDS, the window size should be optimized according to the size of the dataset.

(i) DS should be employed varying the number of transfer samples from 10-100% of the training set from the primary system. The validation set from the secondary instrument should be used to find the optimum number of transfer samples using the misclassification rate as cost function.

(ii) PDS should be employed using the optimum number of samples found with DS. Different window sizes should be tested using the validation set from the secondary system with the misclassification rate as cost function. The window size should vary from 3-29 for a spectral set with resolution of 4 cm^{-1} in the biofingerprint region (235 variables).

(E) Model construction. ● TIMING 1h – 4 d (depending on the data size)

▲ **CRITICAL** Feature extraction (*e.g.*, by means of PCA) or feature selection (*e.g.*, by means of GA or SPA) should be employed to reduce data collinearity and speed up data processing and analysis time. PLS-DA is already a feature extraction method, thus the performance of prior feature extraction is not necessary in this case. The classification technique employed must follow a parsimony order: LDA>PLS-DA>QDA>KNN>SVM>ANN>Random forests>Deep learning approaches.

(i) Apply the feature extraction or selection technique. The optimization of the number of PCs during PCA can be performed using an external validation set (15% of the original data set) or using cross-validation (leave-one-out for small dataset [≤20 samples] or venetian blinds [sample splitting: 10] for large datasets [>20 samples]). GA should be realized three-times starting from different initial populations and the best result using an external validation set (15% of the original data set) should be used. Cross-over probability should be set for 40% and mutation probability should be set for 1-10% according to the size of the dataset.

(ii) The classification method should be employed using optimization with an external validation set or cross-validation, especially for selecting the number of latent variables of PLS-DA and the kernel parameters for SVM. The kernel function for SVM should be RBF kernel, due to its adaptation to different data distributions. To avoid overfitting, cross-validation should be always performed during model construction to estimate the best RBF parameters.

? TROUBLESHOOTING

Spectral acquisition: Spectral resolution, spectral range, SNR and signal aperture should be optimized during experimental setup. Operators using different systems should try to keep these parameters constant to reduce spectral differences.

Data pre-processing: To reduce spectral differences, the same data pre-processing should be applied for spectra acquired in different systems.

Standardization: To improve the prediction capability of the classification model, the primary system used should be the one with highest spectral resolution and smallest noise, since all data from the secondary systems will be standardized to this pattern.

● TIMING

Sample preparation:

(A) Tissue (FFPE): 1-1.5 h

(B) Tissue (Snap-frozen or fresh): 2 h + drying time (3 h)

(C) Cells (fixed or live): 30 min + desiccation time (3 h)

(D) Biofluids (frozen or fresh): 5 min + thawing (20 min) + drying (1-1.5 h)

Spectral acquisition: 1 s – 5 min per spectrum (depending on the instrument and spectral acquisition configurations)

Data pre-processing: 15 min – 4 h

Data analysis:

(A) Exploratory analysis: 1 h – 4 d

(B) Outlier detection: 1 h – 1 d

(C) Standardization: 1 h – 4 d

(D) Model construction: 1 h – 4 d

ANTICIPATED RESULTS

A pilot study was conducted to evaluate the effect of different instrument manufacturers and operators towards spectral acquisition of healthy controls and ovarian cancer samples based on blood plasma (5 healthy controls with 10 spectra per sample; 5 ovarian cancers with 10 spectra per sample) for a binary classification model using ATR-FTIR spectroscopy. All specimens were collected with ethical approval obtained at Royal Preston Hospital UK (16/EE/0010). Table 4 summarizes the experimental conditions in which the experiments were performed.

Table 4. Experimental conditions for pilot study.

Instrument	**Operator**	**Spectral range**	**Number of co-additions**	**Spectral resolution**	**Room temperature**	**Air humidity**
A	1	4000-400 cm^{-1}	32	4 cm^{-1}	23.0ºC	23%
	2	4000-400 cm^{-1}	32	4 cm^{-1}	23.4ºC	26%
B	1	4000-400 cm^{-1}	32	4 cm^{-1}	24.0ºC	26%
	2	4000-400 cm^{-1}	32	4 cm^{-1}	24.9ºC	24%
C	1	4000-400 cm^{-1}	48	4 cm^{-1}	22.5ºC	28%
	2	4000-400 cm^{-1}	48	1 cm^{-1}	22.8ºC	26%

Instrument A and B were Bruker Tensor 27 with an HELIOS ATR attachment while instrument C was an ATR-FTIR Thermo Scientific Nicolet iS10. The spectra were collected for the same types of samples within three different days (operator 1: instrument A in day 1, instrument B in day 3, and instrument C in day 2; operator 2: instrument A in day 2, instrument B in day 1, and instrument C in day 3) and across two different laboratories (instrument A and B in laboratory 1 and instrument C in laboratory 2). Each operator prepared the samples individually from the same bulk, and measured them individually. Spectral acquisition times were around 30 s for instruments A and B, and 40 s for instrument C.

(A) Effect of different instruments

Three different ATR-FTIR spectrometers were used to analyse the samples. Data were pre-processed by truncating at the biological fingerprint region (900-1800 cm^{-1}), followed by Savitzky-Golay smoothing (window of 15 points, 2nd order polynomial function), MSC, baseline correction using automatic weighted least squares and vector normalization (2-Norm, length = 1). Each data set (A, B and C) was pre-processed individually. The raw and pre-processed spectra for healthy controls and ovarian cancer samples are depicted in Supplementary Material 1. All spectra collected by the three instrument maintained the same spectral shape, indicating that the chemical information stayed the same; however, large differences between the absorbance intensity were observed between instrument C and the others (A, B), being caused due to different pressures applied on the sample in the ATR module. The pressure applied to keep the sample in contact with the ATR crystal directly affects the spectral signal intensity, which for instrument A and B (same manufactures) were somewhere controlled by a contra weight, while for instrument C the pressure was set based on a mechanical screw on the device, thus being biased by the operator usage. The absorbance intensity variation between A and B is observed for this same reason, but in a minor scale. Outlier detection was performed using a Hotelling T^2 versus Q residual test (Supplementary Material 1).

(i) Classification. Classification was performed using PCA-LDA (10 PCs, explained variance of 99.21%). Fig. 5a depicts the discriminant function (DF) score plot for PCA-LDA using only the primary system (ATR-FTIR A). As observed, there is an almost perfect separation between the samples from the two classes (accuracy = 100%, sensitivity = 100%, specificity = 100%). However, when the spectra acquired using instruments B and C are predicted using the model for A, the results decreased significantly (accuracy = 66.7%, sensitivity = 83.2%, specificity = 48.9%) (Fig. 5b), necessitating the use of a standardization procedure.

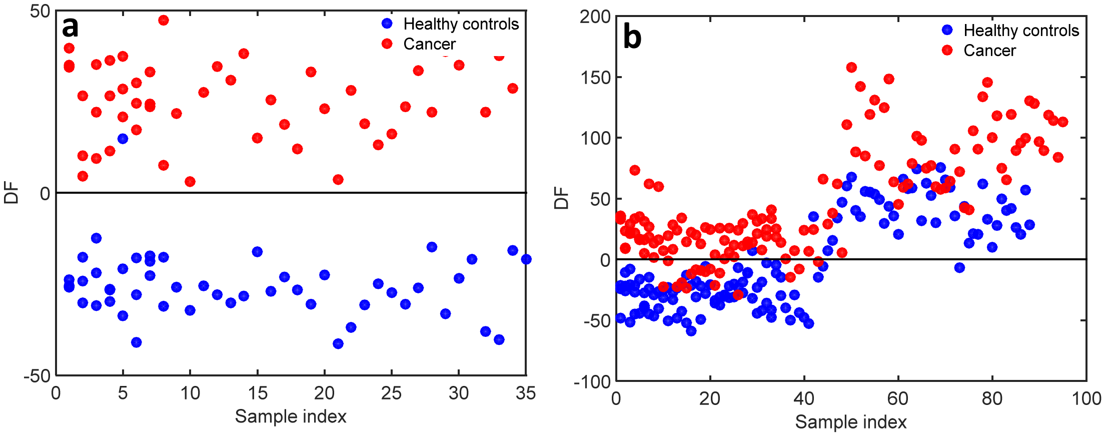


Figure 5. (a) DF plot of the PCA-LDA model for the primary system; (b) DF plot of the PCA-LDA model for the primary system predicting the samples from the secondary systems.

(ii) Standardization. Standardization was employed using both DS and PDS in order to compare the two methods. The number of transfer samples for DS was optimized according to the misclassification rate obtained for the validation set using the secondary system (Fig. 6a). An optimum number corresponding to 80% of the samples in the training set of the primary system (55 transfer samples) was obtained, resulting to a misclassification rate of 22.2% in the validation set of the secondary system. This improved the accuracy (77.8%) and specificity (80.0%). Sensitivity decreased to 75.0%, which is an acceptable value. The results after DS are better balanced than without standardization. Fig. 6b shows the DF plot for the PCA-LDA model using the training of the primary system and prediction with the secondary system after DS.

PDS was also applied. The number of transfer samples was maintained as 55 (80% of the primary training set) and the window size was optimized by using the validation set of the secondary system. An optimum window size of 23 wavenumbers was selected with a misclassification rate of 25.9% (Fig. 6c). The accuracy, sensitivity and specificity using PDS were 74.1%, 71.4% and 75.0%, respectively. The DS presented a slightly higher performance than PDS for this dataset. However, DS generated some outliers not observed before, while PDS did not. Thus, in general, PDS provided a better standardization of the data. The PCA-LDA DF plot after PDS is depicted in Fig. 6d.

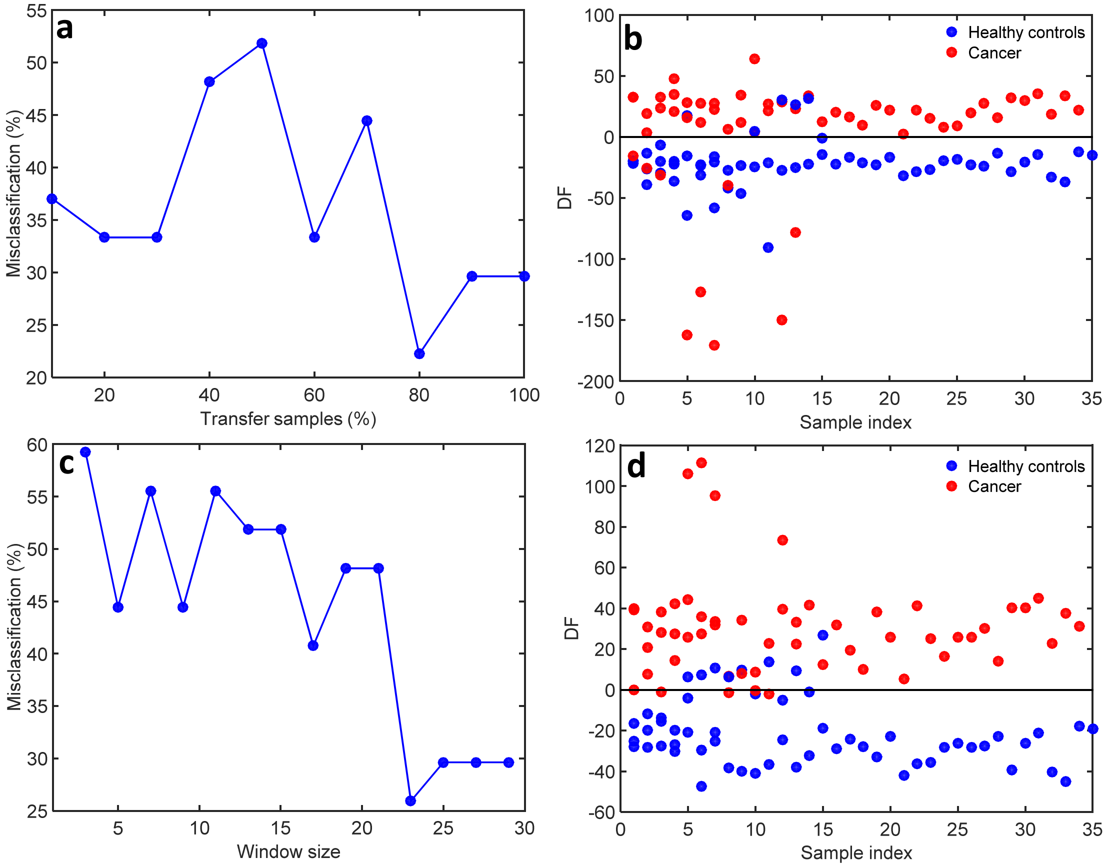


Figure 6. (a) Misclassification rate in % for the validation set of the secondary system varying the number of transfer samples in % from the primary system for DS optimization; (b) DF plot of the PCA-LDA model for the primary system predicting the validation set from the secondary system after DS; (c) Misclassification rate in % for the validation set of the secondary system varying the window size for PDS optimization; (d) DF plot of the PCA-

LDA model for the primary system predicting the validation set from the secondary system after PDS.

(B) Effect of different operators

The effect of different user operators acquiring spectra from the same samples using the same instruments was also evaluated. Similarly to before, data were pre-processed by cutting the biological fingerprint region (900-1800 cm^{-1}), followed by Savitzky-Golay smoothing (window of 15 points, 2nd order polynomial function), MSC, baseline correction using automatic weighted least squares and vector normalization (2-Norm, length = 1). Each dataset was pre-processed individually. All raw and pre-processed spectra varying operators are depicted in Supplementary Material 1. Outlier detection was performed using a Hotelling T^2 versus Q residual test (Supplementary Material 1). The PCA scores plots for the pre-processed spectra are depicted in Supplementary Material 1. The main difference between the operators was observed for instrument C (Supplementary Material 1, Figure S5e), since the spectral resolutions used by them were different, which can cause major data distortion.

(i) Classification. Classification was performed using PCA-LDA (10 PCs, explained variance of 98.62%). Fig. 7a depicts the DF score plot for PCA-LDA using only the primary system (Operator 1). There is a significant separation between the samples from the two classes (accuracy = 88.4%, sensitivity = 77.3%, specificity = 100%). When the spectra acquired by Operator 2 are predicted using the model for Operator 1, the results decreased (accuracy = 75.6%, sensitivity = 66.7%, specificity = 84.6%) (Fig. 7b), which again necessitates the use of a standardization procedure.

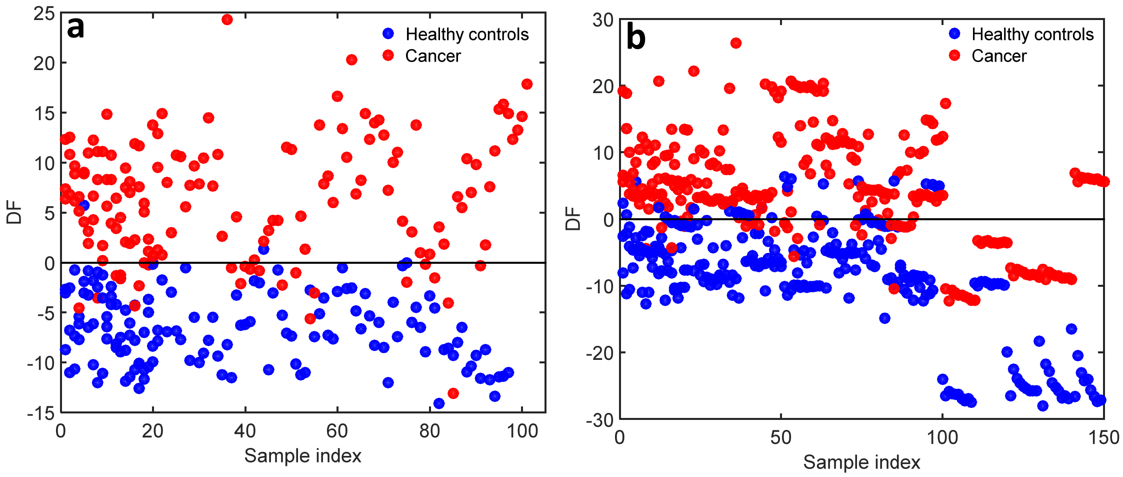


Figure 7. (a) DF plot of the PCA-LDA model for the primary system (Operator 1); (b) DF plot of the PCA-LDA model for the primary system predicting the samples from the secondary system (Operator 2).

(ii) Standardization. DS and PDS were employed as standardization methods. The number of transfer samples for DS was optimized according to the misclassification rate obtained for the validation set using the secondary system (Operator 2) (Fig. 8a). An optimum number of 59 transfer samples (30% of the samples in the training set of the primary system [Operator 1]) was obtained, resulting in a misclassification rate of 17.8% in the validation set of the secondary system. This improved the accuracy (82.2%), sensitivity (69.6%) and specificity (95.5%) compared to the results without DS. Fig. 8b shows the DF plot for the PCA-LDA model using the training of the primary system and prediction with the secondary system after DS.

The number of transfer samples was maintained as 59 for PDS; and the window size was optimized by using the validation set of the secondary system. An optimum window size of 23 wavenumbers was selected with a misclassification rate of 22.2% (Fig. 8c). The accuracy, sensitivity and specificity using PDS were 77.8%, 100% and 54.5%, respectively. Although DS obtained an average better classification performance than PDS for this dataset, it also generated some outliers as mentioned before. For this reason, the results after PDS seem better standardized. The PCA-LDA DF plot after PDS is depicted in Fig. 8d.

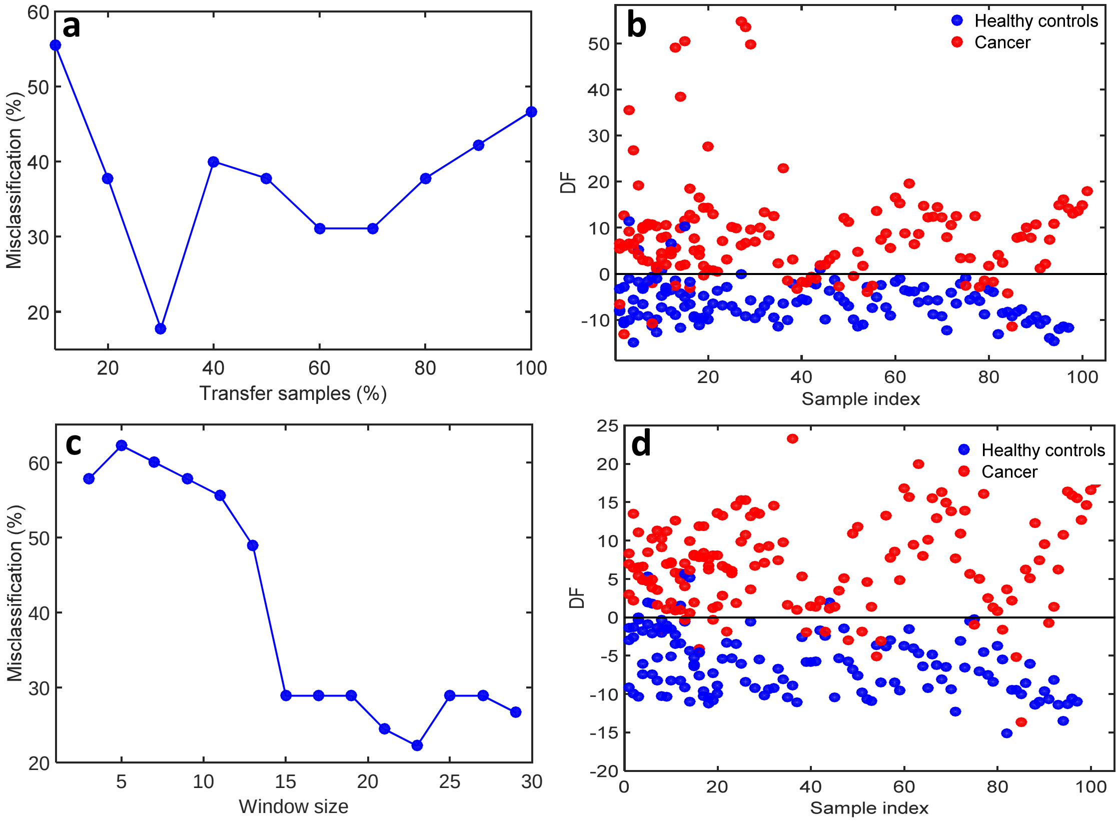


Figure 8. (a) Misclassification rate in % for the validation set of the secondary system (Operator 2) varying the number of transfer samples in % from the primary system (Operator 1) for DS optimization; (b) DF plot of the PCA-LDA model for the primary system predicting the validation set from the secondary system after DS; (c) Misclassification rate in % for the validation set of the secondary system varying the window size for PDS optimization; (d) DF plot of the PCA-LDA model for the primary system predicting the validation set from the secondary system after PDS.

Acknowledgements

CLMM would like to thank CAPES-Brazil (grant Doutorado Pleno no Exterior No. 88881.128982/2016-01) for financial support. MP would like to acknowledge Rosemere Cancer Foundation for funding.

Author contributions

F.L.M. is the principal investigator who conceived the idea for the manuscript; C.L.M.M. and M.P. wrote the manuscript. All co-authors contributed recommendations and provided feedback and changes to the manuscript; and, C.L.M.M., M.P. and F.L.M. brought together the text and finalized the manuscript.

Competing financial interests

The authors declare no competing financial interest.

Data availability statement

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

References

1 Hofmann-Wellenhof, B., Lichtenegger, H. & Collins, J. *Global positioning system: theory and practice*. (Springer Science & Business Media, 2012).

2 Morris, P. & Perkins, A. Diagnostic imaging. *Lancet* **379**, 1525-1533 (2012).

3 Lee, S. S. *et al.* Crohn disease of the small bowel: comparison of CT enterography, MR enterography, and small-bowel follow-through as diagnostic techniques. *Radiology* **251**, 751-761 (2009).

4 Lagleyre, S. *et al.* Reliability of high-resolution CT scan in diagnosis of otosclerosis. *Otol Neurotol* **30**, 1152-1159 (2009).

5 Kalita, J. & Misra, U. Comparison of CT scan and MRI findings in the diagnosis of Japanese encephalitis. *J Neurol Sci* **174**, 3-8 (2000).

6 Schrevens, L., Lorent, N., Dooms, C. & Vansteenkiste, J. The role of PET scan in diagnosis, staging, and management of non-small cell lung cancer. *Oncologist* **9**, 633-643 (2004).

7 Jagust, W., Reed, B., Mungas, D., Ellis, W. & Decarli, C. What does fluorodeoxyglucose PET imaging add to a clinical diagnosis of dementia? *Neurology* **69**, 871-877 (2007).

8 Zhou, M. *et al.* Clinical utility of breast-specific gamma imaging for evaluating disease extent in the newly diagnosed breast cancer patient. *Am J Surg* **197**, 159-163 (2009).

9 Wallace, B. A. *et al.* Biomedical applications of synchrotron radiation circular dichroism spectroscopy: identification of mutant proteins associated with disease and development of a reference database for fold motifs. *Faraday Discuss* **126**, 237-243 (2004).

10 Greenfield, N. J. Using circular dichroism spectra to estimate protein secondary structure. *Nat Protoc* **1**, 2876 (2006).

11 Micsonai, A. *et al.* Accurate secondary structure prediction and fold recognition for circular dichroism spectroscopy. *Proc Natl Acad Sci USA* **112**, E3095-E3103 (2015).

12 Miles, A. J. & Wallace, B. A. Circular dichroism spectroscopy of membrane proteins. *Chem Soc Rev* **45**, 4859-4872 (2016).

13 Brown, J. Q., Vishwanath, K., Palmer, G. M. & Ramanujam, N. Advances in quantitative UV–visible spectroscopy for clinical and pre-clinical application in cancer. *Curr Opin Biotechnol* **20**, 119-131 (2009).

14 Yang, P.-W. *et al.* Visible-absorption spectroscopy as a biomarker to predict treatment response and prognosis of surgically resected esophageal cancer. *Sci Rep* **6**, 33414 (2016).

15 Organization, W. H. *Fluorescence microscopy for disease diagnosis and environmental monitoring*. (2005).

16 Shahzad, A. *et al.* Diagnostic application of fluorescence spectroscopy in oncology field: hopes and challenges. *Appl Spectrosc Rev* **45**, 92-99 (2010).

17 Sieroń, A. *et al.* The role of fluorescence diagnosis in clinical practice. *Onco Targets Ther* **6**, 977 (2013).

18 Shin, D., Vigneswaran, N., Gillenwater, A. & Richards-Kortum, R. Advances in fluorescence imaging techniques to detect oral cancer and its precursors. *Future Oncol* **6**, 1143-1154 (2010).

19 Shahzad, A. *et al.* Emerging applications of fluorescence spectroscopy in medical microbiology field. *J Transl Med* **7**, 99 (2009).

20 Möller-Hartmann, W. *et al.* Clinical application of proton magnetic resonance spectroscopy in the diagnosis of intracranial mass lesions. *Neuroradiology* **44**, 371-381 (2002).

21 Gowda, G. N. *et al.* Metabolomics-based methods for early disease diagnostics. *Expert Rev Mol Diagn* **8**, 617-633 (2008).

22 Frisoni, G. B., Fox, N. C., Jack, C. R., Scheltens, P. & Thompson, P. M. The clinical use of structural MRI in Alzheimer disease. *Nat Rev Neurol* **6**, 67-77 (2010).

23 Chan, A. W. *et al.* 1 H-NMR urinary metabolomic profiling for diagnosis of gastric cancer. *Br J Cancer* **114**, 59 (2016).

24 Palmnas, M. S. & Vogel, H. J. The future of NMR metabolomics in cancer therapy: towards personalizing treatment and developing targeted drugs? *Metabolites* **3**, 373-396 (2013).

25 Patil, P. & Dasgupta, B. Role of diagnostic ultrasound in the assessment of musculoskeletal diseases. *Ther Adv Musculoskelet Dis* **4**, 341-355 (2012).

26 Navani, N. *et al.* Lung cancer diagnosis and staging with endobronchial ultrasound-guided transbronchial needle aspiration compared with conventional approaches: an open-label, pragmatic, randomised controlled trial. *Lancet Respir Med* **3**, 282-289 (2015).

27 Menon, U. *et al.* Sensitivity and specificity of multimodal and ultrasound screening for ovarian cancer, and stage distribution of detected cancers: results of the prevalence screen of the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS). *Lancet Oncol* **10**, 327-340 (2009).

28 Smith-Bindman, R. *et al.* Endovaginal ultrasound to exclude endometrial cancer and other endometrial abnormalities. *Jama* **280**, 1510-1517 (1998).

29 Gajjar, K. *et al.* Diagnostic segregation of human brain tumours using Fourier-transform infrared and/or Raman spectroscopy coupled with discriminant analysis. *Anal Methods* **5**, 89-102 (2013).

30 Bury, D. *et al.* Phenotyping Metastatic Brain Tumors Applying Spectrochemical Analyses: Segregation of Different Cancer Types. *Anal. Lett.*, 1-2 (2018).

31 Hands, J. R. *et al.* Attenuated Total Reflection Fourier Transform Infrared (ATR-FTIR) spectral discrimination of brain tumour severity from serum samples. *J Biophotonics* **7**, 189-199 (2014).

32 Hands, J. R. *et al.* Brain tumour differentiation: rapid stratified serum diagnostics via attenuated total reflection Fourier-transform infrared spectroscopy. *Journal of neuro-oncology* **127**, 463-472 (2016).

33 Walsh, M. J., Kajdacsy-Balla, A., Holton, S. E. & Bhargava, R. Attenuated total reflectance Fourier-transform infrared spectroscopic imaging for breast histopathology. *Vib Spectrosc* **60**, 23-28 (2012).

34 Lane, R. & Seo, S. S. Attenuated Total Reflectance Fourier Transform Infrared Spectroscopy Method to Differentiate Between Normal and Cancerous Breast Cells. *J Nanosci Nanotechnol* **12**, 7395-7400 (2012).

35 Backhaus, J. *et al.* Diagnosis of breast cancer with infrared spectroscopy from serum samples. *Vib Spectrosc* **52**, 173-177 (2010).

36 Wang, J.-S. *et al.* FT-IR spectroscopic analysis of normal and cancerous tissues of esophagus. *World journal of gastroenterology* **9**, 1897 (2003).

37 Maziak, D. E. *et al.* Fourier-transform infrared spectroscopic study of characteristic molecular structure in cancer cells of esophagus: an exploratory study. *Cancer Detect. Prev.* **31** (2007).

38 McIntosh, L. M. *et al.* Infrared spectra of basal cell carcinomas are distinct from non-tumor-bearing skin components. *J Investig Dermatol* **112**, 951-956 (1999).

39 McIntosh, L. M. *et al.* Towards non-invasive screening of skin lesions by near-infrared spectroscopy. *Journal of Investigative Dermatology* **116**, 175-181 (2001).

40 Mostaço-Guidolin, L. B., Murakami, L. S., Nomizo, A. & Bachmann, L. Fourier transform infrared spectroscopy of skin cancer cells and tissues. *Appl Spectrosc Rev* **44**, 438-455 (2009).

41 Mordechai, S. *et al.* Possible common biomarkers from FTIR microspectroscopy of cervical cancer and melanoma. *Journal of microscopy* **215**, 86-91 (2004).

42 Hammody, Z., Sahu, R. K., Mordechai, S., Cagnano, E. & Argov, S. Characterization of malignant melanoma using vibrational spectroscopy. *The Scientific World Journal* **5**, 173-182 (2005).

43 Kondepati, V. R., Keese, M., Mueller, R., Manegold, B. C. & Backhaus, J. Application of near-infrared spectroscopy for the diagnosis of colorectal cancer in resected human tissue specimens. *Vib Spectrosc* **44**, 236-242 (2007).

44 Rigas, B., Morgello, S., Goldman, I. S. & Wong, P. Human colorectal cancers display abnormal Fourier-transform infrared spectra. *Proceedings of the National Academy of Sciences* **87**, 8140-8144 (1990).

45 Yao, H., Shi, X. & Zhang, Y. The Use of FTIR-ATR Spectrometry for Evaluation of Surgical Resection Margin in Colorectal Cancer: A Pilot Study of 56 Samples. *J Spectrosc* **2014**, 4 (2014).

46 Lewis, P. D. *et al.* Evaluation of FTIR Spectroscopy as a diagnostic tool for lung cancer using sputum. *BMC Cancer* **10**, 640 (2010).

47 Akalin, A. *et al.* Classification of malignant and benign tumors of the lung by infrared spectral histopathology (SHP). *Lab Invest* **95**, 406 (2015).

48 Großerueschkamp, F. *et al.* Marker-free automated histopathological annotation of lung tumour subtypes by FTIR imaging. *Analyst* **140**, 2114-2120 (2015).

49 Owens, G. L. *et al.* Vibrational biospectroscopy coupled with multivariate analysis extracts potentially diagnostic features in blood plasma/serum of ovarian cancer patients. *J Biophotonics* **7**, 200-209 (2014).

50 Gajjar, K. *et al.* Fourier-transform infrared spectroscopy coupled with a classification machine for the analysis of blood plasma or serum: a novel diagnostic approach for ovarian cancer. *Analyst* **138**, 3917-3926 (2013).

51 Theophilou, G., Lima, K. M. G., Martin-Hirsch, P. L., Stringfellow, H. F. & Martin, F. L. ATR-FTIR spectroscopy coupled with chemometric analysis discriminates normal, borderline and malignant ovarian tissue: classifying subtypes of human cancer. *Analyst* **141**, 585-594 (2016).

52 Mehrotra, R., Tyagi, G., Jangir, D. K., Dawar, R. & Gupta, N. Analysis of ovarian tumor pathology by Fourier Transform Infrared Spectroscopy. *J Ovarian Res* **3**, 27 (2010).

53 Paraskevaidi, M. *et al.* Potential of mid-infrared spectroscopy as a non-invasive diagnostic test in urine for endometrial or ovarian cancer. *Analyst* (2018).

54 Taylor, S. E. *et al.* Infrared spectroscopy with multivariate analysis to interrogate endometrial tissue: a novel and objective diagnostic approach. *Br J Cancer* **104**, 790-797 (2011).

55 Paraskevaidi, M. *et al.* Aluminium foil as an alternative substrate for the spectroscopic interrogation of endometrial cancer. *J Biophotonics* (2018).

56 Gajjar, K. *et al.* Histology verification demonstrates that biospectroscopy analysis of cervical cytology identifies underlying disease more accurately than conventional screening: removing the confounder of discordance. *PLoS One* **9**, e82416 (2014).

57 Walsh, M. J. *et al.* IR microspectroscopy: potential applications in cervical cancer screening. *Cancer Lett.* **246**, 1-11 (2007).

58 Wood, B. R., Quinn, M. A., Burden, F. R. & McNaughton, D. An investigation into FTIR spectroscopy as a biodiagnostic tool for cervical cancer. *Biospectroscopy* **2**, 143-153 (1996).

59 Podshyvalov, A. *et al.* Distinction of cervical cancer biopsies by use of infrared microspectroscopy and probabilistic neural networks. *Appl Opt* **44**, 3725-3734 (2005).

60 Theophilou, G. *et al.* A biospectroscopic analysis of human prostate tissue obtained from different time periods points to a trans-generational alteration in spectral phenotype. *Sci Rep* **5**, 13465 (2015).

61 Baker, M. J. *et al.* Investigating FTIR based histopathology for the diagnosis of prostate cancer. *J Biophotonics* **2** (2009).

62 Derenne, A., Gasper, R. & Goormaghtigh, E. The FTIR spectrum of prostate cancer cells allows the classification of anticancer drugs according to their mode of action. *Analyst* **136** (2011).

63 Gazi, E. *et al.* A correlation of FTIR spectra derived from prostate cancer biopsies with Gleason grade and tumour stage. *European urology* **50**, 750-761 (2006).

64 Paraskevaidi, M. *et al.* Differential diagnosis of Alzheimer’s disease using spectrochemical analysis of blood. *Proc Natl Acad Sci USA*, 201701517 (2017).

65 Carmona, P. *et al.* Discrimination analysis of blood plasma associated with Alzheimer's disease using vibrational spectroscopy. *J Alzheimers Dis* **34**, 911-920 (2013).

66 Carmona, P., Molina, M., López-Tobar, E. & Toledano, A. Vibrational spectroscopic analysis of peripheral blood plasma of patients with Alzheimer's disease. *Anal Bioanal Chem* **407**, 7747-7756 (2015).

67 Paraskevaidi, M. *et al.* Blood-based near-infrared spectroscopy for the rapid low-cost detection of Alzheimer's disease. *Analyst* (2018).

68 Sitole, L., Steffens, F., Krüger, T. P. J. & Meyer, D. Mid-ATR-FTIR Spectroscopic Profiling of HIV/AIDS Sera for Novel Systems Diagnostics in Global Health. *OMICS* **18**, 513-523 (2014).

69 Coopman, R. *et al.* Glycation in human fingernail clippings using ATR-FTIR spectrometry, a new marker for the diagnosis and monitoring of diabetes mellitus. *Clin Biochem* **50**, 62-67 (2017).

70 Scott, D. A. *et al.* Diabetes-related molecular signatures in infrared spectra of human saliva. *Diabetol Metab Syndr* **2**, 48 (2010).

71 Varma, V. K., Kajdacsy-Balla, A., Akkina, S. K., Setty, S. & Walsh, M. J. A label-free approach by infrared spectroscopic imaging for interrogating the biochemistry of diabetic nephropathy progression. *Kidney Int* **89**, 1153-1159 (2016).

72 Lechowicz, L., Chrapek, M., Gaweda, J., Urbaniak, M. & Konieczna, I. Use of Fourier-transform infrared spectroscopy in the diagnosis of rheumatoid arthritis: a pilot study. *Mol Biol Rep* **43**, 1321-1326 (2016).

73 Canvin, J. *et al.* Infrared spectroscopy: shedding light on synovitis in patients with rheumatoid arthritis. *Rheumatology* **42**, 76-82 (2003).

74 Oemrawsingh, R. M. *et al.* Near-infrared spectroscopy predicts cardiovascular outcome in patients with coronary artery disease. *J Am Coll Cardiol* **64**, 2510-2518 (2014).

75 Wang, J. *et al.* Near-infrared spectroscopic characterization of human advanced atherosclerotic plaques. *J Am Coll Cardiol* **39**, 1305-1313 (2002).

76 Martin, M. *et al.* The effect of common anticoagulants in detection and quantification of malaria parasitemia in human red blood cells by ATR-FTIR spectroscopy. *Analyst* (2017).

77 Khoshmanesh, A. *et al.* Detection and Quantification of Early-Stage Malaria Parasites in Laboratory Infected Erythrocytes by Attenuated Total Reflectance Infrared Spectroscopy and Multivariate Analysis. *Anal Chem* **86**, 4379-4386 (2014).

78 Roy, S. *et al.* Simultaneous ATR-FTIR Based Determination of Malaria Parasitemia, Glucose and Urea in Whole Blood Dried onto a Glass Slide. *Anal Chem* **89**, 5238-5245 (2017).

79 Markus, A. P. J. *et al.* New technique for diagnosis and monitoring of alcaptonuria: quantification of homogentisic acid in urine with mid-infrared spectrometry. *Anal Chim Acta* **429**, 287-292 (2001).

80 Grimard, V. *et al.* Phosphorylation-induced Conformational Changes of Cystic Fibrosis Transmembrane Conductance Regulator Monitored by Attenuated Total Reflection-Fourier Transform IR Spectroscopy and Fluorescence Spectroscopy. *J Biol Chem* **279**, 5528-5536 (2004).

81 Aksoy, C., Guliyev, A., Kilic, E., Uckan, D. & Severcan, F. Bone marrow mesenchymal stem cells in patients with beta thalassemia major: molecular analysis with attenuated total reflection-Fourier transform infrared spectroscopy study as a novel method. *Stem Cells Dev* **21**, 2000-2011 (2012).

82 Graça, G. *et al.* Mid-infrared (MIR) metabolic fingerprinting of amniotic fluid: A possible avenue for early diagnosis of prenatal disorders? *Anal Chim Acta* **764**, 24-31 (2013).

83 Hasegawa, J. *et al.* Evaluation of placental function using near infrared spectroscopy during fetal growth restriction. *J Perinatal Med* **38**, 29-32 (2010).

84 Theelen, T., Berendschot, T. T., Hoyng, C. B., Boon, C. J. & Klevering, B. J. Near-infrared reflectance imaging of neovascular age-related macular degeneration. *Graefe's Archive for Clinical and Experimental Ophthalmology* **247**, 1625 (2009).

85 Semoun, O. *et al.* Infrared features of classic choroidal neovascularisation in exudative age-related macular degeneration. *Br. J. Ophthalmol.* **93**, 182-185 (2009).

86 Peters, A. S. *et al.* Serum-infrared spectroscopy is suitable for diagnosis of atherosclerosis and its clinical manifestations. *Vib Spectrosc* **92**, 20-26 (2017).

87 Afara, I. O., Prasadam, I., Arabshahi, Z., Xiao, Y. & Oloyede, A. Monitoring osteoarthritis progression using near infrared (NIR) spectroscopy. *Sci Rep* **7**, 11463 (2017).

88 Bi, X. *et al.* Fourier transform infrared imaging and MR microscopy studies detect compositional and structural changes in cartilage in a rabbit model of osteoarthritis. *Anal Bioanal Chem* **387**, 1601-1612 (2007).

89 David-Vaudey, E. *et al.* Fourier Transform Infrared Imaging of focal lesions in human osteoarthritic cartilage. *Eur Cell Mater* **10**, 60 (2005).

90 Trevisan, J., Angelov, P. P., Carmichael, P. L., Scott, A. D. & Martin, F. L. Extracting biological information with computational analysis of Fourier-transform infrared (FTIR) biospectroscopy datasets: current practices to future perspectives. *Analyst* **137**, 3202-3215 (2012).

91 Baker, M. J. *et al.* Using Fourier transform IR spectroscopy to analyze biological materials. *Nat Protoc* **9**, 1771-1791 (2014).

92 Andrew Chan, K. L. & Kazarian, S. G. Attenuated total reflection Fourier-transform infrared (ATR-FTIR) imaging of tissues and live cells. *Chem Soc Rev* **45**, 1850-1864 (2016).

93 Pilling, M. & Gardner, P. Fundamental developments in infrared spectroscopic imaging for biomedical applications. *Chem Soc Rev* **45**, 1935-1957 (2016).

94 Martin, F. L. *et al.* Distinguishing cell types or populations based on the computational analysis of their infrared spectra. *Nat Protoc* **5**, 1748-1760 (2010).

95 Butler, H. J. *et al.* Using Raman spectroscopy to characterize biological materials. *Nat Protoc* **11**, 664-687 (2016).

96 Kong, L. *et al.* Characterization of bacterial spore germination using phase-contrast and fluorescence microscopy, Raman spectroscopy and optical tweezers. *Nat Protoc* **6**, 625 (2011).

97 Harmsen, S., Wall, M. A., Huang, R. & Kircher, M. F. Cancer imaging using surface-enhanced resonance Raman scattering nanoparticles. *Nat Protoc* **12**, 1400 (2017).

98 Beckonert, O. *et al.* Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nat Protoc* **2**, 2692 (2007).

99 Felten, J. *et al.* Vibrational spectroscopic image analysis of biological material using multivariate curve resolution–alternating least squares (MCR-ALS). *Nat Protoc* **10**, 217 (2015).

100 Yang, H., Yang, S., Kong, J., Dong, A. & Yu, S. Obtaining information about protein secondary structures in aqueous solution using Fourier transform IR spectroscopy. *Nat Protoc* **10**, 382 (2015).

101 Sreedhar, H. *et al.* High-definition Fourier transform infrared (FT-IR) spectroscopic imaging of human tissue sections towards improving pathology. *J Vis Exp* (2015).

102 Varriale, A. *et al.* Fluorescence correlation spectroscopy assay for gliadin in food. *Anal Chem* **79**, 4687-4689 (2007).

103 Song, X., Li, H., Al-Qadiri, H. M. & Lin, M. Detection of herbicides in drinking water by surface-enhanced Raman spectroscopy coupled with gold nanostructures. *J Food Meas Charact* **7**, 107-113 (2013).

104 Osborne, B. G. & Fearn, T. Near-infrared spectroscopy in food analysis. *Encyclopedia Anal Chem* **5**, 4069-4082 (2000).

105 Qu, J.-H. *et al.* Applications of near-infrared spectroscopy in food safety evaluation and control: A review of recent research advances. *Crit. Rev. Food Sci. Nutr.* **55**, 1939-1954 (2015).

106 Penido, C. A. F., Pacheco, M. T. T., Lednev, I. K. & Silveira, L. Raman spectroscopy in forensic analysis: identification of cocaine and other illegal drugs of abuse. *J Raman Spectrosc* **47**, 28-38 (2016).

107 Ryder, A. G. Classification of narcotics in solid mixtures using principal component analysis and Raman spectroscopy. *J Forensic Sci* **47**, 275-284 (2002).

108 Melin, A. M., Perromat, A. & Déléris, G. Pharmacologic application of Fourier transform IR spectroscopy: in vivo toxicity of carbon tetrachloride on rat liver. *Biopolymers: Original Research on Biomolecules* **57**, 160-168 (2000).

109 Harrigan, G. G. *et al.* Application of high-throughput Fourier-transform infrared spectroscopy in toxicology studies: contribution to a study on the development of an animal model for idiosyncratic toxicity. *Toxicol. Lett.* **146**, 197-205 (2004).

110 Choo-Smith, L.-P. *et al.* Investigating microbial (micro) colony heterogeneity by vibrational spectroscopy. *Appl Environ Microbiol* **67**, 1461-1469 (2001).

111 Helm, D., Labischinski, H., Schallehn, G. & Naumann, D. Classification and identification of bacteria by Fourier-transform infrared spectroscopy. *Microbiology* **137**, 69-79 (1991).

112 Carmona, P., Monzon, M., Monleon, E., Badiola, J. J. & Monreal, J. In vivo detection of scrapie cases from blood by infrared spectroscopy. *J. Gen. Virol.* **86**, 3425-3431 (2005).

113 Cui, L. *et al.* A novel functional single-cell approach to probing nitrogen-fixing bacteria in soil communities by resonance Raman spectroscopy with 15N2 labelling. *Anal. Chem.* **10.1021/acs.analchem.7b05080.** (2018).

114 Lasch, P. & Naumann, D. Infrared spectroscopy in microbiology. *Encyclopedia Anal Chem* (2015).

115 Maquelin, K. *et al.* Identification of medically relevant microorganisms by vibrational spectroscopy. *J Microbiol Methods* **51**, 255-271 (2002).

116 Day, J. S., Edwards, H. G., Dobrowski, S. A. & Voice, A. M. The detection of drugs of abuse in fingerprints using Raman spectroscopy I: latent fingerprints. *Spectrochim Acta A Mol Biomol Spectrosc* **60**, 563-568 (2004).

117 Macleod, N. A. & Matousek, P. Emerging Non-invasive Raman Methods in Process Control and Forensic Applications. *Pharm Res* **25**, 2205 (2008).

118 Lewis, I., Daniel Jr, N., Chaffin, N., Griffiths, P. & Tungol, M. Raman spectroscopic studies of explosive materials: towards a fieldable explosives detector. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **51**, 1985-2000 (1995).

119 Hargreaves, M. D. & Matousek, P. Threat detection of liquid explosive precursor mixtures by Spatially Offset Raman Spectroscopy (SORS). in *Optics and photonics for counterterrorism and crime fighting V.* Vol. **7486** 74860B (International Society for Optics and Photonics).

120 Ali, E. M., Edwards, H. G., Hargreaves, M. D. & Scowen, I. J. Raman spectroscopic investigation of cocaine hydrochloride on human nail in a forensic context. *Anal Bioanal Chem* **390**, 1159-1166 (2008).

121 Vergote, G. J., Vervaet, C., Remon, J. P., Haemers, T. & Verpoort, F. Near-infrared FT-Raman spectroscopy as a rapid analytical tool for the determination of diltiazem hydrochloride in tablets. *Eur. J. Pharm. Sci.* **16**, 63-67 (2002).

122 Eliasson, C. & Matousek, P. Noninvasive authentication of pharmaceutical products through packaging using spatially offset Raman spectroscopy. *Anal Chem* **79**, 1696-1701 (2007).

123 Lohr, D. *et al.* Non-destructive determination of carbohydrate reserves in leaves of ornamental cuttings by near-infrared spectroscopy (NIRS) as a key indicator for quality assessments. *Biosys Eng* **158**, 51-63 (2017).

124 Heys, K. A., Shore, R. F., Pereira, M. G. & Martin, F. L. Levels of Organochlorine Pesticides Are Associated with Amyloid Aggregation in Apex Avian Brains. *Environ Sci Technol* **51**, 8672-8681 (2017).

125 Comino, F., Aranda, V., García-Ruiz, R. & Domínguez-Vidal, A. Infrared spectroscopy as a tool for the assessment of soil biological quality in agricultural soils under contrasting management practices. *Ecol Indicators* **87**, 117-126 (2018).

126 Eliasson, C., Macleod, N. & Matousek, P. Noninvasive detection of concealed liquid explosives using Raman spectroscopy. *Anal Chem* **79**, 8185-8189 (2007).

127 Liu, H.-B., Zhong, H., Karpowicz, N., Chen, Y. & Zhang, X.-C. Terahertz spectroscopy and imaging for defense and security applications. *Proc IEEE* **95**, 1514-1527 (2007).

128 Golightly, R. S., Doering, W. E. & Natan, M. J. Surface-enhanced Raman spectroscopy and homeland security: a perfect match? (ACS Nano, 2009).

129 Sattlecker, M., Stone, N., Smith, J. & Bessant, C. Assessment of robustness and transferability of classification models built for cancer diagnostics using Raman spectroscopy. *J Raman Spectrosc* **42**, 897-903 (2011).

130 Isabelle, M. *et al.* Multi-centre Raman spectral mapping of oesophageal cancer tissues: a study to assess system transferability. *Faraday Discuss* **187**, 87-103 (2016).

131 Guo, S. *et al.* Towards an improvement of model transferability for Raman spectroscopy in biological applications. *Vib Spectrosc* **91**, 111-118 (2017).

132 Luo, X. *et al.* Calibration transfer across near infrared spectrometers for measuring hematocrit in the blood of grazing cattle. *Journal of Near Infrared Spectroscopy* **25**, 15-25 (2017).

133 Vaughan, A. A. *et al.* Liquid chromatography–mass spectrometry calibration transfer and metabolomics data fusion. *Anal Chem* **84**, 9848-9857 (2012).

134 Rodriguez, J. D., Westenberger, B. J., Buhse, L. F. & Kauffman, J. F. Standardization of Raman spectra for transfer of spectral libraries across different instruments. *Analyst* **136**, 4232-4240 (2011).

135 de Andrade, E. W., de Lelis Medeiros de Morais, C., Lopes da Costa, F. S., de Lima, G. & Michell, K. A Multivariate Control Chart Approach for Calibration Transfer between NIR Spectrometers for Simultaneous Determination of Rifampicin and Isoniazid in Pharmaceutical Formulation. *Curr Anal Chem* **14**, 488-494 (2018).

136 Yu, B., Ji, H. & Kang, Y. Standardization of near infrared spectra based on multi-task learning. *Spectroscopy Letters* **49**, 23-29 (2016).

137 Ni, L., Han, M., Luan, S. & Zhang, L. Screening wavelengths with consistent and stable signals to realize calibration model transfer of near infrared spectra. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* (2018).

138 Hu, R. & Xia, J. Calibration transfer of near infrared spectroscopy based on DS algorithm. in *Electric Information and Control Engineering (ICEICE), 2011 International Conference on.* 3062-3065 (IEEE).

139 Forina, M. *et al.* Transfer of calibration function in near-infrared spectroscopy. *Chemom Intellig Lab Syst* **27**, 189-203 (1995).

140 Xiao, H. *et al.* Comparison of benchtop Fourier-transform (FT) and portable grating scanning spectrometers for determination of total soluble solid contents in single grape berry (Vitis vinifera L.) and calibration transfer. *Sensors* **17**, 2693 (2017).

141 Yahaya, O., MatJafri, M., Aziz, A. & Omar, A. Visible spectroscopy calibration transfer model in determining pH of Sala mangoes. *Journal of Instrumentation* **10**, T05002 (2015).

142 Bin, J., Li, X., Fan, W., Zhou, J.-h. & Wang, C.-w. Calibration transfer of near-infrared spectroscopy by canonical correlation analysis coupled with wavelet transform. *Analyst* **142**, 2229-2238 (2017).

143 Monakhova, Y. B. & Diehl, B. W. Transfer of multivariate regression models between high-resolution NMR instruments: application to authenticity control of sunflower lecithin. *Magnetic Resonance in Chemistry* **54**, 712-717 (2016).

144 Zuo, Q., Xiong, S., Chen, Z.-P., Chen, Y. & Yu, R.-Q. A novel calibration strategy based on background correction for quantitative circular dichroism spectroscopy. *Talanta* **174**, 320-324 (2017).

145 Koehler IV, F. W., Small, G. W., Combs, R. J., Knapp, R. B. & Kroutil, R. T. Calibration transfer algorithm for automated qualitative analysis by passive Fourier transform infrared spectrometry. *Anal Chem* **72**, 1690-1698 (2000).

146 Rodrigues, R. R. *et al.* Evaluation of calibration transfer methods using the ATR-FTIR technique to predict density of crude oil. *Chemom Intellig Lab Syst* **166**, 7-13 (2017).

147 Wang, Y., Veltkamp, D. J. & Kowalski, B. R. Multivariate instrument standardization. *Anal Chem* **63**, 2750-2756 (1991).

148 Brouckaert, D., Uyttersprot, J.-S., Broeckx, W. & De Beer, T. Calibration transfer of a Raman spectroscopic quantification method for the assessment of liquid detergent compositions from at-line laboratory to in-line industrial scale. *Talanta* **179**, 386-392 (2018).

149 Andrade, E. V., Morais, C. d. L. M., Costa, F. S. L. & Lima, K. M. G. A Multivariate Control Chart Approach for Calibration Transfer between NIR Spectrometers for Simultaneous Determination of Rifampicin and Isoniazid in Pharmaceutical Formulation. *Curr Anal Chem* **14**, 1-7 (2018).

150 Zamora-Rojas, E., Pérez-Marín, D., De Pedro-Sanz, E., Guerrero-Ginel, J. & Garrido-Varo, A. Handheld NIRS analysis for routine meat quality control: Database transfer from at-line instruments. *Chemom Intellig Lab Syst* **114**, 30-35 (2012).

151 Panchuk, V., Kirsanov, D., Oleneva, E., Semenov, V. & Legin, A. Calibration transfer between different analytical methods. *Talanta* **170**, 457-463 (2017).

152 de Morais, C. d. L. M. & de Lima, K. M. G. Determination and analytical validation of creatinine content in serum using image analysis by multivariate transfer calibration procedures. *Anal Methods* **7**, 6904-6910 (2015).

153 Khaydukova, M. *et al.* Multivariate calibration transfer between two different types of multisensor systems. *Sensors Actuators B: Chem* **246**, 994-1000 (2017).

154 Barreiro, P. *et al.* Calibration Transfer Between Portable and Laboratory NIR Spectrophotometers. *Acta Hortic* (2008).

155 Sulub, Y., LoBrutto, R., Vivilecchia, R. & Wabuyele, B. W. Content uniformity determination of pharmaceutical tablets using five near-infrared reflectance spectrometers: a process analytical technology (PAT) approach using robust multivariate calibration transfer algorithms. *Anal Chim Acta* **611**, 143-150 (2008).

156 Zhang, L., Small, G. W. & Arnold, M. A. Multivariate calibration standardization across instruments for the determination of glucose by Fourier transform near-infrared spectrometry. *Anal Chem* **75**, 5905-5915 (2003).

157 Martens, H., Høy, M., Wise, B. M., Bro, R. & Brockhoff, P. B. Pre ‐ whitening of data by covariance ‐ weighted pre ‐ processing. *J Chemom* **17**, 153-165 (2003).

158 Feudale, R. N. *et al.* Transfer of multivariate calibration models: a review. *Chemom Intellig Lab Syst* **64**, 181-192 (2002).

159 Woody, N. A., Feudale, R. N., Myles, A. J. & Brown, S. D. Transfer of multivariate calibrations between four near-infrared spectrometers using orthogonal signal correction. *Anal Chem* **76**, 2595-2600 (2004).

160 Greensill, C., Wolfs, P., Spiegelman, C. & Walsh, K. Calibration transfer between PDA-based NIR spectrometers in the NIR assessment of melon soluble solids content. *Appl Spectrosc* **55**, 647-653 (2001).

161 Sjöblom, J., Svensson, O., Josefson, M., Kullberg, H. & Wold, S. An evaluation of orthogonal signal correction applied to calibration transfer of near infrared spectra. *Chemom Intellig Lab Syst* **44**, 229-244 (1998).

162 Andrews, D. T. & Wentzell, P. D. Applications of maximum likelihood principal component analysis: incomplete data sets and calibration transfer. *Anal Chim Acta* **350**, 341-352 (1997).

163 Bouveresse, E., Massart, D. & Dardenne, P. Calibration transfer across near-infrared spectrometric instruments using Shenk's algorithm: effects of different standardisation samples. *Anal Chim Acta* **297**, 405-416 (1994).

164 Shenk, J. S. & Westerhaus, M. O. Populations structuring of near infrared spectra and modified partial least squares regression. *Crop Sci.* **31**, 1548-1555 (1991).

165 Paatero, P. & Tapper, U. Positive matrix factorization: A non - negative factor model with optimal utilization of error estimates of data values. *Environmetrics* **5**, 111-126 (1994).

166 Xie, Y. & Hopke, P. K. Calibration transfer as a data reconstruction problem. *Anal Chim Acta* **384**, 193-205 (1999).

167 Goodacre, R. *et al.* On mass spectrometer instrument standardization and interlaboratory calibration transfer using neural networks. *Anal Chim Acta* **348**, 511-532 (1997).

168 Chen, W.-R., Bin, J., Lu, H.-M., Zhang, Z.-M. & Liang, Y.-Z. Calibration transfer via an extreme learning machine auto-encoder. *Analyst* **141**, 1973-1980 (2016).

169 Hu, Y., Peng, S., Bi, Y. & Tang, L. Calibration transfer based on maximum margin criterion for qualitative analysis using Fourier transform infrared spectroscopy. *Analyst* **137**, 5913-5918 (2012).

170 Fan, W., Liang, Y., Yuan, D. & Wang, J. Calibration model transfer for near-infrared spectra based on canonical correlation analysis. *Anal Chim Acta* **623**, 22-29 (2008).

171 Wang, Z., Dean, T. & Kowalski, B. R. Additive background correction in multivariate instrument standardization. *Anal Chem* **67**, 2379-2385 (1995).

172 Kennard, R. W. & Stone, L. A. Computer Aided Design of Experiments. *Technometrics* **11**, 137-148 (1969).

173 Palonpon, A. F. *et al.* Raman and SERS microscopy for molecular imaging of live cells. *Nat Protoc* **8**, 677 (2013).

174 Witze, E. S., Old, W. M., Resing, K. A. & Ahn, N. G. Mapping protein post-translational modifications with mass spectrometry. *Nat Methods* **4**, 798 (2007).

175 Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198 (2003).

176 Pence, I. & Mahadevan-Jansen, A. Clinical instrumentation and applications of Raman spectroscopy. *Chem Soc Rev* **45**, 1958-1979 (2016).

177 Ibrahim, O. *et al.* Improved protocols for pre-processing Raman spectra of formalin fixed paraffin preserved tissue sections. *Anal Methods* **9**, 4709-4717 (2017).

178 Tfayli, A. *et al.* Digital dewaxing of Raman signals: discrimination between nevi and melanoma spectra obtained from paraffin-embedded skin biopsies. *Appl Spectrosc* **63**, 564-570 (2009).

179 Byrne, H. J., Knief, P., Keating, M. E. & Bonnier, F. Spectral pre and post processing for infrared and Raman spectroscopy of biological tissues and cells. *Chem Soc Rev* **45**, 1865-1878 (2016).

180 Meade, A. D. *et al.* Studies of chemical fixation effects in human cell lines using Raman microspectroscopy. *Anal Bioanal Chem* **396**, 1781-1791 (2010).

181 Baker, M. J. *et al.* Developing and understanding biofluid vibrational spectroscopy: a critical review. *Chem Soc Rev* **45**, 1803-1818 (2016).

182 Bonifacio, A., Cervo, S. & Sergo, V. Label-free surface-enhanced Raman spectroscopy of biofluids: fundamental aspects and diagnostic applications. *Anal Bioanal Chem* **407**, 8265-8277 (2015).

183 Mitchell, A. L., Gajjar, K. B., Theophilou, G., Martin, F. L. & Martin-Hirsch, P. L. Vibrational spectroscopy of biofluids for disease screening or diagnosis: translation from the laboratory to a clinical setting. *J Biophotonics* **7**, 153-165 (2014).

184 Lovergne, L. *et al.* Biofluid infrared spectro-diagnostics: pre-analytical considerations for clinical applications. *Faraday Discuss* **187**, 521-537 (2016).

185 Bonifacio, A. *et al.* Surface-enhanced Raman spectroscopy of blood plasma and serum using Ag and Au nanoparticles: a systematic study. *Anal Bioanal Chem* **406**, 2355-2365 (2014).

186 Paraskevaidi, M., Martin-Hirsch, P. L. & Martin, F. L. ATR-FTIR Spectroscopy Tools for Medical Diagnosis and Disease Investigation. *Springer* (2018).

187 Mitchell, B. L., Yasui, Y., Li, C. I., Fitzpatrick, A. L. & Lampe, P. D. Impact of freeze-thaw cycles and storage time on plasma samples used in mass spectrometry based biomarker discovery projects. *Cancer Inform* **1** (2005).

188 Glassford, S. E., Byrne, B. & Kazarian, S. G. Recent applications of ATR FTIR spectroscopy and imaging to proteins. *Biochim Biophys Acta* **1834**, 2849-2858 (2013).

189 Kundu, J., Le, F., Nordlander, P. & Halas, N. J. Surface enhanced infrared absorption (SEIRA) spectroscopy on nanoshell aggregate substrates. *Chem Phys Lett* **452**, 115-119 (2008).

190 Jones, S., Carley, S. & Harrison, M. An introduction to power and sample size estimation. *Emergency Medicine Journal* **20**, 453-458 (2003).

191 Beebe, K. R., Pell, R. J. & Seasholtz, M. B. *Chemometrics: a practical guide*. Vol. **4** (Wiley New York, 1998).

192 Pavia, D. L., Lampman, G. M., Kriz, G. S. & Vyvyan, J. A. *Introduction to spectroscopy*. (Cengage Learning, 2008).

193 Savitzky, A. & Golay, M. J. Smoothing and differentiation of data by simplified least squares procedures. *Anal Chem* **36**, 1627-1639 (1964).

194 Geladi, P., MacDougall, D. & Martens, H. Linearization and scatter-correction for near-infrared reflectance spectra of meat. *Appl Spectrosc* **39**, 491-500 (1985).

195 Barnes, R., Dhanoa, M. S. & Lister, S. J. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Appl Spectrosc* **43**, 772-777 (1989).

196 Brereton, R. G. *Chemometrics: data analysis for the laboratory and chemical plant*. (John Wiley & Sons, 2003).

197 Hastie, T., Tibshirani, R. & Friedman, J. The elements of statistical learning 2nd edition (New York: Springer, 2009).

198 Bro, R. & Smilde, A. K. Principal component analysis. *Anal Methods* **6**, 2812-2831 (2014).

199 Martin, F. L. *et al.* Identifying variables responsible for clustering in discriminant analysis of data from infrared microspectroscopy of a biological sample. *J Comput Biol* **14**, 1176-1184 (2007).

200 Martens, H. & Martens, M. Modified Jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression (PLSR). *Food quality and preference* **11**, 5-16 (2000).

201 Rousseeuw, P. J. & Hubert, M. Robust statistics for outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **1**, 73-79 (2011).

202 Jiang, F., Liu, G., Du, J. & Sui, Y. Initialization of K-modes clustering using outlier detection techniques. *Inf Sci* **332**, 167-183 (2016).

203 Domingues, R., Filippone, M., Michiardi, P. & Zouaoui, J. A comparative evaluation of outlier detection algorithms: Experiments and analyses. *Pattern Recognit* **74**, 406-421 (2018).

204 Bakeev, K. A. *Process analytical technology: spectroscopic tools and implementation strategies for the chemical and pharmaceutical industries*. (John Wiley & Sons, 2010).

205 Kuligowski, J., Quintás, G., Herwig, C. & Lendl, B. A rapid method for the differentiation of yeast cells grown under carbon and nitrogen-limited conditions by means of partial least squares discriminant analysis employing infrared micro-spectroscopic data of entire yeast cells. *Talanta* **99**, 566-573 (2012).

206 Morais, C. L. & Lima, K. M. Comparing unfolded and two-dimensional discriminant analysis and support vector machines for classification of EEM data. *Chemom Intellig Lab Syst* (2017).

207 Dixon, S. J. & Brereton, R. G. Comparison of performance of five common classifiers represented as boundary methods: Euclidean Distance to Centroids, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Learning Vector Quantization and Support Vector Machines, as dependent on data structure. *Chemom Intellig Lab Syst* **95**, 1-17 (2009).

208 Brereton, R. G. & Lloyd, G. R. Partial least squares discriminant analysis: taking the magic away. *J Chemom* **28**, 213-225 (2014).

209 Cover, T. & Hart, P. Nearest neighbor pattern classification. *IEEE Trans Inf Theory* **13**, 21-27 (1967).

210 Cortes, C. & Vapnik, V. Support-vector networks. *Mach Learn* **20**, 273-297 (1995).

211 Abraham, A. Artificial neural networks. *handbook of measuring system design* (2005).

212 Fawagreh, K., Gaber, M. M. & Elyan, E. Random forests: from early developments to recent advancements. *Systems Science & Control Engineering: An Open Access Journal* **2**, 602-609 (2014).

213 LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *nature* **521**, 436 (2015).

214 Seasholtz, M. B. & Kowalski, B. The parsimony principle applied to multivariate calibration. *Anal Chim Acta* **277**, 165-177 (1993).

215 Morais, C. L. & Lima, K. M. Principal Component Analysis with Linear and Quadratic Discriminant Analysis for Identification of Cancer Samples Based on Mass Spectrometry. *J Braz Chem Soc*, 31 (2017).

216 Hibbert, D. B. Vocabulary of concepts and terms in chemometrics (IUPAC Recommendations 2016). *Pure and Applied Chemistry* **88**, 407-443 (2016).

217 McCall, J. Genetic algorithms for modelling and optimisation. *J Comput Appl Math* **184**, 205-222 (2005).

218 Soares, S. F. C., Gomes, A. A., Araujo, M. C. U., Galvão Filho, A. R. & Galvão, R. K. H. The successive projections algorithm. *Trends Anal Chem* **42**, 84-98 (2013).

219 Kamandar, M. & Ghassemian, H. Maximum relevance, minimum redundancy feature extraction for hyperspectral images. in *Electrical Engineering (ICEE), 2010 18th Iranian Conference on.* 254-259 (IEEE).

Supplementary Material 1

Additional results from pilot study

A. Effect of different instruments

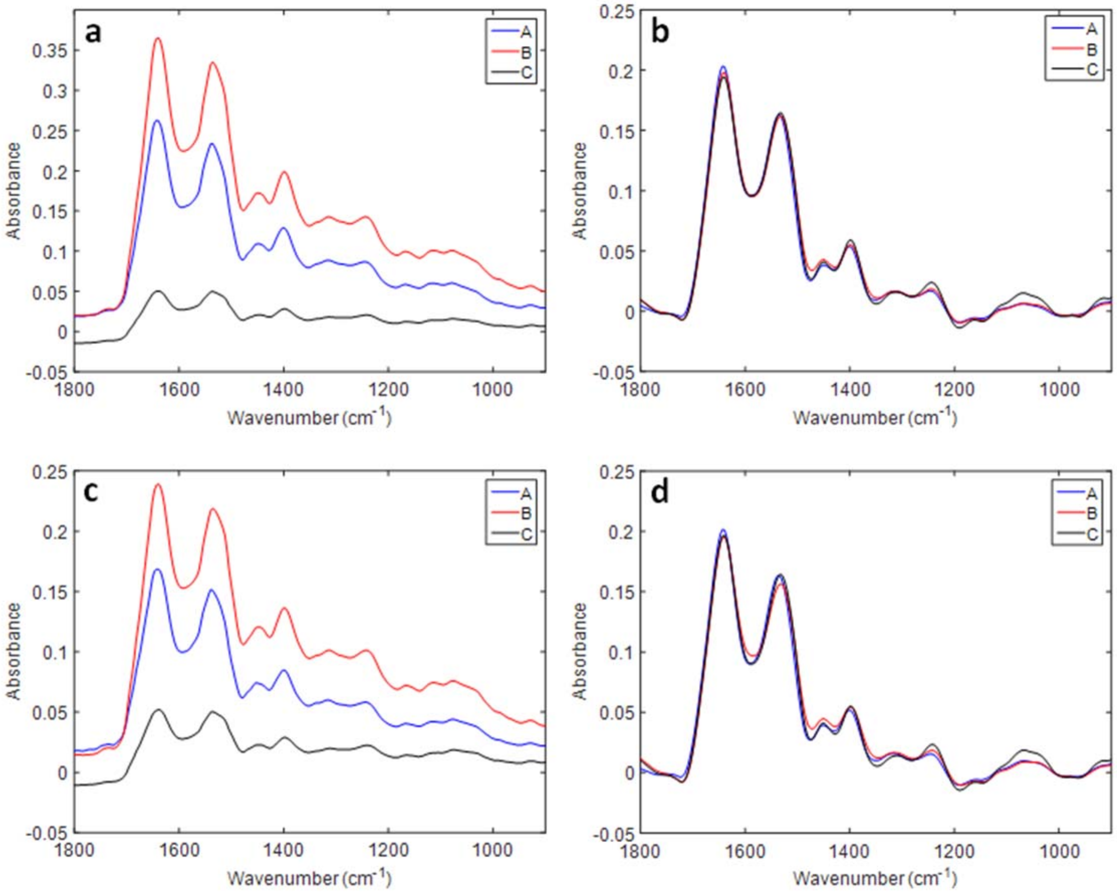


Figure S1. Average (a) raw and (b) pre-processed spectra for healthy controls samples; average (c) raw and (d) pre-processed spectra for cancer samples across three different instruments (A, B and C).

A. Effect of different instruments

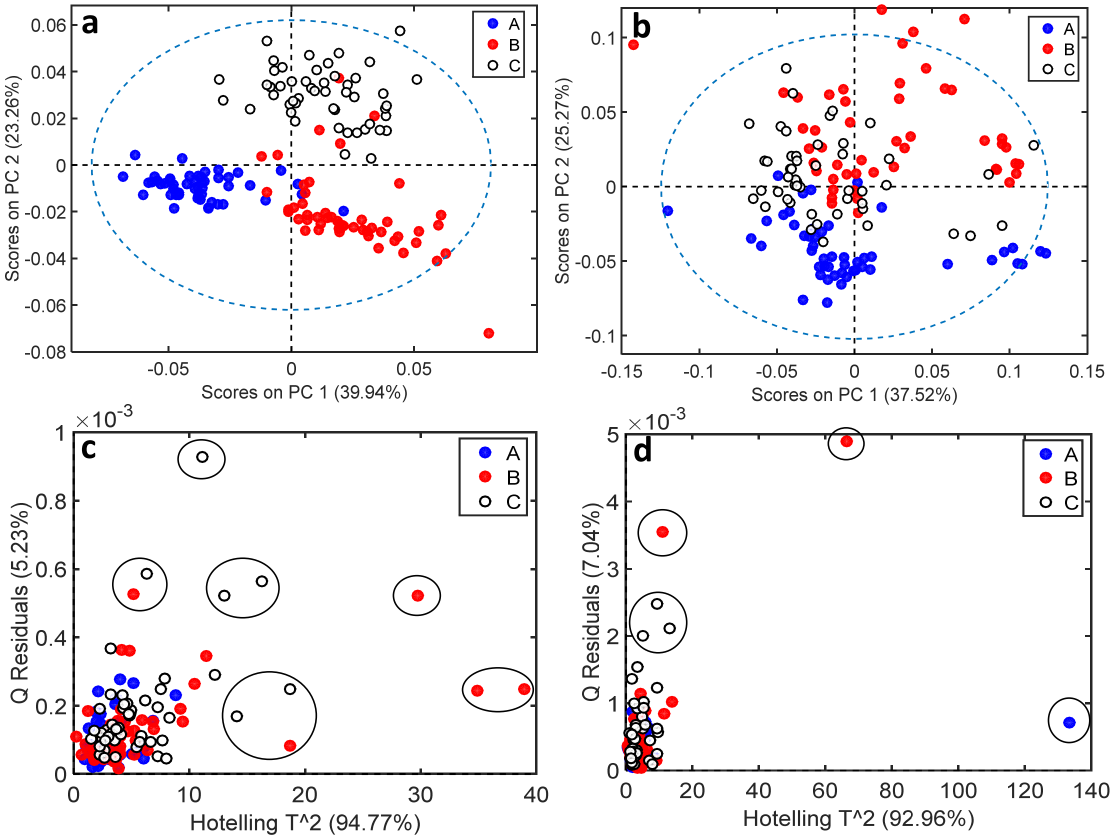


Figure S2. (a) PCA scores for healthy control samples according to the instrument used for spectra acquisition (A, B and C); (b) PCA scores for cancer samples according to the instrument used for spectra acquisition (A, B and C); (c) Hotelling T^2 *versus* Q residual test for healthy control samples according to the instrument used for spectra acquisition (A, B and C) based on a PCA using 5 PCs (94.77% cumulative variance); (d) Hotelling T^2 *versus* Q residual test for cancer samples according to the instrument used for spectra acquisition (A, B and C) based on a PCA using 5 PCs (92.96% cumulative variance). Circled samples in (c) and (d) indicate outliers removed. Confidence ellipse was 95%, depicted in blue in (a) and (b).

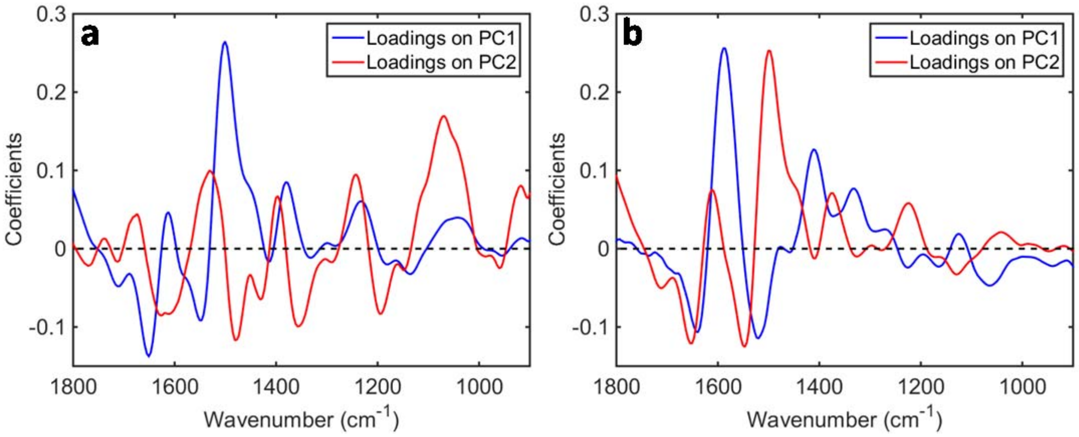


Figure S3. (a) PCA loadings for healthy control samples measured in different instruments (A, B and C); (b) PCA loadings for cancer samples measured in different instruments (A, B and C).

B. Effect of different operators

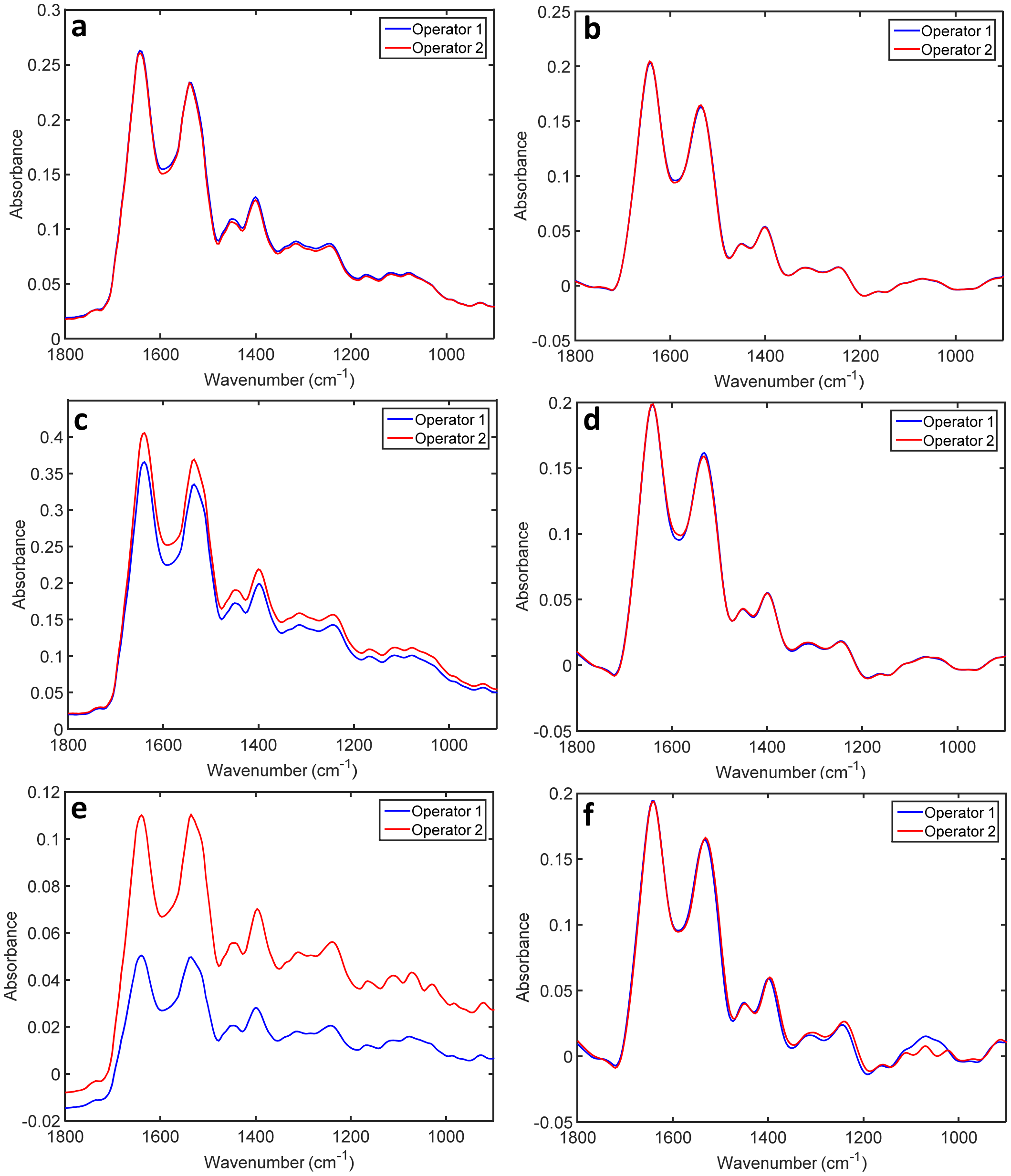


Figure S4. Average (a) raw and (b) pre-processed spectra for healthy control samples acquired with instrument A depending on the operator; average (c) raw and (d) pre-processed spectra for healthy control samples acquired with instrument B depending on the operator; average (e) raw and (f) pre-processed spectra for healthy control samples acquired with instrument C varying the operator.

B. Effect of different operators

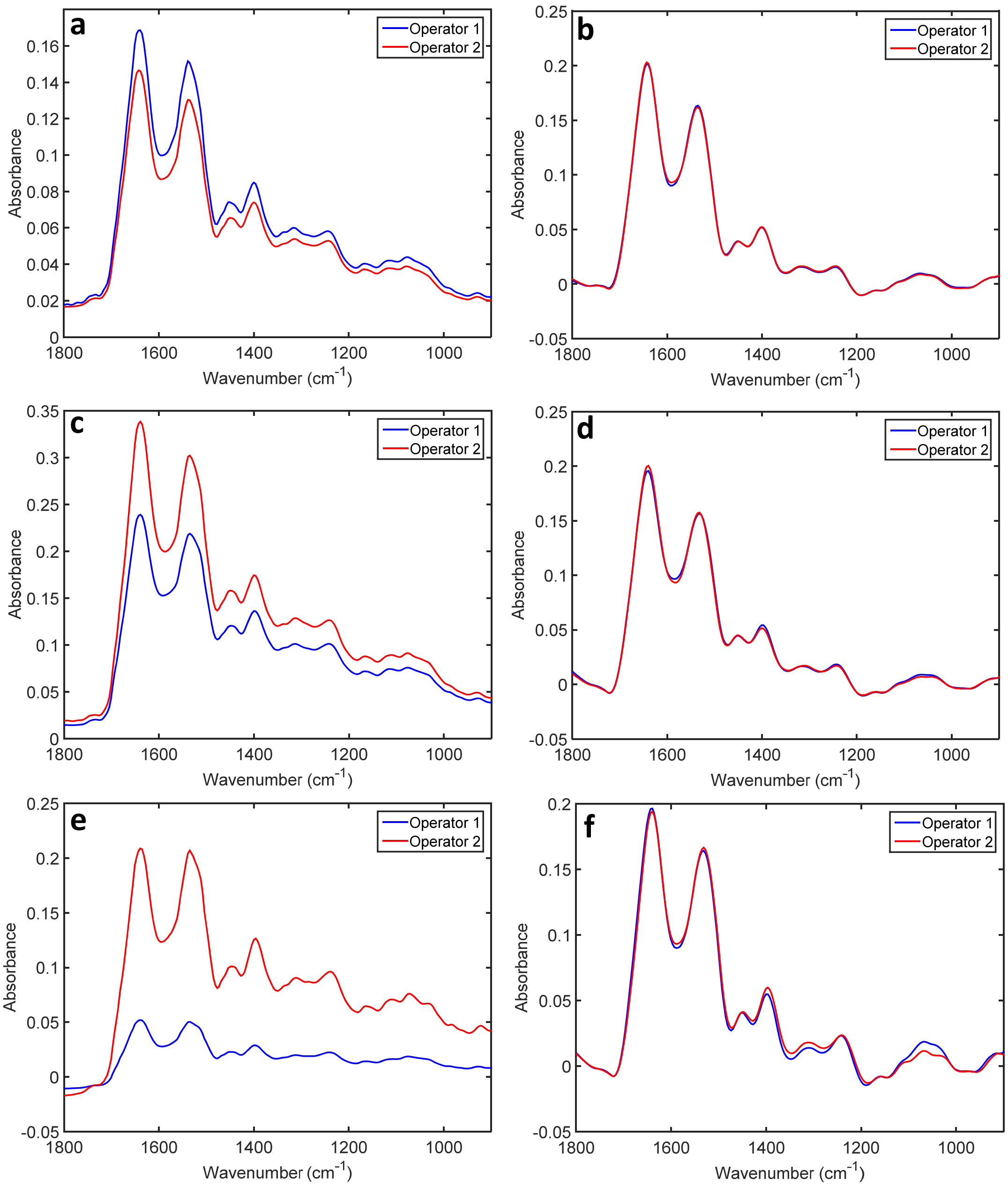


Figure S5. Average (a) raw and (b) pre-processed spectra for cancer samples acquired with instrument A depending on the operator; average (c) raw and (d) pre-processed spectra for cancer samples acquired with instrument B depending on the operator; average (e) raw and (f) pre-processed spectra for cancer samples acquired with instrument C depending on the operator.

B. Effect of different operators

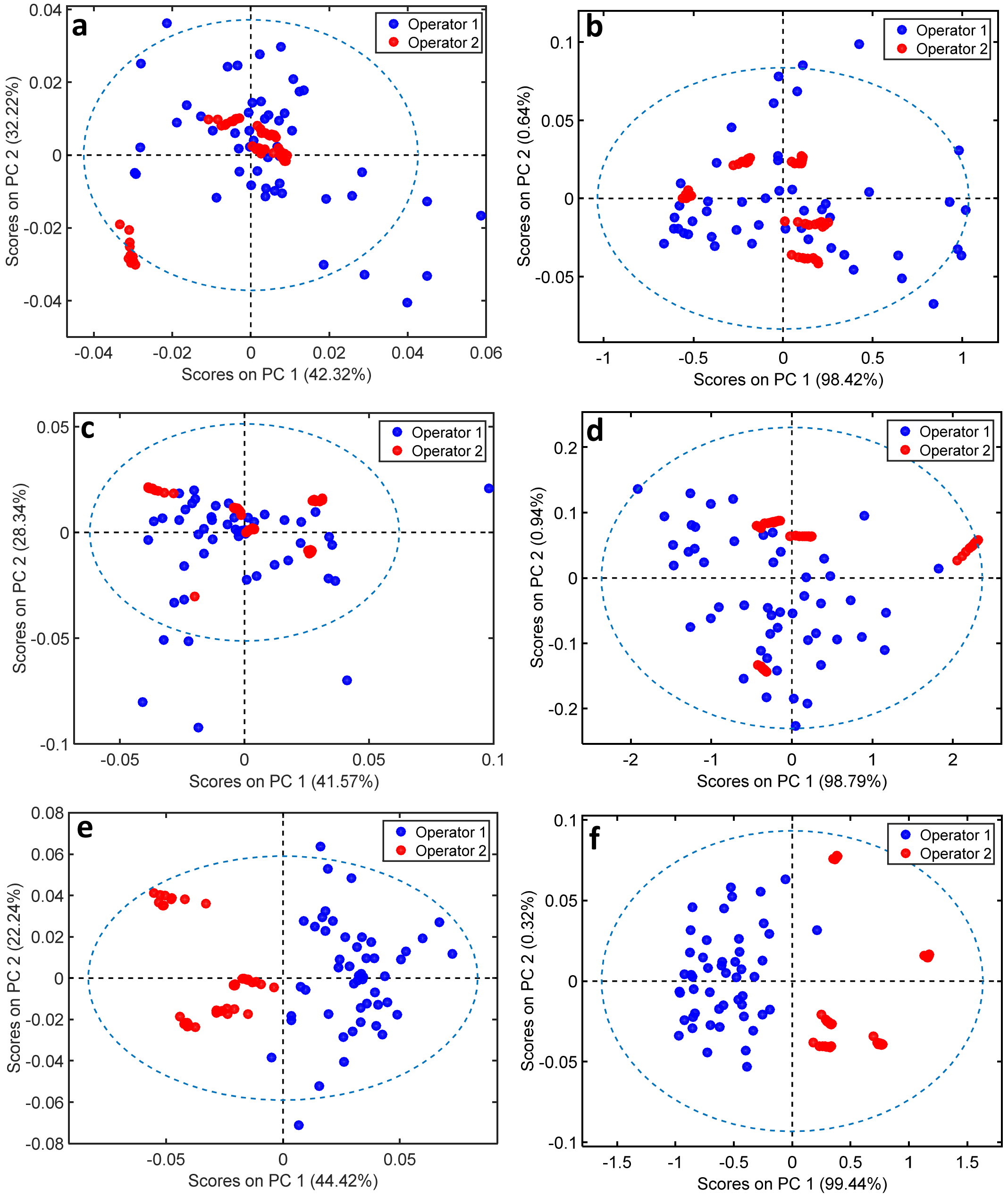


Figure S6. PCA scores for (a) healthy control and (b) cancer samples acquired with instrument A depending on the operator; PCA scores for (c) healthy control and (d) cancer samples acquired with instrument B depending on the operator; PCA scores for (e) healthy control and (f) cancer samples acquired with instrument C depending on the operator. Confidence ellipse was 95%, depicted in blue

C. Effect of different instruments and operators

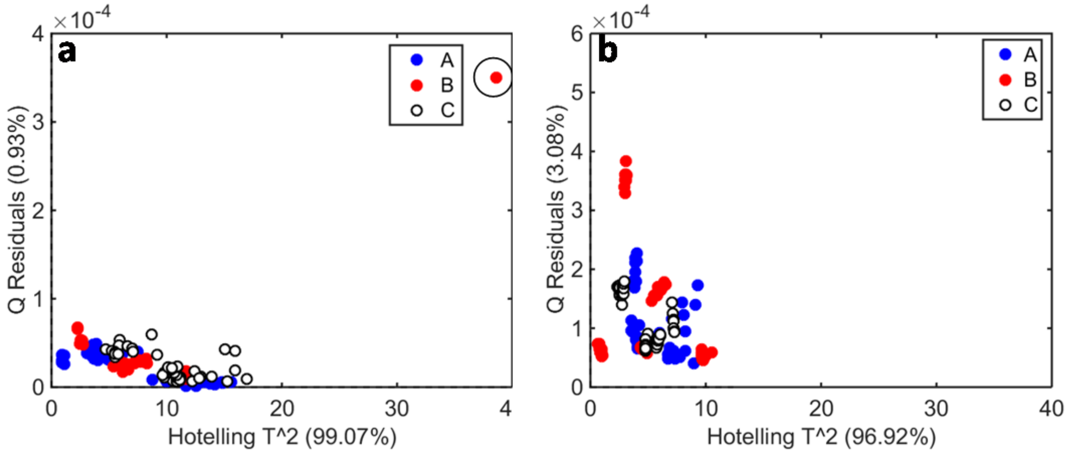


Figure S7. (a) Hotelling T^2 *versus* Q residual test based on a PCA using 8 PCs (99.07% cumulative variance) for healthy control samples depending on the instrument for spectra acquisition (A, B and C) used by Operator 2; (b) Hotelling T^2 *versus* Q residual test based on a PCA using 5 PCs (96.92% cumulative variance) for cancer samples depending on the instrument for spectra acquisition (A, B and C) used by Operator 2. Circled sample in a) indicates an outlier removed. The Hotelling T^2 *versus* Q residual test for Operator 1 is depicted in Fig. S2c-d.

D. Effect of different classes

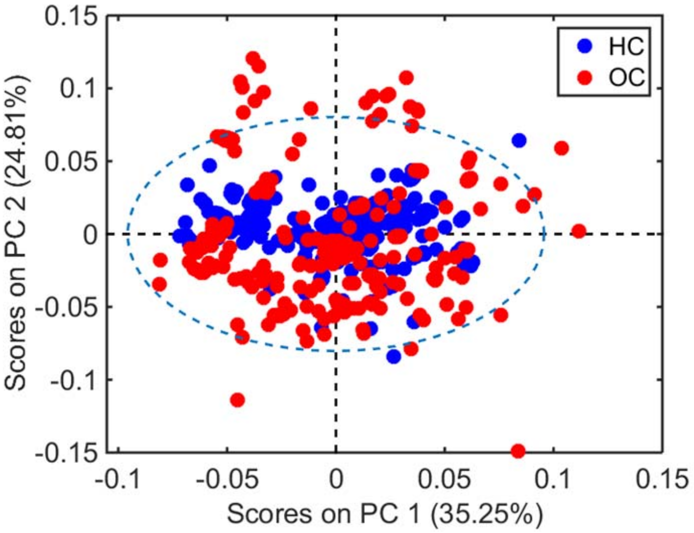


Figure S8. PCA scores for healthy controls (HC) and ovarian cancer (OC) samples based on the spectra acquired by both operators (1 and 2) and by all instruments (A, B and C). Confidence ellipse at a 95% confidence level is depicted in blue

Supplementary Method 1

Protocol for outliers detection

A. Outlier detection using Hotelling T^2 *versus* Q residuals test

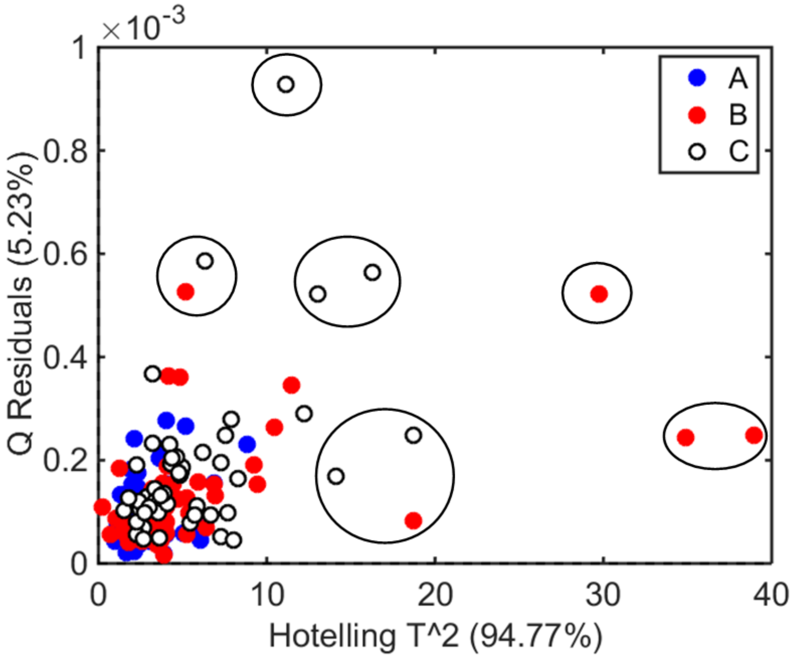
1st step: Build a PCA model.

2nd step: Calculate Hotelling T^2 and Q residuals.

3rd step: Plot Hotelling T^2 *versus* Q residuals

4th step: Select the samples which are most distant to the plot origin (0,0) and remove them one at a time from the data set. This procedure can be performed manually after visual inspection or automatically by algorithms.

Figure S1. Hotelling T^2 *versus* Q residuals for healthy control samples (blood plasma) varying the instrument for spectra acquisition (A, B and C). PCA performed with 5 PCs (94.77% cumulative variance). Circled samples indicate outliers removed.



B. Automatic outlier detection using MATLAB®

Algorithm link to download:

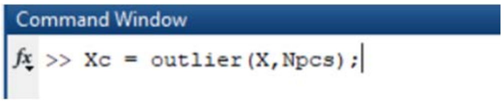
https://doi.org/10.6084/m9.figshare.7066613.v1

1st step: Add the .m files within the file downloaded to the path.

2nd step: Load the spectral data into MATLAB and organize all the spectra into a single matrix "X" containing each spectrum as a row.

3rd step: Perform an initial PCA model to determine the number of principal components (PCs) to work with.

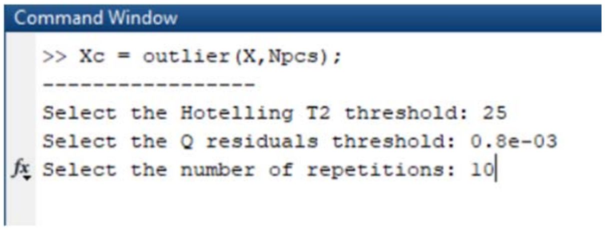
4th step: Run the algorithm as follows:



```
>> Xc = outlier(X,Npcs);
```

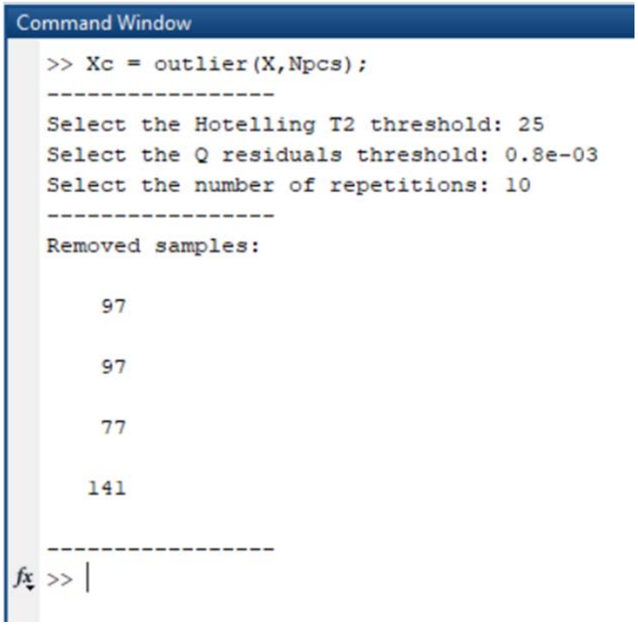
where "Xc" is the spectral matrix without outliers, "X" is the input spectral data, and "Npcs" the number of PCs for PCA.

5th step: Input optimization parameters:



```
>> Xc = outlier(X,Npcs);
------------------
Select the Hotelling T2 threshold: 25
Select the Q residuals threshold: 0.8e-03
Select the number of repetitions: 10
```

In this case, the algorithm will perform a PCA model 10 times removing one sample at a time that follows one of these criteria: Hotteling $T^2 > 25$ or Q residuals > $0.8x10^{-3}$. Then, these samples are automatic excluded from the new dataset (Xc). The list of excluded samples is also displayed in MATLAB. Example:



```
>> Xc = outlier(X,Npcs);
-----------------
Select the Hotelling T2 threshold: 25
Select the Q residuals threshold: 0.8e-03
Select the number of repetitions: 10
-----------------
Removed samples:

    97

    97

    77

   141

-----------------
>> 
```