

Central Lancashire Online Knowledge (CLoK)

Title	Standardization of complex biologically derived spectrochemical datasets
Type	Article
URL	https://clock.uclan.ac.uk/id/eprint/28305/
DOI	https://doi.org/10.1038/s41596-019-0150-x
Date	2019
Citation	Morais, Camilo L M orcid iconORCID: 0000-0003-2573-787X, Paraskevaidi, Maria, Cui, Li, Fullwood, Nigel J, Isabelle, Martin, Lima, Kássio M G, Martin-Hirsch, Pierre L, Sreedhar, Hari, Trevisan, Júlio et al (2019) Standardization of complex biologically derived spectrochemical datasets. Nature Protocols, 14. pp. 1546-1577. ISSN 1754-2189
Creators	Morais, Camilo L M, Paraskevaidi, Maria, Cui, Li, Fullwood, Nigel J, Isabelle, Martin, Lima, Kássio M G, Martin-Hirsch, Pierre L, Sreedhar, Hari, Trevisan, Júlio, Walsh, Michael J, Zhang, Dayi, Zhu, Yong-Guan and Martin, Francis L

It is advisable to refer to the publisher's version if you intend to cite from the work.
<https://doi.org/10.1038/s41596-019-0150-x>

For information about Research at UCLan please go to <http://www.uclan.ac.uk/research/>

All outputs in CLoK are protected by Intellectual Property Rights law, including Copyright law. Copyright, IPR and Moral Rights for the works on this site are retained by the individual authors and/or other copyright owners. Terms and conditions for use of this material are defined in the <http://clock.uclan.ac.uk/policies/>

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 Zhangⁱ, 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*

¹To 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

Spectroscopic techniques, such as Fourier-transform infrared (FTIR) spectroscopy, are used to study the interaction of light with biological materials. This interaction forms the basis of many analytical assays used in disease screening and diagnosis, microbiological studies, forensic and environmental investigations. Advantages of spectrochemical analysis are 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 understand and reduce the effect of random spectral alterations caused by inter-individual, inter-instrument and/or inter-laboratory variations, such as variations in air humidity and CO₂ levels, and the aging of instrumental parts. Thus, it is evident that spectral standardization is crucial for the widespread adoption of these spectrochemical technologies. By using calibration transfer procedures, where the spectral response of a secondary instrument is standardized to resemble the spectral response of a primary instrument, different sources of variations can be normalized into a single model using computational-based methods, such as direct standardization (DS) and piecewise direct standardization (PDS); therefore, measurements performed under different conditions can 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, pharmaceuticals or environmental sciences¹⁻⁵. Fourier-transform infrared (FTIR) spectroscopy is one of the preferred techniques for identification of biomolecules through the study of their characteristic vibrational movements. Another commonly used approach is Raman spectroscopy, which provides complementary spectral information to IR. Raman spectroscopy exploits the inelastic scattering of light whereas IR studies light absorption. Both methods have their benefits and drawbacks. A limitation of IR, for instance, is that water generates undesired peaks at the region of interest, which can mask important biological information, and therefore extra sample preparation and/or spectral processing is necessary. On the contrary, Raman spectroscopy has an inherently weak signal and fluorescence interference, which can, however, be addressed by optimizing the experimental settings or by applying enhancement techniques to increase the Raman signal. For the purposes of this protocol we have used FTIR spectroscopy to demonstrate our standardization model.

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 the 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 bio-spectroscopy.

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)⁶, to image biosensors, such as X-rays⁷⁻¹⁰ and γ -rays¹¹⁻¹³, 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) spectroscopy¹⁴⁻¹⁷, ultraviolet (UV) or visible spectroscopy^{18,19}, fluorescence²⁰⁻²⁴, nuclear magnetic resonance (NMR) spectroscopy²⁵⁻²⁹ and ultrasound (US)^{7,30-33}.

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 methods have focused on cancer investigation with malignancies such as brain³⁴⁻³⁷, breast³⁸⁻⁴⁰, oesophagus^{41,42}, skin⁴³⁻⁴⁷, colorectal⁴⁸⁻⁵⁰, lung⁵¹⁻⁵³, ovarian⁵⁴⁻⁵⁸, endometrial^{55,59,60}, cervical⁶¹⁻⁶⁴ and prostate⁶⁵⁻⁶⁸ cancer being some of them. Non-cancerous diseases have also been examined, namely neurodegenerative disorders⁶⁹⁻⁷², HIV/AIDS⁷³, diabetes⁷⁴⁻⁷⁶, rheumatoid arthritis^{77,78}, cardiovascular diseases^{79,80}, malaria⁸¹⁻⁸³, alkaptonuria⁸⁴, cystic fibrosis⁸⁵, thalassemia⁸⁶, prenatal disorders^{87,88}, macular degeneration^{89,90}, atherosclerosis^{80,91} and osteoarthritis⁹²⁻⁹⁴.

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 present⁹⁵. However, like 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^{4,96,97}. 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^{4,98-105}.

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 industry¹⁰⁶⁻¹⁰⁹, toxicology^{2,110-112}, microbiology¹¹³⁻¹¹⁸, forensics¹¹⁹⁻¹²³, pharmacy^{2,3,124}, environmental and plant science¹²⁵⁻¹²⁷, as well as defence and security¹²⁸⁻¹³⁰. Applications of standardization algorithms vary according to the spectral technique and sample matrix studied, and have been mostly applied to Raman and Fourier-transform near-infrared (FT-NIR) spectroscopy. Table 1 summarizes some standardization applications.

1 **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	131
	Raman	Standardization of spectra acquired in 3 different sites for analysing oesophageal samples based on snap frozen tissues	132
Cells	Raman	Standardization of spectra acquired with 4 different instruments for classification of three different cultured spore species	133
Biofluids	FT-NIR	Standardization of spectra acquired with 3 different instruments for measuring haematocrit in the blood of grazing cattle	134
	LC-MS	Standardization of spectra acquired with 2 different instruments for mapping retention times and matching metabolite features of subjects diagnosed with small cell lung cancer based on blood serum and plasma samples analysis	135
Pharmaceutical materials	Raman	Standardization of spectra acquired with 5 different instruments for analysing various pharmaceutical excipients, active pharmaceutical ingredients (APIs) and common contaminants	136
	FT-NIR	Standardization of spectra acquired with 2 different instruments for simultaneous determination of rifampicin and isoniazid in pharmaceutical formulations	137
	FT-NIR	Standardization of spectra acquired with 2 different instruments for predicting content of 654 pharmaceutical tablets	138
Food	FT-NIR	Standardization of spectra acquired with 3 different instruments for predicting parameters in corn samples	138
	FT-NIR	Standardization of spectra acquired with 2 different instruments for predicting vitamin C in navel orange	139
	FT-NIR	Standardization of spectra recorded in 4 different labs for determining moisture, proteins and oil content in soy seeds	140
	FT-NIR	Standardization of spectra acquired by a benchtop and portable instrument for determining total soluble solid contents in single grape berry	141
	FT-NIR	Standardization of spectra acquired by a benchtop and portable instrument for determining total soluble solid contents in single grape berry	142
Plant	UV-Vis	Standardization of visible spectra acquired with 3 different instruments for measuring pH of Sala mango	143
	FT-NIR	Standardization of spectra acquired with 2 different instruments for predicting baicalin contents in radix scutellariae samples	139
	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	144
Cosmetic	NMR	Standardization of spectra acquired with 3 different instruments for authenticity control of sunflower lecithin	145
	CD spectroscopy	Standardization of spectra acquired between standard and real-world samples for determining Pb ²⁺ in cosmetic samples	146
Inorganic substances	FT-IR	Standardization of interferogram spectra acquired with 2 instruments for classifying acetone and SF ₆ samples	147
Fuel	FT-IR	Standardization of spectra acquired with 2 different instruments for predicting density of crude oil samples	148

2

Model transferability

Transferability models have been previously developed, however this is still an under-investigated field, especially for biomedical applications. These models use computer-based methods to standardize spectral data generated across different experimental settings (*e.g.*, different instruments, operators or laboratories). 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 of the estimated group categories (*e.g.*, whether the sample is classified as healthy or cancerous) 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^{149,150}, direct standardization (DS)¹⁵¹⁻¹⁵⁵, piecewise direct standardization (PDS)^{149,156-158}, piecewise linear discriminant analysis (PLDA)¹⁴⁷, guided model reoptimization (GMR)¹⁵⁸, back-propagation neural network (BNN)¹⁴⁷, generalized least squares weighting (GLSW)¹⁵⁹, model updating (MU)^{160,161}, orthogonal signal correction (OSC)^{162,163}, orthogonal projections to latent structures (OPLS)¹⁴⁸, wavelet hybrid direct standardization (WHDS)¹⁵⁷, maximum likelihood PCA (MLPCA)¹⁶⁴, Shenk and Westerhaus method (SW)^{165,166}, positive matrix factorization (PMF)^{167,168}, artificial neural networks (ANN) drift correction¹⁶⁹, transfer *via* extreme learning machine auto-encoder method (TEAM)¹⁷⁰,

calibration transfer based on the maximum margin criterion (CTMMC)¹⁷¹, calibration transfer based on canonical correlation analysis (CTCCA)¹⁷² and calibration methods, such as wavenumber offset correction, instrument response correction and baseline correction¹³². In this protocol, we use direct standardization (DS) and piecewise direct standardization (PDS), because they are the most common methods for spectral standardization.

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 from the same sample measured by two different instruments¹⁴⁹. 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)¹⁵¹. This is performed based on a linear relationship between the data acquired under different circumstances¹⁶⁰:

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

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

The transformation matrix $\mathbf{F}$ is estimated in a least-squares sense by¹⁷³:

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

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

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

in which T stands for the matrix transpose operation.

Then, when samples are measured under the secondary system, the signals generated $\mathbf{X}$ are transformed to resemble the primary system response by¹⁶⁰:

$$\hat{\mathbf{X}}^T = \mathbf{X}^T \mathbf{F} \quad (04)$$

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

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 method¹⁷³, in which the transformation matrix described in Eq. 02 is calculated with a background correction factor ($\mathbf{F}_b$) and an additive background correction vector $\mathbf{b}_s$ as follows:

$$\mathbf{S}_1 = \mathbf{S}_2 \mathbf{F}_b + \mathbf{1} \mathbf{b}_s^T \quad (05)$$

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

$$\mathbf{b}_s = \mathbf{s}_{1m} - \mathbf{F}_b^T \mathbf{s}_{2m} \quad (06)$$

in which $\mathbf{s}_{1m}$ is the mean vector of $\mathbf{S}_1$ and $\mathbf{s}_{2m}$ 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’ spectra from the primary system ($\mathbf{S}_1$) that will be used to transform the signal obtained using the secondary system ($\mathbf{S}_2$). The transfer samples are obtained from a same cohort of samples (*e.g.*, plasma samples) measured in the two instruments (primary and secondary systems). Usually, the procedure for selecting transfer samples is based on sample selection techniques, such as Kennard-Stone (KS) algorithm¹⁷⁴ or leverage¹⁴⁹. 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 an ill-

conditioned problem, because the number of variables (*e.g.*, wavenumber) 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 zero¹⁴⁹. 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 reduction¹⁴⁹. 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. In this protocol, 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

Any study using vibrational spectroscopy, follows these general steps: 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 mass spectrometry (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^{4,98-100,105,175-177}, suitability of substrates^{4,99}, instrumentation settings^{4,98,99,105,175,177,178} or available software packages (Table 2) and manufacturers^{4,99}.

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.	Commercial

		Contains standardization routines using interpolation, bias and slope correction, and PDS.	
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)⁴, which can alter tissue structures and be inefficient for the complete wax removal¹⁷⁹, or by applying chemometrics (*e.g.*, digital dewaxing)^{180,181}, which keeps the tissue intact but might introduce artefacts due to over- or under-estimation of the wax contribution¹⁷⁹.

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 peak¹⁸². 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 fluid¹⁸³⁻¹⁸⁵. An initial centrifugation step should precede analysis 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 peaks¹⁸⁶⁻¹⁸⁸. Careful planning of experiments as well as consistence throughout a study are of great importance for the generation of robust results. Care should be taken to generate samples that are 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 (a list with appropriate substrates is given in the Materials-Equipment section). Another consideration is the number of freeze-thaw cycles and long-term storage as these could compromise the integrity of the samples^{186,189}. 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^{4,98}. Metal surfaces can also be used to increase the IR signal in a technique known as surface-enhanced IR absorption (SEIRA)^{190,191}. 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 internal humidity of the instrument, 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 uniformly as possible, to ensure that all sources of variation in the samples are stored 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% power¹⁹². Further details regarding sampling methodologies for analysing biological materials using FT-IR spectroscopy can be found in our previous protocols^{4,98}.

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 residuals¹⁹³ can be explored to identify biased patterns, in which heteroscedastic distributions are signs of biased experimental measurements; while homoscedastic distributions are associated with good sampling. SNR can be estimated by dividing the power (P) of signal by the power of noise, that is $SNR = P_{signal}/P_{noise} =$

$(A_{signal}/A_{noise})^2$, 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 the condition number, which is a matrix calculation that measures how sensitive the result is to perturbations in the input data (*i.e.*, spectra) and to roundoff errors made during the solution process. This value is naturally high for spectral data (high collinearity).

Experimental design: pre-processing

Data pre-processing is used to maximise the SNR. This process is fundamental for correcting physical interferences, 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, reduce interferences and possibly correct anomalous samples.

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 spectrum should be truncated to the biofingerprint region (*e.g.*, 900-1800 cm^{-1}) before analysis. This region contains the main absorptions from biochemical compounds and it suffers only minor effects of environmental variability, such as air humidity (free $\nu\text{O-H}$ = 3650–3600 cm^{-1} , hydrogen-bonded $\nu\text{O-H}$ = 3400 – 3300 cm^{-1}) and air CO_2 ($\nu_s\text{CO}_2$ = 2350 cm^{-1})¹⁹⁴. Table 3 summarizes the main pre-processing techniques for correcting noise in biologically-derived datasets.

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

Pre-processing	Interfering	Technique	Advantage	Disadvantage	Optimization
Savitzky-Golay smoothing ¹⁹⁵	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 (<i>e.g.</i> , 2 nd 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) ¹⁹⁶	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) ¹⁹⁷	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	Creates negative signals since the data are centralized to zero (y-scale)	--
Spectral differentiation ¹⁹⁵	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 1 st or 2 nd order differentiation is preferred). The differentiation can be coupled to Savitzky-Golay smoothing
Baseline correction ¹⁹⁸	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 (<i>e.g.</i> , rubber band, automatic weighted least squares, Whittaker filter). The method chosen should be maintained consistent for all systems used

Normalization ⁹⁵	Different sample thickness and concentration	ATR-FTIR, FTIR, Raman	Avoids influence of non-desired signals among the samples	The normalization might hide signal differences between samples at important bands, such as Amide I and Amide II; and also may introduce non-linearities	--
-----------------------------	--	-----------------------	---	--	----

Figure 1 shows 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).

After pre-processing (Table 3), a scaling step should be done, because 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.¹⁹⁹) 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 (*i.e.*, data from each instrumental technique) 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

Further details about these pre-processing steps are provided in “Procedure: Data pre-processing” section. 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. It consists of a data analysis step performed before construction of a chemometric model, in which a portion of the samples are assigned to a training set, while the remaining samples are assigned to a validation and/or test set. The training set is used for model construction, the validation set for model optimization, and the test set for final model evaluation. The process of dividing the samples in three sets can be performed manually or by computer-based methodologies. Manual splitting can generate biased results, therefore we recommend a computational-based split instead. Some examples of these include random selection, leverage¹⁴⁹ or the KS algorithm¹⁷⁴. 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 on²⁰⁰. Mathematically the decomposition takes the form:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \quad (07)$$

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 classes²⁰¹. 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) PDS. Even though the samples in Fig. 2a are pre-processed, three different clusters are still evident. After PDS the samples measured using different systems are normalized into a single cluster.

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-knife²⁰², Z-score²⁰³ or K-modes clustering²⁰⁴ can be utilised among others²⁰⁵. One of the most popular and visually intuitive technique for detecting outliers is the Hotelling T^2 vs Q residual test²⁰⁶. 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 PCs²⁰⁷. 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 PCs²⁰⁷. 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, if the confidence limits is set at a 95% level, certain amount of data-points (5%) should be statistically outside these boundaries.

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)²⁰⁸.

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

Table 4. Classification techniques.

Classification Technique	Advantage	Disadvantage
Linear discriminant analysis (LDA) ²⁰⁹	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) ²⁰⁹	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) ²¹⁰	Fast calculation, high accuracy	Greatly affected by classes having different sizes, needs optimization of the number of latent variables (LVs)
K-Nearest Neighbours (KNN) ²¹¹	Simplicity, non-parametric, suitable for large datasets	Time consuming, needs optimization of the distance calculation method and <i>k</i> value, highly sensitive to the “curse of dimensionality” ¹⁹⁹
Support vector machines (SVM) ²¹²	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) ²¹³	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 ²¹⁴	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 ²¹⁵	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 order²¹⁶, 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 above²¹⁷. PLS-DA is also a feature extraction technique²¹⁰, 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 groups²¹⁸. 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)²¹⁹ and successive projections algorithm (SPA)²²⁰. 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 can also be performed by Minimum Redundancy Maximum Relevance (mRMR) algorithm²²¹, 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.

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.

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 outlined 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.

387 MATERIALS

388 REAGENTS

- 389 • Biological samples (tissue, cells, biofluids)(see Reagent Setup).

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

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

405

406 **Anticoagulants**

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

410

411 **Fixative and preservative agents**

- 412 • Formalin, 10% (vol/vol; Sigma-Aldrich, cat. no. HT501128) ! **CAUTION** Potential
413 carcinogen, irritant and allergenic; use in a fume hood.
- 414 • Ethanol (Fisher Scientific, cat. no. E/0600DF/17)
- 415 • Methanol (Fisher Scientific, cat. no. A456-212) ! **CAUTION** Toxic vapours; use in a fume
416 hood.
- 417 • Acetone (Fisher Scientific, cat. no. A19-1) ! **CAUTION** Acetone vapors may cause
418 dizziness; use in a fume hood.
- 419 • ThinPrep (PreservCyt Solution, Cytoc Corp)
- 420 • SurePath (Becton Dickinson Diagnostics)

421

422 **Dewaxing agents**

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

428

429 **EQUIPMENT**

- 430 • Microtome (Thermo Fisher Scientific, cat. no. 902100A; or cat. no. 956651)
- 431 • Wax dispenser (Electrothermal, cat. no. MH8523B)
- 432 • Sectioning bath (Electrothermal, cat. no. MH8517)
- 433 • Centrifuge (Thermo Fisher Scientific, cat. no. 75002410)
- 434 • Desiccator (Thermo Fisher Scientific, cat. no. 5311-0250)
- 435 • Desiccant (Sigma-Aldrich, cat. no. 13767)
- 436 • Laser power meter (Coherent, cat. no. 1098293)

- Spectrometer
- Computer system

Substrates

▲ **CRITICAL** Substrate should be carefully chosen depending on the spectrochemical approach and the experimental mode that will be used. For more details about the choice of substrate see ref ^{4,99}.

- Low-E slides (Kevley Technologies, CFR)
- BaF₂ slides (Photox Optical Systems)
- CaF₂ slides (Crystran, cat. no. CAF10-10-1)
- 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 N₂.

▲ **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.^{4,98,99}.

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.

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

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

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

490 PROCEDURE

491 Sample preparation

492 **1|** Prepare the biological samples for spectrochemical analysis using the following steps: option
493 A for FFPE tissue samples, option B for snap-frozen or fresh tissue samples, option C for cells
494 and option D for biofluids.

495 ▲ **CRITICAL** Sample preparation is briefly presented in this protocol. More details about
496 sample preparation can be found in refs.^{4,98,99}.

497 **(A) Tissue (FFPE) • TIMING 1-1.5 h**

498 (i) Obtain FFPE tissue blocks.

499 (ii) Section the whole tissue block using a microtome to obtain tissue sections at desired
500 thickness (2-10 μm).

501 ▲ **CRITICAL STEP** Cooling of the tissue on an ice block for 10 min prior to sectioning,
502 hardens the wax and allows easier cutting.

503 (iii) Float the tissue ribbons in a warm H₂O bath (40-44°C) and then deposit onto the
504 substrate of choice.

505 (iv) Allow the tissue sections to dry either at room temperature (30 min) or in a 60°C
506 oven (10 min).

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

509 (v) Dewax the samples by performing three sequential immersions in a dewaxing
510 reagent such as fresh xylene, Histo-Clear solution or hexane (each immersion should last at
511 least 5 min).

512 ▲ **CRITICAL STEP** Thorough dewaxing is important for eliminating all spectral peaks
513 attributed to paraffin.

514 (vi) Immerse the tissue slide in acetone or ethanol (5 min) to remove the xylene and
515 then left to air-dry.

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

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

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

519 ▲ **CRITICAL** For fresh tissue, proceed to step 1B(ii).

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

522 (ii) Use a cryostat to obtain tissue sections at desired thickness (8-10 µm).

523 (iii) Deposit the tissue sections onto an appropriate substrate before spectra are
524 collected (see a list of substrates in the Materials-Equipment section).

525 ▲ **CRITICAL** For FTIR studies the tissue sections need to dry for at least 3 h to remove the
526 H₂O interference from the IR spectra.

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

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

530 ▲ **CRITICAL** If you are working with fixed cells, do step 1C(i) and then proceed to step
531 1C(iii). If you are working with live cells, proceed to step 1C(ii)

532 (i) Wash fixed cells to remove the fixative or preservative solution as these chemicals
533 cause spectral interference in the fingerprint region. Three sequential washes with distilled H₂O
534 or PBS have been shown to remove unwanted peaks.

535 (ii) Detach cultured cells from the growth substrate adding 2-3 mL of fresh warm
536 trypsin/EDTA solution to the side wall of the flask; gently swirl the contents to cover the cell
537 layer. Wash with warmed sterile PBS to remove the medium and trypsin (×3 times; gentle
538 centrifuge at 300 *g* for 7 min).

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

541 (iii) After the final wash, resuspend the remaining cell pellet in distilled H₂O (~50-100
542 µL) and mount onto a substrate of choice; allow sample to dry before analysis.

543 ▲ **CRITICAL STEP** The final suspension of cells (~50-100 µL) should be evenly deposited
544 on the slide either by cytospinning or by micro-pipetting. For cytospinning, take a maximum
545 volume of 200 µL of cells in suspension (spin-fixed cells at 800 *g* for 5 min). After spinning,
546 leave the slide to air-dry.

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

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

549 **▲ CRITICAL** If biofluids are analysed fresh, immediately after collection, continue to step
550 1D(ii).

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

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

553 **▲ CRITICAL STEP** Only a small amount of the biofluid is typically required for
554 spectroscopic studies (1-100 µL). However, this depends and should be tailored according to
555 the study and experimental design. For instance, in case a substrate is used for experiments in
556 the ATR mode, a larger volume is preferred as it allows spectral acquisition from multiple
557 locations of the blood spot. On the contrary, if no substrate is used, such as in the case of the
558 direct deposition of the sample on the ATR crystal, smaller volumes can also be used.

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

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

565 **▲ CRITICAL STEP** For FTIR studies the sample needs to dry adequately before
566 spectroscopic analysis (50 µL dry within approximately 1 h at room temperature). Drying can
567 be sped up by using a gentle stream of air over the sample at a specific flow rate (in a sterile
568 laminar flow hood).

569

570 Spectral acquisition for **FTIR spectroscopy** • **TIMING 2 - 5 min per spectrum**

571 ▲ **CRITICAL** Spectrochemical information can be collected as follows for FTIR
572 spectroscopy.

573 ▲ **CRITICAL** Spectral acquisition is briefly presented in this protocol. More details can be
574 found in refs.^{4,98,99}.

575

576 2 | Optimise the settings before each new study to increase the SNR (see ‘Experimental
577 design: spectral acquisition’).

578 ▲ **CRITICAL STEP** Some of the parameters that need to be adjusted include the
579 resolution, spectral range, co-additions, aperture size, interferometer mirror velocity,
580 and interferogram zero-filling.

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

585 ▲ **CRITICAL STEP** The pressure applied on the sample in the ATR mode affects the
586 signal intensity (*i.e.*, absorbance) between data collected by different instruments and
587 operators. Thus, the pressure applied on the sample should be as similar as possible
588 across different experimental setups to reduce differences between the spectra
589 collected. Depending on the sampling mode that has been chosen (ATR-FTIR,
590 transmission or transflection), deposit the sample onto the appropriate holding
591 substrate.

592 3 | Acquire a background spectrum to account for atmospheric changes.

593 ▲ **CRITICAL STEP** This should be done before every sample.

4 | 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.

■ **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)**

5 | Evaluate the raw data using quality tests to identify anomalous spectra or biased patterns before applying pre-processing. This can be made by visual inspection of the collected spectra followed by Hotelling T^2 versus Q residuals charts (see Supplementary Method 1) using only the mean-centred data, and analysis of PCA residuals. Samples far from the origin of the Hotelling T^2 versus Q residuals chart should be removed, and PCA residuals should be random and close to zero. Further instructions about data quality evaluation can be found at “Experimental Design: data quality evaluation” section.

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

▲ **CRITICAL** Steps 6-11 below can be modified depending on the nature of the dataset. Table 1 provides more details about these pre-processing steps. In case of an ATR-FTIR dataset where samples were acquired and analysed under different experimental conditions, the pre-processing method should follow this order:

6 | **Cutting at biofingerprint region (900-1800 cm^{-1}).** Truncate the spectra to the biofingerprint region, to eliminate atmospheric interference present in other regions of the spectra.

- 7 | **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, since a central data point is required for the smoothing process. Try different window sizes from 3 to 21 and observe how the spectra change (in shape) and how the noise is reduced. Use the smallest window that removes the noise considerably whilst maintaining the original spectral shape. Using a spectral resolution of 4 cm^{-1} , the biofingerprint region ($900\text{-}1800\text{ cm}^{-1}$) 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.
- 8 | **Light scattering correction using either multiplicative scatter correction (MSC), SNV or 2nd derivative.** First try using MSC or SNV, as MSC maintains the spectral scale and both methods maintain the original spectral shape. If the results are not satisfactory (*e.g.*, classification accuracy < 75%), try using the 2nd derivative spectra.
- 9 | **Perform 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.
- 10 | **Normalization** Normalize the spectrum to the amide I peak or amide II peak, or perform a vector normalization (2-Norm, length = 1) to correct different scales across spectra (*e.g.*, due to different sample thicknesses when using FTIR in transmission mode).
- 11 | **Scaling** Mean-centre the data for each variable, and divide this value by the variable standard deviation. In case of data fusion, block-scaling should be used.

Data analysis

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

12 | Determine whether a standardisation procedure is necessary by performing PCA. The PCA scores plot (PC1 vs PC2) should generate a unique clustering pattern for the same type of sample. If two or more clusters are observed for the same type of sample measured under different experimental conditions, then a standardisation procedure is necessary (see Figure 2).

Outlier detection. • TIMING 1h – 1 d (depending on the data size)

13 | 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.

Sample split. • TIMING 1 – 4 h (depending on the data size)

14 | Separate the samples that will be used for the training and the test sets. 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 to separate the samples into each set. KS algorithm is freely available at <https://doi.org/10.6084/m9.figshare.7607420.v1>.

Standardization. • TIMING 1h – 4 d (depending on the data size)

▲ **CRITICAL** Standardization methods should be employed in the following order: DS > PDS (DS should be done before PDS), since the latter is more complex and

requires an additional optimization step (window size optimization). 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.

15 | Use DS to vary the number of transfer samples from 10-100% of the training set from the primary system. Use the validation set from the secondary instrument to find the optimum number of transfer samples using the misclassification rate as cost function.

16 | Perform PDS using the optimum number of samples found with DS. Test different window sizes 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).

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.

17 | 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 dataset) or using cross-validation (leave-one-out for small dataset [ppl 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 dataset) 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.

18 | 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:

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

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

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

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

Steps 2-4, Spectral acquisition: 1 s – 5 min per spectrum (depending on the instrument and spectral acquisition configurations)

Step 5, Data quality evaluation: **15 min – 4 h (depending on the size of the dataset)**

Steps 6-11, Data pre-processing: 15 min – 4 h

Data analysis:

Step 12, Exploratory analysis: 1 h – 4 d

Step 13, Outlier detection: 1 h – 1 d

Step 14, Sample split: 1- 4h (depending on sample size)

Step 15-16, Standardization: 1 h – 4 d

Step 17-18, Model construction: 1 h – 4 d

ANTICIPATED RESULTS

To illustrate how this protocol can be used in practice, we conducted a pilot study 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}	23.0°C	23%
	2	4000-400 cm^{-1}	32	4 cm^{-1}	23.4°C	26%
B	1	4000-400 cm^{-1}	32	4 cm^{-1}	24.0°C	26%
	2	4000-400 cm^{-1}	32	4 cm^{-1}	24.9°C	24%
C	1	4000-400 cm^{-1}	48	4 cm^{-1}	22.5°C	28%
	2	4000-400 cm^{-1}	48	1 cm^{-1}	22.8°C	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.

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 ($900\text{-}1800\text{ cm}^{-1}$), 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 Figure 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 Figure 2).

(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.

(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.

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 Figures 4 and 5. Outlier detection was performed using a Hotelling T^2 versus Q residual test (Supplementary Figure 7). The PCA scores plots for the pre-processed spectra are depicted in Supplementary Figure 6. The main difference between the operators was observed for instrument C Supplementary Figure 5, 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.

(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.

Acknowledgements

CLMM would like to thank Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Brazil (grant 88881.128982/2016-01) for financial support. This work was supported in part in FLM's laboratory by The Engineering and Physical Sciences Research Council (EPSRC; Grant Nos: EP/K023349/1 and EP/K023373/1). MP would like to acknowledge Rosemere Cancer Foundation for funding.

827 Author contributions

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

832 Competing financial interests

833 The authors declare no competing financial interest.

834 Data availability statement

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

837

References

- 1 Baker, M. J. *et al.* Clinical applications of infrared and Raman spectroscopy: state of play and future challenges. *Analyst* **143**, 1735-1757 (2018).
- 2 Melin, A. M., Perromat, A. & Délérís, G. Pharmacologic application of Fourier transform IR spectroscopy: in vivo toxicity of carbon tetrachloride on rat liver. *Biopolymers* **57**, 160-168 (2000).
- 3 Eliasson, C. & Matousek, P. Noninvasive authentication of pharmaceutical products through packaging using spatially offset Raman spectroscopy. *Anal Chem* **79**, 1696-1701 (2007).
- 4 Baker, M. J. *et al.* Using Fourier transform IR spectroscopy to analyze biological materials. *Nat Prot* **9**, 1771-1791 (2014).
- 5 Llabjani, V. *et al.* Polybrominated diphenyl ether-associated alterations in cell biochemistry as determined by attenuated total reflection Fourier-transform infrared spectroscopy: a comparison with DNA-reactive and/or endocrine-disrupting agents. *Environ Sci Technol* **43**, 3356-3364 (2009).
- 6 Hofmann-Wellenhof, B., Lichtenegger, H. & Collins, J. Global positioning system: theory and practice. *Springer Science & Business Media* (2012).
- 7 Morris, P. & Perkins, A. Diagnostic imaging. *Lancet* **379**, 1525-1533 (2012).
- 8 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).
- 9 Lagleyre, S. *et al.* Reliability of high-resolution CT scan in diagnosis of otosclerosis. *Otol Neurotol* **30**, 1152-1159 (2009).
- 10 Kalita, J. & Misra, U. Comparison of CT scan and MRI findings in the diagnosis of Japanese encephalitis. *J Neurol Sci* **174**, 3-8 (2000).
- 11 Schrevers, L., Lorent, N., Doms, C. & Vansteenkiste, J. The role of PET scan in diagnosis, staging, and management of non-small cell lung cancer. *Oncologist* **9**, 633-643 (2004).
- 12 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).
- 13 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).
- 14 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).
- 15 Greenfield, N. J. Using circular dichroism spectra to estimate protein secondary structure. *Nat Protoc* **1**, 2876 (2006).
- 16 Micsonai, A. *et al.* Accurate secondary structure prediction and fold recognition for circular dichroism spectroscopy. *Proc Natl Acad Sci USA* **112**, E3095-E3103 (2015).
- 17 Miles, A. J. & Wallace, B. A. Circular dichroism spectroscopy of membrane proteins. *Chem Soc Rev* **45**, 4859-4872 (2016).
- 18 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).
- 19 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).
- 20 World Health Organization. *Fluorescence microscopy for disease diagnosis and environmental monitoring*. (2005).
- 21 Shahzad, A. *et al.* Diagnostic application of fluorescence spectroscopy in oncology field: hopes and challenges. *Appl Spectrosc Rev* **45**, 92-99 (2010).
- 22 Sieroń, A. *et al.* The role of fluorescence diagnosis in clinical practice. *Onco Targets Ther* **6**, 977 (2013).

888 23 Shin, D., Vigneswaran, N., Gillenwater, A. & Richards-Kortum, R. Advances in fluorescence
889 imaging techniques to detect oral cancer and its precursors. *Future Oncol* **6**, 1143-1154
890 (2010).

891 24 Shahzad, A. *et al.* Emerging applications of fluorescence spectroscopy in medical
892 microbiology field. *J Transl Med* **7**, 99 (2009).

893 25 Möller-Hartmann, W. *et al.* Clinical application of proton magnetic resonance spectroscopy
894 in the diagnosis of intracranial mass lesions. *Neuroradiology* **44**, 371-381 (2002).

895 26 Gowda, G. N. *et al.* Metabolomics-based methods for early disease diagnostics. *Expert Rev*
896 *Mol Diagn* **8**, 617-633 (2008).

897 27 Frisoni, G. B., Fox, N. C., Jack, C. R., Scheltens, P. & Thompson, P. M. The clinical use of
898 structural MRI in Alzheimer disease. *Nature reviews. Neurology* **6**, 67-77 (2010).

899 28 Chan, A. W. *et al.* 1 H-NMR urinary metabolomic profiling for diagnosis of gastric cancer. *Br J*
900 *Cancer* **114**, 59 (2016).

901 29 Palmnas, M. S. & Vogel, H. J. The future of NMR metabolomics in cancer therapy: towards
902 personalizing treatment and developing targeted drugs? *Metabolites* **3**, 373-396 (2013).

903 30 Patil, P. & Dasgupta, B. Role of diagnostic ultrasound in the assessment of musculoskeletal
904 diseases. *Ther Adv Musculoskelet Dis* **4**, 341-355 (2012).

905 31 Navani, N. *et al.* Lung cancer diagnosis and staging with endobronchial ultrasound-guided
906 transbronchial needle aspiration compared with conventional approaches: an open-label,
907 pragmatic, randomised controlled trial. *Lancet Respir Med* **3**, 282-289 (2015).

908 32 Menon, U. *et al.* Sensitivity and specificity of multimodal and ultrasound screening for
909 ovarian cancer, and stage distribution of detected cancers: results of the prevalence screen
910 of the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS). *Lancet Oncol* **10**, 327-
911 340 (2009).

912 33 Smith-Bindman, R. *et al.* Endovaginal ultrasound to exclude endometrial cancer and other
913 endometrial abnormalities. *Jama* **280**, 1510-1517 (1998).

914 34 Gajjar, K. *et al.* Diagnostic segregation of human brain tumours using Fourier-transform
915 infrared and/or Raman spectroscopy coupled with discriminant analysis. *Anal Methods* **5**,
916 89-102 (2013).

917 35 Bury, D. *et al.* Phenotyping Metastatic Brain Tumors Applying Spectrochemical Analyses:
918 Segregation of Different Cancer Types. *Anal Lett*, 1-2 (2018).

919 36 Hands, J. R. *et al.* Attenuated Total Reflection Fourier Transform Infrared (ATR-FTIR) spectral
920 discrimination of brain tumour severity from serum samples. *J Biophotonics* **7**, 189-199
921 (2014).

922 37 Hands, J. R. *et al.* Brain tumour differentiation: rapid stratified serum diagnostics via
923 attenuated total reflection Fourier-transform infrared spectroscopy. *J Neuro-oncol* **127**, 463-
924 472 (2016).

925 38 Walsh, M. J., Kajdacsy-Balla, A., Holton, S. E. & Bhargava, R. Attenuated total reflectance
926 Fourier-transform infrared spectroscopic imaging for breast histopathology. *Vib Spectrosc*
927 **60**, 23-28 (2012).

928 39 Lane, R. & Seo, S. S. Attenuated Total Reflectance Fourier Transform Infrared Spectroscopy
929 Method to Differentiate Between Normal and Cancerous Breast Cells. *J Nanosci Nanotechnol*
930 **12**, 7395-7400 (2012).

931 40 Backhaus, J. *et al.* Diagnosis of breast cancer with infrared spectroscopy from serum
932 samples. *Vib Spectrosc* **52**, 173-177 (2010).

933 41 Wang, J.-S. *et al.* FT-IR spectroscopic analysis of normal and cancerous tissues of esophagus.
934 *World J Gastroenterol* **9**, 1897 (2003).

935 42 Maziak, D. E. *et al.* Fourier-transform infrared spectroscopic study of characteristic
936 molecular structure in cancer cells of esophagus: an exploratory study. *Cancer Detect Prev*
937 **31** (2007).

938 43 McIntosh, L. M. *et al.* Infrared spectra of basal cell carcinomas are distinct from non-tumor-
939 bearing skin components. *J Invest Dermatol* **112**, 951-956 (1999).

940 44 McIntosh, L. M. *et al.* Towards non-invasive screening of skin lesions by near-infrared
941 spectroscopy. *J Invest Dermatol* **116**, 175-181 (2001).

942 45 Mostaço-Guidolin, L. B., Murakami, L. S., Nomizo, A. & Bachmann, L. Fourier transform
943 infrared spectroscopy of skin cancer cells and tissues. *Appl Spectrosc Rev* **44**, 438-455 (2009).

944 46 Mordechai, S. *et al.* Possible common biomarkers from FTIR microspectroscopy of cervical
945 cancer and melanoma. *J Microsc* **215**, 86-91 (2004).

946 47 Hammody, Z., Sahu, R. K., Mordechai, S., Cagnano, E. & Argov, S. Characterization of
947 malignant melanoma using vibrational spectroscopy. *Sci World J* **5**, 173-182 (2005).

948 48 Kondepati, V. R., Keese, M., Mueller, R., Manegold, B. C. & Backhaus, J. Application of near-
949 infrared spectroscopy for the diagnosis of colorectal cancer in resected human tissue
950 specimens. *Vib Spectrosc* **44**, 236-242 (2007).

951 49 Rigas, B., Morgello, S., Goldman, I. S. & Wong, P. Human colorectal cancers display abnormal
952 Fourier-transform infrared spectra. *Proc Natl Acad Sci USA* **87**, 8140-8144 (1990).

953 50 Yao, H., Shi, X. & Zhang, Y. The Use of FTIR-ATR Spectrometry for Evaluation of Surgical
954 Resection Margin in Colorectal Cancer: A Pilot Study of 56 Samples. *J Spectrosc* **2014**, 4
955 (2014).

956 51 Lewis, P. D. *et al.* Evaluation of FTIR Spectroscopy as a diagnostic tool for lung cancer using
957 sputum. *BMC Cancer* **10**, 640 (2010).

958 52 Akalin, A. *et al.* Classification of malignant and benign tumors of the lung by infrared spectral
959 histopathology (SHP). *Lab Invest* **95**, 406 (2015).

960 53 Großerueschkamp, F. *et al.* Marker-free automated histopathological annotation of lung
961 tumour subtypes by FTIR imaging. *Analyst* **140**, 2114-2120 (2015).

962 54 Owens, G. L. *et al.* Vibrational biospectroscopy coupled with multivariate analysis extracts
963 potentially diagnostic features in blood plasma/serum of ovarian cancer patients. *J*
964 *Biophotonics* **7**, 200-209 (2014).

965 55 Gajjar, K. *et al.* Fourier-transform infrared spectroscopy coupled with a classification
966 machine for the analysis of blood plasma or serum: a novel diagnostic approach for ovarian
967 cancer. *Analyst* **138**, 3917-3926 (2013).

968 56 Theophilou, G., Lima, K. M. G., Martin-Hirsch, P. L., Stringfellow, H. F. & Martin, F. L. ATR-
969 FTIR spectroscopy coupled with chemometric analysis discriminates normal, borderline and
970 malignant ovarian tissue: classifying subtypes of human cancer. *Analyst* **141**, 585-594 (2016).

971 57 Mehrotra, R., Tyagi, G., Jangir, D. K., Dawar, R. & Gupta, N. Analysis of ovarian tumor
972 pathology by Fourier Transform Infrared Spectroscopy. *J Ovarian Res* **3**, 27 (2010).

973 58 Paraskevaïdi, M. *et al.* Potential of mid-infrared spectroscopy as a non-invasive diagnostic
974 test in urine for endometrial or ovarian cancer. *Analyst* (2018).

975 59 Taylor, S. E. *et al.* Infrared spectroscopy with multivariate analysis to interrogate
976 endometrial tissue: a novel and objective diagnostic approach. *Br J Cancer* **104**, 790-797
977 (2011).

978 60 Paraskevaïdi, M. *et al.* Aluminium foil as an alternative substrate for the spectroscopic
979 interrogation of endometrial cancer. *J Biophotonics* (2018).

980 61 Gajjar, K. *et al.* Histology verification demonstrates that biospectroscopy analysis of cervical
981 cytology identifies underlying disease more accurately than conventional screening:
982 removing the confounder of discordance. *PLoS One* **9**, e82416 (2014).

983 62 Walsh, M. J. *et al.* IR microspectroscopy: potential applications in cervical cancer screening.
984 *Cancer Lett* **246**, 1-11 (2007).

985 63 Wood, B. R., Quinn, M. A., Burden, F. R. & McNaughton, D. An investigation into FTIR
986 spectroscopy as a biodiagnostic tool for cervical cancer. *Biospectroscopy* **2**, 143-153 (1996).

987 64 Podshyvalov, A. *et al.* Distinction of cervical cancer biopsies by use of infrared
988 microspectroscopy and probabilistic neural networks. *Appl Opt* **44**, 3725-3734 (2005).

989 65 Theophilou, G. *et al.* A biospectroscopic analysis of human prostate tissue obtained from
990 different time periods points to a trans-generational alteration in spectral phenotype. *Sci*
991 *Rep* **5**, 13465 (2015).

992 66 Baker, M. J. *et al.* Investigating FTIR based histopathology for the diagnosis of prostate
993 cancer. *J Biophotonics* **2** (2009).

994 67 Derenne, A., Gasper, R. & Goormaghtigh, E. The FTIR spectrum of prostate cancer cells
995 allows the classification of anticancer drugs according to their mode of action. *Analyst* **136**
996 (2011).

997 68 Gazi, E. *et al.* A correlation of FTIR spectra derived from prostate cancer biopsies with
998 Gleason grade and tumour stage. *Eur Urol* **50**, 750-761 (2006).

999 69 Paraskevasidi, M. *et al.* Differential diagnosis of Alzheimer's disease using spectrochemical
1000 analysis of blood. *Proc Natl Acad Sci USA*, 201701517 (2017).

1001 70 Carmona, P. *et al.* Discrimination analysis of blood plasma associated with Alzheimer's
1002 disease using vibrational spectroscopy. *J Alzheimers Dis* **34**, 911-920 (2013).

1003 71 Carmona, P., Molina, M., López-Tobar, E. & Toledano, A. Vibrational spectroscopic analysis
1004 of peripheral blood plasma of patients with Alzheimer's disease. *Anal Bioanal Chem* **407**,
1005 7747-7756 (2015).

1006 72 Paraskevasidi, M. *et al.* Blood-based near-infrared spectroscopy for the rapid low-cost
1007 detection of Alzheimer's disease. *Analyst* (2018).

1008 73 Sitole, L., Steffens, F., Krüger, T. P. J. & Meyer, D. Mid-ATR-FTIR Spectroscopic Profiling of
1009 HIV/AIDS Sera for Novel Systems Diagnostics in Global Health. *OMICS* **18**, 513-523 (2014).

1010 74 Coopman, R. *et al.* Glycation in human fingernail clippings using ATR-FTIR spectrometry, a
1011 new marker for the diagnosis and monitoring of diabetes mellitus. *Clin Biochem* **50**, 62-67
1012 (2017).

1013 75 Scott, D. A. *et al.* Diabetes-related molecular signatures in infrared spectra of human saliva.
1014 *Diabetol Metab Syndr* **2**, 48 (2010).

1015 76 Varma, V. K., Kajdacsy-Balla, A., Akkina, S. K., Setty, S. & Walsh, M. J. A label-free approach
1016 by infrared spectroscopic imaging for interrogating the biochemistry of diabetic
1017 nephropathy progression. *Kidney Int* **89**, 1153-1159 (2016).

1018 77 Lechowicz, L., Chrapek, M., Gaweda, J., Urbaniak, M. & Konieczna, I. Use of Fourier-
1019 transform infrared spectroscopy in the diagnosis of rheumatoid arthritis: a pilot study. *Mol*
1020 *Biol Rep* **43**, 1321-1326 (2016).

1021 78 Canvin, J. *et al.* Infrared spectroscopy: shedding light on synovitis in patients with
1022 rheumatoid arthritis. *Rheumatology* **42**, 76-82 (2003).

1023 79 Oemrawsingh, R. M. *et al.* Near-infrared spectroscopy predicts cardiovascular outcome in
1024 patients with coronary artery disease. *J Am Coll Cardiol* **64**, 2510-2518 (2014).

1025 80 Wang, J. *et al.* Near-infrared spectroscopic characterization of human advanced
1026 atherosclerotic plaques. *J Am Coll Cardiol* **39**, 1305-1313 (2002).

1027 81 Martin, M. *et al.* The effect of common anticoagulants in detection and quantification of
1028 malaria parasitemia in human red blood cells by ATR-FTIR spectroscopy. *Analyst* (2017).

1029 82 Khoshmanesh, A. *et al.* Detection and Quantification of Early-Stage Malaria Parasites in
1030 Laboratory Infected Erythrocytes by Attenuated Total Reflectance Infrared Spectroscopy and
1031 Multivariate Analysis. *Anal Chem* **86**, 4379-4386 (2014).

1032 83 Roy, S. *et al.* Simultaneous ATR-FTIR Based Determination of Malaria Parasitemia, Glucose
1033 and Urea in Whole Blood Dried onto a Glass Slide. *Anal Chem* **89**, 5238-5245 (2017).

1034 84 Markus, A. P. J. *et al.* New technique for diagnosis and monitoring of alcaptonuria:
1035 quantification of homogentisic acid in urine with mid-infrared spectrometry. *Anal Chim Acta*
1036 **429**, 287-292 (2001).

1037 85 Grimard, V. *et al.* Phosphorylation-induced Conformational Changes of Cystic Fibrosis
1038 Transmembrane Conductance Regulator Monitored by Attenuated Total Reflection-Fourier

1039 Transform IR Spectroscopy and Fluorescence Spectroscopy. *J Biol Chem* **279**, 5528-5536
1040 (2004).

1041 86 Aksoy, C., Guliyev, A., Kilic, E., Uckan, D. & Severcan, F. Bone marrow mesenchymal stem
1042 cells in patients with beta thalassemia major: molecular analysis with attenuated total
1043 reflection-Fourier transform infrared spectroscopy study as a novel method. *Stem Cells Dev*
1044 **21**, 2000-2011 (2012).

1045 87 Graça, G. *et al.* Mid-infrared (MIR) metabolic fingerprinting of amniotic fluid: A possible
1046 avenue for early diagnosis of prenatal disorders? *Anal Chim Acta* **764**, 24-31 (2013).

1047 88 Hasegawa, J. *et al.* Evaluation of placental function using near infrared spectroscopy during
1048 fetal growth restriction. *J Perinatal Med* **38**, 29-32 (2010).

1049 89 Theelen, T., Berendschot, T. T., Hoyng, C. B., Boon, C. J. & Klevering, B. J. Near-infrared
1050 reflectance imaging of neovascular age-related macular degeneration. *Graefes Arch Clin Exp*
1051 *Ophthalmol* **247**, 1625 (2009).

1052 90 Semoun, O. *et al.* Infrared features of classic choroidal neovascularisation in exudative age-
1053 related macular degeneration. *Br J Ophthalmol*. **93**, 182-185 (2009).

1054 91 Peters, A. S. *et al.* Serum-infrared spectroscopy is suitable for diagnosis of atherosclerosis
1055 and its clinical manifestations. *Vib Spectrosc* **92**, 20-26 (2017).

1056 92 Afara, I. O., Prasad, I., Arabshahi, Z., Xiao, Y. & Oloyede, A. Monitoring osteoarthritis
1057 progression using near infrared (NIR) spectroscopy. *Sci Rep* **7**, 11463 (2017).

1058 93 Bi, X. *et al.* Fourier transform infrared imaging and MR microscopy studies detect
1059 compositional and structural changes in cartilage in a rabbit model of osteoarthritis. *Anal*
1060 *Bioanal Chem* **387**, 1601-1612 (2007).

1061 94 David-Vaudey, E. *et al.* Fourier Transform Infrared Imaging of focal lesions in human
1062 osteoarthritic cartilage. *Eur Cell Mater* **10**, 60 (2005).

1063 95 Trevisan, J., Angelov, P. P., Carmichael, P. L., Scott, A. D. & Martin, F. L. Extracting biological
1064 information with computational analysis of Fourier-transform infrared (FTIR)
1065 biospectroscopy datasets: current practices to future perspectives. *Analyst* **137**, 3202-3215
1066 (2012).

1067 96 Andrew Chan, K. L. & Kazarian, S. G. Attenuated total reflection Fourier-transform infrared
1068 (ATR-FTIR) imaging of tissues and live cells. *Chem Soc Rev* **45**, 1850-1864 (2016).

1069 97 Pilling, M. & Gardner, P. Fundamental developments in infrared spectroscopic imaging for
1070 biomedical applications. *Chem Soc Rev* **45**, 1935-1957 (2016).

1071 98 Martin, F. L. *et al.* Distinguishing cell types or populations based on the computational
1072 analysis of their infrared spectra. *Nat Protoc* **5**, 1748-1760 (2010).

1073 99 Butler, H. J. *et al.* Using Raman spectroscopy to characterize biological materials. *Nat Protoc*
1074 **11**, 664-687 (2016).

1075 100 Kong, L. *et al.* Characterization of bacterial spore germination using phase-contrast and
1076 fluorescence microscopy, Raman spectroscopy and optical tweezers. *Nat Protoc* **6**, 625
1077 (2011).

1078 101 Harmsen, S., Wall, M. A., Huang, R. & Kircher, M. F. Cancer imaging using surface-enhanced
1079 resonance Raman scattering nanoparticles. *Nat Protoc* **12**, 1400 (2017).

1080 102 Beckonert, O. *et al.* Metabolic profiling, metabolomic and metabonomic procedures for
1081 NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nat Protoc* **2**, 2692 (2007).

1082 103 Felten, J. *et al.* Vibrational spectroscopic image analysis of biological material using
1083 multivariate curve resolution-alternating least squares (MCR-ALS). *Nat Protoc* **10**, 217
1084 (2015).

1085 104 Yang, H., Yang, S., Kong, J., Dong, A. & Yu, S. Obtaining information about protein secondary
1086 structures in aqueous solution using Fourier transform IR spectroscopy. *Nat Protoc* **10**, 382
1087 (2015).

1088 105 Sreedhar, H. *et al.* High-definition Fourier transform infrared (FT-IR) spectroscopic imaging of
1089 human tissue sections towards improving pathology. *J Vis Exp* (2015).

1090 106 Varriale, A. *et al.* Fluorescence correlation spectroscopy assay for gliadin in food. *Anal Chem*
1091 **79**, 4687-4689 (2007).

1092 107 Song, X., Li, H., Al-Qadiri, H. M. & Lin, M. Detection of herbicides in drinking water by
1093 surface-enhanced Raman spectroscopy coupled with gold nanostructures. *J Food Meas*
1094 *Charact* **7**, 107-113 (2013).

1095 108 Osborne, B. G. & Fearn, T. Near-infrared spectroscopy in food analysis. *Encyclopedia Anal*
1096 *Chem* **5**, 4069-4082 (2000).

1097 109 Qu, J.-H. *et al.* Applications of near-infrared spectroscopy in food safety evaluation and
1098 control: A review of recent research advances. *Crit Rev Food Sci Nut.* **55**, 1939-1954 (2015).

1099 110 Penido, C. A. F., Pacheco, M. T. T., Lednev, I. K. & Silveira, L. Raman spectroscopy in forensic
1100 analysis: identification of cocaine and other illegal drugs of abuse. *J Raman Spectrosc* **47**, 28-
1101 38 (2016).

1102 111 Ryder, A. G. Classification of narcotics in solid mixtures using principal component analysis
1103 and Raman spectroscopy. *J Forensic Sci* **47**, 275-284 (2002).

1104 112 Harrigan, G. G. *et al.* Application of high-throughput Fourier-transform infrared spectroscopy
1105 in toxicology studies: contribution to a study on the development of an animal model for
1106 idiosyncratic toxicity. *Toxicol Lett* **146**, 197-205 (2004).

1107 113 Choo-Smith, L.-P. *et al.* Investigating microbial (micro) colony heterogeneity by vibrational
1108 spectroscopy. *Appl Environ Microbiol* **67**, 1461-1469 (2001).

1109 114 Helm, D., Labischinski, H., Schallehn, G. & Naumann, D. Classification and identification of
1110 bacteria by Fourier-transform infrared spectroscopy. *Microbiology* **137**, 69-79 (1991).

1111 115 Carmona, P., Monzon, M., Monleon, E., Badiola, J. J. & Monreal, J. In vivo detection of
1112 scrapie cases from blood by infrared spectroscopy. *J Gen Virol* **86**, 3425-3431 (2005).

1113 116 Cui, L. *et al.* A novel functional single-cell approach to probing nitrogen-fixing bacteria in soil
1114 communities by resonance Raman spectroscopy with ¹⁵N₂ labelling. *Anal Chem*
1115 **10.1021/acs.analchem.7b05080**. (2018).

1116 117 Lasch, P. & Naumann, D. Infrared spectroscopy in microbiology. *Encyclopedia Anal Chem*
1117 (2015).

1118 118 Maquelin, K. *et al.* Identification of medically relevant microorganisms by vibrational
1119 spectroscopy. *J Microbiol Methods* **51**, 255-271 (2002).

1120 119 Day, J. S., Edwards, H. G., Dobrowski, S. A. & Voice, A. M. The detection of drugs of abuse in
1121 fingerprints using Raman spectroscopy I: latent fingerprints. *Spectrochim Acta A Mol Biomol*
1122 *Spectrosc* **60**, 563-568 (2004).

1123 120 Macleod, N. A. & Matousek, P. Emerging Non-invasive Raman Methods in Process Control
1124 and Forensic Applications. *Pharm Res* **25**, 2205 (2008).

1125 121 Lewis, I., Daniel Jr, N., Chaffin, N., Griffiths, P. & Tungol, M. Raman spectroscopic studies of
1126 explosive materials: towards a fieldable explosives detector. *Spectrochim Acta A* **51**, 1985-
1127 2000 (1995).

1128 122 Hargreaves, M. D. & Matousek, P. Threat detection of liquid explosive precursor mixtures by
1129 Spatially Offset Raman Spectroscopy (SORS). in *Optics and photonics for counterterrorism*
1130 *and crime fighting V*. Vol. **7486** 74860B (International Society for Optics and Photonics).

1131 123 Ali, E. M., Edwards, H. G., Hargreaves, M. D. & Scowen, I. J. Raman spectroscopic
1132 investigation of cocaine hydrochloride on human nail in a forensic context. *Anal Bioanal*
1133 *Chem* **390**, 1159-1166 (2008).

1134 124 Vergote, G. J., Vervaet, C., Remon, J. P., Haemers, T. & Verpoort, F. Near-infrared FT-Raman
1135 spectroscopy as a rapid analytical tool for the determination of diltiazem hydrochloride in
1136 tablets. *Eur J Pharm Sci* **16**, 63-67 (2002).

1137 125 Lohr, D. *et al.* Non-destructive determination of carbohydrate reserves in leaves of
1138 ornamental cuttings by near-infrared spectroscopy (NIRS) as a key indicator for quality
1139 assessments. *Biosys Eng* **158**, 51-63 (2017).

1140 126 Heys, K. A., Shore, R. F., Pereira, M. G. & Martin, F. L. Levels of Organochlorine Pesticides Are
1141 Associated with Amyloid Aggregation in Apex Avian Brains. *Environ Sci Technol* **51**, 8672-
1142 8681 (2017).

1143 127 Comino, F., Aranda, V., García-Ruiz, R. & Domínguez-Vidal, A. Infrared spectroscopy as a tool
1144 for the assessment of soil biological quality in agricultural soils under contrasting
1145 management practices. *Ecol Indicators* **87**, 117-126 (2018).

1146 128 Eliasson, C., Macleod, N. & Matousek, P. Noninvasive detection of concealed liquid
1147 explosives using Raman spectroscopy. *Anal Chem* **79**, 8185-8189 (2007).

1148 129 Liu, H.-B., Zhong, H., Karpowicz, N., Chen, Y. & Zhang, X.-C. Terahertz spectroscopy and
1149 imaging for defense and security applications. *Proc IEEE* **95**, 1514-1527 (2007).

1150 130 Golightly, R. S., Doering, W. E. & Natan, M. J. Surface-enhanced Raman spectroscopy and
1151 homeland security: a perfect match? *ACS Nano* **3**, 2859-2869 (2009).

1152 131 Sattlecker, M., Stone, N., Smith, J. & Bessant, C. Assessment of robustness and transferability
1153 of classification models built for cancer diagnostics using Raman spectroscopy. *J Raman*
1154 *Spectrosc* **42**, 897-903 (2011).

1155 132 Isabelle, M. *et al.* Multi-centre Raman spectral mapping of oesophageal cancer tissues: a
1156 study to assess system transferability. *Faraday Discuss* **187**, 87-103 (2016).

1157 133 Guo, S. *et al.* Towards an improvement of model transferability for Raman spectroscopy in
1158 biological applications. *Vib Spectrosc* **91**, 111-118 (2017).

1159 134 Luo, X. *et al.* Calibration transfer across near infrared spectrometers for measuring
1160 hematocrit in the blood of grazing cattle. *J Near Infrared Spec* **25**, 15-25 (2017).

1161 135 Vaughan, A. A. *et al.* Liquid chromatography–mass spectrometry calibration transfer and
1162 metabolomics data fusion. *Anal Chem* **84**, 9848-9857 (2012).

1163 136 Rodriguez, J. D., Westenberger, B. J., Buhse, L. F. & Kauffman, J. F. Standardization of Raman
1164 spectra for transfer of spectral libraries across different instruments. *Analyst* **136**, 4232-4240
1165 (2011).

1166 137 de Andrade, E. W., de Lelis Medeiros de Morais, C., Lopes da Costa, F. S., de Lima, G. &
1167 Michell, K. A Multivariate Control Chart Approach for Calibration Transfer between NIR
1168 Spectrometers for Simultaneous Determination of Rifampicin and Isoniazid in
1169 Pharmaceutical Formulation. *Curr Anal Chem* **14**, 488-494 (2018).

1170 138 Yu, B., Ji, H. & Kang, Y. Standardization of near infrared spectra based on multi-task learning.
1171 *Spectrosc Lett* **49**, 23-29 (2016).

1172 139 Ni, L., Han, M., Luan, S. & Zhang, L. Screening wavelengths with consistent and stable signals
1173 to realize calibration model transfer of near infrared spectra. *Spectrochim Acta A* **206**, 350-8
1174 (2019).

1175 140 Hu, R. & Xia, J. Calibration transfer of near infrared spectroscopy based on DS algorithm. in
1176 *Electric Information and Control Engineering (ICEICE), 2011 International Conference on*
1177 3062-3065 (IEEE).

1178 141 Forina, M. *et al.* Transfer of calibration function in near-infrared spectroscopy. *Chemom*
1179 *Intellig Lab Syst* **27**, 189-203 (1995).

1180 142 Xiao, H. *et al.* Comparison of benchtop Fourier-transform (FT) and portable grating scanning
1181 spectrometers for determination of total soluble solid contents in single grape berry (*Vitis*
1182 *vinifera* L.) and calibration transfer. *Sensors* **17**, 2693 (2017).

1183 143 Yahaya, O., MatJafri, M., Aziz, A. & Omar, A. Visible spectroscopy calibration transfer model
1184 in determining pH of Sala mangoes. *J Instrum* **10**, T05002 (2015).

1185 144 Bin, J., Li, X., Fan, W., Zhou, J.-h. & Wang, C.-w. Calibration transfer of near-infrared
1186 spectroscopy by canonical correlation analysis coupled with wavelet transform. *Analyst* **142**,
1187 2229-2238 (2017).

1188 145 Monakhova, Y. B. & Diehl, B. W. Transfer of multivariate regression models between high -
1189 resolution NMR instruments: application to authenticity control of sunflower lecithin. *Magn*
1190 *Reson Chem* **54**, 712-717 (2016).

1191 146 Zuo, Q., Xiong, S., Chen, Z.-P., Chen, Y. & Yu, R.-Q. A novel calibration strategy based on
1192 background correction for quantitative circular dichroism spectroscopy. *Talanta* **174**, 320-
1193 324 (2017).

1194 147 Koehler IV, F. W., Small, G. W., Combs, R. J., Knapp, R. B. & Kroutil, R. T. Calibration transfer
1195 algorithm for automated qualitative analysis by passive Fourier transform infrared
1196 spectrometry. *Anal Chem* **72**, 1690-1698 (2000).

1197 148 Rodrigues, R. R. *et al.* Evaluation of calibration transfer methods using the ATR-FTIR
1198 technique to predict density of crude oil. *Chemom Intellig Lab Syst* **166**, 7-13 (2017).

1199 149 Wang, Y., Veltkamp, D. J. & Kowalski, B. R. Multivariate instrument standardization. *Anal*
1200 *Chem* **63**, 2750-2756 (1991).

1201 150 Brouckaert, D., Uyttersprot, J.-S., Broeckx, W. & De Beer, T. Calibration transfer of a Raman
1202 spectroscopic quantification method for the assessment of liquid detergent compositions
1203 from at-line laboratory to in-line industrial scale. *Talanta* **179**, 386-392 (2018).

1204 151 Andrade, E. V., Morais, C. d. L. M., Costa, F. S. L. & Lima, K. M. G. A Multivariate Control
1205 Chart Approach for Calibration Transfer between NIR Spectrometers for Simultaneous
1206 Determination of Rifampicin and Isoniazid in Pharmaceutical Formulation. *Curr Anal Chem*
1207 **14**, 1-7 (2018).

1208 152 Zamora-Rojas, E., Pérez-Marín, D., De Pedro-Sanz, E., Guerrero-Ginel, J. & Garrido-Varo, A.
1209 Handheld NIRS analysis for routine meat quality control: Database transfer from at-line
1210 instruments. *Chemom Intellig Lab Syst* **114**, 30-35 (2012).

1211 153 Panchuk, V., Kirsanov, D., Oleneva, E., Semenov, V. & Legin, A. Calibration transfer between
1212 different analytical methods. *Talanta* **170**, 457-463 (2017).

1213 154 de Morais, C. d. L. M. & de Lima, K. M. G. Determination and analytical validation of
1214 creatinine content in serum using image analysis by multivariate transfer calibration
1215 procedures. *Anal Methods* **7**, 6904-6910 (2015).

1216 155 Khaydukova, M. *et al.* Multivariate calibration transfer between two different types of
1217 multisensor systems. *Sensors Actuators B Chem* **246**, 994-1000 (2017).

1218 156 Barreiro, P. *et al.* Calibration Transfer Between Portable and Laboratory NIR
1219 Spectrophotometers. *Acta Hortic* (2008).

1220 157 Sulub, Y., LoBrutto, R., Vivilecchia, R. & Wabuyele, B. W. Content uniformity determination
1221 of pharmaceutical tablets using five near-infrared reflectance spectrometers: a process
1222 analytical technology (PAT) approach using robust multivariate calibration transfer
1223 algorithms. *Anal Chim Acta* **611**, 143-150 (2008).

1224 158 Zhang, L., Small, G. W. & Arnold, M. A. Multivariate calibration standardization across
1225 instruments for the determination of glucose by Fourier transform near-infrared
1226 spectrometry. *Anal Chem* **75**, 5905-5915 (2003).

1227 159 Martens, H., Høy, M., Wise, B. M., Bro, R. & Brockhoff, P. B. Pre - whitening of data by
1228 covariance - weighted pre - processing. *J Chemom* **17**, 153-165 (2003).

1229 160 Feudale, R. N. *et al.* Transfer of multivariate calibration models: a review. *Chemom Intellig*
1230 *Lab Syst* **64**, 181-192 (2002).

1231 161 Woody, N. A., Feudale, R. N., Myles, A. J. & Brown, S. D. Transfer of multivariate calibrations
1232 between four near-infrared spectrometers using orthogonal signal correction. *Anal Chem* **76**,
1233 2595-2600 (2004).

1234 162 Greensill, C., Wolfs, P., Spiegelman, C. & Walsh, K. Calibration transfer between PDA-based
1235 NIR spectrometers in the NIR assessment of melon soluble solids content. *Appl Spectrosc* **55**,
1236 647-653 (2001).

1237 163 Sjöblom, J., Svensson, O., Josefson, M., Kullberg, H. & Wold, S. An evaluation of orthogonal
1238 signal correction applied to calibration transfer of near infrared spectra. *Chemom Intellig Lab*
1239 *Syst* **44**, 229-244 (1998).

1240 164 Andrews, D. T. & Wentzell, P. D. Applications of maximum likelihood principal component
1241 analysis: incomplete data sets and calibration transfer. *Anal Chim Acta* **350**, 341-352 (1997).

1242 165 Bouveresse, E., Massart, D. & Dardenne, P. Calibration transfer across near-infrared
1243 spectrometric instruments using Shenk's algorithm: effects of different standardisation
1244 samples. *Anal Chim Acta* **297**, 405-416 (1994).

1245 166 Shenk, J. S. & Westerhaus, M. O. Populations structuring of near infrared spectra and
1246 modified partial least squares regression. *Crop Sci* **31**, 1548-1555 (1991).

1247 167 Paatero, P. & Tapper, U. Positive matrix factorization: A non - negative factor model with
1248 optimal utilization of error estimates of data values. *Environmetrics* **5**, 111-126 (1994).

1249 168 Xie, Y. & Hopke, P. K. Calibration transfer as a data reconstruction problem. *Anal Chim Acta*
1250 **384**, 193-205 (1999).

1251 169 Goodacre, R. *et al.* On mass spectrometer instrument standardization and interlaboratory
1252 calibration transfer using neural networks. *Anal Chim Acta* **348**, 511-532 (1997).

1253 170 Chen, W.-R., Bin, J., Lu, H.-M., Zhang, Z.-M. & Liang, Y.-Z. Calibration transfer via an extreme
1254 learning machine auto-encoder. *Analyst* **141**, 1973-1980 (2016).

1255 171 Hu, Y., Peng, S., Bi, Y. & Tang, L. Calibration transfer based on maximum margin criterion for
1256 qualitative analysis using Fourier transform infrared spectroscopy. *Analyst* **137**, 5913-5918
1257 (2012).

1258 172 Fan, W., Liang, Y., Yuan, D. & Wang, J. Calibration model transfer for near-infrared spectra
1259 based on canonical correlation analysis. *Anal Chim Acta* **623**, 22-29 (2008).

1260 173 Wang, Z., Dean, T. & Kowalski, B. R. Additive background correction in multivariate
1261 instrument standardization. *Anal Chem* **67**, 2379-2385 (1995).

1262 174 Kennard, R. W. & Stone, L. A. Computer Aided Design of Experiments. *Technometrics* **11**,
1263 137-148 (1969).

1264 175 Palonpon, A. F. *et al.* Raman and SERS microscopy for molecular imaging of live cells. *Nat*
1265 *Protoc* **8**, 677 (2013).

1266 176 Witze, E. S., Old, W. M., Resing, K. A. & Ahn, N. G. Mapping protein post-translational
1267 modifications with mass spectrometry. *Nat Methods* **4**, 798 (2007).

1268 177 Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198 (2003).

1269 178 Pence, I. & Mahadevan-Jansen, A. Clinical instrumentation and applications of Raman
1270 spectroscopy. *Chem Soc Rev* **45**, 1958-1979 (2016).

1271 179 Ibrahim, O. *et al.* Improved protocols for pre-processing Raman spectra of formalin fixed
1272 paraffin preserved tissue sections. *Anal Methods* **9**, 4709-4717 (2017).

1273 180 Tfayli, A. *et al.* Digital dewaxing of Raman signals: discrimination between nevi and
1274 melanoma spectra obtained from paraffin-embedded skin biopsies. *Appl Spectrosc* **63**, 564-
1275 570 (2009).

1276 181 Byrne, H. J., Knief, P., Keating, M. E. & Bonnier, F. Spectral pre and post processing for
1277 infrared and Raman spectroscopy of biological tissues and cells. *Chem Soc Rev* **45**, 1865-1878
1278 (2016).

1279 182 Meade, A. D. *et al.* Studies of chemical fixation effects in human cell lines using Raman
1280 microspectroscopy. *Anal Bioanal Chem* **396**, 1781-1791 (2010).

1281 183 Baker, M. J. *et al.* Developing and understanding biofluid vibrational spectroscopy: a critical
1282 review. *Chem Soc Rev* **45**, 1803-1818 (2016).

1283 184 Bonifacio, A., Cervo, S. & Sergo, V. Label-free surface-enhanced Raman spectroscopy of
1284 biofluids: fundamental aspects and diagnostic applications. *Anal Bioanal Chem* **407**, 8265-
1285 8277 (2015).

1286 185 Mitchell, A. L., Gajjar, K. B., Theophilou, G., Martin, F. L. & Martin-Hirsch, P. L. Vibrational
1287 spectroscopy of biofluids for disease screening or diagnosis: translation from the laboratory
1288 to a clinical setting. *J Biophotonics* **7**, 153-165 (2014).

1289 186 Lovergne, L. *et al.* Biofluid infrared spectro-diagnostics: pre-analytical considerations for
1290 clinical applications. *Faraday Discuss* **187**, 521-537 (2016).

1291 187 Bonifacio, A. *et al.* Surface-enhanced Raman spectroscopy of blood plasma and serum using
1292 Ag and Au nanoparticles: a systematic study. *Anal Bioanal Chem* **406**, 2355-2365 (2014).

1293 188 Paraskevaïdi, M., Martin-Hirsch, P. L. & Martin, F. L. ATR-FTIR Spectroscopy Tools for Medical
1294 Diagnosis and Disease Investigation. In *Nanotechnology Characterization Tools for*
1295 *Biosensing and Medical Diagnosis*, Springer, Berlin, Heidelberg, 163-211 (2017).

1296 189 Mitchell, B. L., Yasui, Y., Li, C. I., Fitzpatrick, A. L. & Lampe, P. D. Impact of freeze-thaw cycles
1297 and storage time on plasma samples used in mass spectrometry based biomarker discovery
1298 projects. *Cancer Inform* **1** (2005).

1299 190 Glassford, S. E., Byrne, B. & Kazarian, S. G. Recent applications of ATR FTIR spectroscopy and
1300 imaging to proteins. *Biochim Biophys Acta* **1834**, 2849-2858 (2013).

1301 191 Kundu, J., Le, F., Nordlander, P. & Halas, N. J. Surface enhanced infrared absorption (SEIRA)
1302 spectroscopy on nanoshell aggregate substrates. *Chem Phys Lett* **452**, 115-119 (2008).

1303 192 Jones, S., Carley, S. & Harrison, M. An introduction to power and sample size estimation.
1304 *Emerg Med J* **20**, 453-458 (2003).

1305 193 Beebe, K. R., Pell, R. J. & Seasholtz, M. B. Chemometrics: a practical guide. In *Wiley New York*
1306 **4** (1998).

1307 194 Pavia, D. L., Lampman, G. M., Kriz, G. S. & Vyvyan, J. A. Introduction to spectroscopy. In
1308 *Cengage Learning* (2008).

1309 195 Savitzky, A. & Golay, M. J. Smoothing and differentiation of data by simplified least squares
1310 procedures. *Anal Chem* **36**, 1627-1639 (1964).

1311 196 Geladi, P., MacDougall, D. & Martens, H. Linearization and scatter-correction for near-
1312 infrared reflectance spectra of meat. *Appl Spectrosc* **39**, 491-500 (1985).

1313 197 Barnes, R., Dhanoa, M. S. & Lister, S. J. Standard normal variate transformation and de-
1314 trending of near-infrared diffuse reflectance spectra. *Appl Spectrosc* **43**, 772-777 (1989).

1315 198 Brereton, R. G. Chemometrics: data analysis for the laboratory and chemical plant. In *John*
1316 *Wiley & Sons* (2003).

1317 199 Hastie, T., Tibshirani, R. & Friedman, J. The elements of statistical learning 2nd edition. In
1318 *New York: Springer* (2009).

1319 200 Bro, R. & Smilde, A. K. Principal component analysis. *Anal Methods* **6**, 2812-2831 (2014).

1320 201 Martin, F. L. *et al.* Identifying variables responsible for clustering in discriminant analysis of
1321 data from infrared microspectroscopy of a biological sample. *J Comput Biol* **14**, 1176-1184
1322 (2007).

1323 202 Martens, H. & Martens, M. Modified Jack-knife estimation of parameter uncertainty in
1324 bilinear modelling by partial least squares regression (PLSR). *Food Qual Prefer* **11**, 5-16
1325 (2000).

1326 203 Rousseeuw, P. J. & Hubert, M. Robust statistics for outlier detection. *Wiley Interdiscip Rev*
1327 *Data Min Knowl Discov* **1**, 73-79 (2011).

1328 204 Jiang, F., Liu, G., Du, J. & Sui, Y. Initialization of K-modes clustering using outlier detection
1329 techniques. *Inf Sci* **332**, 167-183 (2016).

1330 205 Domingues, R., Filippone, M., Michiardi, P. & Zouaoui, J. A comparative evaluation of outlier
1331 detection algorithms: Experiments and analyses. *Pattern Recognit* **74**, 406-421 (2018).

1332 206 Bakeev, K. A. Process analytical technology: spectroscopic tools and implementation
1333 strategies for the chemical and pharmaceutical industries. In *John Wiley & Sons* (2010).

1334 207 Kuligowski, J., Quintás, G., Herwig, C. & Lendl, B. A rapid method for the differentiation of
1335 yeast cells grown under carbon and nitrogen-limited conditions by means of partial least
1336 squares discriminant analysis employing infrared micro-spectroscopic data of entire yeast
1337 cells. *Talanta* **99**, 566-573 (2012).

1338 208 Morais, C. L. & Lima, K. M. Comparing unfolded and two-dimensional discriminant analysis
1339 and support vector machines for classification of EEM data. *Chemom Intell Lab Syst* **170**, 1-2
1340 (2017).

1341 209 Dixon, S. J. & Brereton, R. G. Comparison of performance of five common classifiers
1342 represented as boundary methods: Euclidean Distance to Centroids, Linear Discriminant

1343 Analysis, Quadratic Discriminant Analysis, Learning Vector Quantization and Support Vector
1344 Machines, as dependent on data structure. *Chemom Intell Lab Syst* **95**, 1-17 (2009).
1345 210 Brereton, R. G. & Lloyd, G. R. Partial least squares discriminant analysis: taking the magic
1346 away. *J Chemom* **28**, 213-225 (2014).
1347 211 Cover, T. & Hart, P. Nearest neighbor pattern classification. *IEEE Trans Inf Theory* **13**, 21-27
1348 (1967).
1349 212 Cortes, C. & Vapnik, V. Support-vector networks. *Mach Learn* **20**, 273-297 (1995).
1350 213 Abraham, A. Artificial neural networks. *handbook of measuring system design* (2005).
1351 214 Fawagreh, K., Gaber, M. M. & Elyan, E. Random forests: from early developments to recent
1352 advancements. *Systems Science & Control Engineering* **2**, 602-609 (2014).
1353 215 LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436 (2015).
1354 216 Seasholtz, M. B. & Kowalski, B. The parsimony principle applied to multivariate calibration.
1355 *Anal Chim Acta* **277**, 165-177 (1993).
1356 217 Morais, C. L. & Lima, K. M. Principal Component Analysis with Linear and Quadratic
1357 Discriminant Analysis for Identification of Cancer Samples Based on Mass Spectrometry. *J*
1358 *Braz Chem Soc*, 31 (2017).
1359 218 Hibbert, D. B. Vocabulary of concepts and terms in chemometrics (IUPAC Recommendations
1360 2016). *Pure Appl Chem* **88**, 407-443 (2016).
1361 219 McCall, J. Genetic algorithms for modelling and optimisation. *J Comput Appl Math* **184**, 205-
1362 222 (2005).
1363 220 Soares, S. F. C., Gomes, A. A., Araujo, M. C. U., Galvão Filho, A. R. & Galvão, R. K. H. The
1364 successive projections algorithm. *Trends Anal Chem* **42**, 84-98 (2013).
1365 221 Kamandar, M. & Ghassemian, H. Maximum relevance, minimum redundancy feature
1366 extraction for hyperspectral images. In *Electrical Engineering (ICEE), 2010 18th Iranian*
1367 *Conference on*. 254-259 (IEEE).

1368

1369

Figure legends

Figure 1. IR spectra of healthy control (absence of disease) samples varying ATR-FTIR instruments and operators. Average (a) raw and (b) pre-processed IR spectra for healthy control samples measured across three different ATR-FTIR spectrometers in the same institute (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).

Figure 2. PCA scores for healthy control (absence of disease) samples varying ATR-FTIR instruments before and after standardization. (a) PCA scores for healthy control samples across three different ATR-FTIR spectrometers in the same institute (A, B and C) after pre-processing but before PDS; (b) PCA scores for healthy control samples across three different ATR-FTIR spectrometers in the same institute (A, B and C) after PDS (model built with 55 transfer samples and window size of 23 wavenumbers). The dotted blue circle shows 95 % confidence ellipse (two-sided). Each measurement observation (circle) corresponds to the data acquired from a unique operator.

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

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

Figure 5. Discriminant function (DF) plots using PCA-LDA to discriminate healthy control (absence of disease) samples from ovarian cancer samples varying the instrument. (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. Sample index represents the number of samples' spectra.

Figure 6. PCA-LDA results for DS and PDS standardisation models for spectra collected by the three different instruments. (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. Transfer samples (%) refer to the percentage of training samples' spectra from the primary instrument that are used to transform the signal obtained using the secondary instrument.

Figure 7. Discriminant function (DF) plots using PCA-LDA to discriminate healthy control (absence of disease) samples from ovarian cancer samples varying the operator.

(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).

Figure 8. PCA-LDA results for DS and PDS standardisation models for spectra collected by two different operators. (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.