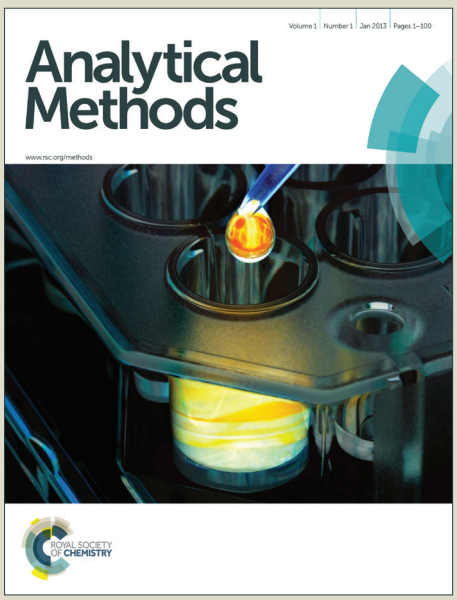
Analytical Methods

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/methods

Infrared spectroscopy with multivariate analysis segregates low-grade cervical cytology based on likelihood to regress, remain static or progress

Nikhil C. Purandare[1,2,], Imran I. Patel[1], Kássio M.G. Lima[1,3], Júlio Trevisan[1], Marwan Ma'Ayeh[2], Anne McHugh[2], Günther Von Bünau[2], Pierre L. Martin Hirsch[1], Walter J. Prendiville[2], Francis L. Martin[1*]

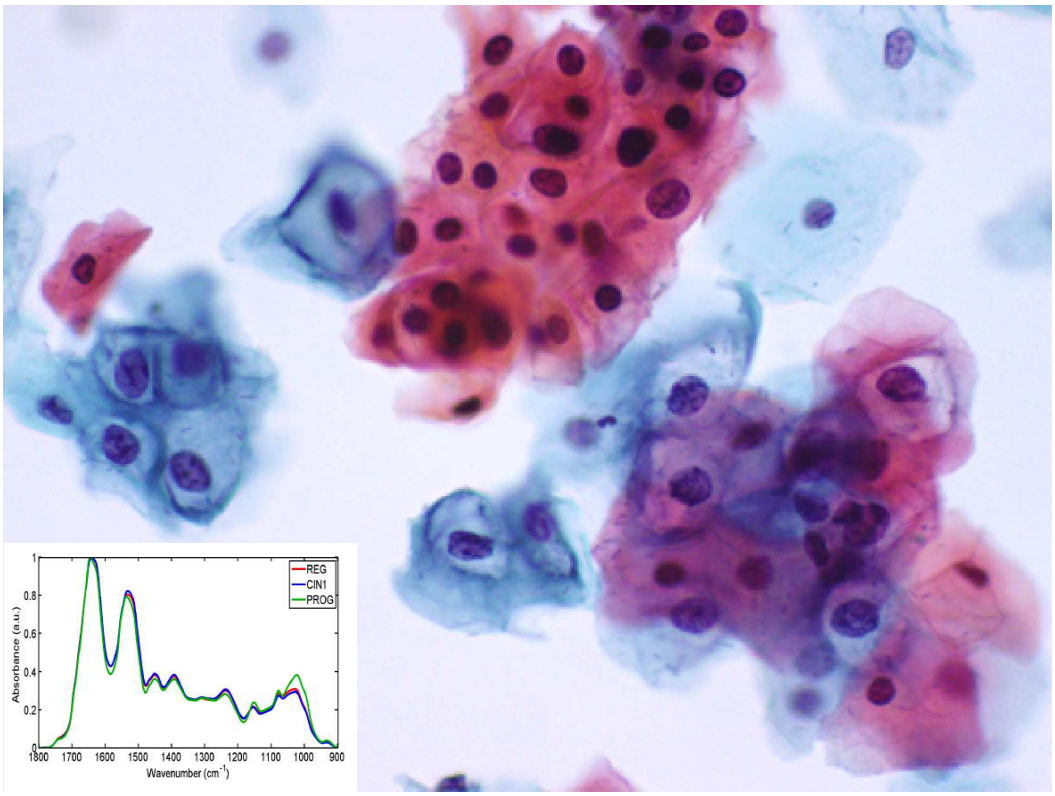
[1]*Centre for Biophotonics, LEC, Lancaster University, Lancaster LA1 4YQ, UK*

[2]*National Clinical Skills Centre, Dublin 8, Ireland*

[3]*Institute of Chemistry, Federal University of Rio Grande do Norte, Natal 59072-970, RN-Brazil*

[*]**Corresponding Author:** Prof Francis L Martin PhD, Centre for Biophotonics, LEC, Lancaster University, Lancaster LA1, 4YQ, UK; Tel.: +44 (0)1524 510206; Email: f.martin@lancaster.ac.uk

ToC graphic



Predicting progressive disease in low-grade cervical cytology

Abstract

Cervical cancer is the 2nd most common female cancer worldwide. However, in the developed world, cervical screening has reduced this cancer burden. Most smear referrals are low-grade, requiring continuous monitoring until they regress. Others need monitoring for static disease, while a few require treatment due to persistent low-grade or progressive disease. The 'Holy Grail' in cervical screening is predicting which patient is likely to have progressive disease. Fourier-transform infrared (FTIR) spectroscopy exploits the fact that an infrared (IR) spectrum represents a "biochemical-cell fingerprint", which can be obtained from a cellular specimen based on a wavenumber-dependent absorption band pattern of constituents' vibrating chemical bonds. Low-grade (CIN1) specimens (*n*=67) diagnosed on cytology were analysed using IR spectroscopy. The *n*=67 study participants were rescreened by conventional cytology after a year whereupon three showed progressive disease and 31 had persistent low-grade atypia; 33 had regressed. Spectra from the initial cytology samples were then analysed using principal component analysis (PCA) with output (10 principal components) being inputted into linear discriminant analysis (LDA) to predict which samples would progress, remain static or regress; this approach was compared with variable selection techniques, namely the successive projection algorithm (SPA) and genetic algorithm (GA). Significant wavenumbers distinguishing regressive *vs.* static disease were 1736 cm^{-1}, 1680 cm^{-1}, 1512 cm^{-1}, 1234 cm^{-1}, 1099 cm^{-1} and 968 cm^{-1}; separating the two categories is difficult due to a significant degree of 'overlap'. Progressive disease can be significantly differentiated from static disease based on wavenumbers 1662 cm^{-1}, 1648 cm^{-1}, 1628 cm^{-1}, 1512 cm^{-1}, 1474 cm^{-1} and 965 cm^{-1}; it can be segregated from regressive disease with 1686 cm^{-1}, 1674 cm^{-1}, 1625 cm^{-1}, 1561 cm^{-1}, 1525 cm^{-1} and 1310 cm^{-1}. The GA-LDA model shows good separation for all categories (*i.e.*, regressive *vs.* static *vs.* progressive disease) using 35 wavenumbers. An ability to predict progressive disease will reduce the need for repeat smears every six months whilst allowing early identification of patients who require treatment.

Keywords: Biospectroscopy; Cervical cytology; Dyskaryosis; Fourier-transform infrared; Low-grade; Multivariate analysis; Progression

Abbreviations: A Randomised Trial of HPV Testing in Primary Cervical Screening, ARTISTIC; Cervical intraepithelial neoplasia, CIN; Cytology that progressed to high-grade disease, PROG; Cytology that regressed after 1 y, REG; Fourier-transform, FT; Genetic algorithm, GA; High-grade squamous intra-epithelial lesion, HGSIL; Human papilloma virus, HPV; Infrared, IR; Kennard-Stone, KS; Large loop excision of transformation zone, LLETZ; Linear discriminant, LD; Linear discriminant analysis, LDA; Low-grade squamous intra-epithelial lesion, LGSIL; Minichromosome maintenance 7 protein, MCM7; Successive projection algorithm, SPA; Principal component, PC; Principal component analysis, PCA; twist-related protein 2, TWIST2

Introduction

Cervical cancer is the 2nd most common cancer in women worldwide.[1] Human papilloma virus (HPV) infection is the cause of almost all cervical cancers.[2] As many as 46% of women are infected with HPV after their first sexual relationship.[3] It is estimated that almost 70% of women will be infected with HPV during their lifetime.[4] Secondary prevention in the form of screening was found to lead to a significant reduction in the incidence of invasive cervical cancer, through early detection and earlier intervention.[5] The current method of screening for disease in the UK is cervical cytology.[6] The recently-introduced vaccination programme against HPV does not provide full protection. Screening programmes must therefore continue, but the challenge in cervical screening is in detecting those individuals who are at higher risk of tumour progression.[7] Cytological and histological results do not reliably distinguish the few with abnormal results who will progress to invasive cancer from the vast majority that will regress or remain unchanged.[8]

Cervical cytology screening has been shown to be associated with poor sensitivity and a poor positive predictive value.[9] Testing for HPV DNA is more sensitive than cervical cytology in detecting pre-cancerous lesions.[10] However, the ARTISTIC trial ("A Randomised Trial of HPV Testing in Primary Cervical Screening") found that over two screening rounds a combined approach (*i.e.*, HPV testing + cytology) did not detect a higher rate of high-grade disease over liquid-based cytology.[11] HPV viruses are classified into high-risk, intermediate-risk and low-risk genotypes. The "High-Risk" HPV subtypes are 16, 18, 31 and 35; types 16 and 18 alone contribute to 70% of all HPV-related cervical cancer.[12] The "Intermediate-Risk" HPV subtypes are 33, 39, 52, 56, 58, 59 and 68. Subtypes 6 and 11 are "Low-Risk" viruses, and account for 90% of genital warts.[13] HPV testing has a role in cervical

screening in women >35 y,[14] but is unable to predict which disease is more likely to progress.[15,16]

Cervical dysplasia may be squamous or glandular. Most abnormalities are squamous in nature and for that reason we will deal with the natural history of squamous disease. Around 80% of cervical intraepithelial neoplasia (CIN)1 is likely to regress spontaneously over a period of 2 y.[17] The published literature suggests that only 11% of CIN1 lesions will progress to high-grade disease.[18] Up to 70% of CIN2 will also regress without treatment within 2 y[19,20] though as many as 24% will progress to CIN3.[20] All women with CIN3 will be treated by an excision procedure. The treatment of CIN2, especially in younger women, is a topic of debate with national guidelines in some countries, including Ireland,[21] advocating treating CIN2 by excision while others suggest it is more likely to regress and should be managed conservatively.[19] Currently the most common method of excision treatment is the Large Loop Excision of the Transformation Zone (LLETZ) procedure. Women with CIN2 or persistent CIN1 are often treated by a LLETZ procedure.

The LLETZ treatment involves excision of the transformation zone using an electrical loop.[22] It is cheap and easy to perform whilst allowing grade of dysplasia and margins to be easily evaluated. The LLETZ procedure is associated with a small risk of bleeding and infection. A recent meta-analysis study has suggested that cervical excision procedures are associated with an adverse pregnancy outcome. This may be due to cervical alteration as a result of the procedure, *i.e.*, loss of cervical tissue volume, which compromises its mechanical function. The scar tissue and the newly-formed collagen may not be as strong. The risk of preterm labour increases with the size of the excision.[23,24] Hence, it would be ideal to develop a screening tool with the ability to predict progression of CIN and avoid unnecessary treatment.

Newer technologies are being developed to detect cervical dysplasia and its progression. Electrical impedance spectroscopy shows promise to reduce the need for a biopsy and has the potential to detect high-grade disease.[25] Chromosomal studies using FISH probes to identify 3q26 gain show promise; absence of 3q26 gain has a 100% negative predictive value for progression but is unable to predict which of the positive cases will progress.[26] Studies on the +874 (T/A) *IFNG* and +1188 (A/C) *IL-12B* genes in cervical smears suggested that the C allele (mutant) may protect against the emergence of CIN and its progression.[27] TWIST2, a basic helix-loop-helix transcription factor has been linked to cervical cancer progression.[28] Ki67, p16 and mini-chromosome maintenance 7 protein (MCM7) are more common in high-grade specimens and have potential in assessing disease progression.[29,30] In reality, there is a need for a cheap and robust test applicable to screening with predictive value.

FTIR spectroscopy is a technique that has been touted as an adjunct to help identify biomarkers of progression. Using this technique, cellular material has been analysed to determine toxin exposure[31], stem cell characterization[32] and to investigate cancer.[33] It has shown potential in the field of cervical cancer screening. Being an inexpensive and robust technique with the ability to segregate grades of cytology, it could potentially be used globally.[34] This technique employs IR to study cellular changes at a molecular level. Molecules absorb the mid-IR region (2.5 μm to 25 μm) at specific wavelengths corresponding to energy levels of the vibrating chemical bonds present, generating a spectrum or a biochemical-cell fingerprint (1800 cm^{-1} - 900 cm^{-1}).[35] This region contains spectral peaks associated with lipids (≈1750 cm^{-1}), Amide I (≈1650 cm^{-1}), Amide II (≈1550 cm^{-1}), methyl groups of lipids and proteins (≈1400 cm^{-1}), Amide III (≈1260 cm^{-1}), asymmetric phosphate stretching vibrations ($\nu_{as}PO_2^-$; ≈1225 cm^{-1}), symmetric phosphate stretching vibrations ($\nu_sPO_2^-$; ≈1080 cm^-

^{1}), C-OH groups of serine, threonine and tyrosine and C-O groups of carbohydrates (≈1155 cm^{-1}), glycogen (≈1030 cm^{-1}) and protein phosphorylation (≈970 cm^{-1}).[35-37]

Of great interest would be the development of an automated algorithm to differentiate between grades of cytology and identify biomarkers of progression. Certain wavenumbers (*i.e.*, particular spectral ranges) may differentiate categories: Amide I (1612 cm^{-1} to 1651 cm^{-1}), Amide II (1512 cm^{-1} to 1551 cm^{-1}), methyl and methylene groups of membrane lipid and proteins (1358 cm^{-1} to 1435 cm^{-1}), $\nu_{as}PO_2^-$ (1192 cm^{-1} to 1261 cm^{-1}) and glycogen/protein phosphorylation (960 cm^{-1} to 1080 cm^{-1}).[36,38-40] A well-developed approach to identify spectral biomarkers is the successive projection algorithm (SPA) or genetic algorithm (GA) in conjunction with linear discriminant analysis (LDA).[41-43] Basically, SPA-LDA and GA-LDA each use a cost function associated with average risk of misclassification in a validation set and can also minimize generalization problems usually associated with collinearity whilst avoiding overfitting. This allows the detection of specific spectral ranges within which specimens differ not only within sample categories but also those that fall within the boundaries between different categories. These 'crossover' regions consist of specimens that are initially misclassified on the basis of spectral similarity. The 'Holy Grail in colposcopy' would be the capability to identify which of such cases are likely to progress.

This study is the first to apply FTIR spectroscopy to identify cases of CIN1 that are more likely to progress. The principle of using biospectroscopy to detect pre-cancer is based on the fact that it may be able to detect underlying disease better than cytology. Our aim was to apply this approach to predict progression as well as to identify wavenumbers as predictive markers, which could assist in predicting disease progression. A secondary aim was to determine if this approach could differentiate

between study participants (patients) whose disease is more likely to regress from those whose disease process remains static. This study analyses spectral data (from cytology specimens) from women who initially presented with a smear suggestive of CIN1 and to retrospectively segregate the three groups (those who regressed *vs.* who remained static *vs.* who progressed to disease) following a repeat smear a year later.

Materials and Methods

This study was conducted during the period from 1st September 2010 to 31st August 2011. Specimens were collected from two separate colposcopy units in Dublin, Ireland. The two centres are the Adelaide and Meath Hospital (Tallaght) and the Coombe Women's and Infant's University Hospital. Ethics committee approval was obtained from both hospitals independently prior to the commencement of the study. All specimens were collected into Thin-Prep® as per routine practice in the two centres. A total of n=67 specimens were collected over a period of one year. Written informed consent was obtained from each study participant (patient). Specimens were sent for spectroscopic analysis after the cytological diagnosis was obtained. Six mL of Thin-Prep® from each specimen was analysed at the Centre for Biophotonics, Lancaster University, UK.

All specimens were centrifuged at 1500 rpm for 5 min. The cell pellet, after discarding the methanol (*i.e.*, fixative in Thin-Prep®), was washed with distilled H_2O and centrifuged; this process was repeated three times. The resulting cell pellet was suspended in 0.5 mL of distilled H_2O. The suspension was applied and then left to dry on an IR-reflective slide (Low-E; Kevley Technologies Inc., OH, USA). Once dry, the specimen was desiccated for a further 24 h. This was to remove any possibility of H_2O contaminating specimen spectra. A Tensor 27 FTIR Spectrometer with Helios

ATR attachment (Bruker Optik GmbH) was used to obtain a total of *n*=670 spectra (10 each from each of 67 specimens). The instrument settings were 32 scans, spectral resolution of 8 cm^{-1}, and interferogram zero-filling of 2×. From each sample analysed, 10 different spectra were objectively obtain from different areas. Prior to analysing each specimen, the diamond crystal within the spectrometer was washed and a background spectrum was obtained to account for atmospheric composition.

All data processing was carried out within MATLAB r2011a (http://www.mathworks.com) using the IRootLab toolbox[44] (http://irootlab.googlecode.com). Raw spectra were pre-processed by cutting between 1,800 and 900 cm^{-1} (469 data points), rubberband baseline-corrected and normalized to the Amide I peak (*i.e.*, around 1,650 cm^{-1}). Acquisition of large datasets with hundreds of spectra, require algorithms to identify subtle but important differences between spectral categories, which are difficult to determine by univariate analysis alone. Therefore, multivariate analysis methods, principal component analysis (PCA) or PCA-LDA,[45] were applied. PCA is an unsupervised data reduction technique generating scores and loadings plots from derived principal components (PCs) of mean-centred spectra.[46] Each PC was examined individually to determine which represented the best segregation of categories. We calculated the variances of the individual PCs and found that the first 10 PCs capture between 99.1% and 99.6% of the total variance of the original dataset (*i.e.*, the sum of the variances of the individual wavenumber absorbance intensities), depending on the analysis case reported below, with PCs of greater order representing mostly noise (only PC1 captured around 76% of the variance in the original dataset). Therefore, input of the first 10 PCs into the supervised technique of LDA was applied. The PCA step prior to LDA is necessary to reduce the number of variables inputted into LDA, as it is

generally accepted that the ratio between the number of spectra and the number of variables (*i.e.,* PCs) should be at least five, for inputting a dataset into a supervised method such as LDA.[47] LDA maximizes inter-category variance in relation to intra-category variance based on pre-set class labels,[46] giving optimal category segregation. A scatter plot ("scores plot") is generated to visualise segregation of the categories, whilst derived loadings plots determine the wavenumbers responsible for segregation between two categories. The loadings Statistical significance of each PC and linear discriminant (LD) contributing to inter-category segregation were determined by unpaired *t*-test and ANOVA.

For SPA-LDA and GA-LDA models, the samples were divided into training (70%), validation (15%) and prediction sets (15%) by applying the classic Kennard-Stone (KS) uniform sampling algorithm[48] to the IR spectra, as shown in Table 1. The training samples were used in the modelling procedure (including variable selection for LDA), whereas the prediction set was only used in the final evaluation of the classification. The optimum number of variables for SPA-LDA and GA-LDA was determined from the minimum of the cost function G calculated for a given validation dataset as:

$$G = \frac{1}{N_V}\sum_{n=1}^{N_V} g_{n,} \tag{1}$$

where g_n is defined as

$$g_n = \frac{r^2(x_n, m_{I(n)})}{\min_{I(m)\neq I(n)} r^2(x_n, m_{I(m)})} \tag{2}$$

where $I(n)$ is the index of the true class for the nth validation object x_n.

The GA routine was carried out during 100 generations with 200 chromosomes each. Crossover and mutation probabilities were set to 60% and 10%,

respectively. Moreover, the algorithm was repeated three times, starting from different random initial populations. The best solution (in terms of the fitness value) resulting from the three realizations of the GA was employed.

Results

A total of n=67 specimens were collected generating 670 spectra to be analysed. Of the 67 study participants with mild dyskaryosis on initial presentation, 33 had a normal smear after one-year follow-up, 31 a diagnosis of low-grade disease, and three a high-grade smear (two with moderate dyskaryosis, one with severe dyskaryosis). This means 49.25% of these CIN1 patients regressed after one year, 46.25% remained low-grade and 4.5% progressed to high-grade disease. Figure 1*A* shows the mean spectra of all three categories. All the spectra are plotted in Figure S1 [see Electronic Supplementary Information (ESI)], whereas the mean spectra with standard deviation bands are shown in ESI Figure S2. It is clearly evident that there are differences in the fingerprint spectra between the three categories depicted; although these is overlap of the error bands in ESI Figure S2; importantly, the mean spectra from the progressive disease category appear to be significantly different from the rest. ESI Figures S3 and S4 show the mean spectrum with standard deviation bands for representative samples/patients (highest and lowest mean variance across all wavenumbers, respectively, in the dataset). CIN1 is characterised by koilocytosis, which is the pathognomonic feature of HPV infection (Fig. 1*B*). It is recognised by the presence of a large sharply-defined, cleared peri-nuclear halo surrounded by a condensed rim of cyanophilic or fuchsia pink cytoplasm. Nuclei are enlarged and hyperchromatic with irregular membranes. Figure 2 is a 2-D PCA-LDA scores plot of the derived spectral points from each category, and demonstrates that progressive disease separates away

from regressive and persistent (static) states. However, there is marked overlap in spectral points of categories that remain static compared to those that regress.

PCA-LDA was subsequently employed to analyse the differences between the three categories (regressive *vs.* static *vs.* progressive disease) taken two-by-two (Figures 3-5). The results are visualized in the form of 1-D scores plots showing segregation of two categories along with estimated distributions of the scores for each category ("B" panels). Furthermore, we plotted the absolute values of the loadings vectors along with their envelope curves ("A" panels). An envelope curve is obtained by joining the peaks of the absolute value of a loadings vector. Such a curve is drawn over the loadings vectors in a thicker line to facilitate the identification of the most important peaks within these vectors. Taking the absolute value is a mathematical operation that discards negative signs.

When PCA-LDA was used to segregate the two categories, progressive disease *vs.* static cytology, the most category-distinguishing wavenumbers were 1662 cm^{-1}, 1648 cm^{-1}, *1628* cm^{-1}, 1512 cm^{-1}, 1474 cm^{-1} and 965 cm^{-1} (Fig. 3*A*). Figure 3*B* is a 1-D scores plot that shows PCA-LDA is able to segregate spectral points derived from specimens that progress from those that remained unchanged. On the other hand, ATR-FTIR spectroscopy did not easily distinguish specimens that regress from those that remain unchanged (*i.e.*, static). This is shown in Figure 4*B* where an 'overlap' (*i.e.*, crossover) between the two categories hints at minimal segregation. However, the prominent wavenumbers distinguishing regressive *vs.* static, using PCA-LDA, are 1736 cm^{-1}, 1680 cm^{-1}, 1512cm^{-1}, 1234 cm^{-1}, 1099 cm^{-1} and 968 cm^{-1} (Fig. 4*A*).

Figure 5*B* shows a 1-D scores plot of specimens that regressed *vs.* those that progressed. Using PCA-LDA one can identify wavenumbers that appear to segregate

these two categories; those that appear important are 1686 cm^{-1}, 1674 cm^{-1}, 1625 cm^{-1}, 1561 cm^{-1}, 1525 cm^{-1} and 1310 cm^{-1} (Fig. 5*A*).

As can be seen in Table 1, the data [*n*=558 spectra = 210/CIN1 (static), 318/REG (regressive) and 30/PROG (progressive)] were divided into training, validation and prediction sets, according to the KS algorithm. The algorithm was applied separately to each category, which is a classic method to extract a representative set of objects from a given dataset. This algorithm works basically in three steps by subset selection: Step 1: for each spectrum *I* not selected in the subset, the Euclidean distances *d(k,i)* between the considered spectrum and each spectrum *k* already selected in the subset are computed; Step 2: for each spectrum *I* not selected in the subset, the smallest Euclidean distance computed between the considered spectrum and the spectra already selected in the subset is found; and, Step 3: the nonselected spectrum *I* that has the highest distance is found and selected in the subset. Then, Steps 1-3 are repeated until the desired number of spectra has been included in the subset.

SPA was applied to the dataset (regressive *vs.* static *vs.* progressive disease) and resulted in the selection of 10 variables, namely 987 cm^{-1}, 1018 cm^{-1}, 1064 cm^{-1}, 1261 cm^{-1}, 1504 cm^{-1}, 1543 cm^{-1}, 1616 cm^{-1}, 1674 cm^{-1}, 1735 cm^{-1} and 1797 cm^{-1}, as shown in Figure 6A. Using these 10 selected wavenumbers, the Fisher scores was obtained and this generated improved segregation between each category (see Figure 6B) when compared with PCA-LDA results. However, then GA was applied to the dataset and resulted in the selection of 35 variables, namely 898 cm^{-1}, 906 cm^{-1}, 952 cm^{-1}, 991 cm^{-1}, 1014 cm^{-1}, 1084 cm^{-1}, 1087 cm^{-1}, 1111 cm^{-1}, 1149 cm^{-1}, 1180 cm^{-1}, 1188 cm^{-1}, 1195 cm^{-1}, 1228 cm^{-1}, 1234 cm^{-1}, 1257 cm^{-1}, 1288 cm^{-1}, 1307 cm^{-1}, 1334 cm^{-1}, 1342 cm^{-1}, 1369 cm^{-1}, 1404 cm^{-1}, 1446 cm^{-1}, 1492 cm^{-1}, 1508 cm^{-1}, 1525 cm^{-1},

1539 cm^{-1}, 1562 cm^{-1}, 1593 cm^{-1}, 1597 cm^{-1}, 1635 cm^{-1}, 1639 cm^{-1}, 1685 cm^{-1}, 1708 cm^{-1}, 1720 cm^{-1} and 1732 cm^{-1}. Using these 35 selected wavenumbers (Figure 6C), the Fisher scores was obtained for all the specimens in the dataset (Figure 6D) whose cost function minimum point was achieved with 35 wavenumbers. As can be seen, there was a good separation for each category, especially for the progressive disease class. However, there is a slight overlap between regressive and static cytology categories. Examination of the selected wavenumbers following PCA-LDA, SPA-LDA and GA-LDA indicates that the main biochemical alterations are associated with lipids, proteins, nucleic acids, carbohydrates and to a lesser extent with DNA vibrations (Table 2).

Discussion

The introduction of the cervical cancer screening programme has reduced the burden of cervical cancer in the developed world.[5] Abnormal cervical smears can cause significant patient anxiety. Most low-grade smears regress, but almost all patients with low-grade abnormal smears continue to be screened until the smear regresses, until it persists long enough to be treated, or until it progresses and is then treated. All this surveillance comes at a great cost, not only to the patient but also to the 'Health Service'. Many CIN2 lesions are treated in women without children because there is a suspicion of underlying CIN3. This in the long-term may increase risks of prematurity and/or dysfunctional labours.[49] With the advent of the cervical cancer screening programme in Ireland, it has been noticeable over the last few years that more women are booking into the Antenatal clinic with a history of some form of cervical excision. Multiple treatments and greater excisions are more likely to cause pre-term labour. This creates much anxiety amongst women. Cervical length scanning and cerclages

do not come without their own risks. Not only does this have implications for the antenatal care for this patient, it also has cost and resource implications in Maternity Hospitals. There is a need for a screening test that has the ability to give a diagnosis as well being capable of predicting likelihood of disease progression in order to reduce the number of unnecessary treatments performed.

In the developing world cervical cytology is the mainstay for cervical cancer screening. Though not universally applied yet, still it has brought about some reductions in cervical cancer, but there is a long way to go. In over-populated countries with low resources, patients who are screened may not return for follow-up. Screening costs money and is laborious. Outreach camps in rural and underprivileged areas attempt to integrate cervical screening into health packages. A test is needed that is cheap, robust, and can produce quick results at screening with the ability to predict progression. This way, in the not so distant future patients can be screened, the risk of progression assessed and they can be treated where necessary so as to avoid losing patients to follow-up. Fewer numbers of patients will require continuous screening making it more cost-effective.

FTIR spectroscopy has shown potential in distinguishing between normal, low-grade and high-grade disease.[39] Certain wavenumbers may underlie the computational segregation between these three categories. Using the same principle, we tried to segregate regressive, static and progressive disease in CIN1 specimens. When n=67 study participants were re-investigated following a year (8 to 14 months) post-initial smear test, 4.5% of patient had progressive disease, almost half the cases regressed and close to half remained low-grade (*i.e.*, static). This data is similar to other published work that suggests that only a small percentage of CIN1 will progress to high-grade disease,[18] while most will regress. A certain percentage will continue to

be abnormal yet not progress and it is this very group that needs to be identified so that they are not over screened or over-treated. It would also be useful to know which cases are more likely to progress so that they may be treated early. There is some suggestion that HPV E6/E7 oncogenic transcripts may be used as a molecular biomarker in women with ASCUS or LGSIL to help predict which women will have disease progression.[50] IR spectroscopy also shows potential in being able to predict which cases are likely to progress.

When ATR-FTIR spectroscopy was employed to predict disease progression, it was observed that using PCA-LDA gives better segregation than PCA alone. If SPA-LDA or GA-LDA following ATR-FTIR spectroscopy analysis was applied to all specimens, it was observed that these latter approaches result in even better segregation of cytology categories than PCA-LDA. SPA-LDA was applied in the dataset using only 10 variables to discriminate all the categories. The variable selection technique of GA with LDA was also performed if even better segregation between category-specific ATR-FTIR spectra could be obtained; the resulting GA-LDA model successfully detected the biochemical alterations in the cytology specimens using only 35 wavenumbers. These wavenumbers should be important contributors to segregation between the three categories.

When distinguishing between regressive *vs.* progressive disease, maximal differences were at the wavenumbers 1686 cm^{-1}, 1674 cm^{-1}, 1625 cm^{-1} (Amide I), 1561 cm^{-1} (Amide II), 1525 cm^{-1} and 1310 cm^{-1}. Differences between progressive *vs.* static disease were observed at 1662 cm^{-1}, 1648 cm^{-1} (Amide I), 1628 cm^{-1}, 1512 cm^{-1} (Amide II), 1474 cm^{-1} and 965 cm^{-1} (protein phosphorylation). When comparing regressive *vs.* static disease, there was significant overlap making it difficult to segregate the two categories but wavenumber differences were noted at 1736 cm^{-1}

(lipids), 1680 cm^{-1}, 1512 cm^{-1}, 1234 cm^{-1} ($\nu_{as}PO_2^-$), 1099 cm^{-1} ($\nu_sPO_2^-$) and 968 cm^{-1} (protein phosphorylation). Several selected wavenumbers (GA-LDA) appear to be of particular interest, namely, the variables at 1334 cm^{-1} and 1342 cm^{-1}, representing the Amide III from proteins. The variables at 1369 cm^{-1} and 1404 cm^{-1} represent the spectral region of fatty acid region and the variables between 1508 cm^{-1}-1597 cm^{-1} correspond of Amide II of proteins.

The above would suggest that wavenumbers 1625 cm^{-1} to 1662 cm^{-1} (Amide I), 1512 cm^{-1} to 1525 cm^{-1} (righthand side of Amide II) and 956 cm^{-1} to 968 cm^{-1} (righthand side of protein phosphorylation) appear to be the three main distinguishing features between these categories. However, 965 cm^{-1}, 968 cm^{-1}, 1014 cm^{-1}, 1099 cm^{-1}, 1234 cm^{-1}, 1334 cm^{-1}, 1342 cm^{-1}, 1508 cm^{-1}, 1512 cm^{-1}, 1562 cm^{-1}, 1628 cm^{-1}, 1648 cm^{-1}, 1685 cm^{-1}, 1708 cm^{-1}, 1720 cm^{-1} and 1736 cm^{-1} summarizes the highlighted variables responsible for separating static, regressive and progressive disease specimens by PCA-LDA, SPA-LDA and GA-LDA algorithms. Larger studies might be able to help distinguish an algorithm to segregate the groups blindly. The 'Holy Grail' in cervical cancer screening is the ability pick up disease that is more likely to progress.[51] Many useful tests such as 3q26 gain, twist-related protein 2 (TWIST2), Ki67, p16 and minichromosome maintenance 7 protein (MCM7) are still under investigation. Most of these are still at a rudimentary phase; some are specialized and expensive. The need is for a cheap test that is easy to perform, robust and cost effective. This technique employs the same sample preparation as is required for conventional liquid-based cytology. Sample preparation only involves washing to get rid of the methanol to avoid it from affecting the spectral signature. The cost lies mostly in the instrumentation (*e.g.*, a Bruker TENSOR27 with a Helios ATR attachment currently costs around £40k). This instrument is the size of a desktop

computer. It is robust with the potential to be made portable and cheaper with the possibility of being automated to increase throughput, making it cost effective.

Essentially, this test has the potential to be cheap and easy to use. The computational process towards data classification needs further development and testing;[52,53] for instance, a systematic assessment of pre-processing methods (*e.g.*, rubberband baseline correction *vs.* derivatization) and classification methods (*e.g.,* LDA *vs.* SVM) could be conducted on a larger dataset. Larger studies on the progression of low-grade disease and studies on conservatively managed CIN2 need to incorporate the use of IR spectroscopy to predict progression. This may also help reduce the screening interval for low-grade disease.

Acknowledgements Funding from Engineering and Physical Sciences Research Council (Grant no.: EP/K023349/1) is gratefully acknowledged.

References

1. D.M. Parkin, and F. Bray, Vaccine **24,** 11-25 (2006).
2. D.M. Parkin, Int J Cancer **118,** 3030-44 (2006).
3. S. Collins, S. Mazloomzadeh, H. Winter, P. Blomfield, A. Bailey, L. Young, and C. Woodman, BJOG **109,** 96-98 (2002).
4. K. Syrjanen, M. Hakama, S. Saarikoski, M. Väyrynen, M. Yliskoski, S. Syrjänen, V. Kataja, and O. Castérn, Sex Transm Dis **17,** 15-19 (1990).
5. J. Peto, C. Gilham, O. Fletcher, and F.E. Matthews, Lancet **364,** 249-56 (2004).
6. D. Luesley, S. Leeson, M. Desai, P. Hadden, H. Kitchener, P. Martin-Hirsch, W. Prendiville, C. Redman, M. Shafi, and J. Tidy, Colposcopy and Programme Management: Guidelines for the NHS Cervical Screening Program. 2nd Edition, NHSCSP Publication No. 20 (2010).
7. N. Wentzensen, and S.J. Klug, Dtsch Arzteblatt Int **105,** 617-22 (2008).
8. C.B.J. Woodman, S.I. Collins, and L.S. Young, Nat Rev Cancer **7,** 11-22 (2007).
9. H.J. Soost, H.J. Lange, W. Lehmacher, and B. Ruffing-Kullmann, Acta Cytol **35,** 8-14 (1991).
10. G. Ronco, and P.G. Rossi, BMC Women's Health **8,** 23 (2008).
11. H. Kitchener, M. Almonte, C. Thomson, P. Wheeler, A. Sargent, B. Stoykova, H. Baysson, C. Roberts, R. Dowie, M. Desai, J. Mather, A. Bailey, A. Turner, S. Moss, and J. Peto, Lancet Oncol **7,** 672-682 (2009).
12. N. Munoz, F. X. Bosch, X. Castellsague, M. Diaz, S. de Sanjose, D. Hammouda, K.V. Shah, and C.J. Meijer, Int J Cancer **111,** 278-85 (2004).

13. M. Schiffman, P. Castle, J. Jeronimo, A.C. Rodriguez, and S. Wacholder, Lancet **370,** 890-907 (2007).

14. M. Leinonen, P. Nieminen, L. Kotaniemi-Talonen, N. Malila, J. Tarkkanen, P. Laurila, and A. Anttila, J Natl Cancer Inst **101,** 1612-1623 (2009).

15. G.Y. Ho, M.H. Einstein, S.L. Romney, A.S. Kadish, M. Abadi, M. Mikhail, J. Basu, B. Thysen, L. Reimers, P.R. Palan, S. Trim, N. Soroudi, R.D. Burk, and The Albert Einstein Cervix Dysplasia Clinical Consortium, J Low Genit Tract Dis **15,** 268-75 (2011).

16. M.K. Yong, J.Y. Park, K.M. Lee, T.W. Kong, S.C. Yoo, W.Y. Kim, J.H. Yoon, S.J. Chang, K.H. Chang, and H.S. Ryu, J Gynecol Oncol **19,** 113-116 (2008).

17. S.S.N. Lee, R.J. Collins, T.C. Pun, D.K.L. Cheng, and H.Y.S. Ngan, Int J Gynecol Obstet **60,** 35-40 (1998).

18. M.A. Duggan, S.C. McGregor, G.C. Stuart, S. Morris, V. Chang-Poon, A. Schepansky, and L. Honore, Eur J Gynaecol Oncol **19,** 338-344 (1998).

19. A.B. Moscicki, M.A. Yifei, C. Wibbelsman, T.M. Darragh, A. Powers, S. Farhat, and S. Shibosk, Obstet Gynecol **116,** 1373-1380 (2010).

20. M.G. Discacciati, C.A. de Souza, M.G. d'Otavianno, L.A. Angelo-Andrade, M.C. Westin, S.H. Rabelo-Santos, and L.C. Zeferino, Eur Jour Obstet Gynecol Reprod Biol **155,** 204-208 (2011).

21. Cervical Check, The National Cervical Screening Programme. http://www.cancerscreening.ie/publications/QA_final_web_version.pdf.

22. W. Prendiville, J. Cullimore, and S. Norman, BJOG **96,** 1054-1060 (1989).

23. S. Khalid, E. Dimitriou, R. Conroy, E. Paraskevaidis, M. Kyrgiou, C. Harrity, M. Arbyn, and W.J. Prendiville, BJOG **119,** 685-691 (2012).

24. J.M. Crane, Obstet Gynecol **102,** 1058-1062 (2003).

25. R. Balasubramani, B.H. Brown, J. Healey, and J.A. Tidy, Gynecol Oncol **115,** 267-271 (2009).

26. A. Rodolakis, I. Biliatis, H. Symiakaki, E. Kershnar, M.W. Kilpatrick, D. Haidopoulos, N. Thomakos, and A. Antsaklis, Int J Gynecol Cancer **22,** 742-747 (2012).

27. V. Do Carmo Vasconcelos de Carvalho, J.L. de Machêdo, C.A. de Lima, M. da Conceicão Gomes de Lima, S. de Andrade Heráclio, M. Amorim, M. de Mascena Diniz Maia, A.L. Porto, and P.R. de Souza, Mol Biol Rep **39,** 7627-7634 (2012).

28. Y. Li, W. Wang, W. Wang, R. Yang, T. Wang, T. Su, D. Weng, T. Tao, W. Li, D. Ma, and S. Wang, Gynecol Oncol **124,** 112-118 (2012).

29. A.J. Kruse, J.P. Baak, E.A. Janssen, K.H. Kjellevoid, B. Fiane, K. Lovslett, J. Bergh, and S. Robboy, Cell Oncol **26,** 13-20 (2004).

30. S. Lobato, A. Tafuri, P.À. Fernandes, M.V. Caliari, M.X. Silva, M.A. Xavier, and A.R. Vago, Gynecol Oncol **23,** 11-15 (2012).

31. X. Bi, M.J. Walsh, X. Wei, G. Sheng, J. Fu, P.L. Martin-Hirsch, G.O. Thomas, K.C. Jones, and F.L. Martin, Environ Sci Technol **41,** 5915-5922 (2007).

32. I.I. Patel, W.J. Harrison, J.G. Kerns, J. Filik, K. Wehbe, P.L. Carmichael, A.D. Scott, M.P. Philpott, M.D. Frogley, G. Cinque, and F.L. Martin, Anal Bioanal Chem **404,** 1745-1758 (2012).

33. J. Babrah, K. McCarthy, R.J. Lush, A.D. Rye, C. Bessant, and N. Stone, Analyst **134,** 763-768 (2009).

34. N.C. Purandare, J. Trevisan, I.I. Patel, K. Gajjar, A.L. Mitchell, G. Theophilou, G. Valasoulis, M. Martin, G. Von Bunau, M. Kyrgiou, E.

Paraskevaidis, P.L. Martin-Hirsch, W.J. Prendiville, and F.L. Martin, Bioanalysis **5,** 2697-2711 (2013).

35. I.I. Patel, J. Trevisan, P.B. Singh, C.M. Nicholson, R.K. Krishnan, S.S. Matanhelia, and F.L. Martin, Anal Bioanal Chem **401,** 969-982 (2011).
36. S. Neviliappan, L. Fang Kan, T.T.L. Walter, S. Arulkumaran, and P.T.T. Wong, Gynecol Oncol **85,** 170-174 (2002).
37. M.J. Walsh, M.J. German, M.N. Singh, H.M. Pollock, A. Hammiche, M. Kyrgiou, H.F. Stringfellow, E. Paraskevaidis, P.L. Martin-Hirsch, and F.L. Martin, Cancer Lett **246,** 1-11 (2007).
38. M.J. Walsh, M.N. Singh, H.M. Pollock, L.J. Cooper, M.J. German, H.F. Stringfellow, N.J. Fullwood, E. Paraskevaidis, P.L. Martin-Hirsch, and F.L. Martin, Biochem Biophys Res Commun **352,** 213-219 (2007).
39. N.C. Purandare, I.I. Patel, J. Trevisan, N. Bolger, R. Kelehan, G. Von Bunau, P.L. Martin-Hirsch, W.J. Prendiville, and F.L. Martin, Analyst **138,** 3909-3916 (2013).
40. B.R. Wood, L. Chiriboga, H. Yee, M.A. Quinn, D. McNaughton, and M. Diem, Gynecol Oncol **93,** 59-68 (2004).
41. Tapp, H. S.; Defernez, M.; Kemsley, E. K. FTIR spectroscopy and multivariate analysis can distinguish the geographic origin of extra virgin olive oils. *J. Agric. Food Chem.* **2003**, *51*, 6110–6115.
42. Oliveira, J. S.; Baia, T. C.; Gama, R. A.; Lima, K. M. G. Development of a novel non-destructive method based on spectral fingerprint for determination of abused drug in insects: An alternative entomotoxicology approach. *Microchem. J.* **2014**, *115*, 39–46.

43. Pontes, M. J. C.; Galvão, R. K. H.; Araújo, M. C. U.; Moreira, P. N. T.; Neto, O. D. P.; José, G. E.; Saldanha, T. C. B. The successive projections algorithm for spectral variable selection in classification problems. *Chemom. Intell. Lab. Syst.* **2005**, *78*, 11–18.

44. J. Trevisan, P.P. Angelov, A.D. Scott, P.L. Carmichael, and F.L. Martin, Bioinformatics **29,** 1095-1097 (2013).

45. F.L. Martin, M.J. German, E. Wit, T. Fearn, N. Ragavan, and H.M. Pollock, J Comput Biol **14,** 1176-1184 (2007).

46. R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification, John Wiley & Sons, New York, 2nd edn. (2001).

47. P. Lasch, and W. Petrich, Data Acquisition and Analysis in Biomedical Vibrational *in* Biomedical Applications of Synchrotron Infrared Microspectroscopy, ed. D. Moss, RSC, Cambridge (2011).

48. R.W. Kennard; L.A. Stone. Computer Aided Design of Experiments. *Technometrics* **1969**, *11*, 137–148.

49. N.C. Purandare, A.F. McHugh, and V. Breschiani, Preterm Labour **2,** 60-66 (2012).

50. P. Paba, C. Ascone, A.A. Criscuolo, F. Marcuccilli, M. Ciccozzi, F. Sesti, E. Piccione, C.F. Perno, and M. Ciotti, Anticancer Res **32,** 1253-1257 (2012).

51. K. Gajjar, A.A. Ahmadzai, G. Valasoulis, J. Trevisan, C. Founta, M. Nasioutziki, A. Loufopoulos, M. Kyrgiou, S.M. Stasinou, P. Karakitsos, E. Paraskevaidis, B. Da Gama-Rose, P.L. Martin-Hirsch, and F.L. Martin, PLoS One **9,** e82416 (2014).

52. K. Gajjar, J. Trevisan, G. Owens, P.J. Keating, N.J. Wood, H.F. Stringfellow, P.L. Martin-Hirsch, and F.L. Martin, Analyst **138,** 3917-3926 (2013).

53. F.L. Martin, J.G. Kelly, V. Llabjani, P.L. Martin-Hirsch, I.I. Patel, J. Trevisan, N.J. Fullwood, and M.J. Walsh, Nat Protoc **5,** 1748-1760 (2010).

Legends to Figures

Figure 1 Predicting progressive disease in low-grade cervical cytology. (***A***) Average spectra acquired from all specimens. The spectra from patients with regressive disease are shown in red; those with static disease (CIN1) are shown in blue; and, those from patients with progressive disease are in green. (***B***) An example of CIN1 following a conventional Papanicolaou stain showing a typical mixture of differing cell types.

Figure 2. Two-D PCA-LDA showing segregation as well as crossover.

Figure 3. Comparison of static and progressive specimens. The panel shows principal component analysis-linear discriminant analysis (PCA-LDA) loadings plots (***A***) alongside one-dimensional scores plots (***B***) showing segregated and crossover specimens.

Figure 4. Comparison of static and regressive specimens. The panel shows principal component analysis-linear discriminant analysis (PCA-LDA) loadings plots (***A***) alongside one-dimensional scores plots (***B***) showing segregated and crossover specimens.

Figure 5. Comparison of progressive and regressive specimens. The panel shows principal component analysis-linear discriminant analysis (PCA-LDA) loadings plots (***A***) alongside one-dimensional scores plots (***B***) showing segregated and crossover specimens.

Figure 6. The application of variable selection techniques to the segregation of retrospectively categorised low-grade cervical cytology specimens. Successive projection algorithm (SPA)-linear discriminant analysis (LDA) results: (***A***) Ten wavenumber variables selected; and, (***B***) DF1 × DF2 discriminant function values calculated by using the variables selected by SPA-LDA from all specimens. Genetic algorithm (GA)-LDA results: (***C***) 35 wavenumbers selected; and, (***D***) DF1 × DF2 discriminant function values calculated by using the variables selected by GA-LDA from all specimens.

Figure 1

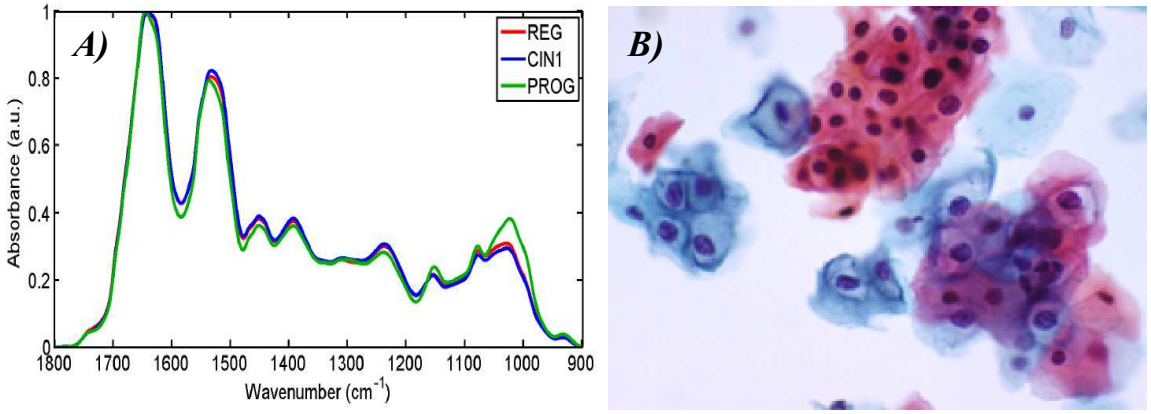


Figure 2

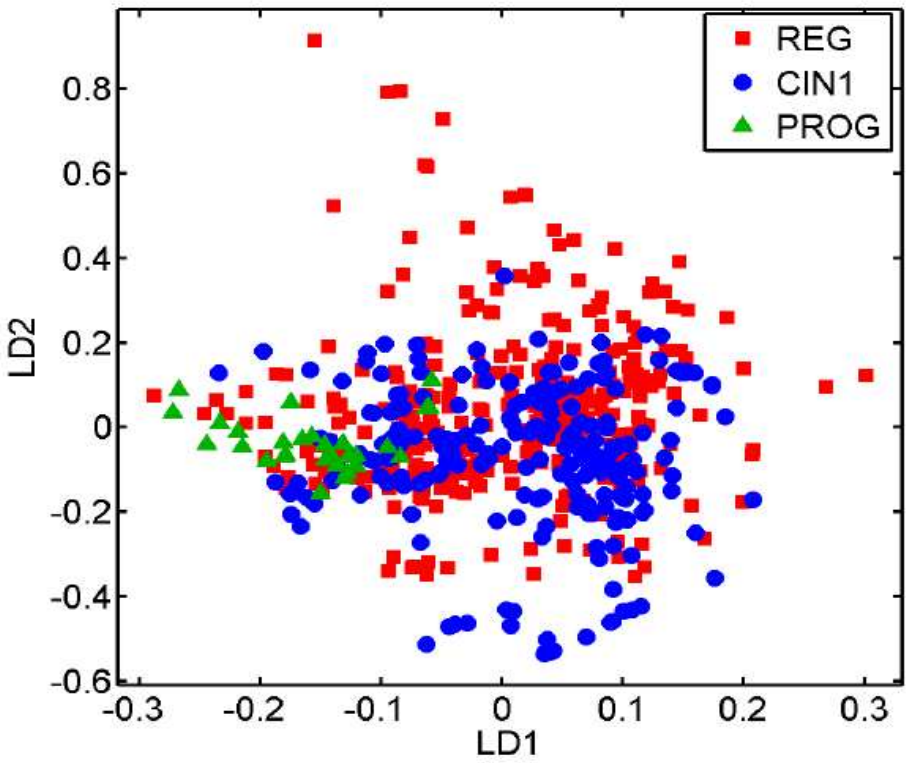


Figure 3

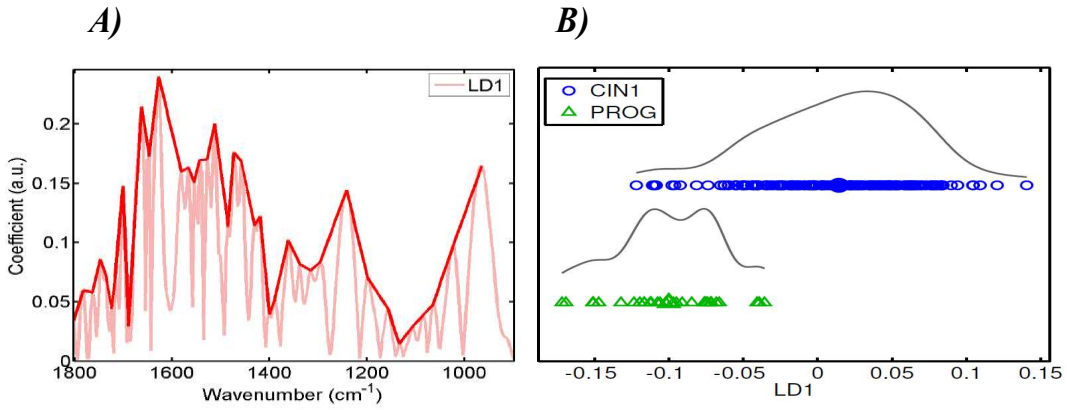


Figure 4

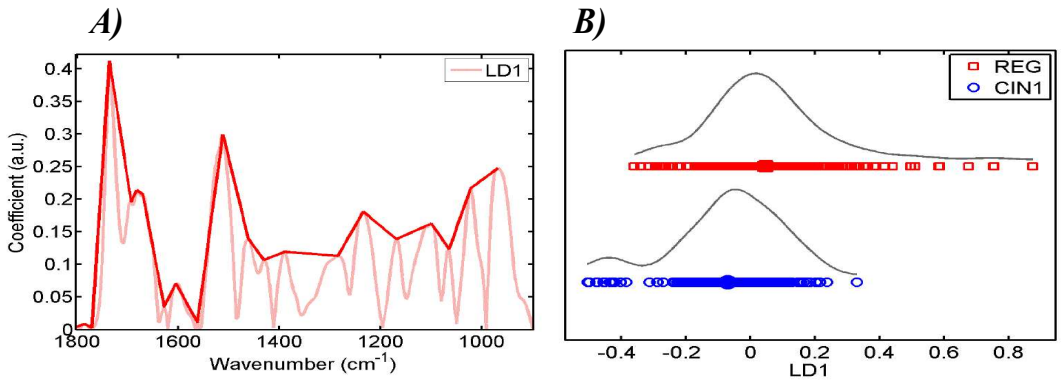


Figure 5

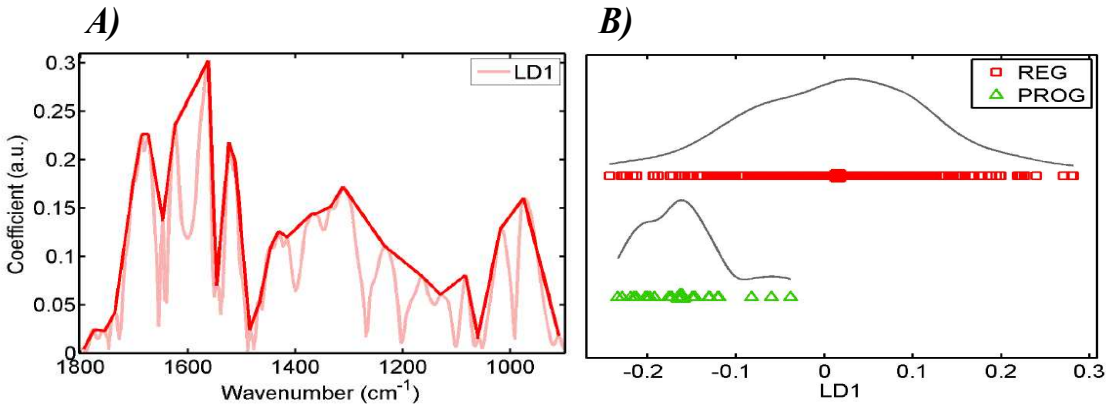


Figure 6

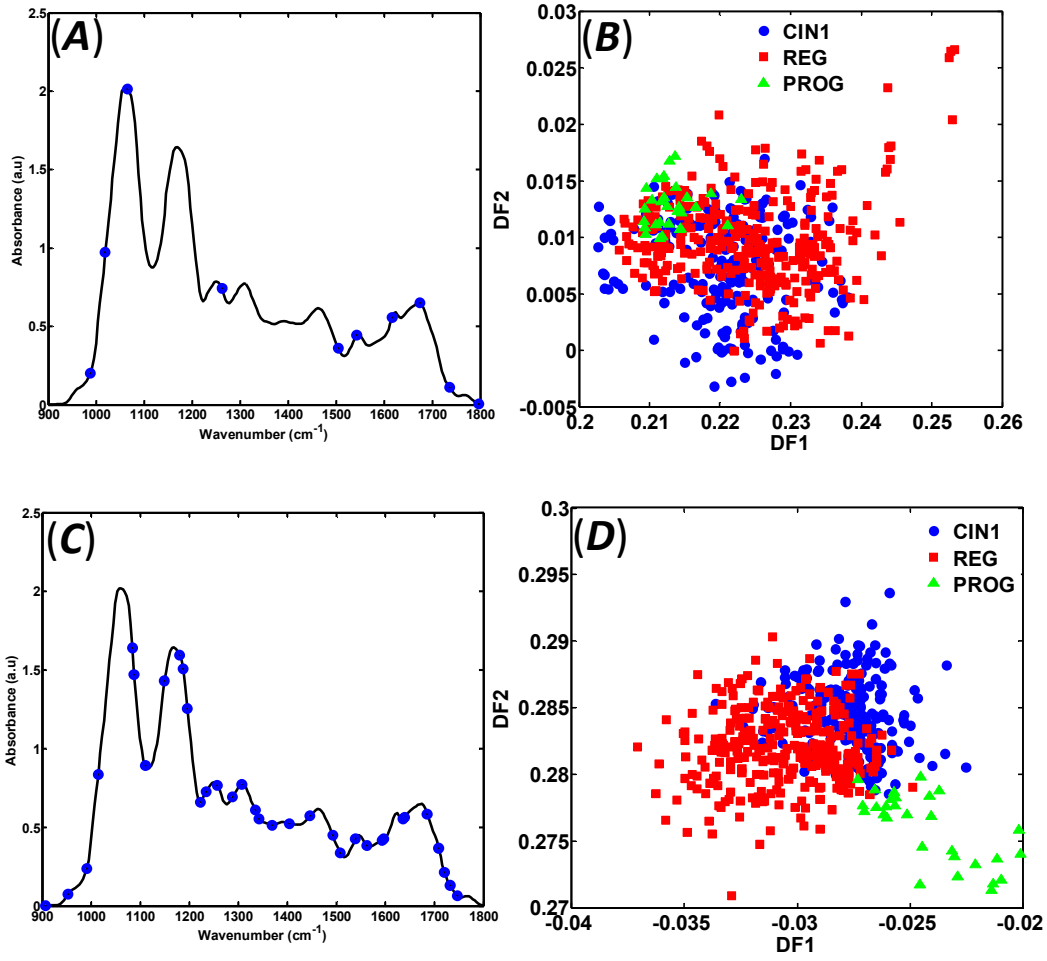


Table 1: Number of training, validation and prediction specimens (or spectra) in each category.

Category	Set training	Validation	Prediction
CIN1	140	35	35
REG	218	50	50
PROG	20	5	5
Total	378	90	90

CIN1, static as cervical intraepithelial neoplasia 1; REG, cytology that regressed after 1 y; and, PROG, cytology that progressed to high-grade disease

Table 2: Highlighted variables responsible for separating CIN1, REG and PROG specimens by PCA-LDA, SPA-LDA or GA-LDA algorithms.

Wavenumbers (cm^{-1})	Tentative Assignments
965	Out-of-plane C-H bending
968	DNA band
1014	C-O and C-C stretching; C-O-H and C-O-C deformation of carbohydrates
1099	$V_{as}PO_2^-$
1234	$V_{as}PO_2^-$
1334	Amide III
1342	Amide III (N-H stretch, C-N stretch of aromatic amines)
1508	Amide II of proteins
1512	νC=O (Amide II)
1562	Amide II of proteins (*e.g.*, side-chain carboxyl groups)
1628	Amide I (C=N; associated with β-sheets)
1648	Amide I (random coil)
1685	Amide I (C=O stretch of ketones; conjugated)
1708	C=O stretching vibrations of ketones
1720	C=O stretching vibrations of aldehydes
1736	Lipid (νCOOH carboxyl groups)