Potential of mid-infrared spectroscopy as a non-invasive diagnostic test in urine for endometrial or ovarian cancer

Maria Paraskevaidi[a,1], Camilo L.M. Morais[a,b], Kássio M.G. Lima[b], Katherine M. Ashton[c], Helen F. Stringfellow[c], Pierre L. Martin-Hirsch[d] and Francis L. Martin[a,1]

[a]School of Pharmacy and Biomedical Sciences, University of Central Lancashire, Preston PR1 2HE, UK; [b]Institute of Chemistry, Biological Chemistry and Chemometrics, Federal University of Rio Grande do Norte, Natal 59072-970, Brazil; [c]Pathology Department, Lancashire Teaching Hospitals NHS Foundation, Preston PR2 9HT, UK; [d]Department of Obstetrics and Gynaecology, Lancashire Teaching Hospitals NHS Foundation, Preston PR2 9HT, UK

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

Abstract

The current lack of an accurate, cost-effective and non-invasive test that would allow for screening and diagnosis of gynaecological carcinomas, such as endometrial and ovarian cancer, signals the necessity for alternative approaches. The potential of spectroscopic techniques in disease investigation and diagnosis has been previously demonstrated. Here, we used attenuated total reflection Fourier-transform infrared (ATR-FTIR) spectroscopy to analyse urine samples from women with endometrial (n=10) and ovarian cancer (n=10), as well as from healthy individuals (n=10). After applying multivariate analysis and classification algorithms, biomarkers of disease were pointed out and high levels of accuracy were achieved for both endometrial (95% sensitivity, 100% specificity; accuracy: 95%) and ovarian cancer (100% sensitivity, 96.3% specificity; accuracy 100%). The efficacy of this approach, in combination with the non-invasive method for urine collection, suggest a potential diagnostic tool for endometrial and ovarian cancers.

Keywords: infrared spectroscopy; chemometrics; non-invasive; ovarian cancer; endometrial cancer; diagnosis

1. Introduction

Endometrial and ovarian cancers are the commonest cancers in post-menopausal women worldwide. In the UK alone, 9,300 women develop endometrial cancer, of whom 2,100 die annually [1]; ovarian cancer affects 7,300 women resulting in 4,100 deaths per year (*i.e.*, ~40% overall survival) [2]. The epidemiology of endometrial and ovarian cancer is closely entwined, histological subtypes of endometrial cancer mirror subtypes found in ovarian cancer and the same risk factors seem to influence both diseases [3]. Endometrial cancer is often symptomatic at an early stage (stage I) when there is still time for treatment [4]. In the case of ovarian carcinoma, however, symptoms present late in most cases and after the cancer has already metastasized within the abdomen, resulting in late-stage disease and poor prognoses [5]. An accurate and early diagnosis of both diseases, and especially ovarian cancer, is of major need as it would permit an early intervention and potentially early-stage diagnosis and consequently improved prognosis.

The gold standard for diagnosis of endometrial cancer is biopsy performed either in an outpatient or inpatient setting after a patient presents with symptomatic bleeding. For ovarian cancer diagnosis in patients with symptoms, women initially undergo a pelvic examination, followed by measurement of serum cancer antigen (CA-125); if symptoms persist in the absence of raised CA-125 levels, an abdominal and transvaginal ultrasound follow [5]. In asymptomatic women for ovarian cancer screening, a combination of these biomarkers is used. In the future, with the escalating incidence of endometrial cancer secondary to obesity, there might be a role for screening for disease. All of the above-mentioned diagnostic approaches have drawbacks, either being invasive (*e.g.*, biopsy) or expensive (*e.g.*, ultrasound). Even though a blood biomarker would be an ideal diagnostic approach, CA-125 has now been found to be unsuitable for early-stage diagnosis as it is only elevated in 50% of the individuals [6]. A

number of research groups are actively investigating the utility of multiple biomarkers for a more accurate diagnosis of endometrial and ovarian cancers [6, 7]. However, current methods of biomarker identification are heavily dependent on multiplex assays and molecular techniques which are costly.

Vibrational spectroscopy has gained increasing attention in the recent years due to its potential as a diagnostic tool for various diseases, by providing chemical and structural information of the sample in use [8, 9]. Both infrared (IR) and Raman spectroscopic techniques have been extensively used for cancer diagnostics using tissue, cells or biofluids, such as blood plasma/serum, urine, bile, ascitic fluid and cerebrospinal fluid [10, 11]. A screening or diagnostic test should be non-invasive to facilitate compliance and, as such, venepuncture and urine analysis are ideal. Blood- and urine-based spectroscopy have already been applied successfully in studies including brain [12], breast [13], gynaecological [14, 15] and other types of cancer. In the present study, attenuated total reflection-Fourier transform infrared (ATR-FTIR) spectroscopy was used to analyse urine samples from women with endometrial and ovarian cancer and facilitate towards their segregation form healthy controls. The classification models that were employed to distinguish between these groups were partial least squares discriminant analysis (PLS-DA), principal component analysis with support vector machines (PCA-SVM) and genetic algorithm with linear discriminant analysis (GA-LDA).

2. Materials and Methods

2.1 Study population and sample collection

All samples were collected at Royal Preston Hospital UK after obtaining ethical approval (16/EE/0010). All experiments were performed in accordance with relevant laws and guidelines, and approved by the ethics committee at University of Central Lancashire (UCLan). Informed consent was obtained from all human subjects. Urine samples were collected from

30 individuals: 10 healthy women with no symptoms of cancer who were used as controls, 10 women with endometrial cancer and 10 women with ovarian cancer. All urine specimens were obtained after patients were administered a general anaesthetic prior to hysterectomy for benign or malignant indications. All patients had undergone a period of at least 6 hours fasting; the suggested preoperative fasting time is 6-8 hours for light meals and 2 hours for fluids [16, 17]. Prolonged fasting (12-16 h) should be avoided as it triggers gluconeogenesis precipitation and increases the organic response to trauma [18]. Specimens were obtained after aseptic preparation of the urethra and after catheterisation, thus avoiding any contamination. All cases were staged according to the guidelines by the International Federation of Gynecologic Oncology (FIGO). Samples were kept at -80°C until the time of spectroscopic analysis. Before analysis, all samples were left to thaw at room temperature and 50 µl were deposited on low-emissivity (low-E) slides (MirrIR Low-E slides, Kevley Technologies, USA); they were then left to air-dry for approximately 45 minutes. All urines were taken pre-operatively on the day of surgery in both controls and disease patients and patients had not received any treatment for the disease prior to surgery. All endometrial and ovarian cancers were high grade cancers. Sub-group analysis for other incidental diseases or factors such as type 2 diabetes, hypertension or medication was not performed in this study.

2.2 Spectroscopic analysis

ATR-FTIR spectroscopy was employed for the analysis of the urine samples. A Tensor 27 FTIR spectrometer with a Helios ATR attachment (Bruker Optics Ltd, Coventry, UK) which contained a diamond crystal, was used for the data collection. The OPUS 7.2 software was used for spectral acquisition. Spectral resolution was set at 8 cm^{-1} and mirror velocity at 2.2 kHz; 32 scans were acquired for each spectrum for optimal signal-to-noise ratio. The diamond crystal was cleaned with distilled water after the use of each sample and a background spectrum was

collected to eliminate atmospheric changes. A CCTV camera was used for visualisation and navigation across the sample's surface; ten spectra were acquired from different locations of each sample to minimize bias.

2.3 Data analysis

After collection, the raw spectra need to be pre-processed in order to account for inconsistencies relating to the experimental procedure and spectral acquisition. Pre-processing and computational analysis of the data was performed using PLS Toolbox version 7.9.3 (Eigenvector Research, Inc., Manson, USA) and an in-house developed IRootLab toolbox (http://trevisanj.github.io/irootlab/) [19]. For the purposes of this study, raw spectra were initially pre-processed as following: cut to the bio-fingerprint region (1800-900 cm^{-1}), rubberband baseline corrected and vector normalised. Rubberband baseline correction is used to correct underlying oscillations on the baseline of the spectra which can be caused by scattering effects, reflection, temperature, concentration, among other instrumental anomalies that render wavenumbers, known for having no absorption, with absorbance values different from zero [20].

For further classification spectra were divided into training (60%, n = 18 patients) and test (40%, n = 12 patients) sets using the Kennard-Stone sample selection algorithm [21]. Partial least squares discriminant analysis (PLS-DA), principal component analysis with support vector machines (PCA-SVM) and genetic algorithm with linear discriminant analysis (GA-LDA) were used as classification methods. PLS-DA is a linear classification technique that uses partial least squares (PLS) to find a straight line that divide the classes spaces [22]. PCA-SVM makes use of principal component analysis (PCA) [23] for data compression; the PCA scores are then used as input variables for a support vector machine (SVM) classifier [24]. The SVM classifier was based on a radial basis function (RBF) kernel which is used to transfer the data to a feature space by means of a non-linear discriminant criterion; a linear decision surface

is then constructed in this feature space to separate the classes analysed [25]. Both PLS-DA and PCA-SVM were optimized using cross-validation venetian blinds in a "leave one patient out" fashion (10 splits with 1 sample per split). GA-LDA was optimized using an external validation data set having half of the samples of the test set. The algorithm was applied three times, using 100 generations with 200 chromosomes, and the best model was selected. Crossover and mutation probabilities were set to 60% and 10%, respectively.

In order to study the differences at specific wavenumbers, we implemented a simple approach, namely difference-between-means (D-B-M) spectra, which subtracts the mean spectra from a reference class (*i.e.*, healthy controls). A peak-detecting algorithm was then used to denote six of the most differentiating peaks.

2.4 Availability of data

All data (raw and pre-processed spectra) along with appropriate code identifiers have been uploaded onto the publicly accessible data repository Figshare (https://figshare.com/articles/Potential_of_mid-infrared_spectroscopy_as_a_non-invasive_diagnostic_test_for_endometrial_or_ovarian_cancer_in_urine/5929516).

2.5 Statistical analysis

The classification performance of the chemometric algorithms was evaluated according to the accuracy, sensitivity and specificity on the test set. The accuracy (AC) represents the number of samples correctly classified considering true and false negatives; sensitivity (SENS) and specificity (SPEC) measure the proportion of positives and negatives that are correctly identified, respectively [26]. These parameters are calculated as follows:

$$\mathrm{AC}(\%) = \left(\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}}\right) \times 100 \qquad (3)$$

$$\mathrm{SENS}(\%) = \left(\frac{\mathrm{TP}}{\mathrm{TP+FN}}\right) \times 100 \quad (4)$$

$$\mathrm{SPEC}(\%) = \left(\frac{\mathrm{TN}}{\mathrm{TN+FP}}\right) \times 100 \quad (5)$$

where TP stands for true positive, TN for true negative, FP for false positive and FN for false negative.

The peaks that were responsible for the differentiation after the D-B-M spectra approach, were imported into GraphPad Prism 7 (GraphPad Software, Inc., La Jolla, CA, 92037, USA) to conduct statistical analyses and calculate the *P*-values. Differences between the two groups (*i.e.*, healthy *vs* cancer) were assessed using a Student's t-test (two-tailed, non-parametric, Mann-Whitney test, 95% confidence interval). The data were expressed as the mean ± standard deviation (SD). A *P*-value of 0.05 or less was considered significant.

3. Results

3.1 Segregation between cancer patients and controls

All spectra were pre-processed before comparison of the cancer patients with the healthy controls. Supplementary Fig. 1A shows the average pre-processed spectra for each class. The most discriminatory peaks for the comparison between endometrial cancer and healthy women were: 1593 cm^{-1} (*P* <0.0001, 95% CI = -0.0179 to -0.012), 1508 cm^{-1} (*P* <0.0001, 95% CI = 0.0038 to 0.0103), 1462 cm^{-1} (*P* <0.0001, 95% CI = -0.0115 to -0.0058), 1400 cm^{-1} (*P* <0.0001, 95% CI = 0.0107 to 0.0161), 1335 cm^{-1} (*P* <0.0001, 95% CI = 0.0053 to 0.0093), 1041 cm^{-1} (*P* <0.0001, 95% CI = 0.0063 to 0.0118) (Supplementary Fig. 1B). Fig. 1 shows the differences in the absorbance of the above-mentioned peaks; a general increase was denoted in the endometrial cancer patients with the exception of the peaks at 1593 cm^{-1} and 1462 cm^{-1} which showed decreased levels. The means and SD values for these peaks were: 1593 cm^{-1} (mean/SD for healthy: 0.18/0.011; mean/SD for cancer: 0.164/0.0098), 1508 cm^{-1} (mean/SD for healthy: 0.0445/0.0115; mean/SD for cancer: 0.0506/0.0104), 1462 cm^{-1}

(mean/SD for healthy: 0.0958/0.0151; mean/SD for cancer: 0.0833/0.0064), 1400 cm^{-1} (mean/SD for healthy: 0.0509/0.007; mean/SD for cancer: 0.0645/0.0095), 1335 cm^{-1} (mean/SD for healthy: 0.0237/0.0053; mean/SD for cancer: 0.0315/0.007), 1041 cm^{-1} (mean/SD for healthy: 0.0286/0.0083; mean/SD for cancer: 0.0383/0.0109).

The peaks responsible for differentiation between healthy and ovarian cancer patients were: 1597 cm^{-1} (P <0.0001, 95% CI = -0.0173 to -0.0114), 1508 cm^{-1} (P <0.0001, 95% CI = 0.0038 to 0.0103), 1408 cm^{-1} (P <0.0001, 95% CI = 0.0102 to 0.0149), 1373 cm^{-1} (P <0.0001, 95% CI = 0.0076 to 0.0122), 1231 cm^{-1} (P <0.0001, 95% CI = 0.0042 to 0.0074), 1041 cm^{-1} (P <0.0001, 95% CI = 0.0063 to 0.0118) (Supplementary Fig. 1B). Similarly, to endometrial cancer, the majority of the peaks showed increased absorbance when cancer was present, apart from the peak at 1597 cm^{-1} (Fig. 2). Means and SD values for each of the abovementioned peaks were: 1597 cm^{-1} (mean/SD for healthy: 0.18/0.0104; mean/SD for cancer: 0.166/0.0098), 1508 cm^{-1} (mean/SD for healthy: 0.0445/0.0115; mean/SD for cancer: 0.0506/0.0104), 1408 cm^{-1} (mean/SD for healthy: 0.0548/0.0067; mean/SD for cancer: 0.0676/0.0085), 1373 cm^{-1} (mean/SD for healthy: 0.0326/0.0075; mean/SD for cancer: 0.0427/0.0085), 1231 cm^{-1} (mean/SD for healthy: 0.0198/0.0079; mean/SD for cancer: 0.0256/0.0079), 1041 cm^{-1} (mean/SD for healthy: 0.0286/0.0083; mean/SD for cancer: 0.0383/0.0109).

3.1 Classification algorithms to calculate diagnostic accuracy

All classification algorithms (PLS-DA, PCA-SVM and GA-LDA) were applied to the data after the same pre-processing (cut to the bio-fingerprint region [1800-900 cm^{-1}], rubberband baseline correction and vector normalisation).

PLS-DA was employed to differentiate healthy, endometrial and ovarian cancer samples using 10 latent variables (LVs), accounting for 96.02% of cumulative variance (Fig. 3A). The number of LVs was selected according to the lowest error of cross-validation (15.4%

for healthy; 16.2% for endometrial cancer; and 12.5% for ovarian cancer) and maximum explained variance (96.02%). PLS-DA loadings are shown in Fig. 3B, where the first three LVs have higher coefficients at ~1041 cm^{-1}, ~1082 cm^{-1}, ~1462 cm^{-1}, ~1547 cm^{-1}, ~1589 cm-1 and ~1670 cm^{-1}. The predicted classes for each sample spectrum analysed by PLS-DA are shown in Fig. 3C (healthy), 3D (endometrial cancer) and 3E (ovarian cancer), where a degree of superposition is observed among the three classes. Only the ovarian cancer dataset showed clearer separation from the other two classes.

PCA-SVM was performed using 10 principal components (PCs), accounting for 97.33% of cumulative variance (Fig. 4A). The PCA loadings (Fig. 4B) had higher coefficients in regions very similar to PLS-DA: ~1042 cm^{-1}, ~1090 cm^{-1}, ~1130 cm^{-1}, ~1462 cm^{-1}, ~1508 cm^{-1}, ~1543 cm^{-1}, ~1589 cm^{-1} and ~1667 cm^{-1}. The predicted classes for each sample are shown in Fig. 4C (healthy), 4D (endometrial cancer) and 4E (ovarian cancer). In comparison to PLS-DA, Fig. 4 C-E shows a clearer separation among the classes, with only a few samples being misclassified.

GA-LDA classified healthy, endometrial and ovarian cancer with a fitness of 1.53 (Fig. 5A). The GA-LDA discriminant function (DF) plot for the three classes is shown in Fig. 5B with clear segregation between the three classes. A total of 20 variables showed differences between the three classes: 922 cm^{-1}, 972 cm^{-1}, 1007 cm^{-1}, 1011 cm^{-1}, 1018 cm^{-1}, 1045 cm^{-1}, 1049 cm^{-1}, 1061 cm^{-1}, 1084 cm^{-1}, 1265 cm^{-1}, 1362 cm^{-1}, 1366 cm^{-1}, 1400 cm^{-1}, 1412 cm^{-1}, 1500 cm^{-1}, 1535 cm^{-1}, 1562 cm^{-1}, 1566 cm^{-1}, 1682 cm^{-1} and 1716 cm^{-1} (Fig. 5B).

Table 1 shows the classification rates achieved by the three algorithms, with PCA-SVM being superior. Both accuracy and sensitivity values for PCA-SVM model ranged from 92.5% (healthy) to 100% (ovarian cancer); and specificity ranged from 96.3% (ovarian cancer) to 100% (endometrial cancer). GA-LDA had accuracy ranging from 90.0% (endometrial cancer)

to 98.3% (ovarian cancer); sensitivity ranging from 70.0% (endometrial cancer) to 100% (healthy/ovarian cancer); and specificity ranging from 87.5% (healthy) to 100% (endometrial cancer). PLS-DA was the worst model as only the ovarian cancer data set had quality parameters as good as the other algorithms. Accuracy ranged from 57.5% (endometrial cancer) to 92.5% (ovarian cancer); sensitivity ranged from 62.5% (endometrial cancer) to 87.5% (ovarian cancer); and specificity ranged from 86.3% (healthy) to 90% (ovarian cancer). The classification rates for the training and test sets using all three algorithms are shown in Supplementary Table 1.

4. Discussion

The wavenumbers that were mostly responsible for segregation between the different classes could facilitate as potential diagnostic biomarkers. In endometrial cancer patients the majority of the IR bands, associated with proteins and nucleic acids, were increased in comparison to healthy individuals. This could potentially be due to an elevated concentration of biomolecules, previously suggested as biomarkers for endometrial cancer, such as human epididymis protein 4 (HE4), CA-125 or carcinoembryonic antigen (CEA) [7, 27, 28]; the increased level of nucleic acid may be caused by the unconstrained proliferation of cells. Only two out of the six discriminatory peaks were lower in the cancer cases; these were attributed to C-C vibrations of phenyl rings of proteins (~1593 cm^{-1}, Amide II) and CH_2 vibrations of lipids (~1462 cm^{-1}). These results could be potentially explained by a number of possible reasons. For instance, preceding research has demonstrated a simultaneous increased degradation of proteins as well as a decreased protein synthesis during cancer cachexia [29]. Another study demonstrated decreased expression of follicle-stimulating hormone (FSH) in endometrial cancer patients when these were compared to healthy controls [30]. The same study also showed decreased levels of matrix metalloproteinases (MMP), which is a family of enzymes implicated

in normal and pathological processes, and previously suggested as novel biomarkers and/or therapeutic targets in human cancer [31]. Also, apolipoprotein-1 (ApoA-1), prealbumin (TTR) have also been shown to be decreased in endometrial cancer patients [7].

With regards to the decrease in the lipid region of endometrial cancer cases, previous work may again justify the results of the current study. After studying a number of lipids in urine, Skotland et. al revealed that increased and/or decreased levels of molecular lipids could be used as non-invasive biomarkers for prostate cancer with high levels of diagnostic accuracy. Therefore, similar conclusions could possibly be extrapolated to endometrial cancer [32]. Previous research has also suggested that lipids, and specifically cholesterol, were lower in blood samples of endometrial cancer than controls [33]; this might explain the lower absorbance in the lipid region (~1462 cm^{-1}) of endometrial cancer patients. More recent studies have further confirmed the increased risk of low cholesterol concentration in other types of cancer as well, such as lung, prostate or colon [34, 35].

When we compared ovarian cancer patients with healthy controls, the discriminatory peaks were mainly attributed to proteins and nucleic acids. Increased levels of these biomolecules were observed in cancerous samples with an exemption of a peak 1597 cm^{-1} which was assigned to C-C phenyl ring of proteins. Continuous research has previously shown that a cancer biomarker can be either upregulated or downregulated. After reviewing several biomarkers, a total of 111 were found significantly altered between ovarian cancer and controls, with ~60% of them being elevated in cancer and ~40% decreased [36]. Some of the biomarkers showing lower levels in cancerous state are, for instance, ApoA-1, FSH, microtubule-associated protein 1 light chain 3 (LC3) and epidermal growth factor receptor (EGFR) [6, 37, 38]. Therefore, this may explain the observed decrease in the Amide II region. On the contrary, increased peaks could potentially be attributed to other established biomarkers such as HE4, previously found to be increased in urine samples of ovarian cancer patients, CA-125, cancer

antigen 15-3 (CA15-3) and others [6, 39]. A relatively recent study, also demonstrated increased levels of the anti-apoptotic protein B-cell lymphoma 2 (Bcl-2) when urine samples from ovarian cancer patients were analysed [40].

Three classification algorithms were employed to calculate the diagnostic accuracy with which spectroscopy identified endometrial and ovarian cancer. The optimal approach was PCA-SVM which identified endometrial cancer with 95% sensitivity and 100% specificity (95% accuracy) and ovarian cancer with 100% sensitivity and 96.3% specificity (100% accuracy), which are exceptionally high in comparison to conventional molecular and imaging methods.

Previously, numerous studies have investigated blood biomarkers as a relatively non-invasive approach towards diagnosis of endometrial cancer. A study using serum HE4 yielded sensitivity of 45.4% and 95% sensitivity [41]; another study developing a multimarker panel for the early detection of endometrial cancer suggested that prolactin could be used as an accurate biomarker with sensitivity and specificity of ~98% [30] . Combination of three different biomarkers (ApoA-1, prealbumin and transferrin) distinguished normal samples from early-stage endometrial cancers with 71% sensitivity and 88% specificity, as well as normal samples from late-stage cancer with 82% sensitivity and 86% specificity [7]. After reviewing 13 studies, Timmermans *et. al*., showed that ultrasonography achieved sensitivity and specificity of 90-98% and 35-54%, respectively [42]. Magnetic resonance imaging (MRI) has been shown to detect early and advanced endometrial cancer with high sensitivity (87-100%) and specificity (90-99%) but stage Ic and stage II disease had significantly reduced sensitivity (19-56%) whereas specificity remained high (86-96%) [43].

Currently, molecular tests measuring serum CA-125 for ovarian cancer, achieve sensitivity of only 50-60% for early-stage disease and specificity of >95% [6, 44]. Moreover,

transvaginal ultrasound (TVS), computed tomography (CT), MRI and power Doppler are of high-cost and achieve sensitivity <90% for early ovarian cases and relatively high false positive results which render them less useful for screening [6]. Combination of different biomarkers has been shown to achieve higher sensitivity and specificity values. For example, two combinations of serum biomarkers for ovarian cancer are CA-125, CA 72-4, CA 15-3 and macrophage colony-stimulating factor (M-CSF) [45], as well as CA-125, ApoA-1, a truncated form of transthyretin and a cleavage fragment of inter-alpha-trypsin inhibitor heavy chain H4 [46]; the above-mentioned combinations of biomarkers improved sensitivity and specificity to 70-73% and 97-98%, respectively. Even though the improved accuracy is acceptable, there is still room for improvement. A different study found that a blood-based assay of 11 analytes could distinguish ovarian cancer from benign case with sensitivity and specificity of 90% [37]. However, an important drawback of molecular methods is their expense and laborious sample preparation and analysis, in contrast to spectroscopic methods which are rapid and label-free.

5. Conclusion

This pilot study demonstrates the efficacy of ATR-FTIR spectroscopy in detecting endometrial and ovarian cancers in urine samples, with high levels of accuracy. Being rapid, non-destructive and at the same time cost-effective, spectroscopy is introduced as an ideal method for studying these types of cancer and could potentially be translated into clinical practise in the future as either a screening or diagnostic test. An adequately powered study will be required to demonstrate the true diagnostic accuracy and validate these preliminary results. Furthermore, the quick and non-invasive nature of urine collection and subsequent analysis has the potential of a preferable vehicle for repeated measurements, thus facilitating monitoring of disease progression/regression/recurrence or even therapeutic response.

Conflict of interest

There are no conflicts of interest to declare.

Acknowledgements

MP was supported by the Rosemere Cancer Foundation (RCF). CLMM would like to acknowledge CAPES for his research grant. KMGL acknowledges CNPq (grant 305962/2014-0) for financial support.

References

1. Cancer Reserach UK, http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/uterine-cancer/incidence (Accessed December 2017).
2. Cancer Research UK, http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/ovarian-cancer/incidence#heading-Six (Accessed October 2017).
3. D. W. Cramer, *Hematol Oncol Clin North Am*, 2012, **26**, 1-12.
4. P. Morice, A. Leary, C. Creutzberg, N. Abu-Rustum and E. Darai, *Lancet*, 2016, **387**, 1094-1108.
5. G. C. Jayson, E. C. Kohn, H. C. Kitchener and J. A. Ledermann, *Lancet*, 2014, **384**, 1376-1388.
6. Z. Yurkovetsky, S. Skates, A. Lomakin, B. Nolen, T. Pulsipher, F. Modugno, J. Marks, A. Godwin, E. Gorelik and I. Jacobs, *J Clin Oncol*, 2010, **28**, 2159-2166.
7. G. Farias-Eisner, F. Su, T. Robbins, J. Kotlerman, S. Reddy and R. Farias-Eisner, *Am J Obstet Gynecol*, 2010, **202**, 73.e71-73.e75.
8. M. Paraskevaidi, C. L. M. Morais, K. M. G. Lima, J. S. Snowden, J. A. Saxon, A. M. T. Richardson, M. Jones, D. M. A. Mann, D. Allsop, P. L. Martin-Hirsch and F. L. Martin, *Proc Natl Acad Sci USA*, 2017, **114**, E7929-e7938.
9. K. L. Andrew Chan and S. G. Kazarian, *Chem Soc Rev*, 2016, **45**, 1850-1864.
10. M. J. Baker, S. R. Hussain, L. Lovergne, V. Untereiner, C. Hughes, R. A. Lukaszewski, G. Thiefin and G. D. Sockalingum, *Chem Soc Rev*, 2016, **45**, 1803-1818.
11. 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 Protocols*, 2010, **5**, 1748-1760.
12. J. R. Hands, K. M. Dorling, P. Abel, K. M. Ashton, A. Brodbelt, C. Davis, T. Dawson, M. D. Jenkinson, R. W. Lea, C. Walker and M. J. Baker, *J Biophotonics*, 2014, **7**, 189-199.
13. J. Backhaus, R. Mueller, N. Formanski, N. Szlama, H.-G. Meerpohl, M. Eidt and P. Bugert, *Vib Spectrosc*, 2010, **52**, 173-177.
14. K. Gajjar, J. Trevisan, G. Owens, P. J. Keating, N. J. Wood, H. F. Stringfellow, P. L. Martin-Hirsch and F. L. Martin, *Analyst*, 2013, **138**, 3917-3926.
15. S. E. Taylor, K. T. Cheung, I. I. Patel, J. Trevisan, H. F. Stringfellow, K. M. Ashton, N. J. Wood, P. J. Keating, P. L. Martin-Hirsch and F. L. Martin, *Br J Cancer*, 2011, **104**, 790-797.
16. E. Søreide, L. Eriksson, G. Hirlekar, H. Eriksson, S. Henneberg, R. Sandin and J. Raeder, *Acta Anaesthesiol Scand*, 2005, **49**, 1041-1047.
17. O. Ljungqvist and E. Søreide, *Br J Surg*, 2003, **90**, 400-406.
18. J. E. de Aguilar-Nascimento and D. B. Dock-Nascimento, *World J Gastrointest Surg*, 2010, **2**, 57.
19. J. Trevisan, P. P. Angelov, A. D. Scott, P. L. Carmichael and F. L. Martin, *Bioinformatics*, 2013, **29**, 1095-1097.

20. J. G. Kelly, J. Trevisan, A. D. Scott, P. L. Carmichael, H. M. Pollock, P. L. Martin-Hirsch and F. L. Martin, *J Proteome Res*, 2011, **10**, 1437-1448.
21. R. W. Kennard and L. A. Stone, *Technometrics*, 1969, **11**, 137-148.
22. R. G. Brereton and G. R. Lloyd, *J Chemometrics*, 2014, **28**, 213-225.
23. R. Bro and A. K. Smilde, *Anal Methods*, 2014, **6**, 2812-2831.
24. C. L. Morais, F. S. Costa and K. M. Lima, *Anal Methods*, 2017, **9**, 2964-2970.
25. C. Cortes and V. Vapnik, *Mach Learn*, 1995, **20**, 273-297.
26. C. L. Morais and K. M. Lima, *Chemometrics Intellig Lab Syst*, 2017.
27. L. Zanotti, E. Bignotti, S. Calza, E. Bandiera, G. Ruggeri, C. Galli, G. Tognon, M. Ragnoli, C. Romani and R. A. Tassi, *Clin Chem Lab Med*, 2012, **50**, 2189-2198.
28. P. B. Panici, G. Scambia, G. Baiocchi, L. Perrone, S. Greggi, F. Battaglia and S. Mancuso, *Gynecol Obstet Invest*, 1989, **27**, 208-212.
29. K. Smith and M. Tisdale, *Br J Cancer*, 1993, **67**, 680.
30. Z. Yurkovetsky, S. Ta'asan, S. Skates, A. Rand, A. Lomakin, F. Linkov, A. Marrangoni, L. Velikokhatnaya, M. Winans and E. Gorelik, *Gynecol Oncol*, 2007, **107**, 58-65.
31. R. Roy, J. Yang and M. A. Moses, *J Clin Oncol*, 2009, **27**, 5287-5297.
32. T. Skotland, K. Ekroos, D. Kauhanen, H. Simolin, T. Seierstad, V. Berge, K. Sandvig and A. Llorente, *Eur J Cancer*, 2017, **70**, 122-132.
33. C. A. Swanson, N. Potischman, R. J. Barrett, M. L. Berman, R. Mortel, L. B. Twiggs, G. D. Wilbanks, R. N. Hoover and L. A. Brinton, *Cancer Epidemiol Biomarkers Prev*, 1994, **3**, 575-581.
34. H. Kikuchi, A. Nanri, A. Hori, M. Sato, K. Kawai, H. Kasai and T. Mizoue, *Nutr Metab*, 2013, **10**, 59.
35. M. Eichholzer, H. B. Stähelin, F. Gutzwiller, E. Lüdin and F. Bernasconi, *Am J Clin Nutr*, 2000, **71**, 569-574.
36. B. M. Nolen and A. E. Lokshin, *Future Oncol*, 2012, **8**, 55-71.
37. S. D. Amonkar, G. P. Bertenshaw, T. H. Chen, K. J. Bergstrom, J. Zhao, P. Seshaiah, P. Yip and B. C. Mansfield, *PLoS One*, 2009, **4**, e4599.
38. Y. Shen, D.-D. Li, L.-L. Wang, R. Deng and X.-F. Zhu, *Autophagy*, 2008, **4**, 1067-1068.
39. J. B. Liao, Y. Y. Yip, E. M. Swisher, K. Agnew, K. E. Hellstrom and I. Hellstrom, *Gynecol Oncol*, 2015, **137**, 430-435.
40. N. S. Anderson, Y. Bermudez, D. Badgwell, R. Chen, S. V. Nicosia, R. C. Bast and P. A. Kruk, *Gynecol Oncol*, 2009, **112**, 60-67.
41. R. G. Moore, A. K. Brown, M. C. Miller, D. Badgwell, Z. Lu, W. J. Allard, C. Granai, R. C. Bast and K. Lu, *Gynecol Oncol*, 2008, **110**, 196-201.
42. A. Timmermans, B. C. Opmeer, K. S. Khan, L. M. Bachmann, E. Epstein, T. J. Clark, J. K. Gupta, S. H. Bakour, T. van den Bosch, H. C. van Doorn, S. T. Cameron, M. G. Giusa, S. Dessole, F. P. Dijkhuizen, G. Ter Riet and B. W. Mol, *Obstet Gynecol*, 2010, **116**, 160-167.
43. O. Ortashi, S. Jain, O. Emannuel, R. Henry, A. Wood and J. Evans, *Eur J Obstet Gynecol Reprod Biol*, 2008, **137**, 232-235.
44. V. Nossov, M. Amneus, F. Su, J. Lang, J. M. T. Janco, S. T. Reddy and R. Farias-Eisner, *Am J Obstet Gynecol*, 2008, **199**, 215-223.
45. S. J. Skates, N. Horick, Y. Yu, F. J. Xu, A. Berchuck, L. J. Havrilesky, H. W. de Bruijn, A. G. van der Zee, R. P. Woolas, I. J. Jacobs, Z. Zhang and R. C. Bast, Jr., *J Clin Oncol*, 2004, **22**, 4059-4066.
46. Z. Zhang, R. C. Bast, Jr., Y. Yu, J. Li, L. J. Sokoll, A. J. Rai, J. M. Rosenzweig, B. Cameron, Y. Y. Wang, X. Y. Meng, A. Berchuck, C. Van Haaften-Day, N. F. Hacker, H. W. de Bruijn, A. G. van der Zee, I. J. Jacobs, E. T. Fung and D. W. Chan, *Cancer Res*, 2004, **64**, 5882-5890.

Figure Legends

Figure 1: Analysis of the top six discriminatory peaks between healthy controls and endometrial cancer patients.

Figure 2: Analysis of the top six discriminatory peaks between healthy controls and ovarian cancer patients.

Figure 3: Cumulative explained variance using PLS-DA (A); PLS-DA loadings on LV1, 2 and 3 (B); predicted healthy class *versus* endometrial and ovarian cancer (C); predicted endometrial cancer class *versus* healthy and ovarian cancer (D); predicted ovarian cancer class *versus* healthy and endometrial cancer (E). Class measured 1 = healthy control; 2 = endometrial cancer; 3 = ovarian cancer. LV: Latent Variable.

Figure 4: Cumulative explained variance using PCA (A); PCA loadings on PC1, 2 and 3 (B); predicted probability of healthy class *versus* endometrial and ovarian cancer (C); predicted probability of endometrial cancer class *versus* healthy and ovarian cancer (D); predicted probability of ovarian cancer class *versus* healthy and endometrial cancer (E). Class measured 1 = healthy control; 2 = endometrial cancer; 3 = ovarian cancer. PC: Principal Component.

Figure 5: Fitness function (A); Discriminant Function (DF) plot (B); and selected variables by GA-LDA (C).