Metabolomics (2014) 10:1113–1120
DOI 10.1007/s11306-014-0650-1

ORIGINAL ARTICLE

Application of gas chromatography mass spectrometry (GC–MS) in conjunction with multivariate classification for the diagnosis of gastrointestinal diseases

Michael Cauchi · Dawn P. Fowler · Christopher Walton · Claire Turner · Wenjing Jia · Rebekah N. Whitehead · Lesley Griffiths · Claire Dawson · Hao Bai · Rosemary H. Waring · David B. Ramsden · John O. Hunter · Jeffrey A. Cole · Conrad Bessant

Received: 22 January 2014 / Accepted: 6 March 2014 / Published online: 20 March 2014

Abstract Gastrointestinal diseases such as irritable bowel syndrome, Crohn's disease (CD) and ulcerative colitis are a growing concern in the developed world. Current techniques for diagnosis are often costly, time consuming, inefficient, of great discomfort to the patient, and offer poor sensitivities and specificities. This paper describes the development and evaluation of a new methodology for the non-invasive diagnosis of such diseases using a combination of gas chromatography mass spectrometry (GC–MS) and chemometrics. Several potential sample matrices were tested: blood, breath, faeces and urine. Faecal samples provided the only statistically significant results, providing discrimination between CD and healthy controls with an overall classification accuracy of 85 % (78 % specificity; 93 % sensitivity). Differentiating CD from other diseases proved more challenging, with overall classification accuracy dropping to 79 % (83 % specificity; 68 % sensitivity). This diagnostic performance compares well with the gold standard technique of colonoscopy, suggesting that GC–MS may have potential as a non-invasive screening tool.

Keywords Gastrointestinal diseases · Irritable bowel syndrome · Crohn's disease · Ulcerative colitis · Gas chromatography mass spectrometry · Partial least squares discriminant analysis

1 Introduction

Gastrointestinal diseases are an increasing cause for concern as incidence rates, particularly in the developed world, are rising annually (Abraham and Cho 2009). Diseases of particular concern are irritable bowel syndrome (IBS), Crohn's disease (CD) and ulcerative colitis (UC), the two latter grouped together as inflammatory bowel disease (IBD), caused by inflammation of the mucosal lining of the

Electronic supplementary material The online version of this article (doi:10.1007/s11306-014-0650-1) contains supplementary material, which is available to authorized users.

M. Cauchi (✉)
Centre for Biomedical Engineering, School of Engineering,
Building 63, Cranfield University, Bedfordshire MK43 0AL, UK
e-mail: m.cauchi@cranfield.ac.uk

D. P. Fowler · C. Walton
School of Applied Sciences, Cranfield University,
Bedfordshire MK43 0AL, UK

C. Turner
The Department of Life, Health and Chemical Sciences, Open
University, Milton Keynes MK7 6AA, UK

W. Jia · R. N. Whitehead · L. Griffiths ·
R. H. Waring · D. B. Ramsden · J. A. Cole
School of Biosciences, University of Birmingham,
Birmingham B15 2TT, UK

C. Dawson · J. O. Hunter
Gastroenterology Research Unit, Addenbrooke's Hospital,
Box 262, Cambridge CB2 0QQ, UK

H. Bai
School of Electronic, Electrical and Computer Engineering,
University of Birmingham, Birmingham B15 2TT, UK

C. Bessant
School of Biological and Chemical Sciences, Queen Mary
University of London, Mile End Road, London E1 4NS, UK

gut. These two have traits that overlap, making them difficult to separate in diagnosis (Nakamura et al. 2003; von Stein et al. 2007; von Stein et al. 2008). In some cases, a patient previously diagnosed with UC can later be re-diagnosed with CD (Moum et al. 1997; Seidman and Deslandres 1997). It is known that CD can affect any part of the gastrointestinal tract whereas UC will only affect the large colon (Nakamura et al. 2003). IBS is a functional disorder in which inflammation in the gut is minimal.

The characteristic trait of IBS is malaise accompanied by abdominal pain and coincides with a change in the frequency and/or consistency of faeces (Nakamura et al. 2003). CD typically affects the lower part of the ileum (Nakamura and Barry 2001), and like UC is a chronic disease with unknown aetiology. The mucosa and submucosa of the large intestine becomes inflamed and a visible indication is diarrhoea mixed with blood. (Nakamura and Barry 2001). It has been reported that surgery can be used to alleviate discomfort (Vella et al. 2007).

1.1 Current diagnostic methods

The current gold standard methods for diagnosing CD and UC are colonoscopy (Fefferman and Farrell 2005; Manes et al. 2009) and sigmoidoscopy (Stange et al. 2008). These methods are invasive, of some discomfort to the patient and are also expensive. Blood tests may reveal the presence of inflammation in the body, particularly in IBD (Makidono et al. 2004). Radiography to image the gut is now less used as it is less accurate than endoscopy and because radiation carries health risks.

Alternative methods of diagnosis are continuously being investigated. Genetic factors are believed to be very important in the pathogenesis of UC and CD (Papadakis and Targen 1999; Farrell et al. 2001), and a multi-gene approach that involves the identification of specific IBD genes originating from colonoscopic biopsies (von Stein et al. 2007; von Stein et al. 2008) has been proposed. By using quantitative polymerase chain reaction (PCR) experiments, seven differentially expressed marker genes were identified. The expression levels of the differentially expressed genes were used in a classification algorithm to distinguish between the diseases. This led to a classification accuracy of 86–92 % which compared well to the clinical approach. IBD specificity and sensitivity ranged from 94–95 to 85–95 %, respectively. However, these tests require initial colonoscopy and are therefore limited in their application.

Enzymes found in faeces can be employed as markers in order to identify CD and UC, in particular calprotectin and lactoferrin via the ELISA (enzyme-linked immunosorbent assay) method (Angriman et al. 2007; Mendoza and Abreu 2009). As well as detecting IBD, they may be used to monitor response to medical therapy, and predict the recurrence of disease after surgery.

Proton transfer mass spectrometry has been employed to analyse the headspace of fluid extracted from the gut during a colonoscopy and to compare with breath samples (Lechner et al. 2005). Significant differences were observed in the exhaled breath and also the headspace, however it was also reported that it was difficult to distinguish IBS from healthy controls particularly in the fluid samples. It was acknowledged that gas chromatography mass spectrometry (GC–MS) would be a better tool for this task.

1.2 GC–MS as a diagnostic tool

Gas chromatography mass spectrometry (GC–MS) has become an important technique in the field of metabolomics, due to its high sensitivity, reproducibility and good peak resolution, all of which results in large information-rich datasets (Pasikanti et al. 2008). Walton et al. employed GC–MS to identify a number of key volatile organic compounds (VOCs) in faeces and showed that differences in abundance could lead to the distinction between CD and other gastrointestinal diseases (Walton et al. 2013).

Extracting useful information from the acquired GC–MS data requires sophisticated multivariate data analysis methods provided by chemometrics (Otto 1999; Brereton 2003; Lavine and Workman 2004; Kussmann et al. 2006). Multivariate analysis of GC–MS data typically involves an exploratory approach followed by pattern recognition (Otto 1999). The former often utilises principal components analysis (PCA) to reduce the dimensionality of the data to assist in recognising trends within the dataset and identifying outlying samples (Wold et al. 1987). Pattern recognition in the form of multivariate classification aims to determine which samples belong to a designated class (Pasikanti et al. 2008). One commonly employed classification technique is partial least squares discriminant analysis (PLS-DA) (Barker and Rayens 2003; Wiklund et al. 2007). This is a supervised technique that permits the separation of samples into different classes, for example cases versus controls. Other machine learning algorithms are also available such as support vector machines (Sattlecker et al. 2010) and artificial neural networks (Hagan et al. 1996), however PLS-DA is often preferred due to the relative ease with which models can be built, tested and optimised. For example, PLS-DA was recently employed to successfully diagnose patients with bladder cancer from GC–MS data, attaining 100 % sensitivity (Pasikanti et al. 2010).

1.3 Application of GC–MS and PLS-DA to diagnosis of gastrointestinal diseases

The hypothesis underlying the work presented in this paper is that a combination of GC–MS and PLS-DA could be used to identify disease-specific metabolomic fingerprints from easily accessible biological samples (blood, urine, faeces and breath). If characteristic profiles can be found for each of the three gastrointestinal diseases (IBS, CD and UC) this could form the basis of a new diagnostic tool. Specifically, this study looked at the ability to distinguish these diseases from healthy controls, but more importantly, from each other. The latter step is of great relevance because it gives an indication of the efficiency of the classification model to discriminate the target disease from not only healthy controls but from other diseases, which is exactly the challenge faced in real clinical applications.

2 Experimental

2.1 Reagents

Analytical grade reagents and solvents were employed throughout, unless otherwise stated.

2.2 Sampling

2.2.1 Selection of candidates

A total of 91 candidates were selected from patients attending the Gastroenterology Department at Addenbrooke's Hospital (Cambridge, UK) whom had been invited to participate in the study, and had provided written consent. Using the conventional diagnostic techniques of radiology, endoscopy and histology of intestinal biopsies, 24 were found to have CD, 19 had UC, and 28 had IBS. The remaining 20 candidates, who had given negative results to questionnaires, were recruited as healthy control subjects. They were friends of members of the gastroenterology department who lived and worked outside the hospital, and who were in good health and taking no medication other than the oral contraceptive and who met the inclusion criteria (age 18–65; fit and healthy; not pregnant or lactating; no history of GI disease; had taken no antibiotics in the previous 6 weeks and were not on chronic medication, e.g. for blood pressure, diabetes, etc.). Every candidate was expected to give a breath, urine, faecal and blood sample.

The study received ethical approval from the National Research Ethics Service in Leeds (West) in July 2007 (07/Q1205/39) and participants' samples were anonymised.

2.2.2 Sample collection

Sample collection was achieved over a period of several months. In order to collect the faecal samples, sample pots were supplied to each candidate to fill. Urine and blood samples were taken on site at Addenbrooke's. The breath samples were also collected on site. A detailed description of the collection of each sample type is given below.

Breath was sampled directly onto thermal desorption (TD) tubes using a device constructed for the purpose. The volunteer wore a facemask with full head harness covering nose and mouth, and breathed normally. The mask incorporated non-return valves and a side-stream sample of exhaled breath was obtained from the exhalation pathway. Sampling flow rate was 200 ml min^{-1}, controlled using a pump and mass flow controller. The device was under software control such that the alveolar portion of a number of breaths could be sampled and accumulated into the TD tube. Sampling was continued until a total volume of 500 ml of breath had been obtained in this way, the process taking typically 5 min.

Faeces were collected into 60 ml sterile polypropylene vials. The samples were refrigerated and then transferred by specialist courier on dry ice to the laboratory whereupon they were immediately frozen at −80 °C prior to analysis.

Urine was collected into 60 ml sterile polypropylene vials and then immediately decanted into three aliquots in polypropylene sterile universal bottles, before being immediately frozen at −80 °C until analysis. It has been reported that the vapour emissions from human urine are not significantly changed by the freezing process on samples to be analysed via GC–MS (Peakman and Elliott 2008).

Blood for this study was collected into vacutainer tubes containing EDTA to prevent clotting. The whole blood was then stored at −80 °C prior to analysis.

2.2.3 Sample preparation

Gas sampling bags for headspace analysis of faeces urine and blood were constructed from Nalophan NA® (65 mm inflated diameter) which were cut into lengths of 500 mm. To the apical end was attached a Swagelok® fitting whilst the basal end was left open so that the sample could be added inside after which the bag was sealed with tie-wraps. The bag was filled with hydrocarbon-free air after the sample was added, and then incubated for 30 min at 40 °C. A volume of 500 ml headspace was then pumped across a pre-packed 1:1 Carbotrap/Tenax thermal desorption (TD) tube using a Flec pump (both Markes International Ltd., Llantrisant, UK).

2.3 GC–MS measurements

An internal standard solution comprising 50 ng d8-toluene (Supelco Cat No. 48593) in methanol was added to each TD tube according to the manufacturer's instructions. Headspace samples were analysed by automated thermal desorption gas chromatography mass spectrometry (ATD–

GC–MS). A Perkin Elmer system was used for analysis, combining a TurboMass MS 4.1, Autosystem XL GC and Automatic Thermal Desorption system ATD 400 (Perkin Elmer, Wellesley, MA, USA). The carrier gas was CP grade helium (BOC gases, Guildford, UK) passed through a combined trap for removal of hydrocarbons, oxygen and water vapour. A wall-coated Zebron ZB624 chromatographic column was used (Phenomenex, Torrance, CA, USA), with dimensions 30 m × 0.4 mm × 0.25 mm (internal diameter), the liquid phase comprising a 0.25 µm layer of 6 % cyanopropylphenyl and 94 % methylpolysiloxane.

TD tubes were initially purged for 2 min in order to remove air and water vapour and then desorbed for 5 min at 300 °C. The ATD valve temperature was set to 180 °C and TD tubes were desorbed onto the secondary cold trap which was initially maintained at 30 °C. Once desorption was complete, the secondary trap was heated to 320 °C using the fastest available heating rate and then maintained for 5 min whilst the effluent was transferred to the GC via a transfer line heated to 210 °C. The GC oven was maintained at 50 °C for 4 min after injection and then raised at a rate of 10 °C per minute until reaching 220 °C and then held for 9 min. The eluted products were transferred by a heated line held at 240 °C to the mass spectrometer where the compounds were subjected to electron ionisation. Full scan mode was selected with mass/charge ratios from 33 to 350 m/z with a scan time of 0.3 and 0.1 s inter-scan delay to produce a total ion count (TIC) chromatogram.

2.4 Data analysis

Data analysis was performed in Matlab (v2008a, Mathworks Inc., USA) employing functions contained within the PLS Toolbox (v3.5, Eigenvector Research Inc., USA).

2.4.1 Data pre-processing

The intensity value of the deuterated toluene ion (m/z 98) peak was obtained and all other intensity values in the GC–MS matrix were normalised against this value. The normalised values were summed across the columns in order to obtain the TIC for that sample. This was repeated for every sample until a data matrix was constructed whose order was the number of samples (rows) by the retention time (columns).

Exploratory data analysis via PCA (Wold et al. 1987) was performed in order to visually identify any specific trends. Hotelling's T^2 (Hotelling 1931), an extension of Student's *t* test, was used to identify outlying samples. Removal of such samples is a prerequisite of the correlation optimised warping (COW) algorithm (Tomasi et al. 2004) that was used to align the chromatographic peaks. The COW algorithm was chosen as it is able to automatically deduce the optimal parameters required for aligning the retention time peaks.

2.4.2 Multivariate classification

The multivariate statistical method of PLS-DA (Barker and Rayens 2003) was used to perform the pattern recognition necessary to separate samples into different disease states according to their GC–MS TICs. The basic workflow underlying the PLS-DA approach is that a set of data from samples of known disease state is used to train a PLS-DA model [the model essentially being a set of linear equations contained within the underlying SIMPLS algorithm (de Jong 1993)] to recognise the characteristic profiles of the different disease states. This model is then used to predict the disease state of a further set of samples for which the disease state is withheld. By comparing the predictions with the clinical diagnosis of those samples it is possible to determine the diagnostic performance of the pattern recognition model. Key performance indicators include sensitivity, specificity, the overall accuracy quoted as percentage of samples correctly classified (%CC)—determined by calculating the ratio of the number of samples correctly classified against the total number of samples analysed—and the area under the received operator characteristics curve (AUROC) (Rahman and Schmeisser 1990; Zweig and Campbell 1993).

2.4.3 Robust performance metrics

To ensure that our results provided a true representation of the diagnostic performance of the GC–MS approach, we applied PLS-DA within a workflow (see Fig. 1) that provides a robust determination of diagnostic performance and an indication of the statistical significance of this performance. This workflow uses a two-stage bootstrapping process. The first stage deduces the optimum classification model after a randomised split of the samples, and the second stage evaluates the performance of the optimum model. The process is repeated 150 times in order to attain an average performance of all models created from different data splits. This ensures that model performance is not overestimated. This is also performed in conjunction with scaling (mean-centring, autoscaling, range-scaling and normalisation) (Hagan et al. 1996; van den Berg et al. 2006) and a heuristic approach to bootstrapping that uses leave-one-out cross-validation (LOO-CV) as an optimisation step. This therefore maximises the relevance of the performance metrics to a large clinical population. The statistical significance of the resulting model's diagnostic performance is determined by performing permutation testing (Westerhuis et al. 2008; Brereton 2009), which involves randomising the class assignations 300 times, and for each assignation, performing the two-

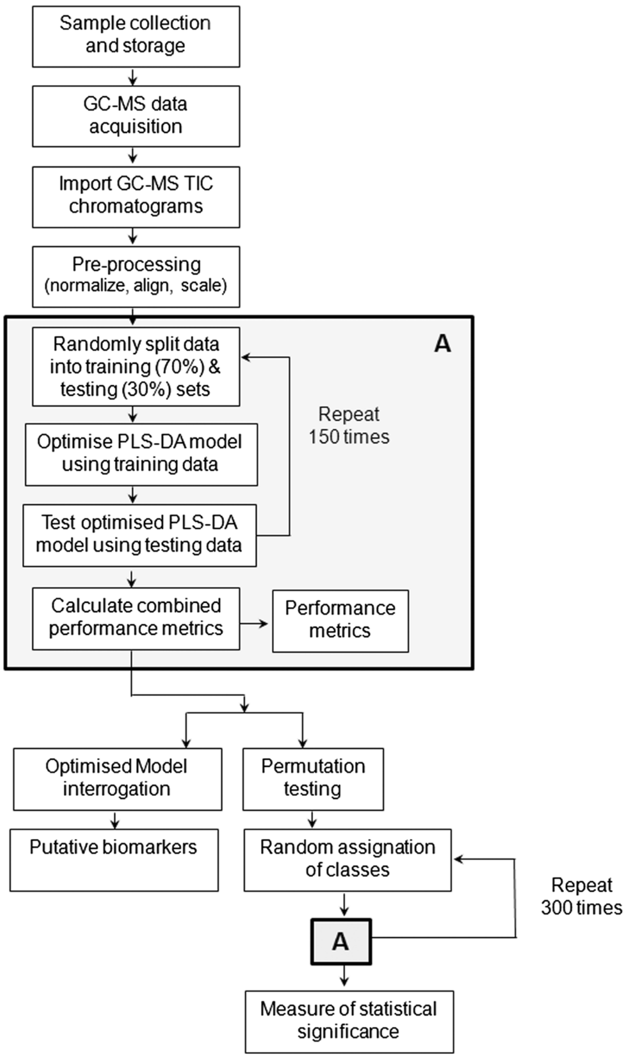
Sample collection and storage
GC-MS data acquisition
Import GC-MS TIC chromatograms
Pre-processing (normalize, align, scale)
A
Randomly split data into training (70%) & testing (30%) sets
Optimise PLS-DA model using training data
Test optimised PLS-DA model using testing data
Repeat 150 times
Calculate combined performance metrics
Performance metrics
Optimised Model interrogation
Putative biomarkers
Permutation testing
Random assignation of classes
A
Repeat 300 times
Measure of statistical significance

Fig. 1 Diagrammatic summary of workflow undertaken in the study. A combination of bootstrap resampling and permutation testing was used to maximise the robustness of diagnostic performance metrics and provide an indication of the statistical significance of the results obtained

stage bootstrapping process already described. This results in a second set of results, with its own distribution. The two distributions are analysed to reveal the amount of overlap and the statistical significance of the mean average performance for the real and permuted datasets calculated using the two-sample two-tailed z-test (Campbell and Machin 1999) (where $\alpha = 0.05$, equivalent to a 95 % confidence level).

3 Results and discussion

3.1 Case against control

Principal components analysis did not reveal any distinct separation between cases and controls. It was however able to assist in identifying a total of seven outlying samples which were subsequently removed prior to peak alignment.

Table 1 summarises the PLS-DA classification performance for each respective case (disease) against the healthy control in all four sample matrices. These results strongly suggest that discrimination between CD and healthy controls in faecal samples is possible having attained an overall %CC of 85 % and correspondingly high sensitivity and specificity. Importantly, the permutation testing showed this result to be highly significant, with a z-test p value of $<1 \times 10^{-6}$. Colonoscopy, generally considered the gold standard, is variably reported to be 79–95 % accurate (Langhorst et al. 2007) suggesting that the GC–MS method is comparable to this traditional approach. The table shows statistically significant results for some other combinations of disease and sample matrix but the %CC and sensitivity in these cases is generally too low to be of diagnostic relevance.

3.2 Case against all

In clinical practice, a patient presenting at a gastroenterology clinic could be suffering from one of a range of diseases so the ability to distinguish between a single disease (e.g. CD) and healthy controls is overly simplistic. The ability to be able to successfully diagnose each disease from a population in which other diseases exist is therefore of paramount importance. The data analysis was therefore repeated, with the aim of determining one case (e.g. CD) against the other cases (e.g. IBS and UC) in addition to the healthy controls.

As in Sect. 3.1, PCA did not reveal any distinct class separations but it did assist in identifying four outlying samples that were subsequently removed prior to peak alignment. Table 2 summarises the optimum results attained for the classification of each respective case (disease) against all other cases (including the healthy controls) in all four sample matrices. These results show that the only case which can be reliably diagnosed in the presence of other cases was CD in faecal samples, achieving a %CC of 79 % and high sensitivity and specificity scores. Figure 2 shows the results of permutation analysis used to determine the statistical significance of this performance. The mean average %CC for the permuted data can be seen to be 62 %. The position of the mean can be attributed to there being more negatively assigned samples (i.e. apparently healthy controls, IBS and UC) than positively-assigned samples (i.e. CD). The best mean %CC observed for the real data was 79 %, which the z-test shows to be significantly different to the 62 % mean %CC of the permuted samples, with a p value of $<1.0 \times 10^{-6}$.

There are other statistically significant results in Table 2. However, compared to the standout results for CD diagnosis from faeces, the %CC is at least 7 % lower in

Table 1 PLS-DA classification performance for different sample matrices after normalization against the deuterated toluene peak on importing the data for disease case versus apparently healthy controls

Sample matrix	Case	Scale	Overall classified (i.e. %CC)	Specificity (%)	Sensitivity (%)	AUROC[a]	*p* value from *z*-test ($\alpha = 0.05$)	SSD[b]
Blood	CD	MC	53	49	59	0.60	0.2147	No
	IBS	MC	53	67	32	0.40	0.3327	No
	UC	MC	54	63	41	0.46	0.1582	No
Breath	CD	Norm	53	51	55	0.54	0.0438	Yes
	IBS	Norm	58	72	41	0.44	$<1.0 \times 10^{-6}$	Yes
	UC	MC	61	86	11	0.28	0.0006	Yes
Faecal	**CD**	**Norm**	**85**	**78**	**93**	**0.97**	$\mathbf{<1.0 \times 10^{-6}}$	**Yes**
	IBS	MC	61	71	51	0.63	$<1.0 \times 10^{-6}$	Yes
	UC	RS2	58	69	43	0.54	$<1.0 \times 10^{-6}$	Yes
Urine	CD	Norm	53	69	27	0.31	0.7410	No
	IBS	RS1	64	80	38	0.53	$<1.0 \times 10^{-6}$	Yes
	UC	Norm	56	79	18	0.27	0.6384	No

The significance of the result for each combination of sample matrix and disease is determined from the permutation testing results using a *z*-test. The complete set of results for accuracies, specificities and sensitivities attained for each case for each of the scaling methods in each of the sample matrices are available in the Supplementary Material (section SM1). Bold highlights best result attained

None no scaling, *AS* auto-scaling, *MC* mean-centring, *RS1* range-scaling (0–1), *RS2* range-scaling (−1 to 1), *Norm* normalised

[a] Area under the receiver operating characteristic (AUROC) curve

[b] SSD denotes whether the *z*-test was statistically significantly different (the difference in means between the two distributions)

Table 2 PLS-DA classification performance for different sample matrices after normalization against the deuterated toluene peak on importing the data for disease case versus all cases (including apparently healthy controls)

Sample matrix	Case	Scale	Overall Classified (i.e. %CC)	Sensitivity (%)	Specificity (%)	AUROC[a]	*p* value from *z*-test ($\alpha = 0.05$)	SSD[b]
Blood	CD	RS1	68	49	75	0.52	$<1.0 \times 10^{-6}$	Yes
	IBS	MC	69	25	87	0.50	$<1.0 \times 10^{-6}$	Yes
	UC	Auto	69	34	79	0.57	0.0011	Yes
Breath	CD	Auto	64	24	83	0.57	0.0005	Yes
	IBS	Norm	64	22	80	0.50	0.1932	No
	UC	None	77	10	90	0.51	0.00002	Yes
Faecal	**CD**	**RS2**	**79**	**68**	**83**	**0.65**	$\mathbf{<1.0 \times 10^{-6}}$	**Yes**
	IBS	RS1	59	22	72	0.44	0.1615	No
	UC	Norm	68	32	78	0.52	0.00006	Yes
Urine	CD	None	72	48	81	0.59	$<1.0 \times 10^{-6}$	Yes
	IBS	Norm	67	29	79	0.48	0.0007	Yes
	UC	Norm	68	11	87	0.26	0.0659	No

The significance of the result for each combination of sample matrix and disease is determined from the permutation testing results using a *z*-test. The complete set of results for accuracies, specificities and sensitivities attained for each case for each of the scaling methods in each of the sample matrices are available in the Supplementary Material (section SM2). Bold highlights best result attained

Key as for Table 1

these other cases and sensitivities are all below 50 %. These performance metrics are far below what would be needed for clinical practice. To help understand what is limiting diagnostic performance, Table 3 shows how samples of each type were classified by the various diagnostic models. As might be expected from the previously reported results, faecal samples from CD sufferers were the only sample type for which the majority of samples were correctly classified (68 %, corresponding to the sensitivity in Table 2). However, even for the faecal analysis, 9 % of IBS patients, 14 % of UC patients and 10 % of controls were also diagnosed as having CD. The confusion between CD and UC is perhaps unsurprising given that they are both IBDs, and diagnosis is often difficult.

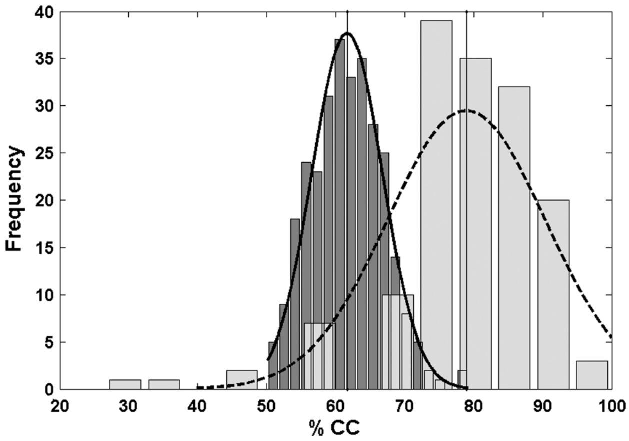

Fig. 2 Distribution of the overall percentage of samples classified for **a** *light grey bars* from prediction of CD versus all other cases from faecal samples (150 evaluation loops, mean average 79 %), **b** *dark grey bars* after randomised assignation (300 times, mean average 62 %)

Table 3 Illustration of the ability of the model developed for each disease and sample matrix to classify the samples

Sample matrix	Percentage of samples positively classified by each classifier			
	Case	CD model (%)	IBS model (%)	UC model (%)
Faecal	CD	**68**	18	18
	IBS	9	**22**	15
	UC	14	15	**32**
	CTRL	10	18	13
Urine	CD	**48**	12	9
	IBS	13	**29**	11
	UC	13	13	**11**
	CTRL	14	15	8
Breath	CD	**24**	13	9
	IBS	9	**22**	7
	UC	10	15	**10**
	CTRL	10	13	8
Blood	CD	**49**	9	16
	IBS	13	**25**	14
	UC	13	9	**34**
	CTRL	19	6	15

Bold values indicate same case comparison, e.g. case IBS versus IBS model (%)

4 Concluding remarks

A wide ranging study was carried out to determine whether a combination of GC–MS and PLS-DA has potential for the diagnosis of three gastrointestinal diseases (IBS, CD and UC) from easily accessible biological samples (blood, breath, urine and faeces). Using thorough statistical evaluation techniques it was found that the GC–MS approach can be used to diagnose CD from faecal samples. The proportion of samples correctly classified when seeking to identify CD patients from among a population representative of those presenting at a gastroenterology clinic was 79 % for CD (68 % sensitivity; 83 % specificity), which is on a par with the current gold standard approach of colonoscopy.

The fact that the diagnostic performance achieved for CD from faeces is so much better than that achieved for the other combinations of diseases and sample matrices further serves to highlight the significance of this finding. Based on these results it is recommended that any future work in this field concentrates on the analysis of faecal samples. Obvious improvements recommended for future studies would be the collection of a larger dataset from a larger patient cohort originating from multiple clinics, and the use of more advanced multivariate classification methods such as support vector machines. Both these measures would be expected to boost diagnostic performance.

The authors stress that this test is not a replacement for colonoscopy. Its role would be to speed up the diagnosis of IBD which currently may be greatly delayed particularly in cases of CD (Schoepfer et al. 2013). The gold standard will always be colonoscopy as this enables minute examination of the intestinal mucosa together with biopsy and polyp removal. The faecal screening may prove a very valuable test for detecting IBD in patients with symptoms but a colonoscopy would still subsequently be necessary to confirm the diagnosis and thus our testing would be offered as an initial simple straightforward screening technique with no risk to the patient (there is a 1 % risk of perforation and 1 % risk of haemorrhage with colonoscopy).

Acknowledgements We gratefully acknowledge the Wellcome Trust for funding the work (Project 080238/Z/06/Z).

References

Abraham, C., & Cho, J. H. (2009). Inflammatory bowel disease. *New England Journal of Medicine, 361*(21), 2066–2078.

Angriman, I., Scarpa, M., et al. (2007). Enzymes in feces: Useful markers of chronic inflammatory bowel disease. *Clinica Chimica Acta, 381*(1), 63–68.

Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics, 17*(3), 166–173.

Brereton, R. G. (2003). *Chemometrics: Data analysis for the laboratory and chemical plant*. Chichester: Wiley.

Brereton, R. G. (2009). *Chemometrics for pattern recognition*. Chichester: Wiley.

Campbell, M. J., & Machin, D. (1999). *Medical statistics: A common sense approach*. Chichester: Wiley.

de Jong, S. (1993). SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems, 18*(3), 251–263.

Farrell, R. J., Banerjee, S., et al. (2001). Recent advances in inflammatory bowel disease. *Critical Reviews in Clinical Laboratory Sciences, 38*(1), 33–108.

Fefferman, D. S., & Farrell, R. J. (2005). Endoscopy in inflammatory bowel disease: Indications, surveillance and use in clinical practice. *Clinical Gastroenterology and Hepatology, 3*, 11–24.

Hagan, M. T., Demuth, H. B., et al. (1996). *Neural network design*. Boston: International Thompson Publishing.

Hotelling, H. (1931). The generalization of student's ratio. *Annals of Mathematics and Statistics, 2*(3), 360–378.

Kussmann, M., Raymond, F., et al. (2006). OMICS-driven biomarker discovery in nutrition and health. *Journal of Biotechnology, 124*(4), 758–787.

Langhorst, J., Kühle, C. A., et al. (2007). MR colonography without bowel purgation for the assessment of inflammatory bowel diseases: Diagnostic accuracy and patient acceptance. *Inflammatory Bowel Diseases, 13*(8), 1001–1008.

Lavine, B., & Workman, J. J. (2004). Chemometrics. *Analytical Chemistry, 76*(12), 3365–3372.

Lechner, M., Colvin, H. P., et al. (2005). Headspace screening of fluid obtained from the gut during colonoscopy and breath analysis by proton transfer reaction–mass spectrometry: A novel approach in the diagnosis of gastro-intestinal diseases. *International Journal of Mass Spectrometry, 243*(2), 151–154.

Makidono, C., Mizuno, M., et al. (2004). Increased serum concentrations and surface expression on peripheral white blood cells of decay-accelerating factor (cd55) in patients with active ulcerative colitis. *Journal of Laboratory and Clinical Medicine, 143*(3), 152–158.

Manes, G., Imbesi, V., et al. (2009). Use of colonoscopy in the management of patients with Crohn's disease: Appropriateness and diagnostic yield. *Digestive and Liver Disease, 41*(9), 653–658.

Mendoza, J. L., & Abreu, M. T. (2009). Biological markers in inflammatory bowel disease: Practical consideration for clinicians. *Gastroentérologie Clinique et Biologique, 33*(Supplement 3), S158–S173.

Moum, B., Ekbom, A., et al. (1997). Inflammatory bowel disease: Re-evaluation of the diagnosis in a prospective population based study in south eastern Norway. *Gut, 40*, 328–332.

Nakamura, R. M., & Barry, M. (2001). Serologic markers in inflammatory bowel disease (IBD). *Medical Laboratory Observations, 33*, 8–15.

Nakamura, R. M., Matsutani, M., et al. (2003). Advances in clinical laboratory tests for inflammatory bowel disease. *Clinica Chimica Acta, 335*(1–2), 9–20.

Otto, M. (1999). *Chemometrics: statistics and computer applications in analytical chemistry*. Germany: Wiley.

Papadakis, K. A., & Targen, S. A. (1999). Current theories of the causes of inflammatory bowel disease. *Gastroenterology Clinics of North America, 28*, 283–296.

Pasikanti, K. K., Esuvaranathan, K., et al. (2010). Noninvasive urinary metabonomic diagnosis of human bladder cancer. *Journal of Proteome Research, 9*(6), 2988–2995.

Pasikanti, K. K., Ho, P. C., et al. (2008). Gas chromatography/mass spectrometry in metabolic profiling of biological fluids. *Journal of Chromatography B, 871*(2), 202–211.

Peakman, T. C., & Elliott, P. (2008). The UK Biobank sample handling and storage validation studies. *International Journal of Epidemiology, 37*(suppl 1), i2–i6.

Rahman, Q., & Schmeisser, G. (1990). Characterization of the speed of convergence of the trapezoidal rule. *Numerische Mathematik, 57*(1), 123–138.

Sattlecker, M., Bessant, C., et al. (2010). Investigation of support vector machines and Raman spectroscopy for lymph node diagnostics. *Analyst, 135*(5), 895–901.

Schoepfer, A. M., Dehlavi, M.-A., et al. (2013). Diagnostic delay in Crohn's disease is associated with a complicated disease course and increased operation rate. *American Journal of Gastroenterology, 108*(11), 1744–1753.

Seidman, E., & Deslandres, C. (1997). *Pitfalls in the diagnosis and management of pediatric IBD*. Lancaster: Kluwer Academic Publishing.

Stange, E. F., Travis, S. P. L., et al. (2008). European evidence-based consensus on the diagnosis and management of ulcerative colitis: Definitions and diagnosis. *Journal of Crohn's and Colitis, 2*(1), 1–23.

Tomasi, G., van den Berg, F., et al. (2004). Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *Journal of Chemometrics, 18*(5), 231–241.

van den Berg, R., Hoefsloot, H., et al. (2006). Centering, scaling, and transformations: Improving the biological information content of metabolomics data. *BMC Genomics, 7*(1), 142.

Vella, M., Masood, M. R., et al. (2007). Surgery for ulcerative colitis. *The Surgeon, 5*(5), 355–362.

von Stein, P., Kouznetsov, N., et al. (2007). P032 multi-gene approach to discriminate for ulcerative colitis, Crohn's disease and irritable bowel syndrome. *Journal of Crohn's and Colitis Supplements, 1*(1), 12.

von Stein, P., Lofberg, R., et al. (2008). Multigene analysis can discriminate between ulcerative colitis, Crohn's disease, and irritable bowel syndrome. *Gastroenterology, 134*(7), 1869–1881.

Walton, C., Fowler, D. P., et al. (2013). Analysis of volatile organic compounds of bacterial origin in chronic gastrointestinal diseases. *Inflammatory Bowel Diseases, 19*(10), 2069–2078.

Westerhuis, J., Hoefsloot, H., et al. (2008). Assessment of PLSDA cross validation. *Metabolomics, 4*(1), 81–89.

Wiklund, S., Johansson, E., et al. (2007). Visualization of GC/TOF-MS-based metabolomics data for identification of biochemically interesting compounds using OPLS class models. *Analytical Chemistry, 80*(1), 115–122.

Wold, S., Esbensen, K., et al. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems, 2*(1–3), 37–52.

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry, 39*(4), 561–577.