Evaluation of a combination of SIFT-MS and multivariate data analysis for the diagnosis of *Mycobacterium bovis* in wild badgers

Andrew D. Spooner [a], **Conrad Bessant** [a], **Claire Turner** *[b], **Henri Knobloch** [a] **and Mark Chambers** [c]

[a]*Cranfield University, Cranfield, Bedfordshire, UK MK43 0AL. E-mail: c.bessant@cranfield.ac.uk*

[b]*Dept. Chemistry and Analytical Sciences, The Open University, Walton Hall, Milton Keynes, UK MK7 6AA. E-mail: c.turner@open.ac.uk*

[c]*TB Research Group, Dept. of Statutory and Exotic Bacteria, Veterinary Laboratories Agency, New Haw, Addlestone, Surrey, UK KT15 3NB. E-mail: m.a.chambers@vla.defra.gsi.gov.uk*

Received 20th March 2009 , Accepted 3rd July 2009
First published on the web 17th July 2009

The currently accepted ‘gold standard’ tuberculosis (TB) detection method for veterinary applications is that of culturing from a tissue sample *post mortem*. The test is accurate, but growing *Mycobacterium bovis* is difficult and the process can take up to 12 weeks to return a diagnosis. In this paper we evaluate a much faster screening approach based on serum headspace analysis using selected ion flow tube mass spectrometry (SIFT-MS). SIFT-MS is a rapid, quantitative gas analysis technique, with sample analysis times of as little as a few seconds. Headspace from above serum samples from wild badgers, captured as part of a randomised trial, was analysed. Multivariate classification algorithms were then employed to extract a simple TB diagnosis from the complex multivariate response provided by the SIFT-MS instrument. This is the first time that such multivariate analysis has been applied to SIFT-MS data. An accuracy of TB discrimination of approximately 88% true positive was achieved which shows promise, but the corresponding false positive rate of 38% indicates that there is more work to do before this approach could replace the culture test. Recommendations for future work that could increase the performance are therefore proposed.

1 Introduction

As well as being a major health problem in the human population,[1] tuberculosis (TB) is also a problem in agriculture, as highlighted by the media in the UK with the publication of the Krebs report (1997)[2] and the Independent Scientific Group (ISG) on Cattle TB report (2007).[3] The bovine TB situation is exacerbated in the UK and Ireland by a reservoir of *Mycobacterium bovis* infection in badgers (*Meles meles*). One of the problems identified is the need for a rapid detection method to decrease the time between test and treatment in humans and to decrease the time cattle that are quarantined waiting for diagnosis. The ISG was commissioned by the UK government to look into the impact of badger culling on the prevalence and transmission of tuberculosis in cattle. As part of this, a large number of samples were collected from culled badgers, providing the basis for evaluating the novel diagnostic method described in this paper.

Historically (and long since abandoned as a diagnostic technique), tuberculosis was known for a characteristic smell on the breath of the infected. This smell must be associated either with the production of specific 'marker' volatiles by the bacteria or an increase in the levels of particular volatiles produced by the host in response to the infection. This has led researchers to investigate the use of gas analysis techniques for rapid TB detection.

1.1 Choice of sample

TB is generally found in the lungs of the host, be it human or animal, and therefore the 'gold standard' technique for detection of the disease is a culture test from lung tissues obtained *post mortem*.[4] Sputum or tracheal aspirate samples from infected individuals can contain high levels of the bacteria, but only when the host respiratory tract is infected, and are either difficult or impossible to obtain from animals, and from children. As volatile organic compounds (VOCs) are known to be emitted in breath this would be the next biologically preferred sample, but the samples are difficult to obtain, store and transport safely. As the lungs provide a large interface between the body and the atmosphere, many VOCs found in breath can also be found in blood serum samples.[5] Such samples are relatively easy to obtain, store and distribute safely and hence are the sample of choice for this study. In addition, previous work has shown that electronic nose technology has been able to discriminate between the serum of cattle infected with *M. bovis* and uninfected control animals.[6]

1.2 SIFT-MS

Selected ion flow tube mass spectrometry (SIFT-MS) was originally developed to study ion-neutral reactions at thermal interaction energies before being adapted to perform volatile gas analysis.[7] More recently it was further developed for life science applications.[8] The technique relies on trace gases reacting with precursor ions (H_3O^+, NO^+ and O_2^+) generated in a microwave discharge through moist air. The precursor ions are selected using a quadrupole mass filter and then injected into a helium carrier gas which then passes down a flow tube. The sample of headspace gas to be analysed is introduced into the flow tube by means of a heated capillary at a known flow rate, and the trace gas components rapidly react with the precursor ions to generate product ions. These are then detected downstream. The SIFT-MS instrument may be operated in one of two modes: the first is the full scan mode where full spectra are obtained over a range of values of mass to charge ratio (m/z) – this is the mode used in this study. The alternative, which is good for quantitative analysis of individual compounds, may be carried out by rapidly switching the detector between pre-selected ions; the ratio of product ions to precursor ions enables the quantification of the analyte of interest.[8] The concentration of compounds may be readily determined through these data and knowledge of the reaction rate coefficient between precursor and product ion. However, due to the relatively small user community, public databases of these reaction coefficients of specific VOCs are only available for a small number of compounds compared to the database of more than 200 000 GC-MS signatures contained in the NIST database (www.nist.gov). Therefore, alternative data analysis approaches are required. Some of the benefits of SIFT-MS over GC-MS are rapid analysis time, lower mass range and the relative simplicity of the spectrum returned. However, complex biological samples still yield highly multivariate responses which are difficult to analyse without computational assistance.

Previous studies involving SIFT-MS analysis of human or animal samples utilised prior knowledge about the spectra and compounds present.[9] Usually, markers are known and the effects of a trial are determined by the change in a small number of selected *m*/*z* peaks. The problem investigating diseases using volatile analysis is well known and two-fold. Firstly, biological samples contain hundreds of volatiles that are naturally occurring and have a naturally high variance. Finding the two or three reliable markers can be difficult and requires much manual work. Secondly, it is not guaranteed that these markers exist. The characteristic smell of the disease may be from the combination of ten or more individual compounds, each with an associated *m*/*z* peak, each of which, as a single entity, is not suitable as a marker. Indeed, they may all be present naturally, but in different relative concentrations.

Chemometrics techniques such as principal components analysis (PCA) and Partial Least Squares Discriminant Analysis (PLSDA[10]) have been applied in similar situations to data from other sources such as electronic nose[11] and GC-MS.[12] These are the approaches applied here.

2 Methodology

2.1 Sample collection and measurement

The samples were all collected from the Defra funded Randomised Badger Culling Trial managed by the ISG on Cattle TB.[3] Wild badgers from different regions across the UK, where TB has been repeatedly found in cattle, were cage trapped. The serum samples for this study were obtained from a subset of these badgers during trapping operations ending in 2005 that were first anaesthetised, then blood samples taken before the animals were killed by lethal injection. Samples from an extensive range of tissues were removed *post mortem* and plated for diagnosis of *M. bovis* infection status. These blood samples have previously been used to evaluate the potential for gamma interferon as a diagnostic tool for TB in badgers.[13] Individual blood samples were processed to produce sera and frozen for analysis at a later date.

The 245 sera samples from the above trial were shipped to Cranfield University where they were thawed. 700 µL of sample was pipetted into a sample bag made from Nalophan NA tubing (Kalle, UK) with diameter 65 mm and length 31 cm. 700 mL of zero grade (hydrocarbon-free) air (BOC) was added and the sample bags incubated at 40 °C. Two-thirds (467 mL) of the headspace was used for other studies and the remaining third (233 mL) analysed using SIFT-MS. Because the serum samples are aqueous, and incubated at 40 °C, the water vapour pressure was inherently high (about 7% water).

The SIFT-MS analysis was performed using 10 s scans per precursor which was repeated ten times across the three precursor ions (total analysis time of 5 min). Full scan mode was employed, over a mass to charge ratio of 10–109 (the limit of the SIFT-MS instrument used in this study at that time). The SIFT-MS instrument used was a Mk2 model from PDZ Europa (UK). The sample was incubated at 40 °C for 10 min prior to and during connection to the SIFT-MS inlet capillary.

Once all the spectra were recorded and saved, the raw data were converted into a standard spreadsheet format (CSV) and then analysed in Matlab R2008b (Mathworks, USA), utilising PLS_Toolbox 3.5 (Eigenvector Research).

2.2 Data preparation and pre-processing

As the precursors are carrier ions added to the sample, they need to be removed from the dataset as they are clearly unrelated to the infection status of the sample. This is performed by zeroing the count for that precursor's ions across all samples. The ions removed are: *m*/*z* of 19, 20, 21, 30, 32, 37, 38, 39, 55, 56, 57, 73, 74, 75, 91. These ions comprise the 19 precursor plus isotopologues at *m*/*z* 20 and 21, plus water clusters (due to the aqueous nature of biological samples) and their isotopologues at 37, 38 and 39; 55, 56, 57; 74, 75 and 76 and 91. Very small amounts of 30 and 32 (NO^+ and O_2^+) are also present in the H_3O^+ spectrum, so are also excluded. Based on similar reasoning, 19, 30, 32, 37, 48, 50, 55, 57, 66, 73, 91, are removed from the NO^+ precursor channel and 19, 30, 32, 33, 34, 37, 50, 55, 56, 57, 73, from the O_2^+ channel. After removal of these precursors, univariate analysis is possible by investigating the spectra by hand. Fig. 1 shows three spectra (one for each precursor) from the analysis of a serum sample from a wild badger classified as infected by the culture test and three spectra from the analysis of serum from an animal classified as uninfected by the culture test.

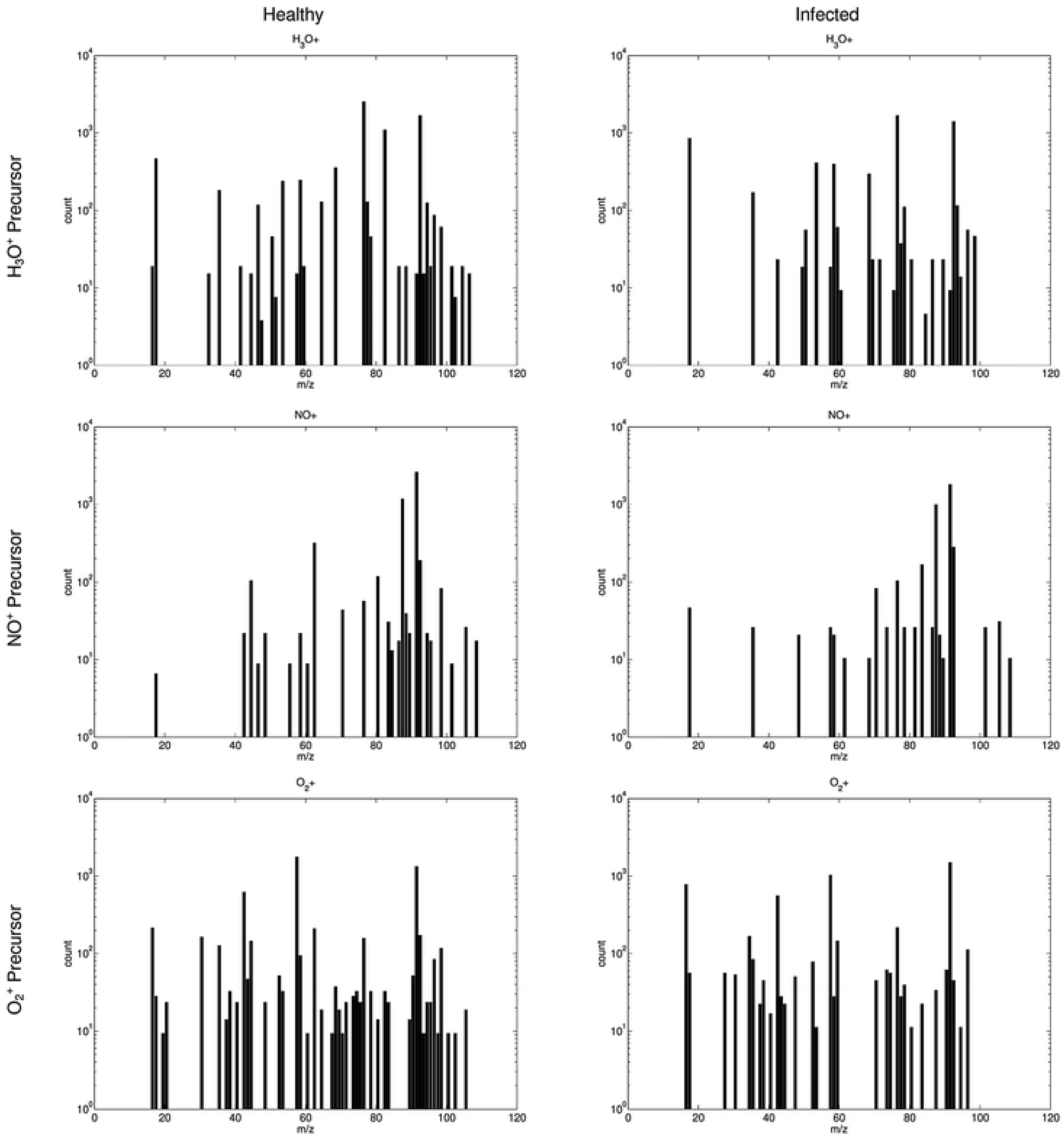

Fig. 1 Spectra using all three precursor ions (H_3O^+, NO^+ and O_2^+) for the headspace above two serum samples (one from a badger later classified as tuberculosis negative, and one from a badger classified as tuberculosis positive using the culture test). The *m*/*z* values linked to the precursors have been zeroed (see text for explanation).

After removal of the peaks caused by the precursors, the three spectra from each sample are concatenated into a single vector of length 300 and all the samples combined into a matrix, ***X***. Hence, vector positions 1–100 correspond to the H_3O^+ spectra from *m*/*z* 10 to *m*/*z* 109, vector positions 101–200 correspond to the NO^+ spectra from *m*/*z* 10 to *m*/*z* 109 and the last 100 vector positions correspond to the O_2^+ spectra from *m*/*z* 10 to *m*/*z* 109. This dataset is then used for all subsequent multivariate analysis.

The dataset contained 194 samples that were culture negative and 51 samples that were culture positive (245 samples in total). It should be noted that although culture is a highly respected diagnostic tool, it is not 100% accurate[13] and a small number of the negative tuberculosis classified badgers may have had an undiagnosed low level infection.

Heatmaps (plotting the intensities of each element of the matrix as a colour) were then used to verify that the data had been correctly imported and combined. The effect of data scaling was also investigated, with the methods considered being mean centring (eqn (1)), auto scaling (eqn (2)), range scaling (eqn (3)) and taking logs (eqn (4)).

$$x_{ij} = x_{ij} - \bar{x}_j \tag{1}$$

$$x_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j} \tag{2}$$

$$x_{ij} = \frac{\left(x_{ij} - \min\left(x_j\right)\right)}{\left(\max\left(x_j\right) - \min\left(x_j\right)\right)} \tag{3}$$

$$x_{ij} = \log_{10}(x_{ij} + 1) \tag{4}$$

2.3 Data analysis

Having pre-processed and checked the data, the first analysis method employed was PCA. This is a common data exploration method[14] which extracts and displays the characteristics that caused the greatest variance in the dataset. It breaks the data into eigenvectors and eigenvalues. The two matrices are then re-arranged and become scores – which indicate the relationships between the samples – and loadings, which reveal how much each measured variable contributes to the scores. The components required are selected and the remaining information ends up as a discarded error matrix. The scores matrix contains dimensions of decreasing variance content that can be plotted against each other to produce the figures.

Next, PLSDA[10] was used to build classification models using the tuberculosis status of the samples. PLSDA is a supervised method, so information about the parameter of interest (in this case TB infection status) is required to train the algorithm to determine which parts of the acquired spectra capture the maximum difference between the class states. PLSDA first separates the data into two groups, one with negatives (or class 1) and the other with positives (class 2). It then continues in a similar fashion to PCA, in that eigenvectors and eigenvalues are obtained. However, with PLSDA the difference between the variance of the two groups (the co-variance), is maximised. Theoretically, the information returned should contain any information that could be caused by the property under investigation.

It is known that PLSDA models can over-estimate the accuracy of classification if not properly validated.[15] Indeed, given sufficient data, any training algorithm should be able to correctly classify all the samples used to build the model. To cope with this problem, the model was optimised using a Leave One Out (LOO) method. In this method, a sample is classified against a model built using the rest of the samples. This process is then repeated with each sample until all the samples have been classified. Information about the number of true positive and true negative identifications was also extracted.

An important question in studies of this type is how many samples are required to build a classification model of suitable diagnostic power.[16] This is important to minimise both the number of samples collected and the number of analyses required. To answer this question, a random selection of 50 samples was taken and leave one out cross-validation performed. The number of correctly classified samples was extracted and stored. The sample size was then increased by one (and randomly re-sampled) and the process repeated until all but one sample had been included. The whole process was then bootstrapped (repeated ensuring that different sets of random samples were used) 50 times to achieve a statistical distribution at each point.

3 Results and discussion

The raw spectra for any single positive sample and any single negative sample show a large number of similarities and a large number of differences. Comparing two positive or two negative spectra also shows a number of similarities and differences. The problem is therefore to track down reliable diagnostic differences between the groups of spectra. Clearly, separating any differences between the 194 negative spectra and the 51 positive spectra without using multivariate techniques would be both time consuming and subject to potential error.

The PCA scores plot derived from the experimental data after the data have been pretreated using mean centring (eqn (1)) is shown in Fig. 2. Although the figure does not show any discrimination between infected and uninfected, there does appear to be some structure. Analysis of the PCA loadings (the weight given to each *m*/*z* peak) shows that the structure seen is caused by acetone (PC1) and ethanol (PC2). The main ions implicated are 77 for acetone using H_3O^+, *m*/*z* 77 is the ion resulting from acetone·H^+·H_2O, *i.e.* one water cluster, which arises because of the high water vapour pressure in the samples. The water cluster of the isomer, propanal, will also produce ions at the same *m*/*z*; however, other work with GC-MS (data not shown) indicates that it is likely to be largely, if not exclusively, due to acetone. A peak at *m*/*z* 88 was also found with NO^+, indicating acetone. Similarly, *m*/*z* 83 is the main ion generated for ethanol, which corresponds to ethanol·H^+·$2H_2O$, also using H_2O^+. These two compounds are associated with diet – ethanol is a natural product of decay in anaerobic conditions and acetone is a naturally occurring systemic compound that varies widely in concentration, depending on diet, blood sugar and individual metabolism.[9] This implies that the natural difference between the animals is larger than the difference caused by TB. This doesn't mean that there is not information in the data pertaining to TB, just that it does not cause the majority (65%) of the variance captured by the first two principal components. However, investigating PCA components of lower variance did not reveal clear disease-related discrimination either.

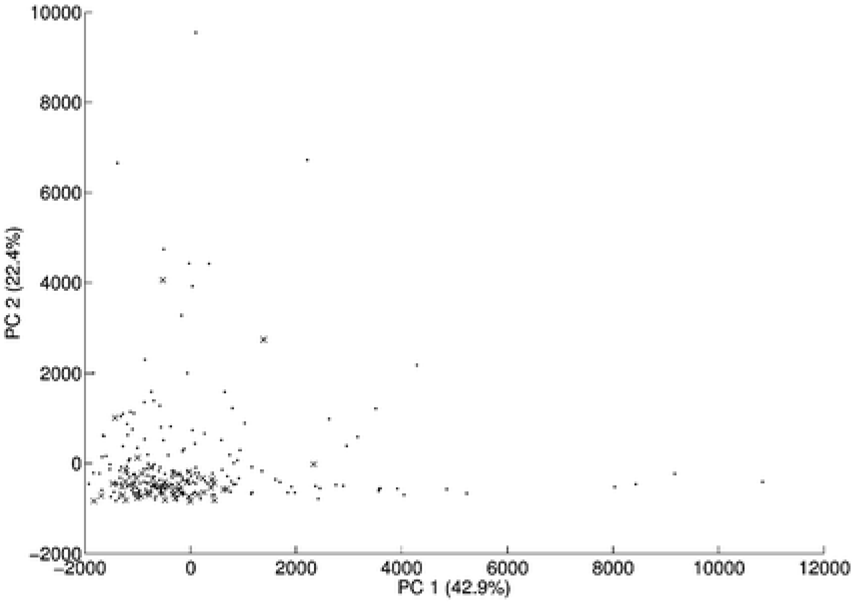

Fig. 2 The PCA scores plot does not appear to show any discrimination between TB (×) and control samples (•). The data were pretreated using mean centring. Investigating the loadings shows that the information on the plot relates to the concentrations of *m*/*z* 77 and *m*/*z* 83, which are likely to be caused by acetone (PC1 positive direction) and ethanol (PC2 positive direction) respectively. This information is unlikely to be related to TB but to another influence such as diet.

PLSDA has the potential to distinguish between samples that could not be separated using PCA because it maximises the co-variance between the acquired data and the sample classifications (control *versus* TB). This gives the possibility of distinguishing between samples. In Fig. 3, an estimate of this ability has been produced (··· line). More important is the result produced using the leave one out algorithm (— line) in which models are built using all but one sample and then the sample is classified. The number of correctly classified positives (−−− line) and correctly classified negatives (−•− line) is also shown as a percentage. The leave one out validation on the complete dataset shows that an overall accuracy of 67% can be achieved with 88% correctly classified positives and 62% correctly classified negatives using the currently available dataset. Analysing the loadings of the PLSDA analysis is more complicated than analysing the loadings of PCA. The discrimination seen is the result of seven latent variables and these all need to be analysed to find which *m*/*z* are the most likely to be responsible for the discrimination. Values of 18, 36, 54 (ammonia), 51 (methanol), 77, 95 (acetone), 83 (ethanol), 93 (toluene from the anaesthetic) on H_3O^+; 18, 45 (ethanol), 88 (acetone), 92 (toluene) 63 (unknown) on NO^+; 17, 43, 45, 58 (acetone), 75, 77, 92 (toluene), 109 (unknown) on O_2^+ are the *m*/*z* that have a high impact on the discrimination. These compounds are responsible for a high proportion of the co-variance used during classification, yet the list is in no way complete. One implication is that the differences between the TB positive and TB negative samples may be subtle and not due to just one or two biomarkers.

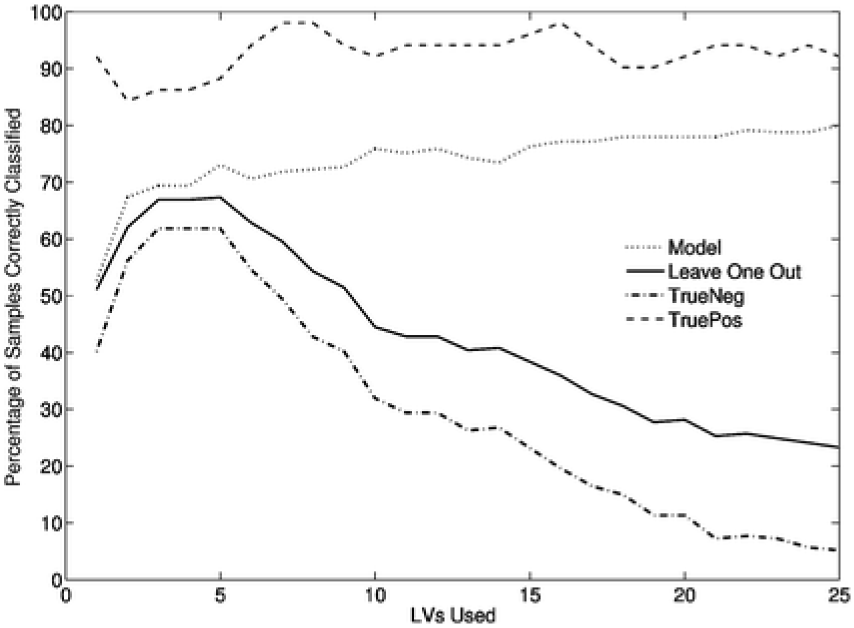

Fig. 3 PLSDA model (···) and leave one out optimisation curves, correctly classified (—), true positives (−−−) and true negatives (−•−) for the first 25 latent variables (LVs) included. The accuracy of the PLSDA models (···) shows the number of samples correctly classified as a percentage. It is well known that this over-estimates the accuracy of the model and therefore the leave one out results are more informative. The total percentage of samples correctly classified by the leave one out algorithm is then broken down into the percentage of true positives and true negatives. The optimal number of latent variables is five and this returns an accuracy (as estimated by LOO) of 67%, with 88% true positive and 62% true negative.

Analysing the PLSDA results for the three precursors separately was also performed with less successful results (results not shown).

The relationship between the accuracy of the PLSDA LOO performance and the size of the training dataset is shown in Fig. 4. It can be see that the maximum accuracy of the PLSDA has yet to be reached, although the average improvement in accuracy reduces as the number of samples increases, suggesting that provision of additional samples might not substantially improve diagnostic performance. It can also be seen that in this specific example a minimum number of samples required to produce a repeatable model is 150. Below this, the accuracy of the model (65%) is stable but the errors associated are very large and increase with fewer samples. It appears that using more than 250 samples is unlikely to substantially increase the accuracy of diagnosis.

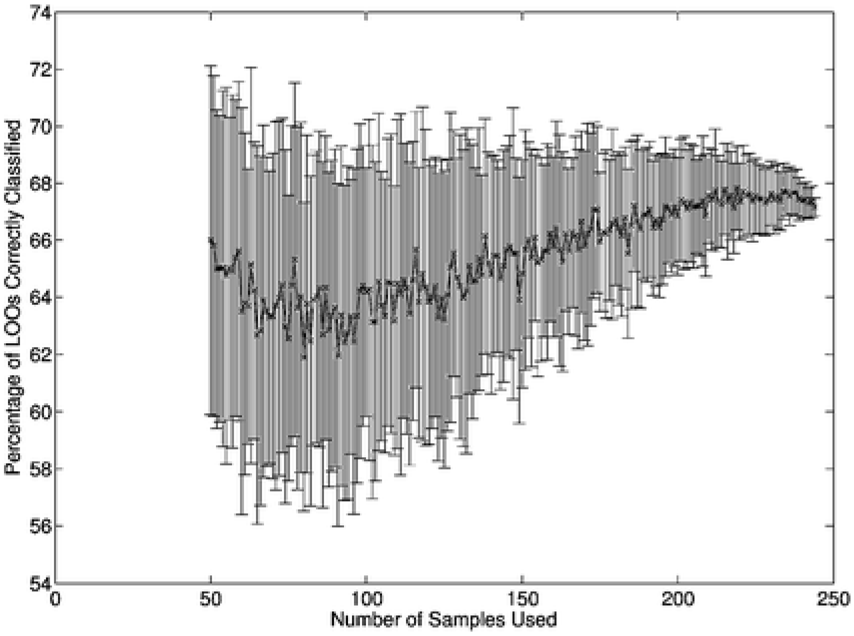

Fig. 4 The mean and standard deviation of PLSDA LOO corrected classified results over a range of dataset sample sizes using 64 bootstraps. Error bars indicate one standard deviation, capturing 66% of the models produced.

Finally, models were built using randomly selected samples and then tested using the samples remaining, to ensure complete independence in the testing of the model. Models were created using 80% (196 samples), 90% (220) and 95% (232) of the total number of samples. The accuracy was determined to be 72% (± 25%), 64% (± 31%) and 76% (± 21%) respectively, with the error being one standard deviation as calculated from 200 bootstraps.

4 Conclusions

It has been shown that by linking SIFT-MS analysis with chemometrics techniques we have made some progress towards a new method of disease detection provided that the dataset sizes are large enough to enable training of statistical models. PLSDA accuracy estimated using a leave one out (LOO) algorithm was 67%, with 88% true positives and 62% true negatives. A dataset size of 150–250 samples was required for the samples under investigation.

This work has demonstrated for the first time a way of analysing large SIFT-MS datasets without trying to analyse individual compounds in individual spectra. Although the approach proposed is much faster than a traditional culture test, the accuracy achieved makes it unsuitable as a replacement for the culture test. The gamma interferon test is the most accurate test for TB in the live badger and achieved a 93.6% true negative and 80.9% true positive rate using the same subset of animals.[13] It is encouraging that the headspace analysis method detected even more badgers with culture-confirmed TB but at the expense of more apparent false positive results. However, the sensitivity of the gamma interferon test could also be increased by changing the test positive cut-off point if a higher false positive rate was accepted. It is believed that the overall accuracy achieved with the SIFT-MS method was limited due to the sample set being subject to large quantities of unknown variables. Factors such as sex, age, diet, location, general health of animal and anaesthetic injection would have all had an impact on the volatiles within the samples, although some of the samples classified as false positive may actually represent badgers with TB undiagnosed by culture. The anaesthetic (a cocktail of ketamine and medetomidine), in particular, appeared to lead to very high levels of toluene in the blood (seen as peaks at *m*/*z* 93 in H_3O^+ spectrum and 92 in NO^+ and O_2^+ spectra).

In terms of further work we recommend looking at a better controlled group of animals as this would introduce the possibility of looking for a group of markers specific to TB. Once these have been identified, it should be easier to remove from the data variance caused by factors other than tuberculosis. In addition, new developments in SIFT-MS instruments mean that better sensitivity and mass ranges are now available. Such an instrument would be able to analyse more compounds and thus would most likely enable better discrimination.

Acknowledgements

This paper forms part of project SE3221 which was funded by the UK's Department for Environment, Food and Rural Affairs and the authors gratefully acknowledge DEFRA's funding.

References

1 P. Onyebujoh and G. A. W. Rook, *Nature Reviews Microbiology*, 2004, **2**, 930–932 [Links].

2 J. R. Krebs, *Bovine Tuberculosis in Cattle and Badgers*, Ministry of Agriculture, Fisheries and Food, 1997.

3 F. J. Bourne, *Bovine TB: The Scientific Evidence*, Independent Scientific Group on Cattle TB, Defra, 2007.

4 R. de la Rua-Domenech, A. T. Goodchild, H. M. Vordermeier, R. G. Hewinson, K. H. Christiansen and R. S. Clifton-Hadley, *Research in Veterinary Science*, 2006, **81**, 190–210.

5 M. Phillips, *Analytical Biochemistry*, 1997, **247**, 272–278 [Links].

6 R. Fend, R. Geddes, S. Lesellier, H. M. Vordermeier, L. A. L. Corner, E. Gormley, E. Costello, R. G. Hewinson, D. J. Marlin, A. C. Woodman and M. A.

Chambers, *Journal of Clinical Microbiology*, 2005, **43**, 1745–1751 [Links].

7 D. Smith and N. G. Adams, *Advances in Atomic and Molecular Physics*, 1987, **24**, 49.

8 D. Smith and P. Spanel, *Mass Spectrometry Reviews*, 2005, **24**, 661–700 [Links].

9 C. Turner, P. Spanel and D. Smith, *Physiological Measurement*, 2006, **27**, 321–337 [Links].

10 M. Barker and W. Rayens, *Journal of Chemometrics*, 2003, **17**, 166–173.

11 S. M. Scott, D. James and Z. Ali, *Microchimica Acta*, 2006, **156**, 183–207 [Links].

12 P. Jonsson, H. Stenlund, T. Moritz, J. Trygg, M. Sjostrom, E. R. Verheij, J. Lindberg, I. Schuppe-Koistinen and H. Antti, *Metabolomics*, 2006, **2**, 135–143 [Links].

13 D. Dalley, D. Dave, S. Lesellier, S. Palmer, T. Crawshaw, R. G. Hewinson and M. Chambers, *Tuberculosis*, 2008, **88**, 235–243 [Links].

14 M. Otto, *Chemometrics: statistics and computer application in analytical chemistry*, WILEY-VCH, Weinheim, 1999.

15 J. A. Westerhuis, H. C. J. Hoefsloot, S. Smit, D. J. Vis, A. K. Smilde, E. J. J. Velzen, J. P. M. Duijnhoven and F. A. Dorsten, *Metabolomics*, 2008, **4**, 81–89 [Links].

16 R. G. Brereton, *Trends in Analytical Chemistry*, 2006, **25**, 1103–1111 [Links].