Evaluation of a gas sensor array and pattern recognition for the identification of bladder cancer from urine headspace

How to cite:

For guidance on citations see FAQs.

© 2011 The Royal Society of Chemistry

Version: Accepted Manuscript

Link(s) to article on publisher’s website:
http://dx.doi.org/doi:10.1039/C0AN00382D

Copyright and Moral Rights for the articles on this site are retained by the individual authors and/or other copyright owners. For more information on Open Research Online's data policy on reuse of materials please consult the policies page.
Evaluation of a Gas Sensor Array and Pattern Recognition for the Identification of Bladder Cancer from Urine Headspace

Christina M. Weber,*Michael Cauchi,*Mitesh Patel,*Conrad Bessant,*Claire Turner,†L. E. Britton†and Carolyn M. Willis*†

Received (in XXX, XXX) XthXXXXXXX 200X, Accepted XthXXXXXXX 200X
First published on the web Xth XXXXXXXXX 200X
DOI: 10.1039/b000000x

Previous studies have indicated that volatile compounds specific to bladder cancer may exist in urine headspace, raising the possibility that headspace analysis could be used for diagnosis of this particular cancer. In this paper, we evaluate the use of a commercially available gas sensor array coupled with a specifically designed pattern recognition algorithm for this purpose. The best diagnostic performance that we were able to obtain with independent test data provided by healthy volunteers and bladder cancer patients was 70% overall accuracy (70% sensitivity and 70% specificity). When the data of patients suffering from other non-cancerous urological diseases were added to those of the healthy controls, the classification accuracy fell to 65% with 60% sensitivity and 67% specificity. While this is not sufficient for a diagnostic test, it is significantly better than random chance, leading us to conclude that there is useful information in the urine headspace but that a more informative analytical technique, such as mass spectrometry, is required if this is to be exploited fully.

Introduction

Transitional cell carcinoma (TCC) of the bladder is the most common form of bladder cancer and simultaneously the second most common malignant tumour of the genito-urinary tract.1 As with many cancers, untreated TCC can be fatal, so early detection is vitally important. Cystoscopy with biopsy is the “gold standard” for bladder cancer detection, but it is expensive, inconvenient and invasive. Urine cytology is a more acceptable alternative, and the most widely applied of the non-invasive procedures available. Although urine cytology has a high specificity for bladder cancer (90-98%), its sensitivity is low (20-50%), especially for low-grade tumours, since those shed proportionally few cells into the urine. Furthermore, the results take several days, requiring an expert to interpret the test.

Utilisation of molecular biomarkers for bladder cancer is a tantalising prospect for making diagnosis more sensitive, rapid and convenient, provided such biomarkers can be discovered in urine and readily identified in a clinical setting. Two FDA approved protein markers, nuclear matrix protein 22 (NMP22) and bladder tumour antigen (BTAstat) have proven effective in this context.2,3 They are more sensitive than urine cytology, having achieved reported sensitivities of 50-85% and 50-70%, respectively. However, their specificities, in the region of 60-70%, are inferior to that of urine cytology.

Recently, it has been suggested that volatile organic compounds (VOCs) present in the headspace of urine samples may be used as biomarkers. In particular, a previous study4 showed that dogs could be trained to identify bladder cancer sufferers from the odour of their urine. The results of this proof of principle study suggested that one or more VOCs present within the urine headspace are indicative of bladder cancer and have the potential to be used as diagnostic biomarkers. As it is not realistic to use dogs in a clinical setting, we substituted the dogs with a more practical and objective gas sensor array, also called an electronic nose (eNose). Developed to mimic olfaction, these instruments have shown promise in lung cancer detection.5

Methods

Participant selection

A total of 30 patients, aged 50-88, presenting at Buckinghamshire Hospitals NHS Trust with new or recurrent transitional cell carcinoma (TCC) of the bladder supplied urine prior to surgical intervention. Grade and stage of the tumour were recorded. Fifty nine control subjects, categorised into one of three groups (Controls 1, 2 and 3, depending upon age and disease status), also provided urine samples. Twenty healthy individuals aged 18-31, with no urine abnormality on dipstick made up Control group 1 (C1). Control group 2 (C2) consisted of 20 subjects between 18 and 32, with any non-cancerous condition or disease, and/or one or more positive dipstick finding. Menstruating women with blood in their urine were included in this group, for example, as were individuals with suspected urinary tract infection, positive for leucocytes, blood and/or protein. Categorized into the Control 3 group (C3) were 19 patients, aged 24 and 89, with confirmed non-cancerous urological disease, with or without urine dipstick abnormalities. Urological conditions included renal and ureteric stones, renal cysts and polypoid cystitis. A summary of the age and gender of the subjects in each cohort is included as supplementary material.

As criteria for inclusion/exclusion, controls over 32 years of age were required to have had recent cystoscopy to exclude visible bladder malignancy. For both controls and the cancer positive group (TCC), men over 50 years were only included if recent cancer-negative prostate histology had been demonstrated. Individuals with pre-malignant urological disease or a history of urological carcinoma other than TCC were excluded. A history of malignancy in other organ...
systems (> 5 years previously) was acceptable, providing the individual was now considered disease-free. All other past and/or present medical conditions were permissible. There were no exclusions on the basis of medication, menstrual cycle, diet, alcohol consumption, or chemical exposure. However, details of all of these factors were recorded for each participant, should their influence on the composition and odour of the urine need to be considered at any stage. Special attention was paid to smoking habits, with one third of those with bladder cancer being current cigarette smokers, as compared to 17 out of the 59 control subjects. A spreadsheet containing all this information is provided as supplementary material.

The study was approved by the Mid and South Buckinghamshire Local Research Ethics Committee, and all participants gave written informed consent.

Analysis and processing of urine samples

Following urinalysis (Multistix 10 SG, Bayer Corporation, NY, USA), fresh urine specimens were refrigerated immediately, and frozen 1-24 hrs later as 0.5ml aliquots in glass vials. They were then stored at -80°C until required.

Headspace Analysis

An electronic nose was used to characterise the VOC (volatile organic compound)content of urine. Measurements were performed using an NST 3320 Lab Emission Analyser. The instrument comprises 12 metal oxide semi-conductor (MOS) sensors and an array of 10 individual metal oxide semi-conductor field-effect transistor (MOSFET) sensors, together with a capacitance-based humidity sensor and an infrared-based CO₂ sensor.

The 89 urine samples were randomised, removed from storage at -80°C and allowed to thaw at room temperature (approximately 21°C). The thawed samples were then aliquoted into two headspace vials, each containing 2.5 ml of urine, providing duplicate samples. The samples were then incubated for 1 hour at 38°C before they were analysed in batches of 12 samples using the eNose. Signals from each of 24 sensors are generated by their response to the different chemical characteristics of the urine headspace. The analysis was performed in cycles, exposing the sensors to fresh reference air between the measurements. A cycle was composed of Baseline, Sample and Recovery. During runtime, all sensor signals were collected in periodic time intervals of one second. From this resulting raw data, eight characteristic signal parameters were estimated by the eNose software as baseline, response, absolute response, on derivative, on integral, off derivative, off integral and difference. This can be seen as data reduction, very likely resulting in a speed up of subsequent pattern recognition analysis. Figure 1 shows the sensor parameters graphically. All samples were measured twice (89 samples and 89 duplicates) and the resulting extracted signal parameters - features - were finally exported in Microsoft Excel spreadsheet format.

Data Analysis

The provided spreadsheet files were analysed using Matlab R2009a (The Mathworks Inc.). Additionally the PLS Toolbox 3.5 (Eigenvector Research Inc.) was employed to classify the data via partial least squares-discriminant analysis (PLS-DA). Before the loaded data was analysed, the effect of data scaling was investigated, with the methods considered being auto-scaling (Equation 1), mean-centring (Equation 2), range-scaling between 0 and 1 (Equation 3) and range-scaling between -1 and +1 (Equation 4).

\[
y_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j} \quad (1)
\]

\[
y_{ij} = x_{ij} - \bar{x}_j \quad (2)
\]

\[
y_{ij} = \frac{\left( x_{ij} - \min(x_j) \right)}{\left( \max(x_j) - \min(x_j) \right)} \quad (3)
\]

\[
y_{ij} = \frac{2 \times \left( x_{ij} - \min(x_j) \right)}{\left( \max(x_j) - \min(x_j) \right) - 1} \quad (4)
\]

Exploratory Data Analysis was accomplished via principal component analysis (PCA), which is the most widely used multivariate statistical technique. This part of the analysis was performed to reveal the characteristics that cause the greatest variance in the dataset.

Before the pre-processed data was sent to the classifier, a crucial step - feature selection - was performed to provide the classifier with the most significant variables, so as to achieve higher classification accuracy. This procedure resembles a filter method that automatically picks out variables within the data that are most distinctive of each class, or from another point of view deselects those variables that are not significant at all. For this purpose, univariate statistics - a t-test - was applied. The t-test is more or less a decision rule to determine if two samples belong to the same population, according to a specified significance level alpha α. Dependent on the significance level α, two variables, in this case, non-cancerous and cancerous samples, were either considered as similar or different. For data analysis, the significance level α was
initially set to 0.05 and varied from 0.1 to 0.9 in steps of 0.1. Features whose probability value did not rise above the threshold were taken as being different and were therefore regarded as relevant for disease-related discrimination. From a statistical point of view high $\alpha$-values are not suitable to select significant features, but can be used to deselect insignificant variables as noise for example.

Next, partial least squares discriminant analysis (PLS-DA) was used to build a classification model using the cancer status of the samples. PLS-DA is a supervised method. This calls for information about the parameter of interest (the cancer status) to be known in order to train the algorithm to identify which sensors capture the molecules that differentiate between the classes. PLS-DA is considered to be a dimensionality reduction method and can be seen as the regression extension of principal component analysis. Unlike PCA, which attempts to describe the maximum variation in the measured data, PLS-DA tends to maximise the covariance between the input data and the output class. The information returned by PCA is that which was caused by the attribute with the biggest variance. In contrast, PLS-DA returns only data that was caused by the property under investigation.

It is known that PLS-DA is prone to overestimate the accuracy of classification if it is not accurately validated. For this reason, a very thorough evaluation process - leave-one-out cross-validation (LOOCV) - was implemented to assess the performance of the PLS-DA-classifier. This method is based on the concept of classifying one sample against a model, built using the rest of the samples. This procedure is repeated with each sample until all samples have been classified. Information about the true positive rate and the true negative rate was established.

As final validation of the results, and to attain an indication of the significance of the results, a Monte Carlo Simulation was used to evaluate the obtained results. This involves repeated random sampling. In this context a null model is generated from a set of data that is statistically similar to the data under study, but for which we do not expect to be able to build a meaningful classification model. A total of 250 datasets of this kind were generated, by simply assigning a random class to each sample (while maintaining the relative number of positive and negative samples). For a disease discriminating model trained on the real sample classes to be considered significant it needs to achieve a classification accuracy towards the extremities of those produced by the null models.

As already mentioned in the Methods section, the 59 specimens belonging to the cancer-negative control group were sub-divided into further subgroups (C1-C3). For this reason, we sectioned our work into four main experiments. Firstly, each of the three control subgroups was classified separately against the cancer group and, finally, all control subgroups were combined and classified together against the TCCs.

In each of the four experiments, we used the full sensor array as well as the MOS sensors only and the MOSFET sensors only. Furthermore we applied the same pre-processing methods to the data, including scaling and feature selection. Four different scaling methods and no scaling were evaluated, with range-scaling between 0 and 1 being found to be the best; therefore, only these results are shown in this paper. Given that we were looking for a pattern within the data and were therefore interested in relative quantities, not absolute quantities, it is not surprising that range-scaling was most appropriate.

Explorative data analysis, by way of PCA, was also applied in each experiment to reveal natural groupings based on the sensor responses of the electronic nose. The PCA scores plot derived from the experimental data is shown in Figure 2. However, the visual output of this method did not disclose any relation to the cancer status of the samples, even though the principal components shown capture over 98% of the variance in the data. Other influences such as age, diet or gender may be responsible for the groupings obtained. However, this does not mean that the data does not contain any information concerning bladder cancer. The PCA only demonstrated that the cancer status is not responsible for the bigger part of the variance, captured by the first two or three principal components (PCs). Nevertheless, investigating principal components of lower variance did not lead to an explicitly disease-related differentiation, either. PLS-DA aims to maximise the co-variation between the measured data and the classification (non-cancerous vs. cancerous). This leads to the capability to discriminate between samples that could not be separated by PCA.
We started by training the PLS-DA-classifier with the two most disparate groups: Control 1 (C1), representing healthy males or females, and the TCC group incorporating people suffering from bladder cancer. Since group C1 possesses the most differences compared to the cancer group, we expected the classification outcome of this sample set to be the best, which proved to be correct. The best result obtained a total accuracy of 70% by range-scaling the data and configuring the significance level $\alpha = 0.8$ for feature selection. A summary of the best classification results of C1 vs. TCC is given in Table 1a).

Next, we trained the classifier with Control 2 (C2) and the cancer group (TCC) data. Urine samples within this control subgroup showed similar abnormalities on dipstick analysis to some cancer samples, such as blood, for example, and were therefore more difficult to distinguish from cancerous samples than Control 1 (C1) samples. Both specificity and the total accuracy, achieved with the full sensor array, reflect this. Only the sensitivity is marginally higher than that obtained for C1 vs. TCC in full sensor array mode. An overview of the results obtained is given in Table 1b).

In the third experiment, the classifier had to distinguish between samples with confirmed non-cancerous urological diseases (Control 3) and cancerous samples (TCC). This was expected to be the most difficult combination, as disease markers not specific to bladder cancer are likely to be present.

Table 1 Best results of leave-one-out cross-validation runs with PLS-DA for each of the four experiments. The presented classification results were obtained after scaling the data between zero and one and performing feature selection.

<table>
<thead>
<tr>
<th></th>
<th>Classification method</th>
<th>Scaling method</th>
<th>Significance level $\alpha$</th>
<th>No. Features</th>
<th>Total Accuracy</th>
<th>Cancer group TCC Accuracy (Sensitivity)</th>
<th>Control group C1 Accuracy (Specificity)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOS sensors a) C1 vs. TCC</td>
<td>Full Sensor Array</td>
<td>PLS-DA</td>
<td>Range-scaling [0 1]</td>
<td>0.8</td>
<td>140</td>
<td>70.0%</td>
<td>70.0%</td>
</tr>
<tr>
<td></td>
<td>MOSFET sensors</td>
<td>PLS-DA</td>
<td>Range-scaling [0 1]</td>
<td>0.8</td>
<td>15</td>
<td>60.0%</td>
<td>57.5%</td>
</tr>
<tr>
<td></td>
<td>b) C2 vs. TCC</td>
<td>Full Sensor Array</td>
<td>PLS-DA</td>
<td>0.1</td>
<td>69</td>
<td>67.0%</td>
<td>69.0%</td>
</tr>
<tr>
<td></td>
<td>MOSFET sensors</td>
<td>PLS-DA</td>
<td>Range-scaling [0 1]</td>
<td>0.1</td>
<td>3</td>
<td>65.0%</td>
<td>65.0%</td>
</tr>
<tr>
<td></td>
<td>c) C3 vs. TCC</td>
<td>Full Sensor Array</td>
<td>PLS-DA</td>
<td>0.8</td>
<td>132</td>
<td>62.2%</td>
<td>52.6%</td>
</tr>
<tr>
<td></td>
<td>MOSFET sensors</td>
<td>PLS-DA</td>
<td>Range-scaling [0 1]</td>
<td>0.8</td>
<td>29</td>
<td>68.3%</td>
<td>50.0%</td>
</tr>
</tbody>
</table>
As far as the full sensor array is concerned sensitivity and specificity as well as total accuracy did not reach the results obtained in the two former experiments. A detailed overview of these results is presented in Table 1c).

Finally, in the fourth experiment, all three control sub-groups (C1, C2, and C3) were merged to form one control group that was used as a counterpart to the TCC samples when training the binary classifier. The results of this experiment can be seen in Table 1d). As expected, the total accuracy, obtained with the entire sensor array, was not as high as in experiments 1 and 2, where only samples from young people with no urological disease (C1, C2) were included. The major contributor to this more moderate classification outcome was the inclusion of the samples from control group 3. All subjects within this subgroup had confirmed non-cancerous urological disease, the pathologicaleffects of which are likely to be similar to the secondary effects of bladder cancer. Within both these groups, varying amounts of metabolic products associated with inflammation, infection and/or necrosis will almost certainly be present.

Because of this, Control 3 samples form the most important control subset and contain the most relevant information.

Training the classifier with this kind of data is therefore fundamental in order to be able to subtract general disease compounds present in the urine from those specific for bladder cancer. Accurate diagnosis of the control subjects is, of course, paramount to this process, since the inclusion of false negative individuals would lead to incorrect classification rules.

Interestingly, within the TCC sample group, the majority of those incorrectly classified as negative were from patients with more advanced tumours. In these cases, it is possible that metabolic products generated secondarily to the tumour may overwhelm or mask the volatile cancer biomarkers within the urine, giving rise to a urine headspace more closely resembling that of Control 3 samples. Canine olfactory studies support this hypothesis; high grade TCCs with a significant level of invasion are missed more frequently by trained dogs than low-grade superficial tumours.

To assess the significance of the presented results we carried out Monte Carlo Simulation. Figure 3 shows the results attained for each of the four experiments each with 250 random runs. A model with an overall accuracy beyond two standard deviations of the mean indicates that it is a significant result at the 95% confidence level. Thus, the first experiment that included totally healthy volunteers and cancer patients and produced a total accuracy of 70% is very significant (Figure 3(a)). The result of the second experiment, comprising samples showing abnormalities at the dipstick analysis, can also be regarded as significant, with an overall accuracy of 67% (Figure 3(b)). The third experiment included samples with confirmed non-cancerous urological diseases and was therefore the most difficult to build classifiers for. Not only did this experiment derive the lowest total accuracy (62%), the accuracy was found to be just below the 62.5% that marks the 95% confidence limit in the null model results (Figure 3(c)). The fourth experiment made use of all available data and attained a statistically significant classification result of 65%, which comfortably exceeds the 95% confidence limit (Figure 3(d)).
Although the specificity achieved using this method was less than that of conventional urine cytology, sensitivity was appreciably higher. This leads us to conclude that there is an anomalous composition to the VOC content of urine of patients suffering from bladder cancer, confirming previous canine olfactory data. However, the gas sensor array may not be the ideal analytical analysis to detect this abnormal composition of urine headspace. We therefore recommend a more revealing analytical technique such as mass spectrometry. More sophisticated pattern recognition techniques as support vector machines (SVMs) or artificial neural networks (ANNs) may also further advance the results obtained so far.

Notes and references

1. Cranfield University, College Road, Cranfield, Bedfordshire, MK43 0AL, UK; E-mail: c.bessant@cranfield.ac.uk
2. Department of Chemistry and Analytical Sciences, Open University, Milton Keynes, MK7 6AA, UK; E-mail: c.turner@open.ac.uk
3. Department of Dermatology, Amersham Hospital, Amersham, Buckinghamshire, HP7 0JD, UK; E-mail: carolyn.willis@bucks.nhs.uk
4. Electronic Supplementary Information (ESI) available: [Details on urine samples (provided in spreadsheet format)]. See DOI: 10.1039/b000000x/