Published in final edited form as:

J Breath Res. ; 10(1): 017106. doi:10.1088/1752-7155/10/1/017106.

The use of a gas chromatography-sensor system combined with advanced statistical methods, towards the diagnosis of urological malignancies

Raphael B. M. Aggio, PhD¹, Ben de Lacy Costello, PhD², Paul White, PhD³, Tanzeela Khalid, PhD^{1,4}, Norman M. Ratcliffe, PhD², Raj Persad, ChM, FRCS(Eng), FRCS(Urol), FEBU², and Chris S. J. Probert, MD, FRCP, FHEA¹

¹Institute of Translational Medicine, Department of Cellular and Molecular Physiology, University of Liverpool, Liverpool, United Kingdom

²Institute of Biosensor Technology, University of the West of England, Bristol, United Kingdom

³Department of Engineering, Design and Mathematics, University of the West of England, Bristol, United Kingdom

⁴Department of Surgery and Cancer, Imperial College London, London, United Kingdom

Abstract

Prostate cancer is one of the most common cancers. Serum prostate-specific antigen (PSA) is used to aid the selection of men undergoing biopsies. Its use remains controversial. We propose a GC-sensor algorithm system for classifying urine samples from patients with urological symptoms. This pilot study includes 155 men presenting to urology clinics, 58 were diagnosed with prostate cancer, 24 with bladder cancer and 73 with haematuria and or poor stream, without cancer. Principal component analysis (PCA) was applied to assess the discrimination achieved, while linear discriminant analysis (LDA) and support vector machine (SVM) were used as statistical models for sample classification. Leave-one-out cross-validation (LOOCV), repeated 10-fold

Contact information for corresponding author: Norman M. Ratcliffe, Institute of Biosensor Technology, University of The West of England, Coldharbour Lane, Frenchay, Bristol BS16 1QY, norman.ratcliffe@uwe.ac.uk, Phone: +44 3282501, Mobile: +44 7796920793.

Author contribution

Aggio, R.B.M. developed the computer algorithm for pattern recognition and sample classification, performed the data analysis, wrote the manuscript and produced the figures. De Lacy Costello, B. developed the GC-sensor device, supported the analysis of urine samples using the GC-sensor and revised the manuscript. White, P. supported the development of the computer algorithm, supported the data analysis and revised the manuscript. Khalid, T. collected urine samples, performed their analysis using the GC-sensor and revised the manuscript. Ratcliffe, N.M. developed the GC-sensor device, supported the analysis of urine samples through the GC-sensor and revised the manuscript. Persad, R. supported the data analysis and revised the manuscript. Probert, C.S.J. developed the GC-sensor device, supported the analysis of urine samples through the GC-sensor, supported the data analysis and revised the manuscript.

Financial disclosure

R.B.M. Aggio, B. de Lacy Costello, N.M. Ratcliffe, Persad, R. and C.S.J. Probert certifies that all conflicts of interest, including specific financial interests and relationships and affiliations relevant to the subject matter or materials discussed in the manuscript (eg, employment/affiliation, grants or funding, consultancies, honoraria, stock ownership or options, expert testimony, royalties, or patents filed, received, or pending), are the following: B. de Lacy Costello, N.M. Ratcliffe and C.S.J. Probert are the inventors of the intellectual property (IP) related to applications of the GC-Sensor. The IP is owned by their employers, the University of Liverpool and the University of West of England. In addition, R.B.M. Aggio and C.S.J. Probert are the inventors of the IP related to the pipeline used here for data analysis. The University of Liverpool owns this IP. The other authors have nothing to disclose.

cross-validation (10FoldCV), repeated double cross-validation (DoubleCV) and Monte Carlo permutations were applied to assess performance.

Significant separation was found between prostate cancer and control samples, bladder cancer and controls and between bladder and prostate cancer samples. For prostate cancer diagnosis, the GC/SVM system classified samples with 95% sensitivity and 96% specificity after LOOCV. For bladder cancer diagnosis, the SVM reported 96% sensitivity and 100% specificity after LOOCV, while the DoubleCV reported 87% sensitivity and 99% specificity, with SVM showing 78% and 98% sensitivity between prostate and bladder cancer samples. Evaluation of the results of the Monte Carlo permutation of class labels obtained chance-like accuracy values around 50% suggesting the observed results for bladder cancer and prostate cancer detection are not due to over fitting.

The results of the pilot study presented here indicate that the GC system is able to successfully identify patterns that allow classification of urine samples from patients with urological cancers. An accurate diagnosis based on urine samples would reduce the number of negative prostate biopsies performed, and the frequency of surveillance cystoscopy for bladder cancer patients. Larger cohort studies are planned to investigate the potential of this system. Future work may lead to non-invasive breath analyses for diagnosing urological conditions.

Keywords

Prostate Cancer; Bladder cancer; Volatile organic compounds; Sensor; Gas chromatography; Metabolomics; Pattern recognition

1 Introduction

Prostate cancer is one of the most common cancers worldwide [1] and one of the leading causes of cancer mortality in males [2]. It is generally associated with poor quality of life, increased morbidity, expensive, invasive and potentially hazardous diagnostic methods and treatments. Despite this, there remains a lack of accurate non-invasive screening methods [3,4]. Serum prostate-specific antigen (PSA) is currently the main biomarker used to help select men needing prostate biopsies [2], although its use worldwide remains controversial [3,5]. It has not been approved for use in screening programs in most countries because of its low specificity [3] (25 to 40%), which results in a high negative biopsy rate with consequent potential adverse psychological impacts and potential overtreatment of clinically insignificant tumours [6]. In 1994, serum PSA was approved by the Food and Drug Administration (FDA) for prostate cancer screening in the USA at a cut-off of 4ng/ml [7]. Recently, the US Preventive Services Task Force (USPSTF) released new recommendations limiting the use of the PSA test [2]. PSA levels are highly associated with the prostate volume, which is known to increase with age [8]. For this reason, age-specific PSA cut-off levels have been proposed [9]. However, other medical conditions such as urinary tract infections, prostatitis and benign prostatic hypertrophy are also known to increase serum PSA levels [10]. As a consequence, the diagnosis of prostate cancer based on PSA is currently associated with a high risk of false positives and false negatives [7,11]. Ultimately,

it results in unnecessary biopsies, potential psychological toll, infections that can occur after biopsies and missing cancer cases [2,12].

Great attention and investment have been applied to investigate and develop tests that are more sensitive and specific than PSA [13–18]. Among them, prostate cancer antigen 3 (PCA3) [15,16], annexin A3 [18], Engrailed-2 protein (EN2) [17], human kallikrein 2 [19], the fusion gene TMPRSS2:ERG [15] and PSA related methods [13,14] are perhaps the most studied. Despite the apparently promising results showed by these potential biomarkers, some require further validation whilst others were unable to be replicated in additional studies [12]. None of these markers are currently in use for screening or diagnosing prostate cancer. Similarly, a few studies have reported the ability of dogs to discriminate prostate cancer from control urine samples with accuracies over 90% [20–23]. However, these results must be validated in larger-scale experiments. In addition, the use of dogs in daily practice is highly questionable [22,23], with high training costs, ethics and hygiene issues that must be carefully discussed.

The apparent ability of dogs to identify prostate cancer from urine samples indicates that volatile organic compounds (VOCs) present in the urine may be potential biomarkers for prostate cancer. VOCs may indicate the metabolic state of a cell or organism [24] and considerable attention has been given to the use of VOCs as biomarkers of specific medical conditions.

Therefore, new technologies have been utilised, e.g. electronic nose technologies, [25–32] to capture differences in VOC levels present in the urine of patients with cancer and cancer-free patients. Roine *et al.* [25] recently reported a sensitivity of 78%, and a specificity of 67% for prostate cancer detection, similar to Khalid *et al.* [26], while Asimakopoulos *et al.* [27] reported 71% sensitivity and 93% specificity.

In the last few years, improvements have been made in sample preparation methods [33], analytical platforms for measuring VOCs' levels in biological samples [34] and in the computer algorithms used for data analysis [35–37]. As a result, many studies have revealed VOCs with potential to be used as biomarkers for many medical conditions e.g. lung cancer [38], breast cancer [39], tuberculosis [40] and gastrointestinal and liver diseases [41]. However, many of the suggested methods are expensive and have limited statistical validation [42].

Breath testing is an attractive idea: it is non-invasive. It assumes that cancer-associated VOCs enter blood and are exhaled by diffusion. The investigation of VOCs in breath samples of patients with a range of cancers has been reported, although there has only been one report for prostate cancer and none for bladder cancer [38,39]. Peng *et al.* [38] included 18 patients with prostate cancer in the study: their findings were presented as PCA plots, there was little overlap between prostate cancers and controls, without any attempt to report accuracy. These promising results have yet to be replicated. However, we chose to investigate VOCs in urine of men with prostate cancer and bladder cancer because of the face-validity of investigating a fluid that has been in direct contact with the organs in which

the cancers are arising. Furthermore, men with urological symptoms routinely provide urine samples at the clinic.

We have developed a system composed of a gas chromatography column coupled to a metal-oxide gas sensor (GC-sensor) [28, 42] and a computer algorithm or pipeline. The pipeline involves chromatogram alignment, data transformation and feature selection algorithms [43, 44] in order to identify the features that best differentiate the medical conditions under study. These features or patterns are then used with preferred statistical modelling techniques to classify unknown samples.

Here, we present a GC-sensor pipeline and evaluate its performance in classifying urine samples from a pilot study involving 58 patients with prostate cancer and 73 patients with urological symptoms. In order to assess how this new GC-sensor pipeline will classify samples of a different urological cancer we have also re-analysed a previously published set of 24 urine samples from patients with bladder cancer [28]. Future work using non-invasive breath analyses may lead to detection of circulating blood borne biomarkers arising from urological diseases.

2 Materials and Methods

2.1 Patient recruitment

Inclusion criteria were men presenting to urology clinics with lower urinary tract symptoms and men scheduled for prostate biopsy based on elevated serum PSA levels (> 4 ng/ml) or abnormal findings on digital rectal examination. The criteria for controls was exactly the same as for the patients with prostate cancer i.e. men presenting to urology clinics with lower urinary tract symptoms. Exclusion criteria were history of urothelial carcinoma or other known malignancies, a urinary tract infection, or a urinary catheter in situ. Subjects were recruited from the Bristol Urological Institute and Bristol Royal Infirmary in Bristol, England. Urine samples from patients with prostate cancer were provided on the morning prior to their trans-rectal ultrasound guided prostate biopsy. Urine samples from patients with bladder cancer were collected as described by Khalid and co-workers [28], where each patient provided a urine sample on the morning prior to their cystoscopy or biopsy. All urine samples were classified after pathological examination of biopsy specimens and/or cystoscopy. Ethical approval for the study was obtained from the Wiltshire Research Ethics Committee (REC reference number 08/H0104/63) and R&D approval from the University Hospitals Bristol NHS Foundation Trust. Participants were recruited over a 13-month period between 2009 and 2011. Each participant reviewed an information sheet and gave written consent.

2.2 Assay method

Aliquots of 0.75ml of fresh urine were stored in septum-topped glass headspace vials (Supelco, Sigma Aldrich, Dorset, UK) and frozen at -20°C .

Urine samples were defrosted by immersing the vial in a water bath at 60°C for 30 seconds, 0.75ml of 1M sodium hydroxide (Fisher Scientific, Leicestershire, UK) was then added to the urine sample and the mixture was re-immersed in the water bath at 60°C for 50 minutes.

Finally, 2cm³ of headspace air was extracted from the vial using an airtight Hamilton 10ml gas syringe (Fisher Scientific Ltd.) and immediately injected into the inlet of the GC-sensor system. The GC system was fitted with a split/splitless injector run in splitless mode, but with a septum purge of 5ml/min.

2.3 Gas chromatography – sensor

The basic characteristics of the GC-sensor have been reported elsewhere [28,42,45,46]. It is composed of a gas chromatography (GC) oven (Clarus 500, Perkin Elmer) fitted with a commercially available capillary column (30 meter long SPB™-1 SULFUR with an inner diameter of 0.32 mm and a film thickness of 4 µm from Sigma Aldrich, Dorset, UK) interfaced to a heated (450°C) metal oxide sensor (MOS chemresistor) used as the detector. The metal oxide sensor itself is a composite of tin oxide and zinc oxide (50:50 by wt.), coated onto a gold inter-digitated alumina square, with a platinum heater on the reverse side [45]. The composite sensor reversibly changes electrical resistance on adsorption of VOCs and is extremely sensitive to a range of volatiles, covering a wide range of chemical classes. For instance it can detect dimethyl disulphide and butanol down to extremely low concentrations, 0.10 and 0.025ppm, respectively (in air, i.e. when not coupled to a GC) [45]; and ethanol, propanol, 3-octanone, diacetyl, butanal, ethylbenzene and octane at 2.5ppm and below [46]. In terms of an SPME-GC-sensor combination, one example shows the detector has a limit of detection for the headspace of aqueous butanol solution at 1.62microgram/L, which is superior to a PE Clarus 500 mass spectral detector, (unpublished data). The injection port of the GC was fitted with a 1 mm quartz liner and heated to a temperature of 150°C. Cylinder air (BOC, Guilford, UK) was used as carrier gas at a pressure of 35 psi (this gives a column flow of 13.4ml/min), which was passed through an air purifier (300 ml Supelcarb™ hydrocarbon trap, Sigma-Aldrich, Dorset, UK). The GC temperature program used in this study was: initial GC oven temperature held at 30°C for 6 minutes, then ramped at a rate of 5°C per min to 100°C with a 22 minute hold, giving a total run time of 42 minutes. The GC sensor system was calibrated daily using a certified gas standard (1% tolerance) of 50 ppm ethanol in blended air (Air Products Plc, Speciality Gases, Crewe CW1 6AP).

The sensor temperature is controlled and measured via an electronic circuit monitored by computer software. The circuit also measures the electrical resistance of the sensor and the computer software records this resistance at 0.5 second intervals. The VOCs present in the headspace of the urine mixture are passed through the GC, where they are separated according to their molecular weight, boiling point, polarity and chemical functionality. When these VOCs reach the metal oxide gas sensor they change the resistance of the metal oxide film, which is then recorded by the computer software. Before applying any data processing steps, the resistance signals generated by the GC-sensor for each sample were inverted to facilitate the use of algorithms applied to the chromatography data. The resistance profile of each sample generated by the GC-sensor was stored in individual files using CDF format.

2.4 The pipeline

In order to facilitate the description of the pipeline we present here, we divided it in 8 different stages described in Figure 1.

2.5.1 Load data – Stage 1—The dataset generated by the GC-sensor is loaded into the R software [47], using R version 3.1.1.

2.5.2 Baseline correction – Stage 2—More accurate results are obtained if the baseline is corrected prior to data normalization, alignment, transformation and classification. Samples generated by the GC-sensor have their baseline removed in two steps (Figure 2). First, iterative restricted least squares [48,49] is applied using the R package baseline [50] (Figure 2 – red line). Then, a resistance threshold is calculated. This threshold is defined by the average resistance value in a sample minus the standard deviation of its resistance values (Figure 2 – orange line). The value of the threshold is subtracted from every value in the chromatogram (Figure 2 – green line). Any negative value, which represent a resistance value lower than the resistance threshold, is set to zero.

2.5.3 Data normalization – Stage 3—A sample is normalized through the division of its resistance values by the highest resistance value registered for this sample. The normalized resistance values of every sample will, therefore, range between 0 to 1.

2.5.4 Alignment – Stage 4—Chromatograms are aligned in a two-step approach. First, the best sample to be used as reference is selected and, then, every sample is aligned to the reference sample. For this, a set of Pearson's correlation coefficients is calculated for each pair of samples (Figure 3). For each pair of samples, 31 Pearson's correlation coefficients are calculated: 1 coefficient calculated with no shifting; 15 coefficients calculated when shifting one of the chromatograms 1 sampling point at each calculation; and 15 coefficients calculated when shifting the same chromatogram -1 sampling point at each calculation. As a result, each chromatogram will be associated with one numeric matrix containing its correlation coefficients in relation to all the other samples from the same experimental condition. Here, each row stores the coefficients for a specific pair of chromatograms at the different sampling points shifted. The maximum coefficient or the highest positive value of Pearson's correlation for each row or pair of chromatograms is obtained and the mean of the maximum coefficients of a matrix are calculated. The sample showing the highest positive correlation coefficient on average is then selected as the reference sample for alignment. The identity of the sample chosen as reference is stored in order to align future unknown samples. This step of selecting the reference sample is performed once, when training the system. Then, in a second step, every sample from every experimental condition is aligned in relation to the selected reference sample using the sampling point that generates the highest positive correlation coefficient between samples. Similarly to correlation optimised warping (COW) [51], the alignment method proposed here makes use of correlation to define the optimal alignment between chromatograms. Our method aims at searching, within a dataset, for the best reference chromatogram for data alignment. All the chromatograms in the dataset are then aligned in relation to this reference. In contrast, the COW method aligns a pair of chromatograms using stretching, compression and correlation.

2.5.5 Data transformation – Stage 5—After chromatogram alignment, a set of transformation steps is applied to datasets generated by the GC-sensor in order to facilitate the discrimination between samples belonging to different medical conditions. First, the resistance values of each sample are converted to the modulus of their wavelet coefficients using the scale 1 of the Mexican hat wavelet [52]. Then, the value 1 is added to each data point of every sample and the whole dataset is log-transformed using its natural logarithm. A range transformation is applied to set the transformed resistance values ranging between 0 and 1. Finally, these values are processed using spatial sign [53], which transforms the data to their spatial signs followed by the computation of a normal covariance matrix. The spatial sign transformation minimizes the potential effect that outliers may have when building classifiers.

2.5.6 Feature selection – Stage 6—All the stages described above are applied with no discrimination between data classes or medical conditions. All samples are equally processed. The data class associated to each sample is only considered at stage 6, where the features that best describe the differences between medical conditions are selected.

The difference in VOC levels across medical conditions is translated into the resistance levels reported for each sample. Each sampling point or resistance measurement performed by the GC-sensor is a feature with the potential to differentiate between samples belonging to different medical conditions. Therefore, the features that best describe the differences between medical conditions must be selected before constructing the statistical models that will be used in the classification of new samples.

At this point, the dataset generated by the GC-sensor has been normalized, aligned and transformed. Two different algorithms, Boruta [43] and recursive feature selection (RFE) [44], based on random forest are applied to the entire dataset in order to select the most important features to be used in sample classification. Ideally, the feature selection algorithms may be applied to part of the dataset, leaving the rest of the dataset reserved for validation. However, this approach requires a large dataset, which is not the case in the pilot study presented here. In summary, the Boruta algorithm involves the development of decision trees based on different sets of samples. Random forest is applied to calculate the loss of accuracy of classification when the values of features are randomly permuted between disease groups. Features associated with loss of accuracy are selected as important features. Boruta was applied using significance level of 0.01, 300 maximal number of importance source runs and multiple comparison adjustment. The RFE algorithm works similarly to Boruta, however, it eliminates features that produce no change in the accuracy level, instead of selecting features that produce loss of accuracy. It was applied to identify the subset of predictors or features that produced the best accuracy after 10 repeats of 10-fold cross-validation. As Boruta and RFE may miss some important features [54], the features reported as important by at least one of these algorithms are then selected for building classifiers. The two feature selection algorithms used here, Boruta and RFE, were chosen because they are well-known tools applied in academia [55] and in commercial data mining [54]; and they are easily implemented in R.

2.5.7 Building classifiers – Stage 7—The features that best describe the differences between medical conditions were selected in Stage 6. These features are used to build the classifiers that will be applied to sample classification. Examples of these classifiers are: linear discriminant analysis (LDA); partial least squares (PLS); random forest (RF); k-nearest neighbour (KNN); support vector machine (SVM) with radial basis function kernel (SVM-R); SVM with linear basis function kernel (SVM-L); and SVM with polynomial basis function kernel (SVM-P) [44, 56–60]. The performance of each classifier depends on the dataset being analysed and the parameters used when building the classifier. The pipeline presented here was developed to identify the features or patterns that best describe the difference between medical conditions. Different statistical modelling techniques may subsequently be constructed using these features to classify or diagnose unknown samples. Here, we used LDA and SVM-P as modelling techniques for sample classification. There is a great amount of material in the literature describing the rational behind classification methods, how they are constructed and how to evaluate their performance [61].

2.5.8 Validation – Stage 8—Building and testing classifiers on the same dataset may produce biased and overoptimistic results due to potential over fitting [62]. Therefore, validation schemes should be used to prevent over fitting. Repeated k-fold cross-validation [63] and repeated double cross-validation [64] are highly recommended. Therefore, the pipeline was developed to validate the performance of classifiers using three validation schemes: leave-one-out cross-validation (LOOCV), 30 repeats of 10-fold cross-validation (10FoldCV), (Figure 4a) and 30 repeats of the 3-fold double cross-validation with an inner loop of 10-fold cross-validation repeated 5 times (DoubleCV), (Figure 4b). In addition, these validation schemes are repeated on the same data sets, however, applying a Monte Carlo random permutation of class labels in each repeat (Figures 4c and 4d). Classifiers are built and tested using the R package caret [44], version 6.0-30. There is no feature selection process involved in the validation step.

2.5.9 Classifying samples—The pipeline above (Figure 1) describes in detail how the chromatogram data is processed, features are selected and classifiers are built. Samples were analysed by the GC-sensor system, a process performed in approximately 90 minutes, which includes sample thawing, sample incubation, separation and data analysis.

2.5.10 Study design and data analysis—In this study, the new pipeline was applied to a pilot dataset containing urine samples from male patients with prostate cancer (n=58), bladder cancer (n=24) and controls with urological symptoms (haematuria and or poor stream) that had negative investigations for cancer (n=73).

The in-house-developed pipeline aligns chromatograms, applies data transformation techniques and uses Mexican hat wavelet to extract their high-frequency content. Two algorithms based on the random forest technique select the features that best describe the differences between medical conditions. Principal component analysis (PCA) on the transformed resistance values was performed to visualise the discrimination achieved by the framework. Statistical analyses were performed solely on the resistance profiles processed by the pipeline using R software [47]. Here, LDA and SVM-P were used as classifiers to diagnose unknown samples. We have applied these two classifiers in order to investigate the

performance of a linear and a non-linear classifier. No prognostic factors were considered for modelling and classifying samples from prostate, bladder cancer and the control groups. No discrimination was used in relation to the grade of cancer diagnosed. The performance of the two classifiers were validated using three validation schemes: LOOCV, 10FoldCV (Figure 4a) and DoubleCV, (Figure 4b). In addition, these three cross-validation schemes were repeated on the same data sets, however, applying a Monte Carlo random permutation of class labels in each repeat (Figures 4c and 4d). Monte Carlo technique helps detect potential bias associated with the data. Receiver operating characteristic (ROC) curves were constructed to visualize the performance of classifiers. Classifiers were built and tested using the R package caret [44]. Confidence intervals (CI) were calculated using a bootstrap technique based on 1000 repetitions of sampling 25 values with replacement.

In order to further compare the results produced by the framework with the PSA test, we calculated the sensitivity and specificity of PSA testing using the routine clinical results of the participants. PSA values were compared with the age-specific cut-offs defined by Burford [65], as recommended by NICE (2015) [66]: these were 3 ng/ml for men aged 50-59, 4 ng/ml those aged 60-69 and 5 ng/ml for those aged 70 and over. In addition, we determined the sensitivity and specificity of PSA based on a cut off of 6 ng/ml.

3 Results

3.1 Classifying prostate cancer and control samples

Urine samples from 58 patients diagnosed with prostate cancer and 73 male controls with urological symptoms (haematuria and or poor stream) were analysed by the GC-sensor-pipeline. Demographics for the patient groups studied are presented in Table 1, while the Gleason scores of patients with prostate cancer are presented in Table 2.

3.1.1 GC-sensor pipeline—The GC-sensor pipeline selected 23 features to discriminate urine samples from patients with prostate cancer and controls (Figure 5a). The discrimination achieved by these selected features is illustrated by the PCA (Figure 5b). The first two components of the PCA performed on the data processed by the GC-sensor pipeline show two main clusters of samples, prostate cancer and control, with a region of overlap in between.

The features selected by the GC-sensor pipeline were applied to build a statistical model using LDA and SVM-P. The results of the LOOCV are presented in Table 3. The average and median accuracy, sensitivity and specificity reported by the 10FoldCV and DoubleCV schemes when classifying prostate cancer and control samples are shown in Table 4, while the Monte Carlo random class labels permutation can be found in the supporting information (Tables S1 and S2).

The classification results reported by LDA and SVM-P were similar, with SVM-P performing slightly better than LDA after every validation scheme (Tables 3 and 4). Using SVM-P, the GC-sensor pipeline framework was able to classify prostate cancer and control samples with 95% sensitivity and 96% specificity after LOOCV. After 10FoldCV, SVM-P reported a mean sensitivity of 88% (95% CI [83.9-92.4]) and mean specificity of 93% (95%

CI [89.9-95.7]). After a more stringent method, the DoubleCV, SVM-P reported a mean sensitivity of 86% (95% CI [82.8-88.1]) and mean specificity of 91% (95% CI [90.0-93.3]). The discrimination achieved by the framework after DoubleCV is well illustrated by the Receiver Operating Characteristic (ROC) curves presented in Figure 6. The Monte Carlo random class labels permutation resulted in accuracies around 50% (see Tables S1 and S2), which is expected when samples are classified simply by chance.

3.1.2 Discrimination obtained solely by PSA levels—The diagnostic ability of the PSA test was assessed by considering only the patients respective measured PSA levels and NICE (2015) recommendations [66]. PSA levels were available for 55 of the 58 prostate cancer samples and 53 of the 74 control samples. The results of sample classification based solely on the measured PSA levels are available in Table 5.

3.1.3 Biomarkers found in the literature and the framework presented here—In order to have a better idea of how the results produced by the framework GC-sensor pipeline differs from biomarkers found in the literature, Figure 7 shows a summary of sensitivities and specificities of these biomarkers.

3.2 Classifying bladder cancer and control samples

Urine samples from 24 patients diagnosed with bladder cancer and 73 male controls with urological symptoms (haematuria and or poor stream) were analysed by the GC-sensor pipeline. The demographics are presented in Table 6.

After data normalization, chromatogram alignment and data transformation, the pipeline identified 21 important features for the differentiation of bladder cancer and control samples (Figure 8a). PCA was applied using these features to visualize the level of discrimination achieved by the pipeline (Figure 8b). The first two components of the PCA show a clear separation between two clusters: a cluster containing samples from patients with bladder cancer; and a cluster containing samples from non-cancerous controls.

The sensitivities and specificities reported by the LOOCV are presented in Table 7, the results of the 10FoldCV and the DoubleCV schemes are reported in Table 8, and their Monte Carlo random class labels permutation can be found as Supplementary Tables S3 and S6.

LDA and SVM-P showed very similar performances (Tables 7 and 8). LDA, for example, was able to classify samples with 96% and 99% sensitivity and specificity, respectively, after LOOCV. The 10FoldCV showed sensitivity of 92% (95% CI [90.4-93.6]) and specificity of 98% (95% CI [95.5-99.5]), while the DoubleCV reported 87% (95% CI [84.0-90.5]) and 96% (95% CI [94.1-97.1]) sensitivity and specificity, respectively. These results are further illustrated by the ROC curve built on the results of the DoubleCV and presented in Figure 9. The Monte Carlo random class labels permutation resulted in accuracies around 50% (see Supplementary Tables S3 and S4), which is expected when samples are classified simply by chance.

For comparison of data from analysing bladder cancer versus prostate cancer urine samples, repeated doubleCV and using 2 classification methods, LDA and SVM-P, gave sensitivities of 77.9% and 83.5% and specificity of 91.9% and 97.6% respectively.

3.2.1 Comparison with existing biomarkers—In order to illustrate how the results reported by the pipeline presented here differ from biomarkers found in the literature, Figure 10 shows a forest plot with the summary of sensitivities and specificities from some of these biomarkers and the pipeline presented here.

4 Discussion

Here, we presented and applied a GC-sensor pipeline to classify a pilot dataset containing urine samples from patients with prostate cancer and patients with urological symptoms. These results indicate that the pipeline was able to successfully detect the patterns of VOCs discriminating prostate cancer and control samples. It is illustrated in the PCA (Figure 5), where samples from prostate cancer and control patients form two visible distinct clusters with a narrow overlap region in between.

The separation reported by the PCA was confirmed by the two different statistical modelling techniques applied, LDA and SVM-P. SVM-P performed slightly better than LDA, with sensitivity and specificity of 95% and 96%, respectively, after LOOCV (Table 3). These results were further validated using repeated 10FoldCV and DoubleCV (Table 4). The 10FoldCV returned 88% sensitivity and 93% specificity, and the DoubleCV returned 86% sensitivity and 91% specificity.

In contrast to the GC-sensor pipeline, the PSA testing performed badly in this cohort. The overall accuracy was 56% using age-specific cut offs and 64% with an arbitrary cut-off of 6 ng/ml; the specificity was 9% and 43% for the two methods, respectively. Clearly, the cancer patients and controls were a biased set of subject by the nature of their referral to an urologist with lower urinary tract symptoms. In this study a significant number of men without cancer (48/54) had undergone a TRUS biopsy due to exceeding their age-specific PSA cut off values, however the implementation of these age-specific thresholds meant that no cancer cases were missed for further investigation. If a cut off of 6 ng/ml had been used, then 30/54 men would have had a negative biopsy, but 9/55 cancers would have been missed.

In order to further test the framework presented here, the same GC-sensor pipeline was applied to classify urine samples from patients with bladder cancer. The pipeline was able to detect the patterns differentiating bladder cancer cases from controls, as illustrated in the PCA (Figure 8), where two distinct clusters were visible. The separation reported by the PCA was confirmed by the two different statistical modelling techniques applied: LDA and SVM-P. Both classifiers showed similar performances, with LDA reporting accuracies of 98%, 96% and 94% after LOOCV, 10FoldCV and DoubleCV, respectively. These results are similar to the results reported by Khalid and co-workers [28], however, here more stringent validation schemes were applied. The comparison of urine samples from patients with bladder and prostate cancer is not commonly performed in clinical practice, however both

LDA and SVM-P classification methods gave significantly different results, indicating that different VOCs are associated with the two urological cancers.

We consider that, ideally, a fraction of the dataset (e.g. 20%) being analysed should be used exclusively for validation. It is recommended to reduce the risk of bias and over fitting. However, this approach of validation requires a large dataset, which can be difficult to obtain in clinical studies. The high costs involved and the restricted access to samples in these experiments are generally the main factors limiting the size of the dataset to be analysed. Here, for example, we report results from datasets containing 58 and 24 samples in some of the data classes analysed. In these cases, cross-validation schemes are highly recommended alternatives [63, 64]. The cross-validation schemes and Monte Carlo methods we applied in this study provide an estimate of out-of-sample predictive accuracy for similar populations and help identify possible over fitting. We have observed high average and median accuracies supported by 95% confidence intervals after DoubleCV, which is considered a stringent method. In addition, the results of the Monte Carlo permutation of class labels obtained chance-like accuracy values around 50%. These results suggest no bias or over fitting. Therefore, although no external validation was applied, the results reported here indicate that the methodology was able to successfully differentiate urine samples from patients with prostate and bladder cancer, from control urine samples from patients presenting with urological symptoms. In addition, it is important to note that the study reported here was designed using, as controls, patients with a mix of lower urinary tract symptoms associated with prostatic and/or urinary abnormalities that needed investigating by invasive tests to rule out the presence of cancer. It makes the discrimination between samples from patients with and without cancer a more difficult task to be performed, however, it also increases ecological validity of the results presented here.

No direct comparison can be made between the results reported here and the results reported for other biomarkers (Figures 7 and 10), as the accuracy of the test depends on the population used. However, Figures 7 and 10 provide a reasonable indication of the potential of the GC-sensor pipeline for diagnosing prostate and bladder cancer. Furthermore, a different threshold can be applied to LDA to define the relation between sensitivity and specificity [67] that fits best with the final application of the model. The ideal sensitivity and specificity for screening and diagnosing prostate cancer are not well defined by the medical community. In fact, this is a subject of great discussion [15], therefore only the results optimized for accuracy are reported here.

With regard to the hardware aspects of the GC-sensor system, air was used as carrier gas because the ultimate goal is a standalone system for diagnostic purposes and use of cylinders hinders this advancement. The air cylinder could readily be replaced by a small air compressor combined with air scrubbers in future work. A low column temperature of 100°C was used to minimise column degradation. A recent published work [68] has shown there is very little degradation under these conditions w.r.t retention times, using standard mixtures (50ppm ethanol/methanol), common peaks from urine samples and a homogenised stool mixture. A blank run between samples showed that carryover of VOCs from previous runs causing suppression of the baseline was not an issue. As this is a relatively new technique (GC-sensor combined), method development is on-going to optimise the long

term performance and stability of the instrument. We run standards and control samples daily to assess the long-term performance and stability of the technique. We are also combining the GC-sensor system with a mass spectrometer so that markers of disease can potentially be identified in future large scale studies.

The use of the GC-sensor pipeline in clinical practices would involve the following steps: an unknown sample would be analysed by the GC-sensor system; it would be aligned to a reference chromatogram previously defined when training the system; and quickly classified based on the previously selected features and built classifiers.

Although the work presented here apparently has produced promising results for the diagnosis of prostate and bladder cancer, a large-scale multicentre study including an individual dataset for external validation is necessary to confirm its performance and speculate about its use in clinical practice. Urine is a highly variable fluid affected by diet, dehydration, medicines and dehydration. In this study we deliberately didn't attempt to filter out or compensate for such factors, as a diagnostic method should ideally be as simple as possible. A new study involving a greater number of samples from patients with prostate and bladder cancer is planned. Factors, such as those mentioned above and ethnicity, age, smoking status and family history will be considered when modelling features selected by the GC-sensor pipeline, which may improve the accuracy reported here.

Conclusion

The performance of an in-house-developed method for the diagnosis of urine samples from patients with prostate cancer and non-cancerous patients with urological symptoms was evaluated and showed promising results. A larger scale study is currently being developed. Future work will also involve use of a mass spectral detector to determine the identity of key VOC metabolites to distinguish the urological cancers. This will be valuable information to help choose sensors targeted to these compounds. It is hoped that VOC based analysis systems will considerably reduce the number of prostate biopsies and follow-up cystoscopies performed for prostate and bladder cancer, respectively.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Contract grant acknowledgement

We would like to acknowledge two funding sources: the Rotary club (<https://www.rotary.org/en>) and Ralph Shackman Trust (Woodfines Llp Lockton House Clarendon Road Cambridge, CB28FH, <http://opencharities.org/charities/287406>). The authors also wish to acknowledge support from the Wellcome Trust.

References

1. [Accessed June 6 2015] Cancer Research UK Web site. <http://www.cancerresearchuk.org/cancer-info/cancerstats/types/bladder/incidence/>
2. Garg V, Gu NY, Borrego ME, Raisch DW. A literature review of cost-effectiveness analyses of prostate-specific antigen test in prostate cancer screening. Expert Review of Pharmacoeconomics & Outcomes Research. 2013; 13(3):327–342. [PubMed: 23763530]

3. Bradley LA, Palomaki GE, Gutman S, Samson D, Aronson N. Comparative Effectiveness Review: Prostate Cancer Antigen 3 Testing for the Diagnosis and Management of Prostate Cancer. *Journal of Urology*. 2013; 190(2):389–398. [PubMed: 23545099]
4. Snyder CF, Frick KD, Blackford AL, Herbert RJ, Neville BA, Carducci MA, Earle CC. How Does Initial Treatment Choice Affect Short-Term and Long-Term Costs for Clinically Localized Prostate Cancer? *Cancer*. 2010; 116(23):5391–5399. [PubMed: 20734396]
5. Etzioni RD. Review of evidence concerning PSA screening for prostate cancer has limitations as basis for policy development. *Evidence-based medicine*. 2013; 18(2):75–6. [PubMed: 22773765]
6. Roddam AW, Duffy MJ, Hamdy FC, Ward AM, Patnick J, Price CP, Rimmer J, Sturgeon C, White P, Allen NE. Management, N. H. S. P. C. R., Use of prostate-specific antigen (PSA) isoforms for the detection of prostate cancer in men with a PSA level of 2-10 ng/ml: Systematic review and meta-analysis. *European Urology*. 2005; 48(3):386–399. [PubMed: 15982797]
7. Catalona WJ, Southwick PC, Slawin KM, Partin AW, Brawer MK, Flanigan RC, Patel A, Richie JP, Walsh PC, Scardino PT, Lange PH, et al. Comparison of percent free PSA, PSA density, and age-specific PSA cutoffs for prostate cancer detection and staging. *Urology*. 2000; 56(2):255–260. [PubMed: 10925089]
8. Oesterling JE, Jacobsen SJ, Chute CG, Guess HA, Girman CJ, Panser LA, Lieber MM. Serum Prostate-Specific Antigen In A Community-Based Population Of Healthy-Men - Establishment Of Age-Specific Reference Ranges. *Jama-Journal of the American Medical Association*. 1993; 270(7):860–864.
9. Liu Z-Y, Sun Y-H, Xu C-L, Gao X, Zhang L-M, Ren S-C. Age-specific PSA reference ranges in Chinese men without prostate cancer. *Asian Journal of Andrology*. 2009; 11(1):100–103. [PubMed: 19050693]
10. Diamandis EP. Early Prostate Cancer Antigen-2: A Controversial Prostate Cancer Biomarker? *Clinical Chemistry*. 2010; 56(4):542–544. [PubMed: 20167690]
11. Wolf AMD, Wender RC, Etzioni RB, Thompson IM, D'Amico AV, Volk RJ, Brooks DD, Dash C, Guessous I, Andrews K, DeSantis C, et al. American Cancer Society Guideline for the Early Detection of Prostate Cancer Update 2010. *Ca-a Cancer Journal for Clinicians*. 2010; 60(2):70–98. [PubMed: 20200110]
12. Tuma RS. New Tests for Prostate Cancer May Be Nearing the Clinic. *Journal of the National Cancer Institute*. 2010; 102(11):752–754. [PubMed: 20498428]
13. Herwig R, Mitteregger D, Djavan B, Kramer G, Margreiter M, Leers MP, Glodny B, Haider DG, Hoerl WH, Marberger M. Detecting prostate cancer by intracellular macrophage prostate-specific antigen (PSA): a more specific and sensitive marker than conventional serum total PSA. *European Journal of Clinical Investigation*. 2008; 38(6):430–437. [PubMed: 18489403]
14. Lazzeri M, Haese A, de la Taille A, Redorta JP, McNicholas T, Lughezzani G, Scattoni V, Bini V, Freschi M, Sussman A, Ghaleb B, et al. Serum Isoform -2 proPSA Derivatives Significantly Improve Prediction of Prostate Cancer at Initial Biopsy in a Total PSA Range of 2-10 ng/ml: A Multicentric European Study. *European Urology*. 2013; 63(6):986–994. [PubMed: 23375961]
15. Leyten GHJM, Hessels D, Jannink SA, Smit FP, de Jong H, Cornel EB, de Reijke TM, Vergunst H, Kil P, Knipscheer BC, van Oort IM, et al. Prospective Multicentre Evaluation of PCA3 and TMPRSS2-ERG Gene Fusions as Diagnostic and Prognostic Urinary Biomarkers for Prostate Cancer. *European Urology*. 2014; 65(3):534–542. [PubMed: 23201468]
16. Auprich M, Chun FKH, Ward JF, Pummer K, Babaian R, Augustin H, Luger F, Gutsch S, Budaus L, Fisch M, Huland H, et al. Critical Assessment of Preoperative Urinary Prostate Cancer Antigen 3 on the Accuracy of Prostate Cancer Staging. *European Urology*. 2011; 59(1):96–105. [PubMed: 20980098]
17. Morgan R, Boxall A, Bhatt A, Bailey M, Hindley R, Langley S, Whitaker HC, Neal DE, Ismail M, Whitaker H, Annels N, et al. Engrailed-2 (EN2): A Tumor Specific Urinary Biomarker for the Early Diagnosis of Prostate Cancer. *Clinical Cancer Research*. 2011; 17(5):1090–1098. [PubMed: 21364037]
18. Schostak M, Schwall GP, Poznanovic S, Groebe K, Mueller M, Messinger D, Miller K, Krause H, Pelzer A, Horninger W, Klocker H, et al. Annexin A3 in Urine: A Highly Specific Noninvasive Marker for Prostate Cancer Early Detection. *Journal of Urology*. 2009; 181(1):343–353. [PubMed: 19012935]

19. Kwiatkowski MK, Recker F, Piironen T, Pettersson K, Otto T, Wernli M, Tscholl R. In prostatism patients the ratio of human glandular kallikrein to free PSA improves the discrimination between prostate cancer and benign hyperplasia within the diagnostic "grey zone" of total PSA 4 to 10 ng/ml. *Urology*. 1998; 52(3):360–365. [PubMed: 9730444]
20. Pickel D, Manucy GP, Walker DB, Hall SB, Walker JC. Evidence for canine olfactory detection of melanoma. *Applied Animal Behaviour Science*. 2004; 89(1-2):107–116.
21. Elliker KR, Sommerville BA, Broom DM, Neal DE, Armstrong S, Williams HC. Key considerations for the experimental training and evaluation of cancer odour detection dogs: lessons learnt from a double-blind, controlled trial of prostate cancer detection. *Bmc Urology*. 2014; 14:22. [PubMed: 24575737]
22. Taverna G, Tidu L, Grizzi F, Torri V, Mandressi A, Sardella P, La Torre G, Cocciolone G, Seveso M, Giusti G, Hurler R, et al. Olfactory System of Highly Trained Dogs Detects Prostate Cancer in Urine Samples. *Journal of Urology*. 2015; 193(4):1382–1387. [PubMed: 25264338]
23. Bjartell AS. Dogs Sniffing Urine: A Future Diagnostic Tool or a Way to Identify New Prostate Cancer Markers? *European Urology*. 2011; 59(2):202–203. [PubMed: 21035944]
24. Shirasu M, Touhara K. The scent of disease: volatile organic compounds of the human body related to disease and disorder. *Journal of Biochemistry*. 2011; 150(3):257–266. [PubMed: 21771869]
25. Roine A, Veskimäe E, Tuokko A, Kumpulainen P, Koskimäki J, Keinänen TA, Hakkinen MR, Vepsäläinen J, Paavonen T, Lekkala J, Lehtimäki T, et al. Detection of Prostate Cancer by an Electronic Nose: A Proof of Principle Study. *Journal of Urology*. 2014; 192(1):230–234. [PubMed: 24582536]
26. Khalid T, Aggio RL, White P, De Lacy Costello B, Persad R, Al-Kateb H, Jones P, Probert CS, Ratcliffe N. Urinary Volatile Organic Compounds for the Detection of Prostate Cancer. *Plos One*. 2015; 10(11):e0143283.doi: 10.1371/journal.pone.0143283 [PubMed: 26599280]
27. Asimakopoulos AD, Del Fabbro D, Miano R, Santonico M, Capuano R, Pennazza G, D'Amico A, Finazzi-Agro E. Prostate cancer diagnosis through electronic nose in the urine headspace setting: a pilot study. *Prostate Cancer and Prostatic Diseases*. 2014; 17(2):206–211. [PubMed: 24686772]
28. Khalid T, White P, Costello BDL, Persad R, Ewen R, Johnson E, Probert CS, Ratcliffe N. A Pilot Study Combining a GC-Sensor Device with a Statistical Model for the Identification of Bladder Cancer from Urine Headspace. *Plos One*. 2013; 8(7)
29. Khalid TY, Costello BDL, Ewen R, White P, Stevens S, Gordon F, Collins P, McCune A, Shenoy A, Shetty S, Ratcliffe NM, et al. Breath volatile analysis from patients diagnosed with harmful drinking, cirrhosis and hepatic encephalopathy: a pilot study. *Metabolomics*. 2013; 9(5):938–948.
30. Van den Velde S, Nevens F, Van Hee P, van Steenberghe D, Quirynen M. GC-MS analysis of breath odor compounds in liver patients. *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences*. 2008; 875(2):344–348.
31. Weber CM, Cauchi Mi, Patel M, Bessant C, Turner C, Britton LE, Willis CM. Evaluation of a gas sensor array and pattern recognition for the identification of bladder cancer from urine headspace. *Analyst*. 2011; 136(2):359–364. [PubMed: 20967397]
32. Filipiak W, Filipiak A, Sponring A, Schmid T, Zelger B, Ager C, Klodzinska E, Denz H, Pizzini A, Lucciarini P, Jamnig H, et al. Comparative analyses of volatile organic compounds (VOCs) from patients, tumors and transformed cell lines for the validation of lung cancer-derived breath markers. *Journal of Breath Research*. 2014; 8(2)
33. Demeestere K, Dewulf J, De Witte B, Van Langenhove H. Sample preparation for the analysis of volatile organic compounds in air and water matrices. *Journal of Chromatography A*. 2007; 1153(1-2):130–144. [PubMed: 17258752]
34. Sethi S, Nanda R, Chakraborty T. Clinical Application of Volatile Organic Compound Analysis for Detecting Infectious Diseases. *Clinical Microbiology Reviews*. 2013; 26(3):462–475. [PubMed: 23824368]
35. Aggio R, Villas-Boas SG, Ruggiero K. Metab: an R package for high-throughput analysis of metabolomics data generated by GC-MS. *Bioinformatics*. 2011; 27(16):2316–2318. [PubMed: 21697128]

36. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. XCMS: Processing mass spectrometry data for metabolite profiling using Nonlinear peak alignment, matching, and identification. *Analytical Chemistry*. 2006; 78(3):779–787. [PubMed: 16448051]
37. Xia J, Wishart DS. Web-based inference of biological patterns, functions and pathways from metabolomic data using MetaboAnalyst. *Nature Protocols*. 2011; 6(6):743–760. [PubMed: 21637195]
38. Peng G, Hakim M, Broza YY, Billan S, Abdah-Bortnyak R, Kuten A, Tisch U, Haick H. Detection of lung, breast, colorectal, and prostate cancers from exhaled breath using a single array of nanosensors. *British Journal of Cancer*. 2010; 103(4):542–551. [PubMed: 20648015]
39. Phillips M, Cataneo RN, Saunders C, Hope P, Schmitt P, Wai J. Volatile biomarkers in the breath of women with breast cancer. *Journal of Breath Research*. 2010; 4(2)
40. Phillips M, Basa-Dalay V, Bothamley G, Cataneo RN, Lam PK, Natividad MPR, Schmitt P, Wai J. Breath biomarkers of active pulmonary tuberculosis. *Tuberculosis*. 2010; 90(2):145–151. [PubMed: 20189456]
41. Probert CSJ, Ahmed I, Khalid T, Johnson E, Smith S, Ratcliffe N. Volatile Organic Compounds as Diagnostic Biomarkers in Gastrointestinal and Liver Diseases. *Journal of Gastrointestinal and Liver Diseases*. 2009; 18(3):337–343. [PubMed: 19795029]
42. Shepherd SF, McGuire ND, de Lacy Costello BPJ, Ewen RJ, Jayasena DH, Vaughan K, Ahmed I, Probert CS, Ratcliffe NM. The use of a gas chromatograph coupled to a metal oxide sensor for rapid assessment of stool samples from irritable bowel syndrome and inflammatory bowel disease patients. *Journal of Breath Research*. 2014; 8(2)
43. Kursa MB, Rudnicki WR. Feature Selection with the Boruta Package. *Journal of Statistical Software*. 2010; 36(11):1–13.
44. Kuhn M. Caret: Classification and Regression Training, R package version 6.0-30. 2014
45. Costello B, Ewen RJ, Jones PRH, Ratcliffe NM, Wat RKM. A study of the catalytic and vapour-sensing properties of zinc oxide and tin dioxide in relation to 1-butanol and dimethyldisulphide. *Sensors and Actuators B-Chemical*. 1999; 61(1-3):199–207.
46. Costello BPD, Ewen RJ, Ratcliffe NM, Sivanand P. Thick film organic vapour sensors based on binary mixtures of metal oxides. *Sensors and Actuators B-Chemical*. 2003; 92(1-2):159–166.
47. Team R.C. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. 2015
48. Lieber CA, Mahadevan-Jansen A. Automated method for subtraction of fluorescence from biological Raman spectra. *Applied Spectroscopy*. 2003; 57(11):1363–1367. [PubMed: 14658149]
49. Ruckstuhl AF, Jacobson MP, Field RW, Dodd JA. Baseline subtraction using robust local regression estimation. *Journal of Quantitative Spectroscopy & Radiative Transfer*. 2001; 68(2): 179–193.
50. Liland KH. baseline: Baseline Correction of Spectra. R package. 2014
51. Nielsen NPV, Carstensen JM, Smedsgaard J. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of Chromatography A*. 1998; 805(1-2):17–35.
52. Du P, Kibbe WA, Lin SM. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics*. 2006; 22(17):2059–2065. [PubMed: 16820428]
53. Serneels S, De Nolf E, Van Espen PJ. Spatial sign preprocessing: A simple way to impart moderate robustness to multivariate estimators. *Journal of Chemical Information and Modeling*. 2006; 46(3): 1402–1409. [PubMed: 16711760]
54. Engelhardt A. Benchmarking feature selection with Boruta and caret. [accessed 24/08/2015]<http://www.cybaea.net/Blogs/Benchmarking-feature-selection-with-Boruta-and-caret.html>
55. Kursa MB, Jankowski A, Rudnicki WR. Boruta - A System for Feature Selection. *Fundamenta Informaticae*. 2010; 101(4):271–286.
56. Ji S, Ye J. Generalized Linear Discriminant Analysis: A Unified Framework and Efficient Model Selection. *IEEE Transactions on Neural Networks*. 2008; 19(10):1768–1782. [PubMed: 18842480]
57. Geladi P, Kowalski BR. PARTIAL LEAST-SQUARES REGRESSION - A TUTORIAL. *Analytica Chimica Acta*. 1986; 185:1–17.

58. Boulesteix A-L, Janitza S, Kruppa J, Koenig IR. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery*. 2012; 2(6):493–507.
59. Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. *American Statistician*. 1992; 46(3):175–185.
60. Vapnik VN. An overview of statistical learning theory. *Ieee Transactions on Neural Networks*. 1999; 10(5):988–999. [PubMed: 18252602]
61. Kuhn M. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*. 2008; 28(5):1–26.
62. Anderssen E, Dyrstad K, Westad F, Martens H. Reducing over-optimism in variable selection by cross-model validation. *Chemometrics and Intelligent Laboratory Systems*. 2006; 84(1-2):69–74.
63. Delen D. Analysis of cancer data: a data mining approach. *Expert Systems*. 2009; 26(1):100–112.
64. Filzmoser P, Liebmann B, Varmuza K. Repeated double cross validation. *Journal of Chemometrics*. 2009; 23(3-4):160–171.
65. Burford, DC.; Kirby, M.; Austoker, J. Prostate Cancer Risk Management Programme: information for primary care. National Health Services; 2009. p. 1-21.
66. National Institute for Health and Care Excellence (NICE). [Accessed 10/7/2015] <http://cks.nice.org.uk/prostate-cancer>
67. Santos F, Guyomarc'h P, Bruzek J. Statistical sex determination from craniometrics: Comparison of linear discriminant analysis, logistic regression, and support vector machines. *Forensic Science International*. 2014; 245:204.e1–204.e8. [PubMed: 25459272]
68. McGuire ND, Ewen RJ, de Lacy Costello B, Garner CE, Probert CSJ, Vaughan K, Ratcliffe NM. Towards point of care testing for C. difficile infection by volatile profiling, using the combination of a short multi-capillary chromatography column with metal oxide sensor detection. *Measurement Science and Technology*. 2014; 25(6)
69. Abd El Gawad IA, Moussa HS, Nasr MI, El Gemae EH, Masooud AM, Ibrahim IK, El Hifnawy NM. Comparative study of NMP-22, telomerase, and BTA in the detection of bladder cancer. *Journal of the Egyptian National Cancer Institute*. 2005; 17(3):193–202. [PubMed: 16799657]
70. Grossman HB, Messing E, Soloway M, Tomera K, Katz G, Berger Y, Shen Y. Detection of bladder cancer using a point-of-care proteomic assay. *Jama-Journal of the American Medical Association*. 2005; 293(7):810–816.
71. Breen V, Kasabov N, Kamat AM, Jacobson E, Suttie JM, O'Sullivan PJ, Kavalieris L, Darling DG. A holistic comparative analysis of diagnostic tests for urothelial carcinoma: a study of Cxbladder Detect, UroVysion (R) FISH, NMP22 (R) and cytology based on imputation of multiple datasets. *Bmc Medical Research Methodology*. 2015; 15
72. Bhuiyan J, Akhter J, O'Kane DJ. Performance characteristics of multiple urinary tumor markers and sample collection techniques in the detection of transitional cell carcinoma of the bladder. *Clinica Chimica Acta*. 2003; 331(1-2):69–77.
73. Halling KC, King W, Sokolova IA, Karnes RJ, Meyer RG, Powell EL, Sebo TJ, Cheville JC, Clayton AC, Krajnik KL, Ebert TA, et al. A comparison of BTA stat, hemoglobin dipstick, telomerase and vysis urovysion assays for the detection of urothelial carcinoma in urine. *Journal of Urology*. 2002; 167(5):2001–2006. [PubMed: 11956427]
74. Krause FS, Rauch A, Schrott KM, Engehausen DG. Clinical decisions for treatment of different staged bladder cancer based on multi-target fluorescence in situ hybridization assays? *World Journal of Urology*. 2006; 24(4):418–422. [PubMed: 16688457]
75. Toma MI, Friedrich MG, Hautmann SH, Jakel KT, Erbersdobler A, Hellstern A, Huland H. Comparison of the ImmunoCyt test and urinary cytology with other urine tests in the detection and surveillance of bladder cancer. *World Journal of Urology*. 2004; 22(2):145–149. [PubMed: 14991320]
76. Dimashkieh H, Wolff DJ, Smith TM, Houser PM, Nietert PJ, Yang J. Evaluation of UroVysion and Cytology for Bladder Cancer Detection. *Cancer Cytopathology*. 2013; 121(10):591–597. [PubMed: 23801650]

Figure 1.
Pipeline for data analysis.

Figure 2.
Chromatograms showing baseline correction using a 2 step procedure

Figure 3.
Schematic for selecting the reference sample for chromatogram alignment.

Figure 4.

Validation schemes used to validate the classifiers built by the pipeline presented here: (a) repeated 10-fold cross-validation; (b) repeated double cross-validation; (c) Repeated 10-fold cross-validation Monte Carlo; and (d) repeated double cross-validation Monte Carlo.

Figure 5. Boxplot (a) and two-factor principal component analysis (b) of data derived from patients with prostate cancer and controls

Figure 6. ROC curves of repeated double cross-validation applied to Prostate and Control urine sample data processed by the GC-sensor pipeline and modelled by linear discriminant analysis (LDA) and support vector machine with polynomial kernel (SVM-P).

Figure 7.

Sensitivity and specificity data of biomarkers for prostate cancer. Grey rows = biomarkers found in the literature; Orange rows = GC-sensor pipeline; VOC = volatile organic compound; Validation = the statistical method used to validate sample classification; LOOCV = leave-one-out cross-validation; 10-FoldCV = 30 times repeated 10-fold cross-validation; DoubleCV = 30 time repeated double cross-validation (outer loop = 3 folds; inner loop = 10 fold cross-validation repeated 5 times); PSA = prostate-specific antigen; p2PSA = serum isoform [-2] proPSA; PHI = prostate healthy index; imPSA = intracellular macrophage PSA; PCA3 = prostate cancer antigen 3; Not validated = no statistical method was applied to validate results of sample classification.

Figure 8. Boxplot (a) and two-factor principal component analysis (b) of data derived from patients with bladder cancer and controls

Figure 9. ROC curves of repeated double cross-validation applied to Bladder and Control urine sample data processed by the GC-sensor pipeline and modelled by linear discriminant analysis (LDA) and support vector machine with polynomial kernel (SVM-P).

Figure 10.

Sensitivity and specificity data of biomarkers for bladder cancer. Grey rows = biomarkers found in the literature; Orange rows = framework presented here; BTA = bladder tumor antigen; NMP-22 = nuclear matrix protein 22; Hb dipstick = hemoglobin dipstick; Cxbladder = Cxbladder Detect; VOC = volatile organic compound; Validation = the statistical method used to validate sample classification; LOOCV = leave-one-out cross-validation; 10-FoldCV= 30 times repeated 10-fold cross-validation; DoubleCV = 30 time repeats of 3-fold double cross-validation with an inner loop of 10-fold cross-validation repeated 5 times; Not specified = a statistical validation method for sample classification was not found in source.

Table 1

Demographics for study participants – prostate cancer.

Diagnosis	Patients	Age range in years (median)	Smoking status
Prostate cancer	58	50 – 88 (69)	10 (17%)
Control	73	29 – 86 (64)	15 (21%)

Table 2

Gleason score of patients with prostate cancer.

Number of patients	Gleason score
4	< 6
30	6
20	7
4	> 7

Table 3

Classification results from leave-one-out cross-validation on urine samples from patients with prostate cancer and patients with haematuria and or prostatic symptoms. LDA = linear discriminant analysis; and SVM-P = support vector machine.

Linear discriminant analysis (LDA)				
	Prostate cancer	Control		
Prostate cancer	54	4	Sensitivity =	93.1%
Control	3	70	Specificity =	95.9%
Support vector machine polynomial (SVM-P)				
	Prostate cancer	Control		
Prostate cancer	55	3	Sensitivity =	94.8%
Control	3	70	Specificity =	95.9%

Table 4

Classification results from repeated 10-fold cross-validation and repeated double cross-validation of urine samples from patients with prostate cancer and patients with haematuria and or prostatic symptoms. LDA = linear discriminant analysis; and SVM-P = support vector machine; and CI = confidence interval of mean values.

Repeated 10-Fold cross-validation									
Classifier	Accuracy			Sensitivity			Specificity		
	Mean	Median	CI	Mean	Median	CI	Mean	Median	CI
LDA	88.4	91.7	[85.7-90.9]	87.1	83.3	[85.9-88.4]	89.4	87.5	[85.8-92.6]
SVM-P	90.8	92.3	[88.4-93.1]	88.3	83.3	[83.9-92.4]	93.0	100.0	[89.9-95.7]
Repeated double cross-validation									
Classifier	Accuracy			Sensitivity			Specificity		
	Mean	Median	CI	Mean	Median	CI	Mean	Median	CI
LDA	87.7	88.6	[86.1-89.3]	85.4	85.0	[82.9-87.8]	89.6	91.7	[87.6-91.6]
SVM-P	88.8	88.6	[87.4-90.1]	85.5	85.0	[82.8-88.1]	91.4	91.7	[90.0-93.3]

Table 5

Performance of PSA testing in this cohort of patients with prostate cancer and with urological symptoms.

	Age-specific PSA (%)	PSA >6 (%)
Accuracy	56	64
Sensitivity	100	84
Specificity	9	43
PPV	53	61
NPV	100	72

Table 6

Demographics for study participants – bladder cancer.

Diagnosis	Patients	Age range in years (median)	Smoking status
Bladder cancer	24	27 – 91 (71)	7 (29%)
Control	73	29 – 86 (64)	15 (21%)

Table 7

Classification results from leave-one-out cross-validation on urine samples from patients with bladder cancer and patients with haematuria and or prostatic symptoms. LDA = linear discriminant analysis; and SVM-P = support vector machine.

Linear discriminant analysis (LDA)				
	Bladder cancer	Control		
Bladder cancer	23	1	Sensitivity =	95.8%
Control	1	72	Specificity =	98.6%
Support vector machine polynomial (SVM-P)				
	Bladder cancer	Control		
Bladder cancer	23	1	Sensitivity =	95.8%
Control	0	73	Specificity =	100.0%

Table 8

Classification results from repeated 10-fold cross-validation and repeated double cross-validation on urine samples from patients with bladder cancer and patients with haematuria and or prostatic symptoms. LDA = linear discriminant analysis; and SVM-P = support vector machine; and CI = confidence interval of mean values.

Repeated 10-Fold cross-validation									
Classifier	Accuracy			Sensitivity			Specificity		
	Mean	Median	CI	Mean	Median	CI	Mean	Median	CI
LDA	96.4	100.0	[94.2-98.3]	92.1	100.0	[90.4-93.6]	97.8	100.0	[95.5-99.5]
SVM-P	97.1	100.0	[95.5-98.8]	88.6	100.0	[82.0-94.7]	100.0	100.0	[99.4-100.0]
Repeated double cross-validation									
Classifier	Accuracy			Sensitivity			Specificity		
	Mean	Median	CI	Mean	Median	CI	Mean	Median	CI
LDA	93.6	93.8	[92.1-95.1]	87.4	87.5	[84.0-90.5]	95.7	95.8	[94.1-97.1]
SVM-P	96.2	96.9	[95.3-97.2]	87.2	87.5	[83.5-91.0]	99.2	100.0	[98.5-99.8]