Using Machine Learning and Urine Cytology for Bladder Cancer Prescreening and Patient Stratification

William Tun, Ognjen Arandjelović, Peter David Caie
University of St Andrews
St Andrews KY16 9SX
Scotland, United Kingdom

Abstract

Bladder cancer presents a major and increasing health risk worldwide with early diagnosis being recognized as key to its successful management and treatment. This is the first work to propose the use of data extracted from immunofluorescence images together with readily available data collected from patients referred to cystoscopy as a means of stratification and in particular for the reduction in the number of unnecessary further cystoscopies. We present a thorough analysis of the problem and thus a carefully designed machine learning based solution that achieves this goal. Using a real-world data corpus and thorough statistical analysis, we demonstrate that our method is capable of distinguishing patients at high risk of having bladder cancer with an error rate of less than 5%, who can be prioritized for follow-up examination and treatment, while at the same time reducing unnecessary financial and resource burden, as well as substantial patient discomfort, by correctly identifying 66% of low risk patients with cystoscopy associated morbidities.

Introduction

Bladder cancer is one of the most common cancers worldwide (Ploeg, Aben, and Kiemeneij 2009). The scale of the problem is illustrated well by noting that in 2015 alone approximately 430,000 new cases were diagnosed and 188,000 deaths recorded globally. In addition to the human cost, that is, the effect that the disease has on the individuals directly and indirectly affected (including patients on surveillance who due to the high reoccurrence rate of the disease often undergo unnecessary cytoscopies), the financial burden is major. In the United Kingdom alone, the estimated annual cost imposed upon the National Health Service (NHS) for the treatment and care of patients with bladder cancer is estimated to be over 50 million GBP.

Early diagnosis of bladder cancer is of utmost importance for a successful outcome (Oosterlinck et al. 2002). Currently, cystoscopy (see Figure 1) is universally regarded as the gold standard for the diagnosis and monitoring of the disease (Wiener et al. 1998). Unfortunately the procedure is expensive, uncomfortable to the patient, not always accurate, and comes with the risk of causing infection. Furthermore, there is a significant risk of bladder cancer reoccurrence (35% within the first 3 years), thus creating the requirement for life-long surveillance. The idea explored in the present paper is that of using data already collected by clinics during the course of treatment of patients, as a means of decreasing the cost of treatment (by eliminating unnecessary diagnostic procedures and clinical staff time) while improving patient outcomes by increasing the speed at which high risk cases are identified.

Background and context

In this section we present the relevant aetiological, medical, and clinical background needed to understand the motivation behind our work, and the context of the problem addressed by it.

Aetiology

Cancer is a complex family of over a hundred diseases which share six major hallmarks:
- Abnormal growth of cells due to deregulation of growth signals...
The developmental path of normal cells becoming a tumor and eventually a cancer is called malignant progression. Though not all tumors are cancerous, all cancers arise from tumors. For males and females respectively, bladder cancer is ranked respectively as the fourth and the ninth most common form of cancer. Approximately 66% of bladder cancers are diagnosed in the elderly population above the age of 65 and the disease is estimated to be 3 times more prevalent in men than in women. Genetic studies have identified a few key genes which when deregulated or mutated are associated with the emergence of bladder cancer, e.g. FGR3, RB1, HRAs, TP53 and SC1. Many of these genes are involved in cellular processes such as cell growth, proliferation, division, apoptosis, and angiogenesis. Environmental carcinogens are another key concern. Indeed, smoking and alcohol intake are two major contributing factors to the development of bladder cancer. Toxins contained in the corresponding diet can be safely removed surgically before it has invaded the surrounding tissue, it can be removed surgically without the need for excising the entire bladder, and the patient can be considered cured. Thus the prognosis for excision of cancer at an early stage is positive. Once malignant, the cancer cells can escape the local tissue and enter either the circulatory blood or lymphatic systems. The risk is that the cancer escapes the vasculature and forms a secondary tumour in a distant organ, which can prove lethal. The current prognosis for such late stage cancers is very poor. Thus it is of utmost importance to achieve early diagnosis in order to maximize the likelihood of a positive prognosis. If caught early, the developing potentially cancerous tumor can simply be removed surgically before the development of malignancy.

**Clinical context**

A key symptom of bladder cancer is hematuria i.e. the presence of blood in urine. Gross hematuria refers to when the blood can be seen with the human eye whilst microhematuria refers to when the blood can only be detected with sensitive medical equipment. Unfortunately, urine sample tests for haematuria has low sensitivity due to the intermittent nature of the phenomenon (Shirodkar and Lokeshwar 2008). Furthermore many other diseases, particularly urological diseases (e.g. urolithiasis or severe urinary tract infections) can lead to a high false positive error rate. Consequently, urine sample tests for hematuria alone are not sufficient for detecting bladder cancer. However, blood in urine does permit a degree of stratification, triggering patient referral to a hospital where more in-depth checks can be performed. For the last 80 years, cystoscopy has been the gold standard in detecting bladder cancer (van der Aa et al. 2010). This procedure involves inserting a thin tube with a camera and a light at the end, up through the urethra and into the bladder in order to detect tumors by visual means. If a tumor is found, it can be removed with a surgical procedure known as transurethral resection of the bladder (TURB) (Crawford 2008). Unfortunately, even after cystoscopic resection, there is a high recurrence rate, warranting further surgery. Bladder cancer itself is estimated to have a 35% reoccurrence rate within the first 3 years, thus requiring life-long surveillance. This is burdensome for the patients and NHS on many levels. On the side of the patients, repeated rounds of cystoscopy become inconvenient, uncomfortable, and due to its invasive nature, infection is risked per cystoscopic procedure. Since most bladder cancer patients are elderly, they are more likely to have trouble travelling to hospitals. On top of this, the elderly are more susceptible to developing urinary infections compared to the rest of the population. For the NHS, repeated rounds of cystoscopy on healthy patients believed to potentially have bladder cancer incur a burden on NHS finances, resources and staff time. Thus, there is a demand to have a system that can detect bladder cancer in a way that is cost effective, non-invasive and accurate.

**Premise**

Universal, non-targeted testing of the population for bladder cancer at large is impractical. This is clearly evident even if the financial argument alone is considered. Moreover such an approach would incur unreasonable distress and potential physical harm to a disproportionately large number of individuals which are not at risk of the disease. Indeed, our primary goal is to reduce the number of unnecessary cystoscopies. The proposed idea is to use features extracted from immunofluorescence images together with data already collected by clinics, to this end. In addition to the aforesaid primary aim, we also achieve a decrease in the cost of treatment (by eliminating unnecessary diagnostic procedures and clinical staff time) while improving high risk patient outcomes by early identification. In particular, we stratify patients into three strata: (i) bladder cancer highly probable, (ii) bladder cancer highly unlikely, and (iii) in need of specialist follow-up (bladder cancer presence not establishable with a sufficiently high confidence).
Technical details

In this section we describe the key contributions of the present work, both in terms of clinical and data acquisition methodology, as well as the proposed machine learning approach.

Experimental data

Patient cohort In the present work we used a real-world data of 250 patients. Out of these, 222 have not had a diagnosis of bladder cancer and 28 did. All data was collected according to the standard ethical guidance, with informed consent, and approved for research use.

Input features The types of input features we adopted in this work can be categorized into three groups by their type. The first of these comprise general patient data readily available from their medical records, such as sex and age, etc. The second group can be described as coarse urinalysis information. Specifically:

- **Blood in urine** As mentioned earlier, hematuria is a major indicator of potential presence of cancer. We used a categorical feature, with three possible values: (No) no detectable blood, (MH) microhematuria, blood is present but is only detectable using a microscope, and (GH) gross hematuria, blood can be detected with the naked eye.
- **Clinic type** The reason why a patient was sent for urinalysis also provides important information, in that it captures the relevant medical history. This information too was captured as a categorical variable with three possible values: (CS) patient is on cystoscopic surveillance, (GH) no prior cancer, urinalysis performed due to symptoms such as hematuria, (NC) cystoscopy performed for other reasons.

The last feature group, greatest in feature count, includes a range of relevant information extracted from medical imaging. In particular, the raw data comprises cytological slides generated via immunofluorescence, a technique which uses antibodies that target specific epitopes in cells and allows these targets to be visualized via the incorporation of specific fluorescent fluorophore dyes attached to the antibody. For the experiments described in the present paper, two antibodies were used in the acquisition of immunofluorescence images:

- **Monoclonal Anti-MCM2**: a marker for proliferation. This antibody attaches to proteins on the nucleus. Since cells of the healthy urothelial lining are fully differentiated, Anti-MCM2 can help detect abnormal proliferation which are correlated with the development of cancer. The problem is if this antibody is used alone to detect cancer, it flags many false positives since other cells can be dividing at a rapid rate including healthy regenerating bladder lining after injury caused by e.g. stones or stent insertion.
- **Polyclonal PanCytokeratin (PanCK)**: complements Anti-MCM2 by reducing the false positive problem by being an additional marker for urothelial cells. Thus PanCK can specifically differentiate rapidly dividing urothelial cells from other highly proliferative cells such as those of the immune system. The stain lights up the membrane but can also at times light up the cytoplasm.

The acquired immunofluorescence images (see Figure 2 for an example of a composite image showing both anti-MCM2 and PanCK channels) were used to extract a number of salient sample characteristics, adding information to other readily available clinical data. The signal to noise ratio in the raw data is extremely high, leading to simple follow-up image analysis and feature extraction. In particular, by applying the well-known watershed algorithm (Lin et al. 2006) and segmentation based on region connectedness and ‘blobness’ (see Figure 3), we extracted a range of cytological feature statistics (mean and standard deviation across the slide) for MCM2 expressing and PanCK expressing cell, such as cell eccentricity and circumference, nuclear size and texture (variation in the intensity of staining), etc.

In total, 127 features were used. This high number of features, which in principle having the potential to provide more discriminative information, should also be noted as presenting a challenge in the context of ensuring statistical robustness and classifier stability, as we will elaborate on in some detail in the next section.

Algorithmic approach

In this section we summarize the key aspects of the problem at hand which guided our design of a suitable machine learning methodology, and explain the ultimate technical decisions used for the subsequent evaluation. However, before we proceed, we wish to start with an important general note on our methodological approach.

Recent decades have seen a major progress in machine learning and the development of a broad range of different
algorithmic approaches, each with its specific practical and theoretical advantages and disadvantages. This is clearly a most welcome development and as a consequence our understanding of the limitations, strengths, and challenges of different method has increased dramatically. Nevertheless there is a potential danger associated with these developments, which often gets overlooked. In particular, given the spectrum of approaches which have all proven useful in certain problem areas, the risk is that of blindly applying off-the-shelf approaches and by chance rather than by its fundamental appropriateness, obtaining good experimental results. Being keenly aware of this danger, we took particular care not to fall into this trap. Instead, as we will shortly explain, we analysed the particular aspects of the problem at hand and let these guide our design choices. Hence, rather than comparing a series of different method, through analysis and informed decision-making (driven both by clinical and computer science domain knowledge), arrived at a single hypothesis which we finally evaluated.

Dealing with unbalanced data As noted earlier, a major challenge which presents itself in the context of the problem considered in this work is that of high class imbalance (approximately 1:9 as elaborated in the next section). In particular, given the seriousness of consequences associated with false negative diagnosis (i.e. dismissing a patient with cancer with a diagnosis suggesting the absence of cancer), the conservative approach understandably adopted by health care professionals is to err on the cautious side and screen many individuals who in fact do not have cancer. This leads to the available data containing far fewer cancerous cases than non-cancerous ones.

To address the aforementioned class imbalance in this work we adopt the well-known Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al. 2002). SMOTE is widely used due to its simplicity on the one hand, and its good performance on the other. The approach combines an over-sampling of the minority class (herein, cases positive for cancer) and an under-sampling of the majority class (herein, cases negative for cancer) [19]. Unlike many other methods which perform over-sampling with replacement, SMOTE uses bootstrap sampling followed by \(k\)-nearest neighbour based generation of synthetic observations, see Figure 4. The key advantage of this approach is that the chance of duplication is reduced, resulting in a blurring of the class decision boundary, which improves generalization performance.

Dealing with feature heterogeneity Another technically interesting and challenging aspect of the problem at the heart of the present work concerns the heterogeneity of feature types and distributions used. As we already explained in the previous section, a number of our input features are inherently of a categorical nature. Others are continuous but exhibit vastly different statistical properties both in terms of the underlying probability density functions, as well as the intervals over which they are observed. Attempting to impose probabilistic models into these is problematic for two key
reasons. Firstly, there is little in terms of available evidence which would guide such probabilistic modelling. Moreover, this approach is rather impractical because of the number of features used (over 100). These considerations guided our choice of the classification methodology. In particular, it led us away from otherwise successful classifiers such as those based on polynomial or logistic regression, or support vector machines (Burges 1998), and towards tree-based approaches which have the advantage of dealing with features independently at the level of “local” decisions. In particular, we decided on the use of a random forest (Breiman 2001; Karsten and Arandjelović 2017; Birkett, Arandjelović, and Humphris 2017), with a further robustness added in the form of gradient boosting (Mason et al. 2000).

**Experimental analysis**

For our evaluation we adopted the usual framework for assessing the performance of supervised learners whereby training and testing were performed on non-overlapping subsets of the available data. This process was guided by two types of considerations. The first of these concerns the robustness of statistical results. In particular, in order to facilitate a deeper understanding of the stability of the overall classifier performance as well as the stability of features, we adopted the $n$-fold cross validation approach. The second type of considerations was of a clinical nature. In particular, it is important to observe the asymmetry between the false positive and false negative error rates, that is, the errors caused by incorrectly inferring the presence of cancer vs. missing out on the detection of cancer. Clearly, the consequences of the latter are far graver, with a delayed (or unperformed) treatment leading to seriously worse patient outcomes and possibly death. To account for this during the training of our classifier we also learn the weighting of the two types of errors, as guided by clinical need. In particular, we ensure that the false negative rate is below 0.05 (i.e. at most 1 in 20 cases of cancers is missed).

Following the methodology summarized above and the approach described in the previous section, we obtained the following results. Firstly, as desired the average (of $n$-fold cross validation) false negative error rate was below 0.05 (5%). The corresponding false positive error rate was 0.34 (34%). Considering the nature of the training data, that is, the non-specificity of the test, these are remarkable results. In particular, they show that patients at high risk can be accurately and promptly prioritized for further treatment while at the same time being able readily to discount 66% of the patients at low risk, thus dramatically reducing the associated cost of unnecessary follow-up examinations, clinical staff time, and undue distress. A graphical summary is provided in Figure 5. Across different values of $n$ in $n$-fold cross validation, we uniformly found the standard deviation of the aforementioned key statistics to be low. This finding demonstrates that the obtained results are indeed based on real, relevant information, rather than being caused by chance, and that the performance is stable regardless of the specific cohort used in training. It should be noted that, as expected from theoretical principles, the standard deviation rises as the number of folds in $n$-fold cross validation is reduced – this is a consequence of increased overlap (correlation) between training sets for different runs for small values of $n$. Nevertheless, the deviation remains extremely low ($< 2.5\%$) across all values of $n$, as shown in Figure 6.

In order to facilitate an additional understanding of the decision process learnt by our method, as well as provide additional reassurance as regards the soundness of the process, we also examined in detail the role played by individual features. The contributions of different features to the model (the so-called gain), computed by aggregating the contributions across individual trees in the trained forest, is shown in Figure 7 – for clarity, only the top 15 features are shown and their importance re-scaled relative to the top ranked feature. It can be readily observed that these features dominate. The first of these indicates a previous case of bladder cancer. This is highly reassuring given the ample evidence that bladder cancer is characterized by a high reoccurrence rate (Mitrakas et al. 2015). As noted earlier, the importance of the presence of blood in urine is another already widely recognized indicator of potential bladder cancer. Hence in this respect too, the structure of the decision process learnt by our algorithm is consistent with clinical evidence. Lastly, the feature described by the string MCM2panCK.mean.std.panCK characterizes the heterogeneity of panCK immunofluorescence appearance of nuclei associated both with cells which express both MCM2 and panCK biomarkers. Such heterogeneity, combined with the expression itself is indicative of high and abnormality proliferating cells, again reaffirming the soundness of the learnt decision making process.

**Summary and conclusions**

Bladder cancer presents a major and increasing health challenge. With some 430,000 new cases diagnosed annually and 188,000 deaths, the human cost alone is worrying. Moreover, the financial cost (both direct and indirect in terms of staff time) associated with treatment and care of patients is significant. In this paper we tackled both sides of the problem by proposing to use features extracted from immunofluorescence images together with other readily available data collected from patients referred to cystoscopy. We presented a thorough analysis of different challenges posed by clinical, practical, and technical requirements, and thus described a machine learning algorithm that performs patient stratification. On a real-world data corpus we demonstrated that our method is able to distinguish patients at high risk with an error rate of less than 5%, who can be prioritized for follow-up examination and treatment, while at the same time reducing unnecessary financial and resource burden by correctly identifying 66% of low risk patients. We hope that the present findings will contribute to a broadening range of emerging systems aimed at supporting clinicians and patients (Li and Arandjelović 2017; Vasiljeva and Arandjelović 2017; Barracliffe, Arandjelović, and Humphris 2017).

**References**

Figure 5: A graphical summary of key results of our \( n \)-fold cross validation. Shown are the mean and standard deviation values for the two clinically relevant variables: the false positive and the false negative rates. The performance can be readily seen to be extremely impressive, all the more considering the nature of the training data, that is, the non-specificity of the test. In particular, our findings show that patients at high risk can be accurately and promptly prioritized for further treatment while at the same time being able readily to discount most of the patients at low risk, thus dramatically reducing the associated cost of unnecessary follow-up examinations, clinical staff time, and undue distress.

Figure 6: Dependence of the variance in the false positive rate (vertical axis) as a function of \( n \), the number of folds in our \( n \)-fold cross validation. As expected from theoretical principles, the standard deviation rises as the number of folds in \( n \)-fold cross validation is reduced, which is a consequence of increased overlap (correlation) between training sets for different runs for small values of \( n \). Nevertheless, as the plot readily shows, the deviation remains extremely low (< 2.5%) across all values of \( n \).


Figure 7: The importance of the 15 most significant features: the contributions of different features to the model (the so-called gain) was computed by aggregating the contributions across individual trees in the trained forest. The gain is shown relative to the gain of the feature deemed most important by the metric.


