A novel approach for accurate prediction of spontaneous passage of ureteral stones: Support vector machines

F Dal Moro[1,3], A Abate[2,3], GRG Lanckriet[2,3], G Arandjelovic[1], P Gasparella[1], P Bassi[1], M Mancini[1] and F Pagano[1]

[1]*Department of Urology, University of Padova, Padova, Italy and* [2]*Electrical Engineering and Computer Sciences Department, University of California at Berkeley, Berkeley, California, USA*

The objective of this study was to optimally predict the spontaneous passage of ureteral stones in patients with renal colic by applying for the first time support vector machines (SVM), an instance of kernel methods, for classification. After reviewing the results found in the literature, we compared the performances obtained with logistic regression (LR) and accurately trained artificial neural networks (ANN) to those obtained with SVM, that is, the standard SVM, and the linear programming SVM (LP-SVM); the latter techniques show an improved performance. Moreover, we rank the prediction factors according to their importance using Fisher scores and the LP-SVM feature weights. A data set of 1163 patients affected by renal colic has been analyzed and restricted to single out a statistically coherent subset of 402 patients. Nine clinical factors are used as inputs for the classification algorithms, to predict one binary output. The algorithms are cross-validated by training and testing on randomly selected train- and test-set partitions of the data and reporting the average performance on the test sets. The SVM-based approaches obtained a sensitivity of 84.5% and a specificity of 86.9%. The feature ranking based on LP-SVM gives the highest importance to stone size, stone position and symptom duration before check-up. We propose a statistically correct way of employing LR, ANN and SVM for the prediction of spontaneous passage of ureteral stones in patients with renal colic. SVM outperformed ANN, as well as LR. This study will soon be translated into a practical software toolbox for actual clinical usage.

Kidney International (2006) **69,** 157–160. doi:10.1038/sj.ki.5000010

KEYWORDS: urolithiasis; ureteral calculi; support vector machine; artificial intelligence; statistical methods; neural networks

Correspondence: *FD Moro, Department of Urology, University of Padova Medical School, Via Giustiniani, 2, Padova I-35128, Italy. E-mail: fabrizio.dalmoro@unipd.it*

[3]These authors contributed equally to this work.

Received 7 November 2004; revised 16 May 2005; accepted 8 July 2005

Referring to the statistics on the incidence of kidney stone disease in industrialized countries, we understand how important it is to correctly analyze this pathology in order to predict accurately which patients need what sort of intervention. Everybody agrees that considering the stone size is the most important factor for predicting the spontaneous passage of calculi.[1,2] However, this does not seem discriminative enough when calculi are of mid-size dimensions. At this stage, the urologist needs more information in order to take a valid clinical decision, but there is no demonstrated result as to which factor should be considered first and what are the actual interactions between all the factors.[3]

In literature, the statistical methodologies employed have been the multivariate logistic regression (LR)[4] and the artificial neural network (ANN).[5] In this work, we propose to use the recently developed support vector machines (SVM),[6–8] an instance of kernel methods, for classification, as well as linear programming SVM (LP-SVM);[9] these are in general believed to outperform the ANN.[8,10]

The paper will unfold as follows: along with a critical analysis of the results presented in medical literature – with a special focus on the ANN – we describe how the statistical tests are performed. Critical results follow, and a discussion on their significance, both technically and clinically, is developed. Conclusions mark the state of the art of our work, and define some future directions of our research.

RESULTS

Figure 1 plots the achievable true positive (TP) rate (i.e., sensitivity) versus true negative rate (TN) (i.e., specificity) for the different learning algorithms. Each of the four plots corresponds to a different learning algorithm. Each dot within a plot corresponds to the average test-set performance obtained for a certain setting of the algorithm's 'hyper-parameters', that is, parameters that are *a priori* chosen and are endogenous to the actual training procedure. The choice of SVM and LP-SVM reflects the relative importance the training algorithm should give to false positives versus false negatives. For the ANN and LR, these parameters are, respectively, related to the actual structure of the network or

to more technical training issues (weights and thresholds, for instance).

The best results in prediction accuracy were singled out, picking up a point at the upper-right-most part of each of the four plots (see arrows); using the old method of multivariate LR, the outcome showed 90.3% sensitivity and 69.7% specificity (Figure 1a). The ANN matched this performance with 94.9% sensitivity and 62.9% specificity (Figure 1b). When using an SVM, 84.5% sensitivity and 86.9% specificity could be obtained (Figure 1d). LP-SVM presented results that were on the upper rim of the SVM performance (Figure 1c). Again, it was possible to associate to each and every point of this plot a single combination of all the hyper-parameters of the respective algorithm.

With respect to our second objective, ranking the input factors, Table 1 shows the ranking obtained using Fisher scores and LP-SVM weights, respectively. As both ranking approaches are essentially different, we should not necessarily expect the rankings to be similar. However, when inspecting the results, we saw a rather high overlap within the top five values of both rankings (the factor identified as most significant being the same and three factors from the top five overlapping in both results). This certainly advocates the robustness and significance of the obtained outcomes. Moreover, these rankings were validated by simulations using only the more relevant inputs. More precisely, we set up and ran the training/testing procedure first on the most prominent input, then on the two most influential inputs and finally on the first five in the ranking. For both rankings, similar results were obtained. Using stone size only led to acceptable results. Using more inputs increased the performance, whereas using just the five most important inputs was qualitatively equivalent to the results obtained using all inputs. Therefore, we concluded that the remaining four clinical factors introduce spurious information and, in this specific setting, can be regarded as redundant.

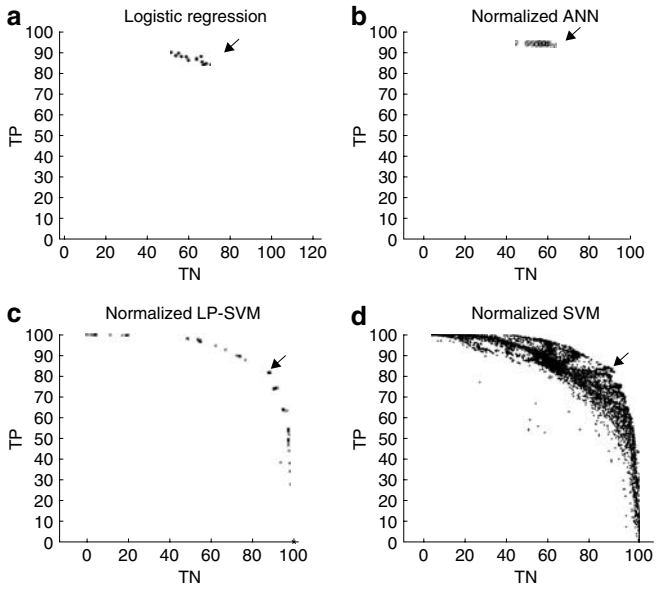

Figure 1 | Comparison of the average test-set performances for the four learning algorithms run on normalized data. (a–d) The axes represent specificity and sensitivity. Each dot within a plot corresponds to the average test-set performance obtained for a certain setting of algorithm hyper-parameters that are endogenous to the actual training procedure. As stated in the literature, ANN slightly improves the results obtained through LR, while the kernel algorithms outperform the other two methods.

DISCUSSION

Let us first list and highlight the main pitfalls of the results presented in literature, which have mostly been obtained with the aid of ANN.[11] First of all, the used data sets are often of relatively low cardinality, a condition that is more likely to provide poor results or unstable prediction algorithms.[12,13]

Second, the ANN results in literature are based on using only one hold-out test set and hence so the reported performance depends heavily on the particular test set that is used. Therefore, training and testing should be performed more than once and the test-set performances averaged out, to reduce the variance of the performance estimate. Whereas most literature ignores this fact, we applied cross-validation and averaged the performance over 30 randomly chosen test sets, as mentioned before. Therefore, we performed statistically more accurate tests for all our learning algorithms, including ANN.

Third, we strongly question the ANN results concerning the input rankings: it is known that networks with a structure that is more complex than that of a perceptron (i.e., with one or more hidden layers), offer no clear connection between their weights and the relative relevance of their inputs.[14,5] Also, it is wrong to look at the absolute values of the weights of even a perceptron when the inputs are not normalized.[12,15] We resolve the pitfall of ANN not allowing the determination of the relative importance of the clinical factors by using Fisher scores and the LP-SVM approach.

Table 1 | Classes of importance of the spontaneous stone expulsion factors second to different methods

Classes of importance	Fischer score	LP-SVM
I	Stone size	Stone size
	Symptoms' duration before check-up	Stone position
	Previous urological treatment	Symptoms' duration before check-up
II	Sex	Previous stone expulsion
	Stone position	Age
	Fever	Fever
III	Previous stone expulsion	Sex
	Body mass index	Body mass index
	Age	Previous urological treatment

As for the particular strength of the SVM approach, we first pointed out the broad range of performances that could be achieved in the specificity/sensitivity plane (Figure 1), by varying the SVM hyper-parameter settings. This gave rise to a curve that was similar to a receiver operating characteristic curve, although more specialized. A usual receiver operating characteristic curve would be obtained from one set of classifier weights, using the known testing-threshold shift. In this case, each dot corresponds to a different set of classifier weights, obtained from SVM training for a specific hyper-parameter setting. These plots show how flexible the SVM is in terms of specificity/sensitivity trade-off. The ANN and LR offer a lot less flexibility.

In the case of ANN, we varied several training parameters, resulting in only a small variation in TP and TN rates, although enough to still improve on the prediction accuracy of the LR (Figure 1).

The points referring to the SVM, being widely spread through the TP/TN plot, show how this method can be more descriptive than the former two.

The SVM prediction improves LR and ANN significantly along the specificity axis. This, important from a statistical standpoint, also has a sharp clinical meaning: a wrong prediction in terms of specificity would result in the patient missing an invasive intervention, which would effectively be needed. Thus, it is clear that the best prediction of spontaneous stone passage will be one that combines an outstanding sensitivity with a remarkable specificity. The SVM approach offers a great variety of predictive sensitivity/specificity combinations, depending on the setting of its hyper-parameters. If we consider a possible optimal operation point (corresponding to a specific hyper-parameter setting), that is, 84.5% sensitivity and 86.9% specificity, the SVM approach shows significantly better results than those obtained with LR and ANN.

Focusing on the problem of input ranking, we notice how the results obtained with the Fisher scores make sense from a clinical point of view. In earlier work, the ranking, computed with ANN, gave questionable results.[12,15] The classification obtained with LP-SVM was similar to the first. We compared the results obtained by those two methods by splitting the spontaneous passage factors in three groups of decreasing importance according to the weights we obtained, so that we could ponder over their clinical value.

Simulations with an increasing number of input features improved until the 'heaviest' five inputs were used, the latter leading to results equivalent to those obtained when using all input factors. This means that the last four inputs do not add any further information to the prediction problem and can be regarded as 'redundant'.

The hydration and the medical therapy can increase the rate of spontaneous stone passage, but they were not taken into consideration as parameters. That is because it is common praxis, when hydration is considered, to advise each patient with renal colic a minimum 2–3 l of water intake a day. Patients who underwent treatment with Ca antagonists, cortisone or alpha-blocker agents (i.e., Tamsulosin), prior to and/or after the colic episode, were already excluded before: in fact, the efficacy of these treatments has been proven by several studies.[16,17]

The fact that the stone size is by and large the most influential factor explains why the LR (linear) results are not too far from those obtained with the (nonlinear) ANN. Nevertheless, the SVM approach is still able to infer deeper relationships between inputs and outputs, resulting in a better performance, and therefore represents the method of choice in tackling this problem.

CONCLUSIONS

This work proposes the application of the SVM to drastically improve the prediction results for intervention on renal colic obtained in the literature. The new results, which outclass those obtained via LR and the ANN approach, are particularly interesting from a clinical perspective, as they maintain the ANN level of sensitivity (i.e., correctly predicting that no intervention is needed) while improving significantly on the specificity (i.e., correctly predicting the need for an intervention). The authors are willing to translate these algorithms into a software toolbox, which would then help physicians on their fieldwork. This is the first time an instance of kernel methods, that is, the SVM, has been applied with success to such clinical data. Intelligent systems such as this could markedly reduce costs of therapeutical approaches and recoveries for kidney stone disease. Given the outstanding performance of SVMs, their application in other fields of urology, such as the oncological field, is imminent.

MATERIALS AND METHODS

We gathered and sorted the information collected from 1163 patients who were treated for an episode of renal colic in the period from January to December 2003 in the Urology Institute of the Hospital of Padova, Italy. A focused selection of the patients was made on the basis of some important criteria. The patients excluded were as follows:

- patients in whom the colic episode was due to renal calculi;
- patients in whom the actual show-up or expulsion of the calculi could not be detected;
- patients treated with Ca antagonists, cortisone or a-litics in the 3 months previous and/or after the colic episode;
- patients with anatomic malformations of the excretory tract;
- transplanted or mono-kidney patients, under more aggressive therapy;
- patients with more than one ureteral calculi;
- patients in whom the rigorous follow-up at the 3-month check-up from the episode was not possible; and
- patients who, after the axcess to emergency unit underwent extracorporeal shock wave litothripsy (ESWL), endourological or surgical procedures for stone removal.

Out of 1163 patients with pieloureteral colic, 402 were found valuable for experiment, as summarized in Figure 2.

Furthermore, for the actual statistical tests, we considered diagnostic criteria for the renal colic such as spontaneous expulsion (as reported by the patient), colic treatments together with ESWL,

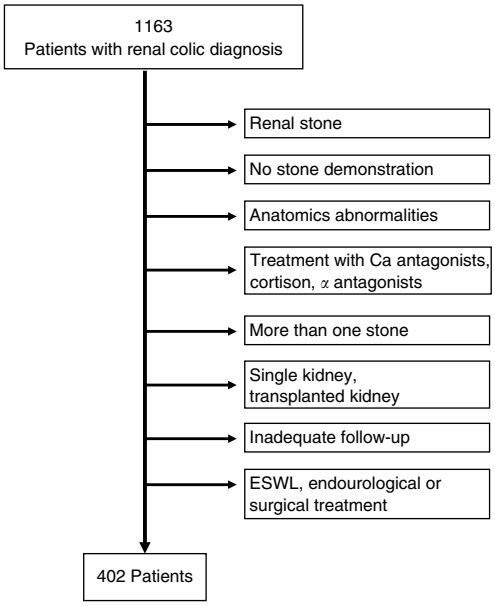

Figure 2 | Scheme for the selection of patients.

imaging showing ureteral calculi and clinical findings of the physician during the colic episode. As already mentioned, all the patients who, after the excess in the emergency unit, underwent ESWL, endourological or surgical treatment for the stone removal were excluded. The interval between first renal colic and stone passage was 6 months. In total, we considered nine clinically important factors (i.e., 'inputs') for each of the 408 patients (i.e., 'data points'). We selected the factors among those referred to as most influential in medical literature: age, sex, body mass index, fever, previous urological treatments, previous expulsion of stones, duration of the symptoms (in hours), dimension and position of the stone.[18] With each patient is also associated a 'binary output' value, corresponding to two classes of patients, that is, those ones with actual spontaneous expulsion of the stone (0) and those needing an intervention (1).

Experiments were performed using the learning algorithms LR, ANN, SVM and LP-SVM. Performance was evaluated using cross-validation, a well-known statistical methodology: 50 of the 402 data points (i.e., patients) were randomly selected and not used for training. After training with LR, ANN and SVM and LP-SVM on the 352 training data points, the accuracy of the trained classifier was tested on the hold-out test set of the 50 data points, by reporting the percentage of correctly predicted spontaneous expulsions (true negatives) and the percentage of correctly predicted cases needing intervention (true positives). This procedure was repeated 30 times, resulting in 30 different random splits in training and test sets. Finally, the average true positive and true negative rate on the 30 test sets was reported. Also, all simulations were performed both on the original data set that was not normalized, as well as on a data set with covariates normalized to have zero mean and unit variance.

LR and LP-SVM[9] are linear classification methods: they work best if both classes of data can be separated reasonably well in a linear way, that is, using a hyper-plane. If this is not the case, a nonlinear separating function is needed. This can be established with ANN or a standard SVM. The latter is based on a methodology known as kernel-based learning,[7,8] which allows one to come up with nonlinear versions of many well-known linear statistical algorithms. In the case of SVM, the kernel methodology is used to obtain the nonlinear SVM algorithm, derived from a linear maximal margin classifier. The algorithms were implemented using MATLAB® and commercial optimization software Mosek®.

The second objective, ranking the clinical factors according to their importance, was addressed in two ways. First, by using Fisher scores: these scores are computed as the difference in means of the factor values, computed for each class (i.e., input), corrected by their variance within each class; these scores therefore analyze the importance of every input factor independently. Second, we used the explicit LP-SVM feature weights:[8] these weights were obtained from the training algorithm, looking at all factors simultaneously and thus taking the dependence between the different inputs into account.

REFERENCES

1. Segura JW, Preminger GM, Assimos DG *et al.* Ureteral stones clinical guidelines panel summary report on the management of ureteral calculi. *J Urol* 1997; **158**: 1915–1921.
2. Anagnostu T, Tolley D. Management of ureteric stones. *Eur Urol* 2004; **45**: 714–721.
3. Miller OF, Kane CJ. Time to stone passage for observed ureteral calculi: a guide for patient education. *J Urol* 1999; **162**: 688–690.
4. Parekattil SJ, White MD, Moran ME, Kogan BA. A computer model to predict the outcome and duration of ureteral or renal calculous passage. *J Urol* 2004; **171**: 1436–1439.
5. Ramesh AN, Kambhampati C, Monson JR, Drew PJ. Artificial intelligence in medicine. *Ann R Coll Surg Engl* 2004; **86**: 334–338.
6. Cristianini N, Schoelkopf B. Support vector machines and kernel methods. *AI Mag* 2002; **23**: 31–41.
7. Boser BE, Guyon I, Vapnik V. A Training algorithm for optimal margin classifiers. *Proc Comput Learn Theory* 1992: 144–152, ACM Press.
8. Cristianini N, Shawe-Taylor J. *An Introduction to Support Vector Machines*. Cambridge University Press: Cambridge; 2000.
9. Bradley PS, Mangasarian OL, Street WN. Feature selection via mathematical programming. *INFORMS J Comput* 1998; **10**: 209–217.
10. Tu JV. Advantages and disadvantages of using artificial neural networks vs logistic regression for predicting medical outcomes. *J Clin Epidemiol* 1996; **49**: 1225–1231.
11. Batuello JT, Gamito EJ, Crawford E *et al.* Artificial neural network model for the assessment of lymph node spread in patients with clinically localized prostate cancer. *Urology* 2001; **57**: 481–485.
12. Cummings JM, Boullier JA, Izenberg SD *et al.* Prediction of spontaneous ureteral calculous passage by an artificial neural network. *J Urol* 2000; **164**: 326–328.
13. Bagli DJ, Agarwal SK, Venkateswaran S *et al.* Artificial neural networks in pediatric urology: prediction of sonographic outcome following pyeloplasty. *J Urol* 1998; **160**: 980–983.
14. Russel S, Norvig P. *Artificial Intelligence, A Modern Approach*. 2nd edn, Prentice-Hall: Englewood Cliffs, NJ.
15. Leane MM, Cumming I, Corrigan OI. The use of artificial neural networks for the selection of the most appropriate formulation and processing variables in order to predict the *in vitro* dissolution of sustained release minitablets. *PharmSciTech* 2003; **4**: E26.
16. Porpiglia F, Ghignone G, Fiori C *et al.* Nifedipine versus Tamsulosin for the management of lower ureteral stones. *J Urol* 2004; **172**: 568–571.
17. Dellabella M, Milanese G, Muzzonigro G. Efficacy of Tamsulosin in the medical management of juxtavesical ureteral stones. *J Urol* 2003; **170**: 2202–2205.
18. Gomha MA, Sheir KZ, Showky S *et al.* Can we improve the prediction of stone-free status after extracorporeal shock wave lithotripsy for ureteral stones? A neural network or a statistical model? *J Urol* 2004; **172**: 175–179.