The contribution of machine learning to predicting cancer outcome

R.G.P.M. van Stiphout [ab] E.O. Postma [c] V. Valentini [d] P. Lambin [a]

[a] *Maastro Clinic, P.O.Box 3035, 6202 NA , Maastricht*
[b] *Department of Knowledge Engineering, P.O.Box 616, 6200 MD, Maastricht*
[c] *Tilburg centre for Creative Computing (TiCC), P.O.Box 90153, 5000 LE, Tilburg*
[d] *Universita Cattolica S.Cuore, Largo A.Gemelli 8, 00168 Rome, Italy*

Abstract

Artificial intelligence methods may aid physicians to predict long-term outcome of individualized treatments of cancer. Hitherto, in the clinical literature on outcome prediction, traditional statistical methods prevail. This paper addresses the contribution of machine learning as compared to traditional statistical methods in the prediction of the long-term outcome of cancer treatment. Using a dataset of 1552 patients with clinical and pathological features a model was induced using a traditional statistical method (logistic regression) and a state-of-the-art machine learning method (proximal support vector machine). The models were trained to predict three outcome after five years: (1) local recurrence of the cancer, (2) metastases, and (3) overall survival of the patient. The performances of the models were evaluated using the Area-Under-the-Curve (AUC) of the Receiver Operating Characteristic (ROC) curve in combination with 10-fold cross-validation. The results reveal that both models perform on a par with mean AUCs between 0.72 and 0.78. No significant difference in performance could be established between the two methods. We conclude that proximal support vector machines do not improve the long-term cancer outcome prediction as compared to logistic regression. Further research is needed to establish if our result generalizes to other state-of-the-art methods in machine learning.

1 Introduction

In the context of cancer treatment, reliable prognosis is pivotal to clinical decision making. Prognosis is defined as the prediction of the future course and outcome of the disease process. Currently, cancer treatment is becoming more individualized, meaning that an assessment has to be made of the risks and benefits of a certain treatment based on the characteristics of the patient and the disease. Individualized treatments are expected to improve the final outcome of these patients when compared to the one-to-treat-all concept [14]. Improvements in patient examinations and the ability to record all data digitally, has resulted in large amounts of medical data. These data include information on demographics, imaging, blood biomarkers, tumor tissue markers, pathology (specimen evaluation after surgery), toxicities, genomics, and proteomics. The increasing volume and quality of the data, enables physicians to use computational techniques to assist in the complex clinical decision making process. In cancer research the use of traditional statistics, like multivariate regression or correlation analysis, is well established to identify prognostic factors for outcome prediction. During the last 10 years an increase in applying machine learning techniques in medicine and cancer research can be detected. Figure 1 illustrates the increased use of artificial intelligence, pattern recognition, and machine learning in the medical domain and the sub-domains of cancer and rectal cancer. The medical domain is well suited for these machine learning methods because of the large, noisy, incomplete and complex data sets [5]. Most applications of machine learning in the medical domain of cancer treatment are concerned with the diagnosis and detection of cancer, rather than with the response to treatment and prediction of outcome. This paper focuses on the contribution of machine learning to outcome prediction in cancer treatment.

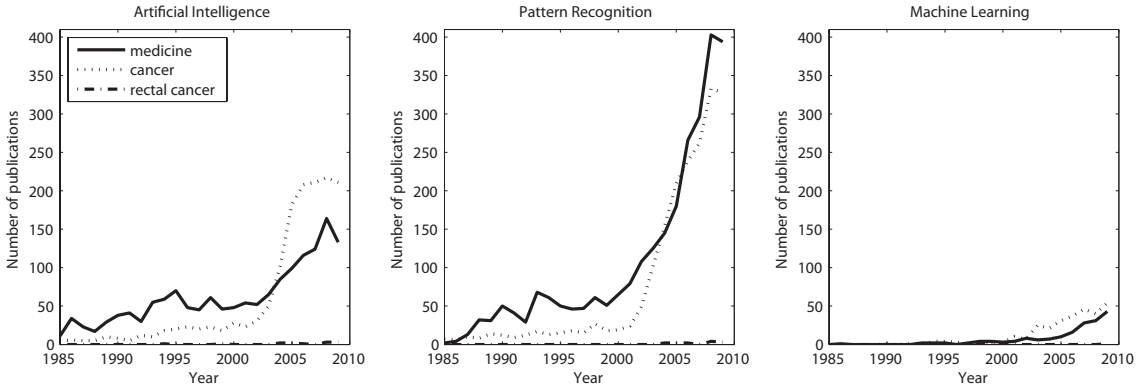

Figure 1: Number of publications for each year since 1985 using techniques in artificial intelligence, pattern recognition and machine learning for the domains of medicine, cancer and rectal cancer

1.1 Research question and approach

The goal of this study is to evaluate the value of machine learning techniques for outcome classification in cancer treatment when compared to traditional statistics. The research question addressed in this paper reads: *To what extent does machine learning contribute to the prognosis in cancer treatment?*

This study uses a published pooled dataset for the development of a prediction model for rectal cancer. This type of cancer has a very high prevalence and is worldwide the third most diagnosed type of cancer. In literature, the studies detecting prognostic factors for rectal cancer are numerous, but studies combining those factors to a prediction model for clinical decision making are sparse. Most studies detect prognostic factors by applying a Cox regression, in which the effect of several variables on the time to a specified event is calculated [2, 16, 1]. Kaplan-Meier curves describe the rate of outcome, for example survival or metastases, over time after the treatment is given. An example of three Kaplan-Meier curves is shown in Figure 2. From top to bottom, the three curves correspond to low, medium, and low risk, respectively.

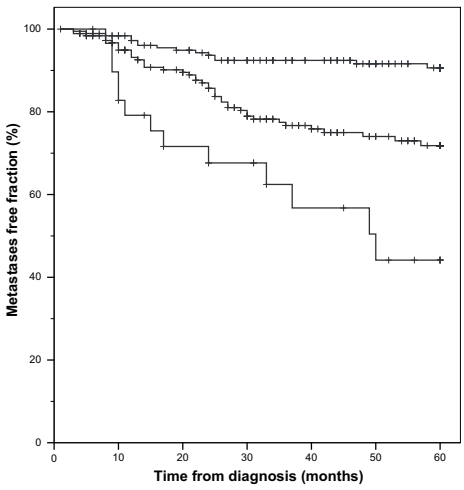

Figure 2: Example of Kaplan-Meier curve showing the rate of metastases-free patients over time in a population. Three different risk groups (from top to bottom, low, medium and high risk) are distinguished.

Although the Kaplan-Meier curve has become a clinical standard, multivariate methods that provide probabilities for the outcomes are to be preferred. A popular and powerful multivariate method is logistic regression and has been applied in many clinical studies [7, 13, 3, 8]. Unfortunately, none of these studies provides any classification performance measure for their models. The probable reason is that the identification of the prognostic factors is deemed most important.

Therefore, in this study we compare logistic regression with a state-of-the-art machine learning method on a medical data set. The support vector machine is the method of our choice, because it is known to have very good discrimination and generalization performance for binary classification problems [15].

2 Methods

This section describes the data (2.1), the statistical and machine learning methods (2.2 and 2.3), and the experimental set-up (2.4).

2.1 Data

This study uses data from patients submitted to a week of high dose radiotherapy followed by immediate surgery. The data includes 1552 patients and originates from three different published trials in Europe:

- The Dutch TME trial [11]: 913 patients,
- the Swedish rectal cancer trial [9]: 495 patients, and
- the Polish rectal cancer trial [4]: 144 patients.

Only those features that were present in all three data sets were included. Table 1 lists these features with the corresponding description and possible values. Most of the features are binary or ordinal, except for the continuous features age and tumor distance. The evaluated outcomes are binary (no/yes) and involve the occurrence of these events within 5 years: local recurrence, metastases and survival.

Nr.	Feature	Description	Unit/values
1	Age	The age of the patient	[years]
2	Gender	The gender of the patient	{male, female}
3	Distance	Distance of the tumor to the anal verge	[cm]
4	Surgery type	Type of performed surgical procedure	{Conventional surgery, TME}
5	Surgery group	Categorized surgical procedure	{No surgery, LAR, APR}
6	Resisual	Presence of residual disease	{no, yes}
7	pT	Pathological tumor stage	{pT0, pT1, pT2, pT3/4}
8	pN	Pathological nodal stage	{pN0, pN1, pN2}
9	PA stage	Overall pathological stage	{0, I, II, III, IV}
10	psurgcom	Post surgical complications	{no, yes}

Table 1: Description of the features in the dataset

2.2 Logistic regression

Logistic regression is used for the prediction of the probability of an occurrence of an event by fitting the data to a logit function. The logit function z is a linear combination of regression coefficients b_i, input variables X_i and intercept constant b_0:

$$z = b_0 + b_1 X_1 + b_2 X_2 + ... + b_k X_k$$

By rewriting the logit function, the probability of an event occurence is defined as:

$$P(event) = \frac{1}{1 + e^{-z}}$$

2.3 Proximal support vector machine

The proximal support vector machine (pSVM) [12] is a computationally efficient alternative of the standard SVM. While the standard SVM maximizes the margin between support vector datapoints, the pSVM classifies points depending on proximity to one of two parallel planes that are pushed as far apart as possible (see

Figure 3). For the matrix A with size n (number of features) $\times$ m (number of datapoints), the formulation of the optimization problem for a SVM with a linear kernel is defined as:

$$\min_{(w,\gamma,y)\epsilon\mathbb{R}^{n+1+m}} \nu\frac{1}{2} \parallel y \parallel + \frac{1}{2}(w'w + \gamma^2)$$

The parameters are the normal to the bounding plane w, the constant determining the location of the bounding planes relative to the origin γ, the outcome label y and the sensitivity parameter ν. In the pSVM, the condition $D(Aw - e\gamma) + y \geq e$ (with diagonal matrix D and the error margin e) of the SVM, is modified into $D(Aw - e\gamma) + y = e$, which reduces the optimization problem to an explicit exact solution which is faster to compute. Computational studies demonstrated that the pSVM classifier performs on a par with the SVM [12].

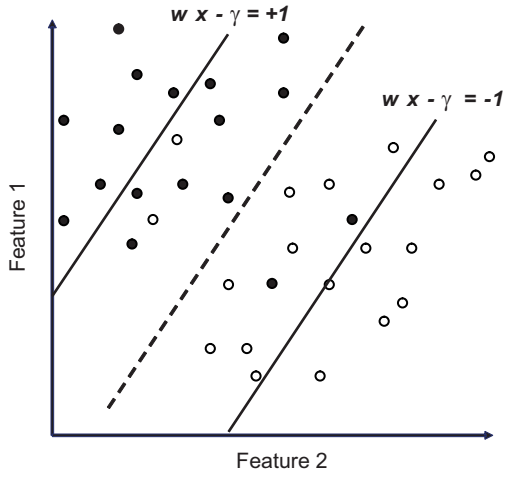

Figure 3: Illustration of the data fitting on two parallel planes in the proximal support vector machine

2.4 Experimental set-up

Figure 4 presents an overview of the experimental set-up. Our study centers on the difference between a traditional statistical method (logistic regression) and a state-of-the-art machine learning method (support vector machine). As can be seen in the diagram, in our study there is also a difference in the feature selection. More specifically, for the machine learning method we employ an exhaustive feature search, whereas for the traditional statistical method univariate analysis is used. This implies that in case we find a superior performance for the support vector machine as compared to logistic regression, this may be due to the exhaustive feature search.

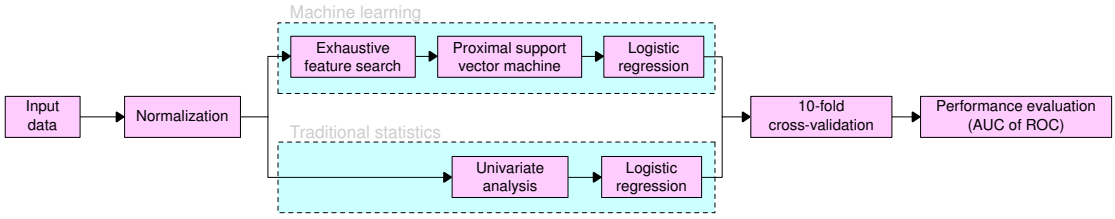

Figure 4: The flowchart of the performed experiments for both machine learning and traditional statistics

The three data sets were pooled and any missing values (1.6%) were substituted by the average value (continuous) or most common value (ordinal). Features were normalized into z-scores ($z = (X - \mu)/\sigma$) to allow comparison of the contribution of each feature (coefficients). In the logistic regression experiment, a univariate test was performed with a Wilcoxon signed-ranks test to select admissible features. Features with a p-value < 0.05 were used as an input for the logistic regression. The pSVM experiment involves feature selection by an exhaustive feature search, meaning that all possible combinations of input features were

tested for the area-under-the-curve (AUC) of the receiver operating characteristic (ROC) curve. Feature selection was based on the presence of features in the highest 5% of AUCs. The pSVM classifier was tested and its output was converted to probabilities using the method proposed by Jakulin [10]. Both the logistic regression and pSVM methods were tested for their generalization capacity by 10-fold cross-validation (CV) and the compared performance measure was the AUC of the ROC curve. All methods were implemented in Matlab 7.1 (Mathworks Inc., Natick, MA).

3 Results

The results of the experiment are presented in Table 2 and Figure 5. The mean AUC and the standard deviation (SD) were calculated as part of the 10-fold cross-validation procedure.

Outcome	Phase	AUC	pSVM	LR
Local recurrence	Training	Mean	0.7597	0.7669
		SD	0.0090	0.0127
	Test	Mean	0.7703	0.7473
		SD	0.0853	0.1049
Metastases	Training	Mean	0.7789	0.7798
		SD	0.0043	0.0054
	Test	Mean	0.7777	0.7757
		SD	0.039	0.0562
Survival	Training	Mean	0.7308	0.7454
		SD	0.0057	0.0034
	Test	Mean	0.7272	0.7391
		SD	0.0520	0.029

Table 2: Results of the proximal support vector machine (pSVM) scheme and the logistic regression (LR) for the three cancer outcomes: local recurrence, metastases, and survival.

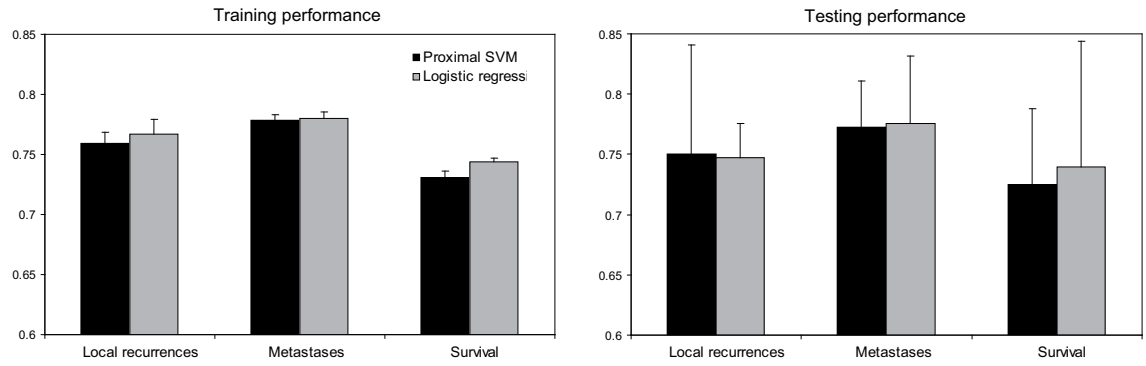

Figure 5: Visual representation of the results in Table 2. The bars represent the means and the error bars represent the standard deviations.

Below we list the linear model equations that include the features that are identified as predictors with the logistic regression scheme (z_{LR}) and the proximal SVM scheme (z_{pSVM}), respectively. The model equations for each predictor and the offset are given for the three outcomes, separately.

Local recurrence:

- $z_{LR} = -0.5943 \cdot \text{Surgery type} - 0.3060 \cdot \text{Time to surgery} + 0.5654 \cdot \text{Residual} + 0.5928 \cdot \text{PAstage} - 3.1277$

- $z_{pSVM} = -0.2298 \cdot \text{Distance} - 0.5141 \cdot \text{Surgery type} + 0.6509 \cdot \text{Residual} + 0.3667 \cdot \text{PAstage} - 3.0334$

Metastases:

- $z_{LR} = -0.1411 \cdot \text{Age} - 0.1893 \cdot \text{Surgery type} + 0.2913 \cdot \text{Residual} + 0.5721 \cdot \text{pT} + 0.6531 \cdot \text{pN} - 1.4245$
- $z_{pSVM} = -0.1134 \cdot \text{Age} - 0.1527 \cdot \text{Surgery type} + 0.3396 \cdot \text{Residual} + 0.3560 \cdot \text{pT} + 0.7564 \cdot \text{pN} - 0.0675 \cdot \text{PAstage} - 1.3684$

Survival:

- $z_{LR} = 0.4269 \cdot \text{Age} - 0.1392 \cdot \text{Gender} - 0.2353 \cdot \text{Surgery type} + 0.3632 \cdot \text{Residual} + 0.3227 \cdot \text{pT} + 0.5139 \cdot \text{pN} + 0.1667 \cdot \text{psurgcom} - 0.7944$
- $z_{pSVM} = 0.4176 \cdot \text{Age} + 0.4054 \cdot \text{Residual} + 0.3791 \cdot \text{pN} + 0.2329 \cdot \text{PAstage} - 0.7602$

4 Discussion

We performed a comparative evaluation of logistic regression and the support vector machine on an extensive medical data set. Table 2 and Figure 5 show that both methods perform at an equal level. An important reason that may explain why the pSVM does not outperform logistic regression is the use of a linear kernel for the proximal SVM. However, preliminary experiments with another dataset and a non-linear multivariate model did not improve performance.
The shapes of the ROC curves show no difference, implying that no gain in sensitivity or specificity can be obtained by adopting either method. Also, the model equations show also high similarity for the found predictors and corresponding weights. For local recurrence prediction only two features differed (time to surgery in logistic regression and tumor distance in pSVM). For metastases, the pSVM method selects PA-stage as an extra predictor. The model equations for survival prediction differed the most because LR found seven predictors and pSVM only four. The larger number of predictors for LR may be due to overfitting. LR seems to suffer more from overfitting because the performances on the test set are relatively lower compared to the training performances than for pSVM.
Despite the disappointing results of the pSVM as compared to LR, the obtained AUCs are good in clinical practice. These models deal with 5-year predictions based on only clinical and pathological data. Physicians tend to predict these long-term outcomes with very poor accuracy (AUC around 0.5) when only raw data is provided [6].

5 Conclusions and future work

No clear difference between the performances of the standard statistical method and the machine learning method could be established. In future research, other kernels for the SVM and other machine learning classifiers and feature selection schemes will be studied to improve performance for follow-up outcome prediction. Also the addition of expert knowledge in combination with Bayesian networks is expected to achieve this improvement.

References

[1] I. Bedrosian, G. Giacco, L. Pederson, M. A. Rodriguez-Bigas, B. Feig, K. K. Hunt, L. Ellis, S. A. Curley, J. N. Vauthey, M. Delclos, C. H. Crane, N. Janjan, and J. M. Skibber. Outcome after curative resection for locally recurrent rectal cancer. *Dis Colon Rectum*, 49(2):175–82, 2006.

[2] H. Bouzourene, F. T. Bosman, M. Matter, and P. Coucke. Predictive factors in locally advanced rectal cancer treated with preoperative hyperfractionated and accelerated radiotherapy. *Hum Pathol*, 34(6):541–8, 2003.

[3] A. Bufalari, C. Boselli, G. Giustozzi, and L. Moggi. Locally advanced rectal cancer: a multivariate analysis of outcome risk factors. *J Surg Oncol*, 74(1):2–10, 2000.

[4] K. Bujko, MP Nowacki, L. Kepka, J. Oledzki, M. Bebenek, and M. Kryj. Postoperative complications in patients irradiated pre-operatively for rectal cancer: report of a randomised trial comparing short-term radiotherapy vs chemoradiation. *Colorectal Disease*, 7(4):410–416, 2005.

[5] J.A. Cruz and D.S. Wishart. Applications of machine learning in cancer prediction and prognosis. *Cancer Informatics*, 2:59–77, 2006.

[6] C. Dehing-Oberije, S. Yu, D. De Ruysscher, S. Meersschout, K. Van Beek, Y. Lievens, J. Van Meerbeeck, W. De Neve, B. Rao, H. van der Weide, et al. Development and External Validation of Prognostic Model for 2-Year Survival of Non–Small-Cell Lung Cancer Patients Treated With Chemoradiotherapy. *International journal of radiation oncology, biology, physics*, 74(2):355–362, 2009.

[7] J. A. Diaz-Gonzalez, F. A. Calvo, J. Cortes, J. L. Garcia-Sabrido, M. Gomez-Espi, E. Del Valle, F. Munoz-Jimenez, and E. Alvarez. Prognostic factors for disease-free survival in patients with t3-4 or n+ rectal cancer treated with preoperative chemoradiation therapy, surgery, and intraoperative irradiation. *Int J Radiat Oncol Biol Phys*, 64(4):1122–8, 2006.

[8] R. C. Dresen, E. E. Peters, H. J. Rutten, G. A. Nieuwenhuijzen, T. B. Demeyere, A. J. van den Brule, A. G. Kessels, R. G. Beets-Tan, J. H. van Krieken, and I. D. Nagtegaal. Local recurrence in rectal cancer can be predicted by histopathological factors. *Eur J Surg Oncol*, 35(10):1071–7, 2009.

[9] J. Folkesson, H. Birgisson, L. Pahlman, B. Cedermark, B. Glimelius, and U. Gunnarsson. Swedish rectal cancer trial: long lasting benefits from radiotherapy on survival and local recurrence rate. *Journal of Clinical Oncology*, 23(24):5644–5650, 2005.

[10] A. Jakulin, M. Mozina, J. Demsar, I. Bratko, and B. Zupan. Nomograms for visualizing support vector machines. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 108–117, 2005.

[11] E. Kapiteijn, C.A.M. Marijnen, I.D. Nagtegaal, H. Putter, W.H. Steup, T. Wiggers, H.J.T. Rutten, L. Pahlman, B. Glimelius, J. Van Krieken, et al. Preoperative radiotherapy combined with total mesorectal excision for resectable rectal cancer. *The New England journal of medicine*, 345(9):638–646, 2001.

[12] O.L. Mangasarian and E.W. Wild. Proximal support vector machine classifiers. *Proceedings KDD-2001: Knowledge Discovery and Data Mining*, pages 77–86, 2001.

[13] C. Massacesi, A. Norman, T. Price, M. Hill, P. Ross, and D. Cunningham. A clinical nomogram for predicting long-term survival in advanced colorectal cancer. *Eur J Cancer*, 36(16):2044–52, 2000.

[14] DR Parkinson and J. Ziegler. Educating for personalized medicine: a perspective from oncology. *Clinical Pharmacology & Therapeutics*, 86(1):23–25, 2009.

[15] V.N. Vapnik and V. Vapnik. *Statistical learning theory*. Wiley New York, 1998.

[16] J. N. Wiig, S. G. Larsen, S. Dueland, and K. E. Giercksky. Preoperative irradiation and surgery for local recurrence of rectal and rectosigmoid cancer. prognostic factors with regard to survival and further local recurrence. *Colorectal Dis*, 10(1):48–57, 2008.