A Robust Machine Learning Protocol for Prediction of Prostate Cancer Survival at Multiple Time-Horizons

Wojciech Lesiński[1] 0000-0001-6628-9466 and Witold R. Rudnicki[1,2,3]
0000-0002-7928-4944

[1] Institute of Informatics, University of Białystok, Białystok, Poland
w.lesinski@uwb.edu.pl
[2] Computational Centre, University of Białystok, Białystok, Poland
[3] Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, Warsaw, Poland

Abstract. Prostate cancer is one of the leading causes of cancer death in men in Western societies. Predicting patients' survival using clinical descriptors is important for stratification in the risk classes and selecting appropriate treatment. Current work is devoted to developing a robust Machine Learning (ML) protocol for predicting the survival of patients with metastatic castration-resistant prostate cancer. In particular, we aimed to identify relevant factors for survival at various time horizons. To this end, we built ML models for eight different predictive horizons, starting at three and up to forty-eight months. The model building involved the identification of informative variables with the help of the MultiDimensional Feature Selection (MDFS) algorithm, entire modelling procedure was performed in multiple repeats of cross-validation. We evaluated the application of 5 popular classification algorithms: Random Forest, XGBoost, logistic regression, k-NN and naive Bayes, for this task. Best modelling results for all time horizons were obtained with the help of Random Forest. Good prediction results and stable feature selection were obtained for six horizons, excluding the shortest and longest ones. The informative variables differ significantly for different predictive time horizons. Different factors affect survival rates over different periods, however, four clinical variables: ALP, LDH, HB and PSA, were relevant for all stable predictive horizons. The modelling procedure that involves computationally intensive multiple repeats of cross-validated modelling, allows for robust prediction of the relevant features and for much-improved estimation of uncertainty of results.

Keywords: prostate cancer, feature selection, machine learning, random forest, xgboost, logistic regression, knn.

1 Introduction

Prostate cancer is the sixth most common cancer in the world (in the number of new cases), the third most common cancer in men, and the most common

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_12

cancer in men in Europe, North America, and some parts of Africa [7]. Prostate cancer is a form of cancer that develops in the prostate gland. As with many tumours, they can be benign or malignant. Prostate cancer cells can spread by breaking away from a prostate tumour. They can travel through blood vessels or lymph nodes to reach other body parts. After spreading, cancer cells may attach to other tissues and grow to form new tumours, causing damage where they land. Metastatic castrate-resistant prostate cancer (mCRPC) is a prostate cancer with metastasis that keeps growing even when the amount of testosterone in the body is reduced to very low levels. Many early-stage prostate cancers need normal testosterone levels to grow, but castrate-resistant prostate cancers do not.

One of the essential goals of research on prostate cancer is the development of predictive models for patients' survival. The classical approach was proposed by Halabi [9], [10] and coworkers, who used proportional hazard models. Halabi's model is based on eight clinical variables: lactate dehydrogenase, prostate-specific antigen, alkaline phosphatase, Gleason sum, Eastern Cooperative Oncology Group performance status, haemoglobin, and the presence of visceral disease. Another notable work, [18], used a joint longitudinal survival–cure model based mainly on PSA level. In Mahapatra et al. [12] survival prediction models were built using DNA methylation data.

More recently, the Prostate Cancer DREAM Challenge was organised in 2015 by DREAM community to improve predictive models [1]. The DREAM Challenges [4] is a non-profit, collaborative community effort comprising contributors from across the research spectrum, including researchers from universities, technology companies, not-for-profits, and biotechnology and pharmaceutical companies. Many independent scientists made survival models and predictions within the Prostate Cancer DREAM Challenge. The challenge resulted in developing multiple new algorithms with significantly improved performance compared to a reference Halabi model, see [8].

Both the reference model and best model developed within DREAM Challenge, are variants of the proportional hazard model. This model has a relatively rigid construction - the influence of variables used for modelling is identical for different predictive horizons. The current study explores an alternative approach, where a series of independent models is developed for different predictive horizons. Identification of informative variables is performed independently at each horizon. This approach allows a more fine-grained analysis of the problem.

What is more, the format of the DREAM Challenges has two significant limitations, namely, inefficient use of data and inefficient evaluation of modelling error. The main problem is the strict division into training and testing sets. Solutions were evaluated based only on test sets' results, and such setup necessarily results in random biases arising due to the data split. Moreover, not all available data is used for the development of the model, and estimates of the error bounds of the model are based on the single data split between the training and validation set.

The current study applies a robust modelling protocol for building a series of predictive models with different time horizons. This leads to obtaining compara-

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_12

ble predictions compared to aggregate models from the DREAM Challenge but also allows for detailed analysis of the influence of various factors for survival in different time horizons.

2 Materials and Methods

2.1 Data

The publicly released data from the DREAM Challenge was used for the current study. The data set comprises 1600 individual cases collected in three clinical trials: VENICE [17], MAILSAIL [15] and ENTHUSE [6]. Data contains five groups of variables and two clinical indicators of prostate cancer progression corresponding to the patient's medical history, laboratory values, lesion sites, previous treatments, vital signs, and basic demographic information. The final record for the patient consists of 128 descriptive variables and two decision variables: the patient's status (alive or deceased) and the time of last observation. Based on this data nine binary decision variables were created for different predictive horizons: 3 months, 6 months, 1 year, 18 months, 2, 3, and 4 years.

2.2 Modelling

Machine learning methods often produce models biased towards the training set. In particular, the selection of hyper-parameters of the algorithms and the selection of variables that will be used for modelling can introduce strong biases. What is more - a simple selection of the best-performing model also can lead to a bias. Finally, dividing the data set into training and validation sets involves bias by creating two partitions with negative correlations between fluctuations from the actual averages. To minimize the influence of biases and estimate the variance of the model-building process, the entire modelling procedure was performed within multiple repeats of the cross-validation loop.

A single iteration of our approach is based on the following general protocol:

- Split the data into training and validation set;
- Identify informative variables in the training set;
- Select most informative variables;
- Build model on training set;
- Estimate models on validation set,

This protocol was repeated 150 times for each classifier – time horizon pair (30 iterations of 5-fold cross-validation). The series of predictive models were constructed for various horizons of prediction. A binary classification model of survival beyond this horizon was built at each horizon.

The data set was imbalanced, especially in the short time horizons. While some algorithms, e.g. Random Forest, are relatively robust when dealing with imbalanced data, in many cases, imbalanced data may cause many problems for machine learning methods [11]. The main difficulty lies in finding properties

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_12

that discern the minority class from the much more numerous majority class. The simple downsampling of the majority class can deal with this. It involves randomly removing observations from the majority class to prevent its signal from dominating the learning algorithm. Downsampling was used for prediction in 3 and 6 months time horizon.

Two measures used to evaluate models' quality are suitable for unbalanced data. Matthews Correlation Coefficient (MCC) [13], MCC measures the correlation of distribution of classes in predictions with actual distribution in the sample. The area under the receiver operating curve (AUROC or AUC) is a global measure of performance that can be applied to any classifier that ranks the binary prediction for objects.

Feature selection Identification of informative variables was performed with the help of the Multidimensional Feature Selection (MDFS) algorithm [16, 14]. The method is based on information theory and considers synergistic interactions between descriptive variables. It was developed in our laboratory and implemented in the R package *MDFS*. The algorithm returns binary decisions about variables' relevance and ranking based on Information Gain and p-value.

Classification For modelling, we used five popular classifiers: **Random Forest** algorithm [2], **XGboost** [3] **k-Nearest Neighbours (k-NN), Logistic regression,** and **Naive Bayes.** Random Forest and XGboost are based on decision trees and work well *out of the box* on most data sets [5]. Logistic regression represents generalized linear models, whereas k-NN is a simple and widely known method based on distances between objects. Finally, Naive Bayes is a simple algorithm that may work well for additive problems. All tests were performed in 30 repeats of the 5-folds cross-validation. Both feature selection and ML model building were performed within cross-validation. All ML algorithms were applied to the identical folds in the cross-validation to ensure fair comparisons.

Table 1. Cross-validated quality of predictions for five classification algorithms at 8 predictive horizons.

time horizon	Random Forest		XGboost		logistic regression		k-NN		Naive Bayes	
	MCC	AUC	MCC	AUC	MCC	AUC	MCC	AUC	MCC	AUC
3 months	0.23	0.66	0.26	0.65	0.17	0.62	0.16	0.57	0.23	0.64
6 months	0.34	0.72	0.30	0.70	0.27	0.68	0.27	0.62	0.23	0.66
9 months	0.43	0.78	0.36	0.74	0.36	0.74	0.33	0.65	0.28	0.71
12 months	0.44	0.79	0.39	0.76	0.40	0.76	0.35	0.66	0.30	0.72
18 months	0.44	0.77	0.38	0.74	0.39	0.75	0.34	0.67	0.28	0.71
24 months	0.45	0.80	0.41	0.77	0.38	0.74	0.40	0.68	0.43	0.76
36 months	0.42	0.77	0.37	0.74	0.32	0.71	0.26	0.62	0.37	0.72
48 months	0.21	0.65	0.24	0.66	0.20	0.61	0.18	0.58	0.17	0.66

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_12

3 Results and Discussion

Survival predictions were made for eight different time horizons with the help of the five classifiers mentioned before.

As can be expected, the worst-performing models were built for the shortest and longest horizons, most likely due to a small number of cases in the minority class (non-survivors) for the shortest horizon and an overall small number of non-censored cases for the longest horizon, see Table 1. The Random Forest classifier obtained the best results for all predictive horizons. At the shortest horizon, we obtained $AUC = 0.66$ and $MCC = 0.23$. The quality of predictions increases with an increased horizon. The best results were obtained for horizons between 9 and 24 months, with acceptable results for 6 and 36 months. The prediction quality falls at 48 months horizon back to $AUC = 0.65$ and $MCC = 0.21$. AUC curves for all Random Forest models for all examined time horizons are displayed in Fig. 1.

The XGboost produced slightly worse models than Random Forest. The difference is insignificant for a single horizon but significant when all horizons are considered together. The logistic regression generally produced slightly worse models than XGboost. The two simplest methods produced significantly worse models – differences from the best model were significant for almost all predictive horizons.

Evaluation of feature importance was built based on 30 repeats of 5 folds cross-validation procedure. The cumulative ranking of importance for each period was obtained as a count of occurrences of variables in the sets of the top ten most relevant variables in all repeats of the cross-validation procedure. In cases when the feature selector reports fewer than ten relevant variables, the highest-ranked descriptors were included. The feature selection results were volatile in

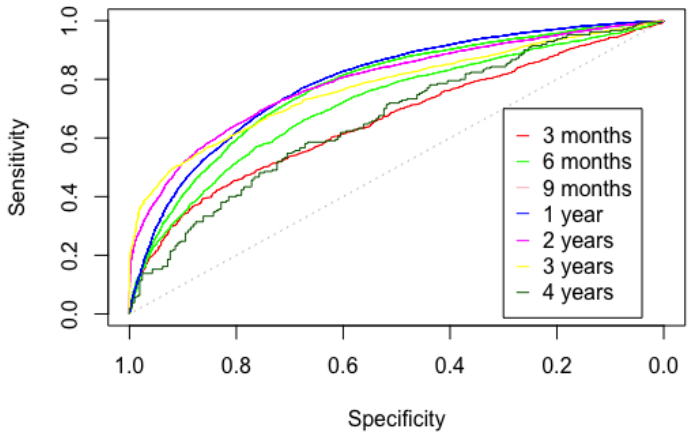

Fig. 1. ROC plots for Random Forest model.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_12

Table 2. Feature selection ranking in given period. Importance for short-, medium-, and long-term prediction was computed as the geometric mean of ranking. If a variable has not been ranked in a given horizon ranking, equal to 20 was used.

Descriptor	Predictive horizons											
	Months						short		medium		long	
	6	9	12	18	24	36	imp.	score	imp.	score	imp.	score
ALP	1	1	1	1	1	3	1	1	1	1	1.7	1
PSA	4	5	5	3	3	8	2.2	2	3.9	3	4.9	5
HB	2	3	3	4	6	5	2.4	3	3.5	4	5.5	6
LDH	3	4	4	2	2	7	3.5	4	2.8	2	3.7	4
ECOG-C	9	9	-	8	10	9	9	6	12.6	10	9.5	8
CCRC	-	8	6	-	7	6	12.6	10	11.0	8	6.5	4
NA	-	7	9	-	-	10	11.8	8	13.4	-	14.1	10
AST	6	2	2	5	8	-	3.5	4	3.2	5	12.6	9
ANALGESICS	-	10	10	-		-	14.1	-	14.1	6	-	-
ALB	5	6	8	-	-	-	5.5	4	12.6	6	-	-
TBILI	8	-	-	-	-	-	12.6	10	-	-	-	-
REGION-C	-	-	-	6	5	1	-	-	11.0	7	2.2	2
MHNEOPLA	7	-	-	-	-	-	11.8	8	-	-	-	-
ALT	10	9	-	-	-	-	9.5	7	-	-	-	-
NUE	-	-	7	7	-	-	-	-	7.0	9	-	-
PLT	-	-	-	9	-	-	-	-	13.4	-	-	-
BMI	-	-	-	10	-	-	-	-	14.1	-	-	-
SMOKE	-	-	-	-	4	2		-	-	-	2.8	3
TSTAG-DX	-	-	-	-	9	4	-	-	-	-	6.0	7

the shortest and the longest time horizons. In particular, none of the variables was present within the top ten variables in all 150 repeats of cross-validation. In comparison, at 12 months horizon, six variables were included in the top ten in all 150 cases. Therefore, the two extreme horizons were removed from further analyses. Other horizons were divided into three groups: short-term (6 and 9 months), medium-term (12 and 18 months) and long-term (24 and 36 months). The results of the feature selection procedure for these horizons are displayed in Table 2

Nineteen variables were included in top ten most relevant variables in six predictive horizons. Only four of them, alkaline phosphatase level (ALP), lactate dehydrogenase level (LDH), prostate-specific antigen (PSA) and haemoglobin level (HB) were important for all predictive horizons. One should note, that these four variables were the most relevant ones for short- and medium-term predictions, and were also quite important for the long-term predictions. In particular ALP was the most relevant variable for all but one predictive horizons. Four other variables, namely ECOG-C, calculated creatinine clearance (CCRC), sodium (NA) and aspartate aminotransferase (AST) were relevant for short-, medium-, and long-term predictions, but with lower ranks. Three variables, ANALGESICS, total bilirubin level (TBILI) and albumin level (ALB) were rel-

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_12

evant for the short- and medium-term predictions. Appearance of other sites of neoplasms (MHNEOPLA) and level of alanine transaminase (ALT) are important only for short-term predictions. Region of the world (REGION-C) is relevant for medium- and long-term predictions. Neutrophils (NUE), body mass index (BMI) and platelet count (PLT) appear only in medium prediction. Finally two variables (smoking status (SMOKE) and primary tumor stage score (TSTAG-DX)) are relevant for the medium and long-term predictions only.

What is interesting, the variables corresponding to socioeconomic status and behaviour (REGION-C and SMOKE), while irrelevant for the short- and medium-term predictions become the most important ones for prognosis of the long-term survival, in particular for the 36 months prediction.

These results agree very well with basic medical knowledge. The bad results of medical tests showing the overall health of the patient, such as total bilirubin level or platelet count level are strong indicators of a bad prognosis in the short period but are not very important for long-term prediction. On the other hand presence of other neoplasms may not be an indicator of immediate threat, but are very serious risk factors in the mid- and long-term horizon.

4 Conclusions and Future Works

The approach presented in the current study relies on computationally intensive procedures. The multiple repeats of cross-validation and inclusion of feature selection within cross-validation allow for the removal of biases that are observed for a single division between training and validation set, or even for a single run of cross-validation. What is more, it gives the opportunity to estimate variance that results from both performing feature selection and model building on finite and relatively small samples. For future work, we would like to combine our methods with classical survival models. Finding new datasets also seems like a good idea.

References

1. Abdallah, K., Hugh-Jones, C., Norman, T., Friend, S., Stolovitzky, G.: The prostate cancer dream challenge: a community-wide effort to use open clinical trial data for the quantitative prediction of outcomes in metastatic prostate cancer (2015)
2. Breiman, L.: Random forests. Machine Learning **45**, 5–32 (2001)
3. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. pp. 785–794 (08 2016). https://doi.org/10.1145/2939672.2939785
4. Costello, J., Stolovitzky, G.: Seeking the wisdom of crowds through challenge-based competitions in biomedical research. Clinical pharmacology and therapeutics **93** (02 2013). https://doi.org/10.1038/clpt.2013.36
5. Fernández-Delgado, M., et al.: Do we need hundreds of classifiers to solve real world classification problems. J. Mach. Learn. Res **15**(1), 3133–3181 (2014)
6. Fizazi, K., et al.: Phase iii, randomized, placebo-controlled study of docetaxel in combination with zibotentan in patients with metastatic castration-resistant prostate cancer. Journal of Clinical Oncology **31**(14), 1740–1747 (2013)

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_12

7. Grönberg, H.: Prostate cancer epidemiology. The Lancet **361**(9360), 859 – 864 (2003)
8. Guinney, J., et al.: Prediction of overall survival for patients with metastatic castration-resistant prostate cancer: development of a prognostic model through a crowdsourced challenge with open clinical trial data. The Lancet Oncology **18**(1), 132–142 (2017). https://doi.org/https://doi.org/10.1016/S1470-2045(16)30560-5, https://www.sciencedirect.com/science/article/pii/S1470204516305605
9. Halabi, S., et al.: Prognostic model for predicting survival in men with hormone-refractory metastatic prostate cancer. Journal of clinical oncology : official journal of the American Society of Clinical Oncology **21**, 1232–7 (05 2003). https://doi.org/10.1200/JCO.2003.06.100
10. Halabi, S., et al.: Updated prognostic model for predicting overall survival in first-line chemotherapy for patients with metastatic castration-resistant prostate cancer. Journal of clinical oncology : official journal of the American Society of Clinical Oncology **32** (01 2014). https://doi.org/10.1200/JCO.2013.52.3696
11. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering **21**(9), 1263–1284 (Sep 2009)
12. Mahapatra, S., et al.: Global methylation profiling for risk prediction of prostate cancer. Clinical Cancer Research **18**(10), 2882–2895 (2012). https://doi.org/10.1158/1078-0432.CCR-11-2090
13. Matthews, B.: Comparison of the predicted and observed secondary structure of t4 phage lysozyme. Biochimica et Biophysica Acta **405**(2), 442—-451 (1975)
14. Mnich, K., Rudnicki, W.R.: All-relevant feature selection using multidimensional filters with exhaustive search. Information Sciences **524**, 277 – 297 (2020). https://doi.org/https://doi.org/10.1016/j.ins.2020.03.024
15. Petrylak, D., et al.: Docetaxel and prednisone with or without lenalidomide in chemotherapy-naive patients with metastatic castration-resistant prostate cancer (mainsail): a randomised, double-blind, placebo-controlled phase 3 trial. The lancet oncology **16**(4), 417–425 (2015)
16. Piliszek, R., Mnich, K., Migacz, S., Tabaszewski, P., Sułecki, A., Polewko-Klim, A., Rudnicki, W.: MDFS: MultiDimensional Feature Selection in R. The R Journal (2019). https://doi.org/10.32614/RJ-2019-019, https://doi.org/10.32614/RJ-2019-019
17. Tannock, I.F., et al.: Aflibercept versus placebo in combination with docetaxel and prednisone for treatment of men with metastatic castration-resistant prostate cancer (venice): a phase 3, double-blind randomised trial. The lancet oncology **14**(8), 760–768 (2013)
18. Yu, M., et al.: Individual prediction in prostate cancer studies using a joint longitudinal survival–cure model. Journal of the American Statistical Association **103**(481), 178–187 (2008)

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_12