Machine learning based predictive model of early mortality in stage III and IV prostate cancer

Robert Chen[1]
[1]Emory University School of Medicine

Abstract

Prostate cancer remains the third highest cause of cancer-related deaths. Metastatic prostate cancer could yield poor prognosis, however there is limited work on predictive models for clinical decision support in stage III and IV prostate cancer.

We developed a machine learning model for predicting early mortality in prostate cancer (survival less than 21 months after initial diagnosis). A cohort of 10,303 patients was extracted from the Surveillance, Epidemiology and End Results (SEER) program. Features were constructed in several domains including demographics, histology of primary tumor, and metastatic sites. Feature selection was performed followed by regularized logistic regression. The model was evaluated using 5-fold cross validation and achieved 75.2% accuracy with AUC 0.649. Of the 19 most predictive features, all of them were validated to be clinically meaningful for prediction of early mortality.

Our study serves as a framework for prediction of early mortality in patients with stage II and stage IV prostate cancer, and can be generalized to predictive modeling problems for other relevant clinical endpoints. Future work should involve integration of other data sources such as electronic health record and genomic or metabolomic data.

Keywords: prostate cancer, machine learning, predictive modeling, electronic health records, biomarkers

Introduction

Prostate cancer is the cancer with highest prevalence among men and remains the third highest cause of cancer-related deaths.[1] The median 5-year survival across all patients with prostate cancer is over 98%. However, only 30% of patients with advanced cancer have 5-year survival rate. The median overall survival of patients with distant prostate cancer is only 31%. [2]

There is limited work on the capabilities of predictive modeling approaches for early detection of mortality risk in patients with distant prostate cancer. One approach sought to stratify patients into subgroups with Cox multivariable regression [3], but used a limited feature set including solely clinical stage and Gleason score. Furthermore, while the prostate specific antigen [4]biomarker has been used in practice clinically as a prognostic marker, there is limited evidence of its predictive value due to low specificity.

Meanwhile, machine learning has shown to have strong potential for early detection of clinical endpoints in applications such as disease prediction[5–8], readmission prediction [9–11], drug adverse event prediction [12], among others.

In this study we developed a machine learning model based on logistic regression for prediction of early mortality from retrospective real-world data and evaluated their performance. Our model leverages predictive features in a variety of domains including demographics, histology, staging, tumor spread and metastatic status.

Methods

A cohort of patients was selected from the The Surveillance, Epidemiology and End Results (SEER) program public retrospective dataset[13].

Cohort Construction

Patients were selected from The Surveillance, Epidemiology and End Results (SEER) program was used to identify a cohort of 51,354 male patients who were diagnosed with prostate cancer between 2010 and 2015. Of these patients, inclusion and exclusion criteria were applied, resulting in a cohort of 10,303 patients to be used in the machine learning model.

Inclusion criteria: Patients were included if they had the following associated ICD-9 codes for prostate cancer: C619. The patient is required to have a group stage of III or IV. Furthermore, patients are included if they have a date of diagnosis between 2010 and 2015.

Exclusion criteria: Patients are excluded if they did not have follow-up information following their initial diagnosis.

Table 1 shows descriptive statistics of the cohort.

	Survival < 21 mos	**Survival >= 21 mos**	**Total**
Characteristic	*n*= 2,576	*n* = 7,668	*N*= 10,244
Age (mean)	66.6	64.3	64.8
Sex (% male)	100%	100%	100%
Race (%)			
American Indian/Asian	7.5%	7.2%	7.3%
Black	15.3%	13.3%	13.8%
White	76.3%	79.0%	78.3%
Other	1.0%	0.55%	0.67%

Group Stage (%)			
III	46.9%	68.5%	63.1%
IV	53.0%	46.5%	36.9%
Tumor Size (mean, cm)	2.3cm	3.0cm	2.84cm
Gleason Score (mean)	4.34	4.10	4.16
Histology (%)			
Unspecified	0%	0.03%	0.02%
Epithelial neoplasm	0.85%	0.09%	0.28%
Acinar cell neoplasm	0.39%	0.15%	0.21%
Complex epithelial neoplasm	0.31%	0.04%	0.11%
Adenomas and adenocarcinomas	98.06%	99.1%	98.8%
Cystic, mucinous, serous neoplasm	0.04%	0.08%	0.07%
Ductal and lobular neoplasms	0.03%	0.53%	0.49%

Table 1: Baseline characteristics of the study cohort.

Features corresponding to several domains were extracted for the cohort. These include demographics, AJCC staging criteria, metastatic sites, and prostate-specific criteria including Gleason score and biopsy-related findings. Table 2 shows the description of features used in the model.

Feature	**No. of features**	**Example**	**Aggregation**
Age	1	76.4	Continuous
Race (White, Black, American Indian/Asian, Other)	3	Black	Categorical
Race - Hispanic	1	Hispanic	Categorical
Group Stage	2	III	Categorical
AJCC staging (T,N,M)	19	N0	Categorical
Gleason Score	2	Score on Needle Core Biopsy	Continuous

Histology	7	Epithelial neoplasm	Categorical
Percentage of positive cores on biopsy	1	30%	Continuous
Metastatic sites	4	Metastasis to Bone	Categorical
Tumor Size	1	1.5 cm	Continuous
Lymph Node Involvement[1]	8	Regional lymph nodes removed for examination with pre-surgical systemic treatment or radiation, but lymph node evaluation based on clinical evidence.	Categorical
AJCC Metastasis Eval [2]	7	Meets criteria for AJCC pathologic staging of distant metastasi	Categorical

Table 2: Features used in the model, as well as aggregation used in data preprocessing before model training.

Kaplan Meier Analysis

A Kaplan Meier[14] analysis was performed with all patients in the cohort. The 25th percentile of overall survival time in months, was used as a threshold to determine class labels of patients:

0: patient survived at least the 25th percentile of overall survival
1: patient death occurred before the 25th percentile of overall survival

Machine Learning Model

Feature Construction

Categorical features were one-hot encoded into separate features. For example, the feature *race* which includes categories (*white, black, other*) would be one hot encoded for a patient as (1,0,0) for *white*, (0,1,0) for *black*, and (0,0,1) for *other*. Continuous features were used in the current form. Standardization was performed on all features.

[1] See NCI SEER definition: https://training.seer.cancer.gov/schema/rp_ureter/reg_ln_eval.html
[2] See NCI SEER definition: https://staging.seer.cancer.gov/cs/input/02.05.50/prostate/mets_eval/?breadcrumbs=(~schema_list~),(~view_schema~,~prostate~)

Predictive Modeling

Principal component analysis was performed on the cohort to reduce dimensionality. Feature selection was performed using ANOVA F-value for the. A logistic regression[15] model was trained using the features constructed, with the target labels

0: patient survived at least the 25th percentile of overall survival

1: patient death occurred before the 25th percentile of overall survival.

Grid search was performed to learn the most optimal set of modeling parameters from the following set: number of features {all}, regularization {*l1, l2*}, C {1e-2, 1e-1, 1, 1e1, 1e2}. The model was evaluated via 5-fold cross validation. The scikit-learn [16] Python package was used to implement the analysis.

Results

Kaplan Meier Analysis

The median overall survival was 37 months. The 25th percentile of survival time was 21 months, which was used in the definition of patient classes for the machine learning problem (Figure 1).

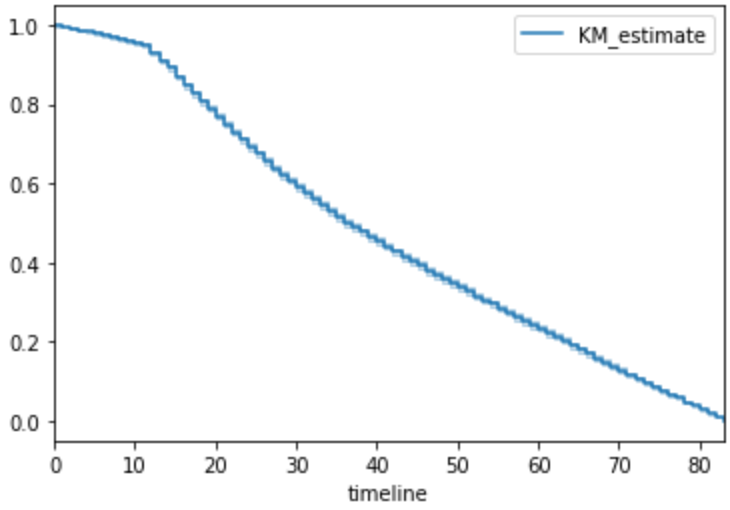


Figure 1: Kaplan Meier survival curve of all patients in the analytical cohort.

Principal Component Analysis

We utilized principal components analysis (PCA) to visualize the variation in the cohort of patients. Figure 2 shows the patients projected onto the first two principal components.

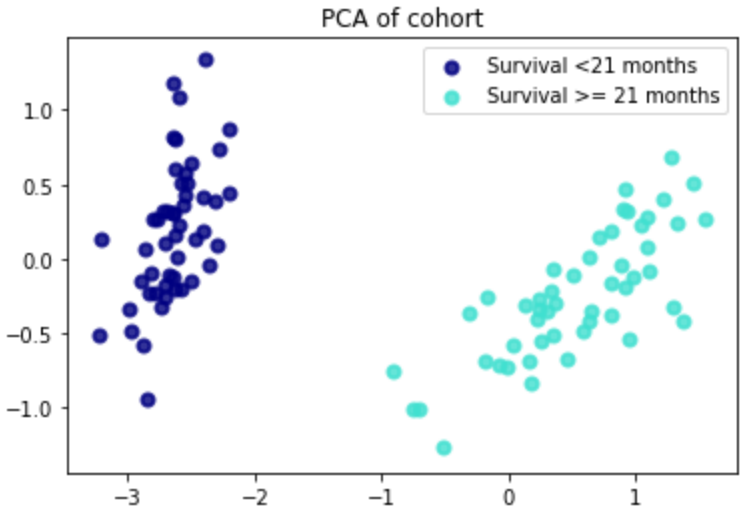


Figure 2: PCA plot of patients. Patients surviving less than XX years are shown in red; otherwise green.

Predictive Model Performance Metrics

Across all folds of cross validation, the model achieved AUC of 0.649, accuracy of 0.752, precision of 0.623, recall of 0.051, and F1 score 0.094. The receiver operating curve is shown in figure 3.

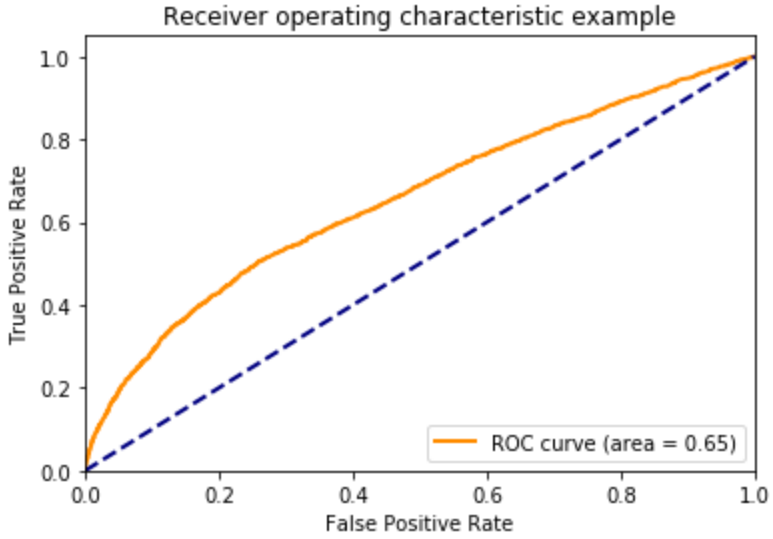
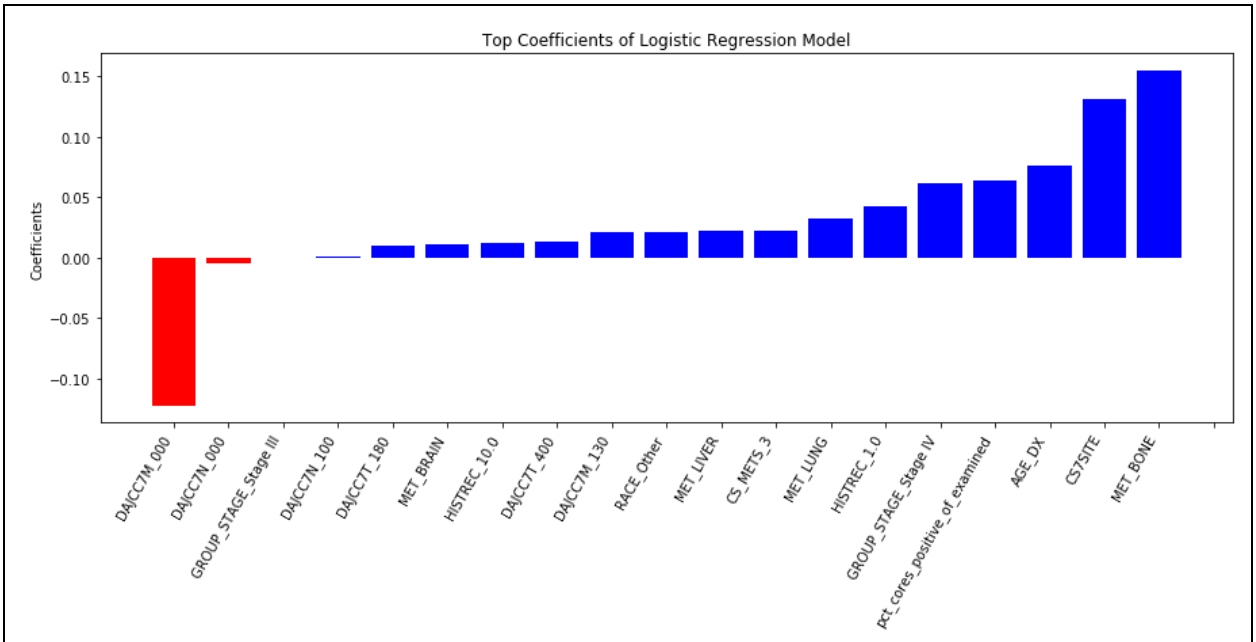


FIgure 3: Receiver operating curve for the logistic regression model.

Feature Importance

Of the 19 most predictive features with non-zero weights learned from the logistic regression model, all of them conveyed clinical meaningfulness in the application of early mortality prediction. Metastases to the bone, lung, liver and brain were positively predictive of early mortality, with bone being the metastatic site with strongest clinical predictiveness. Features are visualized in terms of predicted weights in Figure 4.



Label	Feature Description	Weight
MET_BONE	Metastasis to bone	0.15517728
CS7SITE	Gleason score	0.1316249
AGE_DX	Age at diagnosis	0.07634129
pct_cores_positive_of_examined	Percentage of positive cores of all examined cores in biopsy	0.06389444
GROUP_STAGE_Stage IV	Group stage IV	0.06112921
HISTREC_1.0	Histology: epithelial neoplasms	0.04209245
MET_LUNG	Metastasis to lung	0.03206458
CS_METS_3	Specimen from metastatic site microscopically positive	0.02235561
MET_LIVER	Metastasis to liver	0.0218343
RACE_Other	Race, not white or black	0.02173258
DAJCC7M_130	Staging: M1c	0.02170331
DAJCC7T_400	Staging: T4	0.01364619
HISTREC_10.0	Histology: acinar cell neoplasms	0.01211546
MET_BRAIN	Metastasis to brain	0.01115691
DAJCC7T_180	Staging: T1c	0.01021806
DAJCC7N_100	Staging: N1	0.00047948
GROUP_STAGE_Stage III	Group stage III	-7.08E-05

DAJCC7N_000	Staging: N0	-0.0044827
DAJCC7M_000	Staging: M0	-0.1220388

Figure 4: top most predictive features, including all features with a positive weight or negative weight learned from the model. Positive weights indicate positive correlation with early mortality, while negative weights indicate negative correlation with early mortality.

Discussion

In this study, we built a machine learning model to predict early mortality (less than 21 months, the median overall survival of patients with stage III and IV prostate cancer).

The most predictive features were arrived at via extraction of weights learned from the logistic regression model. It is important to note that the most predictive features were determined from feature coefficients in the model. In our logistic regression model, feature importances were interpreted based on coefficients learned from the model. In clinical applications, interpretability is important for downstream usage of the model in personalized treatment plans for patients. Methods such as LIME [17] have been successfully used in healthcare applications including prediction of mortality in ICU patients [18].

It is important to note that there are methods for identifying risk factors as correlated to mortality, such as Cox proportional hazards model. We did not implement a Cox model in this scenario because the problem was setup as a prediction problem.

There are three main areas for future work. First, the incorporation of additional features such as medications, comorbidities, behavioral factors (e.g., ECOG), genomics, insurance type, health system encounters (e.g., inpatient encounters, hospitalizations) should be explored. Prior work on machine learning for early detection of disease has shown that inpatient encounters have been strongly predictive of early mortality. [5] While the SEER database does not include such features, a study could be performed from electronic health records (EHRs) as a data source.

The second area of future work involves leveraging temporal models that take into account the relative temporal relationship between occurrences of features. Models that take into account temporal relationships between features include autoencoders for learning phenotypic features from temporal features such as labs[19], as well as recurrent neural network models which learn predictive representations that capture temporal relationships between features[20].

The third area of future work involves validation of the model trained on SEER data, in other data sources. Transfer learning methods can be employed. There are various ongoing efforts of standardization of EHR data models such as Health Level 7 (HL7) standard Fast Healthcare Interoperability Resources (FHIR) [21]and Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers (OHDSI)[22], which will enable improved reproducibility of scientific research.

Conclusion

We developed a machine learning model that predicts early mortality in prostate cancer patients with 75.2% accuracy with AUC 0.649, and is able to identify clinical features predictive of early mortality.

Future work should include integration of additional data sources, as well as explore temporal modeling strategies to account for clinical changes over time.

References

1. Grönberg H. Prostate cancer epidemiology. Lancet. 2003;361: 859–864. doi:10.1016/S0140-6736(03)12713-4

2. Surveilance E. End Results (SEER) Program. Cancer Stat Fact Sheet: Acute Myeloid Leulemia National Institutes of Health, National Cancer Institute. 2016.

3. Joniau S, Briganti A, Gontero P, Gandaglia G, Tosco L, Fieuws S, et al. Stratification of high-risk prostate cancer into prognostic categories: a European multi-institutional study. Eur Urol. 2015;67: 157–164. doi:10.1016/j.eururo.2014.01.020

4. Dhanasekaran SM, Barrette TR, Ghosh D, Shah R, Varambally S, Kurachi K, et al. Delineation of prognostic biomarkers in prostate cancer. Nature. 2001;412: 822–826. doi:10.1038/35090585

5. Ng K, Steinhubl SR, deFilippi C, Dey S, Stewart WF. Early Detection of Heart Failure Using Electronic Health Records: Practical Implications for Time Before Diagnosis, Data Diversity, Data Quantity, and Data Density. Circ Cardiovasc Qual Outcomes. 2016;9: 649–658. doi:10.1161/CIRCOUTCOMES.116.002797

6. Chen R, Stewart WF, Sun J, Ng K, Yan X. Recurrent Neural Networks for Early Detection of Heart Failure From Longitudinal Electronic Health Record Data: Implications for Temporal Modeling With Respect to Time Before Diagnosis, Data Density, Data Quantity, and Data Type. Circ Cardiovasc Qual Outcomes. 2019;12: e005114. Available: https://www.ahajournals.org/doi/abs/10.1161/CIRCOUTCOMES.118.005114

7. Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. J Am Med Inform Assoc. 2017;24: 361–370. doi:10.1093/jamia/ocw112

8. Kallenberger SM, Schmidt C. Forecasting the development of acute kidney injury using a recurrent neural network. Cardiovasc Res. 2019;115: e155–e157. doi:10.1093/cvr/cvz279

9. Chen R, Su H, Khalilia M, Lin S, Peng Y, Davis T, et al. Cloud-based Predictive Modeling System and its Application to Asthma Readmission Prediction. AMIA Annu Symp Proc. 2015;2015: 406–415. Available: https://www.ncbi.nlm.nih.gov/pubmed/26958172

10. Desautels T, Das R, Calvert J, Trivedi M, Summers C, Wales DJ, et al. Prediction of early unplanned intensive care unit readmission in a UK tertiary care hospital: a cross-sectional machine learning approach. BMJ Open. 2017;7: e017199. doi:10.1136/bmjopen-2017-017199

11. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med. 2018;1: 18. doi:10.1038/s41746-018-0029-1

12. Cheng F, Zhao Z. Machine learning-based prediction of drug–drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties. J Am Med Inform Assoc. 2014;21: e278–e286. doi:10.1136/amiajnl-2013-002512

13. Hankey BF, Ries LA, Edwards BK. The surveillance, epidemiology, and end results program: a national resource. Cancer Epidemiol Biomarkers Prev. 1999;8: 1117–1121. Available:

https://www.ncbi.nlm.nih.gov/pubmed/10613347

14. Kaplan EL, Meier P. Nonparametric Estimation from Incomplete Observations. J Am Stat Assoc. 1958;53: 457–481. doi:10.1080/01621459.1958.10501452

15. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. Springer Science & Business Media; 2009. Available: https://play.google.com/store/books/details?id=tVIjmNS3Ob8C

16. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12: 2825–2830. Available: http://www.jmlr.org/papers/v12/pedregosa11a.html

17. Ribeiro MT, Singh S, Guestrin C. " Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM. 2016. Available: https://dl.acm.org/doi/abs/10.1145/2939672.2939778

18. Katuwal GJ, Chen R. Machine Learning Model Interpretability for Precision Medicine. arXiv [q-bio.QM]. 2016. Available: http://arxiv.org/abs/1610.09045

19. Lasko TA, Denny JC, Levy MA. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. PLoS One. 2013;8: e66341. doi:10.1371/journal.pone.0066341

20. Ma F, Chitta R, Zhou J, You Q, Sun T, Gao J. Dipole: Diagnosis Prediction in Healthcare via Attention-based Bidirectional Recurrent Neural Networks. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: Association for Computing Machinery; 2017. pp. 1903–1911. doi:10.1145/3097983.3098088

21. Bender D, Sartipi K. HL7 FHIR: An Agile and RESTful approach to healthcare information exchange. Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems. ieeexplore.ieee.org; 2013. pp. 326–331. doi:10.1109/CBMS.2013.6627810

22. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform. 2015;216: 574–578. Available: https://www.ncbi.nlm.nih.gov/pubmed/26262116