Informatics and Technology in Clinical Care and Public Health
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210860

Machine Learning Approaches for Early Prostate Cancer Prediction Based on Healthcare Utilization Patterns

Joseph FINKELSTEIN [1], Wanting CUI, Tiphaine C. MARTIN and Ramon PARSONS
Icahn School of Medicine at Mount Sinai, New York, NY, USA

Abstract. The goal of this study was to build a machine learning model for early prostate cancer prediction based on healthcare utilization patterns. We examined the frequency and pattern changes of healthcare utilization in 2916 prostate cancer patients 3 years prior to their prostate cancer diagnoses and explored several supervised machine learning techniques to predict possible prostate cancer diagnosis. Analysis of patients' medical activities between 1 year and 2 years prior to their prostate cancer diagnoses using XGBoost model provided the best prediction accuracy with high F1 score (0.9) and AUC score (0.73). These pilot results indicated that application of machine learning to healthcare utilization patterns may result in early identification of prostate cancer diagnosis.

Keywords. Prostate Cancer, Big Data Analytics, Machine Learning

1. Introduction

Early discovery of cancer has crucial ramifications on the disease prognosis. In the recent meta-analysis of seven major cancers, even a 4-week delay in cancer treatment was associated with increased mortality across systemic treatment, surgical and radiotherapy indications [1]. A number of observational studies reported differing patterns in healthcare utilization before and after cancer diagnosis [2-3]. Significant differences were found in healthcare utilization patterns preceding cancer diagnosis as compared to matched non-cancer patients [4] including patients with prostate cancer [5]. Previous studies demonstrated utility of claims-based approaches for building predictive models in prostate cancer [6]. Based on these resent reports, we hypothesized that healthcare consumption preceding a diagnosis of prostate cancer may exhibit specific patterns which can be used for early cancer prediction. As machine learning approaches have been shown to be particularly instrumental in identifying characteristic data patterns in building predictive models, supervised machine learning techniques have been employed in this project. Thus, the goal of this pilot study was to analyze the frequency and pattern changes of patients' medical activities 3 years prior to their prostate cancer diagnoses, and to build machine learning models based on their medical activities to predict possible prostate cancer diagnosis within the near future.

[1] Joseph Finkelstein, Icahn School of Medicine at Mount Sinai, 1425 Madison Ave, New York, NY, USA, 10029; E-mail: Joseph.Finkelstein@mssm.edu.

2. Method

A de-identified analytical dataset has been constructed from electronic health record at Mount Sinai Health System in New York City comprising patients who were diagnosed with prostate cancer between 01/2009 and 12/2019. Since we aimed to monitor patients' medical activity frequency and pattern prior to their prostate cancer diagnoses, we extracted these patients' medical activities such as medical procedures, lab tests and radiology, 3 years prior to their cancer diagnoses. We only included patients who have at least 3 procedures or tests.

In predictive modeling, we used 84 predictive variables, which spanned through 3 groups: lab test, radiology and procedures representing the entire spectrum of care utilization. All predictive variables were continuous, and the value indicated the number of times a patient had done a test. XGBoost model was used to select most informative features out of the initial 187 types of lab tests. We included all features where the feature importance was over 0.05. In the end, 67 variables were selected. We also summarized lab tests into 4 parent groups using standard Logical Observation Identifier Names and Codes (LOINC) codes. The 4 parents groups are: Chem_drug_tox_chal_sero_allergy, urine, MassMolConc and CellDiffCoun. There were 5 variables related to radiology: X-Ray, MRI, CT, PET/CT and ultrasound. In addition, we also categorized all procedures into 8 groups using Current Procedural Terminology (CPT) codes. The procedure groups we included were: chemistry procedures, hematology procedures, organ disease panel, urinalysis procedures, immunology procedures, microbiology procedures, therapeutic drug assays and cardiovascular procedures.

We constructed 4 datasets with different time cut off points for target variables. The target variable was binary with indication of cancer. In the first dataset, we defined lab tests or procedures performed greater than 2 years and less than 3 years prior to diagnoses, as the timeframe for no cancer. We defined tests or procedures performed within 1 year of cancer diagnoses as time period with cancer. In the second dataset, we defined lab tests or procedures performed greater than 1 year and less than 3 years prior to diagnoses, as the timeframe for no cancer; and tests or procedures performed within 1 year of cancer diagnoses as time period with cancer. In the third dataset, we defined lab tests or procedures performed greater than 2 year and less than 3 years prior to diagnoses, as the timeframe for no cancer; and tests or procedures performed within 2 years of cancer diagnoses as time period with cancer. In the last dataset, we defined lab tests or procedures performed greater than 1 year and less than 2 years prior to diagnoses, as the timeframe for no cancer; and tests or procedures performed within 1 year of cancer diagnoses as time period with cancer.

In model training, each dataset was randomly divided into 80% training and 20% testing. We compared 3 machine learning models: Support Vector Machine (SVM), Random Forest (RF) and XGBoost (XGB). We tuned hyper parameters and performed 3-fold cross validation to choose the best hyper parameters using the training set. In the end, we applied the best-tuned model for each algorithm to the testing set and calculated accuracy, precision, recall, F1 score and area under the curve (AUC) accordingly.

3. Results

There are 2916 number of records in the first dataset; 724 patients had records of lab tests and procedures that are greater than 2 years and less than 3 years prior to the

diagnoses; and 2174 patients had records of tests and procedures within 1 year of prostate cancer diagnoses. XGBoost model performed the best (Table 1), since it has the highest F1 score (0.9) and AUC score (0.73).

In the second dataset, 1108 patients had medical activities greater than 1 years and less than 3 years prior to the prostate cancer diagnoses. In contrast 2159 patients had medical activities within 1 year of cancer diagnoses. Both XGBoost model and random forest model performed well in this subset (Table 1), as they both have high AUC score (0.69) and F1 score (0.82).

In the third dataset, 742 patients had medical activities greater than 2 years and less than 3 years prior to their cancer diagnoses, and 2196 patients had medical activities within 2 years of their diagnoses. According to Table 1, although SVM model produced the highest AUC score (0.74), the F1 score (0.83) and recall (0.77) were both low, compared to the other 2 models (F1 score = 0.88, recall 0.92).

In the fourth dataset, 918 patients had medical activities greater than 1 year and less than 2 years prior to their diagnoses, and 2174 patients had medical activities within 1 year of their cancer diagnoses. All 3 models produced the AUC score (0.68), with random forest generated the highest F1 score (0.86).

Table 1. Results of predictive models.

	Accuracy	Precision	Recall	F1	AUC
Dataset 1					
XGB	0.84	0.87	0.93	0.90	0.73
SVM	0.72	0.89	0.73	0.80	0.72
RF	0.83	0.85	0.95	0.89	0.70
Dataset 2					
XGB	0.75	0.80	0.84	0.82	0.69
SVM	0.67	0.83	0.65	0.73	0.68
RF	0.75	0.80	0.84	0.82	0.68
Dataset 3					
XGB	0.8	0.84	0.92	0.88	0.69
SVM	0.76	0.89	0.77	0.83	0.74
RF	0.8	0.84	0.92	0.88	0.69
Dataset 4					
XGB	0.77	0.80	0.90	0.85	0.68
SVM	0.69	0.83	0.70	0.76	0.68
RF	0.79	0.80	0.94	0.86	0.68

4. Discussion

Overall, models in the first dataset has the highest AUC scores, and models in the fourth dataset had the lowest AUC scores. The time cut off point for the first dataset was within 1 year for cancer and greater than 2 years and less than 3 years prior to the diagnosis for no cancer in this time period. In contrast the time cut off point for the last dataset was within 1 year for cancer and greater than 1 year and less than 2 years prior to the diagnosis for no cancer. Thus, patterns of patients' medical activities were more likely to change

between 1 year and 2 years prior to their prostate cancer diagnoses. In addition, the frequency of patients using medical services increased significantly when closer to the diagnoses.

XGBoost model performed well for all four datasets. Although the SVM model from the third dataset had the highest AUC score, this model produced relatively low F1 score and recall. Since we aimed to find the best overall model, we selected the XGBoost model from the first dataset. It has the highest F1 score (0.9) and second highest AUC score (0.73) among all models. By examining the feature importance of this model, we found that PSA (Prostate-Specific Antigen) test, microbiology procedures, chemistry procedures, organ disease panel and aPTT (activated partial thromboplastin time) blood test were the top 5 factors, which indicates that a change of these tests and procedures' frequency and patterns was highly associated with possible prostate cancer diagnoses within 1 or 2 years.

In future studies, we plan to optimize our predictive features. We will explore various methodologies to summarize LOINC codes and CPT codes. And we will also examine medication intake prior to cancer diagnoses. In addition, we will expand our studies to 5 years prior to patients' prostate cancer diagnoses and explore various time cut off points in relationship to the pattern of patients' medical activities.

5. Conclusion

In this study, we examined the frequency and pattern changes of patients' medical activities 3 years prior to their prostate cancer diagnoses and built machine learning models based on their medical activities to predict possible prostate cancer diagnosis within the near future. XGBoost model from the first dataset performed the best, with high F1 score (0.9) and AUC score (0.73). Frequency and patterns of patients' medical activities would change between 1 year and 2 years prior to their prostate cancer diagnoses. The results indicated that further exploration of this approach is warranted.

Reference

[1] Hanna TP, King WD, Thibodeau S, Jalink M, Paulin GA, Harvey-Jones E, O'Sullivan DE, Booth CM, Sullivan R, Aggarwal A. Mortality due to cancer treatment delay: systematic review and meta-analysis. BMJ. 2020;371:m4087.

[2] Shen C, Dasari A, Xu Y, Zhou S, Gu D, Chu Y, Halperin DM, Shih YT, Yao JC. Pre-existing Symptoms and Healthcare Utilization Prior to Diagnosis of Neuroendocrine Tumors: A SEER-Medicare Database Study. Sci Rep. 2018;8(1):16863.

[3] Jones LE, Doebbeling CC. Primary care utilization patterns before and after lung cancer diagnosis. Eur J of Cancer Care 2009;18:165-173.

[4] Park J, Look KA. Health Care Expenditure Burden of Cancer Care in the United States. Inquiry. 2019;56:46958019880696.

[5] Sun M, Marchese M, Friedlander DF, et al. Health care spending in prostate cancer: An assessment of characteristics and health care utilization of high resource-patients. Urol Oncol. 2021 Feb;39(2):130.e17-130.e24.

[6] Riviere P, Tokeshi C, Hou J, Nalawade V, Sarkar R, Paravati AJ, Schiaffino M, Rose B, Xu R, Murphy JD. Claims-Based Approach to Predict Cause-Specific Survival in Men with Prostate Cancer. JCO Clin Cancer Inform. 2019;3:1-7.