Envisioning the Future of Health Informatics and Digital Health
J. Mantas et al. (Eds.)

doi:10.3233/SHTI250037

Predicting Prostate Cancer Diagnosis Using Machine Learning Analysis of Healthcare Utilization Patterns

Wanting CUI[a,1], Ahmad HALWANI[a,b], Chunyang LI[a] and Joseph FINKELSTEIN[a]
[a] *University of Utah, Salt Lake City, Utah, USA*
[b] *Salt Lake City Veterans Affairs Medical Center*

Abstract. This study investigated healthcare utilization patterns prior to prostate cancer diagnoses, aiming to develop machine learning models for early prediction of cancer diagnosis. Data from the All of Us Research Program was used, focusing on adult patients diagnosed with prostate cancer between 2010 and 2019. Key variables were derived from procedure, measurements, and condition records, including PSA values, comorbidity index, and symptoms. Multiple machine learning models were tested to predict prostate cancer 3, 6, 9, and 12 months ahead of time. The dataset included 1,276 cancer patients and 1,232 non-cancer patients. The XGBoost model performed best at 3 months, achieving an accuracy and F1 score of 0.73 and an AUC of 0.82. At 6 months, the model had an accuracy and F1 score of 0.71 and an AUC of 0.78. Performance declined with longer prediction windows. PSA values were consistently the most important predictor across all timeframes, along with other factors like triglyceride and creatinine levels.

Keywords. Prostate cancer prediction, Machine Learning, Real-world data, healthcare utilization pattern

1. Introduction

Early cancer detection plays a critical role in improving patient prognosis, as even short delays in treatment can significantly increase mortality and limit treatment options, especially in high-risk and aggressive cancers. A four-week delay in cancer treatment can raise mortality rates by 6-8% for surgery and 9-13% for certain radiotherapies and systemic treatments [1]. These risks rise further with delays of eight and twelve weeks, emphasizing the importance of early diagnosis and timely treatment.

Observational studies have highlighted differences in healthcare utilization between cancer patients and non-cancer patients. A study using the SEER-Medicare database revealed that, in the 12 months before diagnosis, cancer patients had more outpatient visits, twice the emergency room admissions, and 10% higher hospitalization rates than non-cancer controls [2]. Additionally, healthcare expenditures for cancer patients, particularly those with prostate cancer, were significantly higher, with prostate cancer patients averaging annual costs over $15,000, largely driven by hospital services and ambulatory care visits [3].

[1] Corresponding Author: Wanting Cui, wanting.cui@utah.edu.

In recent years, machine learning (ML) and real-world data (RWD) have increasingly been applied to enhance early cancer detection and improve cancer care management [4-8]. ML has been used for screening, diagnosis, prognosis, and treatment selection. Traditional models like lasso regression and decision trees, along with neural networks, have been employed successfully [4]. By analyzing vast datasets, ML models can uncover subtle patterns in patient behaviors and healthcare utilization that may not be apparent through traditional methods. Such insights can inform personalized care strategies, optimize resource allocation, and reduce healthcare costs by identifying patients at higher risk earlier. Furthermore, ML has the capacity to integrate complex data types—such as electronic health records (EHR), genomic data, and medical notes—further improving diagnostic accuracy and treatment personalization for conditions like prostate cancer. For example, a time-series neural network model achieved high accuracy in predicting pancreatic cancer three months before diagnosis [5].

Inspired by prior research, this study aimed to explore healthcare utilization patterns in the three years before a prostate cancer diagnosis, with the goal of building machine learning models for early prostate cancer prediction.

2. Methods

The dataset was extracted from the All of Us Research Program [9]. Prostate cancer patients were first identified. The inclusion criteria were patients diagnosed with 'primary malignant neoplasm of prostate (SNOMED = 93974005) between 2010 and 2019. Patients with missing age, gender, or prior cancer diagnoses were excluded. Medical activities and lab tests within 3 years prior to diagnosis were gathered, and only patients with 2 or more activities were included. A cohort of non-cancer male patients was also selected, with an arbitrary prediction date, age-matched to the cancer group. Similarly, only non-cancer patients with 2 or more activities were included for comparability.

Multiple machine learning models were developed to predict whether a patient has prostate cancer at various time intervals. Extreme Gradient Boosting (XGBoost) was selected as the primary machine learning model due to its superior performance in previous studies [10]. XGBoost is an advanced implementation of Gradient Boosted Decision Trees (GBDT), using an ensemble learning method that aggregates the predictions from multiple decision trees to deliver a more accurate prediction. This technique uses the concept of boosting, where a series of decision trees are trained in succession. Each tree focuses on the errors or residuals left by its predecessors and seeks to minimize these through gradient descent, which will increase the predictive accuracy of the ensemble. In addition, other predictive models such as random forest, SVM, and neural networks were tested to compare the predictive results.

Predictive variables were generated by analyzing data from procedures, measurements, and condition tables. CPT4-coded procedures were grouped using the Clinical Classifications Software (CCS), and surgical procedures were flagged as narrow, broad, or neither [11]. Two additional variable sets were created using AAPC criteria [12]. The first set focused on evaluation and management services, covering outpatient, inpatient, emergency care, and other services. The second set categorized new patient visits based on their duration, dividing them into four groups: visits under 30 minutes, 30-44 minutes, 45-59 minutes, and 60-74 minutes, which provided a more detailed breakdown of patient interactions. The number of occurrences within each timeframe was calculated, with only categories having more than 5 occurrences included. Lab tests were identified using LOINC codes, grouped into 80 subgroups and 5 parent categories. Abnormal lab test results for key markers such as hemoglobin, white blood cell count, platelets, AST, ALT, CRP, calcium, and creatinine were tracked, and their frequencies calculated. PSA test results were summarized into four key variables: most recent, maximum, minimum, and mean difference. Although some variables may be correlated, XGBoost is effective at mitigating issues related to highly correlated variables. When constructing each individual tree, XGBoost randomly samples a subset of features (variables) instead of using all variables. This feature subsampling method reduces the likelihood that any single correlated variable dominates the splitting criteria across all trees, which reduces the impact of multicollinearity.

The Charlson comorbidity index was computed for each patient. Genitourinary symptoms were captured using ICD codes (R30-R39), with variables created for hematuria, urinary retention, and total symptom count. The number of occurrences of each variable within the predictive window was calculated for all patients.

The target variable was binary, indicating a positive cancer diagnosis. Four predictive datasets were constructed for 3, 6, 9, and 12 months before diagnosis, based on medical activities within specific timeframes. For the 3-month prediction, activities from 4 to 15 months prior to diagnosis were analyzed, while for the 6-, 9-, and 12-month predictions, data from the respective 12 months prior to each prediction window were used. The dataset was split into 70% for training and 30% for testing. Parameter tuning and 5-fold cross-validation were performed to optimize the models. Evaluation metrics included accuracy, F1 score, and ROC AUC score. The best models for each timeframe were then applied to the testing set, and their performance was assessed. All analyses were conducted in Python 3.8 within Anaconda Jupyter Notebook.

Table 1. Demographics of cancer and non-cancer patients.

	Cancer (n = 1276)		**Non-Cancer (n = 1232)**	
	Mean	**Std**	**Mean**	**Std**
Age	66.67	8.03	65.67	10.59
Comorbidity Index	3.46	2.11	3.87	2.61
	Count	**Percent**	**Count**	**Percent**
Race				
White	1005	78.76%	927	75.24%
Black	235	18.42%	231	18.75%
Other	36	2.82%	74	6.01%

3. Results

There were 1276 prostate cancer patients and 1232 non-cancer patients in the analytic dataset. These patients generated over 78,000 procedure records, over 500,000 measurement records, and 250,000 condition records. The average age of these patients was 66.67 (±8.03) years old, which was similar to that of the non-cancer patients (65.57±10.59). The average comorbidity index at the prediction date for the cancer group was 3.46±2.11, whereas the comorbidity index for the non-cancer group was 3.87±2.61. In terms of race, the majority of patients in both groups were White: 78% in the cancer group and 75% in the non-cancer group. And around 18% of patients were black or African American in both groups. In addition, all patient's genders were male.

In machine learning, XGBoost was the best-performing model as it showed consistently high performance across all timeframes and the fastest training time. Table 2 shows the performance of the XGBoost model in predicting prostate cancer at different time intervals before diagnosis (3, 6, 9, and 12 months ahead). The model's accuracy, F1 score, and AUC values declined as the predictive timeframe increased. For predictions made 3 months before diagnosis, the model achieved its best performance with an accuracy of 0.73, F1 score of 0.73, and AUC of 0.82. At 6 months ahead, the performance decreased slightly, with an accuracy of 0.71, F1 score of 0.71, and AUC of 0.78. For predictions 9 months before diagnosis, the accuracy and F1 score dropped to 0.66, and the AUC fell to 0.73. At 12 months ahead, the model's performance was the lowest, with an accuracy of 0.67, F1 score of 0.68, and AUC of 0.71.

Table 2. XGBoost model performances for all timeframes.

Time Periods	Accuracy	F1	AUC
3 months ahead	0.73	0.73	0.82
6 months ahead	0.71	0.71	0.78
9 months ahead	0.66	0.66	0.73
12 months ahead	0.67	0.68	0.71

We further analyzed the important variables generated by the XGBoost model. Across all timeframes, PSA test values, both the most recent and maximum PSA values, were consistently ranked as the most important variables. In the 3- and 6-month windows, other significant factors include routine chest X-rays, levels of sodium, urea nitrogen, and various lab measurements (e.g., creatinine, triglycerides, calcium). At the 9- and 12-month intervals, PSA values remained key predictors, but other variables such as epithelial cell presence in urine, nonoperative urinary system measurements, and blood levels of bicarbonate, phosphate, and creatinine showed importance.

4. Discussion

The study focused on predicting prostate cancer diagnoses using machine learning models, with the best performance seen 3 months before diagnosis (accuracy and F1 score of 0.73, AUC of 0.82). Performance decreased as the prediction window increased, with 6-month predictions also showing promise (accuracy 0.71, AUC 0.78).

The demographics of both cancer and non-cancer patients showed that they were older adults with several underlying health conditions. PSA values consistently emerged as key predictors across all timeframes, along with variables like triglyceride, sodium, creatinine, and routine chest X-rays. These findings showed the importance of both cancer-specific biomarkers and broader health indicators. Our results corroborate previous reports demonstrating the potential value of EHR and RWD for precision medicine [13-15] and the generation of real-world evidence [16-18]. The study had limitations, such as not using time series analysis and lacking symptoms and detailed medical history. Future research will explore time series modeling, additional coding systems, and medication information. In addition, we will use large language models to extract symptoms and other useful healthcare utilization patterns from unstructured notes to improve prediction accuracy further.

5. Conclusion

In conclusion, this study showed the potential of machine learning to monitor healthcare utilization patterns and enhance early detection of prostate cancer. The models predicting cancer 3 and 6 months ahead showed good performance. This showed the value of predictive tools in improving diagnostic efficiency and patient care. A major opportunity for enhancing these models is to integrate large language models (LLMs). By incorporating LLMs, future models could analyze unstructured medical data, such as clinical notes and physician narratives, which often contain vital information unavailable in structured datasets. This would enable a more comprehensive understanding of patient symptoms, family history, race, and other risk factors, uncovering hidden patterns and improving the models' predictive power. LLMs could also enhance feature extraction, making the models more sensitive to subtle indicators of disease progression. Thus, additional research using LLMs is warranted.

References

[1] Hanna TP, King WD, Thibodeau S, et al. Mortality due to cancer treatment delay: systematic review and meta-analysis. BMJ. 2020;371:m4087.
[2] Shen C, Dasari A, Xu Y, et al. Pre-existing Symptoms and Healthcare Utilization Prior to Diagnosis of Neuroendocrine Tumors: A SEER-Medicare Database Study. Sci Rep. 2018;8(1):16863.
[3] Park J, Look KA. Health Care Expenditure Burden of Cancer Care in the United States. Inquiry. 2019;56:46958019880696.
[4] Matthew Nagy et al., Machine Learning in Oncology: What Should Clinicians Know?. JCO Clin Cancer Inform 2020;4:799-810.
[5] Placido D, Yuan B, Hjaltelin JX, et al. A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories. Nat Med. 2023;29(5):1113-1122.
[6] Huo X, Finkelstein J. Prostate cancer prediction using classification algorithms. Journal of Clinical Oncology 2022;40(16_suppl):e13590.
[7] Adrianna Janik et al., Machine Learning–Assisted Recurrence Prediction for Patients with Early-Stage Non–Small-Cell Lung Cancer. JCO Clin Cancer Inform 2023;7:e2200062.
[8] Riviere P, Tokeshi C, Hou J, et al. Claims-Based Approach to Predict Cause-Specific Survival in Men with Prostate Cancer. JCO Clin Cancer Inform. 2019;3:1-7.
[9] All of Us Research Program. https://allofus.nih.gov/
[10] Finkelstein J, Cui W, Martin TC, Parsons R. Machine Learning Approaches for Early Prostate Cancer Prediction Based on Healthcare Utilization Patterns. Stud Health Technol Inform. 2022;289:65-68.
[11] HCUP CCS-Services and Procedures. Healthcare Cost and Utilization Project (HCUP). May 2021. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/toolssoftware/ccs_svcsproc/ccssvcproc.jsp.
[12] Codify by AAPC. https://www.aapc.com/codes/cpt-codes-range/
[13] Cui W, Finkelstein J. Identifying Determinants of Survival Disparities in Multiple Myeloma Patients Using Electronic Health Record Data. Stud Health Technol Inform. 2024 Jan 25;310:956-960.
[14] Finkelstein J, Zhang F, Levitin SA, Cappelli D. Using big data to promote precision oral health in the context of a learning healthcare system. J Public Health Dent. 2020 Mar;80 Suppl 1(Suppl 1):S43-S58.
[15] Hickin MP, Shariff JA, Jennette PJ, Finkelstein J, Papapanou PN. Incidence and Determinants of Dental Implant Failure: A Review of Electronic Health Records in a U.S. Dental School. J Dent Educ. 2017 Oct;81(10):1233-1242. doi: 10.21815/JDE.017.080.
[16] Cui W, Finkelstein J. Impact of COVID-19 Pandemic on Use of Telemedicine Services in an Academic Medical Center. Stud Health Technol Inform. 2021 May 27;281:407-411. doi: 10.3233/SHTI210190.
[17] Shah-Mohammadi F, Cui W, Bachi K, Hurd Y, Finkelstein J. Using Natural Language Processing of Clinical Notes to Predict Outcomes of Opioid Treatment Program. Annu Int Conf IEEE Eng Med Biol Soc. 2022 Jul;2022:4415-4420. doi: 10.1109/EMBC48229.2022.9871960.
[18] Cui W, Finkelstein J. Identifying Determinants of Disparities in Lung Cancer Survival Rates from Electronic Health Record Data. Stud Health Technol Inform. 2022 May 25;294:715-716.