APAMI2020 General Oral Presentation Session | APAMI 2020 | General Oral Presentation Session
Artificial Intelligence in Healthcare and Clinical Practice (3)
Mon. Nov 23, 2020 10:30 AM - 12:00 PM Room E-1 (Congress center 5F - Conference Room 52)

[AP3-E1-2-03] Artificial Intelligence for Prostate Cancer Prediction Using Electronic Health Record Data

*Tahmina Nasrin Poly[1,2], Md. Mohaimenul Islam[1,2], Hsuan-Chia Yang[1,2], Phung-Anh Nguyen[1,2], Yu-Hsiang Wang[1,3], Yu-Chuan (Jack) Li[1,2] （1. Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taiwan, 2. International Center for Health Information Technology, Taipei Medical University, Taiwan, 3. College of Medicine, Taipei Medical University, Taiwan）
Keywords: Prostate Cancer, Deep Learning, Convolutional Neural Network, Electronic Health Record

Earlier detection of prostate cancer (PCA) patients could improve patient outcomes through the increased scope to follow-up and treatment to the patient at high risk of PCA since healthcare professionals are struggling to correctly identify and successfully treating PCA cancer. To improve patient outcomes requires the prediction of patients at high risk correctly for proper clinical decision making. We, therefore, used a deep learning technique for accurate prediction of PCA patients one year earlier with minimal features from electronic health records. We retrieved data of 4,071 PCA patients from the Taiwan National Health Insurance database (NHID) between 1999 and 2013. Patients' age, sex, comorbidities, and medication history were used to developed and test a convolutional neural network (CNN) based prediction model. The area under the receiver operating curve for the prediction of PCA was 0.94. The sensitivity and specificity of CNN were 0.87 and 0.88, respectively. In this evaluation of data with minimal features from electronic health records, CNN had high sensitivity and specificity for identifying PCA. Although accurate and earlier detection of PCA is known to be challenging, our deep learning-based prediction approach may offer great benefits for correctly stratifying the patients at high risk of PCA that enables earlier decision making for proper treatments.

Artificial Intelligence for Prostate Cancer Prediction Using Electronic Health Record Data

Tahmina Nasrin Poly[a, b], Md. Mohaimenul Islam[a, b], Hsuan-Chia Yang[a, b], Phung-Anh Nguyen[a, b], Yu-Hsiang Wang[a, c] and **Yu-Chuan (Jack) Li[a, b]**

[a] *Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taiwan*
[b] *International Center for Health Information Technology, Taipei Medical University, Taiwan*
[c] *College of Medicine, Taipei Medical University, Taiwan*

Abstract

Healthcare professionals are struggling to correctly identify and successfully treating PCA cancer. Earlier detection of prostate cancer (PCA) patients could improve patient outcomes through the increased scope to follow-up and treatment to the patient at high risk of PCA. To improve patient outcomes requires the prediction of patients at high risk correctly for proper clinical decision making. We, therefore, used a deep learning technique for accurate prediction of PCA patients one year earlier with minimal features from electronic health records. We retrieved data of 4,071 PCA patients from the Taiwan National Health Insurance database (NHID) between 1999 and 2013. Patients' age, sex, comorbidities, and medication history were used to developed and test a convolutional neural network (CNN) based prediction model. The area under the receiver operating curve for the prediction of PCA was 0.94. The sensitivity and specificity of CNN were 0.87 and 0.88, respectively. In this evaluation of data with minimal features from electronic health records, CNN had high sensitivity and specificity for identifying PCA. Although accurate and earlier detection of PCA is known to be challenging, our deep learning-based prediction approach may offer great benefits for correctly stratifying the patients at high risk of PCA that enables earlier decision making for proper treatments.

***Keywords**:*

Prostate Cancer, Deep Learning, Convolutional Neural Network, Electronic Health Record

Introduction

Prostate cancer (PCA) begins in the gland cells of the prostate and commonly diagnosed cancer in men. It is the second leading cause of death in the USA and the fifth leading cause of death worldwide [1]. The incidence rate of PCA varies worldwide; the higher and lower rate of incidence is observed in Oceania (79.1 per 100,000 people) and Asia (11.5) [2]. The incidence rate of PCA increases with age, and it reaches approximately 60% in men aged more than 60 years [3]. Earlier identification and treatment of patients with PCA help to reduce morbidity and mortality. Prostate-specific antigen (PSA) is considered as a gold standard test to screen patients for PCA. However, the performance of the PSA test is not satisfactory yet due to low specificity. It often refers to further unnecessary biopsy testing to decide whether or not the patients have PCA. Moreover, the sensitivity of the PSA test in current clinical practice is still not up to the required standard to make a fruitful decision about intervention, and it always remains nearly 50 percent with the cutoff value of 4 ng/ml [4]. Due to the low performance of the PSA test (low sensitivity and specificity), physicians are struggling to correctly identify the right patients at the right time. Although, several approaches of PSA (e.g. PSA density, PSA velocity, calculation of free PSA, and age-specific PSA reference) have already been introduced to improve the sensitivity and specificity [5-6]; these tests still failed to show the strong diagnostic performance.

Data from electronic health records (EHR) offers a unique opportunity to develop prediction models of improving patients care at the population level. We, therefore, developed a novel prediction model using the CNN algorithm with minimal features from EHR that may help to stratify high-risk patients before a diagnosis of PCA. To our knowledge, it is the first study to develop a prediction model of identifying PCA patients one year earlier.

Methods

Data Source

This study was conducted using the National Health Insurance Database (NHID) that represents over 99% of the 23 million people of Taiwan [7-8]. The database consists of patient demographics information, in- and out-patients claim data, medication history. We collected two million sample data between January 01, 1999, and December 31, 2013, to develop our prediction model. This study was approved by the institutional review board committee of Taipei Medical University.

Study Population

We included all patients if they were ≥20 years old and diagnosed with prostate cancer between January 01, 1999, and December 31, 2013. We identified PCA patients by using ICD-9 codes and records in the Registry of Catastrophic Illness Patient (RCIP), a subpart of the NHI database [9]. To be included in the RCIP database, patients must have pathological confirmation and treatment history of PCA. The index data was defined as the date of PCA diagnosis during the study period. All the patients were classified into two groups as a case (having PCA) and control (no having PCA). Patients were excluded if they were aged less than 20 years at the date of diagnosis of PCA and did not have any outpatient claims over the four-year before PCA diagnosis. Each PCA case was then matched with control by sex, age at cancer diagnosis, and the time of PCA diagnosis.

Features Selection

EHR database contains a variety of clinical features. In this study, we only considered a disease and medication history of all cases and control. However, 3-digits of ICD-9-CM codes such as 001-999, V01-V91 were considered in including all disease history. Although, ICD-9-CM code E000-E999 was not considered in this current study. Furthermore, the 4-digit of ATC code was used to retrieve all patient's medication history. A total of 1931 variables (demographic, disease and medication history, etc.) were collected to develop our prediction model.

Data Preprocessing

3-year (1095 days) patient's data were collected and we considered those patients if they visited the hospital at least three times during the 3-year. We then checked the total number of prescription and length of prescription for those patients. Afterward, 7-days clinical information of all patients was summed up continuously. A total of 157 grid was constituted of 1095-days patients' information.

Model Development

In this study, we developed a CNN algorithm for the prediction of PCA. CNN model is now widely used because of its capability of combining a variety of features and generalization with less feature engineering [10]. It always considers a very useful algorithm for analyzing more heterogeneous EHR data. A total of 1929 features (e.g. 1099 disease feature and 830 medications features) from all included participants were used to categorize into 19 groups. However, 2 convolutional layers with 32 features maps were created for each group. 1*3 size filter was applied with two max-pooling layers for reducing the number of input parameters. Moreover, age and gender variables were used in the flattening layer. A hidden layer with 400 neurons was applied in a fully-connected layer in our CNN model. As we used a large volume of data that's why mini-batch size (32) was applied to optimize our algorithm. A Rectified Linear Unit (ReLU) was used in input and hidden layers and Softmax was used in the output layer.

We employed a 5-fold cross-validation technique in this study. In this process, our total dataset was equally divided into 5 fold. 4/5 fold was used in the training of CNN while the remainder of the 1/5 fold was used to validate the performance of our proposed model. This approach was iterated 5 times by shuffling the test data. The performance of this process was evaluated in each iteration by accuracy, sensitivity, and specificity. Finally, an average of the performance of all five iterations was considered in this study. The area under the receiver operating curve, Sensitivity, specificity was used to measure the performance of our model.

Results

We included 4,071 cases and 16,284 controls in our study. The mean age of case and control patients were 70 and 47.5 years, respectively. All the patients diagnosed with PCA were male. The average number of diagnoses in patients with PCA was higher than controls (44 vs 24) when we considered PCA and non-PCA patients one year before the index date. However, over the three years, the average number of diagnoses in PCA and non-PCA patients was 41.23 and 22.66.

The performance of the deep learning model for predicting PCA patients one year ahead is shown in figure 1. The mean AUC of the CNN model in the minimal electronic health record feature was 0.94, with a sensitivity of 0.87 and specificity of 0.88 at the threshold value of 0.4.

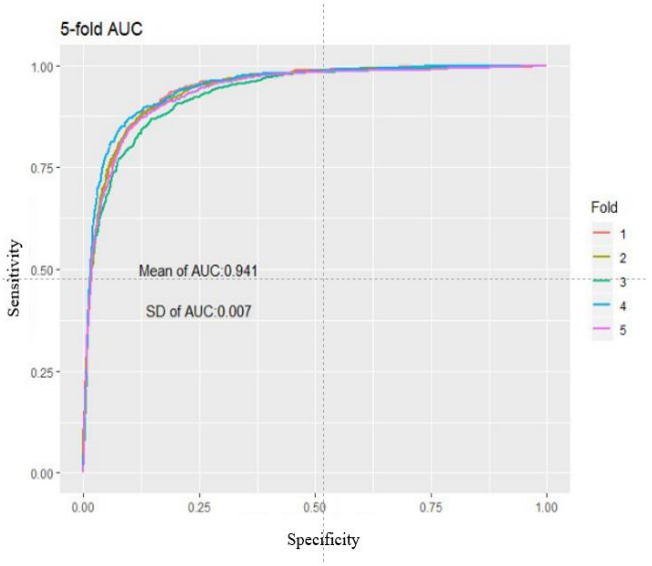

Figure 1- The performance of a deep learning model for predicting PCA.

Discussion

In this current study, we used the CNN model to predict PCA patients one year earlier using the minimal feature of electronic health records. Our prediction model showed high sensitivity and specificity for predicting PCA patients. Findings of our study indicate that the CNN model could provide physicians immense opportunity to improve patient outcomes through earlier PCA classification and subsequent follow-up high-risk PCA patients. CNN-based prediction models may help to improve survival rate, minimize treatment costs, and reduce the risk of mortality [11].

Patients with PCA is always asymptomatic in the early stages, this is the primary challenge to timely recognition and management of PCA [12]. Additionally, low awareness of both patients and physicians regarding PCA reduces the quality of care and early recognition. Our prediction model will help to improve the timely recognition of PCA based on their medical history that may assist physicians to make correct diagnosis decisions. The feature selection process that combined both comorbidities and co-medications history of each patient. Several traditional risk factors (e.g. age, gender) of PCA was top the most important. However, our study identified that patients with hyperplasia of prostate, tinea cruris, inflammatory diseases of the prostate, cataract, cystitis were a high risk of PCA. Ørsted et al. reported that establishing the role of prostate hyperplasia for developing PCA might improve the accuracy of prognostication, enhance intervention, reduce the risk of mortality [13]. Furthermore, a meta-analysis showed that prostate hyperplasia was associated with an increased risk of prostate cancer and the risk of PCA in a patient with prostate hyperplasia was higher in Asians [14]. Moreover, medical history of long-term use of terfenadine, antacids with antiflatulents, anti-inflammatory and anti-rheumatic, nitroxoline, Alpha-adrenoreceptor antagonists were associated with increased risk of PCA. Doat et al. demonstrated that the use of anti-inflammatory medications was associated with an increased risk of prostate cancer [15].

This prediction model of PCA can benefit healthcare professionals in various ways. For example, healthcare policymakers could have stratified the whole population by their risk score and could make budget planning as patients with a high risk of PCA need more treatment resources. This prediction model may also help physicians as assistance tools for correct decision making of PCA risk patients. Patients with

high-risk of PCA can be transferred to the screening test to confirm whether or not the patients have PCA. Our prediction model could help to give a primary idea for treating a patient with a high risk of PCA whether the patients need any additional clinical treatments or not.

Conclusion

Our prediction model can correctly classify PCA patients one year earlier using minimal features of EHR. The utilization of CNN model may facilitate optimal candidate selection and progression of patients at high risk of PCA.

Conflict of Interest

None

References

[1] Rawla P. Epidemiology of Prostate Cancer. *World journal of oncology*. 2019; 10(2):63.

[2] Ferlay J, Ervik M, Lam F, Colombet M, Mery L, Piñeros M, *et al.* Global cancer observatory: cancer today. Lyon, France: International Agency for Research on Cancer 2018.

[3] Miller KD, Siegel RL, Lin CC, Mariotto AB, Kramer JL, Rowland JH, *et al.* Cancer treatment and survivorship statistics, 2016. *CA: a cancer journal for clinicians*. 2016; 66(4):271-89.

[4] Nitta S, Tsutsumi M, Sakka S, Endo T, Hashimoto K, Hasegawa M, *et al.* Machine learning methods can more efficiently predict prostate cancer compared with prostate-specific antigen density and prostate-specific antigen velocity. Prostate International 2019.

[5] Gann PH, Hennekens CH, Stampfer MJ. A prospective evaluation of plasma prostate-specific antigen for detection of prostatic cancer. *Jama*. 1995; 273(4):289-94.

[6] Djavan B, Zlotta A, Kratzik C, Remzi M, Seitz C, Schulman CC, *et al.* PSA, PSA density, PSA density of transition zone, free/total PSA ratio, and PSA velocity for early detection of prostate cancer in men with serum PSA 2.5 to 4.0 ng/mL. *Urology*. 1999; 54(3):517-22.

[7] Yang H-C, Nguyen PAA, Islam M, Huang C-W, Poly TN, Iqbal U, *et al.* Gout drugs use and risk of cancer: A case-control study. *Joint Bone Spine*. 2018; 85(6):747-53.

[8] Liang C-W, Islam MM, Poly TN, Yang H-C, Jack LY. Association between gout and cardiovascular disease risk: A nation-wide case-control study. Joint, bone, spine. *revue du rhumatisme*. 2019; 86(3):389.

[9] Wu C-C, Yu Y-Y, Yang H-C, Nguyen PA, Poly TN, Islam MM, *et al.* Levothyroxine use and the risk of breast cancer: a nation-wide population-based case–control study. *Archives of gynecology and obstetrics*. 2018; 298(2):389-96.

[10] Islam MM, Poly TN, Li Y-CJ. Retinal Vessels Detection Using Convolutional Neural Networks in Fundus Images. bioRxiv. 2019:737668.

[11] Islam MM, Nasrin T, Walther BA, Wu C-C, Yang H-C, Li Y-C. Prediction of sepsis patients using machine learning approach: a meta-analysis. *Computer methods and programs in biomedicine*. 2019; 170:1-9.

[12] Cordon-Cardo C, Kotsianti A, Verbel DA, Teverovskiy M, Capodieci P, Hamann S, *et al.* Improved prediction of prostate cancer recurrence through systems pathology. *The Journal of clinical investigation*. 2007; 117(7):1876-83.

[13] Ørsted DD, Bojesen SE. The link between benign prostatic hyperplasia and prostate cancer. *Nature Reviews Urology*. 2013; 10(1):49.

[14] Dai X, Fang X, Ma Y, Xianyu J. Benign prostatic hyperplasia and the risk of prostate cancer and bladder cancer: a meta-analysis of observational studies. *Medicine* 2016; 95(18).

[15] Doat S, Cénée S, Trétarre B, Rebillard X, Lamy PJ, Bringer JP, *et al.* Nonsteroidal anti-inflammatory drugs (NSAID s) and prostate cancer risk: results from the EPICAP study. *Cancer medicine*. 2017; 6(10):2461-70.

Address for correspondence:

Yu-Chuan (Jack) Li, MD, Ph.D.
Graduate Institute of Biomedical Informatics
College of Medical Science and Technology
Taipei Medical University
250-Wuxing Str., Xinyi Dist., Taipei 11031, Taiwan
E-mail: jaak88@gmail.com, jack@tmu.edu.tw