MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190507

Development of Integrated Data and Prediction System Platform for the Localized Prostate Cancer

Sun Jung Lee[a,b], Sung Hye Yu[a], Yejin Kim[c], Jun Hyuk Hong[d], Choung-Soo Kim[d], Seong Il Seo[e], Chang Wook Jeong[f], Seok-Soo Byun[g], Byung Ha Chung[h], Ji Youl Lee[h], In Young Choi [a,b]

[a]Department of Medical Informatics, College of Medicine, Catholic University of Korea, Seoul, Republic of Korea,
[b]Department of Biomedicine & Health Sciences, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea
[c]Department of Creative IT Engineering, POSTECH, Korea
[d] Department of Urology, University of Ulsan College of Medicine, Seoul, Republic of Korea
[e]Department of Urology, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea
[f] Department of Urology, Seoul National University College of Medicine, Seoul, Republic of Korea
[g]Department of Urology, Seoul National University Bundang Hospital, Seongnam, Republic of Korea
[h]Department of Urology, Yonsei University College of Medicine, Seoul, Republic of Korea

Abstract

In this study, we built a multi-center integrated database platform of localized prostate cancer and developed biochemical recurrence (BCR) prediction system with Gradient Boosted Regression model using Korean Prostate Cancer Registry (KPCR) database. This platform will facilitate clinical management of patients with prostate cancer, and it will also help develop appropriate treatment of prostate cancer.

Keywords:

Prostatic Neoplasms, Biochemical Recurrence Prediction, Database Management Systems

Introduction

Prostate cancer is the fifth most common cancer in men in South Korea, and the incidence is increasing [1]. Prostate cancer is a high incidence of disease and it is important to prevent and manage recurrence after curative treatment. By managing the integrated data of prostate cancer and predicting biochemical recurrence (BCR), progress and pattern of prostate cancer is possible to analyze and may help to choose a treatment in the future. Currently, there are prostate cancer data portals for scientists and clinical doctors, but most of the studies were based on Americans and Europeans data. The cancer incidence rates are different among the countries [2], and there is also racial difference both in cancer incidence and survival rate [3]. Therefore, in order to analyze and predict the prostate cancer data of Koreans, Korean Prostate Cancer Database System is required. The purpose of this study is to construct a multi-center integrated database platform using the Korean Prostate Cancer Registry (KPCR) data and develop a biochemical recurrence (BCR) prediction system to facilitate clinical management of prostate cancer.

Methods

The KPCR Database has collected 7,394 prostate cancer data of Korean from six domestic EMR-based medical institutions. We implemented a system platform to visualize and manage this data. We implemented a web-based platform using Java language and Tomcat Server. In order to predict BCR, we extracted 5,119 analytic data from the KPCR database, generated train dataset for BCR prediction and trained the dataset. We used 15 statistical analysis models to predict BCR including decision tree, logistic regression, and neural networks.

Results

This platform has been implemented features of integrated data management and BCR prediction. The AUC (Area Under the Curve) of BCR prediction is shown in Table 1. 3 years post-surgery, the highest AUC of BCR prediction is achieved by Gradient Boosted Regression, and the value is 0.8419. 5 years after surgery, the highest AUC is achieved by Ridged regression with a value of 0.8071. Finally, we implemented a BCR prediction system using a Gradient Boosted Regression model which reaches the highest average of AUC.

The 3-year and 5-year BCR prediction system with Gradient Boosted Regression model was implemented using the Python language, and we used the GradientBoostingClassifier of the sklearn module in Python. In order to integrate the Web-based platform implemented in Java with the BCR prediction system implemented in Python, we used the Django framework, a Python web framework.

The variables used for BCR prediction include age, initial PSA, clinical stage, pathology Gleason score, pathology stage, surgical margin, perineural invasion, seminal vesicle, extracapsular extension, and lymphovascular invasion. We used the feature_importances_method of GradientBoostingClassifier to output the variable importance, and the result is shown in Table 2.

The result shows that age, initial PSA, and pathology Gleason score are important variables in predicting BCR with gradient boosting regression model.

We developed biochemical recurrence (BCR) prediction system using data of Korean prostate cancer patients. In the input screen, enter the value of age, initial PSA, clinical stage, pathology Gleason score, pathology stage, surgical margin,

perineural invasion, seminal vesicle, extracapsular extension, and lymphovascular invasion. The input value uses the predict_proba method of GradientBoostingClassifier to determine the probability of occurrence of BCR and the probability of non-occurrence of BCR after 3-year and 5-year of surgery. In the result screen, BCR prediction is implemented by outputting the probability of BCR occurrence after 3-year and 5-year of prostate cancer surgery using Gradient Boosting Regression model.

Table 1 – BCR Prediction AUC

Analysis	Method	3-year	5-year
Trees	Decision Tree	0.8186	0.7561
	Random Forest(ntrees=20)	0.8320	0.7956
	Random Forest(ntrees=50)	0.8349	0.8047
	Random Forest(ntrees=80)	0.8362	0.8050
Logistic Regression	Ridged Regression(L2)	0.8288	0.8071
	Lasso(L1)	0.8319	0.7993
	Gradient Boosted	0.8419	0.8031
Neural Networks	1hidden, dropout=0.3 at input	0.7939	0.7895
	1hidden dropout=0.1 at input	0.8027	0.7977
	1hidden dropout=0.1 at input, hidden	0.7984	0.7923
	2hidden, dropout=0.3 at input	0.7978	0.7941
	2hidden dropout=0.1 at input	0.7988	0.7967
	2hidden dropout=0.1 at input, hidden	0.8016	0.7943
Survival Regression	Cox PH	0.7944	0.7816
	Random survival forest	0.7645	0.6920

Table 2 – Variable Importance

Variable	3-year	5-year
Age	0.1601	0.1870
Initial PSA	0.2813	0.2284
Clinical T stage	0.1357	0.1343
Pathology Gleason Score Sum	0.1484	0.1546
Pathology T stage	0.0656	0.0445
Surgical margin	0.0321	0.0525
Perineral invasion	0.0380	0.0483
Seminal vesicle	0.0470	0.0247
Extracapsular extension	0.0055	0.0083
Lymphovascular invasion	0.0454	0.0360

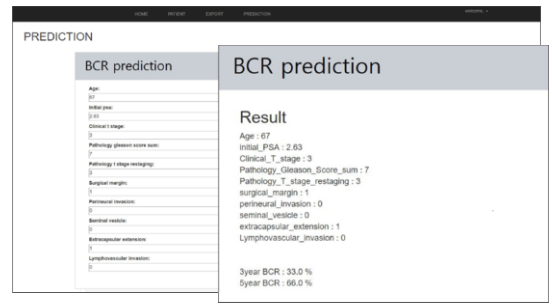

Figure1 – BCR Prediction System

Conclusions

In this study, we designed a management system platform for Korean prostate cancer patients. The platform was implemented on web-based to facilitate data access, management and BCR prediction. It will further facilitate clinical management of patients with prostate cancer and it will also help develop appropriate treatment of prostate cancer in future.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIP,2016R1A2B4015922)

References

[1] 2016 cancer statistics. National Cancer Information Center. [accessed on November 12, 2018]. https://www.cancer.go.kr/lay1/S1T639C641/contents.do.

[2] F. Bray, J. Ferlay, I. Soerjomataram, R.L. Siegel, L.A. Torre, and A. Jemal, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA Cancer J Clin 68 (2018), 394-424.

[3] R.M. Hoffman, F.D. Gilliland, J.W. Eley, L.C. Harlan, R.A. Stephenson, J.L. Stanford, P.C. Albertson, A.S. Hamilton, W.C. Hunt, and A.L. Potosky, Racial and ethnic differences in advanced-stage prostate cancer: the Prostate Cancer Outcomes Study, J Natl Cancer Inst 93 (2001), 388-395.

Address for correspondence

In Young Choi
222 Banpo-daero, Seocho-gu, Seoul 06591, Korea
Tel: +82-2-2258-7648, E-mail: iychoi06@gmail.com