Time-optimized Prediction of Late Cancer Diagnosis

Huan-ju Shih, MHA, Janusz Wojtusiak, PhD, Hua Min, PhD
George Mason University, Fairfax, VA, USA

Introduction Late diagnosis of medical conditions, particularly prostate and breast cancer, presents a significant challenge in healthcare, often leading to advanced disease stages and compromised patient outcomes. This study underscores the critical importance of understanding factors contributing to delayed diagnosis, such as limited awareness and socioeconomic disparities, to inform effective interventions. It proposes leveraging machine learning models to predict late-stage cancer, focusing on gender-specific differences. The study optimizes the time between data collection and reliable cancer stage prediction to achieve this goal. **Methods** This study utilizes the SEER-MHOS (Surveillance, Epidemiology, and End Results-Medicare Health Outcomes Survey) database to identify variables potentially related to prostate cancer diagnosis[1]. A survey response immediately preceding cancer diagnosis was selected. Demographic data and functional indicators, along with MHOS survey dates, primary tumor site, and cancer stage, were extracted. A complex study design was used, and two main stages were applied after dividing the cohort into training and testing sets. The first stage was discovering an optimal lookback window defining the maximum allowed time between survey completion and cancer diagnosis. Iterative multiple imputation was used to fill in missing values. Logistic regression was used to predict missing values. Then, 10-fold cross-validation was used within the training set on lookback windows ranging from 90 days to 3 years, with loop increments of 10 days. For each iteration, gradient boosting, random forest, and logistic regression were used to calculate AUCs, and the results were averaged. Further, the results were smoothened with a moving average to avoid noise resulting from small sample sizes. Finally, the optimal lookback window was found. In the second stage of the experiment, final models were built based on the previously found lookback window. Missing data were re-imputed based only on data from that period. Final models were constructed using multiple classification methods for comparison. Final evaluation of results was completed on the previously set aside test set. The above two-stage process was completed for both prostate cancer and breast cancer data, as well as one dataset that combines both. Models were cross-compared to calculate the performances of the three types of models on prostate and breast datasets independently. Precision, recall, and AUC metrics were computed. **Results** The training dataset contains 3,693 patients, and the testing data contains 1,583 patients. The optimal window size for prostate cancer data was 870 days (2,092 patients), and for breast cancer data was 1270 days, as exemplified in Figure 1 for breast cancer. The figure shows the performance of the three models (right axis) and the amount of data available (left axis) in relation to the lookback window (bottom axis). Points indicate the results of individual experiments, and lines are smoothened AUC values. With larger lookback windows, the amount of data grows, contributing to higher accuracies. However, the inclusion of data that is too old creates noise that decreases performance. When applied to the final test set, lasso regression achieved the highest AUC at 0.61, which is too weak for any practical application of the system but indicates the existence of a small signal that links the input variables to the predicted cancer stage. When cross-applied between prostate and breast cancer datasets, model performances are similar with AUC 0.60. **Discussion and Conclusions** The results indicate that different cancers require varying cutoff points for predicting late-stage diagnosis. The study highlights challenges in late cancer diagnosis, particularly for prostate and breast cancer, due to factors like low awareness and socioeconomic disparities. However, the prediction quality is too low for any practical use as a decision support system. Further analysis indicates that the main limiting factor in the construction of the models was a very small sample size after processing SEER-MHOS data. Despite the current limitations in accuracy, the models could inform the development of more precise screening and follow-up protocols. For instance, patients flagged by the model could be enrolled in more frequent screening programs or placed under more stringent monitoring, ensuring that any signs of cancer are caught at an earlier, more treatable stage. It is likely that the obtained accuracies would be higher with larger datasets.

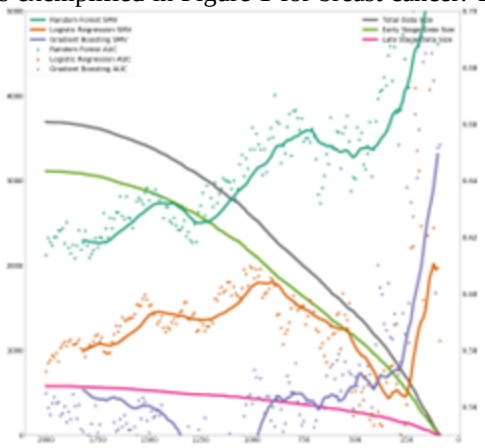

Figure 1: Time window optimization results.

References

1. SEER-Medicare Health Outcomes Survey (SEER-MHOS) Linked Data Resource [Internet]. [cited 2024 Mar 18]. Available from: https://healthcaredelivery.cancer.gov/seer-mhos/