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

doi:10.3233/SHTI190480

Machine Learning Methods to Predict Lung Cancer Survival Using the Veterans Affairs Research Precision Oncology Data Commons

Nhan V Do[a,b], Jaime C Ramos[a], Nathanael R Fillmore[a], Robert L Grossman[c], Michael Fitzsimons[c], Danne C Elbers[a], Frank Meng[a,b], Brett R Johnson[a], Samuel Ajjarapu[a], Corri L DeDomenico[a], Karen E Pierce-Murray[a], Robert B Hall[a], Andrew F Do[a], Kelly Gaynor[a], Peter L Elkin[d], Mary T. Brophy[a,b]

[a]MAVERIC, Department of Veterans Affairs Medical Center, Boston, MA, USA
[b]Boston University School of Medicine, Boston, MA, UA
[c]University of Chicago, Chicago, IL, USA
[d]University at Buffalo, Buffalo, NY, USA

Abstract

We completed a pilot study to guide the development of the VA Research Precision Oncology Data Commons infrastructure as a collaboration platform with the greater research community. Our results using a small subset of patients from the VA's Precision Oncology Program demonstrate the feasibility of our data sharing platform to build predictive models for lung cancer survival using machine learning, as well as highlight the potential of target genome sequencing data.

Keywords:

Machine Learning, Lung Neoplasms, Precision Medicine

Introduction

The Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC) launched the Research for the Precision Oncology Program (RePOP) and the Precision Oncology Data Repository (PODR) in 2016 to acquire the knowledge necessary to improve care for Veterans with cancer. In the fall of 2017, MAVERIC entered a partnership with the University of Chicago to establish a VA Research Precision Oncology Data Commons (RePODC) to provide greater access to both VA data and robust computing resources to the research community external to the VA. We conducted a pilot study to guide the development of the RePODC infrastructure for data sharing and to demonstrate the feasibility of building predictive models with both clinical and genomic data. .

Methods

The exploratory dataset contained 60 instances, each of them a patient record containing a preselection of 153 features based on targeted genome sequencing, an extract from the VA's comprehensive clinical information on each patient, as well as basic demographic data. Following previous work by Lynch et al [1], a small number of candidate features were selected based on their potential relationship with survival time, including both continuous and categorical variables, for a total of 16 attributes. Among the selected features we included the presence of tumor variants for CDKN2A, KRAS and TP53, the number of primary tumors, tumor stage and location, histological findings, exposure to radiation protocols, and the sequence of radiation with surgery. Logistic regression and random forest algorithms were used to predict survival status at 24 months after diagnosis. Ten-fold cross validation was used to assess accuracy.

Results

Cross-validated accuracy was 0.886 for for both the logistic regression and random forest classifiers. In order of importance, the random forest algorithm identified primary tumor location, age, surgery primary site and the variant CDKN2A as the most relevant features to predict survival stage, with higher scores than other features of known importance like stage or histology.

Conclusions

Our results, using a small subset of patients from the VA Research in Precision Oncology Data Commons, demonstrate the feasibility of using this data to build predictive models for lung cancer survival using machine learning methods, and particularly highlight the potential of targeted tumor sequencing data for building predictive models of survival. This work suggests that by expanding access to VA data, we enable use of modern machine learning methods and advanced analytics to contribute to scientific knowledge and improve Veterans' healthcare.

Acknowledgements

This material is based upon work supported by the Department of Veterans Affairs, Veterans Health Administration, Office of Research and Development, Cooperative Studies Program. The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the United States government.

References

[1] C.M. Lynch, B. Abdollahi, J. D. Fuqua, A. R. de Carlo, J. A. Bartholomai, R. N. Balgemann, V. H. van Berkel, H. B. Frieboes, Prediction of lung cancer patient survival via supervised machine learning classification techniques, *International Journal of Medical Informatics* **108** (2017), 1-8.

Address for correspondence

Nhan Do, MD
150 S. Huntington Ave
Boston, MA 02130; nhan.do@va.gov