The impact of time censoring on machine learning models which identify patients with undiagnosed cardiac amyloidosis

Greg Lee BS,1, Sam Fielden PhD,2, Brendan Carry MD,2, Alvaro Ulloa Cerna PhD,2, Linyuan Jing PhD,2, Arun Nemani PhD,1, Dan Rocha MS,2, Kyle Flick BS,2, Jeffrey Ruhl MS,2, Jagadish Venkataraman PhD,1, Noah Zimmerman PhD,2, RuJun Chen MD,1, Brandon Fornwalt MD., PhD,1, Christopher M. Haggerty PhD.2
1Tempus Labs Inc, Chicago, IL, USA, 2Geisinger, Danville, PA, USA

INTRODUCTION

Cardiac amyloidosis (CA) is a common cause of progressive heart failure. New therapies can improve outcomes but most patients remain undiagnosed and untreated. Machine learning models trained and deployed on electronic health record (EHR) data can find CA patients in retrospective analysis. Certain features may be more prevalent following CA diagnosis and therefore may not be available in a prospective setting (Figure 1). To date, most models have focused on identification of undiagnosed CA using uncensored data. We hypothesized that lack of post-diagnosis censoring when training CA models leads to poor performance in predicting patients with undiagnosed cardiac amyloidosis.

METHODS

Labels were generated by aggregating 96 CA patients from the Geisinger amyloidosis multi-disciplinary clinic registry with 19,200 matched controls (age, sex, encounter frequency and EHR timespan). We used 8 EHR data elements (age, BMI, creatinine, triglycerides, proBNP, PR interval, QTc, QRS duration and IVSd) to train a boosted decision tree ensemble. Retrospective performance was evaluated by 20-fold cross-validation. Models were prospectively deployed to ~100,000 patients who were alive and over the age of 60, who had one or more echocardiogram(s), and had a clinical encounter within 2 years. We randomly sampled predicted positives and negatives and evaluated performance compared to a chart review by a trained clinician. We compared our findings to a web-based CA model that was publicly available in 2020.

RESULTS

CA registry cohort feature prevalence and mean time to diagnosis (days), separated by model. The public model dates represent the first instance of a relevant grouping of ICD codes while Tempus model dates represent the first instance of recorded measurement.

- The Tempus model had moderate performance on at-risk, time-censored patients when trained with and without time censoring (Figure 2).
- When trained and tested on temporally uncensored data the Tempus model showed higher performance which may be unrepresentative of deployment scenarios where post-diagnostic features are unavailable for model use.
- In chart review, the Tempus model demonstrated higher PPV and lower mean age for true positive predictions suggesting increased actionability and yield (Figure 3).
- The publicly available model demonstrated a similar trend when tested on uncensored data as compared to an appropriately censored feature set.

CONCLUSIONS

- Models should be evaluated on temporally censored data so that post-diagnostic features do not artificially inflate performance estimates and negatively impact real-world deployment.
- EHR models can be trained on censored data to find actionable patients with high risk of undiagnosed CA.