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.[1], RuiJun Chen MD.[1], Brandon Fornwalt MD., PhD.[1], Christopher M. Haggerty PhD.[2]

[1]Tempus Labs Inc, Chicago, IL, USA, [2]Geisinger, 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

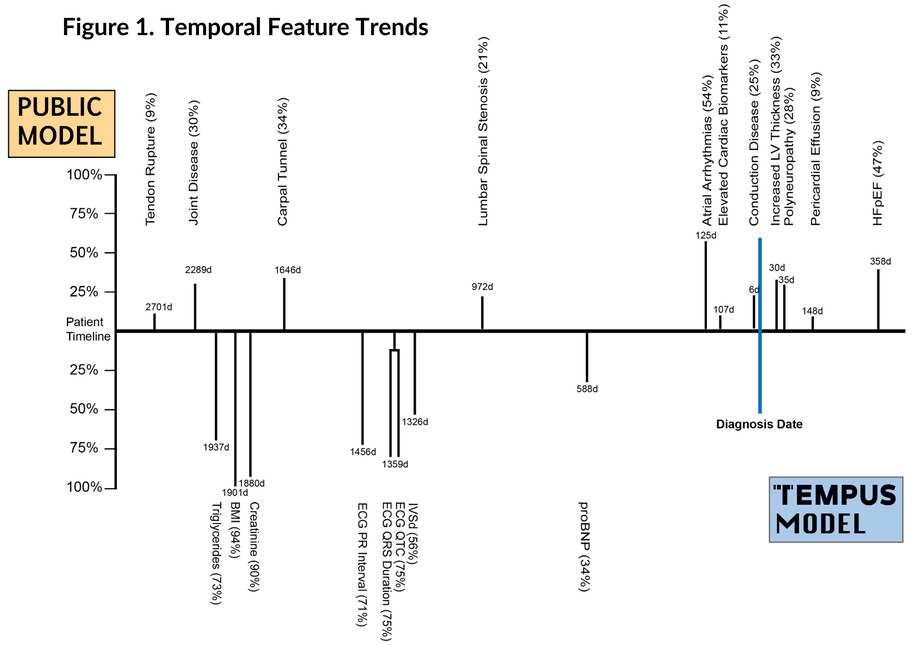
- Labels were generated by aggregating 96 CA patients from the Geisinger amyloidosis multi-disciplinary clinic registry with 19,200 matched controls (age, sex, encounter frequency and EHR timespan)
- We used 8 EHR data elements (age, BMI, creatinine, triglycerides, proBNP, PR interval, QTC, QRS duration and IVSd) to train a boosted decision tree ensemble with and without time censoring
- 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

Acknowledgements: We thank Amrita Iyer Ph.D. for help with poster design

Disclosures: Geisinger receives funding from Tempus for ongoing development of predictive modeling technology and commercialization. None of the Geisinger authors have ownership interest in any of the intellectual property resulting from the partnership. Greg, Arun, Noah, Ruijun and Brandon are Tempus employees.

RESULTS

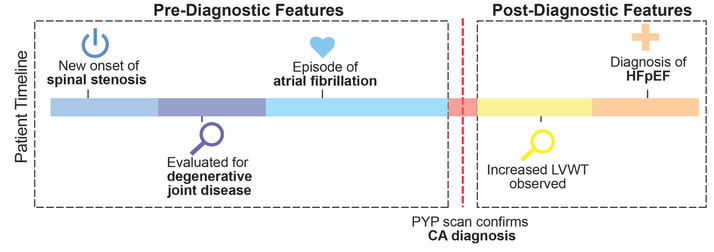
Figure 1. Temporal Feature Trends



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

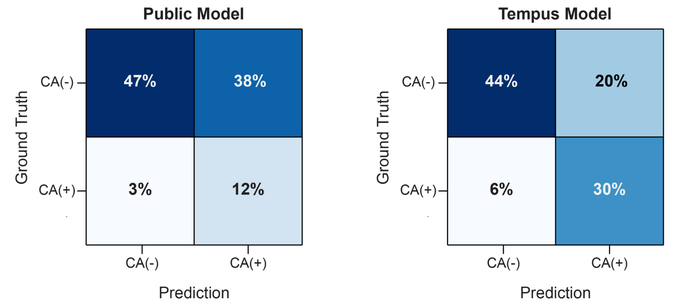
Figure 2. Retrospective Performance



Train/Val Features	Test Features	Public Algo ROC-AUC	Tempus Algo ROC-AUC
uncensored	uncensored	0.86 ± 0.03	0.91 ± 0.03
uncensored	censored	0.54 ± 0.04	0.80± 0.04
censored	censored	n/a	0.84 ± 0.07

A common CA patient timeline with pre-diagnostic features in cool colors and post-diagnostic features in warm colors. Retrospective performance was evaluated by training and testing models on features with and without time censoring.

Figure 3. Prospective Chart Review



Model	Predicted Positive Prevalence	Sample Size	Positive Predicted Value (PPV)	Mean True(+) Age
Public	50%	60 patients	0.15	85.21 ± 8.84
Tempus	50%	100 patients	0.60	78.23 ± 8.18

Performance was evaluated by blinded chart review to distill ground truth by a CA specialist. Due to the spectral presentation of the disease, CA suspicion was ranked on a scale (0 - no CA, 1 - CA highly unlikely, 2 - CA unlikely, 3-non conclusory evidence, 4-50% chance of CA and 5-CA likely). A score from 0-3 is considered CA(-) and a score from 4-5 is considered CA(+).

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

Correspondence