Development of a Novel Approach to Improve Binary Classification Prediction Tasks When Encountering Time-To-Event Data: Performance Comparison of Four Alternative Approaches To Predict Stimulant Use Disorder Among Persons Authorized To Purchase Medical Marijuana

UAMS University of Arkansas for Medical Sciences

MSR34

Allen M. Smith[1], Horacio Gomez-Acevedo[2], Corey Hayes[1], Melody Greer[2], Bradley C. Martin[1]

[1]Division of Pharmaceutical Evaluation and Policy, Department of Pharmacy Practice, University of Arkansas for Medical Sciences, Little Rock, AR, USA; [2]Department of Biomedical Informatics, University ofArkansas for Medical Sciences, Little Rock, AR, USA

BACKGROUND

- Compared to a more traditional continuous-time survival modeling approach, Suresh et al. (2022) demonstrated a modest improvement in model discrimination can often be achieved when utilizing a discrete-time framework, where time-to-event (TTE) data is converted into a person-period dataset that splits subject follow-up time into equally-spaced intervals.[1]
  - However, this discrete-time framework may be further improved by informing prediction in each time interval with updated features from prior intervals rather than just utilizing time-independent features.

Objective: Contrast model performance of four approaches to organizing TTE outcomes: 1. a simple binary classification, 2. continuous-time, 3. discrete-time intervals, and 4. discrete-time-updating approach to predict stimulant use disorder (StUD) risk among Arkansas medical marijuana (MMJ) cardholders

METHODS

Data Source

- Time-to-event datasets were constructed using statewide health insurance claims data between November 2018 – December 2023 from the **Arkansas All-Payer Claims Database (AR-APCD).[3]**

Study Sample

- **Subjects:** Insured (medical + pharmacy benefits), adult (≥ 18 years old) Arkansas MMJ Cardholders without a recent history of StUD in the past 6 months.
- **Index Date:** May 11th, 2019 (opening date of 1st Arkansas MMJ dispensary) or receipt date of MMJ eligibility card, whichever came last
- **Follow-up:** Index date until 1st occurrence of one of the following: 1) New StUD diagnosis, 2) study end date (Dec. 31st , 2023), 3) health plan disenrollment, 4) death from any cause

Engineered Features [n=202]

- Included **demographics**, **acute + chronic comorbidities**, **prescription characteristics**, and **healthcare utilization characteristics.**
- **Feature selection:** Two-pronged recursive feature elimination approach using Random Forest-derived feature importance scores & Cox proportional hazards-derived p-values

Model Training/Testing

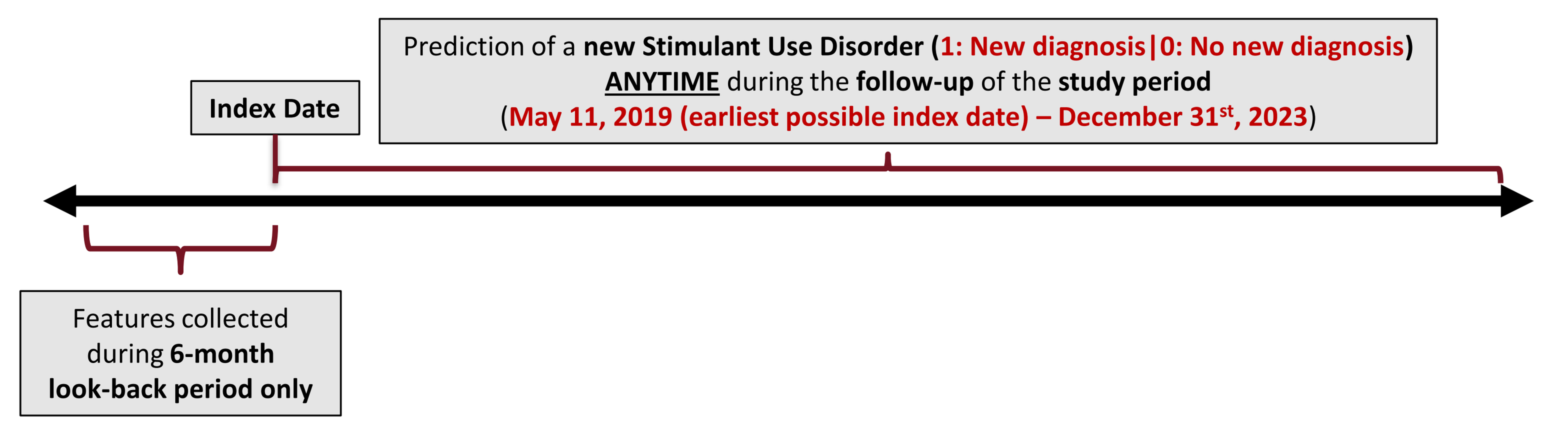
- **Train/test split:** Randomized 50:50 split at person level
- **Data balancing:** 1:25 random undersampling (RUS) of the majority class
- **Hyperparameter tuning:** 90 iterations with 5-fold cross validation
- **Performance Metrics:** Cumulative sensitivity/dynamic specificity area under the receiver-operating characteristic (C/D AUC), Brier Score, Sensitivity, Specificity, Positive Predictive Value (PPV), Negative Predictive Value (NPV)

Data Structure

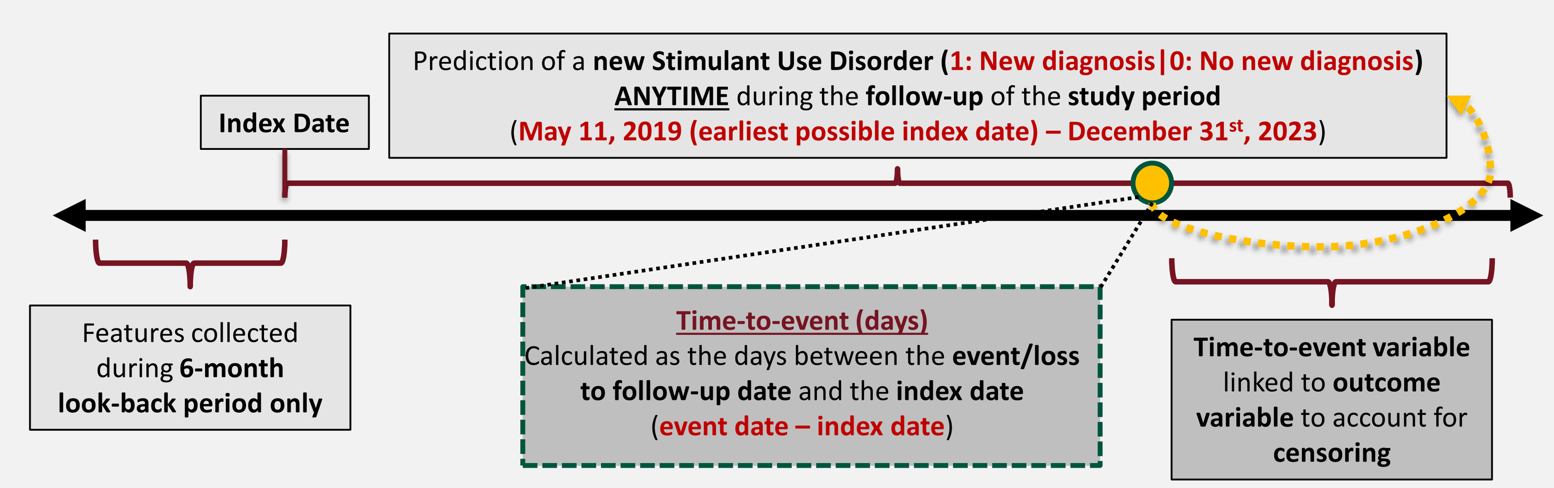
1. **Simple Binary Classification**
   - Person-level dataset that included feature values at baseline with an event indicator (1: StUD|0: No StUD)
   - **Trained Classifiers (2):** Random Forest (RF), Logistic Regression (LR)
2. **Continuous-Time**
   - Person-level dataset that included feature values at baseline, the event indicator, and a continuous variable indicating the length of the follow-up period in days.
   - **Trained Classifiers (4):** Random Survival Forest (RSF), Gradient Boosting (GB), Support Vector Machine (SVM), Cox Proportional Hazards Survival Model (CPH)
3. **Discrete-Time**
   - Person-period-level dataset where each row represented a 90-day time interval that the subject remained at risk
   - StUD prediction for each time interval was informed by feature values collected at baseline.
   - **Trained Classifiers (6):** RF, LR, RSF, GB, SVM, CPH
4. **Discrete-Time-Updating**
   - Person-period-level dataset where each row represented each 90-day time interval that the subject remained at risk
   - StUD prediction for each time interval was informed by the prior 6 months of feature values.
   - **Trained Classifiers (6):** RF, LR, RSF, GB, SVM, CPH

METHODS

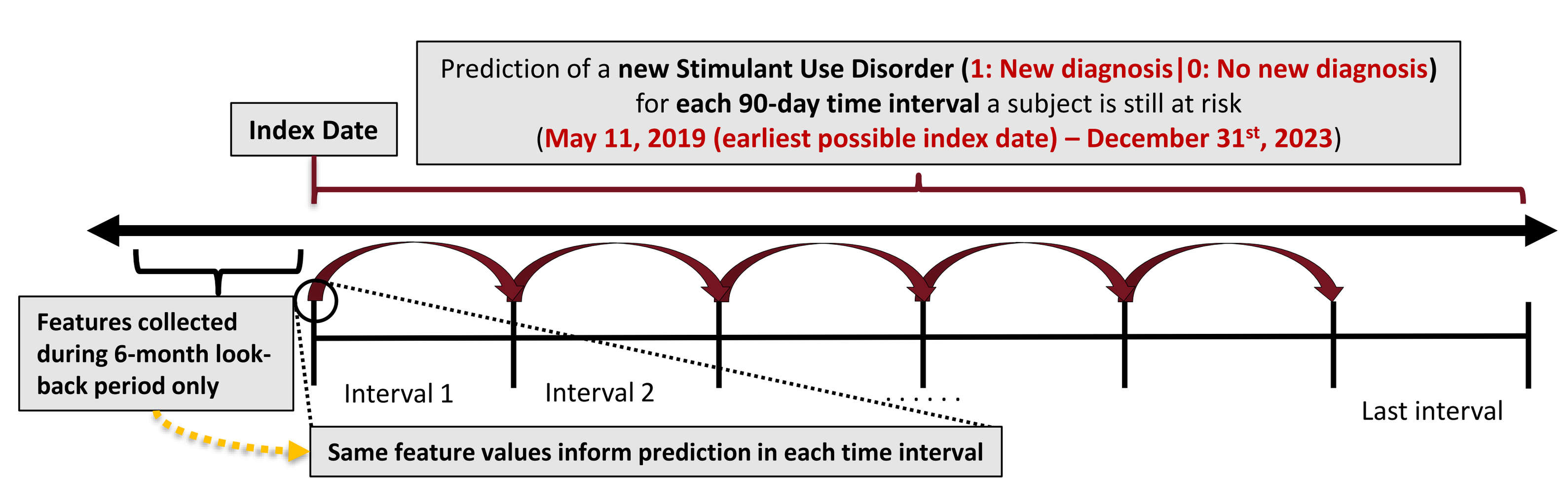
Approach 1: Simple Binary Classification



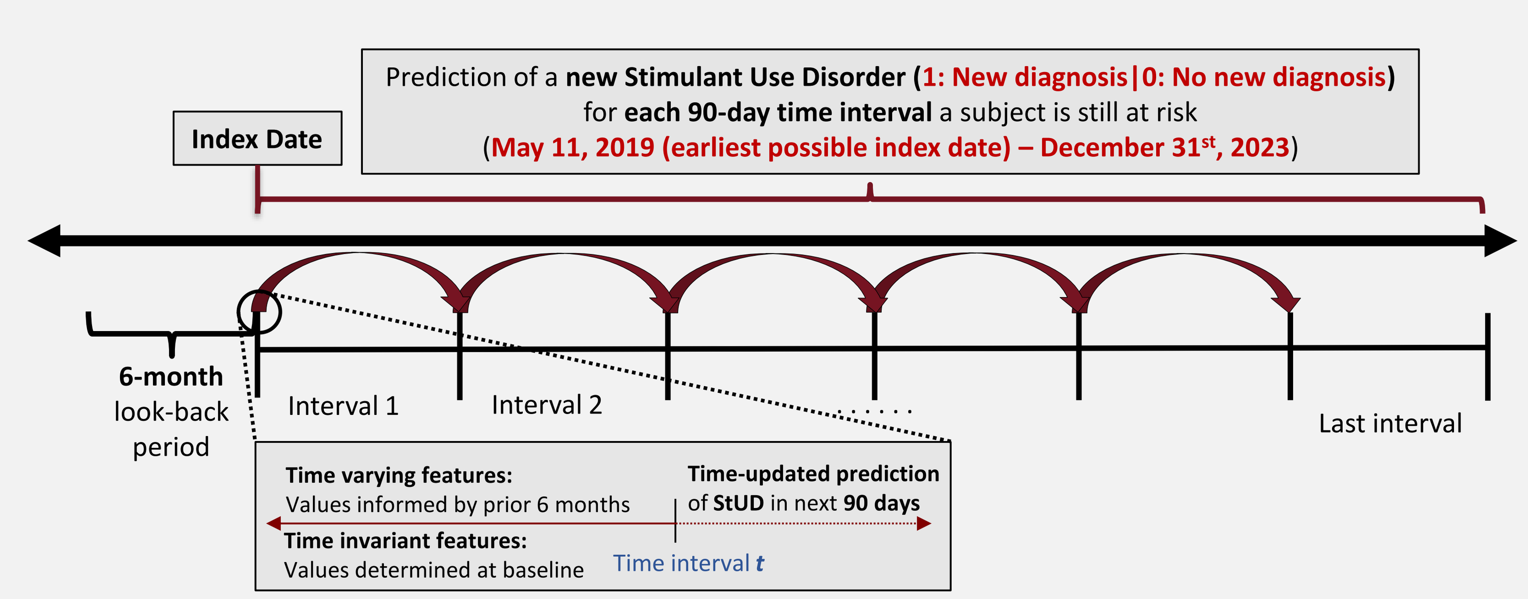
Approach 2: Continuous-Time



Approach 3: Discrete-Time



Approach 4: Discrete-Time-Updating



RESULTS

A total of **54,758 Arkansas MMJ cardholders** met eligibility criteria, of which **857 (1.57%)** received a new StUD diagnosis during the follow-up period.

Prediction of Stimulant Use Disorder within the next 90 days among Arkansas Medical Marijuana Cardholders: Model Performance Comparison Between Four Alternative Approaches for Handling Time-To-Event Data

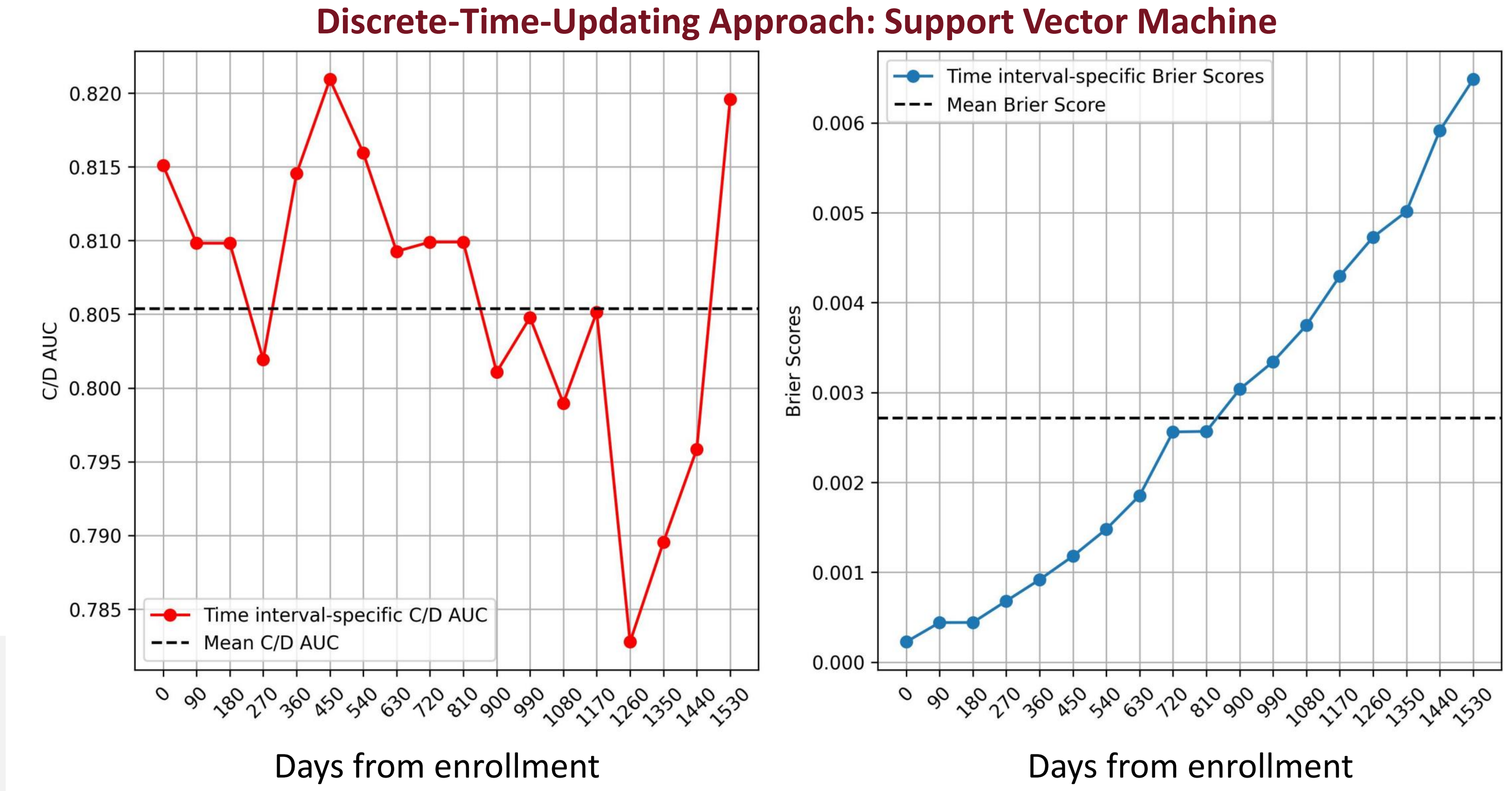
	Binary		Continuous-Time		Discrete-Time		Discrete-Time-Updating	
Classifier	AUC	Brier Score	Mean C/D AUC	Mean Brier Score	Mean C/D AUC	Mean Brier Score	Mean C/D AUC	Mean Brier Score
Random Survival Forest[a]	-	-	0.7516	**0.0143**	0.7419	0.0031	0.7793	0.0030
Gradient Boosting[a]	-	-	**0.7580**	0.0147	**0.7480**	0.0031	0.7878	0.0031
Support Vector Machine[a]	-	-	0.7175	0.0147	0.7277	**0.0027**	**0.8054**	**0.0027**
Cox Proportional Hazards[a]	-	-	0.7201	0.0165	0.7298	0.0036	0.8044	0.0032
Random Forest[b]	**0.7426**	**0.0157**	-	-	0.7468	0.0030	0.7386	0.0030
Logistic Regression[b]	0.7120	0.0188	-	-	0.7229	0.0032	0.7932	0.0031
Random Forest[c]	-	-	-	-	0.7435	0.0030	0.7323	0.0030
Logistic Regression[c]	-	-	-	-	0.7266	0.0032	0.7491	0.0032

AUC = area under the receiver-operating characteristic, C/D AUC = Cumulative sensitivity/dynamic specificity area under the receiver-operating characteristic

[a]Time interval included for calculation of time-dependent survival probabilities
[b]Time interval included as feature
[c]Time interval excluded from model training

Prediction of Stimulant Use Disorder within the next 90 days: AUC and Brier Score of Top Performing Model

Discrete-Time-Updating Approach: Support Vector Machine



CONCLUSION

- ✓ **Compared to the other approaches, the discrete-time-updating approach achieved the best discrimination overall with the support vector machine-trained survival model (C/D AUC=0.8054).**
  - ✓ **However, both the discrete-time and discrete-time-updating approach achieved similar calibration across most trained classifiers, with the support vector machine-trained models in each strategy achieving the best Brier score of 0.0027**
- ✓ **Thus, this novel discrete-time-updating approach that ingests updated features should be considered alongside more traditional approaches for other time-to-event classifications tasks.**

References

1. Suresh K, Severn C, Ghosh D. Survival prediction models: an introduction to discrete-time modeling. BMC Med Res Methodol. 2022;22(1):207. doi:10.1186/s12874-022-01679-6
2. Therneau TM. A Package for Survival Analysis in R. 2020. R package version 3.2-7. https://CRAN.R-project.org/package=survival.
3. Arkansas All-Payer Claims Database. Welcome to the Arkansas All-Payer Claims Database (APCD). https://www.arkansasapcd.net/Home/.