Comparison of Sampling Strategies to Address Severe Data Imbalance and Computational Benchmarking for Time-to-Event Predictive Model Development Across Local and High-Performance Computing Environments: Prediction of Alcohol Use Disorder and Opioid Use Disorder Among Arkansas Medical Marijuana Cardholders

MSR157

Allen M. Smith[1], Horacio Gomez-Acevedo[2], Corey J. Hayes[1], Melody L. Greer[2], Chenghui Li[1], 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 of Arkansas for Medical Sciences, Little Rock, AR, USA

BACKGROUND

- Severe class imbalance is a common challenge in clinical predictive modeling given the rarity of many health outcomes[1, 2]:
  - Data sampling & hyperparameter tuning can improve discrimination, but introduce tradeoffs between predictive performance, information retention, & computational burden.[1, 2]
- Computational demands are further amplified by modern time-to-event modeling frameworks that leverage longitudinal data to support dynamic risk prediction[4]:
  - High-performance computing (HPC) environment[1] and parallelization tools (e.g. Apache Spark)[2] offer potential solutions, but comparative evidence across computing environments remains limited.

Objective: Using real-world prediction of opioid use disorder (OUD) and alcohol use disorder (AUD) among medical cannabis (MC) cardholders, landmark supermodels were developed to compare:

1. Random undersampling (RUS) and random oversampling (ROS) across class ratios to identify the performance-optimized sampling strategy under systematic hyperparameter tuning
2. Compare model training runtimes across three computing environments: (a) **standard local server, (b) Apache Spark-enabled local server, (c) HPC environment**

METHODS

Data Source & Study Sample

- 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).**[5]
- **Subjects:** Insured (medical + pharmacy benefits), adult (≥ 18 years old) Arkansas MC Cardholders without a recent history of the substance use disorder (SUD) of interest in the past 6 months.
- **Index Date:** May 11th, 2019 (opening date of 1st Arkansas MC dispensary) or receipt date of MC eligibility card, whichever came last
- **Follow-up**: Index date until 1st occurrence of one of the following: 1) New OUD|AUD diagnosis, 2) study end date (Dec. 31st , 2023), 3) health plan disenrollment, 4) death from any cause

Engineered Features [OUD (n=170) | AUD (n=169)]

- Included **demographics**, **acute + chronic comorbidities**, **prescription characteristics**, and **healthcare utilization.**

Model Training/Testing

- **Train/test split:** Randomized 70:30 split at person level
- **RUS & ROS class ratios: (OUD:** 1:1, 1:3, 1:10, 1:25, 1:50, 1:100, full-data | **AUD:** 1:1, 1:3, 1:10, 1:25, full-data**)**
- **Hyperparameter tuning:** 90 iterations with 5-fold cross validation
- **Classifiers:** Random Survival Forest (RSF), Support Vector Machine Survival (SVMS), Cox Proportional Hazards (CPH), Random Forest (RF), Logistic Regression (LR)

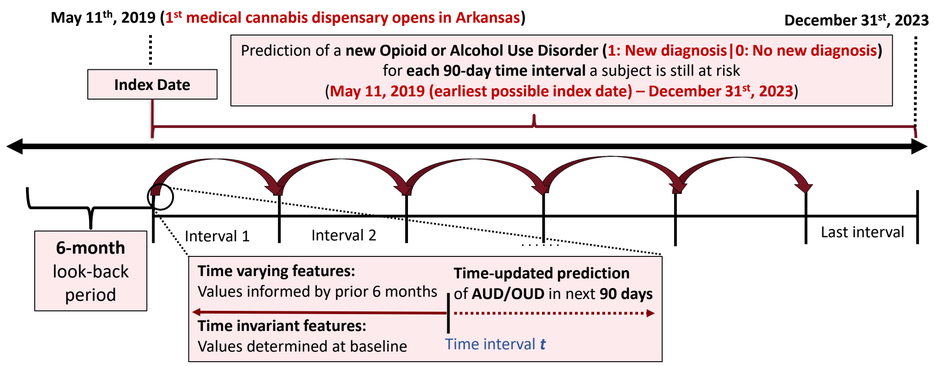
Predictive Performance Evaluation

- **Model discrimination:** cumulative sensitivity/dynamic specificity AUC (**C/D AUC**)
- **Model calibration:** horizon-aligned Inverse probability of censor weighting (**IPCW**) Brier scores

Computational Benchmarking

- **Absolute training times (minutes)** and **relative speedup** were assessed across three computing environments.
  - Local server environments (w/ & w/o Apache Spark v3.5.4) containerized & ran within HPC environment (96 CPU cores, 768 GB RAM), w/ resources restricted (16 CPU cores/160 GB RAM) to emulate typical clinical analytics setting.
- This work was supported with computing resource allocations (MED230043) awarded through the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program.[6]

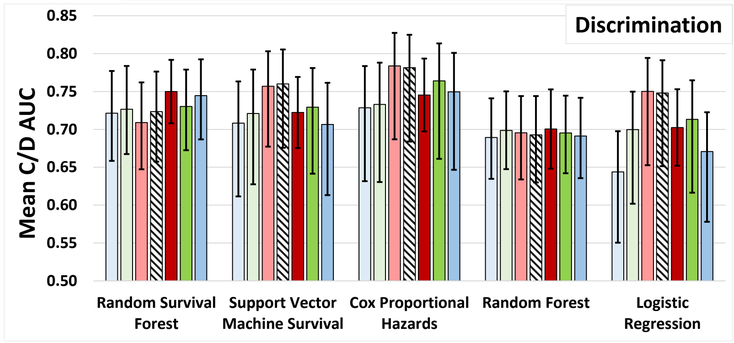
Landmarking Approach

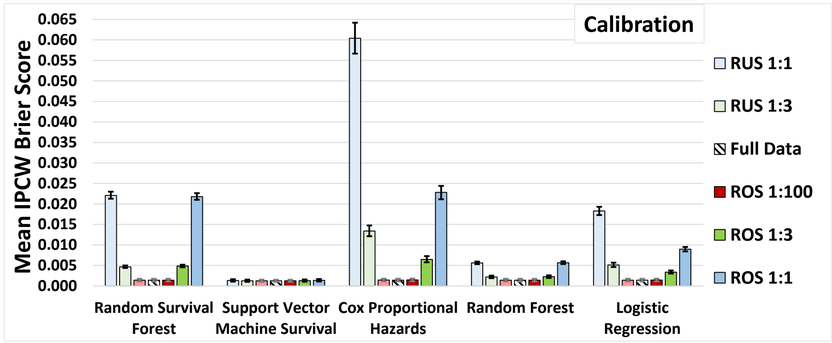


RESULTS

Predictive Performance Across Sampling Strategies: Opioid Use Disorder

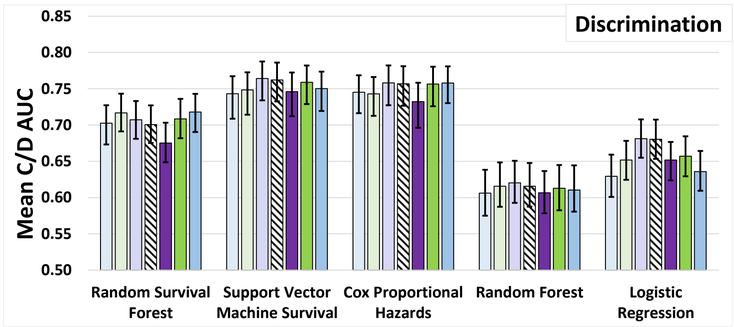
A total of **53,426 Arkansas MC cardholders** met eligibility criteria, of which **364 (0.68%)** received a new **Opioid Use Disorder** diagnosis during the follow-up period.

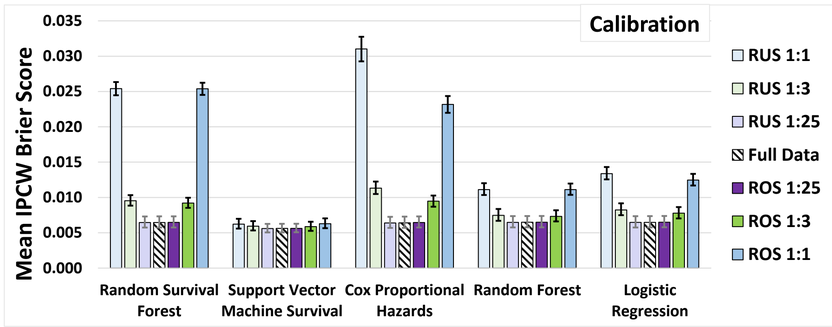




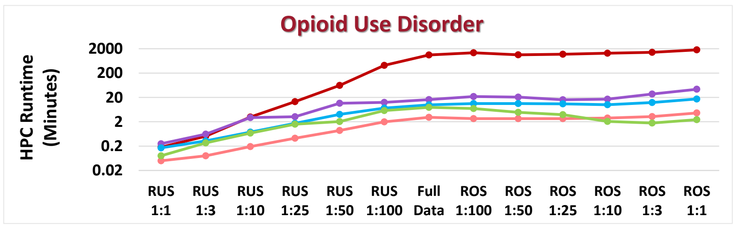
Predictive Performance Across Sampling Strategies: Alcohol Use Disorder

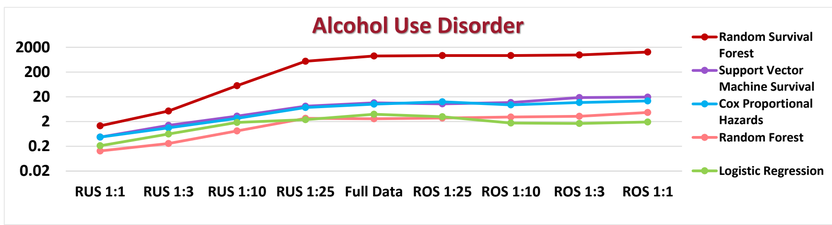
A total of **53,877 Arkansas MC cardholders** met eligibility criteria, of which **1,680 (3.12%)** received a new **Alcohol Use Disorder** diagnosis during the follow-up period.



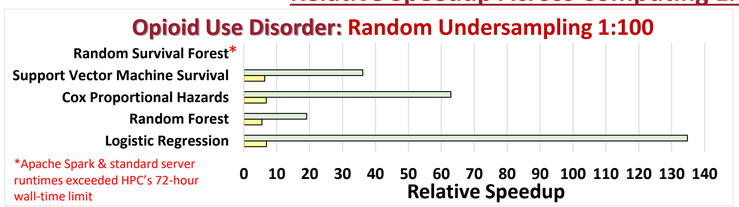


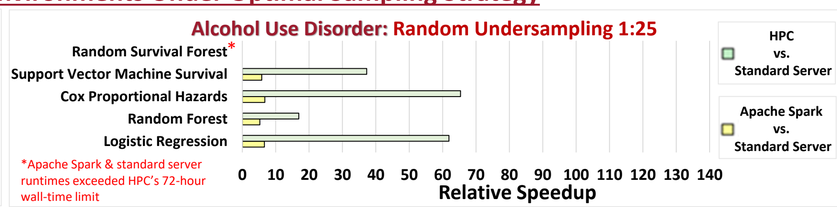
HPC Training Runtimes Across Sampling Strategies





Relative Speedup Across Computing Environments Under Optimal Sampling Strategy





CONCLUSION

- ✓ Sampling strategy had minimal impact on discrimination after hyperparameter tuning.
- ✓ Model choice dominated discriminative performance across all sampling configurations
- ✓ Calibration deteriorated at 1:1–1:3 sampling ratios but stabilized at ≥1:10 imbalance
- ✓ Parallelization substantially improves development speed, particularly when training computationally intensive models
  - ✓ Apache Spark achieved up to ~7× speedup compared to standard computing
  - ✓ High performance computing achieved up to ~190× speedup compared to standard

References

1. Zhang F, Petersen M, Johnson L, Hall J, O'Bryant SE. Hyperparameter Tuning with High Performance Computing Machine Learning for Imbalanced Alzheimer's Disease Data. Appl Sci. 2022 Jul;12(13):6670–6670. doi:10.3390/app12136670 PubMed PMID: 36381541.
2. Hasanin T, Khoshgoftaar TM, Leevy JL, Bauder RA. Severely imbalanced Big Data challenges: investigating data sampling approaches. J Big Data. 2019 Nov;6(1). doi:10.1186/s40537-019-0274-4
3. Kim M, Hwang KB. An empirical evaluation of sampling methods for the classification of imbalanced data. PLoS ONE. 2022 Jul;17(7). doi:10.1371/journal.pone.0271260 PubMed PMID: 35901023.
4. Swart WK de, Loog M, Krijthe JH. A comparative study of methods for dynamic survival analysis. Front Neurol. 2025 Feb;16:1504535–1504535. doi:10.3389/fneur.2025.1504535 PubMed PMID: 40040908.
5. Arkansas All-Payer Claims Database. Welcome to the Arkansas All-Payer Claims Database (APCD). https://www.arkansasapcd.net/Home/.
6. Boerner T, Deems S, Furlani TR, Knuth SL, Towns J. ACCESS: Advancing Innovation. Pract Exp Adv Res Comput. 2023 Jul;173–6. doi:10.1145/3569951.3597559