Developing a Feature Selection Workflow for Variable-Rich Data: A Case Study Utilizing Claims Data to Build Classifiers for the Prediction of Opioid Use Disorder Among Persons Authorized to Purchase Medical Cannabis

MSR2

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

- **High-dimensional (variable-rich) data** in predictive analytics is prone to **overfitting** and makes prediction tasks more difficult due to **data sparsity** and **increased computational complexity**.[1]
- In recent years, more **feature selection workflows** have been developed and proposed as tools to help **optimize the feature space** for prediction tasks in **big healthcare data analytics**.[2, 3]
- However, to our knowledge, no big healthcare data feature selection workflow has been proposed for **time-dependent prediction** informed by the most recent feature values prior to each prediction time window (i.e. **a time-updating feature space**) and most feature selection approaches do not prioritize features based prior evidence

Objective: To develop and apply a feature selection workflow to a high-dimensional, person-time period dataset to select features for opioid use disorder (OUD) risk prediction within 90 days.

METHODS

Data Source

- Statewide health insurance claims data was utilized from the **Arkansas All-Payer Claims Database (AR-APCD)** between November 2018 – December 2023.[4]

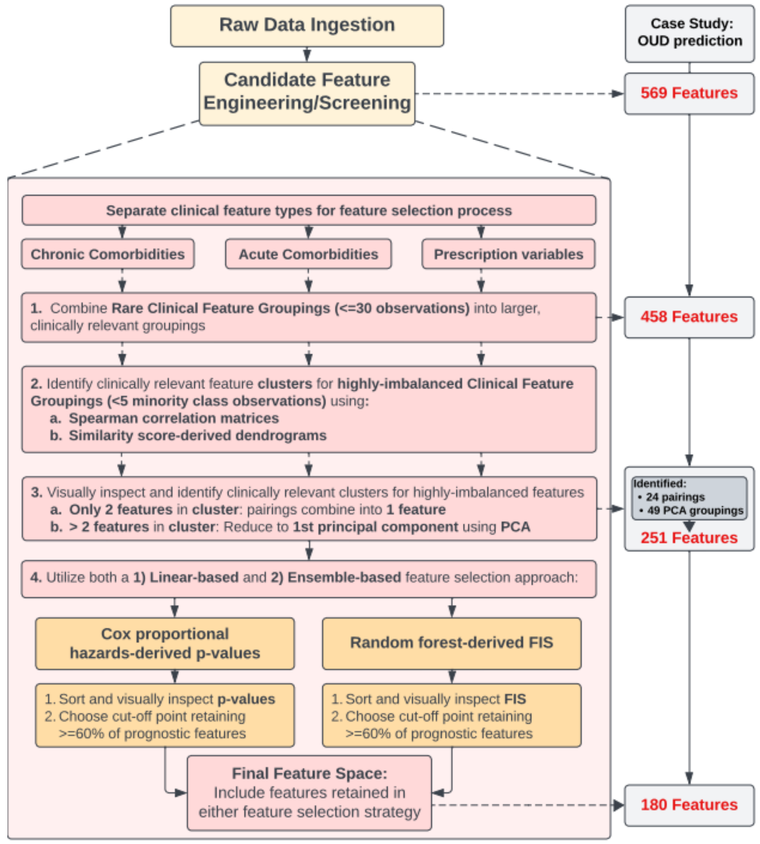
Case study Sample

- **Subjects:** Insured (medical + pharmacy benefits), adult (≥ 18 years old) Arkansas MMJ Cardholders without a recent history of OUD in the past 6 months.
- **Data structure:** Person-period dataset (subject follow-up split into 90-day time intervals), where OUD prediction for each time interval is informed by prior 6 months of features

Engineered Feature Categories

- **Demographics**
  - age, sex, insurance payer type
- **Healthcare Utilization**
  - E.g. primary care provider visit count, cumulative out-of-pocket costs
- **Clinical features**
  - Labeled **prognostic** if evidenced by prior literature**,** labeled **agnostic** otherwise
  - **Prescription Characteristics** (Categorized using First Databank (**FDB**) therapeutic classes)[5]
  - **Comorbidities** (Categorized using Clinical Classifications Software Refined (**CCSR**))[6]
    - Utilized Chronic Condition Indicator Refined (CCIR) to identify "**acute**" and "**chronic**" CCSR-based groupings[7]
      - **Acute condition**
        - **< 50%** ICD-10-CM codes in **CCSR** groupings with **CCIR** flag)
        - Only count in the time-interval(s) the condition was identified
      - **Chronic condition**
        - **≥ 50%** ICD-10-CM codes in **CCSR** groupings with **CCIR** flag)
        - Count in the time-interval the condition was initially identified and carry forward to all future time-intervals.

FEATURE SELECTION WORKFLOW

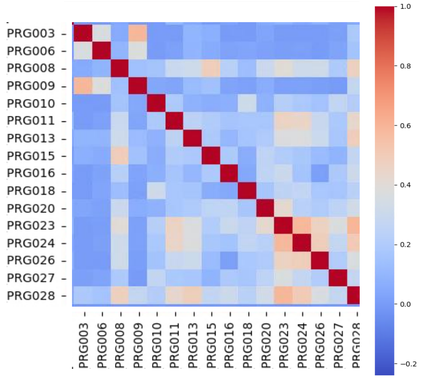


FIS = Feature Importance Scores, OUD = Opioid Use Disorder, PCA = Principal Component Analysis

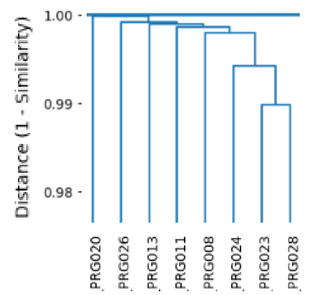
Variable Clustering Demonstration (Ex. Acute Pregnancy Conditions)

- **Identified CCSR categories:** PRG003, PRG006, PRG008, PRG009, PRG010, PRG011, PRG013, PRG015, PRG016, PRG018, PRG020, PRG023, PRG024, PRG026, PRG027, PRG028
- Each category contained **> 30 observations (step 1)** in the cohort overall but contained **<5 observation in the minority class (step 2)**
- After viewing clustering results, **principal component analysis (PCA)** was used to reduce these features to their 1st principal component (labeled "Acute pregnancy conditions") **(step 3)**

Spearman Correlation Matrix



Similarity Score-Derived Dendrogram



Cut-point visualization: Linear and Ensemble-based feature selection results

1. Random forest-derived feature importance scores

Prognostic variables retained: 38 (79.17%)		
Total variables retained: 156 (63.16%)	**FIS cut-point:**	0.0020
Prognostic variables excluded: 10 (20.83%)		
Total variables excluded: 91 (36.84%)		

2. Cox proportional hazards-derived p-values

Prognostic variables retained: 32 (66.70%)		
Total variables retained: 93 (37.65%)	**p-value cut-point**	0.6714
Prognostic variables excluded: 12 (33.30%)		
Total variables excluded: 154 (62.35%)		

Final feature space
(features retained in either feature selection strategy)

Prognostic variables retained: 45 (93.75%)	
Total variables retained: 180 (72.87%)	
Prognostic variables excluded: 3 (6.25%)	
Total variables excluded: 247 (27.13%)	

CONCLUSION

- ✓ **The feature count of the case study dataset was reduced from 569 to a final feature space of 180 while maintaining clinical interpretability for each feature.**
- ✓ **A feature selection workflow leveraging clinical expertise with a comprehensive sequential dimensionality reduction approach is an effective way to reduce high-dimensionality while maintaining a clinically meaningful feature space.**

References

1. Berisha V., Krantsevich C., Hahn P.R., et al. Digital medicine and the curse of dimensionality. NPJ Digit Med. 2021;4(1):153. doi:10.1038/s41746-021-00521-5
2. Wang, H., Zhang, M., Mai, L. et al. An effective multi-step feature selection framework for clinical outcome prediction using electronic medical records. BMC Med Inform Decis Mak. 2025; 25 (84). doi:10.1186/s12911-025-02922-y
3. Mahajan, A., Kaushik, B., Rahmani, M. K. I., and Banga, A. S. A Hybrid Feature Selection and Ensemble Stacked Learning Models on Multi-Variant CVD Datasets for Effective Classification. IEEE Access. 2021; 12: 87023-87038. doi: 10.1109/ACCESS.2024.3412077.
4. Arkansas All-Payer Claims Database. Welcome to the Arkansas All-Payer Claims Database (APCD). https://www.arkansasapcd.net/Home/.
5. First Databank. Drug claims processing: Decisions for financial success. https://www.fdbhealth.com/applications/drug-claims-processing.
6. Agency for Healthcare Research and Quality. Clinical Classifications Software Refined (CCSR). https://hcup-us.ahrq.gov/toolssoftware/ccsr/ccs_refined.jsp.
7. Agency for Healthcare Research and Quality. Chronic Conditions Indicator Refined (CCSR). https://hcup-us.ahrq.gov/toolssoftware/ccsr/ccs_refined.jsp.