

## Apache Superset for rapid exploration tools

Exploring Stroke and Cognitive Impairment in the TMU-CRD OMOP CDM Dataset Using Apache Superset

**Background:** This study utilizes Apache Superset, an open-source BI tool as an alternative to Shiny App, to explore stroke and cognitive impairment data within the OMOP OHDSI framework, specifically leveraging the TMU Clinical Research Database (TMU-CRD), and to explore the OHDSI's PLP study result as well. **Methods:** Exploratory data analysis conducted with Superset's intuitive drag-and-drop interface and SQL capabilities for dashboard creation. Data queries were built using OHDSI's QueryLibrary, facilitating analysis of comorbidities and demographic trends, extracted from TMU-CRD CDM database. Visualization for PLP Results were derived from OHDSI's PLP package result, including XGBoost and LASSO regression, revealing predictive insights from patient features.

10 rows

ndardized\_mean\_diff

옮 후 🚺 🚦



**Results:** Some tabular and charts were

presented interactively, and some

adjustments could be easily facilitated

race_concept_id	INTEGER	Simple partial seizure	9		CONDITIONA	L FORMATTING
ethnicity_concept_id	INTEGER	Furuncle of face	54		+ Add nei	w color formatter
provider_id	INTEGER	Aphthous ulcer of mouth	627			
care_site_id person_source_value	INTEGER NVARCHAR	Cystoid macular retinal degeneration	31			
gender_source_value gender_source_concept_id	NVARCHAR INTEGER	Closed fracture proximal tibia, lateral condyle	4			
race_source_value	NVARCHAR	Intervertebral disc disorder of thoracic region with myelopathy	12			
ethnicity_source_value	NVARCHAR	Open wound of toe with complication	41			
condition era	INTEGER	Acute perichondritis of right external ear	2			
condition ora id	INITEOED	A Constant of Constant Constant	0000			UPDATE CHAR



Superset Dashboards	Charts	Datasets SQL +					+ - Settings	
Evaluation Metrics 🕸 👳	Added to 1	dashboard 🥂 admin user 🖉 2 hours ago					SAVE	
Chart Source		DATA CUSTOMIZE		6 rov	6 rows 00:00:00.39			
main.evaluation_metrics		TREEMAP 1/2 Jul 1 H III 4k (	XGBoost: 2.52 Train: 0.954			Lasso LR: 2.1 Train: 0.953 Test: 0.5631		
Search Metrics & Columns		View all chart	<u>s</u>					
Metrics Showing 1 of 1	^	Query /						
f(x) COUNT(*) 🗄		DIMENSIONS						
Columns Showing 21 of 21	^	× abc model   × abc evaluation   + Drop columns here or click	Test: 0.8017			CV: 0.5791		
abc model		METRIC						
abc evaluation		× f(x) AUROC >						
# population_size		ROWLIMIT		RESULTS SAMPLES				
# outcome_count		1000	RESULTS SAMPLES					
# AUROC     ::       # lower_AUROC     ::			O Search				6 rows	
		SORT BY METRIC						
# upper_AUROC		FILTERS	model 😂	evaluation 💲		AUROC		
# AUPRC	::	x covariate_name IN ('drug_era grou 🔺 >	Lasso LR	CV		0.579140488886542		
# brier_score	::	+ Drop columns/metrics here or click	Lasso LR	Test		0.56309066864497		
# brier_score_scaled	::		Lasso LR	Train		0.952985586822238		
# Eavg			XGBoost	cv		0.761193105997025		
# E90			XGBoost	Test		0.801696610005483		
# Emax		UPDATE CHART	XGBoost	frain		0.953965945067169		

through internal SQL Lab across multiple databases.

**Conclusion:** This approach showcases

Apache Superset's flexibility and

accessibility for exploring large-scale

health datasets. Its BI capabilities

empower researchers to visualize and interpret complex patterns in stroke and cognitive impairment, paving the way for

further clinical insights.

Evaluation Metrics		0 0 0	Model Covariates	:	-	Evaluation Metrics				
XGBoost: 2.52 Train: 0.954	Test: 0.8017	Lasso LR: 2.1 Train: 0.953	CV: 0.5791	model © covariate_name © -Noboost - comantor_era group during day -505 through o days relative to index. Ear, nose and throat 	standardized_mean_diff =	Ŧ	XGBoost: 2.52 Train: 0.954	Test: 0.8017	Lasso LR: 2.1 Train: 0.953	CI
				XGBoost condition_era group during day -365 through 0 days relative to index: Organic mental disorder	0.2457266773943049					
			XGBoost condition_era group during day -365 through 0 days relative to index: Type 2 diabetes mellitus without complication	0.4291332450634706						
		XGBoost condition_era group during day -30 through 0 days relative to index: Osteoarthritis	0.3242586816708152							
		XGBoost condition_era group during day -30 through 0 days relative to index: Intracranial hemorrhage following injury	0.6013523907962188							
		XGBoost race = Unknown	0.4035913566909756							
CV: 0.7612		XGBoost race = Taiwanese	-0.28826389345417847		CV: 0.7612		Tort: 0 5621			
	Test: 0.5631		XGBoost index year: 2019	0.632047443645858	345858			1651. 0.3031		
				XGBoost index year: 2020	0.8201535204561665					
				XGBoost drug_era group during day -30 through 0 days relative to index: Insulins and analogues for	0.6634759365049383					
Covariate Boxplo	t				:		<b>Irug_era group d</b> Max: 0.1195 Brd Quartile: 0.257 Mean: 0.2571 Median: 0.2571	during day -30 through 0 da	ys relative to index: Ins	ulins a
2.5						1 N	st Quartile: 0.257 /lin: 0.3947	1		
2						#	Observations: 2 Outliers: 2			
1.5							0.25 —			







**Limitation:** While Apache Superset excels in flexibility and ease of use, it faces challenges in replicating

standard statistical and evaluation metrics curves (e.g., ROC or calibration curves) commonly produced in

dedicated statistical software. These limitations may require external tools or programming to supplement

Superset's functionality for advanced model evaluations.



Muhammad Solihuddin Muhtar<sup>1</sup>, Nguyen Phung-Anh<sup>2,3,4,5</sup> Jason C. Hsu<sup>1,3,4,5</sup> Min-Huei

- 1. International Ph.D. Program in Biotech and Healthcare Management, College of Management, Taipei Medical University, Taipei, Taiwan
- 2. Graduate Institute of Data Science, College of Management, Taipei Medical University, Taipei, Taiwan
- 3. Clinical Data Center, Office of Data Science, Taipei Medical University, Taipei, Taiwan
- 4. Research Center of Health Care Industry Data Science, College of Management, Taipei Medical University, Taipei, Taiwan
- 5. Clinical Big Data Research Center, Taipei Medical University Hospital, Taipei Medical University, Taipei, Taiwan