Gleaning Novel Insights From Real-World Data: A Machine-Learning Guided Analytical Framework

Poster number: PT20

Scan QR

Handing Xie,[1] Wei-Hsuan "Jenny" Lo-Ciganic,[2] Jing Wang,[3] Marko Mychaskiw,[1] Ying Zhang,[1] Marc Tian[1]

[1]Teva Branded Pharmaceutical Products R&D, Inc., West Chester, PA, United States; [2]University of Pittsburgh, Pittsburgh, PA, United States; [3]KMK Consulting, Inc., Morristown, NJ, United States

Objective: To develop an analytical framework leveraging machine learning (ML) to identify reliable predictors and provide novel clinical insights using real-world data (RWD)

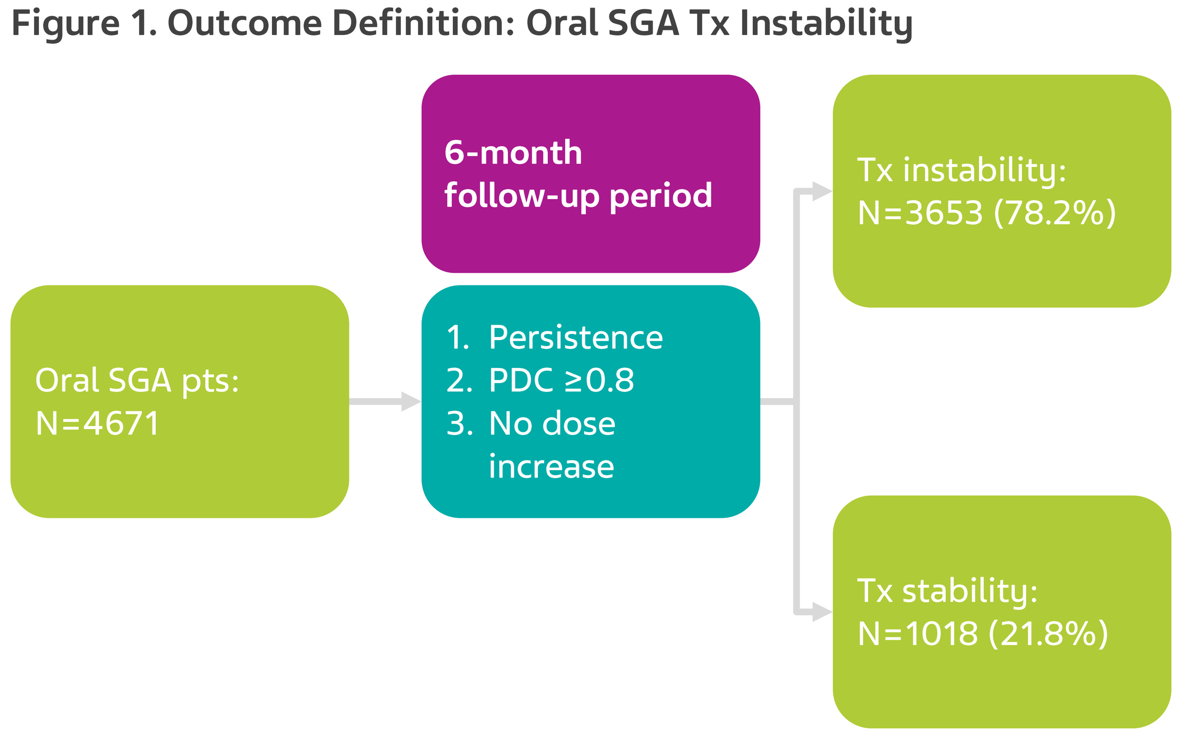
Background and Case Study Background

- Healthcare claims databases contain vast amount of data and insights so applying a fit-for-purpose analytical strategy plays a crucial role to leverage the full potential of real-world data to identify hidden and novel insights. Machine Learning (ML) is becoming an essential tool to analyze large number of variables, identify predictors, etc. However,
  - Different ML models give **inconsistent** top predictors for the same outcome
  - ML models' feature "importance metric" is **hard to interpret**
  - The data engineering and analytical flow in claims databases are **highly diverse** and **hard to follow** for researchers
- **Case study background:** Identify predictors for oral, second-generation antipsychotic (SGA) treatment instability
  - Schizophrenia (SCZ) is highly disabling. It affects 1% of the population worldwide
  - Oral SGAs are commonly prescribed for SCZ. However, poor treatment (Tx) stability leads to relapse
  - Long-term injectable SGAs are available, but underused.
  - There is limited knowledge of the risk factors and related mechanism
- **Case study objectives:**
  - Identify **predictors** during pre-treatment period for Tx instability of oral SGAs
  - Understand the **effect of each predictor** on the Tx instability of oral SGAs
  - Build **analytical framework** for ML guided predictors identification/interpretation

Cohort Definition

- Patients with SCZ who initiated oral SGAs from January 2013 to June 2021 in Marketscan® US claims data
- Index event = first oral SGA
- Data eligibility/insurance enrollment during:
  - 1-year pre-index period (baseline period) to extract predictors
  - 6-month post-index period (follow-up period) for outcome measurement

Figure 1. Outcome Definition: Oral SGA Tx Instability



Understanding Models' Performance

Table 2. ROC AUC Comparison Across Models

Features used	Features	Model name	Train data	Test data
Initial Features	1956	Elastic_net	0.64	0.59
Initial Features	1956	Lasso	0.63	0.56
Initial Features	1956	Random_forest	0.63	0.59
Initial Features	1956	XGBoost	0.82	0.60
Route 1 Features	20	Elastic_net	0.61	0.58
Route 1 Features	20	Lasso	0.62	0.58
Route 1 Features	20	Random_forest	0.59	0.59
Route 1 Features	20	XGBoost	0.64	0.59
Route 2 Features	15	Elastic_net	0.61	0.61
Route 2 Features	15	Lasso	0.61	0.60
Route 2 Features	15	Random_forest	0.60	0.60
Route 2 Features	15	XGBoost	0.67	0.60

Logistic Regression for Predictor Interpretation

Figure 5. Odds Ratio (OR) for 15 Identified Predictors

Predictor	OR	95% CI
Drug abuse, Elixhauser comorbidity, binary	1.58	(1.19, 2.10)
Cannabis-related disorder, binary	1.28	(0.86, 1.90)
Emergency room visits	1.08	(1.03, 1.13)
Amphetamine salt combination use	1.04	(0.99, 1.10)
SNRI use	1.04	(0.98, 1.10)
Sleep disorders	1.00	(0.94, 1.06)
Cannabis-related disorder	1.00	(0.96, 1.03)
Schizophrenia	0.99	(0.98, 1.00)
Schizoaffective disorders	0.98	(0.96, 1.00)
Establised patient office visits	0.97	(0.95, 1.00)
Obsessive-compulsive disorder	0.97	(0.91, 1.03)
Collection of venous blood by venipuncture	0.94	(0.90, 0.98)
Benztropine mesylate use	0.93	(0.88, 0.97)
Psychotherapy office visits	0.92	(0.86, 0.99)
Psychotherapy inpatient visits	0.92	(0.87, 0.97)

Conclusion

An analytical framework is developed to:

- Identify novel/reliable predictors for outcome in claims database
- Explain the effects of each predictor to the outcome.

Top 3 significant predictors oral SGA Tx instability:

- Drug abuse (aOR=1.58)
- More frequent emergency department visits (aOR=1.08)
- Less frequent psychotherapy (aOR=0.92)

Future efforts: In discussion with expert psychiatrists to better understand clinical implication and potentially build a prediction tool to improve real-world clinical practice

Table 1. Feature Engineering Structure

Feature category	Feature (n)
Demographic	12
Diagnosis, Inpatient, 3-digit ICD-10	502
Diagnosis, Outpatient, 3-digit ICD-10	2386
Drug utilization	12
Elixhauser comorbidity index	31
Generic name drug usage	1684
Healthcare utilization	10
Procedure, Inpatient, CPT/HCPCS	2498
Procedure, Outpatient, CPT/HCPCS	7030
Grand total	**14,165**

Each Dx, Px, and Rx use has 2 features created:
Numeric version to evaluate frequency of use and binary version to evaluate any use at all. **Initial features** for model training = Dx, Px, Rx features with ≥1% prevalence in training data + other features.

Figure 2. ML Models' Development

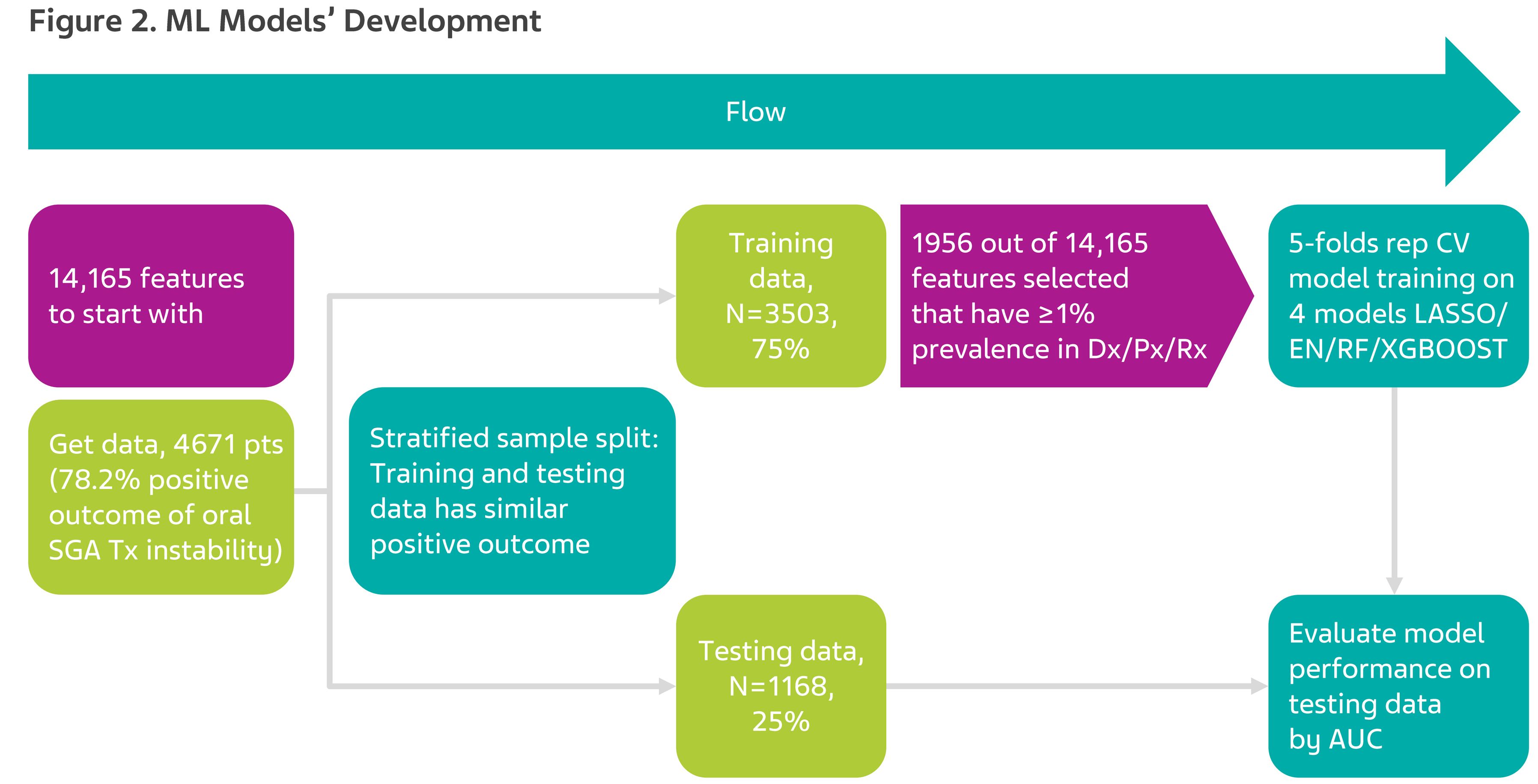


Figure 3. ROC Comparison Among Route 2 Models

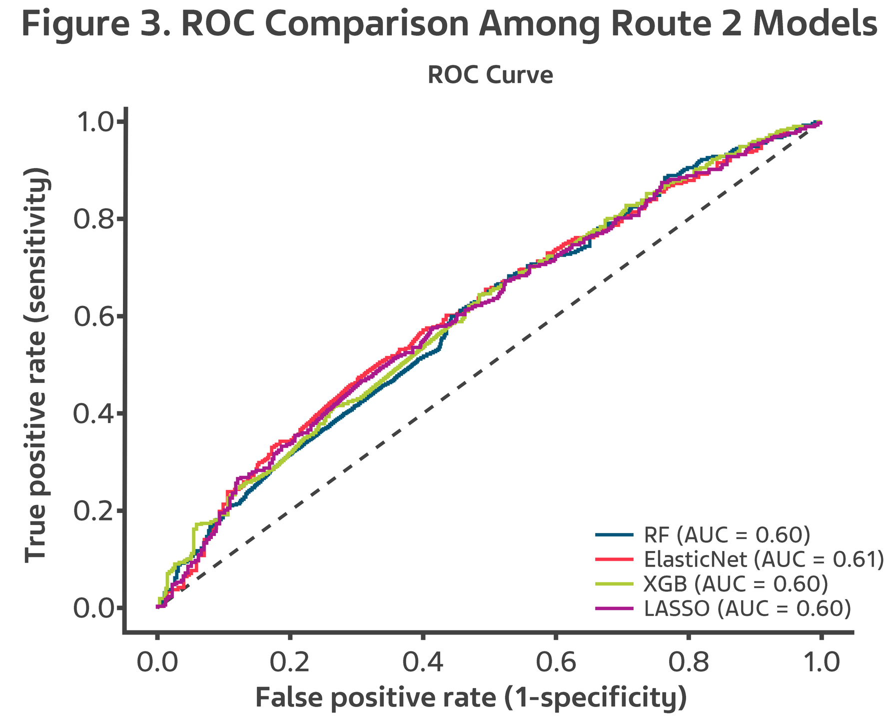
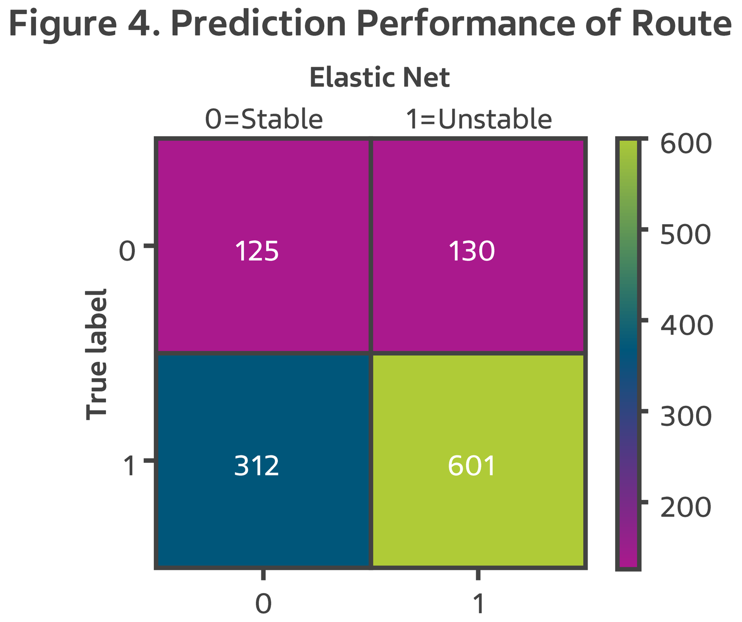


Figure 4. Prediction Performance of Route 2 Elastic Net Model



ROC AUC:	0.61
Sensitivity:	0.66
Specificity:	0.49
Precision (or PPV):	0.82
NPV:	0.29
Accuracy:	0.62
Total patients (test dataset):	1168 (100%)
Total actual positive (unstable):	913 (78%)
Total actual negative (stable):	255 (22%)
Total predicted positive (unstable):	731 (63%)
Total predicted negative (stable):	437 (37%)

The model showed high precision (0.82) and sensitivity (0.66) in predicting oral SGA Tx instability

Models Tuning & Feature Selection

Round 1:

Train 4 ML models: LASSO, elastic net, random forest, and XGBoost with initial feature input. Then,

- **Route 1:** Identify the top 20 features from best performing individual model – XGBoost
- **Route 2:** Identify features that showed up as top 20 features from at least 2 of the 4 ML models (15 features selected)

Round 2:

- Re-train each ML model using the reduced list of features identified by route 1 & 2, then evaluate/compare the performance through AUC

Identify the model from rounds 1 & 2 with highest AUC in the testing dataset. Identify the associated predictors:

Elastic net model fit with 15 features selected from route 2.

1. Report model diagnostic metrics
2. Sequentially fit univariate and multivariate logistic model with 15 features using whole dataset
3. Identify the predictors that have significant odds ratio observed in both univariate and multivariate logistic models. Work with clinicians to interpret the predictors identified

Acknowledgements
The authors would like to thank Steven Thompson, Sanjay Gandhi, Kelli Franzenburg, Mark Suett from Teva Pharmaceuticals for analytical and clinical insights regarding the interpretation of the study results.

Disclosures
This study was supported by funding from Teva Branded Pharmaceutical Products R&D, Inc. Wei-Hsuan "Jenny" Lo-Ciganic has received consulting fees and/or honoraria from Teva Pharmaceuticals. Handing Xie, Marko Mychaskiw, Ying Zhang, and Marc Tian are employees and stockholders of Teva Pharmaceuticals. Jing Wang is employees of KMK Consulting, Inc.

Abbreviations
aOR = adjusted odds ratio, AUC = area under the curve, CPT = current procedural terminology, CV = cross-validation, Dx = diagnosis, HCPCS = Healthcare Common Procedure Coding System, ICD = International Classification of Diseases, ML = machine learning, NPV = negative predicted value, OR = odds ratio, PDC = proportion of days covered, PPV = positive predictive value, Pts = patients, Px = procedure, RF = random forest, ROC = receiver operating characteristic, Rx = prescription drug, SCZ = schizophrenia, SGA = second-generation antipsychotic, SNRI = serotonin-noradrenaline reuptake inhibitor, Tx = treatment

References

1. Padula WV, et al. Value Health. 2022;25(7):1063-1080.
2. Reps JM, et al. J Am Med Inform Assoc. 2018;25(8):969-975.

Presented at the Annual Meeting of the International Society for Pharmacoeconomics and Outcomes Research; May 5–8, 2024; Atlanta, GA.