

Clinical Decision Making using Real World Data

Samiul Haque PhD, Sherrine Eid MPH

SAS Institute Inc.

Samiul.Haque@sas.com



Background

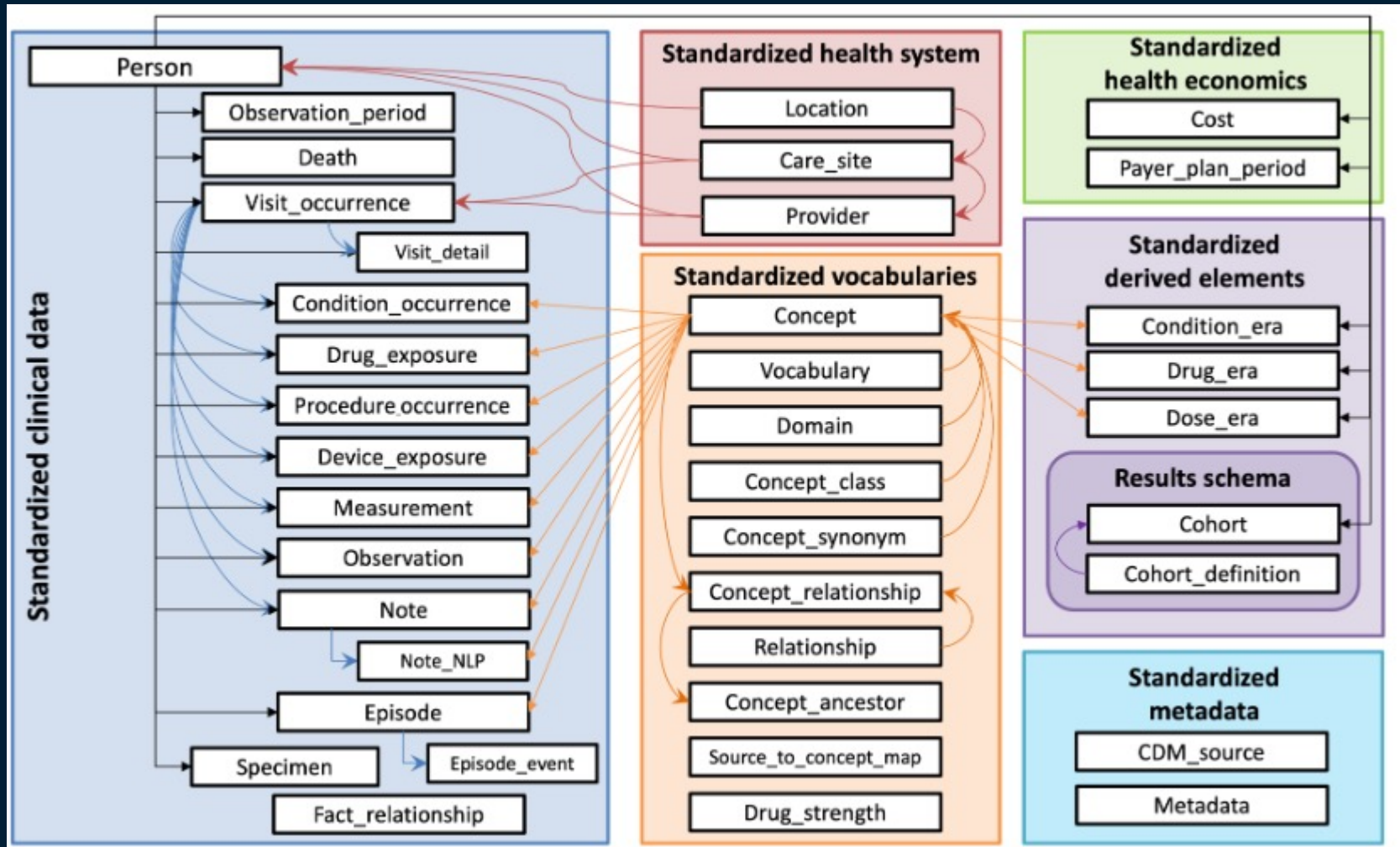


Background

- Patient-level predictions are critical for clinicians to make risk-based and time-sensitive decisions.
- Machine learning-based models can assist clinicians by providing a probabilistic measure of the patient's outcome.
- Training models can be difficult due to lack of data, disparate data sources, and heterogeneity of the patient data.
- OMOP Common Data Model allows systematic analyses of disparate observational database

OMOP - *Observational Medical Outcomes Partnership*

OMOP Common Data Model



<https://ohdsi.github.io/CommonDataModel/>

Clinical Decision-Making Use Case

- In patients that started using NSAIDs for the first time, predict who will develop a gastrointestinal (GI) bleed in the next year.
- The book of OHDSI
- Build AI/ ML model on a High Performance Computing (HPC) environment
- Synthetic dataset (comes with Eunomia R package)
- OMOP common data model (CDM)

NSAIDs = *Non-Steroidal Anti-Inflammatory Drugs*

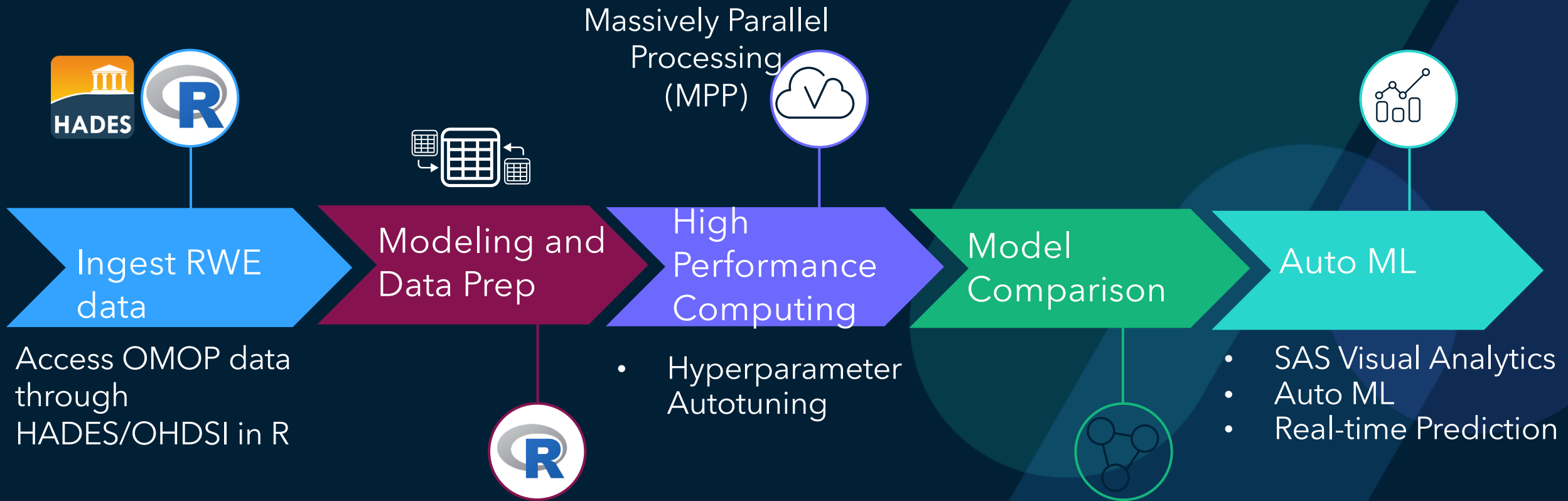
Observational Health Data Sciences and Informatics (OHDSI)

- Multi-stakeholder, interdisciplinary collaborative to bring out the value of health data through large-scale analytics.
- Open-source solutions
 - Extract CDM data
 - Build models

SAS Viya

- Numerous out of the box ML methods
- Supports HPC / MPP (Massively Parallel Processing)
- Integrates seamlessly with open source

Workflow: Clinical Decision Making Use Case



Ingest OMOP Data



##	rowId	covariateId	covariateValue	outcomeId	daysToEvent
## 1	1	4285898210	1	4	NA
## 2	1	372328210	1	4	NA
## 3	1	8507001	1	4	NA
## 4	1	9201923	1	4	NA
## 5	1	1002	33	4	NA
## 6	1	1124957409	1	4	NA

Long table. Not always suitable for ML models

Create Analytics Ready Table

##	rowId	covariateId	covariateValue	outcomeId	daysToEvent
## 1	1	4285898210	1	4	NA
## 2	1	372328210	1	4	NA
## 3	1	8507001	1	4	NA
## 4	1	9201923	1	4	NA
## 5	1	1002	33	4	NA
## 6	1	1124957409	1	4	NA

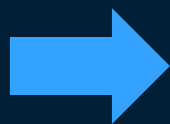
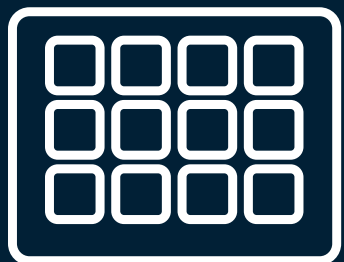


```
cas_df=spread(cas_df,covariateId,covariateValue)
```

##	rowId	outcomeId	daysToEvent	CovariateID_1002	CovariateID_8507001
## 1	1	4	NA	33	1
## 2	2	4	NA	36	0
## 3	3	3	52	41	1
## 4	4	4	NA	41	1
## 5	5	4	NA	42	0
## 6	6	4	NA	40	0

Load Data into Distributed Memory

```
swat::as.casTable(conn, cas_df, casOut = list(name="NSAIDS_GI_BLEED",promote=TRUE,caslib="PUBLIC"))
```



CAS (Cloud Analytic Services)

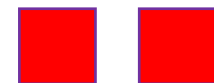
Control



Worker



Worker



...

Worker



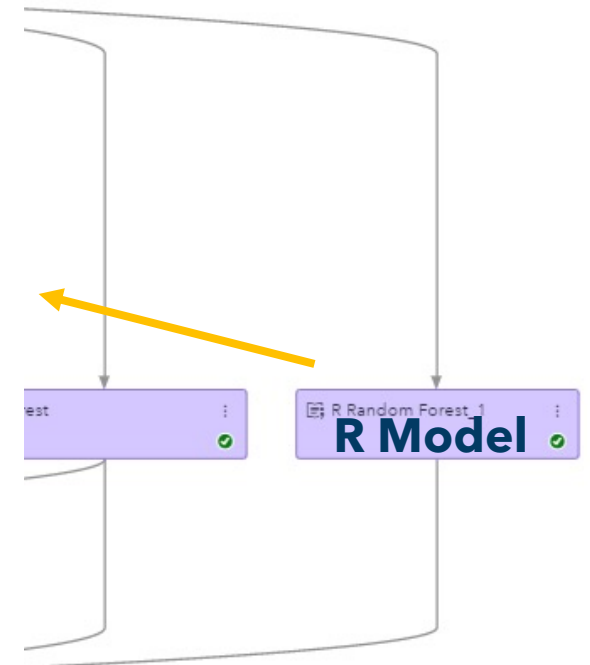
Why ?

- Train Model on Big Data
- Model Tournaments
- Model Auto Tuning
- Model Governance

Model Tournament: Train, Validate & Compare Models

```
1
2 # Builds Random Forest model with 100 trees
3
4 library(randomForest)
5
6 # Fit RandomForest model w/ training data
7 dm_model <- randomForest(dm_model_formula, ntree=100, data=dm_traindf, importance=TRUE)
8
9 # Save MSE plot to PNG
10 png("rpt_forestMsePlot.png")
11 plot(dm_model, main='randomForest MSE Plot')
12 dev.off()
13
14 # Save VariableImportance to CSV
15 write.csv(importance(dm_model), file="rpt_forestIMP.csv", row.names=TRUE)
16
17 # Score full data
18 dm_scoreddf <- data.frame(predict(dm_model, dm_inputdf, type="prob"))
19 colnames(dm_scoreddf) <- c("P_TARGET_GI_BLEED0", "P_TARGET_GI_BLEED1")
```

orted data table



Train & Validate Models

Data Pipelines Pipeline Comparison Insights				
Filter		Data: Validate ▾		Compa
☐	Champion ↓	Registered	Name	Algorithm Name
☒		☒	Forward Logistic Regression	Logistic Regression
☐		☒	Forest	Forest
☐			Neural Network	Neural Network

Identify Important Comorbidities



Real Time Decision Support GUI for Clinician

What values for the most important factors should be used to predict?

age in years _ 1002

39

Ulcerative colitis _ 81893209

0

Peptic ulcer _ 4027663209

1

Esophagitis _ 30753209

1

Polyp of colon _ 4285898209

0

What is the prediction for Patient?

No_GI_Bleed

The predicted Patient, No_GI_Bleed, is the most common Patient value in observed cases. Most observed cases (81.79%) are No_GI_Bleed. The prediction is based on an automatically selected Decision Tree model.

Explanation in Natural Language

Real Time Decision Support GUI for Clinician

What values for the most important factors should be used to predict?

age in years 1002

50

Ulcerative colitis _ 81893209

1

Peptic ulcer _ 4027663209

1

Esophagitis _ 30753209

1

Polyp of colon _ 4285898209

0

What is the prediction for Patient?

GI_Bleed

The predicted Patient, GI_Bleed, is the second most common Patient value in observed cases. Most observed cases (81.79%) are No_GI_Bleed, while 18.21% are GI_Bleed. The prediction is based on an automatically selected Decision Tree model.

Explanation in Natural Language

Summary

- We can leverage RWD for building ML models
- ML models can be used by clinicians for decision support
- Massively parallel processing allows training advanced models by leveraging big data

Lessons Learned

- Do not start from scratch, augment your existing process and incorporate change where needed
- Managing and tracking OS packages + dependencies can be difficult
- Select / design a platform that is robust, open, modular, and governed
- Time is money. Consider the total cost of your analytics life cycle.
 - Data federation & ETL
 - Data quality
 - Model Training
 - Model Deployment
 - Model Management



Thank You!
Samiul.Haque@sas.com