

Applications of ML in Real World Data in Personalized Medicine for Non-Small Cell Lung Cancer Patients

PHUSE US Connect 2025

Samiul Haque (samiul.haque@sas.com)

Sherrine Eid (sherrine.eid@sas.com)



Introduction

Non-Small Cell Lung Cancer

- Lung cancer is the most commonly diagnosed cancer and was identified as leading cause of cancer death in 2018.
 - Most cases are Non-Small Cell Lung Cancer (NSCLC).
- Machine learning methods (ML) and Real World Data (RWD) make personalized medicine a viable possibility to save patient lives.
- Chronic Obstructive Pulmonary Disease (COPD) commonly persists during cancer treatment.
 - COPD requires interventions that could jeopardize the effectiveness of cancer treatment and the patient's prognosis.

Overview

- This work showcases application of advanced analytics and modeling on Real World Data (RWD)
- We have used Symphony Health Data to implement this work
- This work showcases how we can identify personalized treatment plans / interventions by leveraging big data analytics

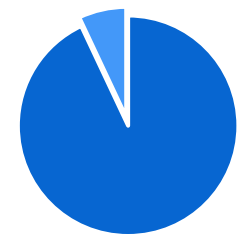
Symphony Health and Data

Symphony Health

- Data Vendor to Healthcare and Life Science Organizations
- Over 317 Million Patient Lives
 - Physician-linked prescription information
 - Entire lifecycle of the claim
 - Tokenized data
 - Patient journeys
 - RWE analytics

Symphony Health

Additional Statistics



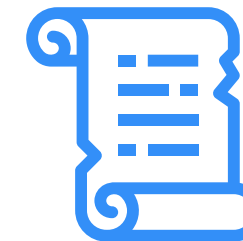
Over **93%**
Rx dispensed in US and territories



Over **12K**
payers tracked thru IDV



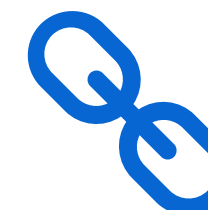
1.9M
unique active practitioners



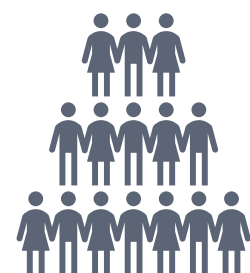
Over **17y**
historical data



5.3 years
average continuous history
for active patients



95% of Rx assigned
to individual health plans



Over **317M**
active patients

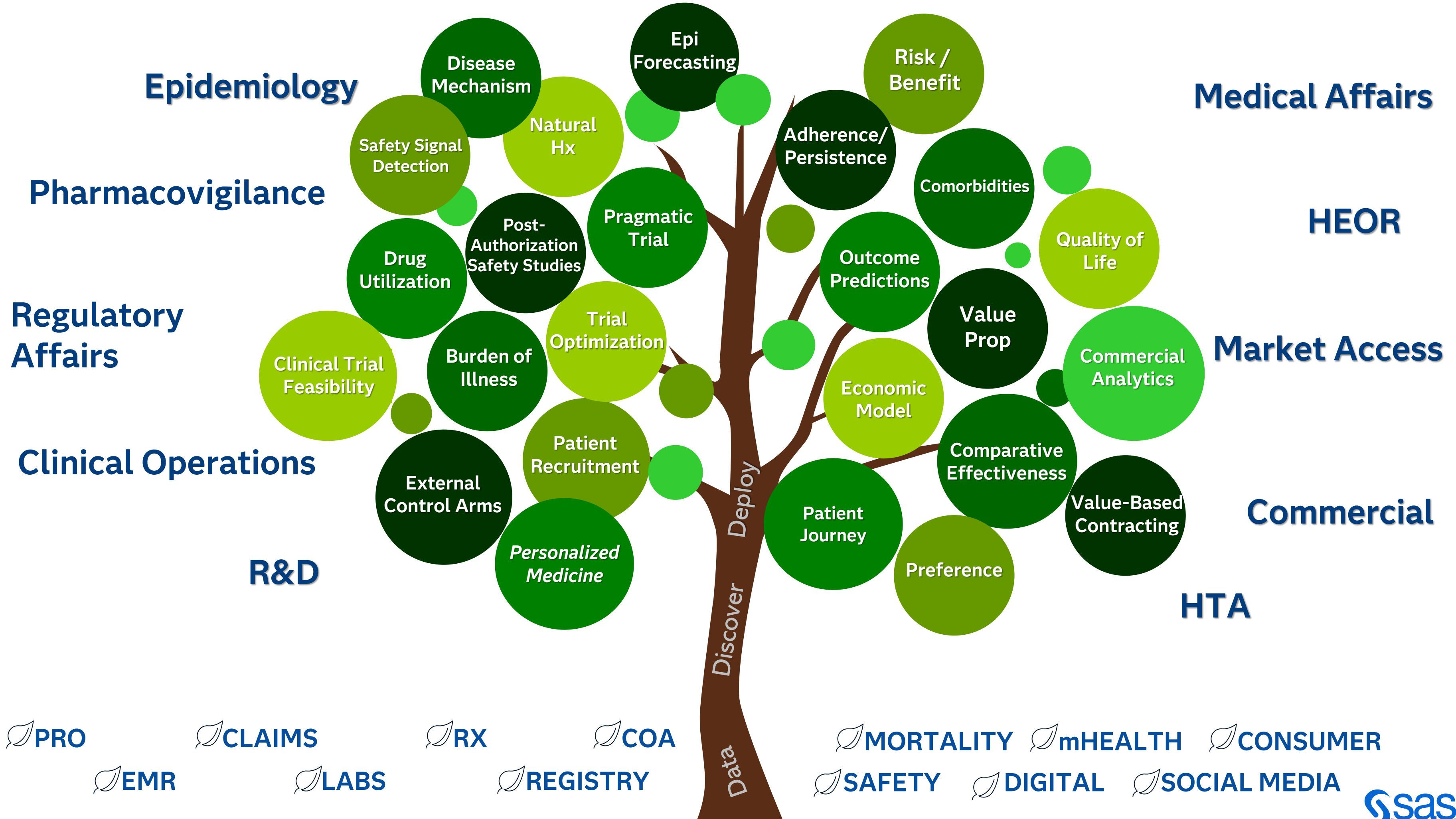
Details of the Data

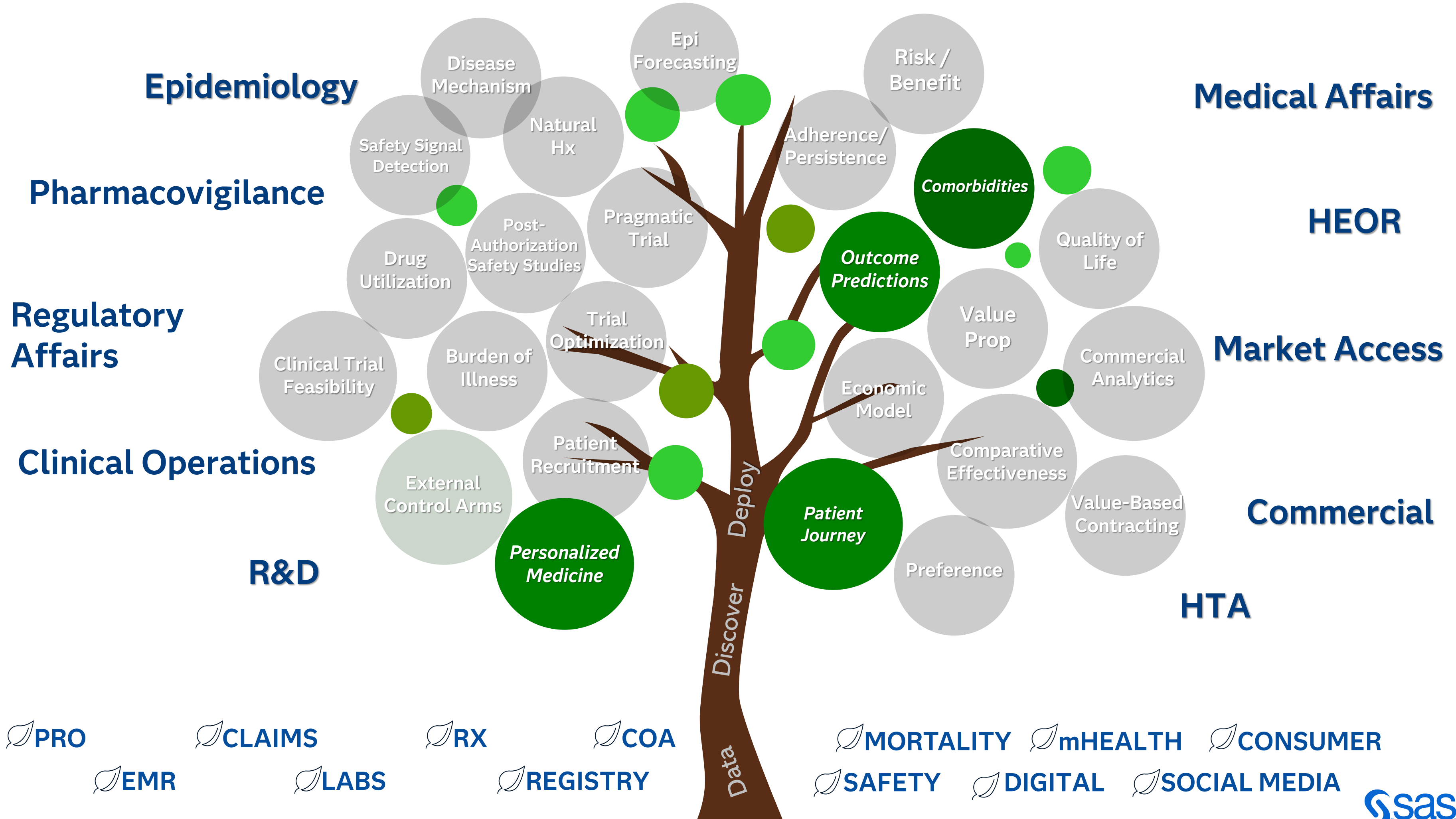
Claims Data ~22.5TB

Integrated Dataverse (IDV®) - most comprehensive source of longitudinal healthcare claims data in industry

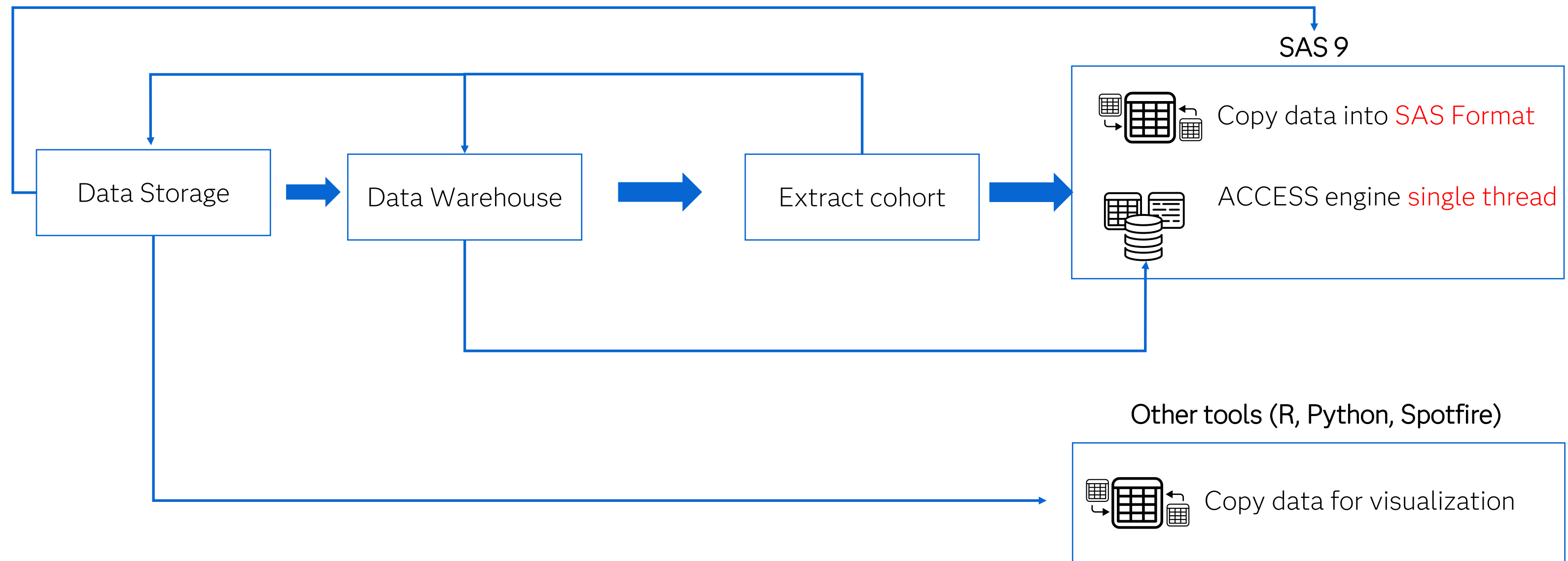
- Demographic, medical, hospital, point-of-sale prescriptions and non-retail invoice data

Over *80 therapeutic areas* covering the most important classes for pharmaceutical manufacturers.



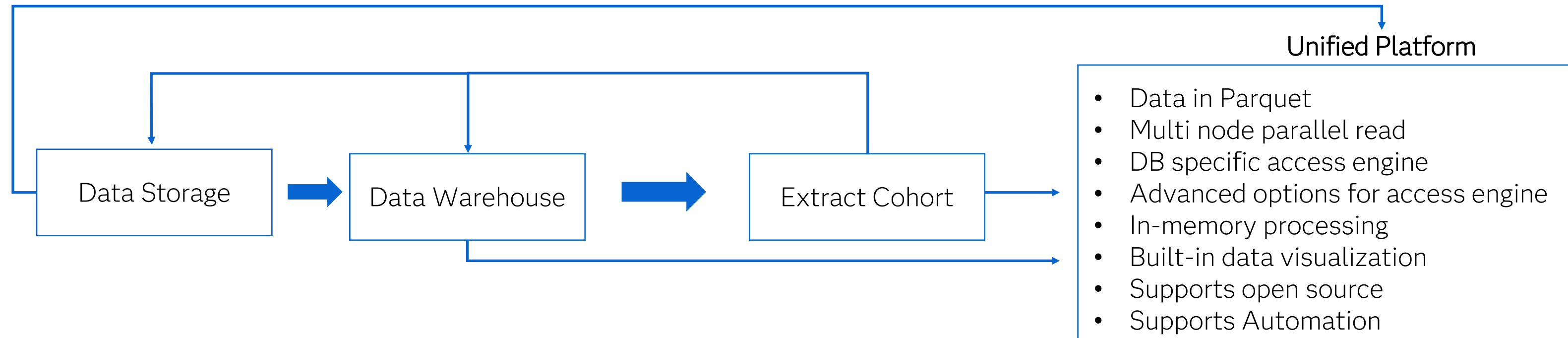


Typical Workflow



- Multiple copies of data
- Security and governance challenges
- Higher compute cost in data warehouse
- Longer run time for data transfer and downstream analyses

Enhanced Workflow



Value:

- Reduce copies of data
- Managed security and governance through Platform
- Maintain data lineage
- Optimize compute cost by leveraging data warehouse + in memory distributed processing
- Faster data transfer and downstream analyses
- Automate end-to-end processes
- Better collaboration
- Application of AI/ML

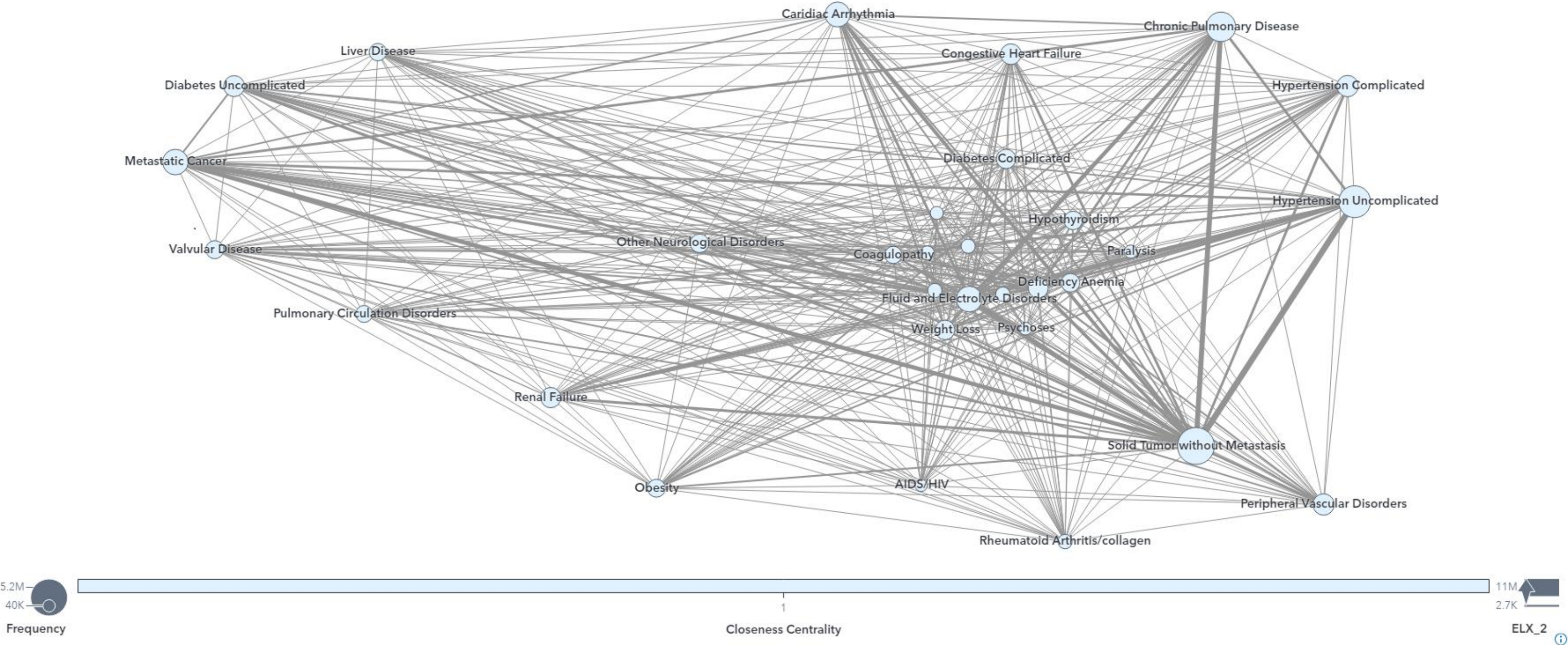
Methods

Analysis Performed Using SAS Viya 4.0

- 1.2 Million NSCLC (ICD-10 C34.*) patients from 2019-2020 using Symphony Health claims data
- Dependent variable: **Metastatic cancer** (MC).
- Independent variables:
 - Comorbidities** (Elixhauser comorbidity risk factor categories)
 - Age**
 - Sex**
 - Drug product.**
- Analytic Methods
 - Automated explanation
 - Path Analysis
 - Logistic regression, gradient boosting, decision tree, random forest and Bayesian networks
 - Network analysis

Results

Network Analysis



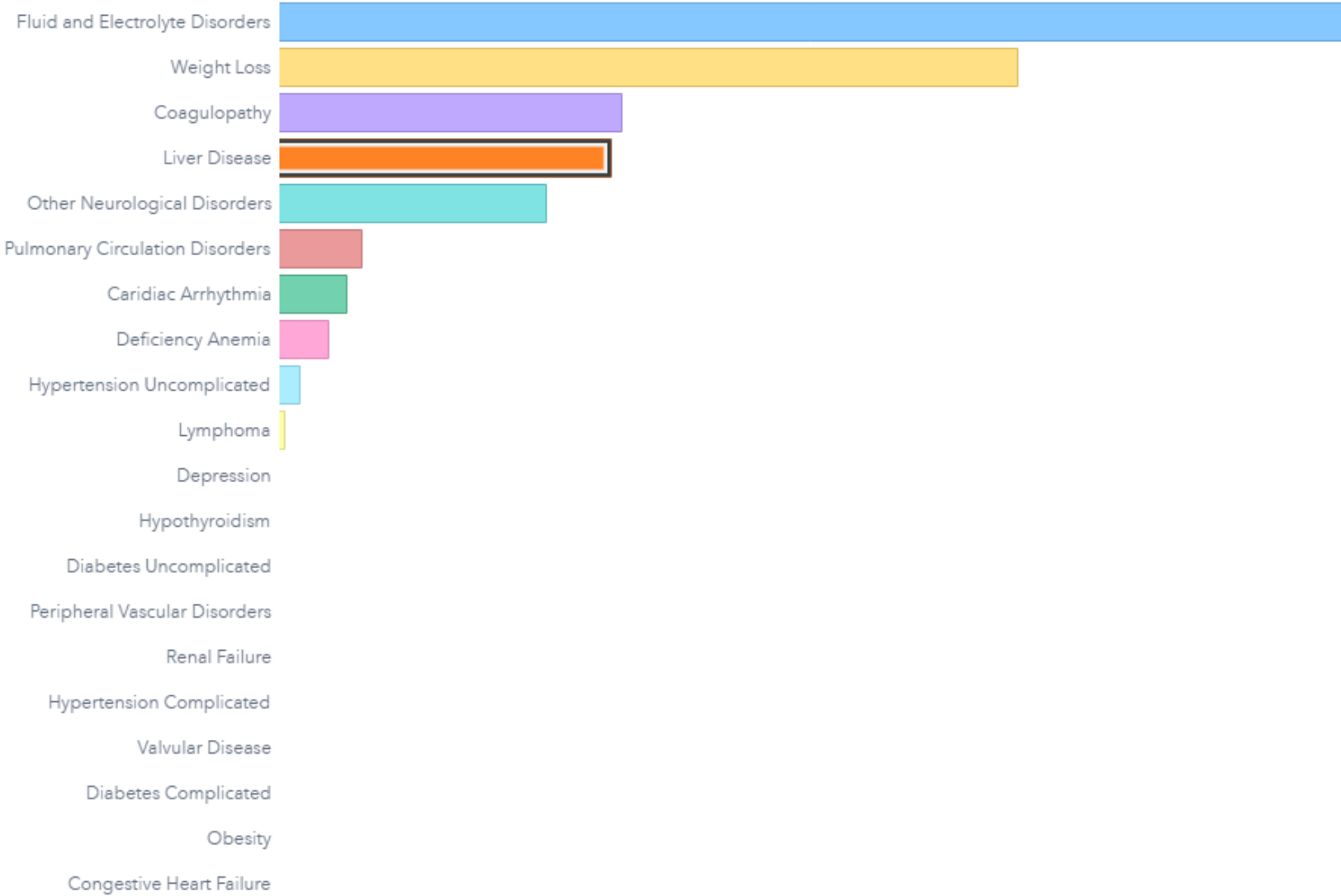
Automated Explanation

What are the characteristics of Metastatic Cancer?

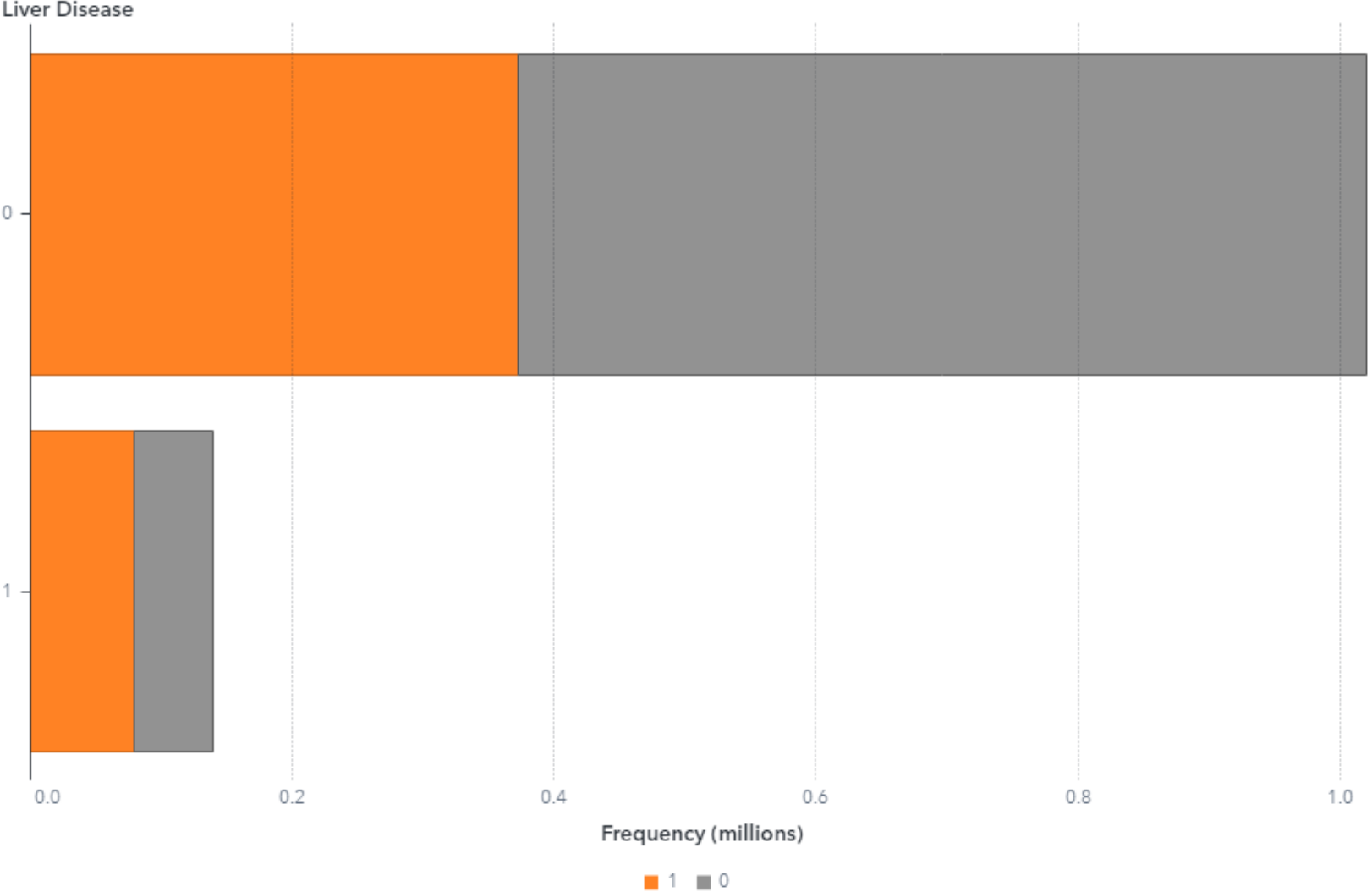
1 is less common at 39.01% (452K of 1.2M). 0 is more common at 60.99%. The three most related factors are Fluid and Electrolyte Disorders, Weight Loss, and Coagulopathy.



What factors are most related to Metastatic Cancer?



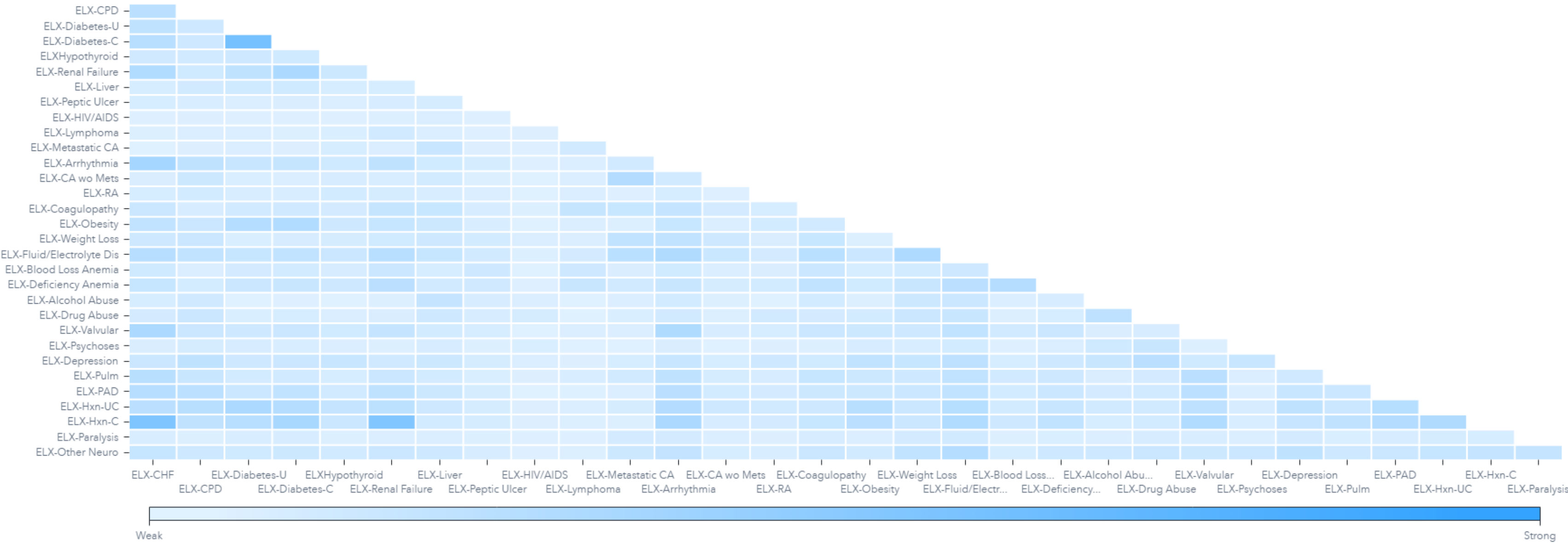
What is the relationship between Metastatic Cancer and Liver Disease?



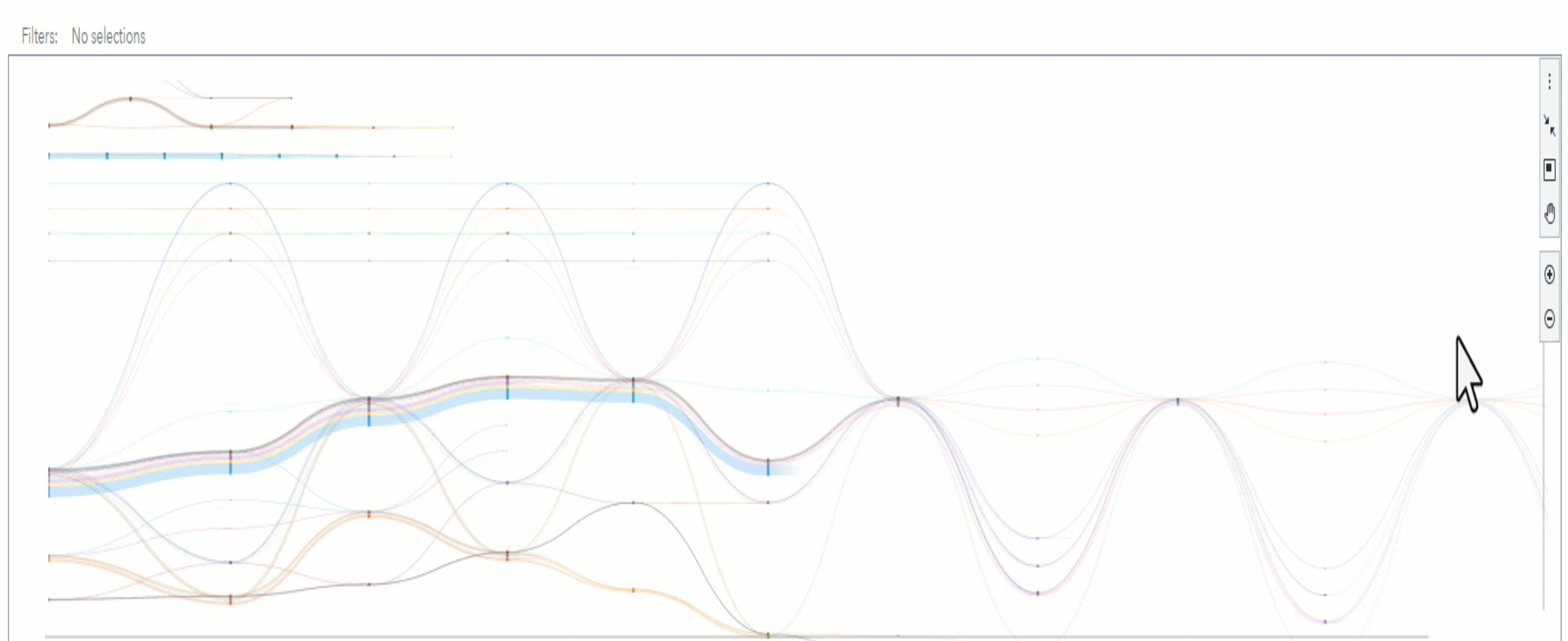
When Liver Disease is 0, the total count of 1 is 373K. When Liver Disease is 1, the total count of 1 is 80K. The most common Liver Disease value is 0.

Correlation Matrix

Correlation of Selected Measures



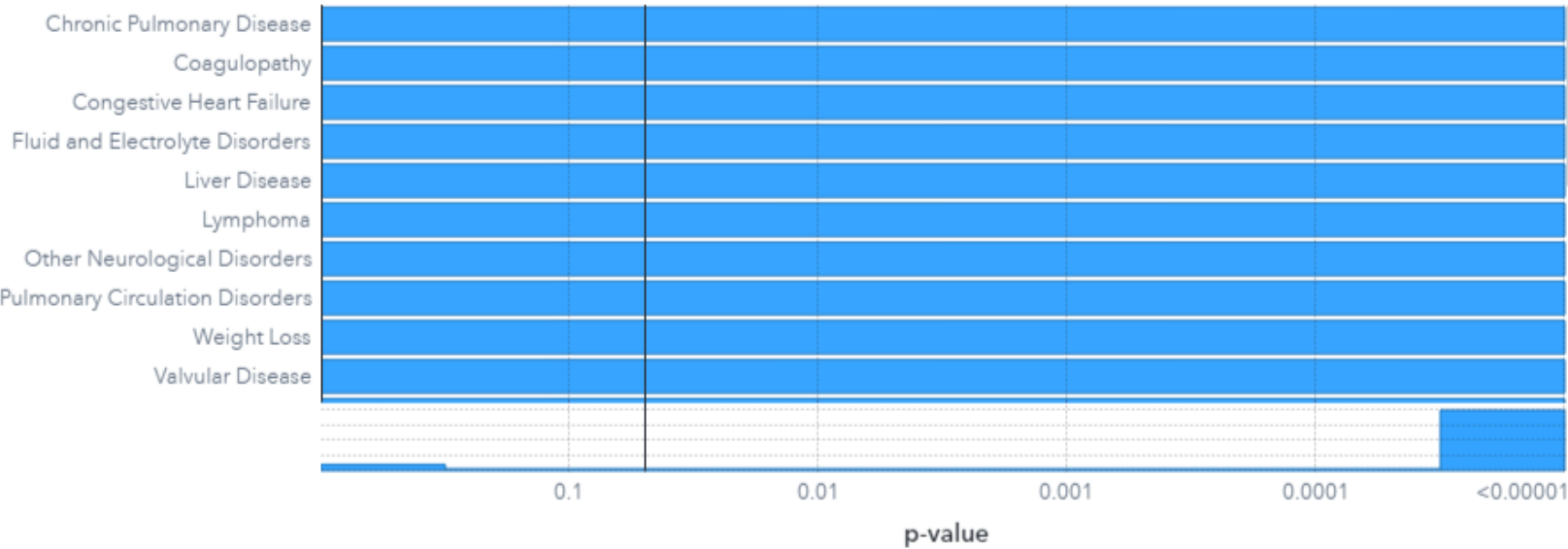
Patient Journey



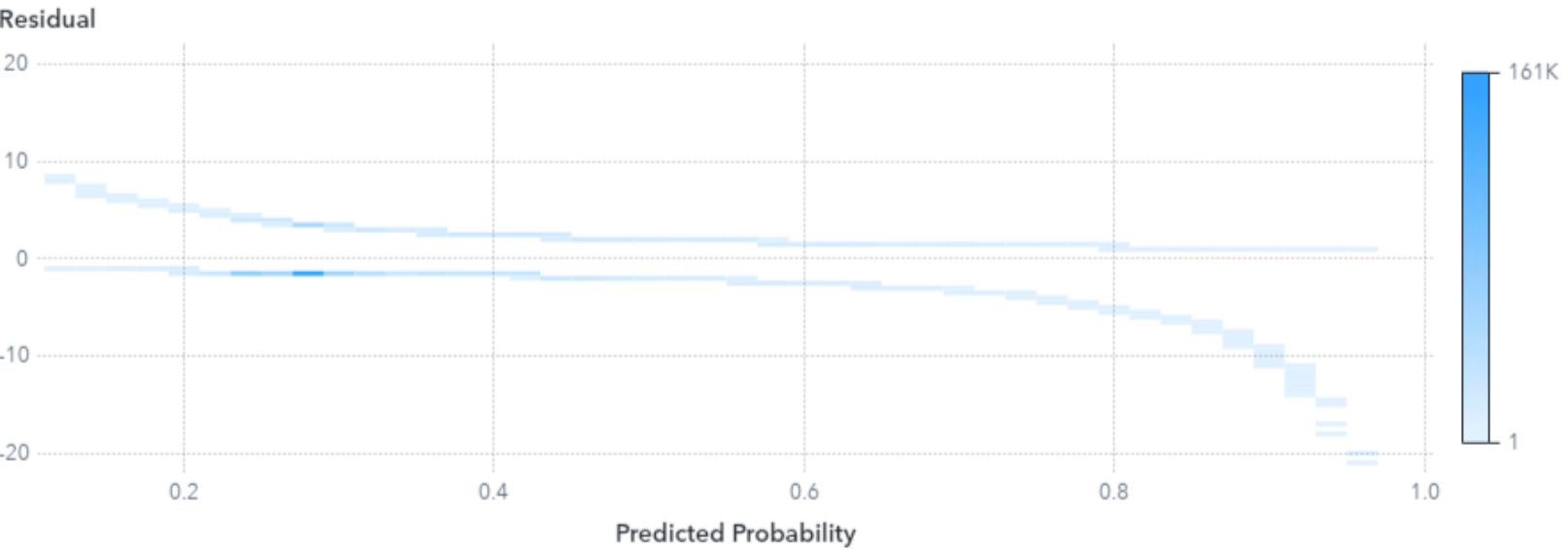
Logistic Regression

Logistic Regression **Metastatic Cancer** Event: 1 ▾ Fit: KS (Youden) 0.2770 ▾ Observations: 1.2M of 1.2M

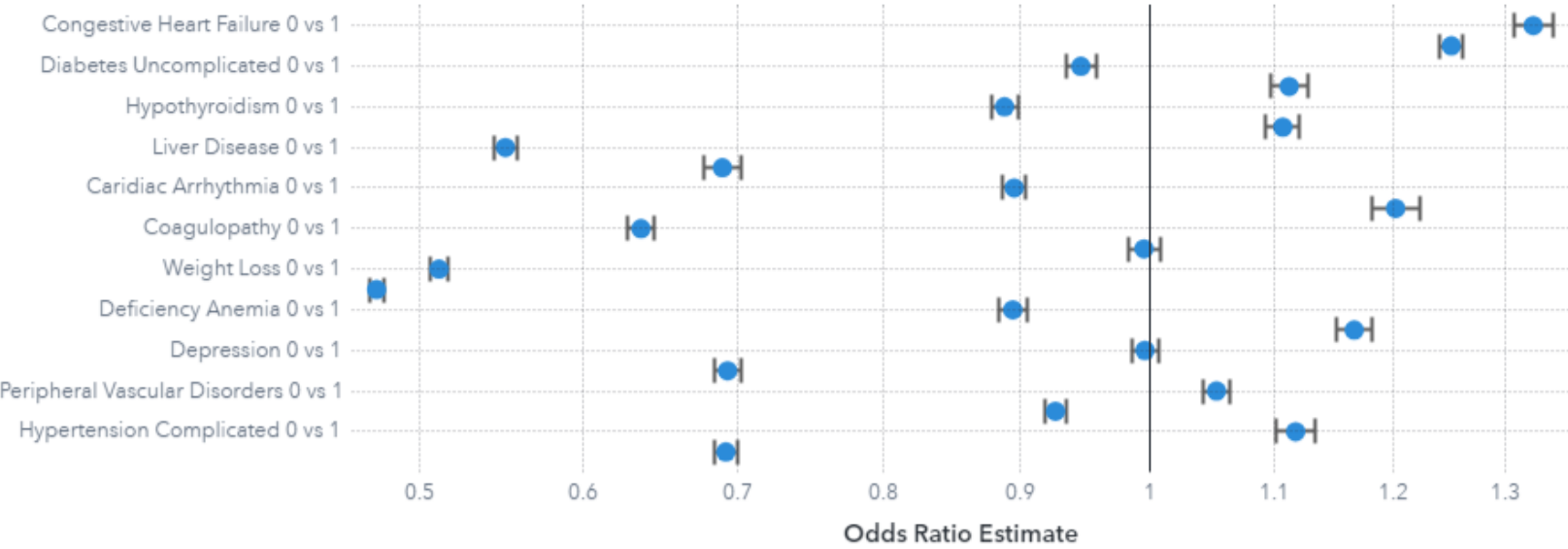
Fit Summary



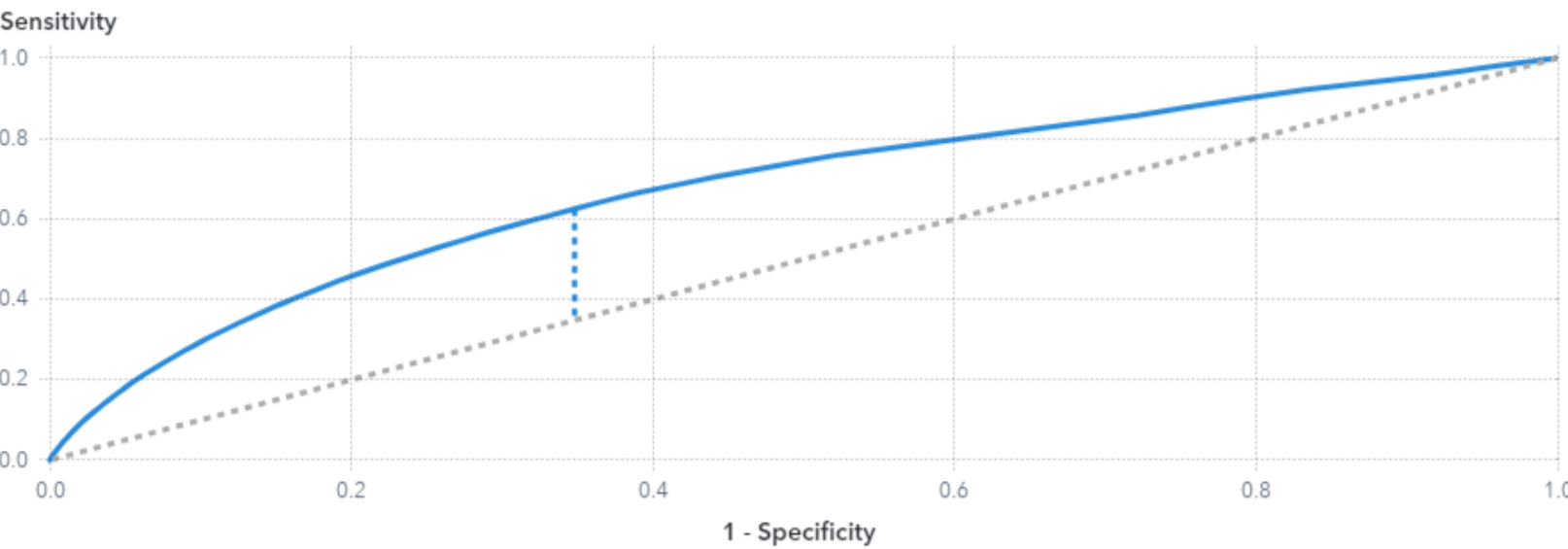
Residual Plot



Odds Ratio Plot



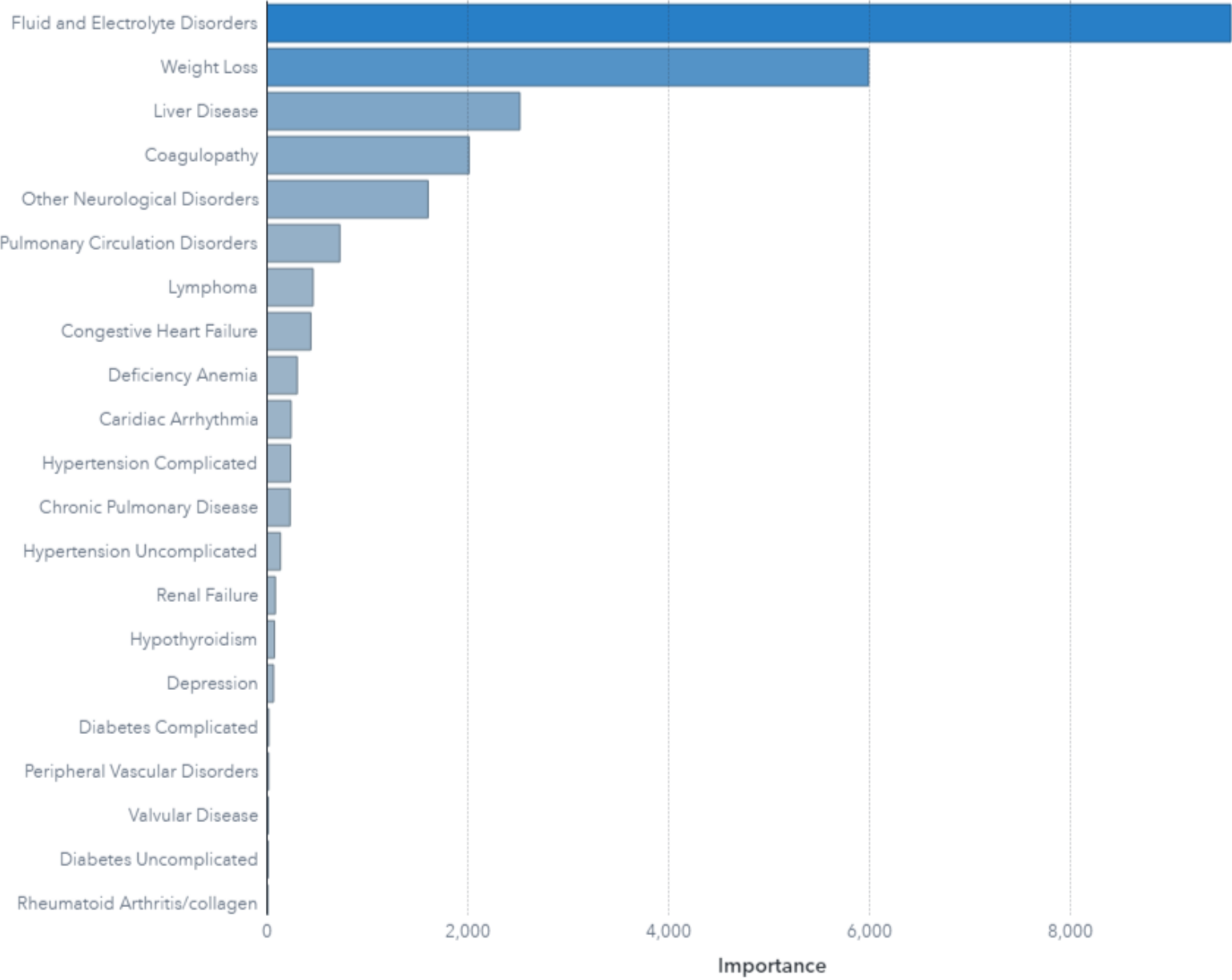
ROC



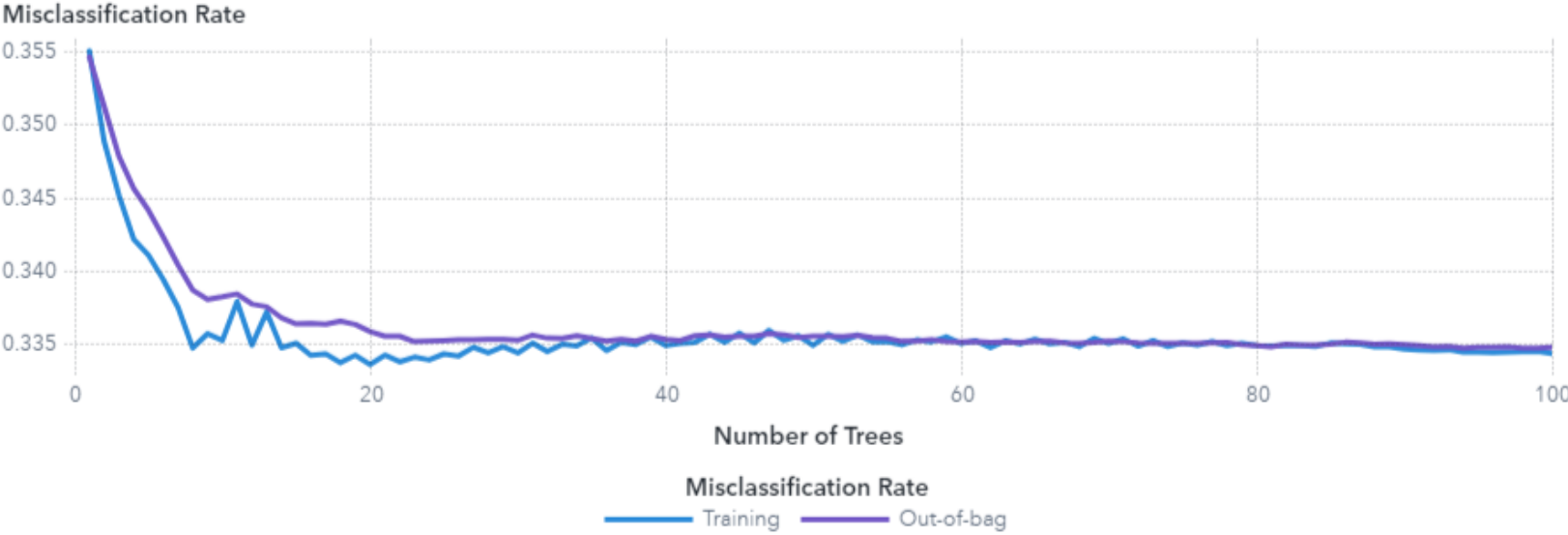
Random Forest Plot

Forest **Metastatic Cancer** Event: 1 ▾ Fit: KS (Youden) 0.2791 ▾ Observations: 1.2M of 1.2M

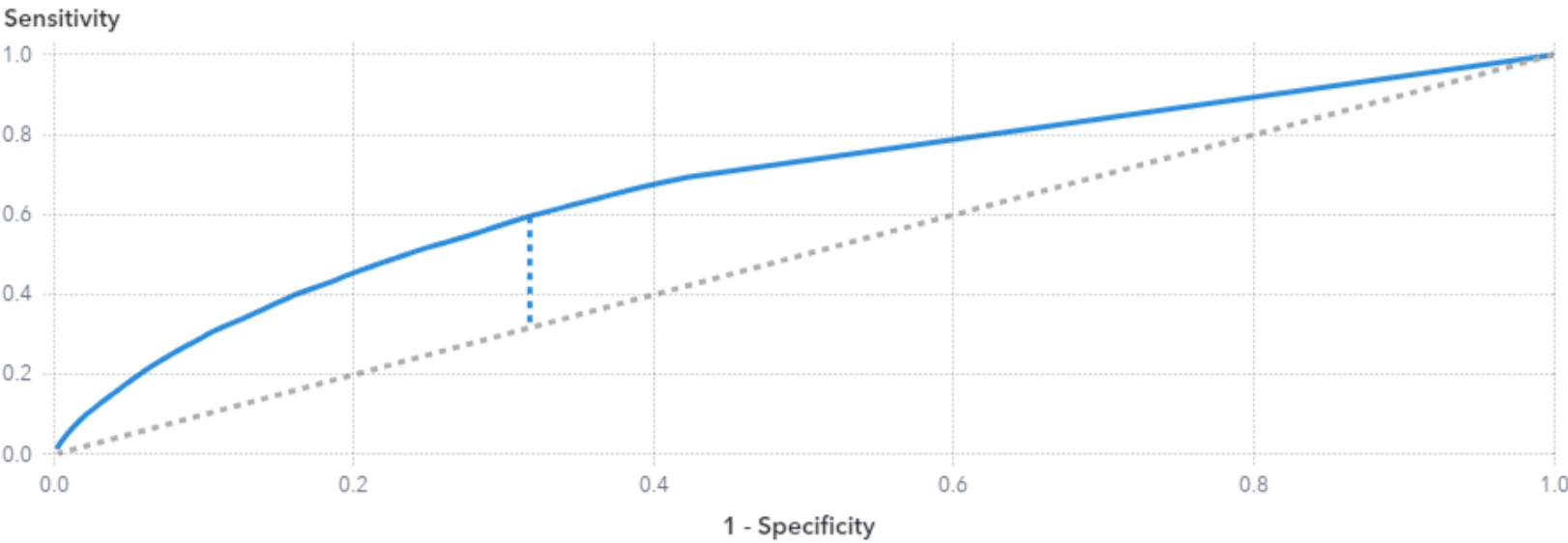
Variable Importance



Error Plot



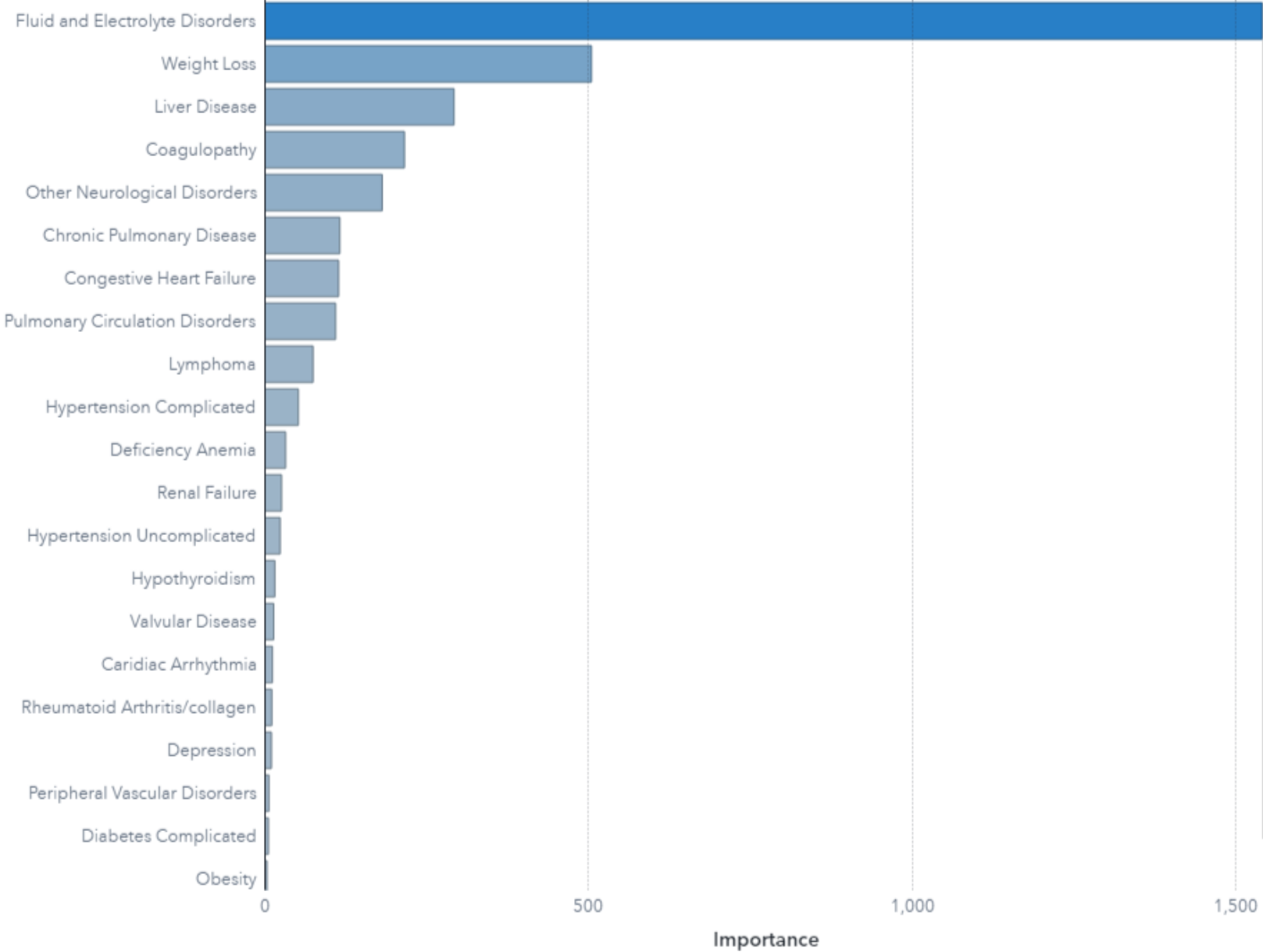
ROC



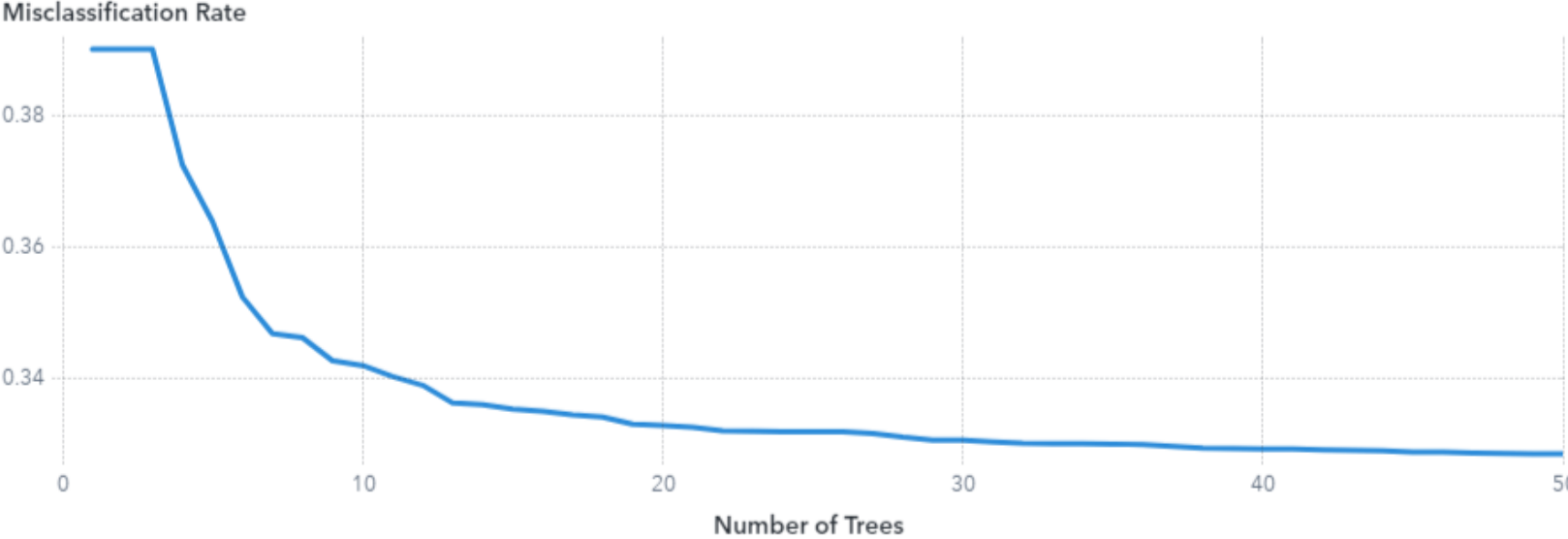
Gradient Boosting

Gradient Boosting **Metastatic Cancer** Event: 1 ▾ Fit: KS (Youden) 0.2854 ▾ Observations: 1.2M of 1.2M

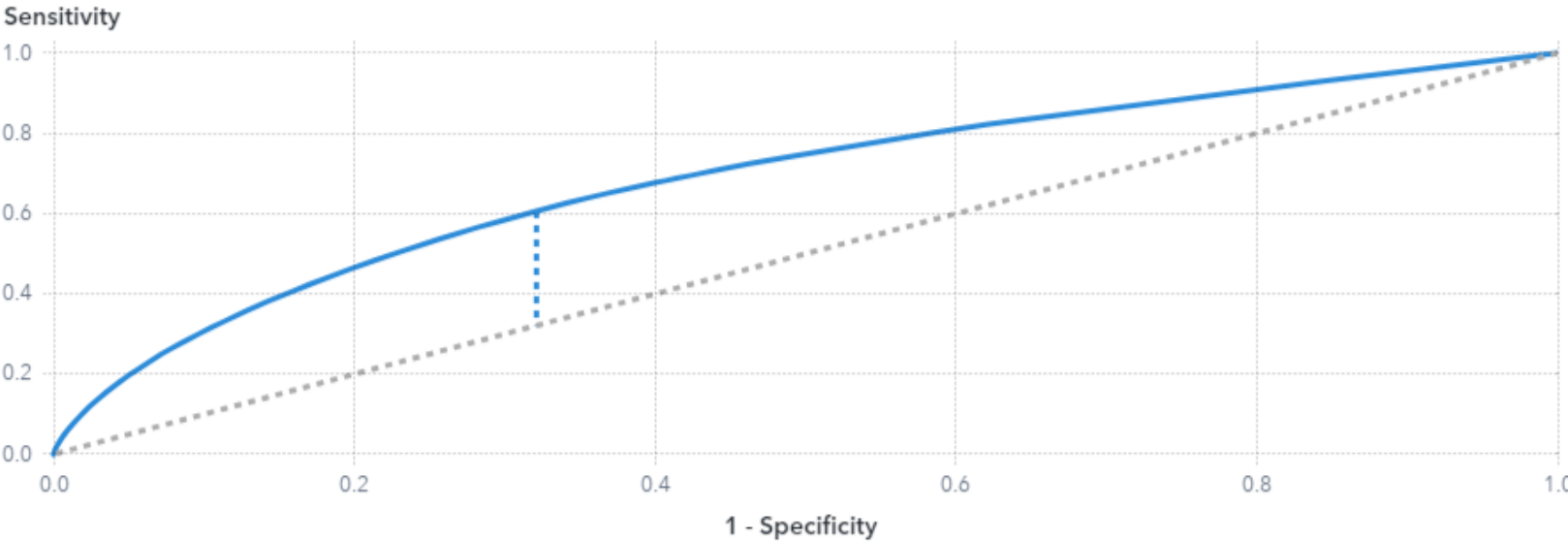
Variable Importance



Iteration Plot



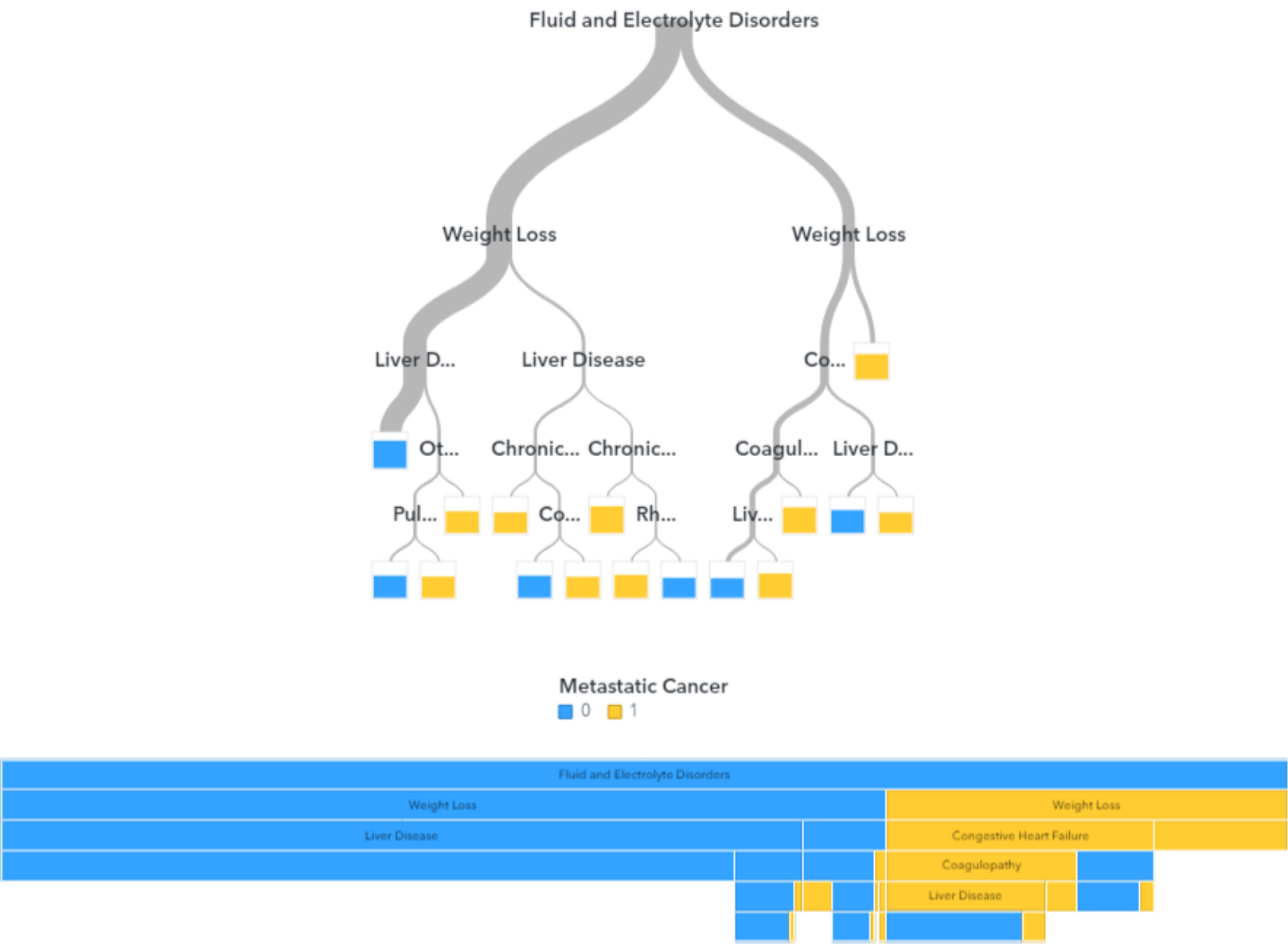
ROC



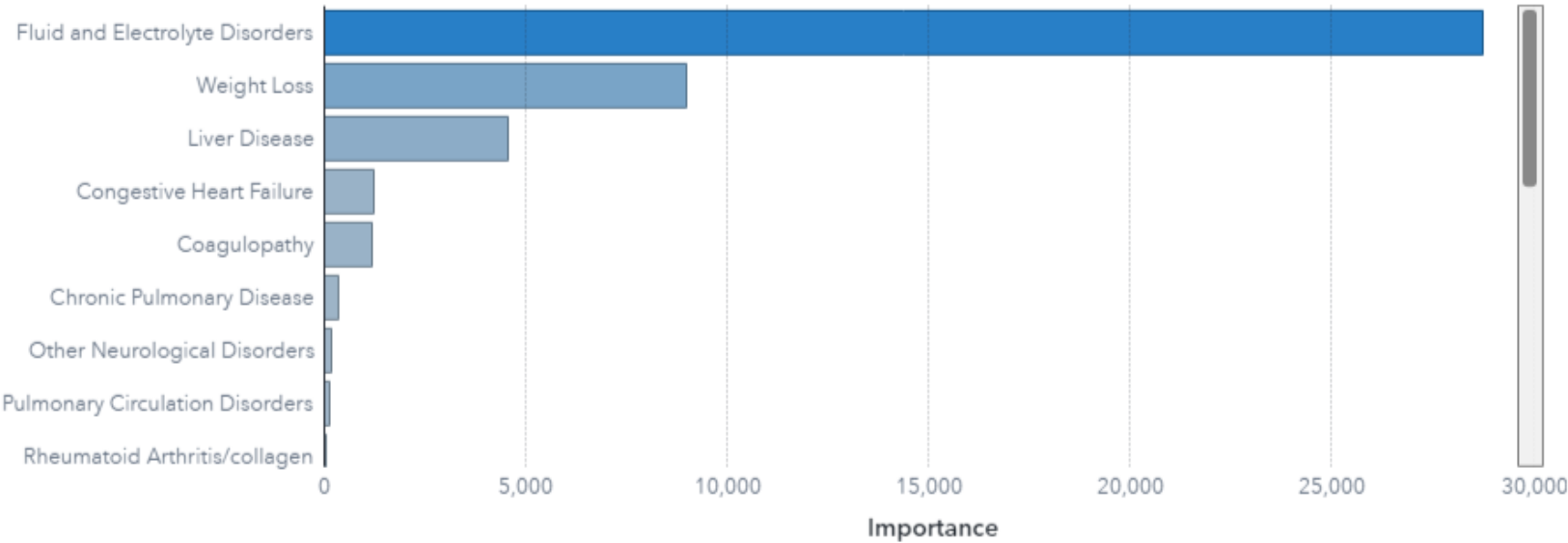
Decision Tree

Decision Tree Metastatic Cancer Event: 1 Fit: KS (Youden) 0.2595 Observations: 1.2M of 1.2M

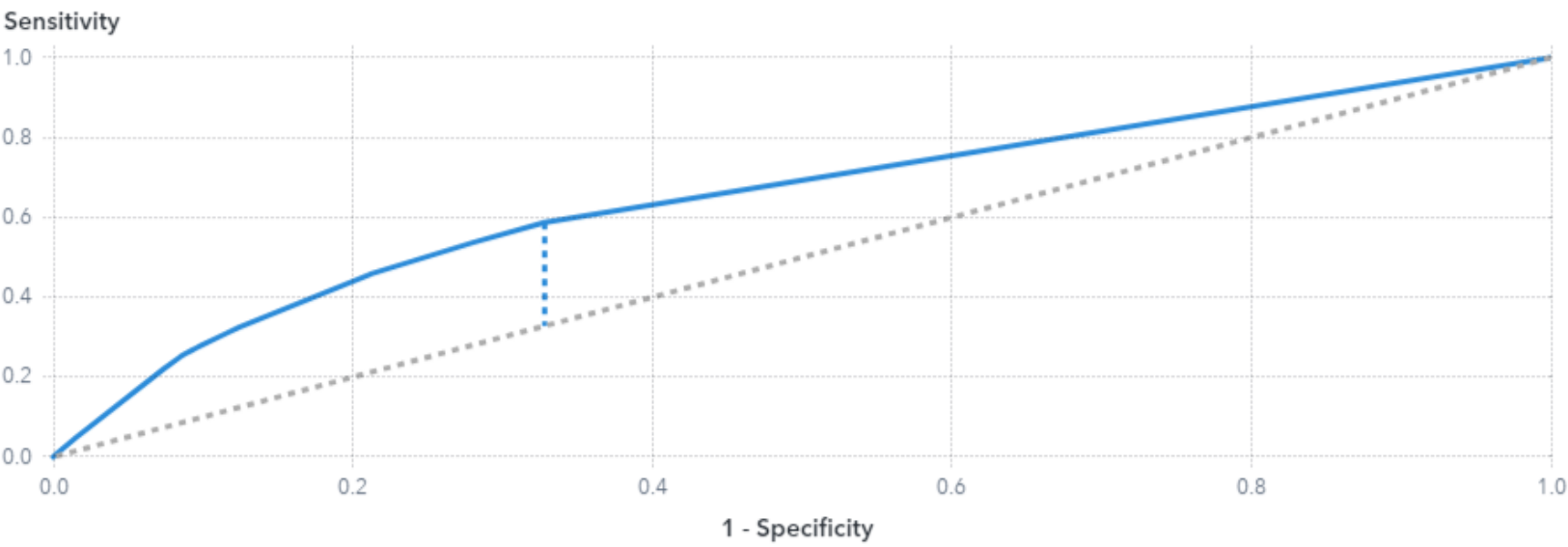
Tree



Variable Importance



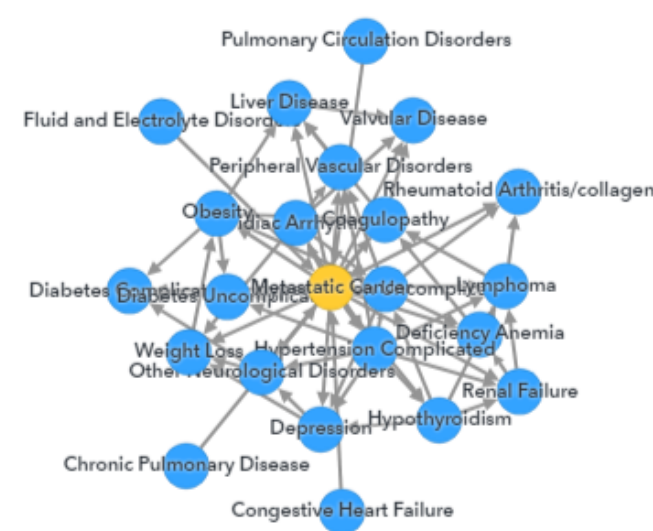
ROC



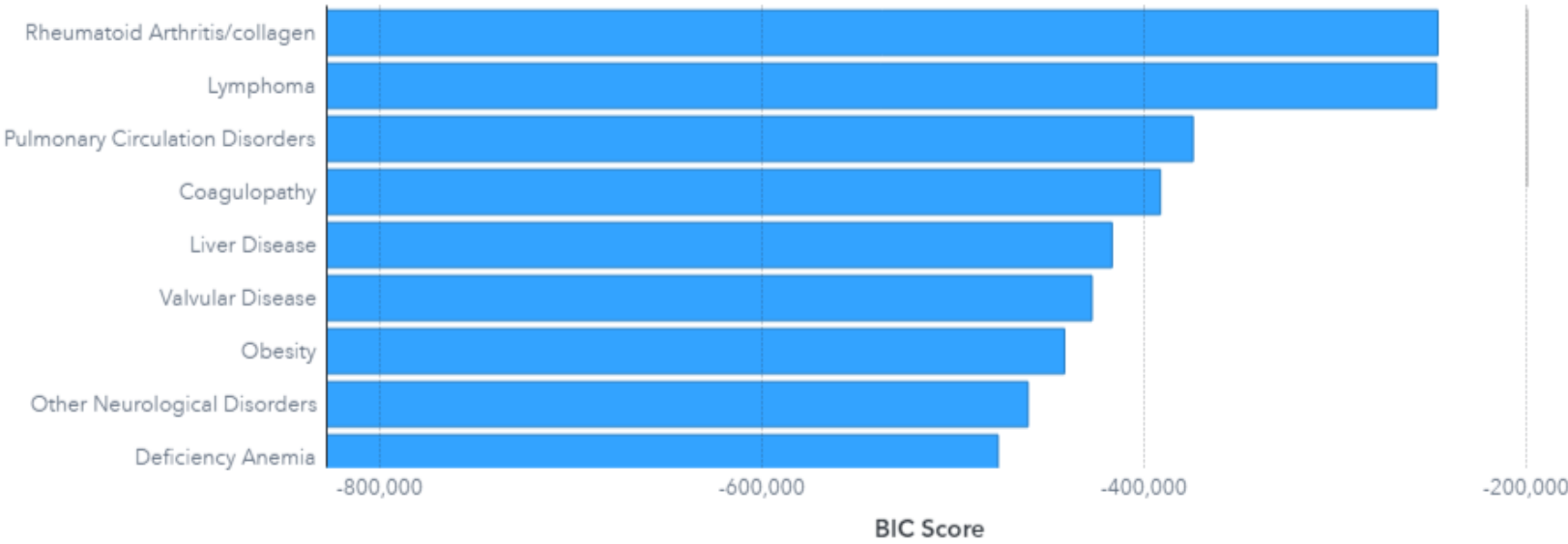
Bayesian Network

Bayesian Network Metastatic Cancer Event: 1 Fit: KS (Youden) 0.2744 Observations: 1.2M of 1.2M

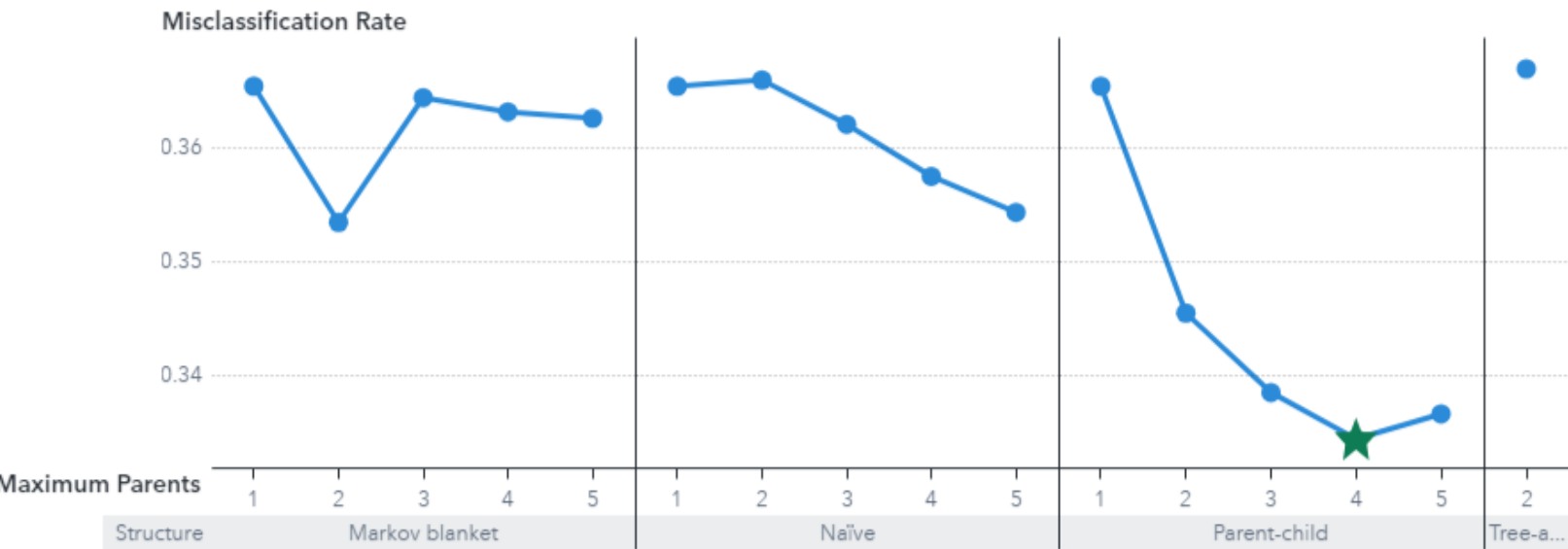
Network



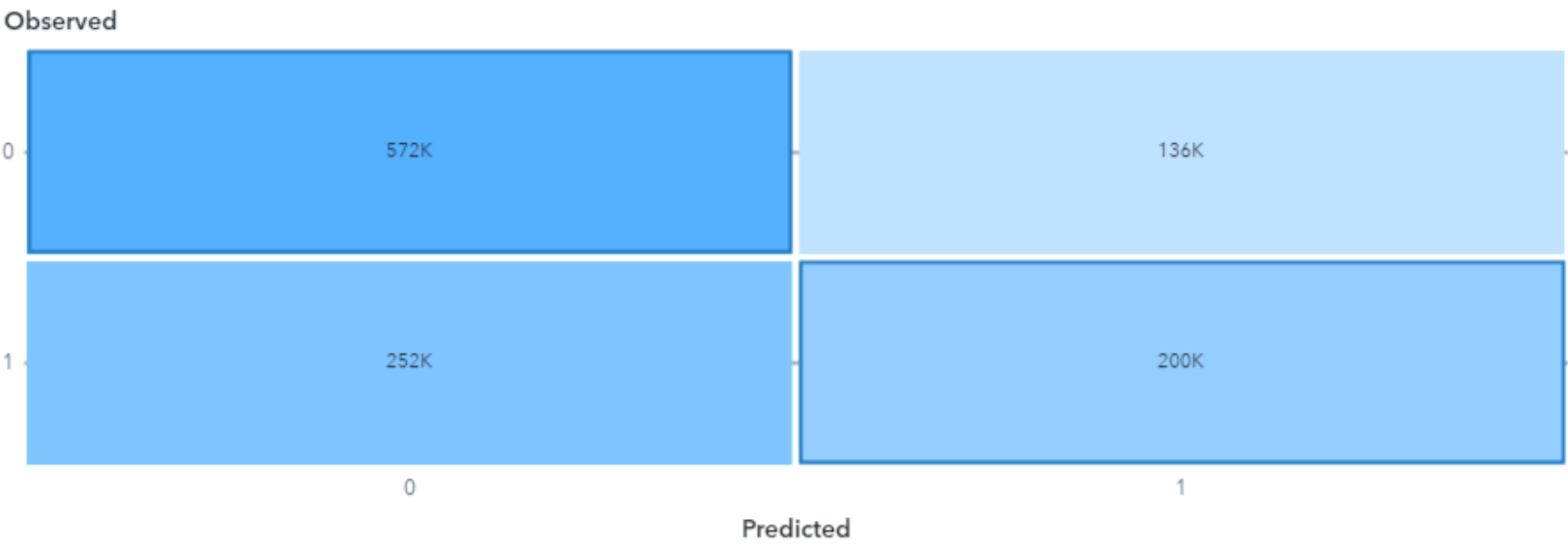
Variables in Network



Model Selection



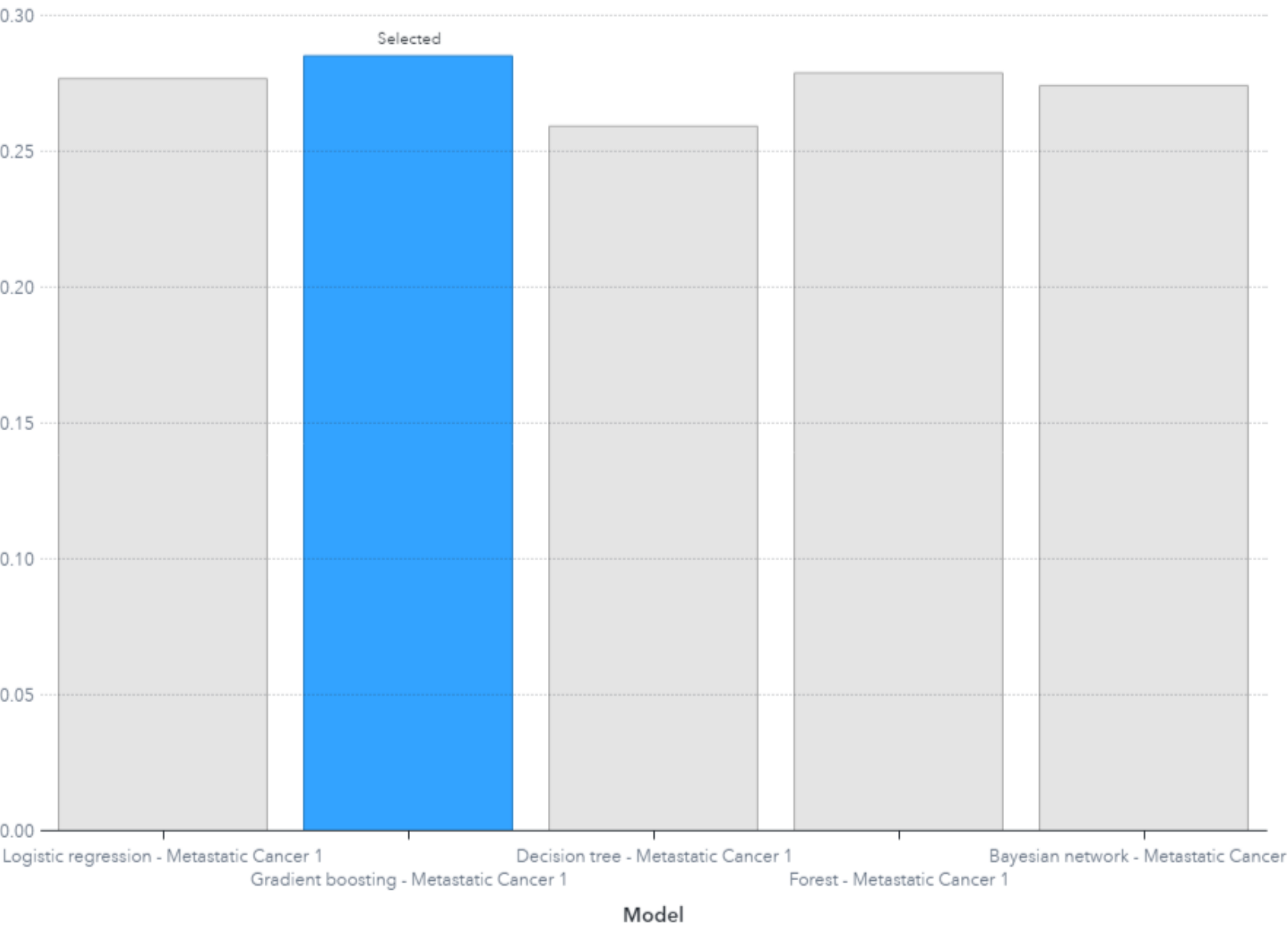
Confusion Matrix



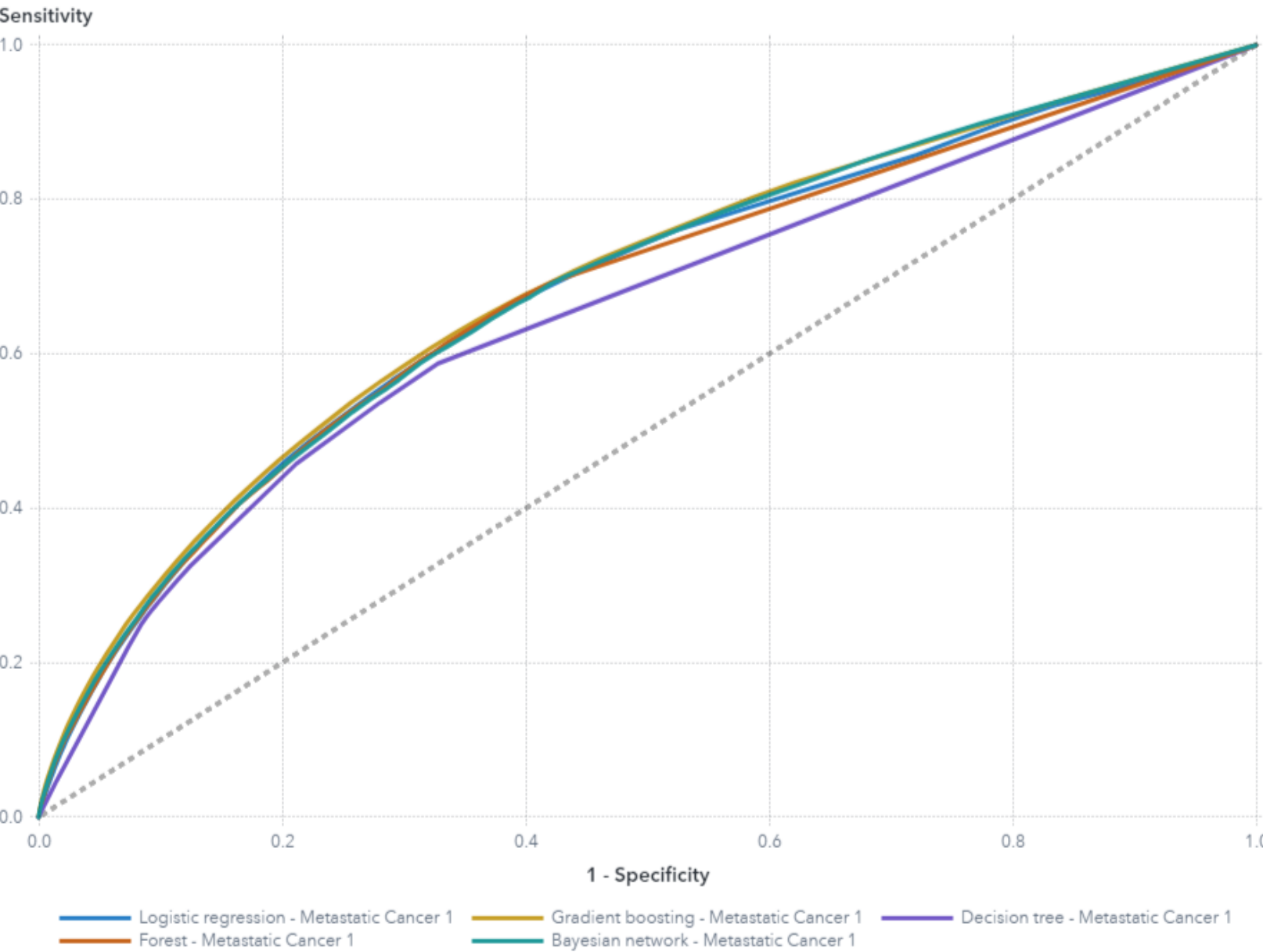
Model Comparison

Model Comparison Metastatic Cancer Event: 1

Fit Statistic 
KS (Youden)



ROC



Conclusion

- More comprehensive analyses are warranted to improve patient outcomes and care.
- COPD affects a patient's quality of life.
 - COPD is also a contributor to mortality in patients diagnosed with NSCLC.
- Personalized medicine could contribute to better prognoses for these patients.
 - Specifically, NSCLC patients should be assessed for comorbidities to select secondary treatments that minimize impact on cancer treatment and/or reduce overall morbidity and mortality.
- We owe our patients more comprehensive approaches which leverage improvements in compute power, scalable environments and ease of access to advanced analytics to assist in optimizing their outcomes.

Wrap Up

Thank You

- Samiul Haque (samiul.haque@sas.com)
- Sherrine Eid (sherrine.eid@sas.com)

