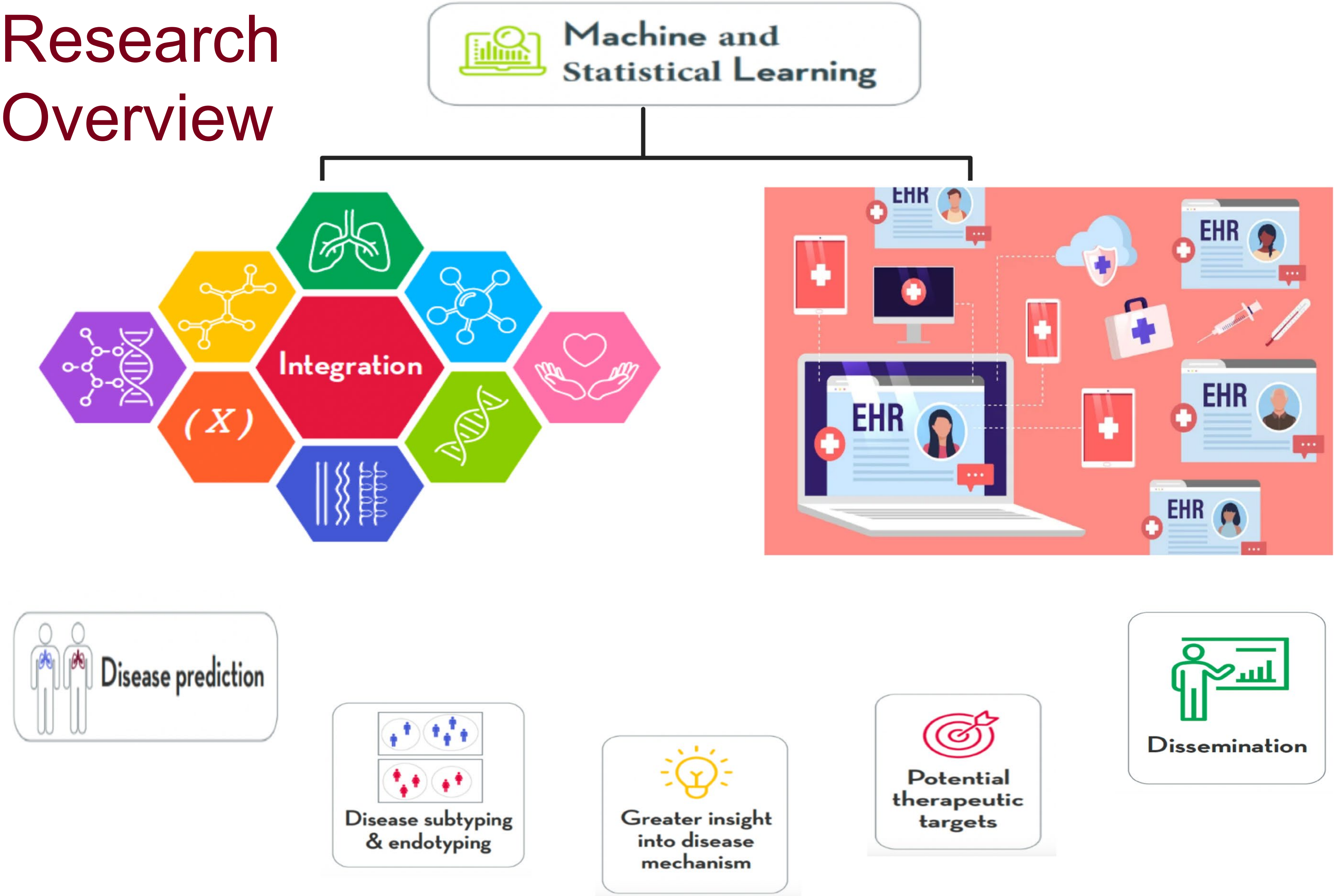
# Applications of Machine Learning in EHR for HIV Subphenotyping
# @
# COERE 2025 Annual Methods Symposium

Sandra Safo (ssafo@umn.edu; www.sandraesafo.com)
Division of Biostatistics and Health Data Science

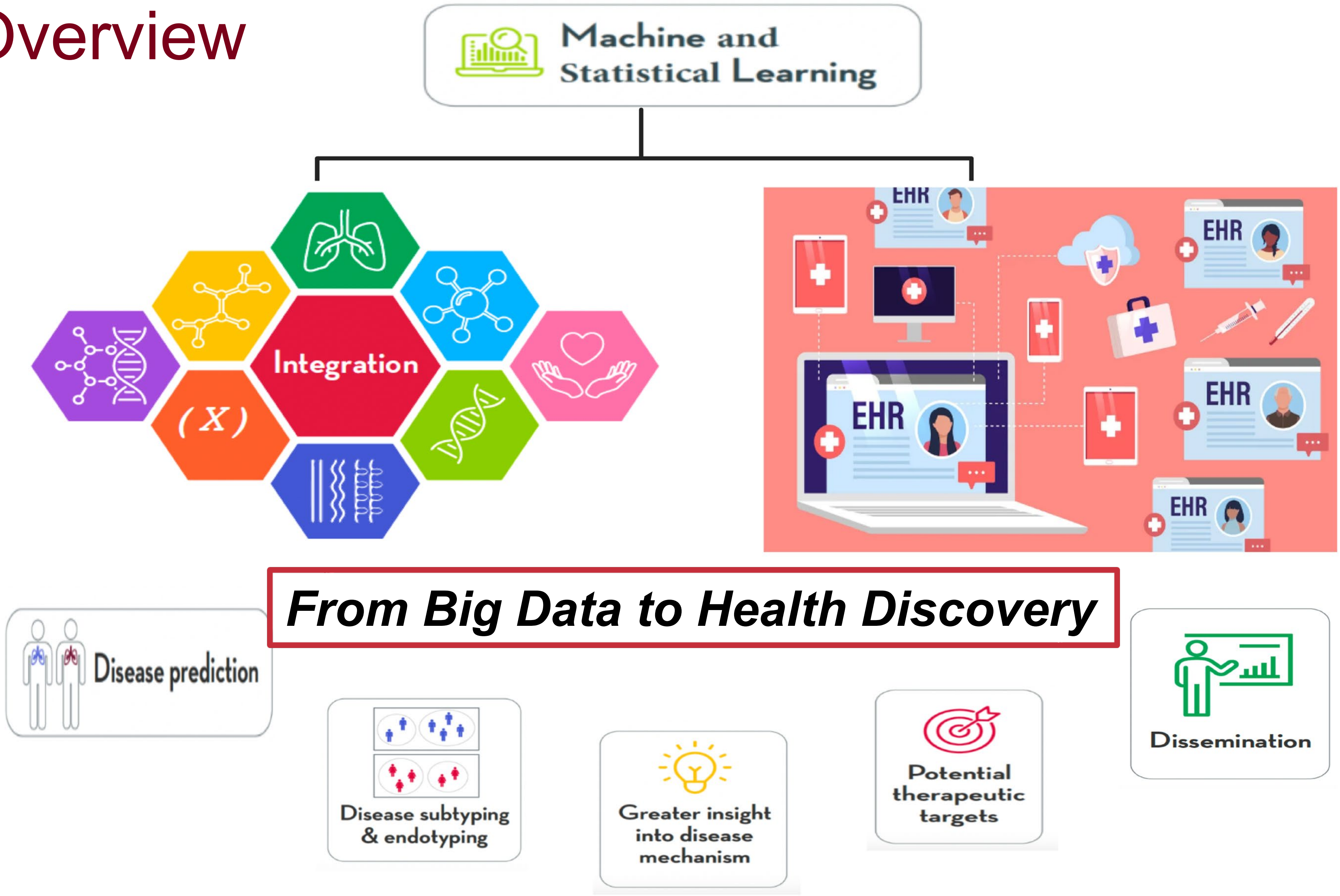
April 10, 2025
Financial Disclosure: None

University of Minnesota
Driven to Discover®

# Research Overview



# Research Overview



# Roadmap

Summary and
Key Takeaways
How can you get started?
Challenges and Considerations
ML in Action
A Gentle Intro to ML
And EHR

# Machine Learning

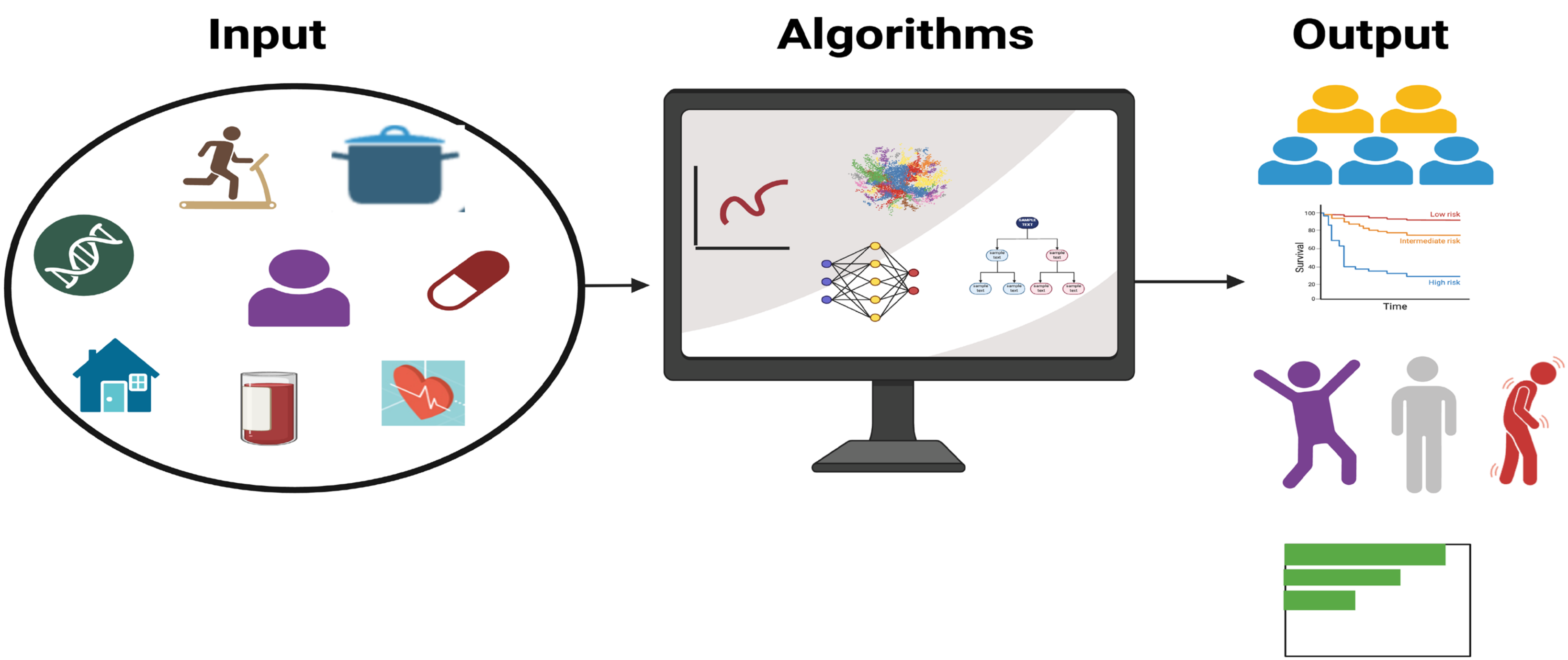



Figure generated with BioRender

# Supervised ML

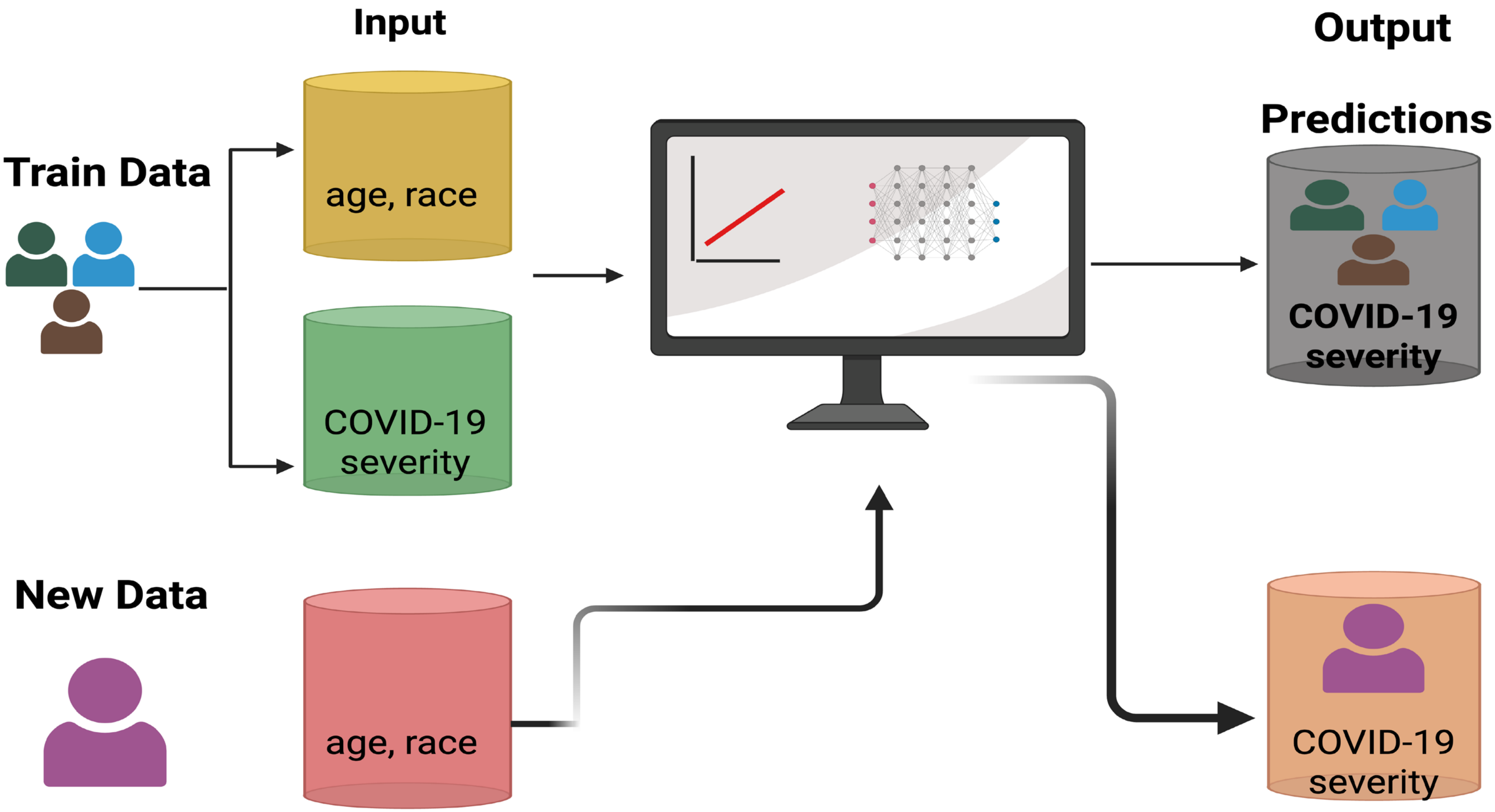



Figure generated with BioRender

# Unsupervised ML

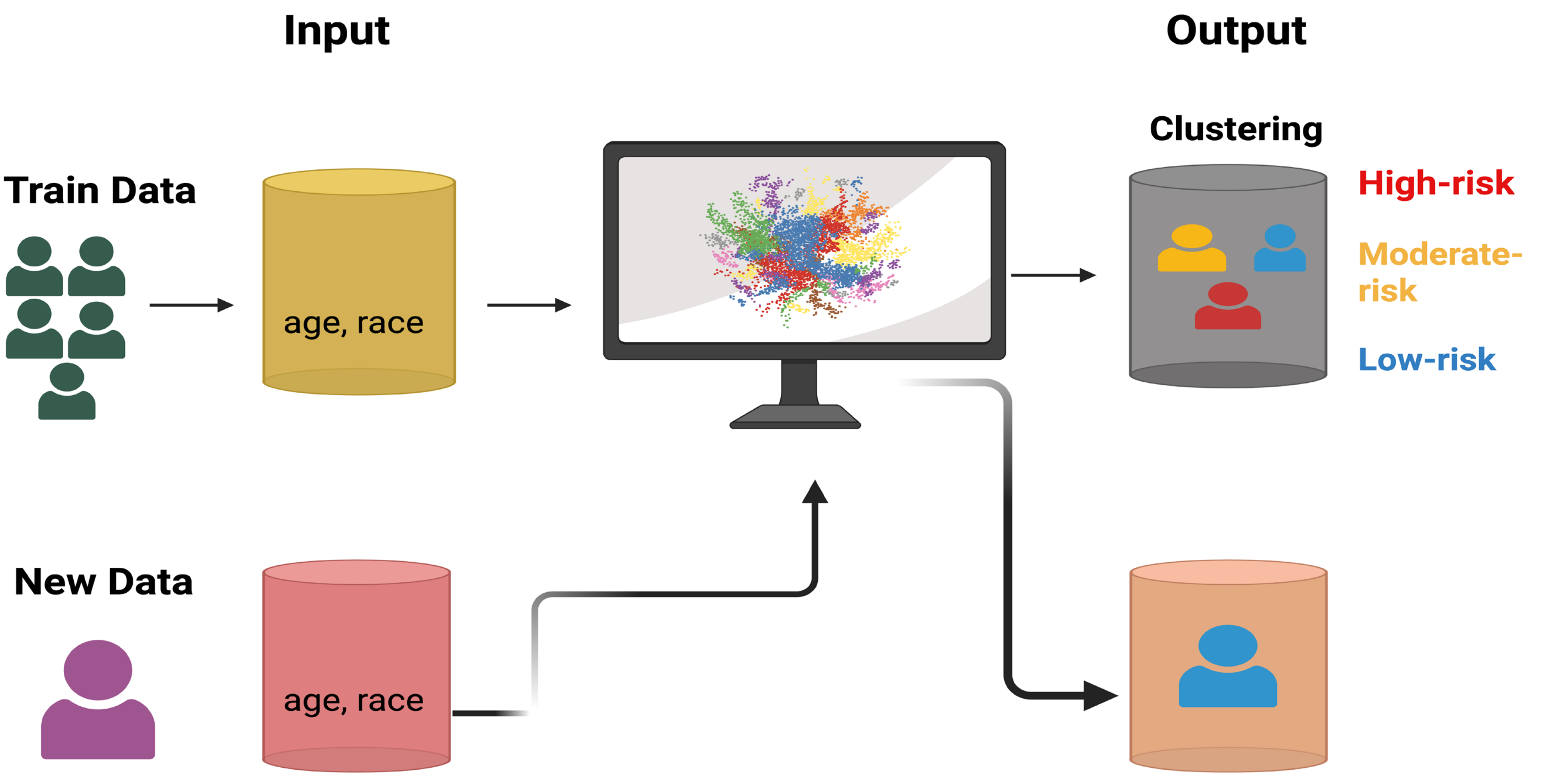



Figure generated with BioRender

# Why care about ML?

**Some Challenges in Health Care**

- Diagnosing diseases early
- Predicting patient outcomes (e.g., risk of severe COVID-19, increased hospitalizations, risk of Long COVID)
- Personalizing treatments based on patient data
- Patient selection for clinical trials

# Why care about ML?

**Opportunities with ML**

- Better diagnostic tools
- Improved decision support
- Data-driven research eliminating guestimates
- Help uncover hidden patterns in large and complex data (e.g., **EHR**)

# Challenges and Opportunities of EHR

**EHR- digitized version of a patient's chart and clinical records over time**

Structured and Unstructured Data

Demographics, Medications, vital signs, past medical history, Comorbidities, laboratory data, doctor's note etc

**Challenges**

Large, multidimensional, and unstructured

Missingness

Data quality (e.g., misclassification, measurement error, selection bias)

**Opportunities**

Can be cost effective to maintain

Data retrieval is quick

Offers opportunities to answer research questions and improve patient outcomes

# Machine Learning in Action: COVID-19 Subphenotyping among Persons Living with HIV (PLWH)

## Background and Rationale of Study

Persons living with HIV (PLWH) are disproportionately impacted by COVID-19

Older age, male sex, a history of smoking and comorbidities, is associated with risk of severe clinical outcomes or death

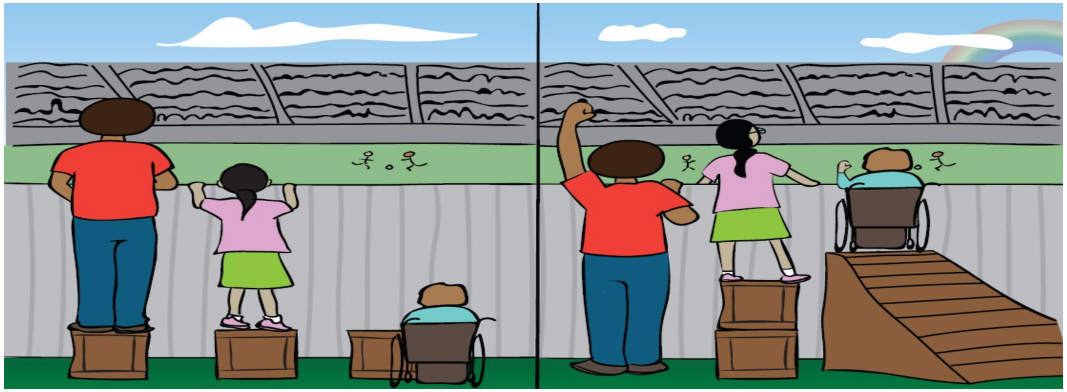
PLWH are more likely to have multiple comorbidities

Impact of factors, e.g. clinical comorbidities, predictive of severe COVID-19 among PLWH varies by race/ethnicity

Limited work exists that comprehensively investigate multiple data modalities including demographics, social determinants of health (SDoH), comorbidities, to better understand the characteristics of PLWH who tend to have worse outcomes.




- To **discover** and **validate** clusters with varying risks (COVID-19 severity, death, hospitalizations, Long COVID)
- To investigate the variables characterizing these clusters
- Our findings could
  - Facilitate targeted policy interventions
  - Enable a deeper understanding of health disparities in PLWH and promote health equity



# N3C Enclave: One of the largest public HIPAA- limited data set in US history 10/10/24

National COVID Cohort Collaborative

**Sites:** 84

**Persons:** 22.8 million

**COVID+ Cases:** 8,914,402

**# of Rows:** 33.9 billion

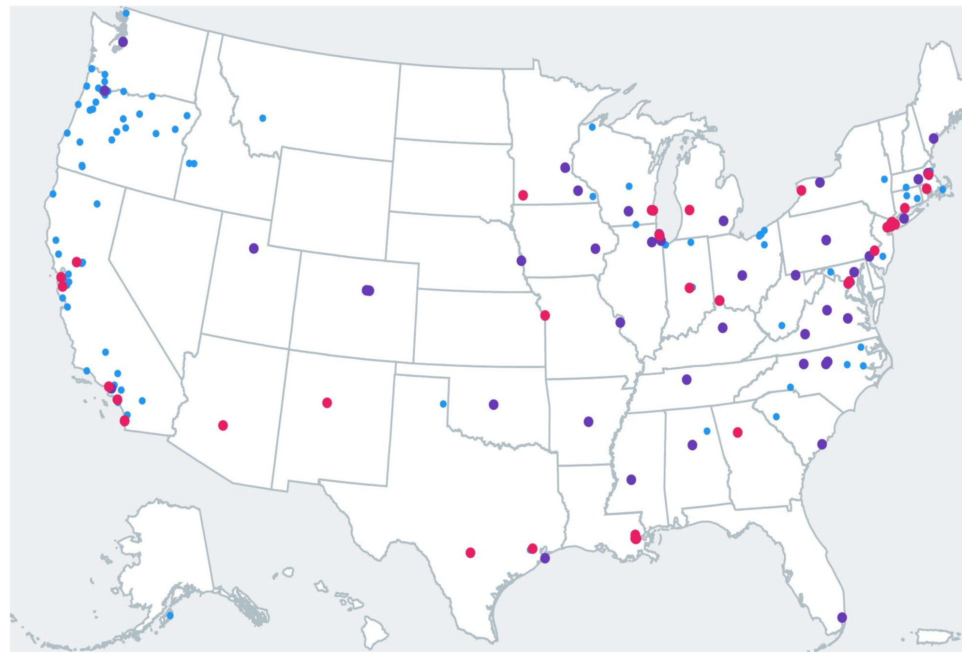
**Clinical Observations:** 3.3 billion

**Lab Results:** 16.3 billion

**Medication Records:** 5.3 billion

**Procedures:** 1.2 billion

**Visits:** 2.0 billion



# Data

Data from N3C (National COVID Cohort Collaborative)

Spans multiple states and hospitals

January 1, 2020 to November 2, 2023

Exclusions

Age < 18

Individuals taking medications for HIV prevention or treatment of chronic hepatitis B infection

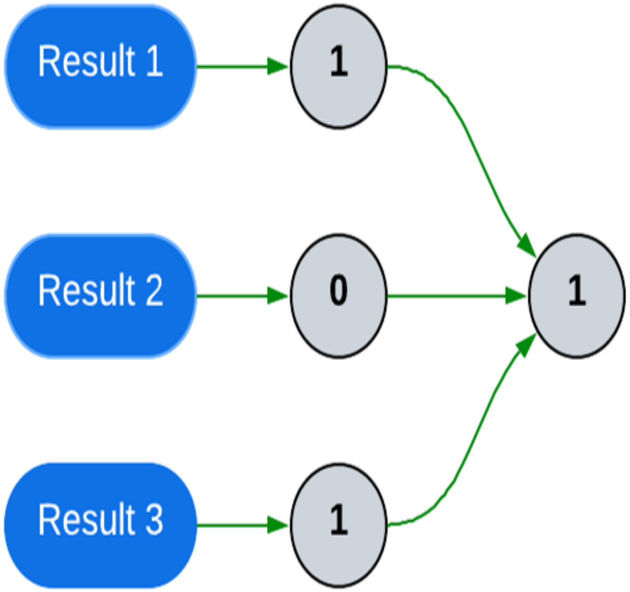
**Outcome**: COVID-19 Severity

**Not Severe**: Asymptomatic to symptomatic assistance needing hospitalization

**Severe**: Hospitalized with invasive mechanical ventilation, extracorporeal membrane oxygenation [ECMO], discharge to hospice or death

Input Variables: demographics, comorbidities, lab measurements, social determinants of health (SDoH)

# Methods




## Data splitting (Focus on PLWH)

- Discovery set: to detect the clusters
- Testing set: to reproduce the clusters

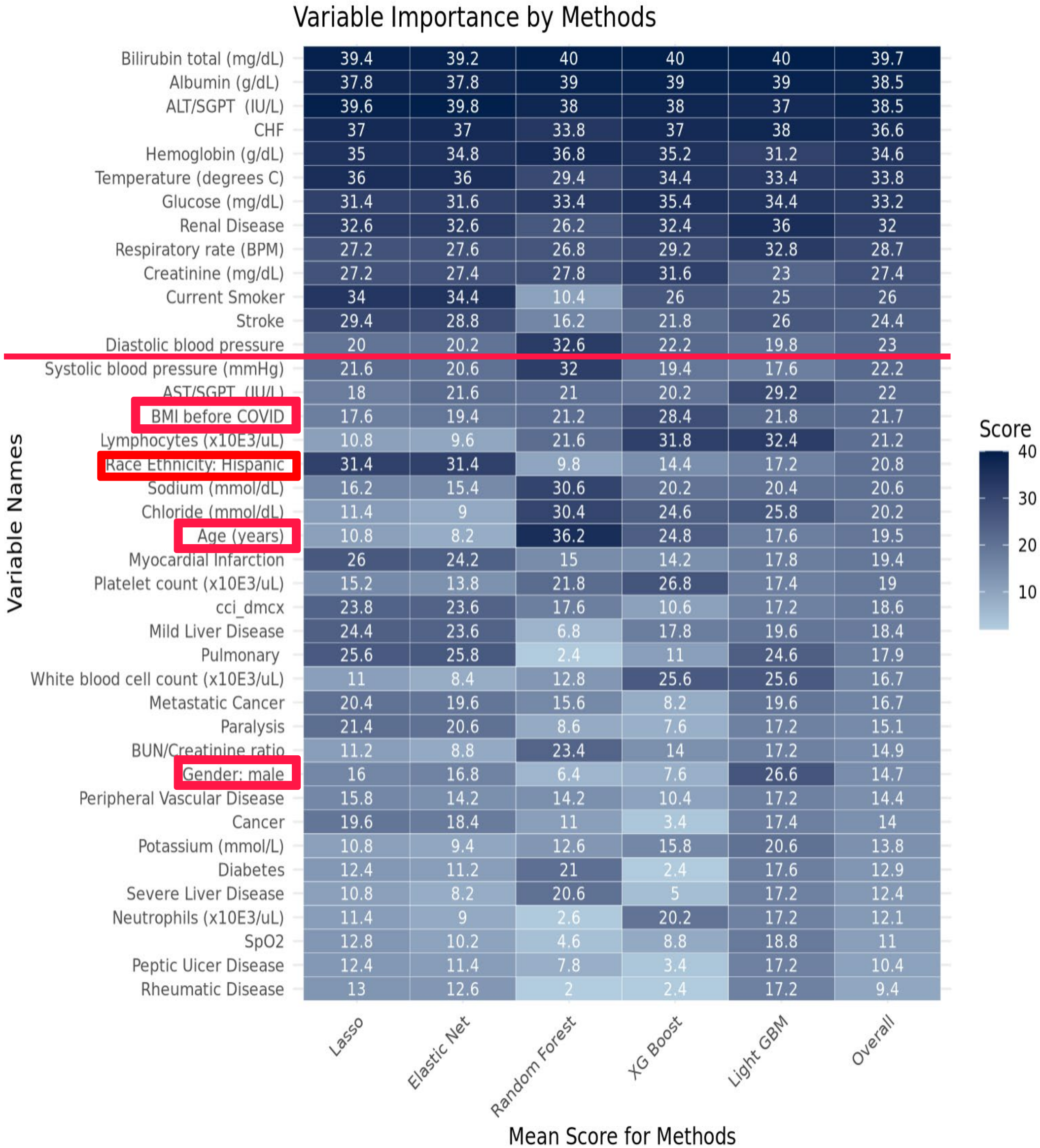
## Variable selection

- Identify which input variables discriminate COVID-19 severity group
- Use these variables to cluster individuals
- Incorporate resampling techniques for statistical rigor and address missingness

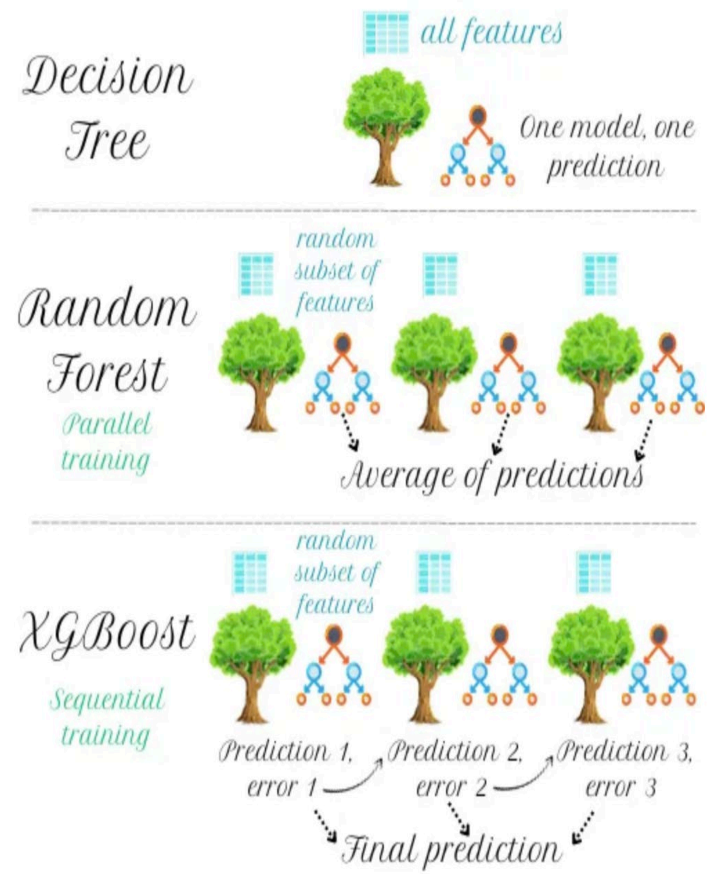
## Cluster Detection

## Reproduce Clusters

# Variable Ranking via Ensemble of Methods




LASSO (Least absolute **shrinkage** and **selection** operator), Elastic Net, Random Forest, XG Boost, Light GBM




Source: Medium

# Baseline Characteristics

| | All N3C (n=5622302) | Persons without HIV (n=5547210) | Persons with HIV (n=75092) |
|---|---|---|---|
| **COVID-19** | | | |
| Not Severe | **5527166 (98%)** | **5453251 (98%)** | **73,915 (98%)** |
| Severe | **95136 (1.7%)** | **93959 (1.7%)** | **1.177 (1.6%)** |
| Smoking Status | | | |
| Non smoker | **5230971 (93%)** | **5166174 (93%)** | **64697 (86%)** |
| Current or former smoker | **391431 (7.0%)** | **381036 (6.9%)** | **10395 (14%)** |
| Sex | | | |
| Females | **3231360 (57%)** | **3195925 (58%)** | **35435 (47%)** |
| Males | **2390942 (43%)** | **2351285 (42%)** | **39657 (53%)** |
| Race/Ethnicity | | | |
| Hispanic or Latino Any Race | **688110 (12%)** | **679102 (12%)** | **9008 (12%)** |
| Black or African American Non-Hispanic | **661217 (12%)** | **646374 (12%)** | **14843 (20%)** |
| White Non-Hispanic | **3403103 (61%)** | **3359522 (61%)** | **43581 (58%)** |
| Other Non-Hispanic | **869872 (15%)** | **862212 (16%)** | **7660 (10%)** |
| Comorbidities > 4 | **152187 (2.7%)** | **147787 (2.7%)** | **4400 (5.9%)** |

# Executive Summary

**Cluster 1:**
Low-risk
Lowest age
Hispanic or Latino
Other Non-Hispanic

No or lowest number of comorbidities

Relative healthier indicated by lab-related variables

**Cluster 2:**
Moderate-risk
Female

Older age

White Non-Hispanic
Lowest BMI

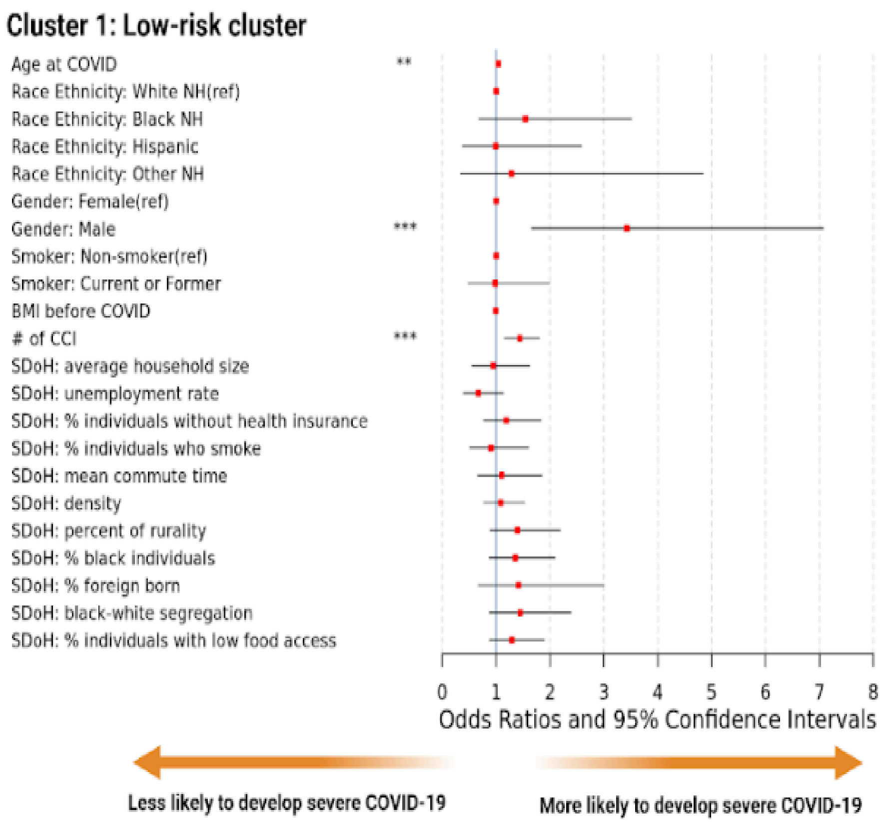
Highest CD4 count percent

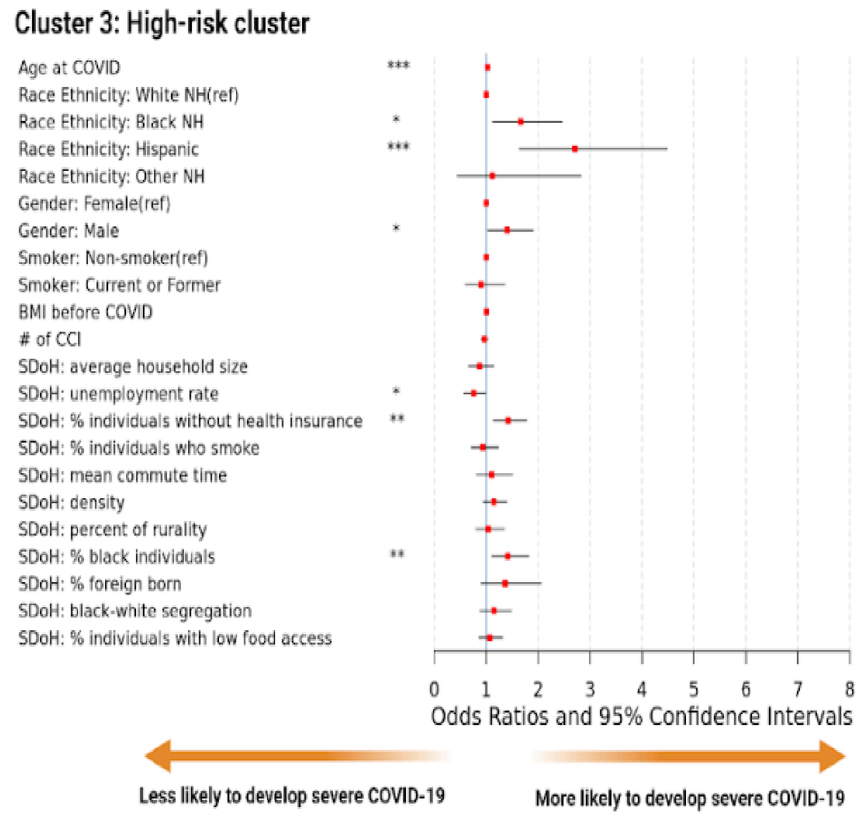
**Cluster 3:**
**High-risk**
Male
Current or former smoker
Black Non-Hispanic
Highest BMI
Highest HIV viral load

Most number of comorbidities
Higher proportions of death, long COVID & hospitalization

| Cluster Comparison | Odds of COVID-19 Severity | Odds of Death | Odds of Long COVID-19 |
|---|---|---|---|
| Cluster 3 vs Cluster 1 | 20 times more likely | 11 times more likely | 1.3 times more likely |
| Cluster 2 vs Cluster 1 | 3 times more likely | 2 times more likely | Similar odds |
| Cluster 3 vs Cluster 2 | 6 times more likely | 6 times more likely | 1.23 times more likely |

# Cluster 1 vs Cluster 3- Severe COVID-19



# Comparison of SDoH across Clusters

## Cluster 1

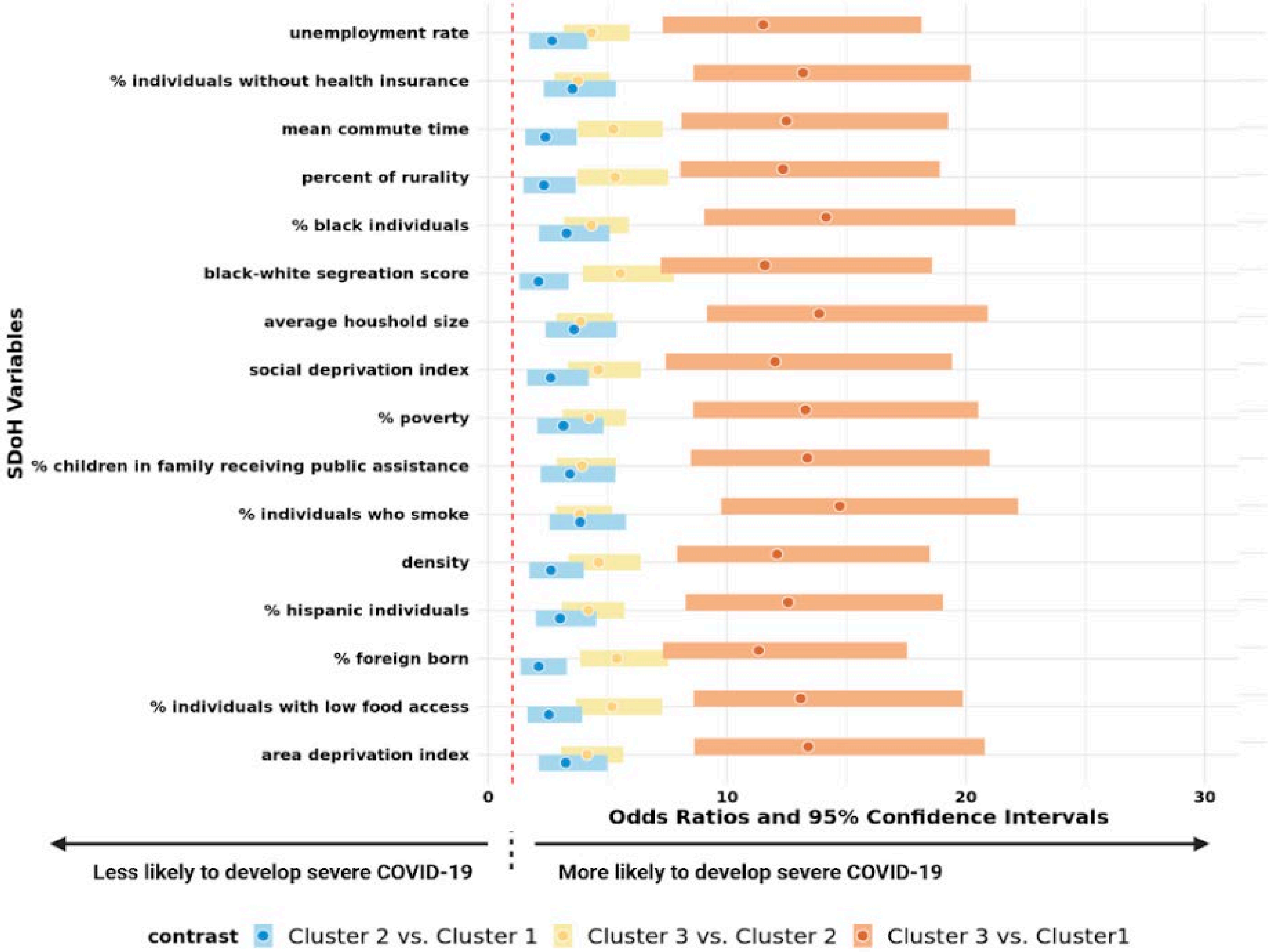
- Few significant associations between SDoH and health outcomes
- Larger household sizes may serve as a protective factor against hospitalization.

## Cluster 2

- Population density and racial disparities (e.g., percentage of Black individuals, Black-White segregation) increased odds of deaths

## Cluster 3

- Lack of health insurance resulted in worse outcomes
- Factors like rurality, foreign-born status, and Black-White segregation appeared to reduce hospitalization odds



SDoH Variables
unemployment rate
% individuals without health insurance
mean commute time
percent of rurality
% black individuals
black-white segreation score
average houshold size
social deprivation index
% poverty
% children in family receiving public assistance
% individuals who smoke
density
% hispanic individuals
% foreign born
% individuals with low food access
area deprivation index
0
10
20
30
Odds Ratios and 95% Confidence Intervals
Less likely to develop severe COVID-19
More likely to develop severe COVID-19
contrast
Cluster 2 vs. Cluster 1
Cluster 3 vs. Cluster 2
Cluster 3 vs. Cluster1

# Implications of findings

Offer deeper insights into characteristics of PLWH with worse COVID-19 outcomes

Could enable targeted policy interventions and improve health equity

Provides a deeper perspective of the intersections of social determinants of health and COVID-19 clusters in PLWH

Approach could be adopted to other clinical outcomes in HIV

# Limitations of Work

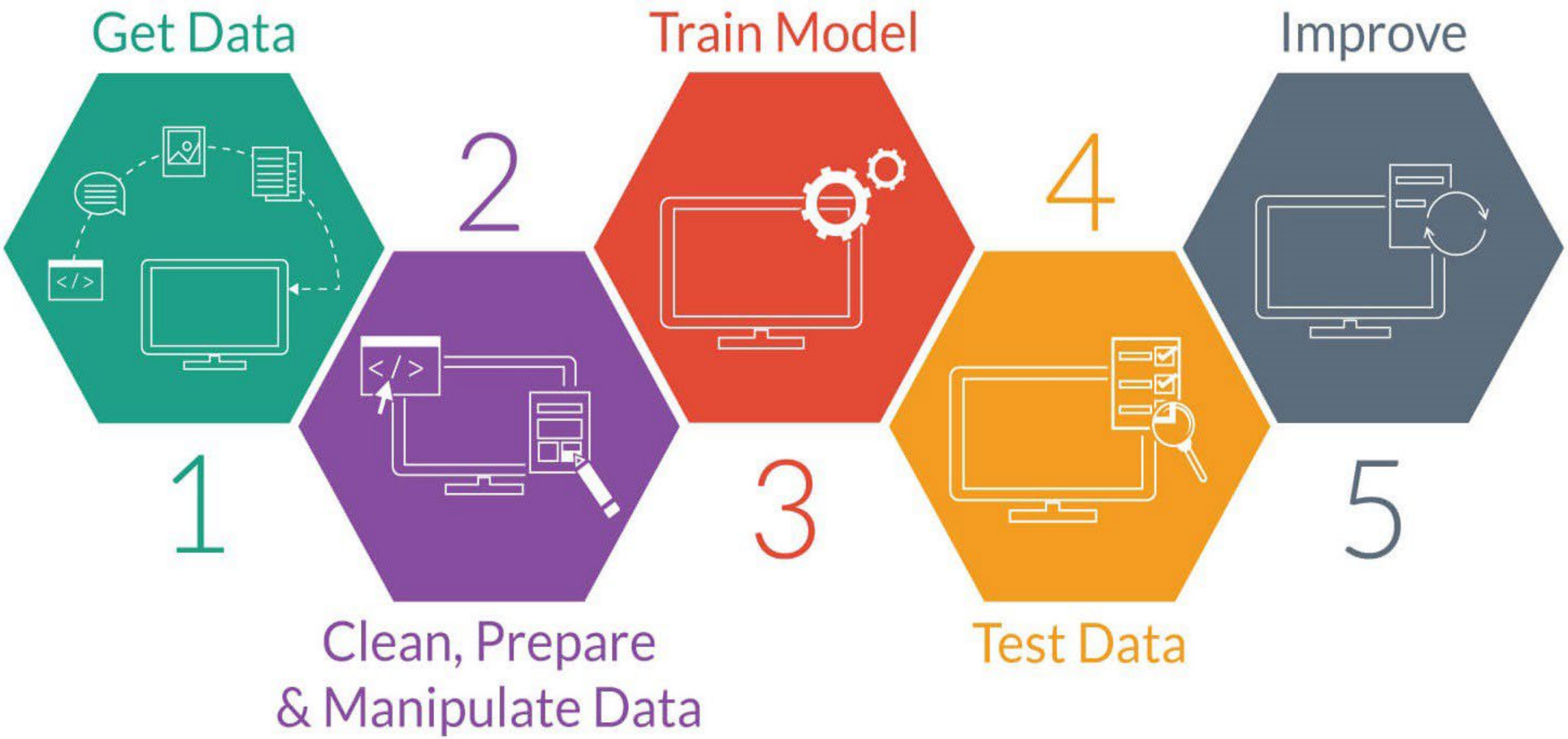
Missing data in lab measurements

Resampling approach might affect our findings

Limited sample size for HIV-specific variables

No adjustment for vaccination

# How can you get started with ML?



# Some Challenges and Considerations of ML

## Ethical concerns

Data privacy (HIPAA-compliance)

Algorithmic fairness

## Limitations of ML

Clinical judgment vs machine learning predictions

The need for transparent, interpretable models (e.g., why a prediction was made)

# Conclusions and Takeaways


## Summary

ML is an exciting tool that can improve healthcare, particularly in EHR

Can improve precision medicine, detect disease early, reduce human error, identify patients for clinical trials and many more

## Call to action

Consider exploring simple ML models in your work

Get involved with others using ML and learn to do the dirty work

Learn to code in R/Python (Coursera etc.)

## Thank You

Tiankai Xie

N3C HIV-Subdomain team

NIH Funding

UNIVERSITY OF MINNESOTA

**Driven to Discover®**

Crookston Duluth Morris Rochester Twin Cities