Artificial Intelligence for Population Health

Kristin P. Bennett

Associate Director Institute for Data Exploration and Applications

Professor Mathematics, Computer Science, and Industrial Engineering

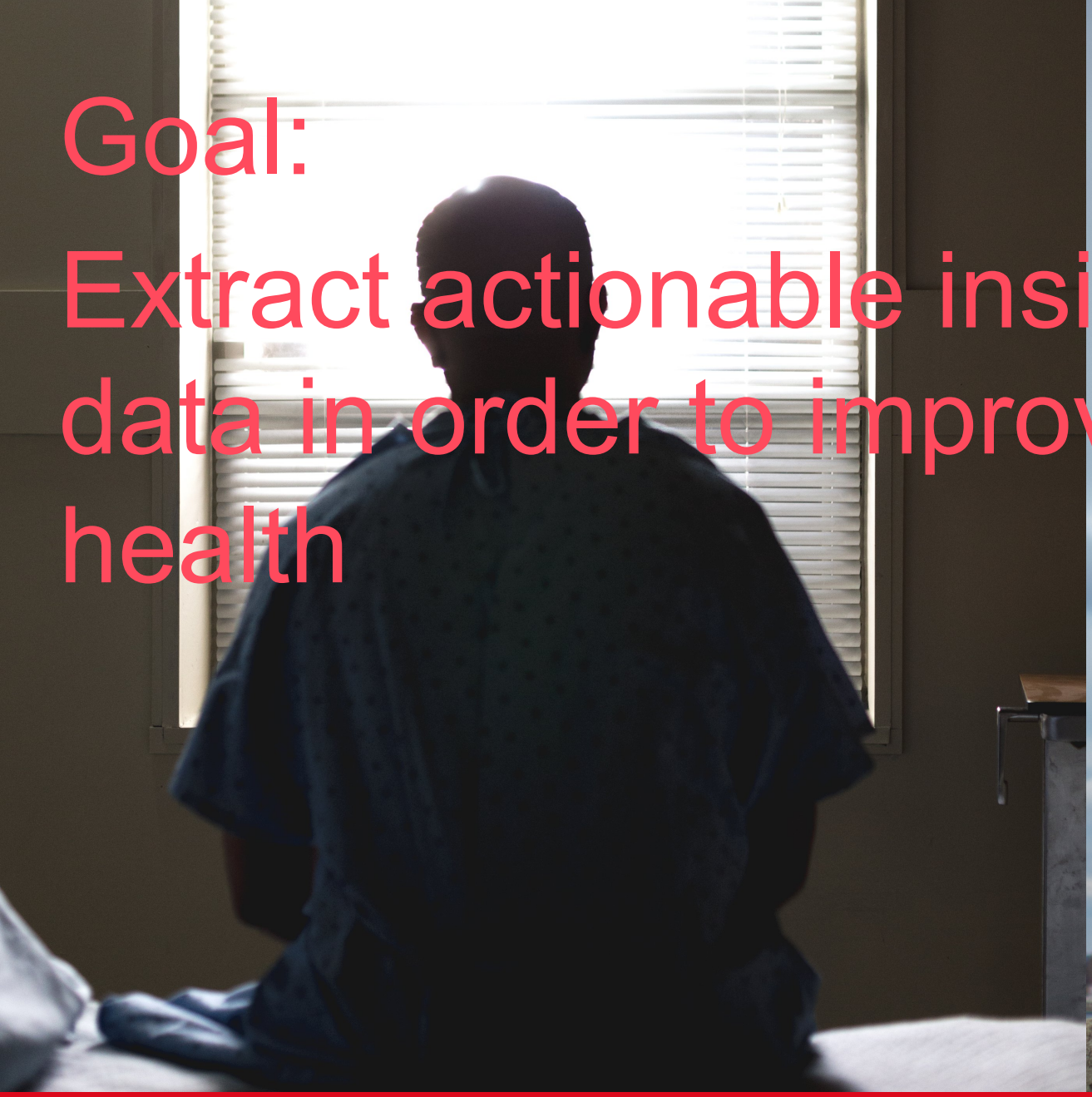
IEEE BIBM Keynote, November 2019

Rensselaer

why not change the world?®

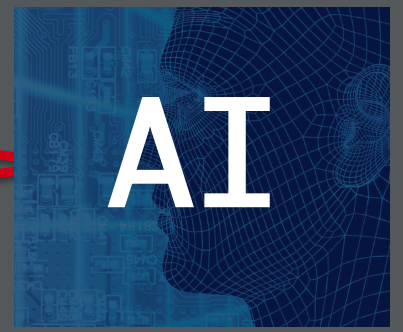
Goal:

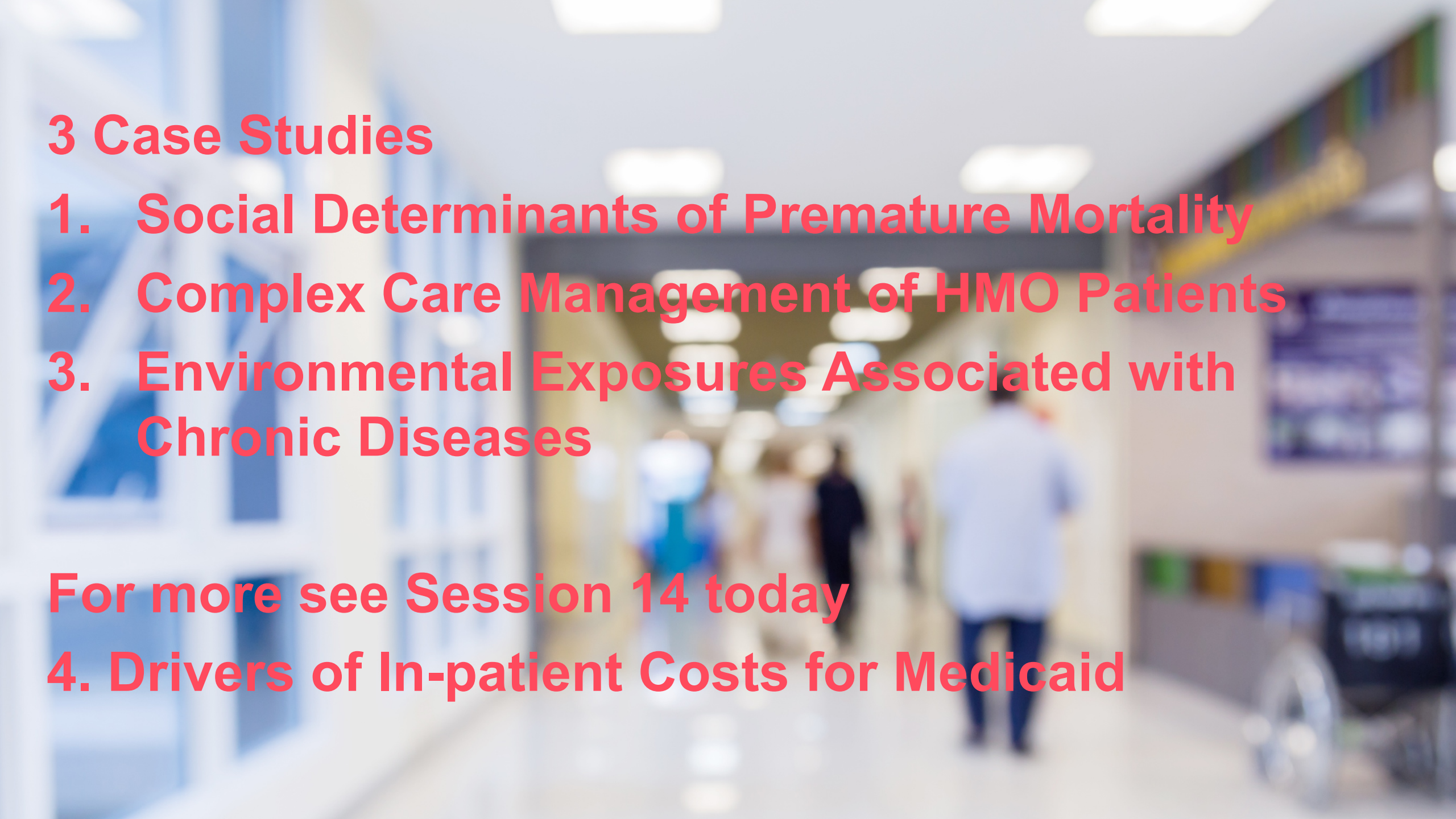
Extract actionable insights from health care data in order to improve overall population health



Population Health

- Surveillance - detect problems
- Obtain actionable insights
  - Who is at risk?
  - What are risks factors/determinants?
- Create policies, programs, and interventions
- Address population health at national, state, community, health care organization and/or patient levels





3 Case Studies

1. Social Determinants of Premature Mortality
2. Complex Care Management of HMO Patients
3. Environmental Exposures Associated with Chronic Diseases

For more see Session 14 today

4. Drivers of In-patient Costs for Medicaid

Study 1: Community–Level Social Determinants of Mortality

Mortality rates for adults ages 25-64 are rising in the United States for many causes of death and geographic locations

MORTALITYMINDER

Select cause of death and state:

Deaths of Despair | California

Mortality rates are rising in the United States depending on where you live.

MortalityMinder analyzes trends of premature death in the United States which are caused by:

- Deaths of Despair
- Cardiovascular Disease
- Cancer
- Assault Deaths
- All Causes

MortalityMinder is a four-page interactive presentation that examines county-level factors associated with mortality trends.

Pick a cause of death and state on the menu bar at the top of the page to see how mortality rates in the United States have changed from 2000 to 2017.

Click right and left at the edges of your screen to investigate further.

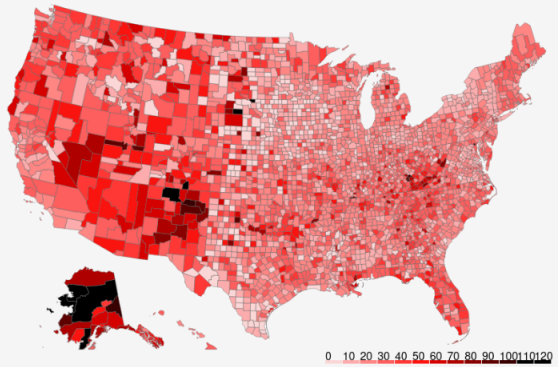
Rensselaer

Deaths of Despair Rates Over Time

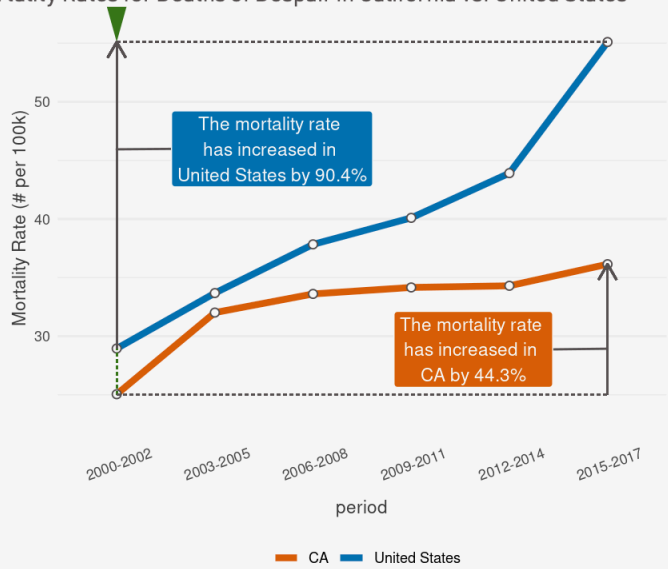
"Deaths of Despair" are deaths due to suicide, overdose, substance abuse and poisonings

2000-02 | 2003-05 | 2006-08 | 2009-11 | 2012-14 | 2015-17

Nationwide Deaths of Despair Rates for 2000-2002



Mortality Rates for Deaths of Despair in California vs. United States



Mortality rate per 100,000 for people ages 25-to 64 due to Deaths of Despair for three year periods for counties (left) and state and nation (right) . Darker colors indicate higher rates.

Source: CDC WONDER

Analysis: Institute for Data Exploration and Applications at Rensselaer Polytechnic Institute

MORTALITYMINDER

Select cause of death and state:

Deaths of Despair | California

Mortality rates are rising in the United States depending on where you live.

MortalityMinder analyzes trends of premature death in the United States which are caused by:

- Deaths of Despair
- Cardiovascular Disease
- Cancer
- Assault Deaths
- All Causes

MortalityMinder is a four-page interactive presentation that examines county-level factors associated with mortality trends.

Pick a cause of death and state on the menu bar at the top of the page to see how mortality rates in the United States have changed from 2000 to 2017.

Click right and left at the edges of your screen to investigate further.

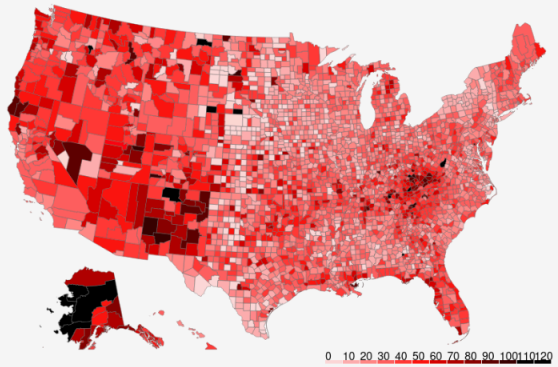
Rensselaer

Deaths of Despair Rates Over Time

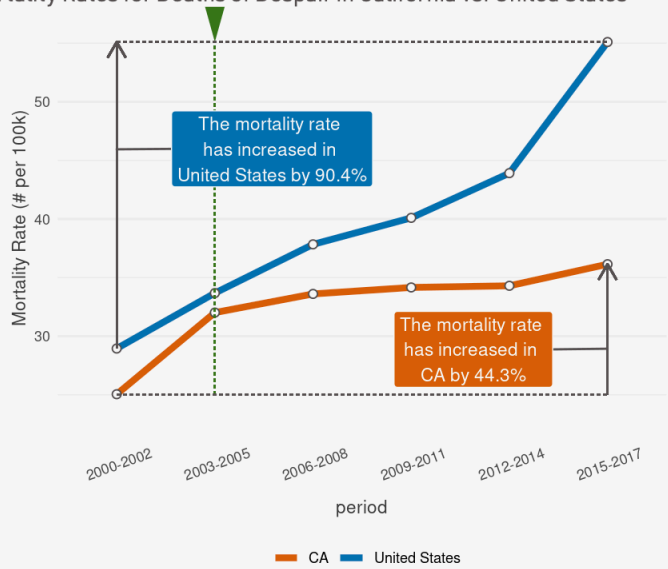
"Deaths of Despair" are deaths due to suicide, overdose, substance abuse and poisonings

2000-02 | 2003-05 | 2006-08 | 2009-11 | 2012-14 | 2015-17

Nationwide Deaths of Despair Rates for 2003-2005



Mortality Rates for Deaths of Despair in California vs. United States



Mortality rate per 100,000 for people ages 25-to 64 due to Deaths of Despair for three year periods for counties (left) and state and nation (right) . Darker colors indicate higher rates.

Source: CDC WONDER

Analysis: Institute for Data Exploration and Applications at Rensselaer Polytechnic Institute

MORTALITYMINDER

Select cause of death and state:

Deaths of Despair | California

Mortality rates are rising in the United States depending on where you live.

MortalityMinder analyzes trends of premature death in the United States which are caused by:

- Deaths of Despair
- Cardiovascular Disease
- Cancer
- Assault Deaths
- All Causes

MortalityMinder is a four-page interactive presentation that examines county-level factors associated with mortality trends.

Pick a cause of death and state on the menu bar at the top of the page to see how mortality rates in the United States have changed from 2000 to 2017.

Click right and left at the edges of your screen to investigate further.

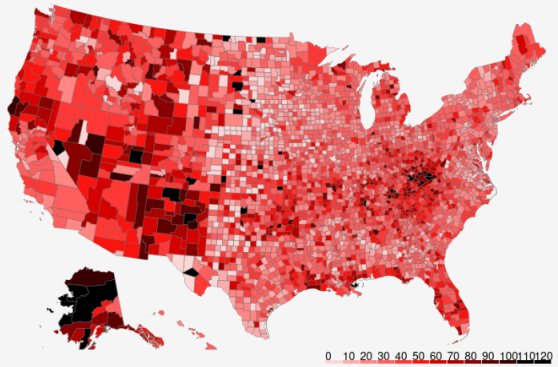
Rensselaer

Deaths of Despair Rates Over Time

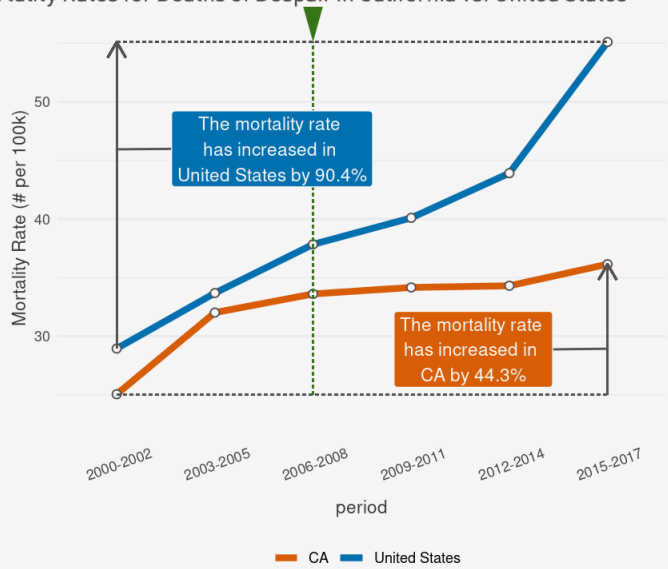
"Deaths of Despair" are deaths due to suicide, overdose, substance abuse and poisonings

2000-02 | 2003-05 | **2006-08** | 2009-11 | 2012-14 | 2015-17

Nationwide Deaths of Despair Rates for 2006-2008



Mortality Rates for Deaths of Despair in California vs. United States



Mortality rate per 100,000 for people ages 25-to 64 due to Deaths of Despair for three year periods for counties (left) and state and nation (right) . Darker colors indicate higher rates.

Source: CDC WONDER

Analysis: Institute for Data Exploration and Applications at Rensselaer Polytechnic Institute

MORTALITYMINDER

Select cause of death and state:

Deaths of Despair | California

Mortality rates are rising in the United States depending on where you live.

MortalityMinder analyzes trends of premature death in the United States which are caused by:

- Deaths of Despair
- Cardiovascular Disease
- Cancer
- Assault Deaths
- All Causes

MortalityMinder is a four-page interactive presentation that examines county-level factors associated with mortality trends.

Pick a cause of death and state on the menu bar at the top of the page to see how mortality rates in the United States have changed from 2000 to 2017.

Click right and left at the edges of your screen to investigate further.

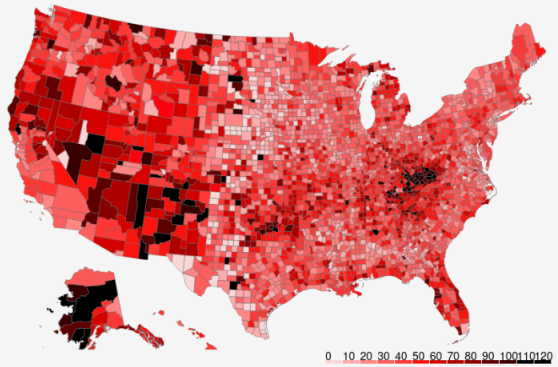
Rensselaer

Deaths of Despair Rates Over Time

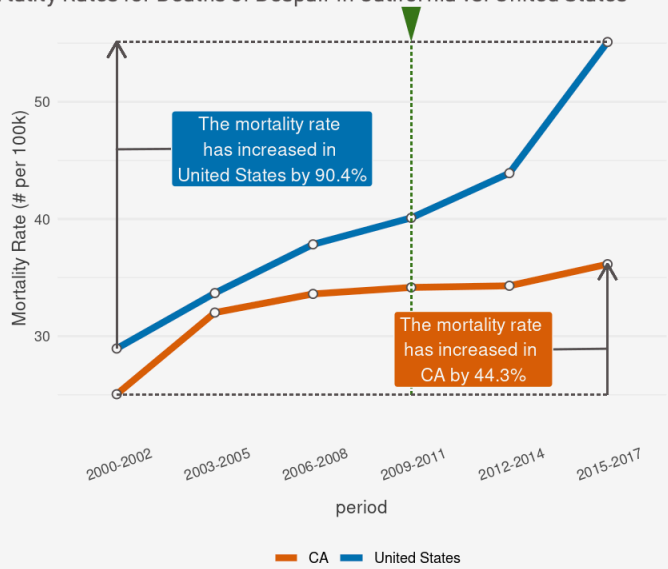
"Deaths of Despair" are deaths due to suicide, overdose, substance abuse and poisonings

2000-02 | 2003-05 | 2006-08 | **2009-11** | 2012-14 | 2015-17

Nationwide Deaths of Despair Rates for 2009-2011



Mortality Rates for Deaths of Despair in California vs. United States



Mortality rate per 100,000 for people ages 25-to 64 due to Deaths of Despair for three year periods for counties (left) and state and nation (right) . Darker colors indicate higher rates.

Source: CDC WONDER

Analysis: Institute for Data Exploration and Applications at Rensselaer Polytechnic Institute

MORTALITYMINDER

Select cause of death and state:

Deaths of Despair | California

Mortality rates are rising in the United States depending on where you live.

MortalityMinder analyzes trends of premature death in the United States which are caused by:

- Deaths of Despair
- Cardiovascular Disease
- Cancer
- Assault Deaths
- All Causes

MortalityMinder is a four-page interactive presentation that examines county-level factors associated with mortality trends.

Pick a cause of death and state on the menu bar at the top of the page to see how mortality rates in the United States have changed from 2000 to 2017.

Click right and left at the edges of your screen to investigate further.

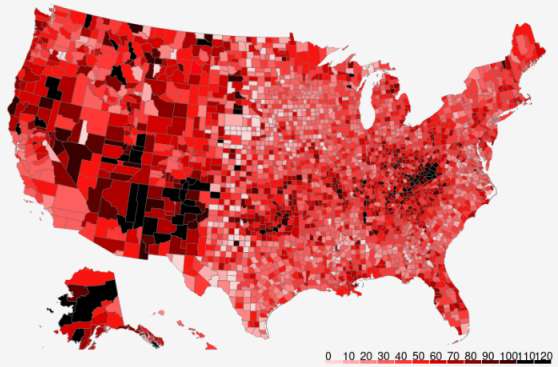
Rensselaer

Deaths of Despair Rates Over Time

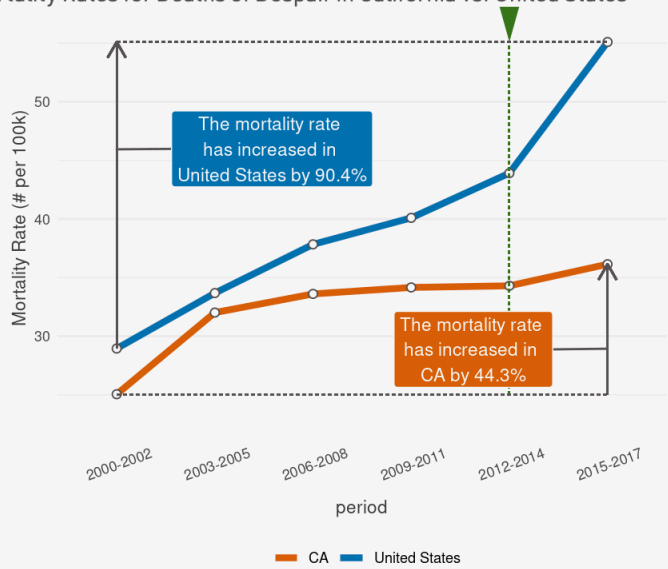
"Deaths of Despair" are deaths due to suicide, overdose, substance abuse and poisonings

2000-02 | 2003-05 | 2006-08 | 2009-11 | **2012-14** | 2015-17

Nationwide Deaths of Despair Rates for 2012-2014



Mortality Rates for Deaths of Despair in California vs. United States



Mortality rate per 100,000 for people ages 25-to 64 due to Deaths of Despair for three year periods for counties (left) and state and nation (right) . Darker colors indicate higher rates.

Source: CDC WONDER

Analysis: Institute for Data Exploration and Applications at Rensselaer Polytechnic Institute

MORTALITYMINDER

Select cause of death and state:

Deaths of Despair | California

Mortality rates are rising in the United States depending on where you live.

MortalityMinder analyzes trends of premature death in the United States which are caused by:

- Deaths of Despair
- Cardiovascular Disease
- Cancer
- Assault Deaths
- All Causes

MortalityMinder is a four-page interactive presentation that examines county-level factors associated with mortality trends.

Pick a cause of death and state on the menu bar at the top of the page to see how mortality rates in the United States have changed from 2000 to 2017.

Click right and left at the edges of your screen to investigate further.

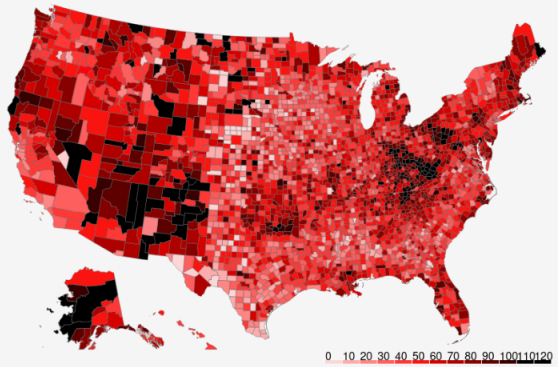
Rensselaer

Deaths of Despair Rates Over Time

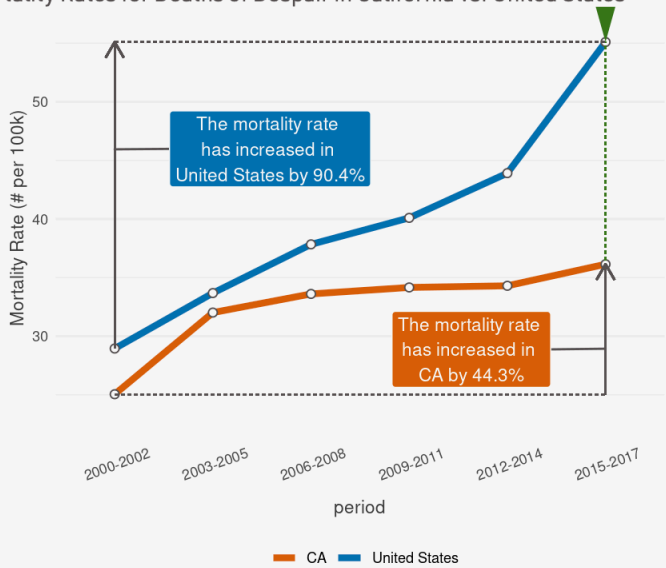
"Deaths of Despair" are deaths due to suicide, overdose, substance abuse and poisonings

2000-02 | 2003-05 | 2006-08 | 2009-11 | 2012-14 | **2015-17**

Nationwide Deaths of Despair Rates for 2015-2017



Mortality Rates for Deaths of Despair in California vs. United States

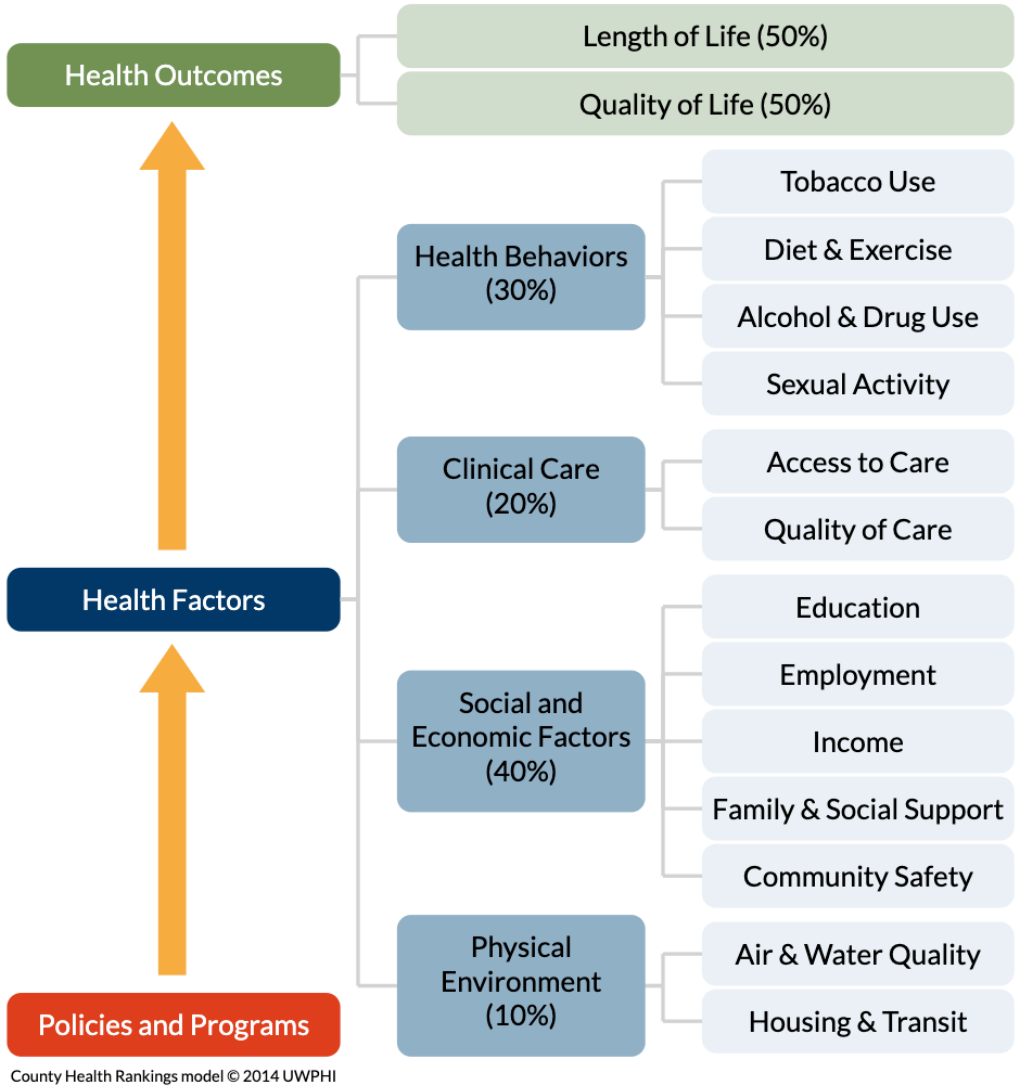


Mortality rate per 100,000 for people ages 25-to 64 due to Deaths of Despair for three year periods for counties (left) and state and nation (right) . Darker colors indicate higher rates.

Source: CDC WONDER

Analysis: Institute for Data Exploration and Applications at Rensselaer Polytechnic Institute

Social Determinants/Factors affect Population Health



County-Level Data:

Mortality rates from CDC Wonder for 2000-2017

147 health factors for 2015-2017 gathered by CountyHealthRankings.org

MORTALITYMINDER

Select cause of death and state:

Deaths of Despair | California

Exploring Causes of Premature Death

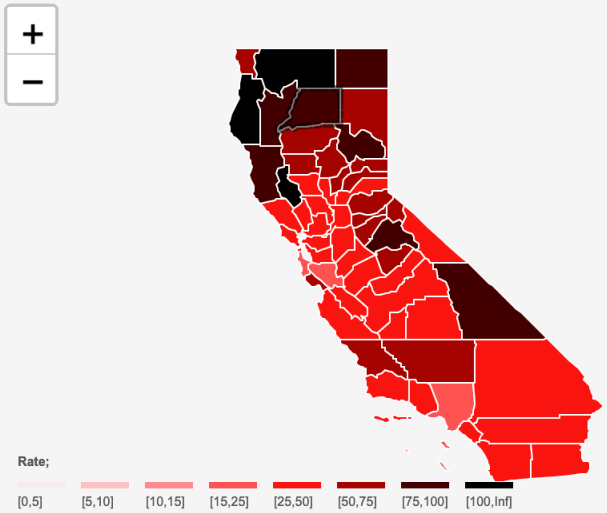
Mortality rates for Deaths of Despair for the State of California

"Deaths of Despair" are deaths due to suicide, overdose, substance abuse and poisonings

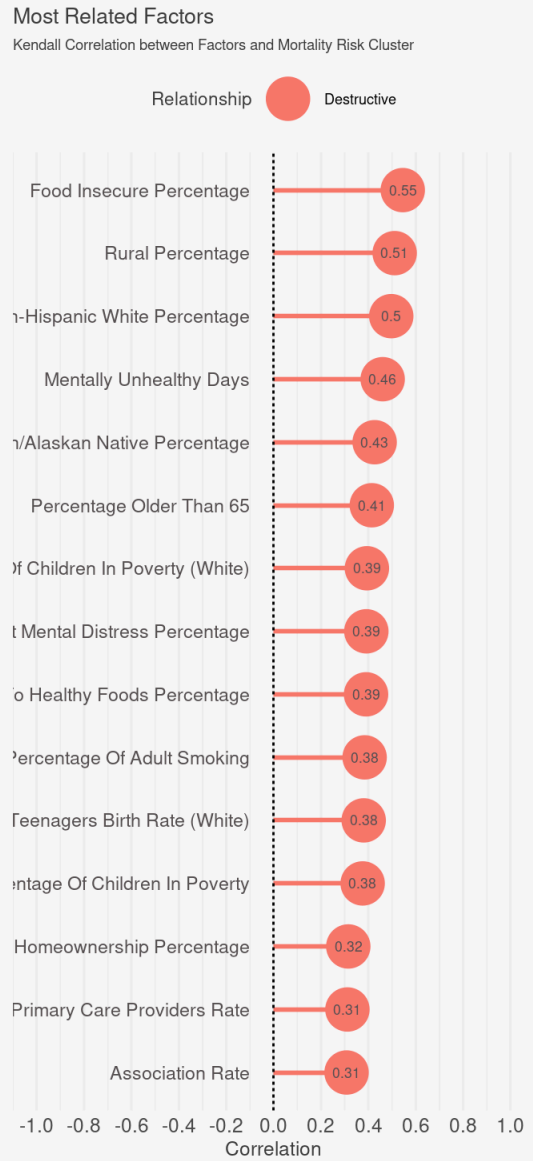
Select year range to see statewide mortality rate distribution for that period. Mouse over maps to identify individual counties. Zoom map with mouse wheel or zoom buttons.

Deaths of Despair Mortality rates for California for 2015-2017

2000-02
2003-05
2006-08
2009-11
2012-14
2015-17

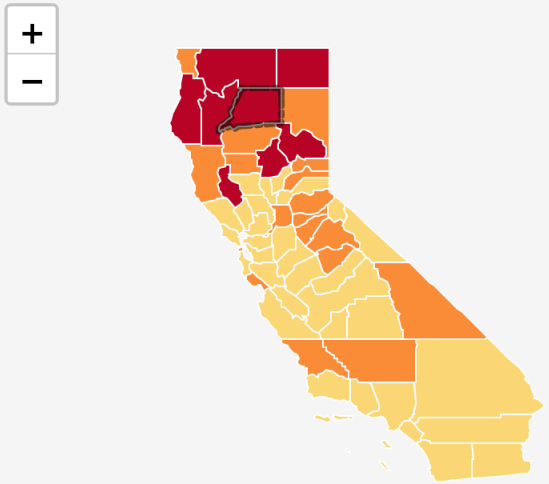


Factors related to Deaths of Despair for California

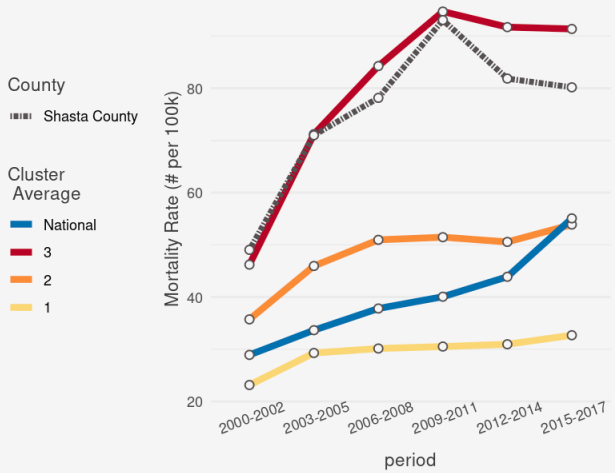


Data Source:
1.CDCWONDER Multi-Cause of Death
2.County Health Ranking 2019

Deaths of Despair Clusters for California



Deaths of Despair Trends for California



MORTALITYMINDER

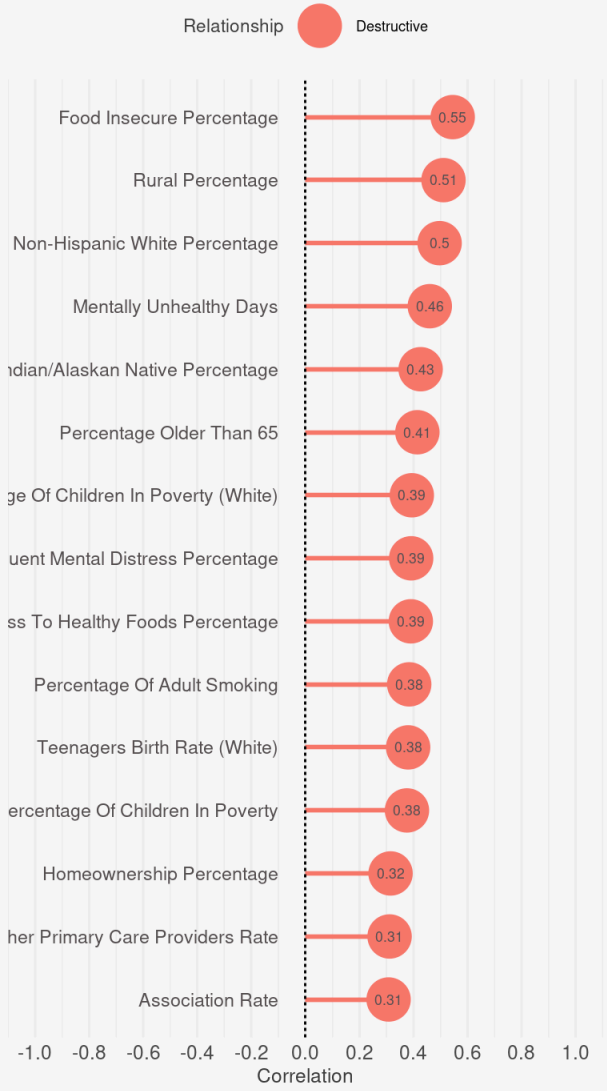
Select cause of death and state:

Deaths of Despair | California

Factors related to Deaths of Despair for California

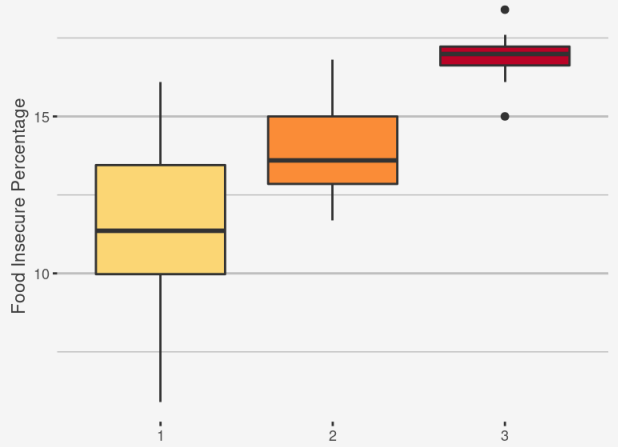
Most Related Factors

Kendall Correlation between Factors and Mortality Risk Cluster

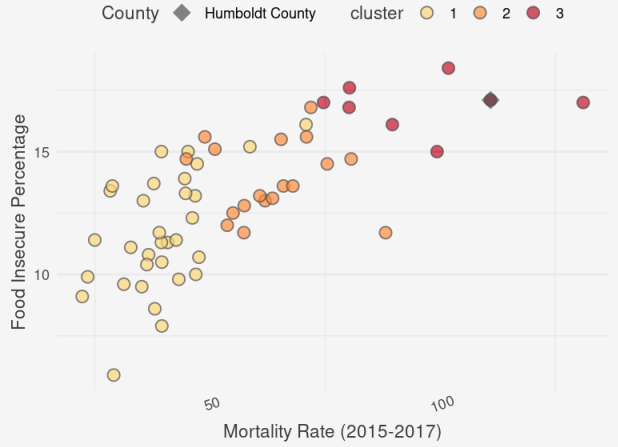


Data Source:
1.CDCWONDER Multi-Cause of Death
2.County Health Ranking 2019

Food Insecure Percentage and Risk Cluster Relationship



Food Insecure Percentage and Mortality Relationship



Select a determinant:

Food Insecure Percentage

Food Insecure Percentage

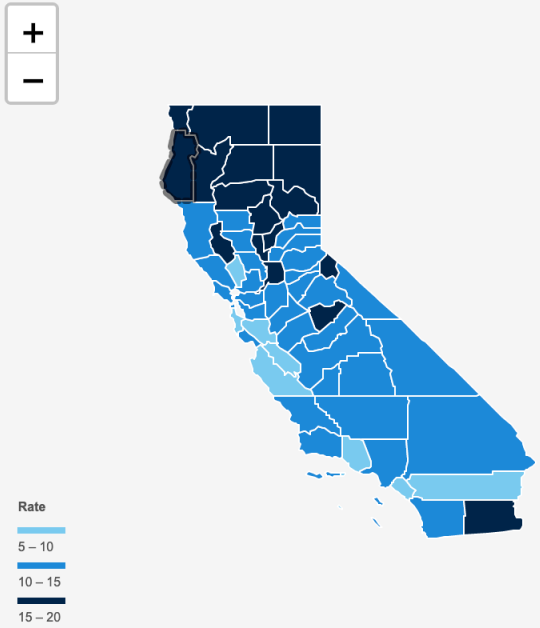
Percentage of population who lack adequate access to food.

Kendal Correlation with Despair mortality: **0.5452**

Statistically significant **destructive** relationship with mortality (p-value = 1.8e-09)

Click here for more information

Food Insecure Percentage Distribution for California



Select a county below or by clicking the map:

Humboldt

MORTALITYMINDER

Select cause of death and state:

Deaths of Despair | New Mexico

Exploring Causes of Premature Death

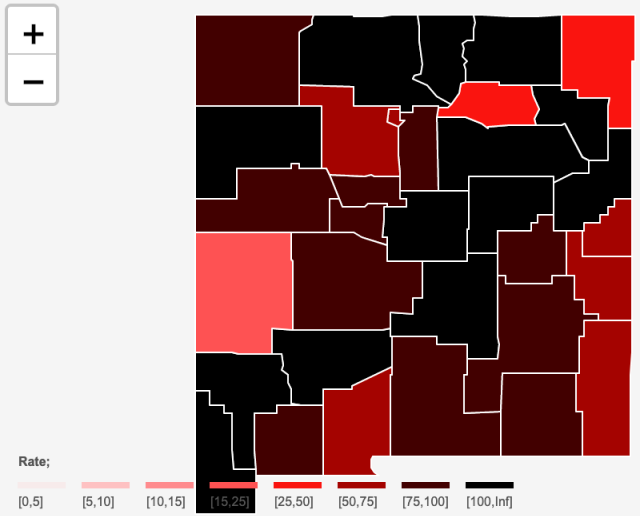
Mortality rates for Deaths of Despair for the State of New Mexico

"Deaths of Despair" are deaths due to suicide, overdose, substance abuse and poisonings

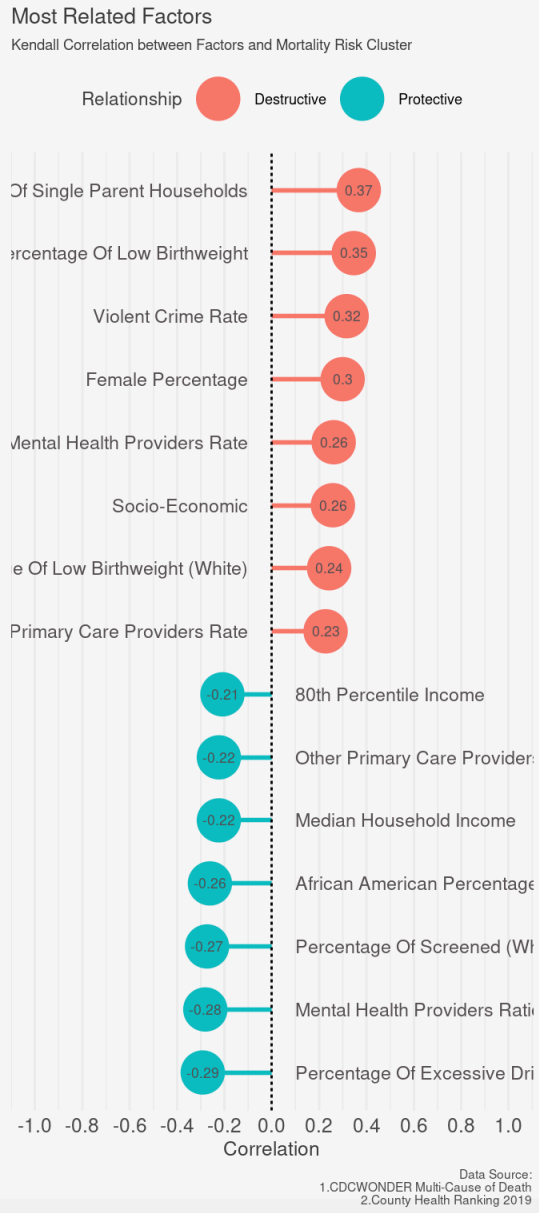
Select year range to see statewide mortality rate distribution for that period. Mouse over maps to identify individual counties. Zoom map with mouse wheel or zoom buttons.

Deaths of Despair Mortality rates for New Mexico for 2015-2017

2000-02
2003-05
2006-08
2009-11
2012-14
2015-17

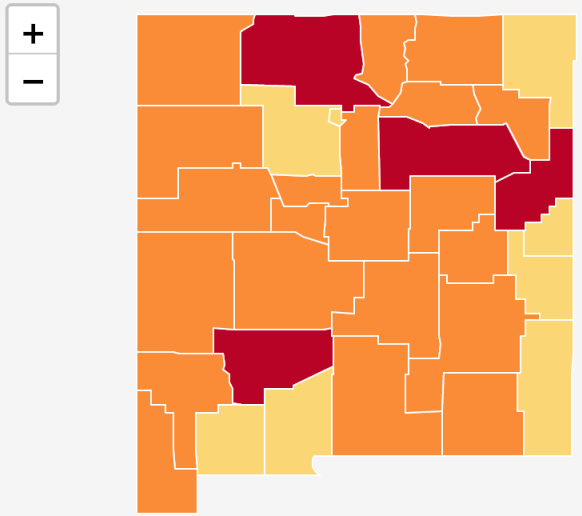


Factors related to Deaths of Despair for New Mexico

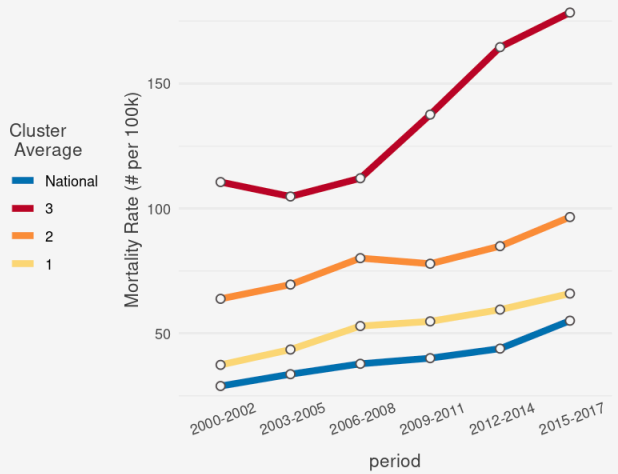


Data Source:
1.CDCWONDER Multi-Cause of Death
2.County Health Ranking 2019

Deaths of Despair Clusters for New Mexico



Deaths of Despair Trends for New Mexico



MORTALITYMINDER

Select cause of death and state:

Cardiovascular Disease | Massachusetts

Exploring Causes of Premature Death

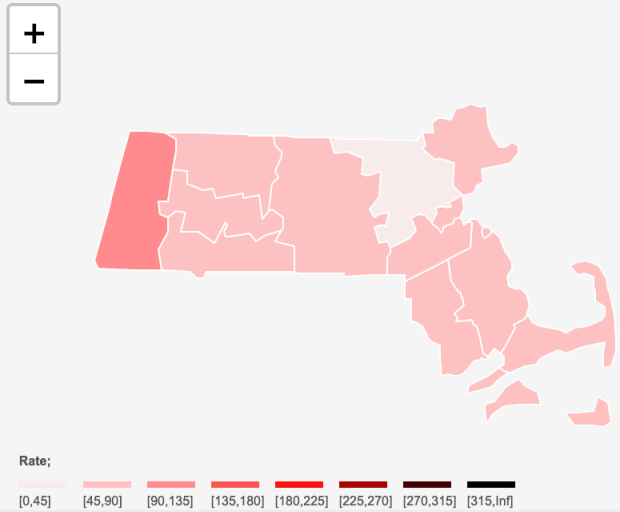
Mortality rates for Cardiovascular Disease for the State of Massachusetts

"Cardiovascular Disease" are deaths due to diseases of the circulatory systems such as heart disease and stroke

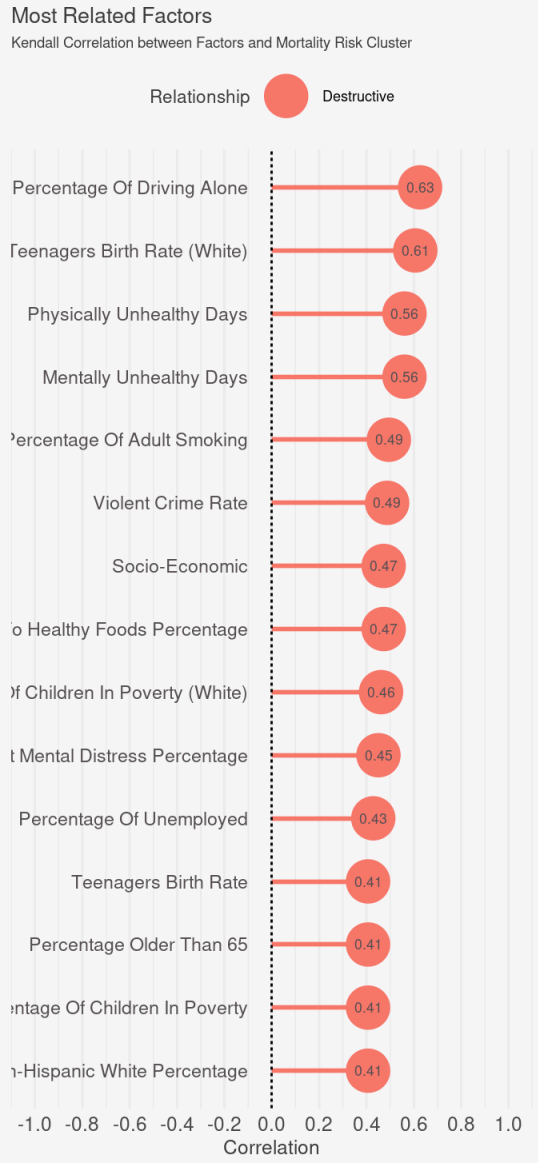
Select year range to see statewide mortality rate distribution for that period. Mouse over maps to identify individual counties. Zoom map with mouse wheel or zoom buttons.

Cardiovascular Disease Mortality rates for Massachusetts for 2015-2017

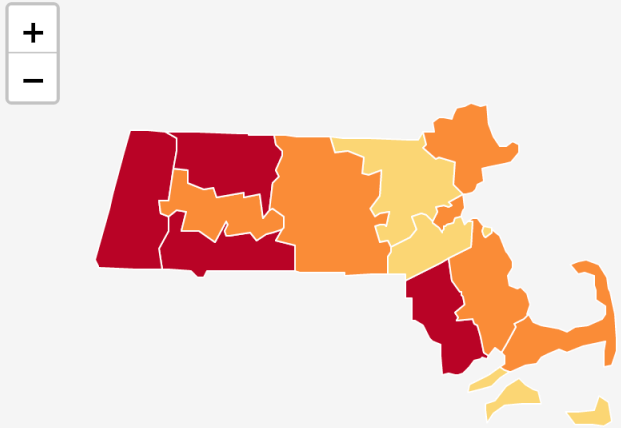
2000-02 | 2003-05 | 2006-08 | 2009-11 | 2012-14 | 2015-17



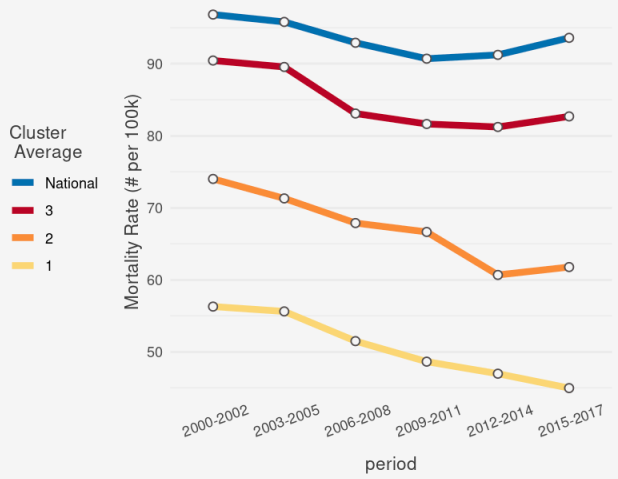
Factors related to Cardiovascular Disease for Massachusetts



Cardiovascular Disease Clusters for Massachusetts



Cardiovascular Disease Trends for Massachusetts



Data Source:
1.CDCWONDER Multi-Cause of Death
2.County Health Ranking 2019

MORTALITYMINDER

Select cause of death and state:

Cardiovascular Disease | South Carolina

Exploring Causes of Premature Death

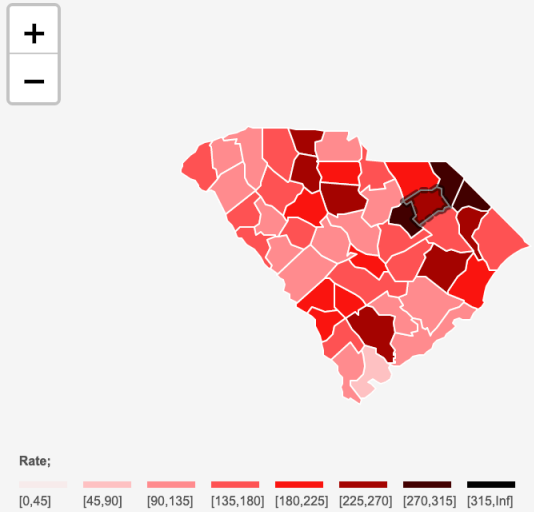
Mortality rates for Cardiovascular Disease for the State of South Carolina

"Cardiovascular Disease" are deaths due to diseases of the circulatory systems such as heart disease and stroke

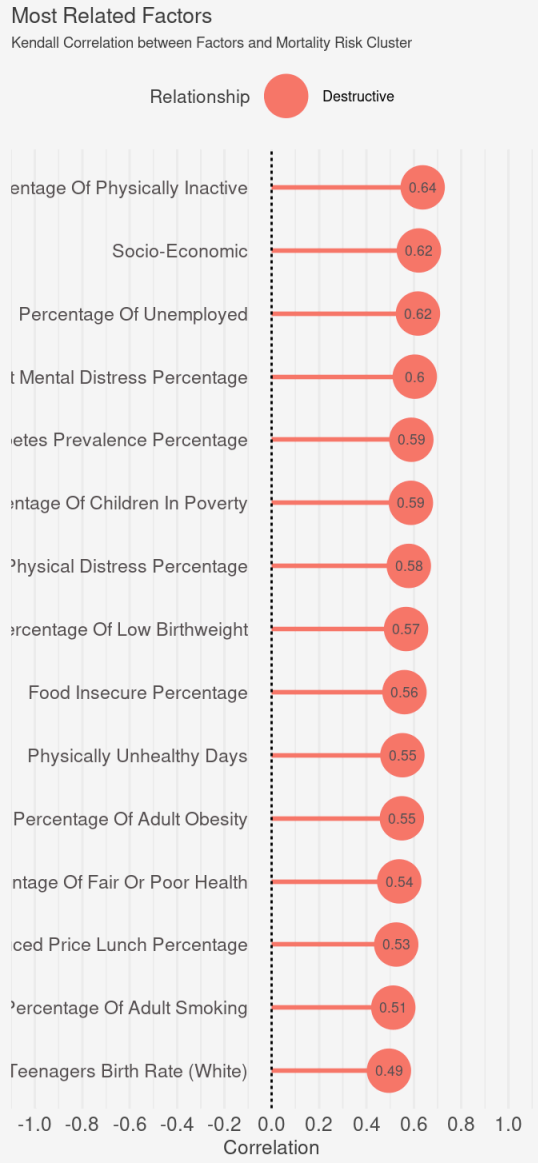
Select year range to see statewide mortality rate distribution for that period. Mouse over maps to identify individual counties. Zoom map with mouse wheel or zoom buttons.

Cardiovascular Disease Mortality rates for South Carolina for 2015-2017

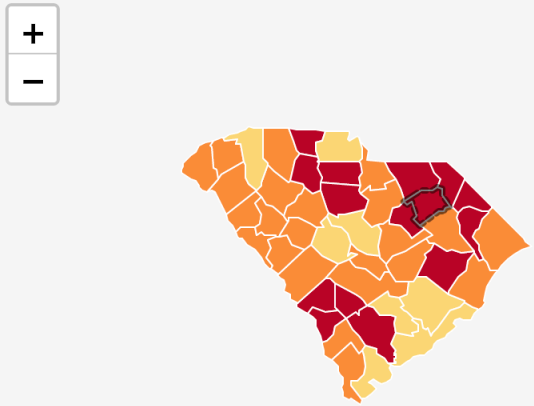
2000-02 | 2003-05 | 2006-08 | 2009-11 | 2012-14 | 2015-17



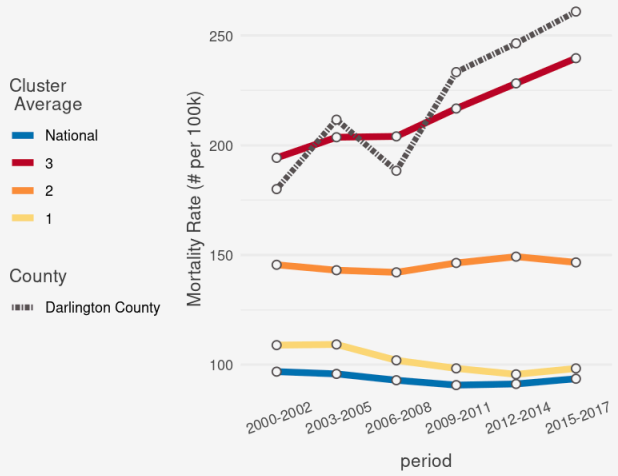
Factors related to Cardiovascular Disease for South Carolina



Cardiovascular Disease Clusters for South Carolina



Cardiovascular Disease Trends for South Carolina



Data Source:
1.CDCWONDER Multi-Cause of Death
2.County Health Ranking 2019

MortalityMinder

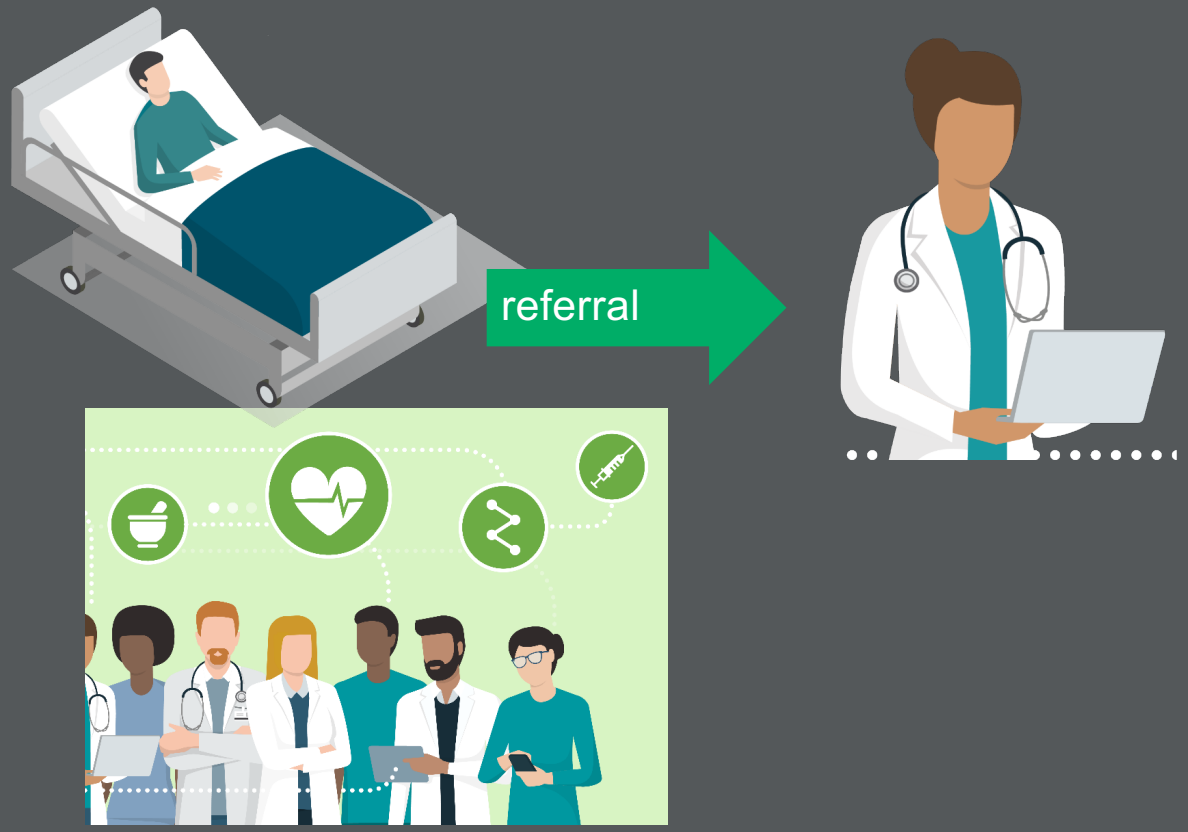
Open source R-Shiny App created by students in Data INCITE Lab

Designed for Decision Makers

Phase 1 winner of

AHRQ Visualization Resources of Community-Level Social Determinants of Health Challenge

Study 2: Referrals to Complex Care Management in HMO



Model finds patients for referral missed by current referral practices

Predicts referral using:

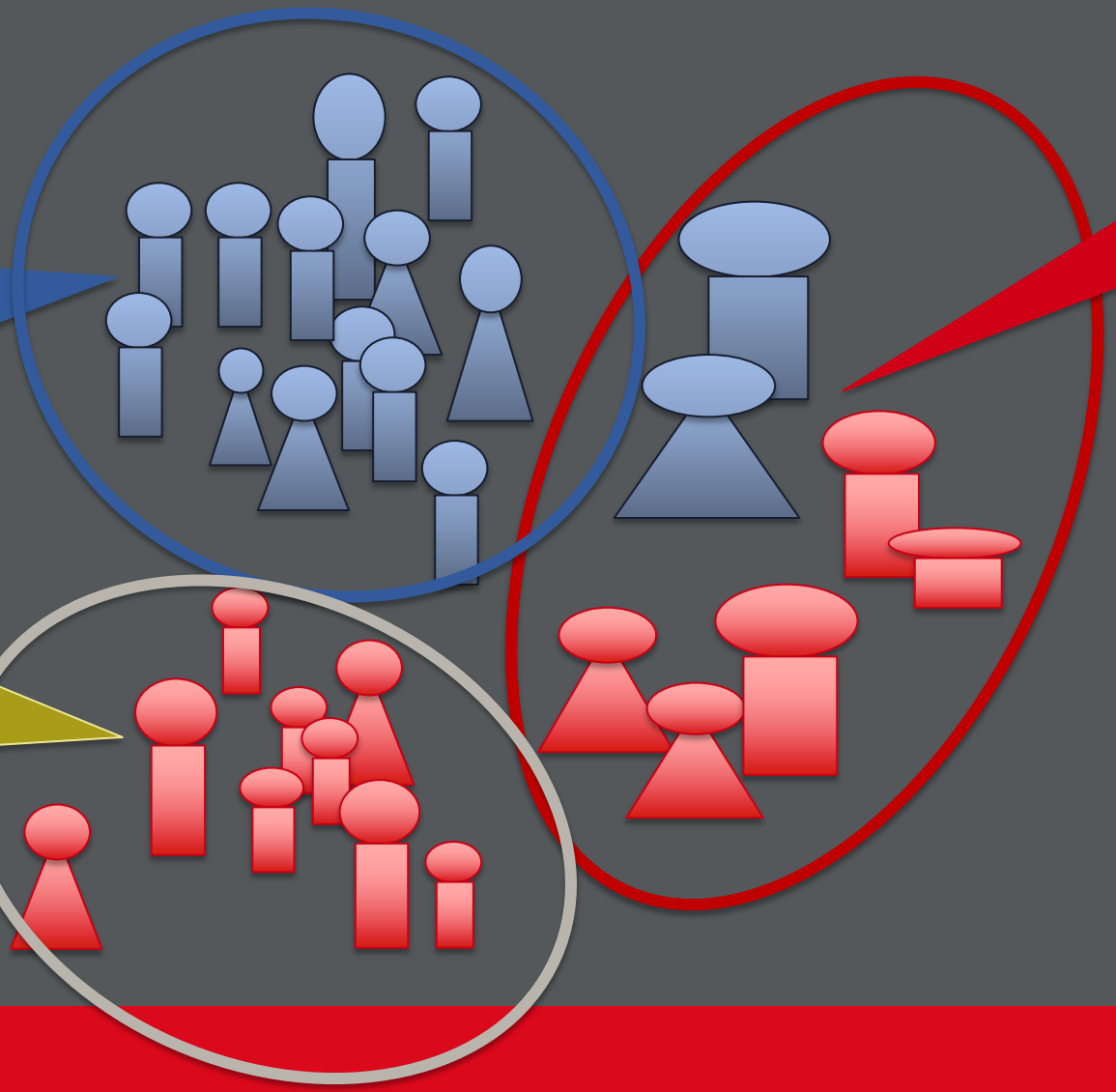
- 90 features: demographics, diagnoses, utilization from Claims and Electronic Medical Records
- Past referrals by doctors

Rensselaer Polytechnic Institute: Georgios Mavroudeas, Xiao Shou, Jason Kuruzovich, Malik Magdon-Ismail, Kristin Bennett **CDPHP:** Matt Vielkind, Mouad Seridi

Simultaneously Discover and Predict Referral Subpopulations

Low Risk
- Low expenditures
- Few comorbidities
- …

Medium Risk
- Mental Health
- ER admission
- ….



High Risk
- Multiple Comorbidities
- Hepatitis
- In-patient

…

Cadre Machine Learning

Supervised Gaussian Mixtures for Risk Analysis

- Simultaneously divides observations into subpopulations (clusters) and learns subpopulation-specific risk models
- E.g. subgroups *specific to the* target dependent variable, referral

$$p(x \mid M) = \sum_{k=1}^{K} \pi_k N(x \mid \mu_k, \Sigma_k)$$

The total probability of a point x to be generated by a model M

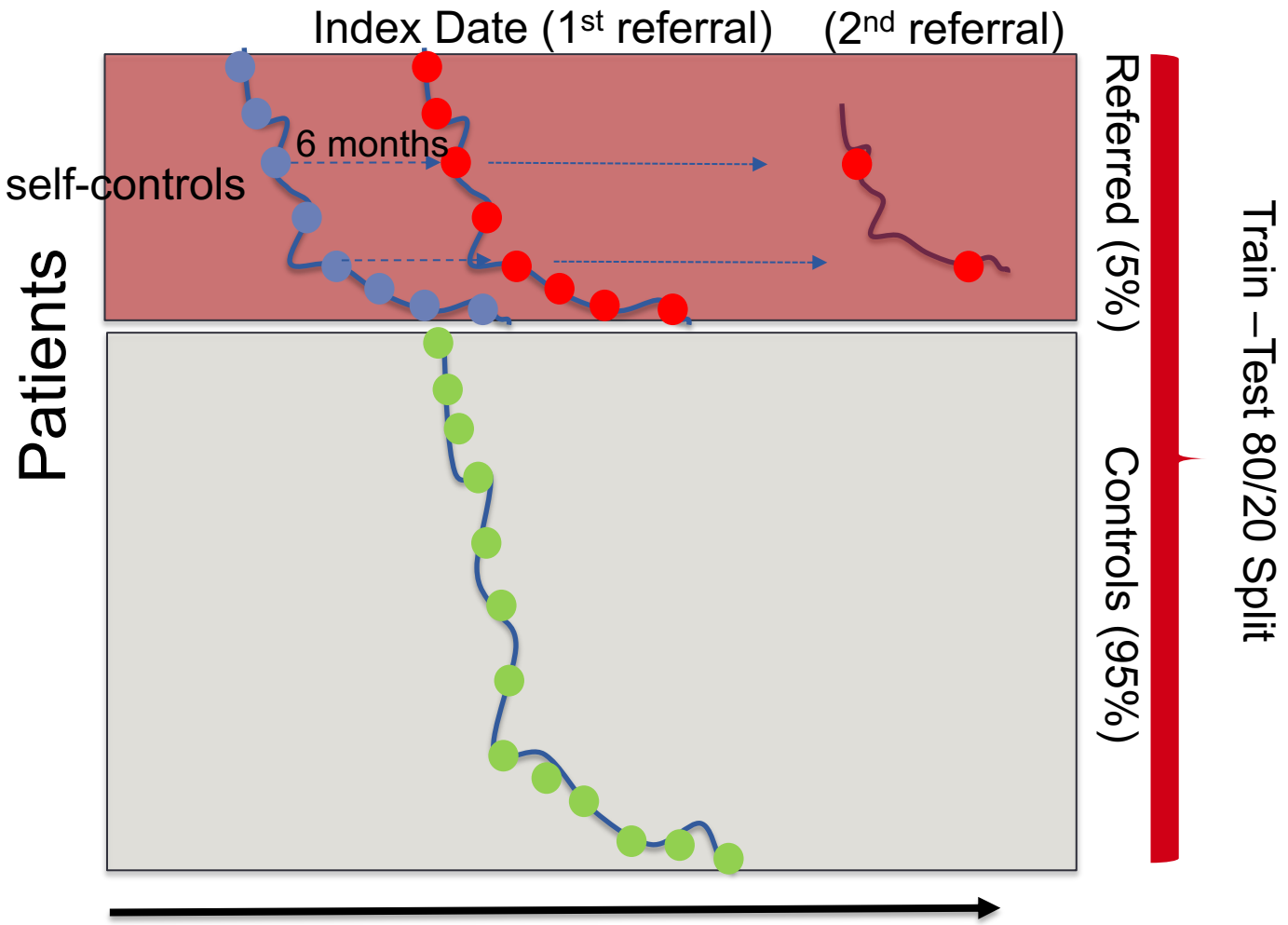
$$p(y = 1 \mid x, M) = \sum_{k=1}^{K} \frac{\pi_k N(x \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x \mid \mu_j, \Sigma_j)} \sigma(w_k^T x)$$

The probability that a point x has the label y = 1 under M

membership classifier

- L1-regularized logistic regression

Study Design



Feature Construction

Utilization(Continuous Features)

Diagnosis Group(Binary Features)

PERSON	Year_Month	allow_amt	allow_snf	…	schizophrenia	hepatitis	…	Age	Referral (index month)
12345678	201501	xxx	xxx	x	x	x	x	x	0
12345678	201502	xxx	xxx	x	1	1	x	x	0
12345678	201503	xxx	xxx	x	1	1	x	x	1

Additional features constructed as

allow_amt, …

allow_amt_prior1m, ….

allow_amt_prior2m, ….

Index month

Prior 1 month

Prior 2 month

Diagnosis group only has index month

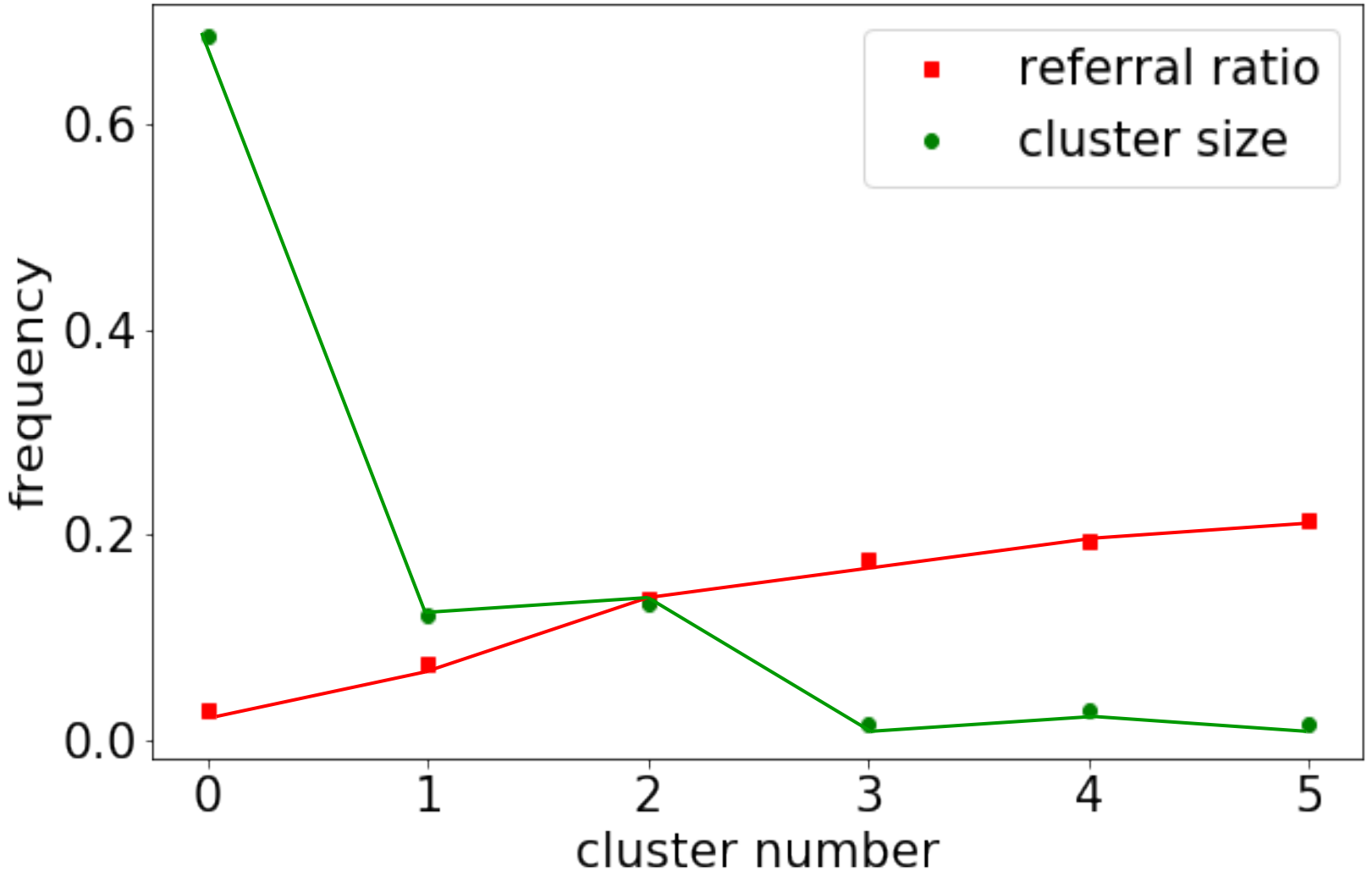
Prediction Test Accuracy

Supervised Gaussian Mixture Model (SGMM) is interpretable and predicts well

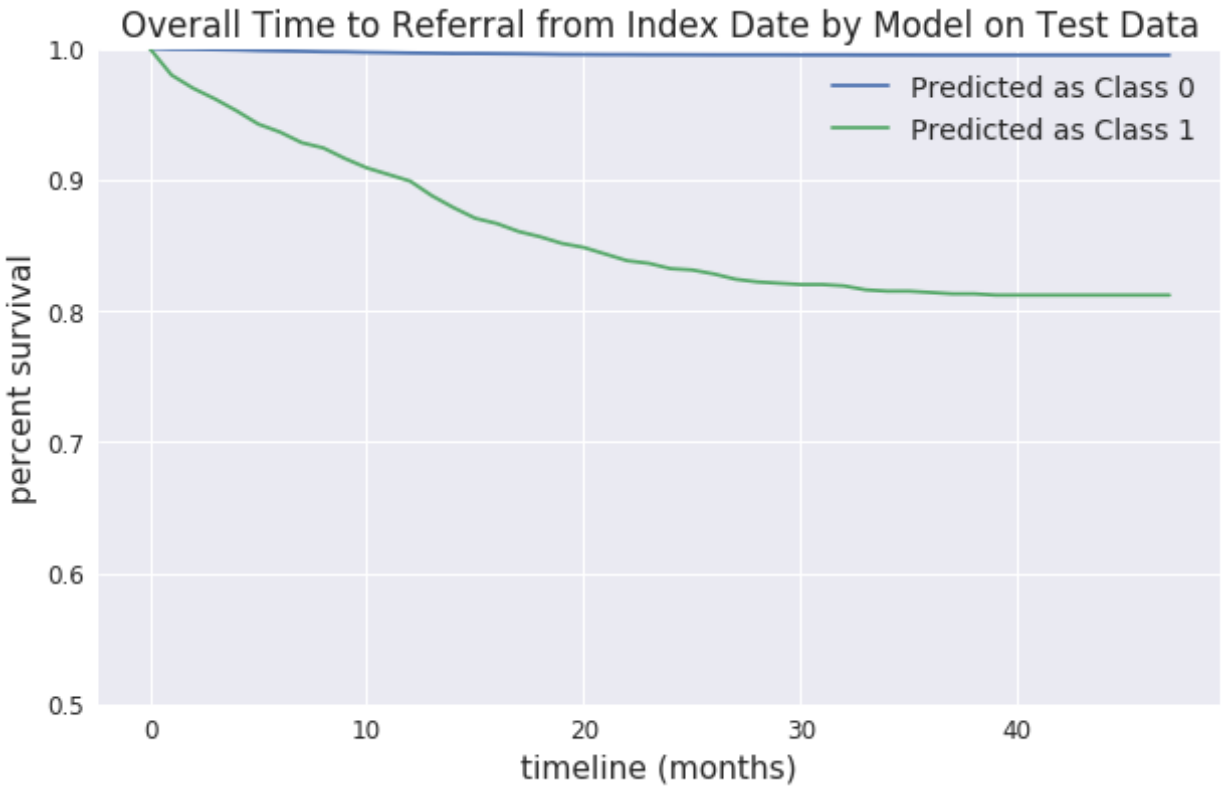
Neural Network has higher accuracy, but is not explainable and yields worse prediction in larger longitudinal analysis

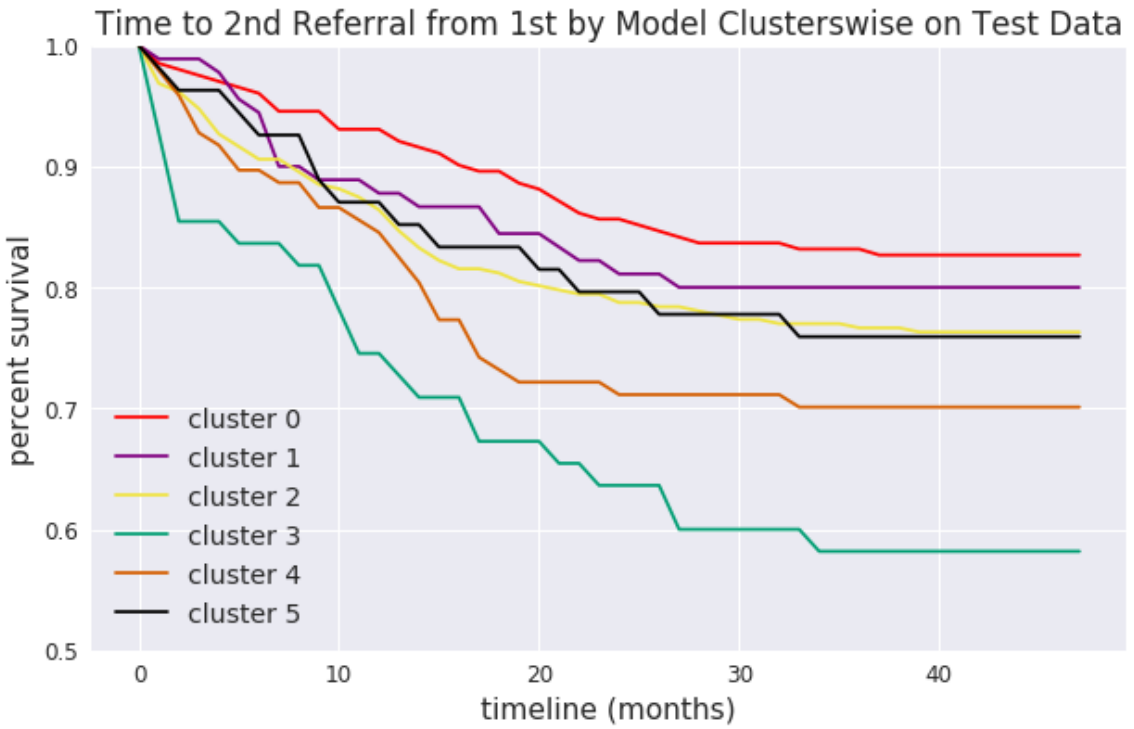
Methods	Recall	Precision	AUC
Log Reg (dx_cnt)	**0.37**	**0.23**	**0.79**
L1 Log Reg	0.42	0.42	0.83
Neural Net	**0.77**	**0.54**	**0.95**
Random Forest (RF)	0.48	0.53	0.91
Adaboost	0.47	0.32	0.86
GradBoost	0.47	0.59	0.91
SGMM	**0.45**	**0.50**	**0.86**
SGMM + RF	**0.45**	**0.52**	**0.90**

SGMM Stratifies Patient Risk Within and Across Clusters

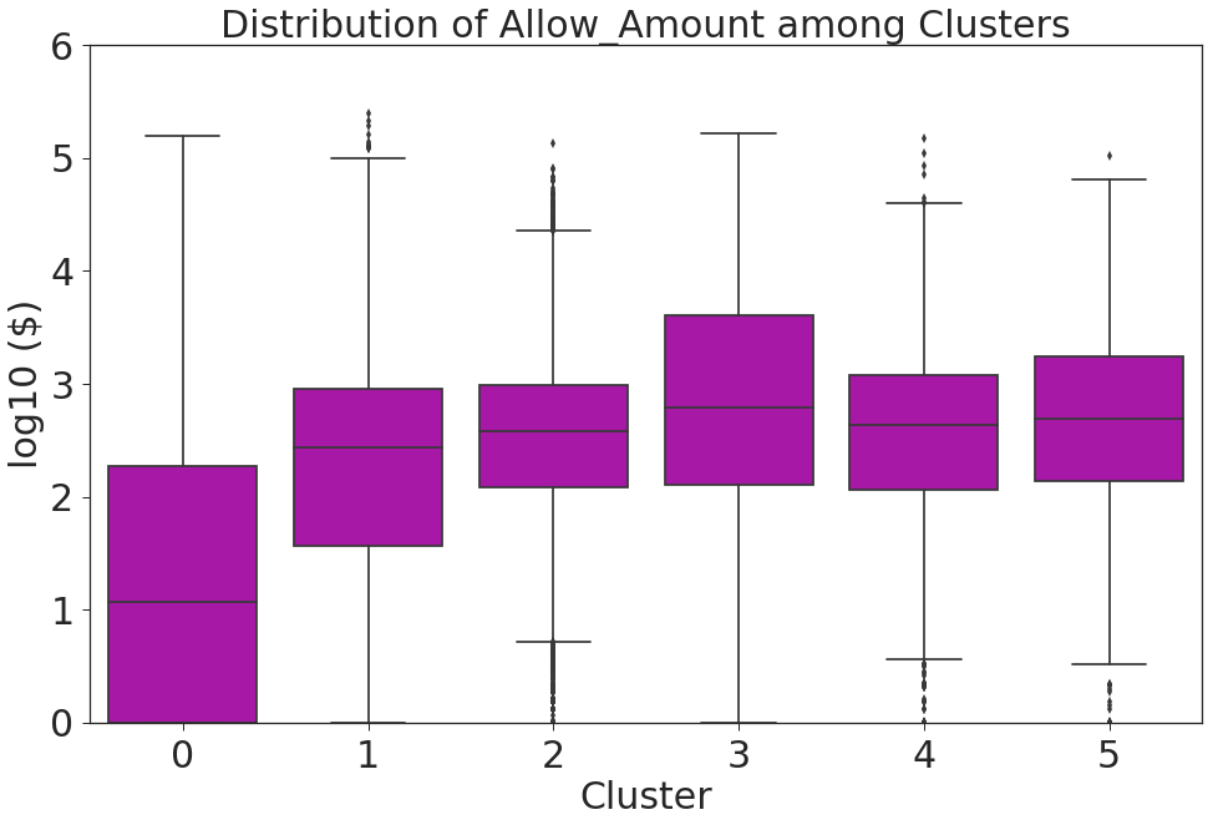


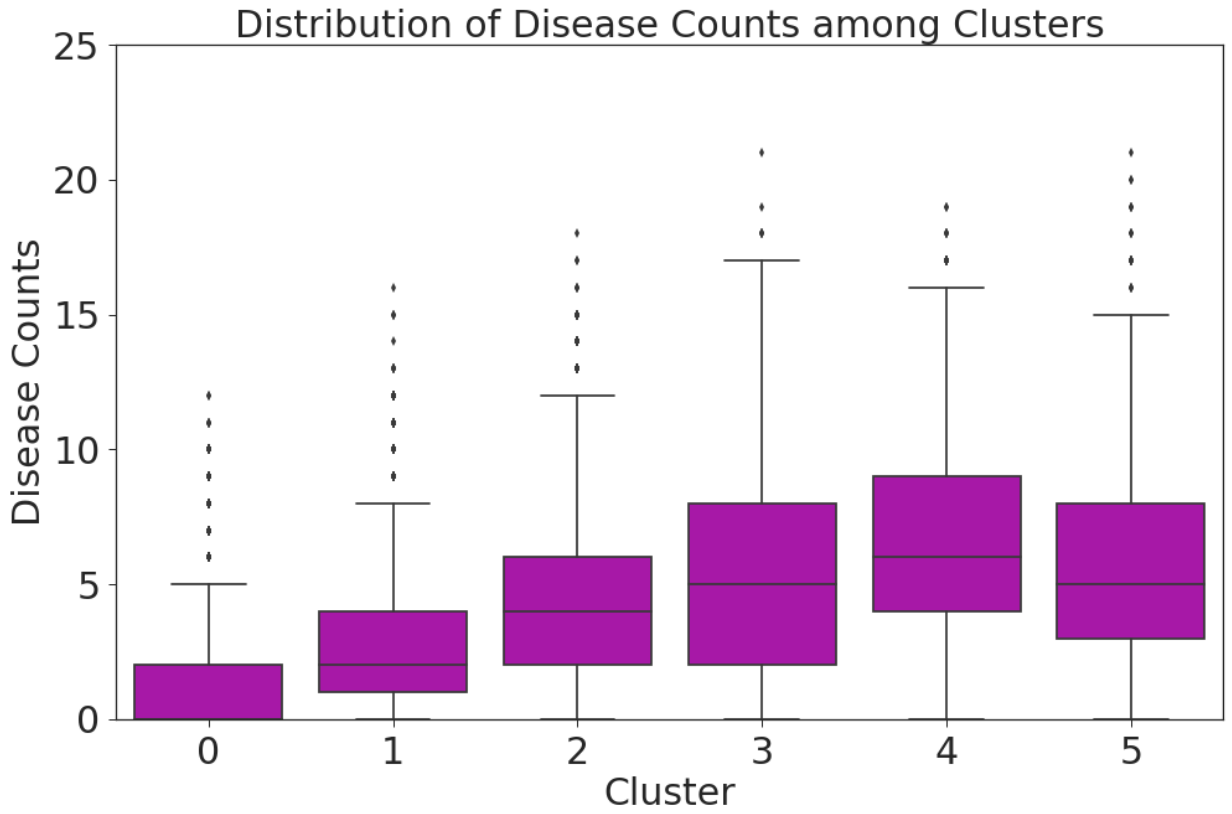
Cadres Stratify Risk Within and Across Cadres





Costs and Comorbidities Vary by Cluster

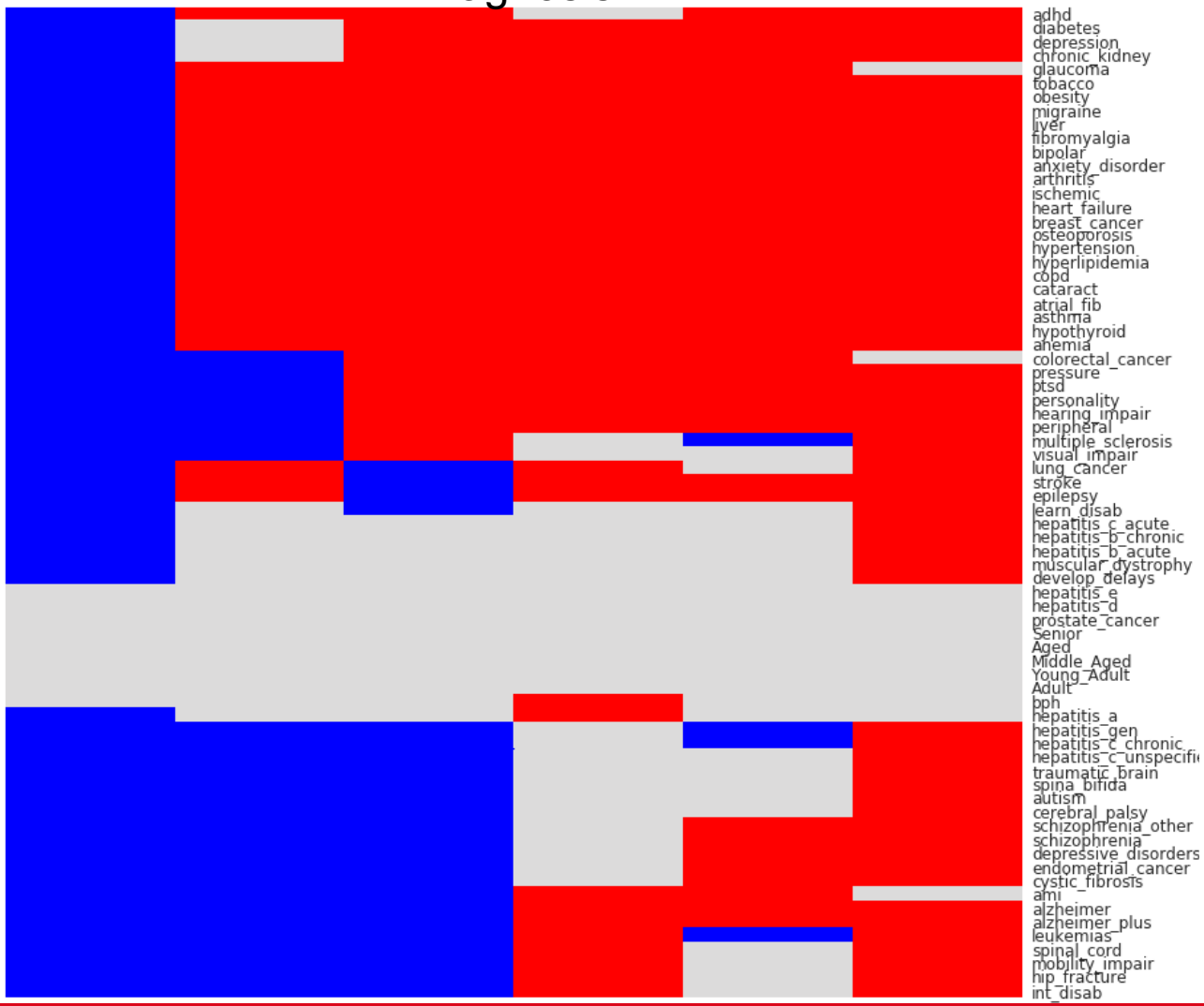




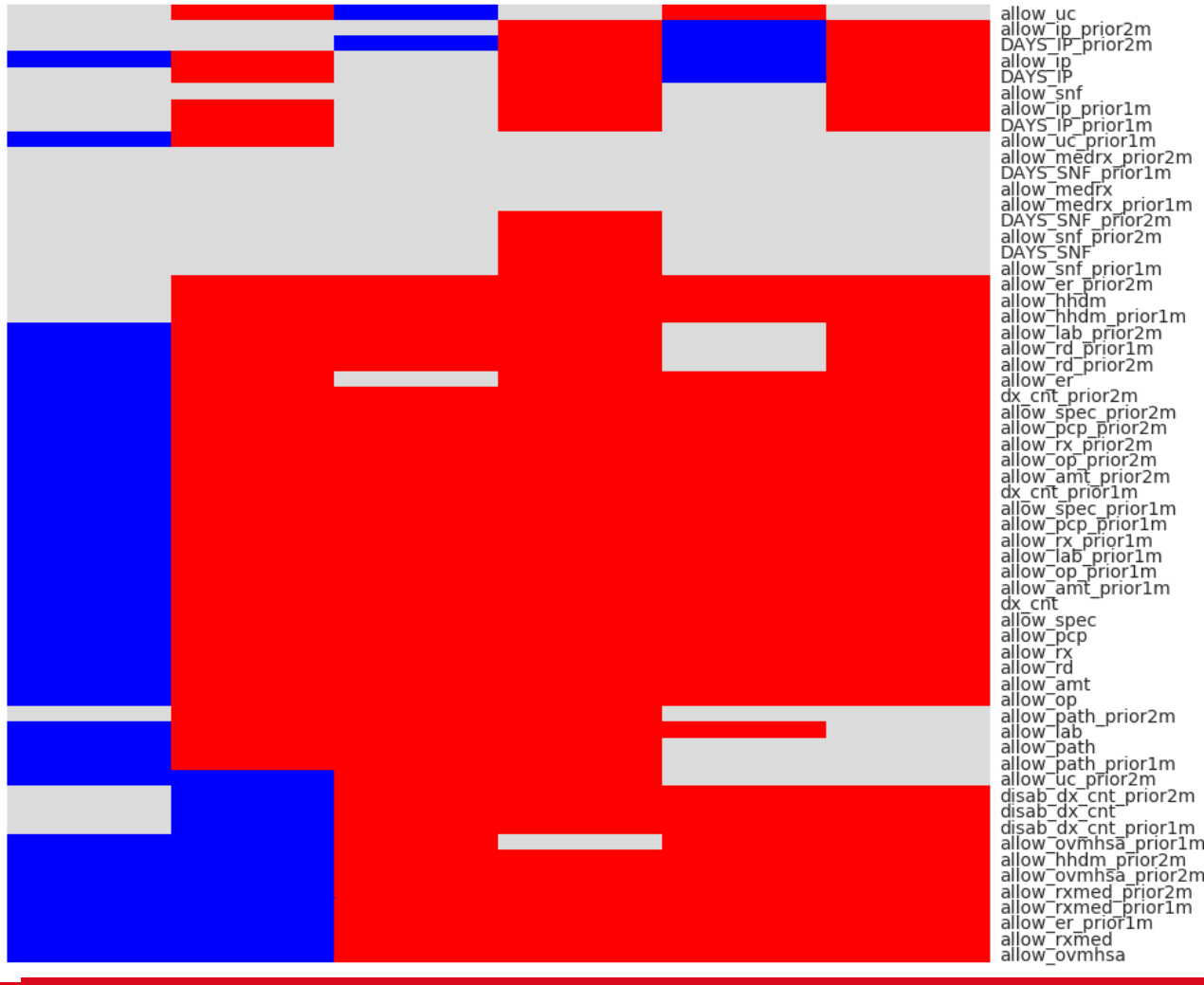
Explaining Subpopulations

Negative Significant Positive Significant

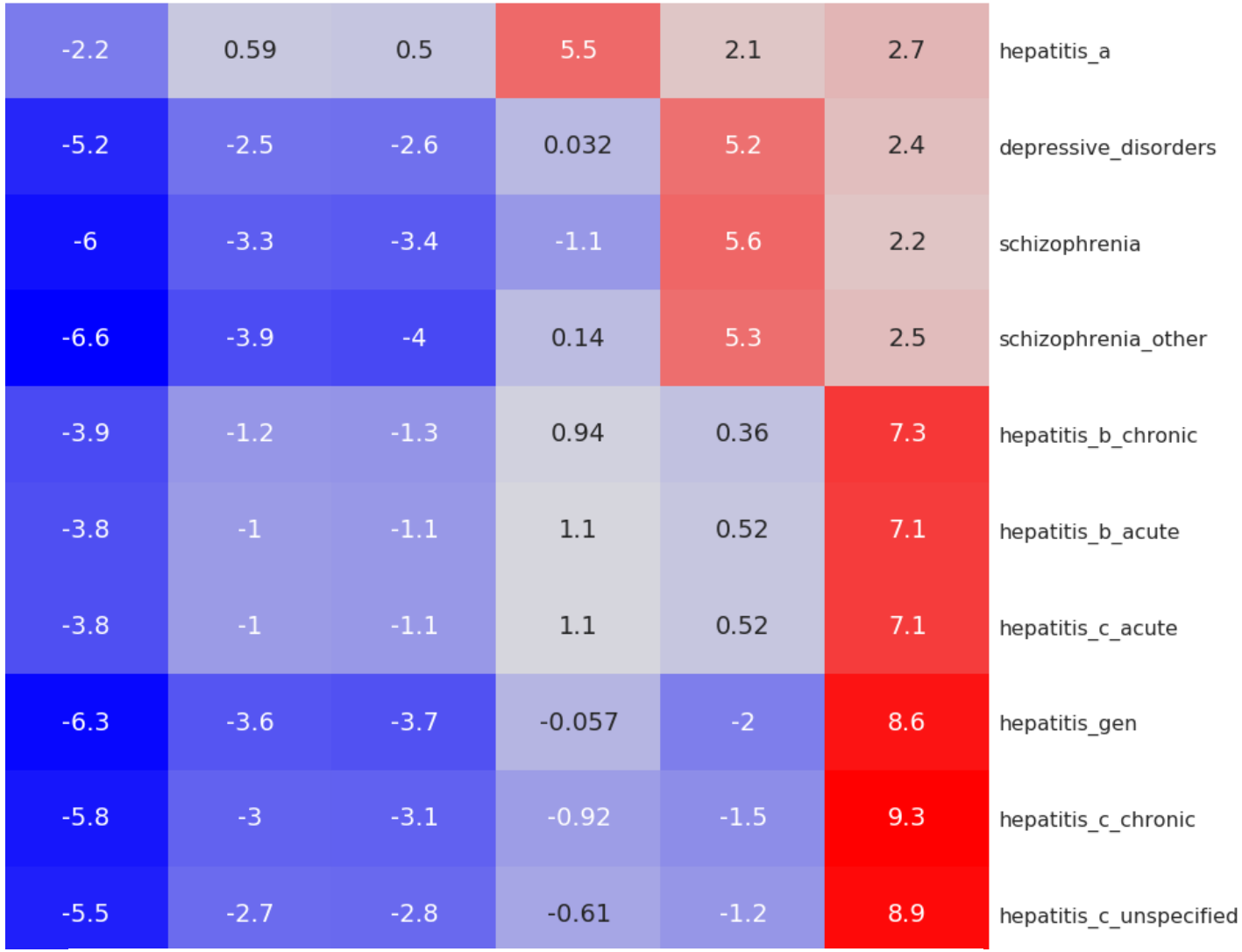
Diagnosis



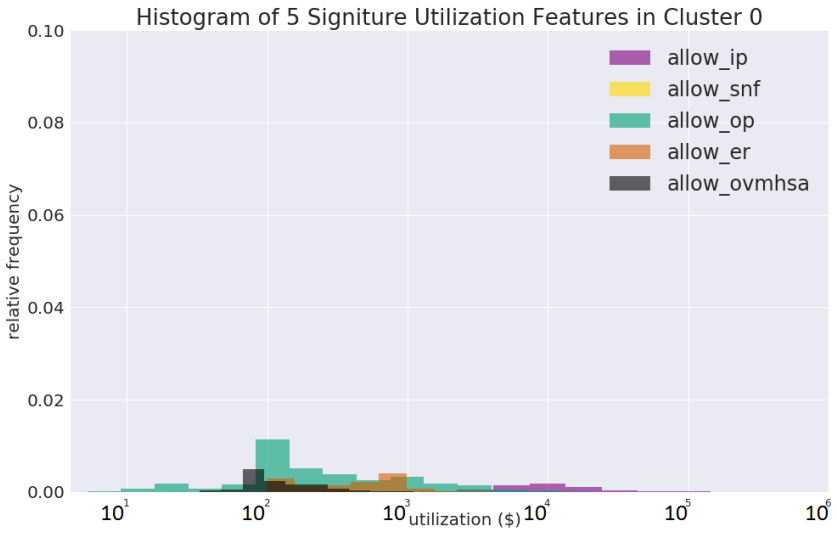
Utilization

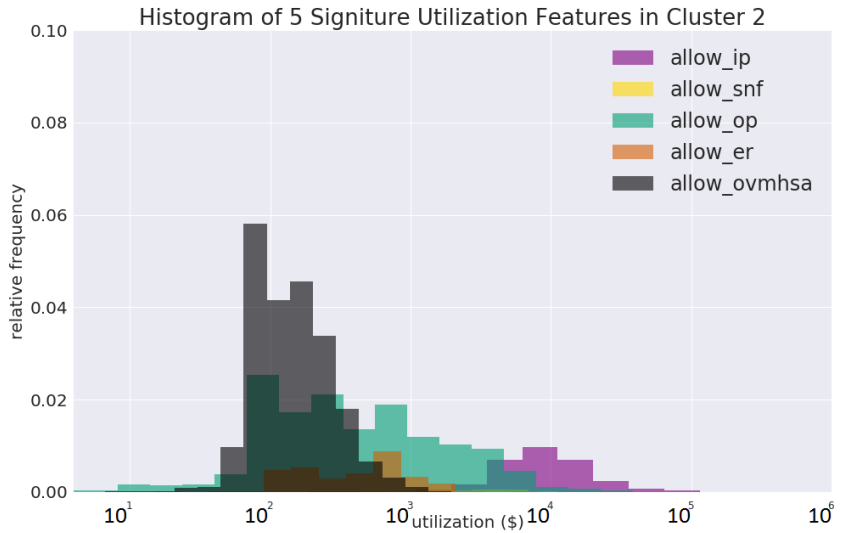


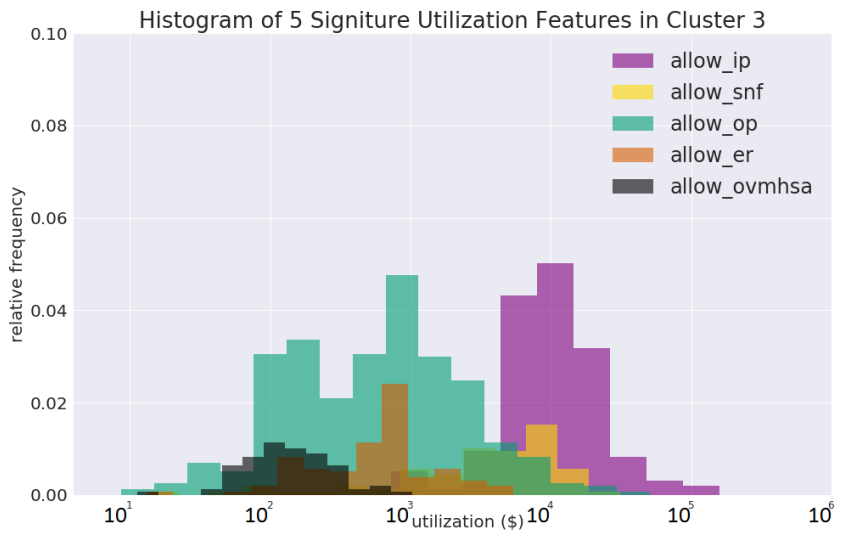
Clusters Have Distinct Diagnosis Groups



Clusters Have Distinct Place of Treatment Profiles







Low-risk cluster patients use few services

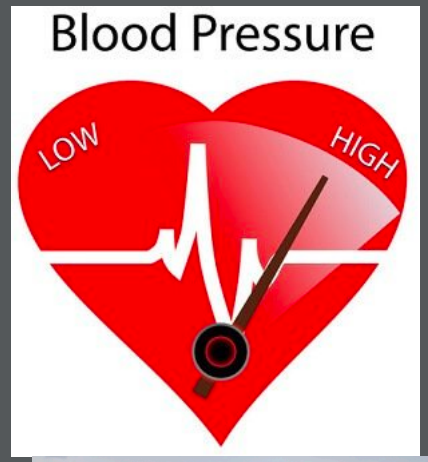
Medium-risk cluster patients with higher mental health services and low in-patient

Medium-risk cluster patients with combination of inpatient and outpatient services

Status: Performing longitudinal evaluation in preparation for deployment

Study 3: Environmental Risks of Chronic Disease

For *[given subpopulation]* in [*data source]*, does *[risk factor]* have a significant association with *[chronic health condition]*?



Motivation

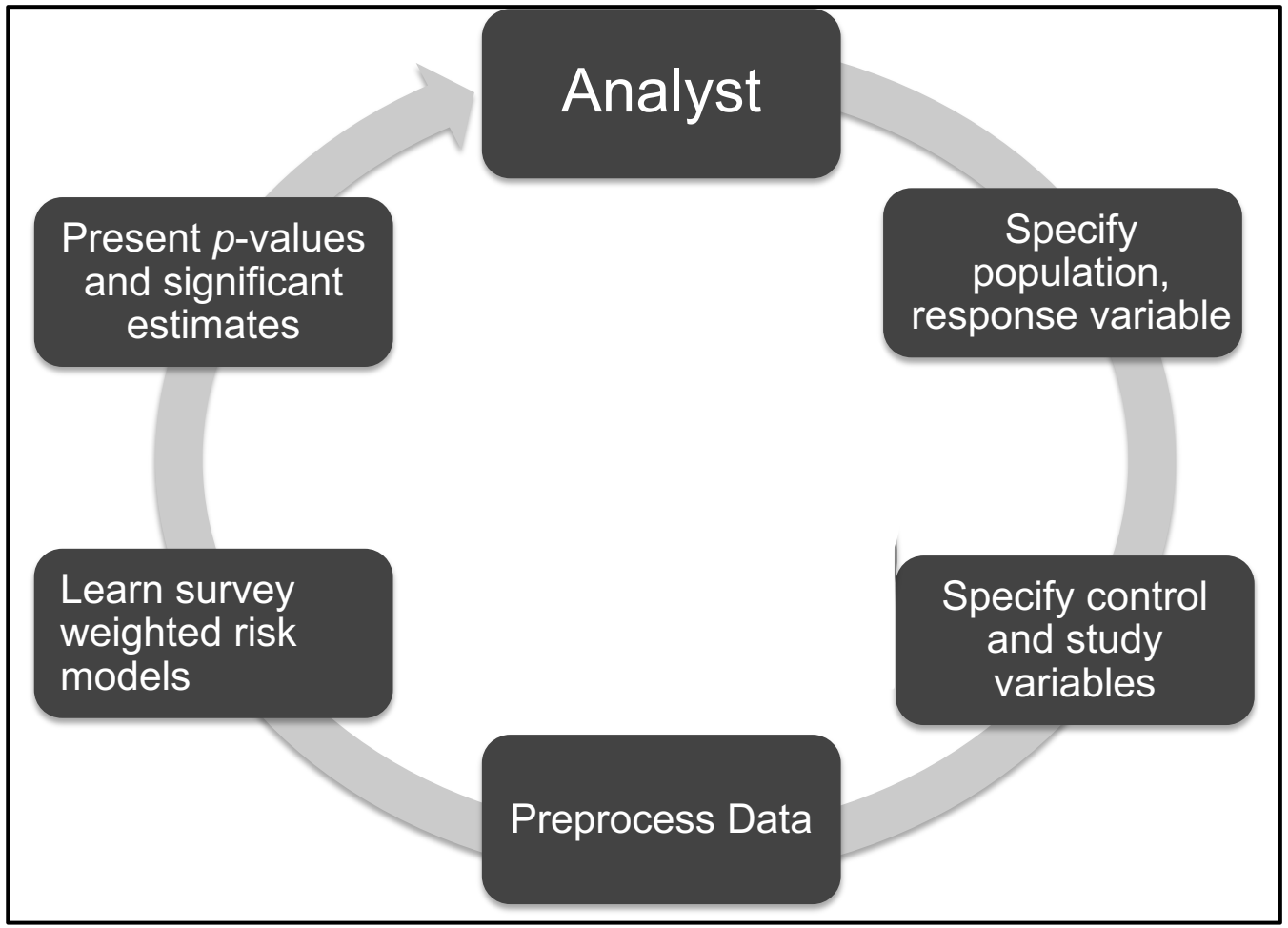
- **Continuous National Health and Nutrition Examination Survey (NHANES)**
- Surveys ~10,000 people every two years from 1999 to present
- Demographics, Dietary, Examination, Laboratory, Questionnaire
- Can look for associations between environmental exposures, lifestyle habits, harmful conditions …

Over 6000 populations health association studies performed on NHANES Datasets using "multivariate logistic regression" to determine risk factors for various conditions and subpopulations

Motivation

- A researcher is interested in some harmful condition such as High Cholesterol
- Given a data set of possible risk factors such as pesticide exposures or lifestyle habits:
  - Are these risk factors associated with the condition on a population level?
  - Are these risk factors associated with condition for some subpopulations?
- *Can we use semantics to enable understandable, scientifically-rigorous risk analysis dynamically as specified by that researcher*?
- **Risk Analyzer** - tool implementing dynamic population health risk analyses
- **Semantically Targeted Analytics** driven by Knowledge Graph and Health Analytics Ontology

Typical Association Study on NHANES



Default Analysis Goal: After controlling for known confounders, estimate association between response variable and study variables

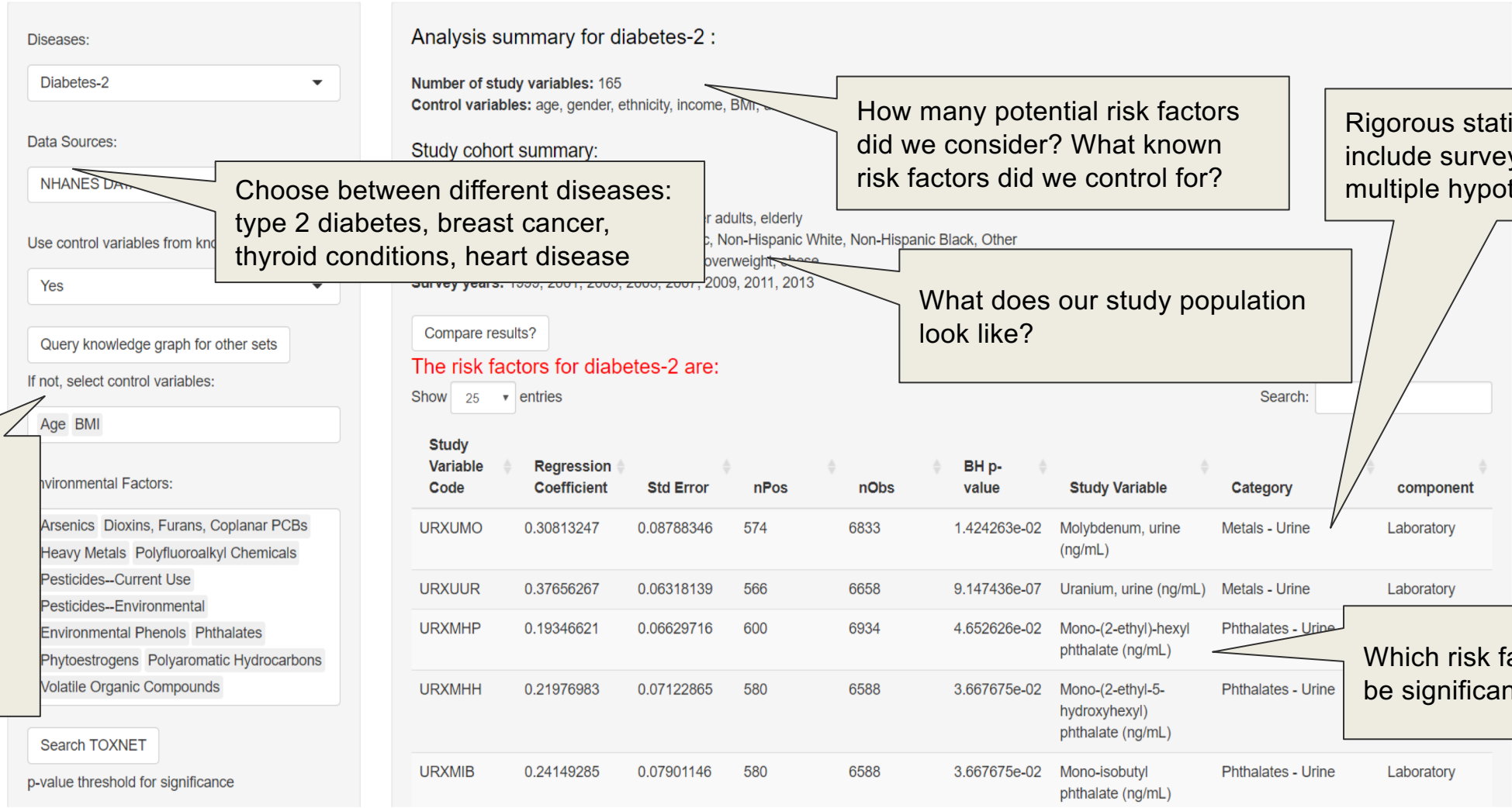
Using NHANES is complicated!

- complex survey design
- survey weighted modeling
- structural and random missingness
- changing factors over time

Applies to other surveillance datasets

Behavior Risk Factor Surveillance Survey

interactive Risk Analyzer



Choose between different diseases: type 2 diabetes, breast cancer, thyroid conditions, heart disease

How many potential risk factors did we consider? What known risk factors did we control for?

What does our study population look like?

Rigorous statistical design include survey weighting and multiple hypothesis testing

User dynamically constructs study by choosing between options

Which risk factors were found to be significant?

Explainable AI - Supervised Cadre Models For Subpopulation-discovery

- Simultaneously divides observations into subpopulations (cadres) and learns subpopulation-specific risk models
  - E.g., subjects below a threshold based on age and BMI have a significant association between blood cadmium and systolic blood pressure

$$f(x) = g(x_{F_C})^T e(x_{F_T})$$

- Risk score function (e.g., for having hypertension)

$$e^m(x_{F_T}) = (W_m)^T x_{F_T}$$

- Risk score function for cadre *m*

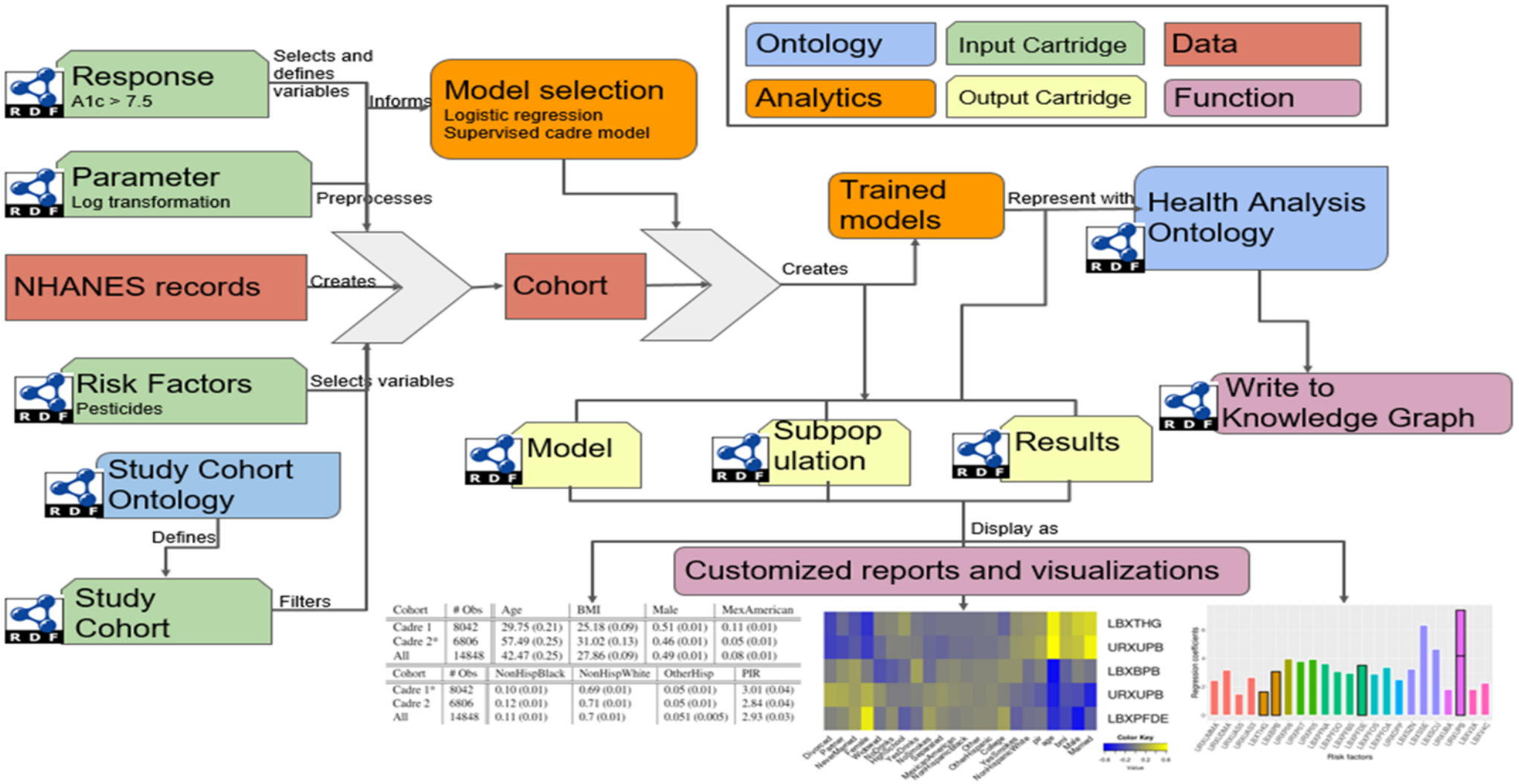
$$g_m(x_{F_C}) = \frac{e^{-\gamma||x_{F_C} - c^m||_d^2}}{\sum_{m'} e^{-\gamma||x_{F_C} - c^{m'}||_d^2}}$$

- Probability that observation *x* belongs to cadre *m*

$$||z||_d = \left(\sum_p |d_p|(z_p)^2\right)^{1/2}$$

- Semimetric used for cadre-assignment

Semantically Targeted Analytics (STA) Captures Analytics Pipeline

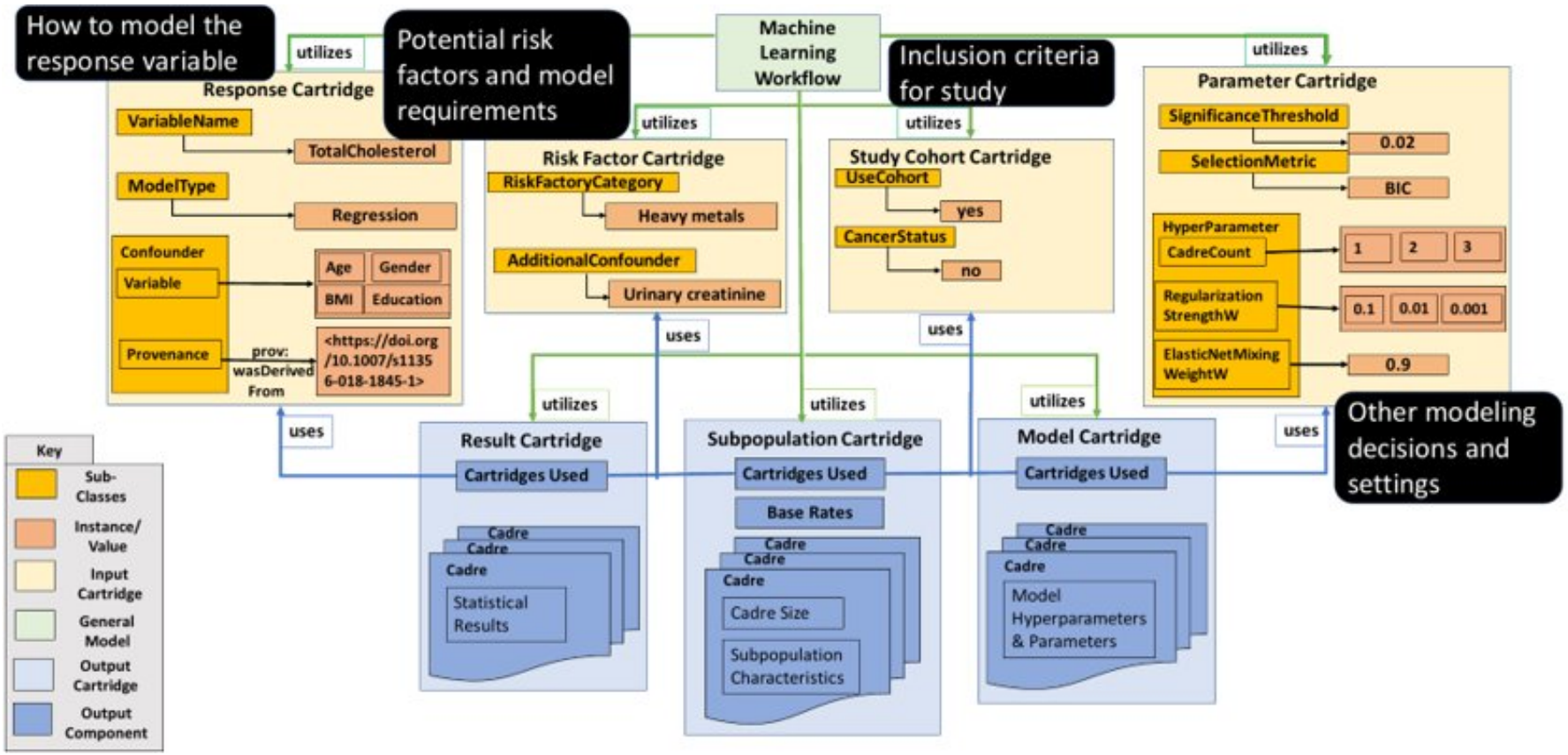


Health Analysis Ontology (HAO)

- It supports modeling of processes, components, models, variables and factors involved in a health analysis pipeline
- It provides a vocabulary necessary to model the reusable components of an analysis (sio:Analysis) implemented by an analysis workflow (hao:AnalysisWorkflow) that we store in cartridges (hao:Cartridge).
- Ontologies currently used in STA

Ontology
Health Analysis Ontology
Study Cohort Ontology
Children's Health Exposure Analysis Resource
The Statistical Methods Ontology
Semanticscience Integrated Ontology
National Cancer Institute Thesaurus
Ontology for Biomedical Investigations
The PROV Ontology
Ontology of Biological and Clinical Statistics
DC Terms
Simple Knowledge Organization System

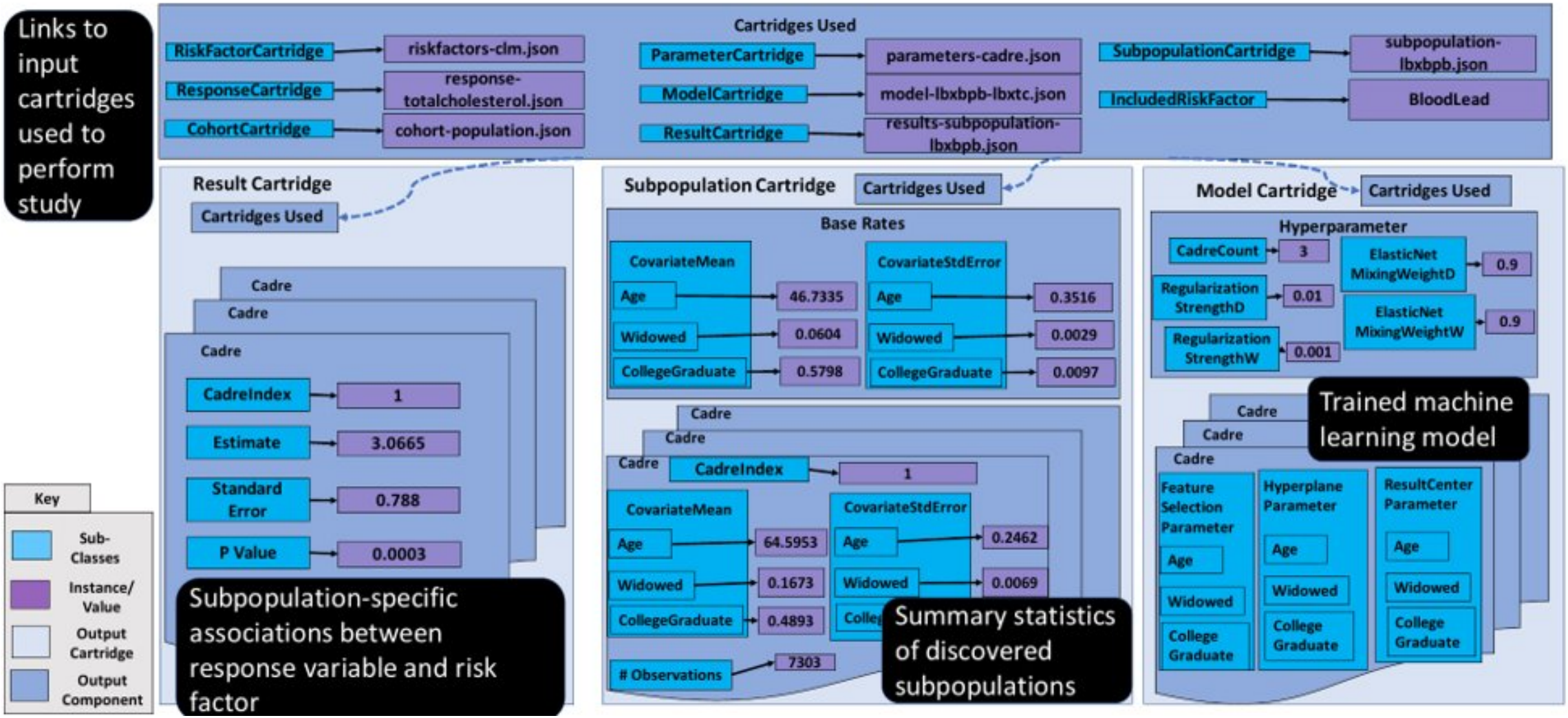
Input Cartridges (Yellow): Define Components Of A Risk Study



Cartridges encode best practices for both analytics modeling and specific domains

This allows rigorous studies to be dynamically constructed, represented, interpreted, and reproduced.

Output Cartridges (Light Blue) Store Findings Uncovered by Machine Learning



Example: Identify Risk Factors Associated With High Total Cholesterol

Response
Total cholesterol is a continuous response variable.

Response
Control for subjects' age, Body Mass Index (BMI), Poverty Income Ratio (PIR),smoking habits, drinking habits, gender, marital status, and education level.

Study cohort
NHANES 1999-2016 excluding cancer

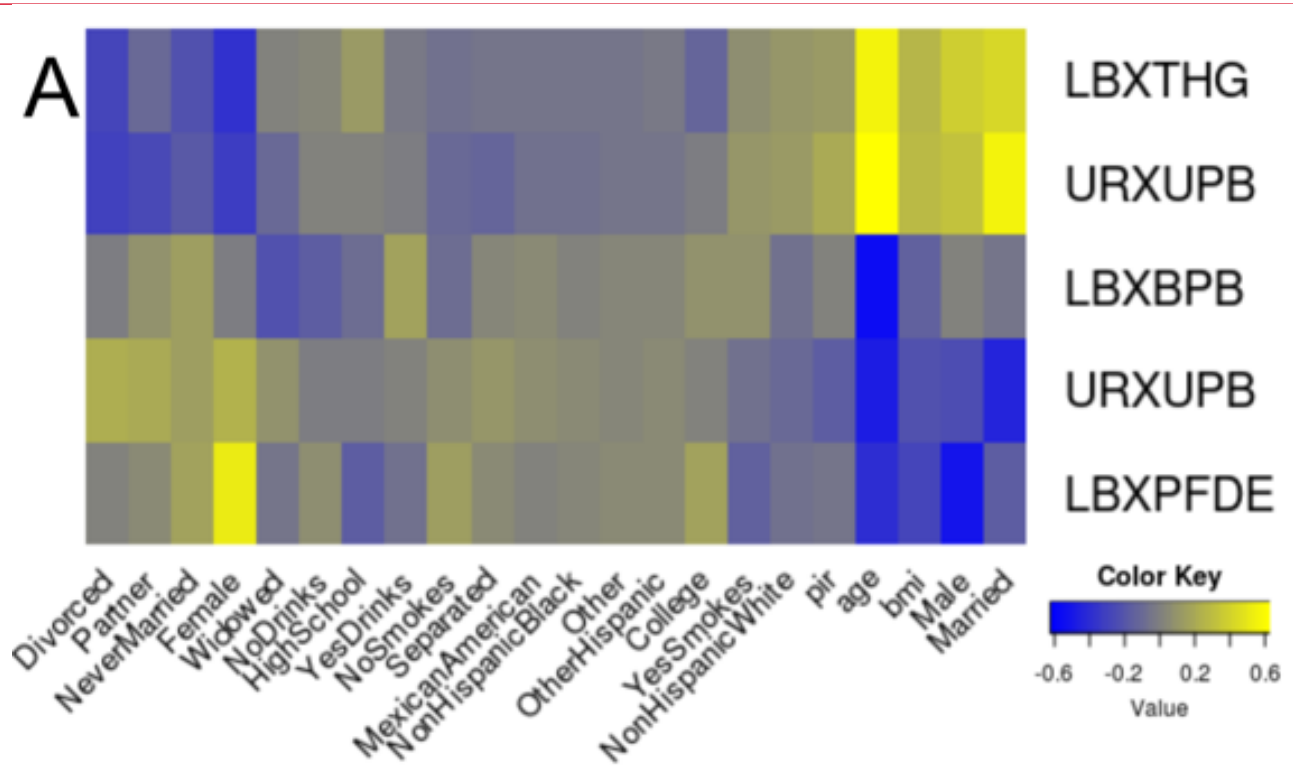
Risk Factor
201 environmental exposure risk factors divided into 17 categories

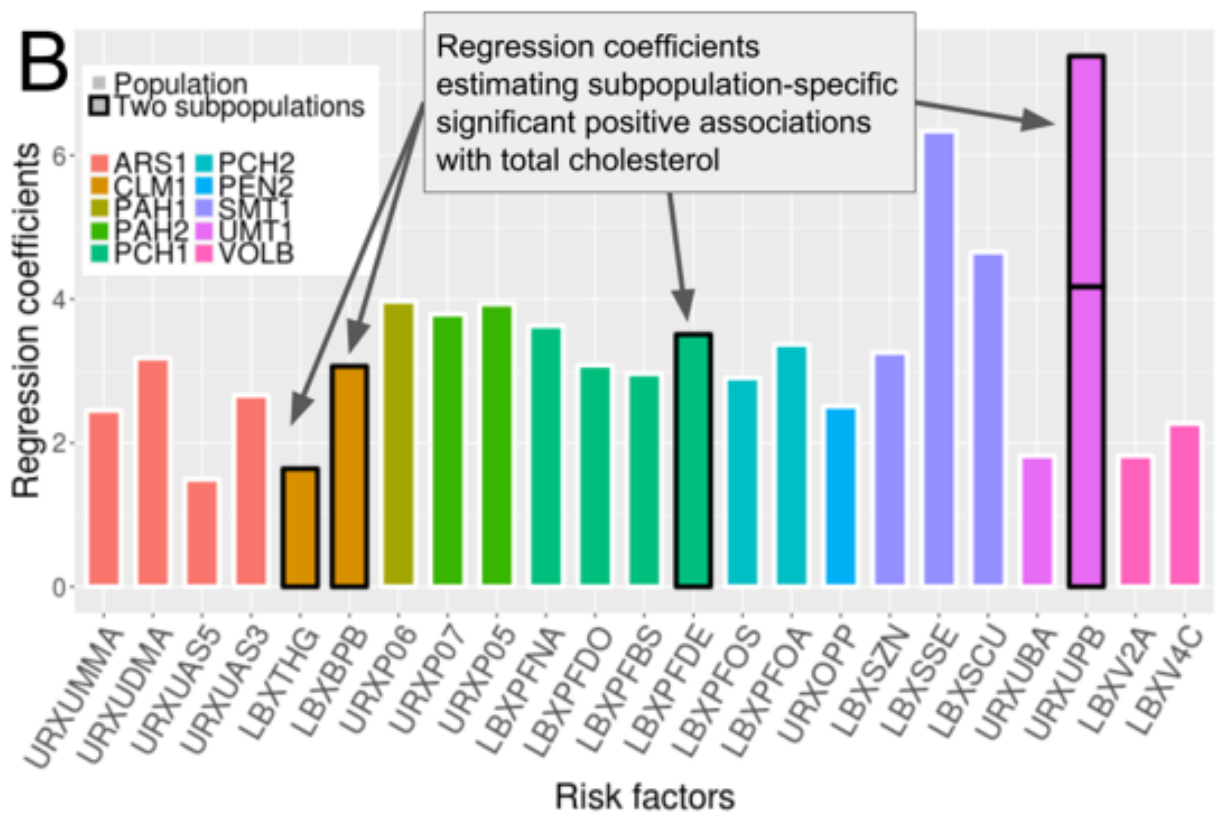
Parameter
Train models with M = 1, 2 and 3 cadres and choose best one using BIC for model selection

Parameter
Standardize risk factor measurements

Parameter
Significance threshold of $\alpha = 0.01$ in multiple hypothesis tests

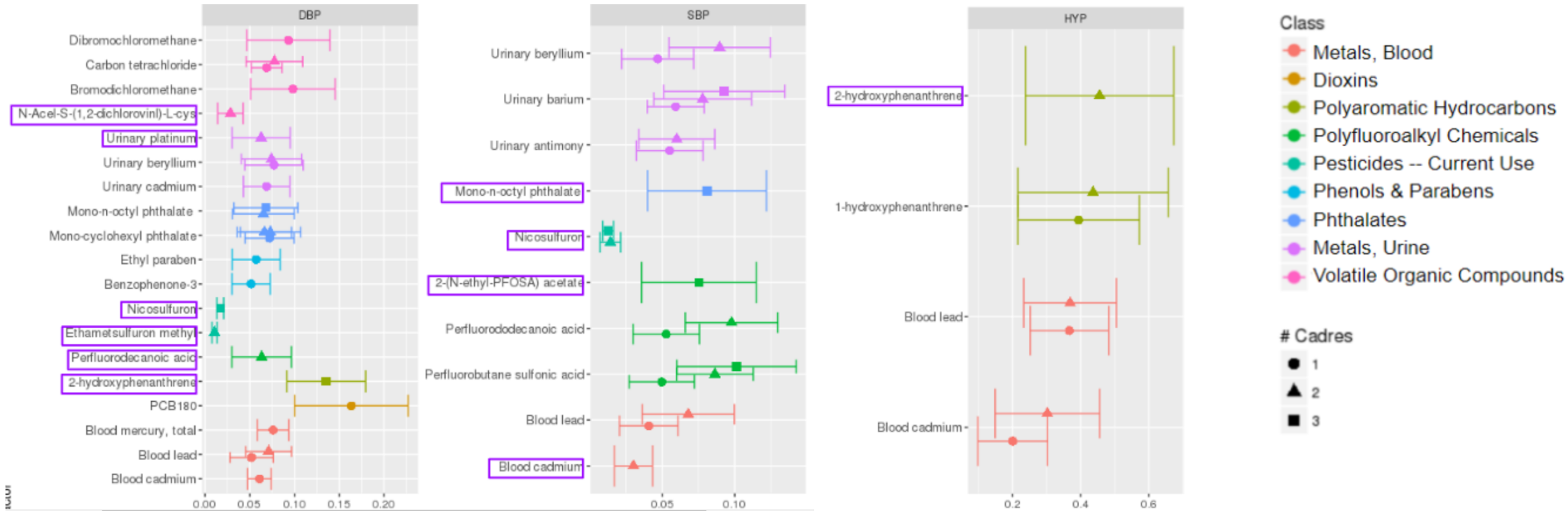
Identify Risk Factors Associated With High Total Cholesterol





- Heatmap of subpopulation means that have significant risk factor associated with high total cholesterol
- Significant positive regression coefficients associated with high total cholesterol

Identify Risk Factors Associated With Hypertension



- Of 218 potential risk factors, 25 had a significant positive association with at least one response variable (α = 0.02, Benjamini-Hochberg FDR correction)
- Eleven significant positive associations are discovered because subpopulation-specific models (the SCM) were used

Improving Population Health using AI

- Exploit Surveillance and/or EMR data
- Think beyond predictive accuracy
- Create actionable insights through Explainable AI
- Jointly discover disparate subpopulations and predictive models
- Use AI/semantics to do automated rigorous analysis with domain knowledge
- Put dynamic analyses apps in the hands of researchers and decision makers

Thank You! Questions?

Kristin P. Bennett, bennek@rpi.edu

Alexander New, Miao Qi, Shruthi Chari, Sabbir M. Rashid, Oshani Seneviratne, James P. McCusker, John S. Erickson, Deborah L. McGuinness, Xiao Shou, Georgios Mavroudeas, Kofi Arhin, Jason N. Kuruzovich, Malik Magdon-Ismail,

John Erickson,

Students in Data INCITE Lab

This work done with support of United Health Foundation®

Prediction Test Accuracy

Supervised Gaussian Mixture Model (SGMM) is interpretable and predicts well
Neural Network higher accuracy, but is not explainable and yields worse prediction in larger longitudinal analysis

Methods	Recall	Precision	AUC
Log Reg (dx_cnt)	**0.37**	**0.23**	**0.79**
L1 Log Reg	0.42	0.42	0.83
Neural Net	**0.77**	**0.54**	**0.95**
Random Forest (RF)	0.48	0.53	0.91
Adaboost	0.47	0.32	0.86
GradBoost	0.47	0.59	0.91
SGMM	**0.45**	**0.50**	**0.86**
SGMM + RF	**0.45**	**0.52**	**0.90**