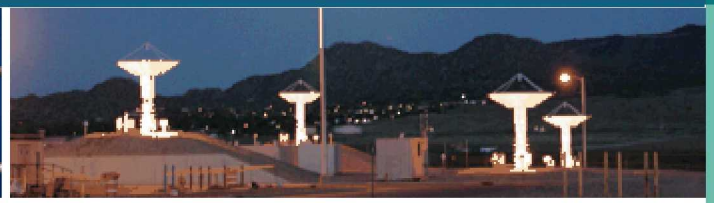
SAND2020-10206PE

COVID-19 Response Machine Learning Sub-Task May Progress Report



Carianne Martinez, Jessica Jones, Drew Levin

Sandia COVID-19 Machine Learning Team

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Disclaimer

Funding Acknowledgement

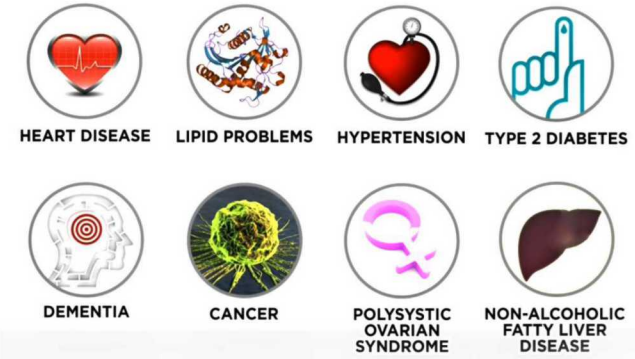
Research was supported by the DOE Office of Science through the National Virtual Biotechnology Laboratory, a consortium of DOE national laboratories focused on response to COVID-19, with funding provided by the Coronavirus CARES Act.

Disclaimer

Questions We Will Answer

How do comorbidities affect infection severity?

- Our county-level model fits will reveal the effects of individual demographic and comorbidity features on infection outcomes.

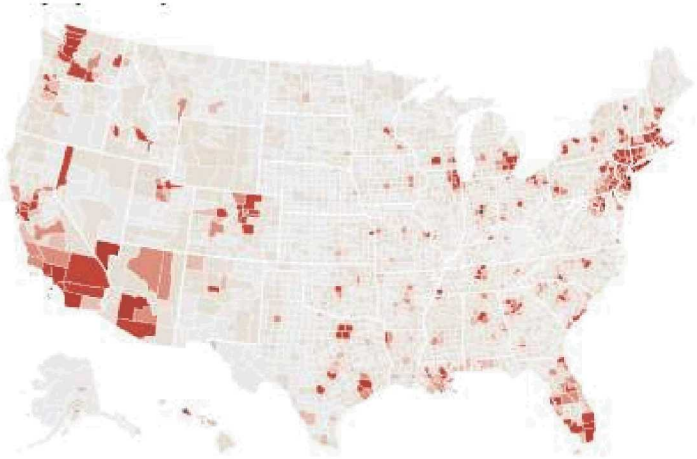


Which patients are most at risk?

- Longitudinal EHR analysis will train an improved deep CNN/RNN model to more accurately predict infection severity based on a patient's full medical history.
- We will produce an interpretable model based on our findings for use in clinical settings.

How will disease progression differ by US county?

- Local county-level models will be extensible to all US counties as broad demographic and comorbidity prevalence data is available.
- Nuanced effects can then be incorporated into our epidemiological models.



Phase One - Project Overview

Machine learning and data science applied to EHR medical data to improve risk prediction for severe COVID-19 symptoms.

Project Tasks:

- Data Source Identification and Formatting
- Examination of Data Types
- Pursue External Data Sources
- Assist other Sub-Tasks
- Analyses and Early Results

Data Source Identification

Disease severity prediction requires patients' medical records, yet there exists no single source of high-quality medical data.

Public data sources come in different formats with different fields and biases.

We examined multiple datasets at the country, province/state, and county level.

Most datasets were not useful, but we were able to identify several that showed promise.

Identified Foreign Country Datasets

China

Hong Kong

Singapore

- Aggregated together
- Early patient records from Wuhan – ages, dates
- Some patient notes, most empty
- Outcomes unreliable and often empty

South Korea

- Ages, dates, locations
- Shows who infected who (very cool)
- No medical outcomes
- Aggregated together – decent consistency
- Shows symptoms for some, but not many

Canada

- Each patient, age, location, date
- No outcomes

Italy

- In Italian!
- Province level data
- No individual outcomes\

Mexico

- Patient ages and location
- No outcomes

Other Countries

- Mostly just provincial counts

Very little useful to predict outcomes!

Domestic Data Sources – Country Level

Primary

John Hopkins Repository
- Our source of choice
- County level positives and deaths

NYTimes
- Also good
- Avoids a few pitfalls of the JH repo

Covid Tracking Project
- State level
- Contains Hospital and ICU
- Shows % positive tests by State
- Contains links to State websites

IHME
- Well known source of projections

Other

Covid Projections
- Estimation of Ro by State
- Estimation of latent total infection by State

Lab Testing Dashboard
- Shows % positive tests by county
- Requires account (.gov allowed)
- Available through MITRE collab.

Covid Projects Dashboard
- List of open-source data analysis projects.
- Good for brainstorming and scanning current methodology

Domestic Data Sources – State Level

New Mexico
- Best source of data
- Full patient records with a few gaps
- Coarse comorbidities, race/ethnicity
- NM has low #s – not much to go on

California
- Hospital & ICU by county
- Some counties also report hospital capacity info
- No patient level data

Ohio
- Again hospital and ICU counts by county
- Again, no patient data
- Reached Out via email for more demographic info – no response

Florida
- Some hospitalization info, not great

New York / New Jersey
- Surprisingly little given the severity

NYC
- Much better than NY State
- Zip code level
- No outcomes

Oregon
- ICU data at state level

Georgia
- Individual patient list with outcome
- Only shows age of each patient and yes/no comorbidity
- (Currently Offline)

Domestic Data Sources - Demographics

Due to lack of patient outcome data, we attempted to secure demographic information by US county as a proxy for patient-level comorbidity values.

Demographic data from 2018 American Community Survey, U.S. Census Bureau

Health data

- 2018 California Health Interview Survey, UCLA Center for Health Policy Research
- 2018 Behavioral Risk Factor Surveillance System Survey, NM DOH

Social Vulnerability Index (community resilience measure), CDC

Nursing home population by county

- Nursing Home Compare by Centers for Medicare & Medicaid Services

Walgreens Prescription Fills Risk Score

IHME Global Health Data Exchange – US county

- Health Care Spending
- Infectious Disease Mortality Rates
- Substance Use Disorders
- Chronic Respiratory Disease
- Cardiovascular Disease
- Life Expectancy
- Cancer Mortality Rates
- Diabetes Prevalence
- Smoking
- Hypertension
- Obesity

External Collaboration to Obtain Medical Data

MITRE
- Joined the Private Health Care Coalition
- Bi-Weekly Meetings / Seminars
- No public data sources, many private data analyses shielded behind DUAs

UNM
- CRADA nearing completion
- Have access to two full medical data sets, each approximately 5 Terabytes in size
  - CERNER (Hospital Records)
  - IBM Watson (Insurance Claim Data)
- Also have real-time access to UNMH system with IRB approval

Department of Veteran's Affairs
- Access to full VA medical data, pending IRB approval to expand DOE/VA research to COVID
- Does have COVID cases
- Drew has access
- Cari and Jessica have gone through DUA training, access pending

Assist Other Sub-Tasks

Walt – Modeling Task
- Fit simple models to estimate length of hospital and ICU stay
- Assisted team in model design and parameterization

Vanessa – Economics Task
- Provided references to relevant data sources as requested

Jaideep – Projections
- Answered questions as needed
- Provided article references with desired parameter values
- Assisted Erin with data access

Analyses and Early Results

Analyses Attempted:

Risk by Age

Linear Regression

Logistic Regression

Length of Stay

Statistical

PyMC3 Bayesian

- Naïve Bayesian
- Hierarchical (looked into)

Symptoms -> Outcomes

Data formatting

Outcome Examination

Demographic/Co-Morbidity Risk

Marginal -> Joint distribution methods

CA county demographic -> hospitalization outcomes

NM linelist analysis

Fit to mortality and positive test slopes

Exemplar - Regression to Parameterize Risk By Age

Data taken from public sources:

- China
- South Korea
- Hong Kong
- United States

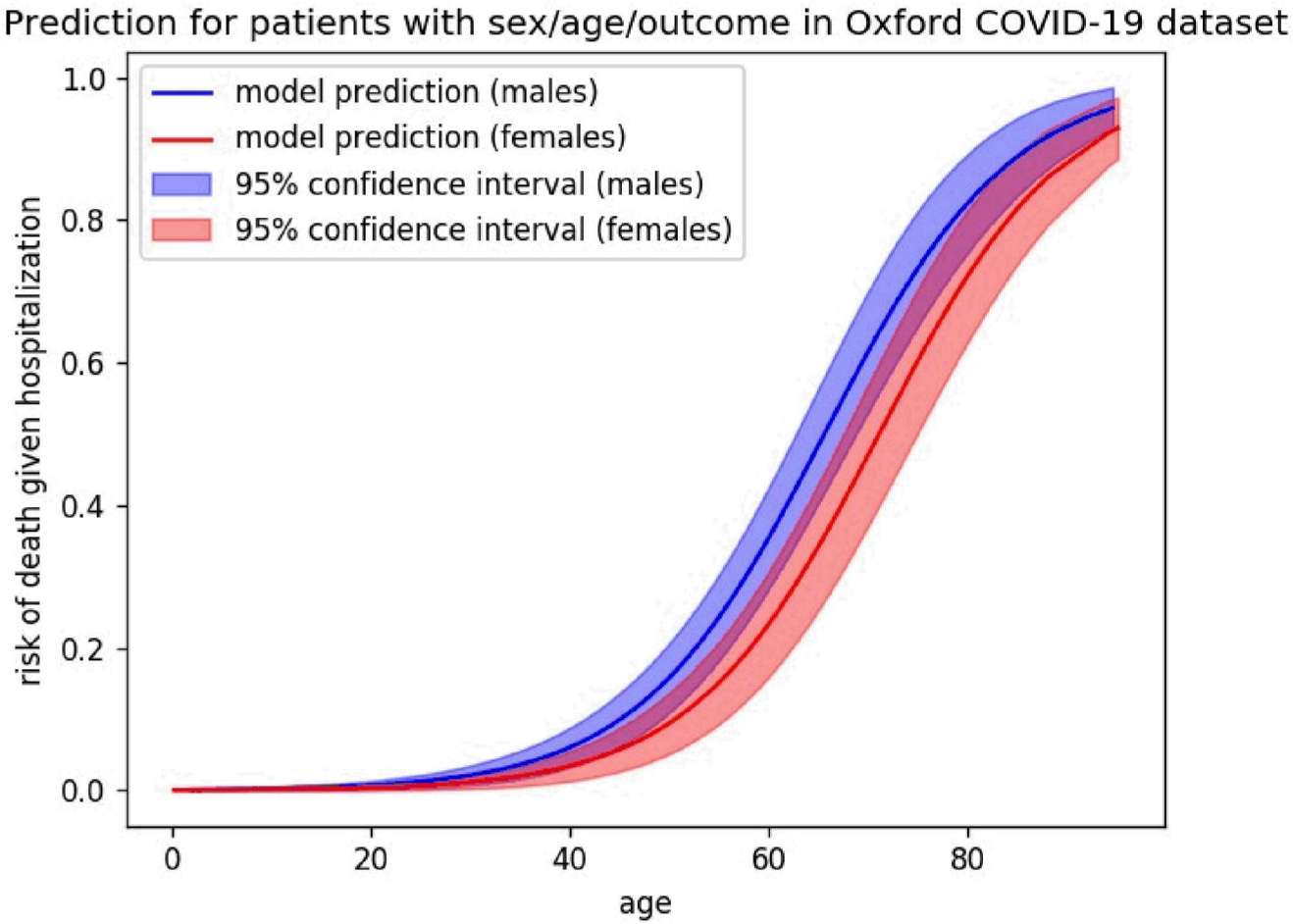
Logistic Regression Fit

- Reveals risk curves with uncertainty
- Also provides model parameters for use in other projections

Can be extended to explore comorbidity risk with the right data

Mortality risk increases with age

Men at higher risk than women



Hospital Length of Stay Analysis

Hospital Length of Stay

ICU length of Stay

- The two most sensitive parameters in project models

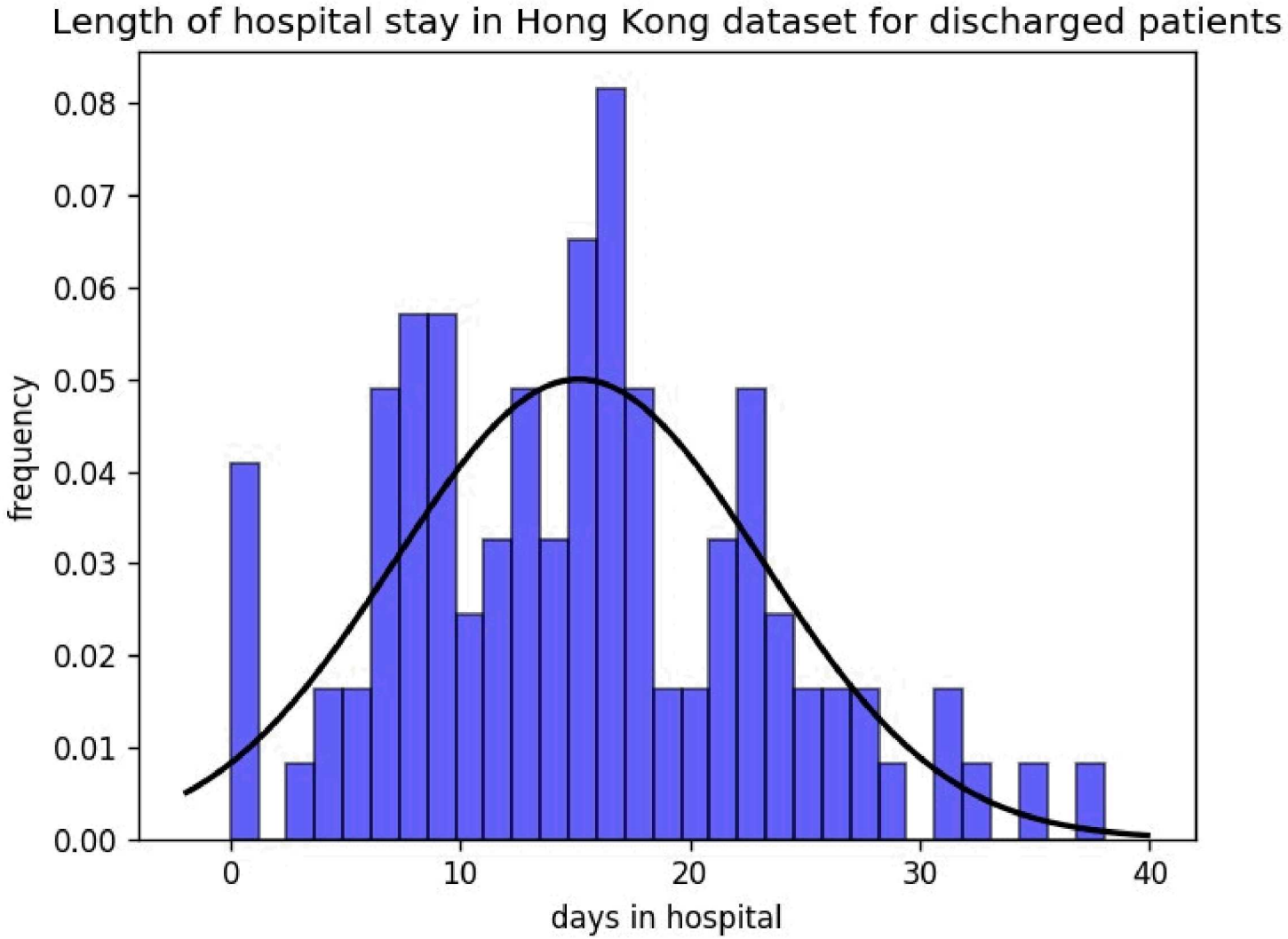
Use publicly available data

- China / South Korea / HK
- United States

Statistically fit to distributions

Example - Gaussian Fit

- Hospital Length of Stay
- **15.1 days +/- 8.0 std**



Bayesian Models of Hospital Trajectories

Hospital and ICU length of stay are ***conditional*** on many factors:

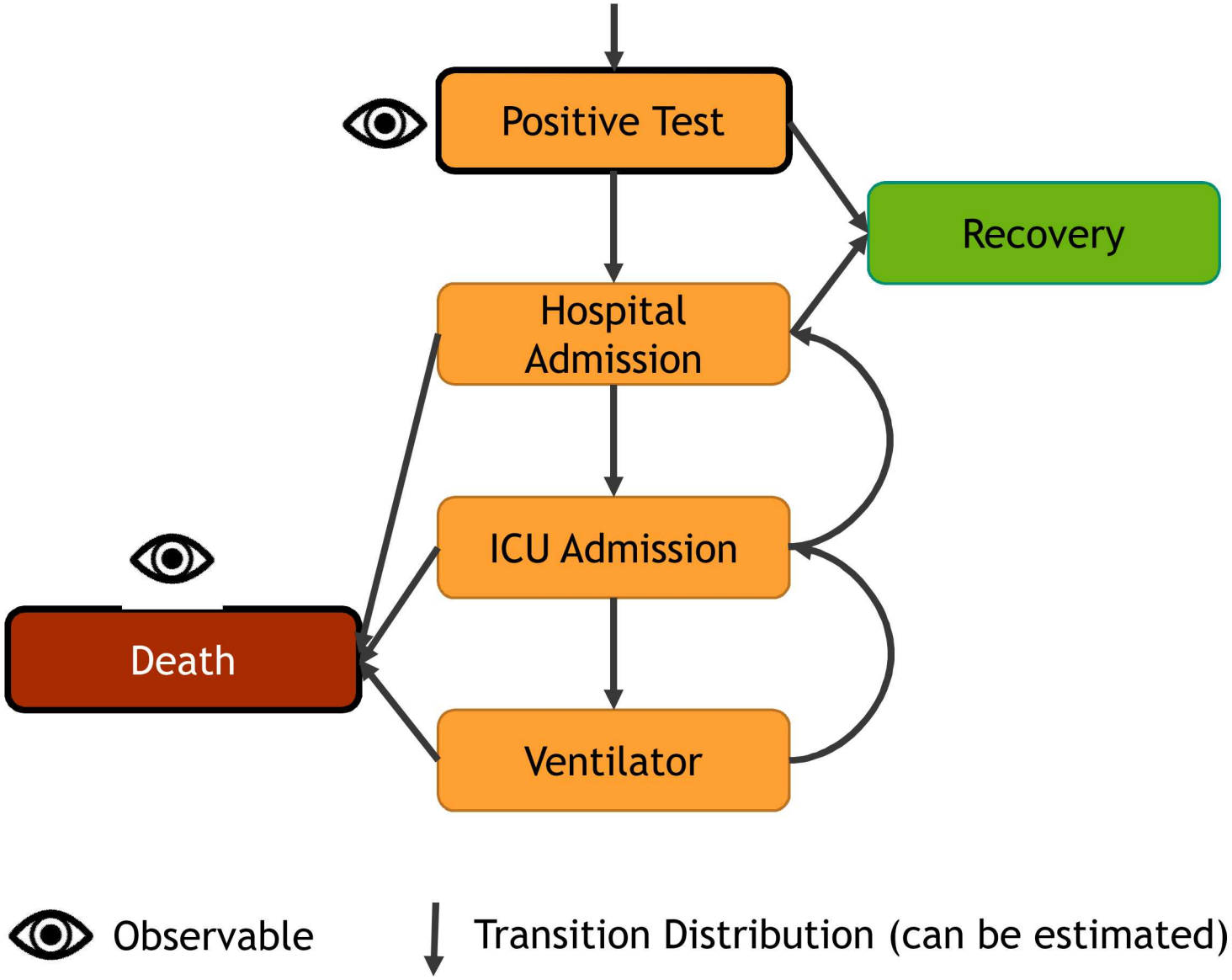
- Patient demographics
- Comorbidities
- Reported Symptoms
- Disease Severity

Further, length of stay distributions are not necessarily symmetrical Gaussians.

How do we estimate length of stay distributions from datasets with no explicit length of stay data?

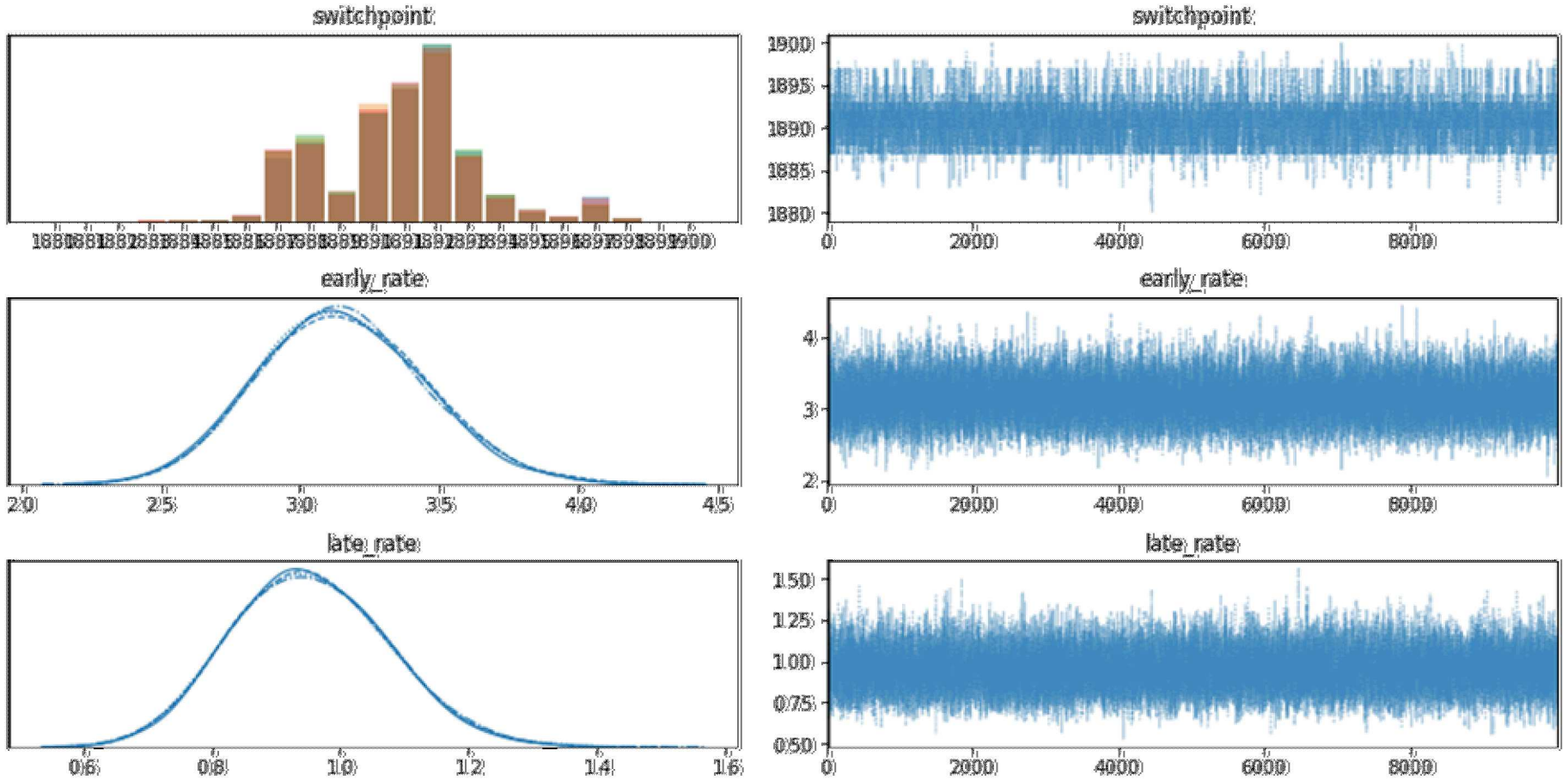
Bayesian maximum likelihood parameter estimation of a compartmental model of infection

Find parameterization that maximizes likelihood of observable values.



Bayesian Models of Hospital Trajectories

Maximum Likelihood Estimation using Python and Probabilistic Programming (PyMC3)



Simple example: Model fits give full distribution over estimated parameters.

Unfortunately, we still lack the necessary data to perform this type of analysis.

Reported Symptoms as a Predictor of Outcome

Many patients report specific symptoms upon hospital admission:

- Fever
- Soreness
- Cough
- Vertigo
- Nausea
- Diarrhea
- Shortness of Breath
- Chest Pains
- Congestion
- Headache
- And many more (300 unique values)

Approach

1) Categorize free text reported symptom field to bin symptoms into categories

2) Fit model predicting outcome severity given symptom types

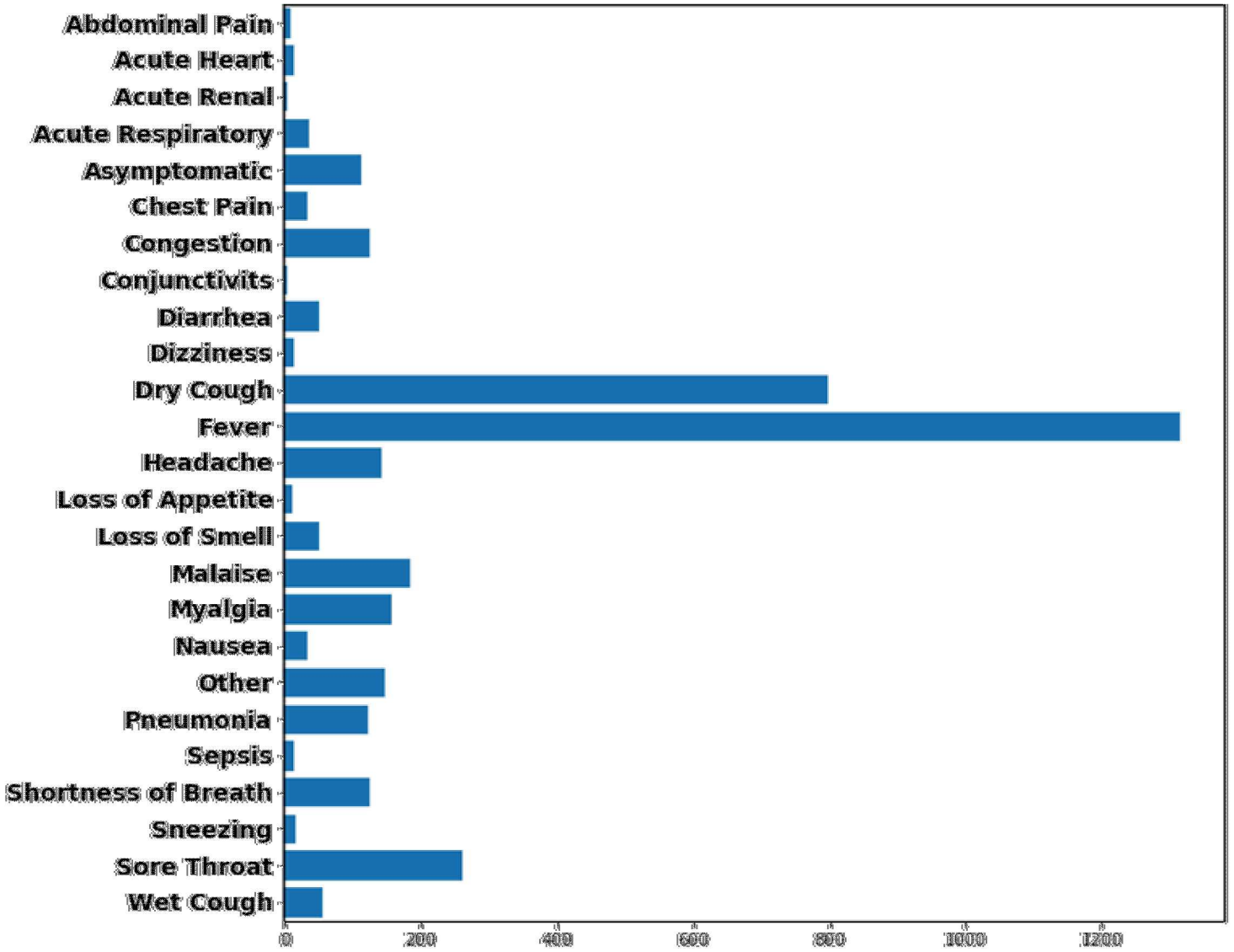
3) Model parameterization reveals symptom effect

3,843 Listed Symptoms

300 Unique Values

25 Symptom Categories

Distribution of Reported Symptoms



One of the earliest data points we have!

3,843 Listed Symptoms

300 Unique Values

25 Symptom Categories

Only *103* Patients with both symptoms and outcomes.

A larger sample would allow us to use symptoms to predict severity!

New Mexico / California Demographic and Comorbidity Risk

Problem: We don't have sufficient patient data with both comorbidities and medical outcomes.

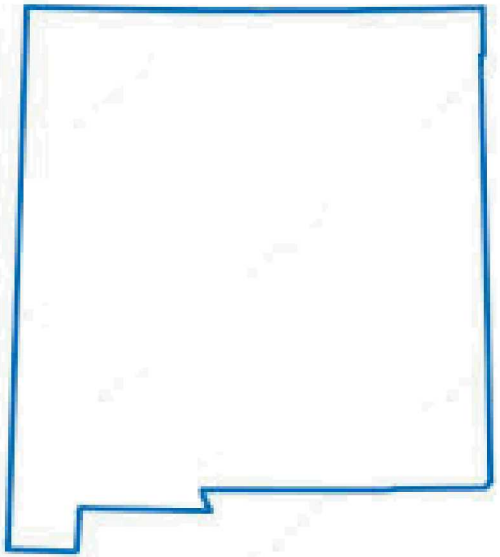
Solution: Estimate both by aggregating results at the US County level.

Approach

1. Obtain the full set of marginal demographic and comorbidity distributions we intend to use.
   - Must be by US county
   - Demographics: Age, sex, race/ethnicity, SES, substance abuse
   - Comorbidities: County-level prevalence of diabetes, smoking, hypertension, etc.
2. Find any sources that may suggest joint distributions so we can build a model to produce expected joint distributions
3. Find all states with good hospitalization and ICU data.
4. Train the full model on the states that have data.
5. Examine comorbidity effects.
6. Use the model to project outcomes for states and counties that don't have hospitalization data.

New Mexico and California Report Hospitalization by County

New Mexico



Full list of patients

Includes

- Location
- Hospitalization Dates
- Outcomes
- Demographics
- Comorbidities

California

County-level aggregate counts only

- Total testing positive
- In Hospital
- In ICU
- Deaths

No patient outcomes

Marginals to Joint Distributions

County-level data obtained by sources listed earlier (and next slide)

Most data describes prevalence of a single type (marginal distribution)

We require full joint distributions to best fit our models.

It is naïve to assume we can multiply the marginals through to obtain the joint distribution (distributions not IID).

For now we rely on sources that provide joint distributions. Example ->

Controlling for				
		Gender (SEX1)	=	Female
		Demographic Information	=	White only, Non-Hispanic

Sample Size
Row % (95% Confidence Interval)
Column % (95% Confidence Interval)
Total % (Weighted) (95% Confidence Interval)

Calculated variable for 6-level imputed age category (_AGE_G, 18-24, 25-34,35-44,45-54,55-64,65+)

Ever told you have Chronic Obstructive Pulmonary Disease or COPD, emphysema or chronic bronchitis (CHCCOPD1)

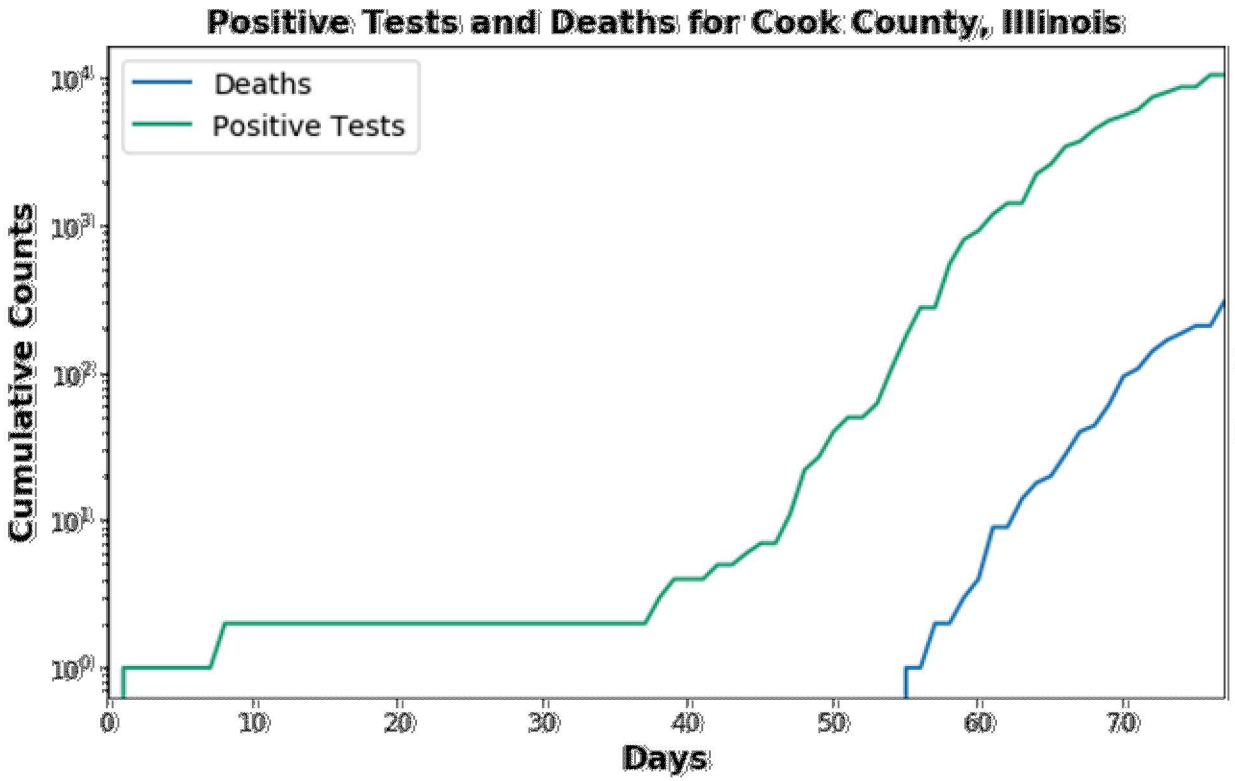
		Age 18 to 24		Age 25 to 34		Age 35 to 44		Age 45 to 54		Age 55 to 64		Age 65 Or older		Total	
Yes	n	0		1		9		16		54		122		202	
	Row%	0.0%	(0.0 - 0.0)	*		*		11.0%	(4.7 - 17.4)	26.5%	(18.5 - 34.6)	56.6%	(47.3 - 65.9)	100.0%	(100.0 - 100.0)
	Col%	0.0%	(0.0 - 0.0)	*		*		*		12.7%	(8.6 - 16.8)	14.6%	(11.1 - 18.1)	9.4%	(7.7 - 11.1)
	%	0.0%	(0.0 - 0.0)	*		*		*		2.5%	(1.6 - 3.3)	5.3%	(4.0 - 6.6)	9.4%	(7.7 - 11.1)
No	n	66		126		138		218		383		755		1,686	
	Row%	8.0%	(5.9 - 10.1)	10.5%	(8.3 - 12.7)	12.4%	(10.1 - 14.8)	15.8%	(13.2 - 18.3)	18.9%	(16.4 - 21.4)	34.4%	(31.2 - 37.5)	100.0%	(100.0 - 100.0)
	Col%	100.0%	(100.0 - 100.0)	99.9%	(99.8 - 100.0)	95.4%	(92.1 - 98.8)	93.2%	(89.2 - 97.3)	87.3%	(83.2 - 91.4)	85.4%	(81.9 - 88.9)	90.6%	(88.9 - 92.3)
	%	7.3%	(5.3 - 9.2)	9.5%	(7.5 - 11.5)	11.3%	(9.1 - 13.4)	14.3%	(12.0 - 16.6)	17.1%	(14.8 - 19.4)	31.1%	(28.3 - 34.0)	90.6%	(88.9 - 92.3)
Total	n	66		127		147		234		437		877		1,888	
	Row%	7.3%	(5.3 - 9.2)	9.5%	(7.5 - 11.5)	11.8%	(9.6 - 14.0)	15.3%	(13.0 - 17.7)	19.6%	(17.2 - 22.0)	36.4%	(33.4 - 39.4)		
	Col%	100.0%	(100.0 - 100.0)	100.0%	(100.0 - 100.0)	100.0%	(100.0 - 100.0)	100.0%	(100.0 - 100.0)	100.0%	(100.0 - 100.0)	100.0%	(100.0 - 100.0)		
	%	7.3%	(5.3 - 9.2)	9.5%	(7.5 - 11.5)	11.8%	(9.6 - 14.0)	15.3%	(13.0 - 17.7)	19.6%	(17.2 - 22.0)	36.4%	(33.4 - 39.4)		

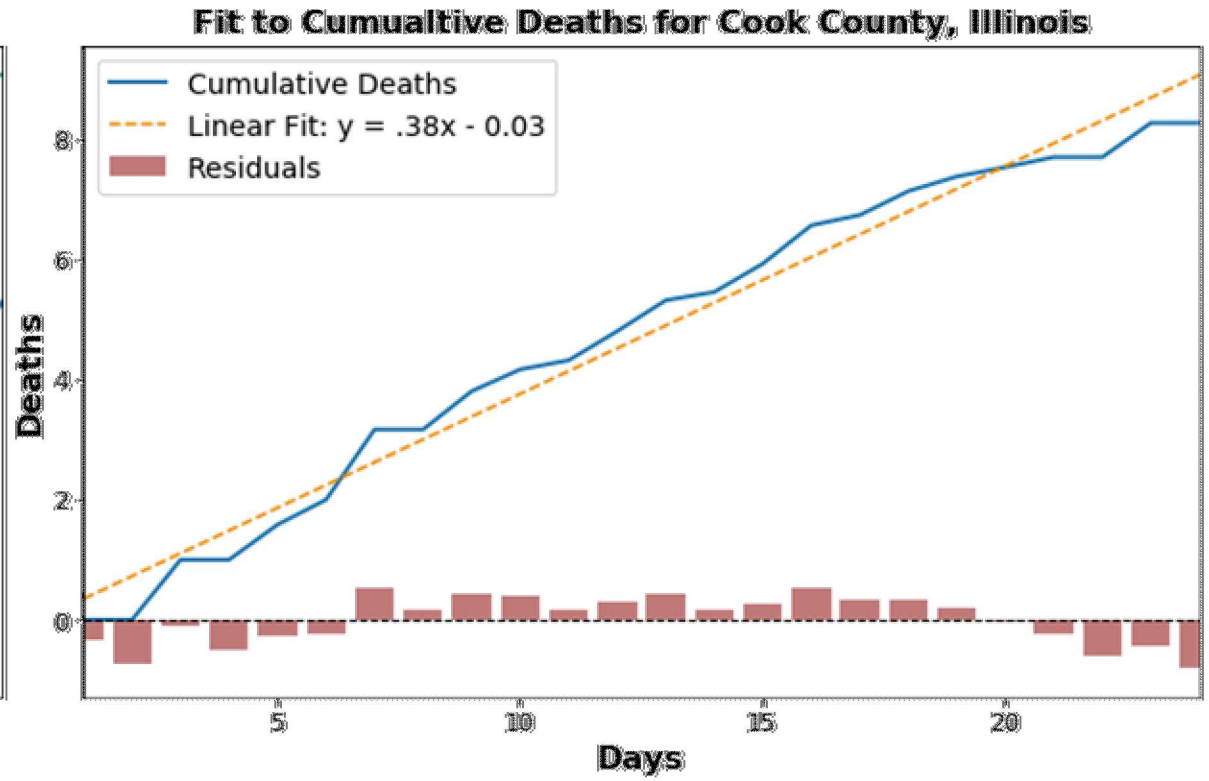
Wald Chi-Square Value	Degrees of Freedom	p-value
77.36	5	<0.0001

* Estimate not available if the unweighted sample size for the denominator was < 50 or the Relative Standard Error (RSE) is > 0.3.

County-Level Outcome Estimation

Models require something to predict. For a county-level model we require county level outcomes.





Solution: Linear regression fits to the logarithmic growth trajectories of county-level positive test results and deaths. Done for each US county

Example: Chicago doubling exponential 'Death Slope' of 0.38. Deaths double every 1/0.38, or **2.63 days**

Estimating Outcomes from Data

We are leveraging publicly available data from the State of California, the U.S Census and Johns Hopkins to develop statistical models to fit COVID-19 outcomes at a county level.

To represent each county, we generate a set of features from these various data sources, and train a regression model as follows:

- Use cross validation with all training to select best parameter (alpha) for a Lasso regression model. This model selects the most relevant features to the prediction and combines highly correlated features to drastically reduce the number of features used in the final model.
- Using the optimal alpha value, we re-train the Lasso regression model with 5-fold cross validation again using all training data to get statistics around the coefficients for each feature.
- We observed the relationship between the features with the highest coefficient in the trained model and the outcome, and we recorded the training and cross validation R^2 scores for each model.
- When we have adequate examples to hold out for testing, we test the model on held out data.

For county level features, we use:

- CA health survey features (joint probabilities between age/race and health conditions)
- Subsets of health survey features. (i.e. only adults with heart disease information)
- 2018 US census data features (age/race/sex)

The COVID-19 outcomes we modeled are:

- Mortality growth rate (from Johns Hopkins) – Fit by previous slide
- Positives slope (from Johns Hopkins) - Fit by previous slide
- Average % of hospitalized patients suspected of having COVID that are in the ICU on a given day (California only - from CA DOH)

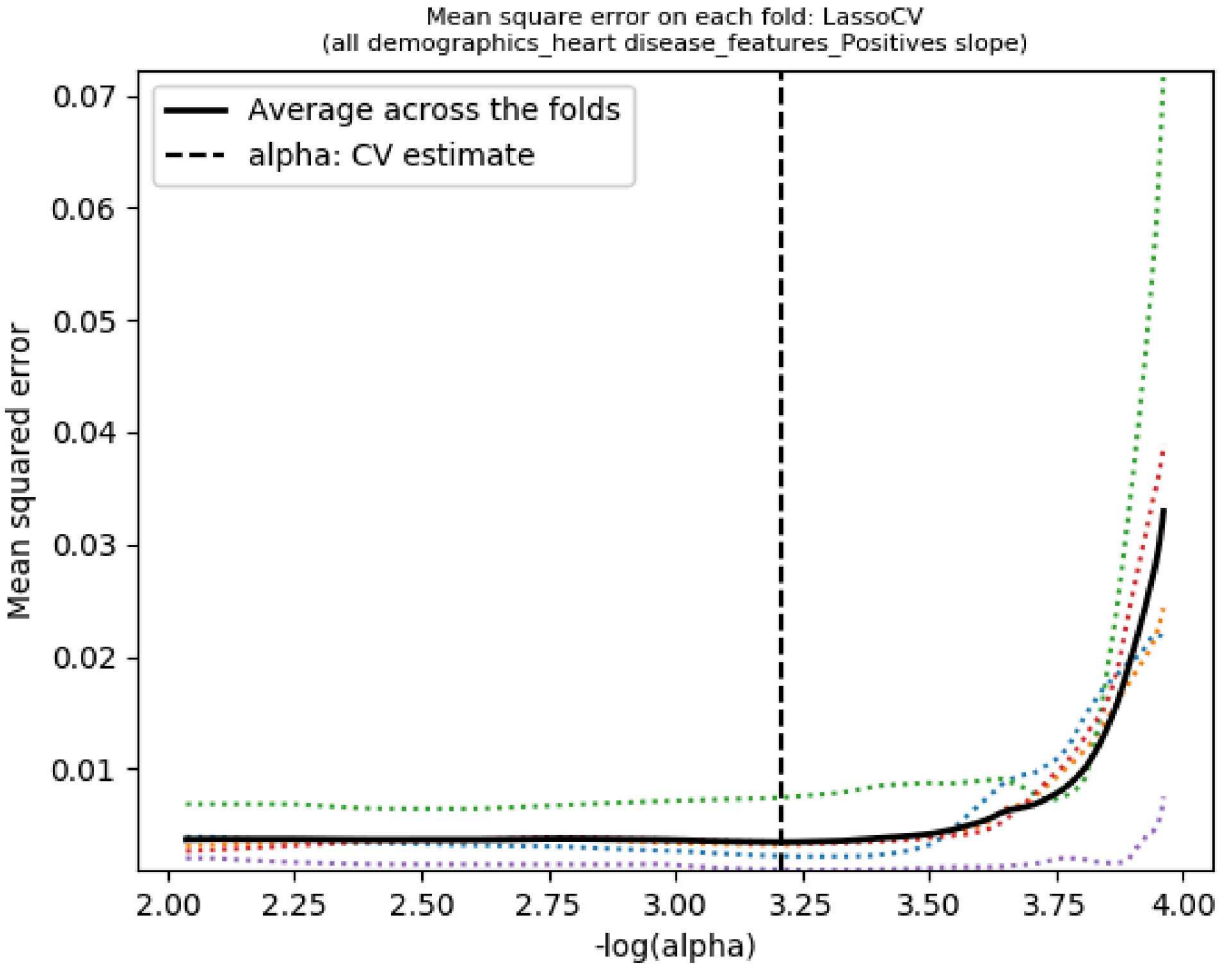
Model Tuning: CA Dataset

Using the CA health survey data, we trained a model with only percentages of adults broken down by age group, race, and heart health to fit the COVID positives slope.

The figure illustrates the 5 cross validation (CV) runs used to pick the best alpha parameter for the Lasso CV model.

Over 38 training examples:

- Training R^2 Score: 0.51
- Mean Absolute Error: 0.031
- CV Mean Test Error: 0.045 +/- 0.038
- CV Mean Train Score: 0.51 +/- 0.03



Preliminary Results – SES Proxy Predicts County Mortality Growth

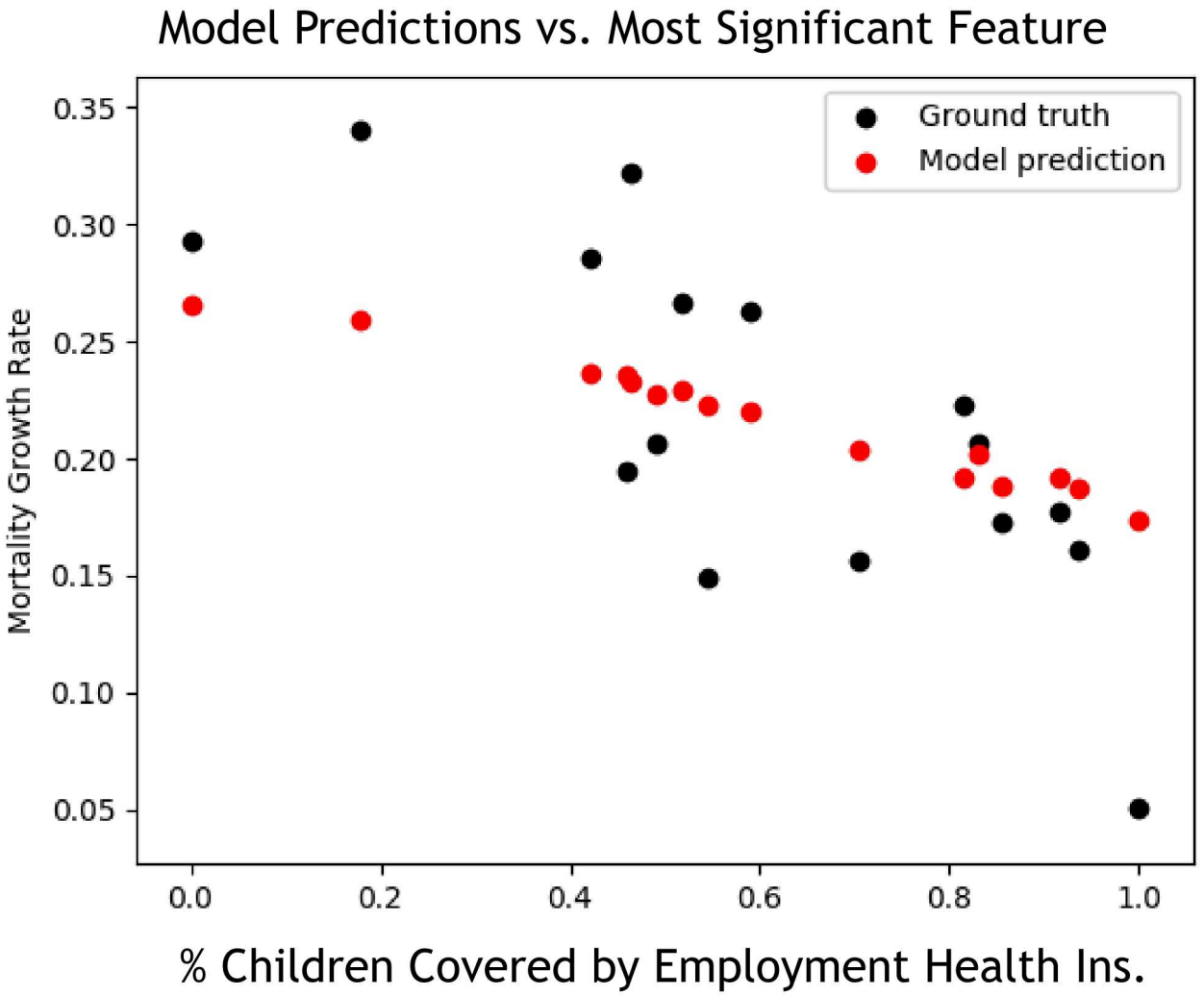
Using the CA health survey data, we trained a model with all available features to fit the COVID Mortality growth rate.

Over 16 training examples:

- Training R^2 Score: 0.43
- Training Mean Absolute Error: 0.045
- CV Mean Test Error: 0.061 +/- 0.040
- CV Mean Train Score: 0.45 +/- 0.07

This figure shows the relationship between ***the most significant feature*** in the model (the highest coefficient in the trained model) to the outcome.

Lasso regression merges similar features into one, **it is likely that the feature shown is representative of the % of county residents with employment based insurance.**



Preliminary Results - Continued

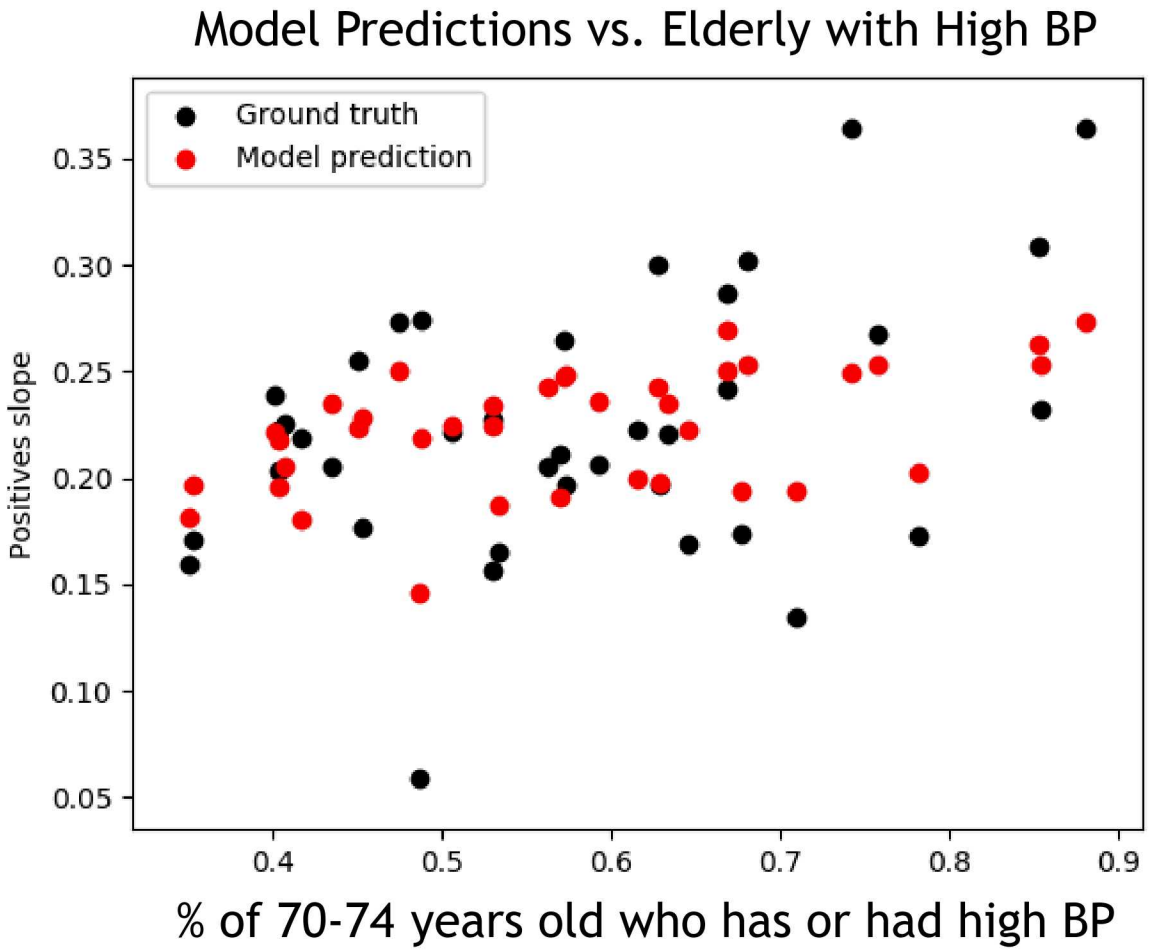
The best model fits we observed were those obtained by isolating particular health features:

- Heart Disease
- Blood Pressure
- BMI
- Asthma

And when using all available features.

Example: California counties' true values (black dots) and model predictions (red dots) when isolating proportion of elderly residents with high blood pressure.

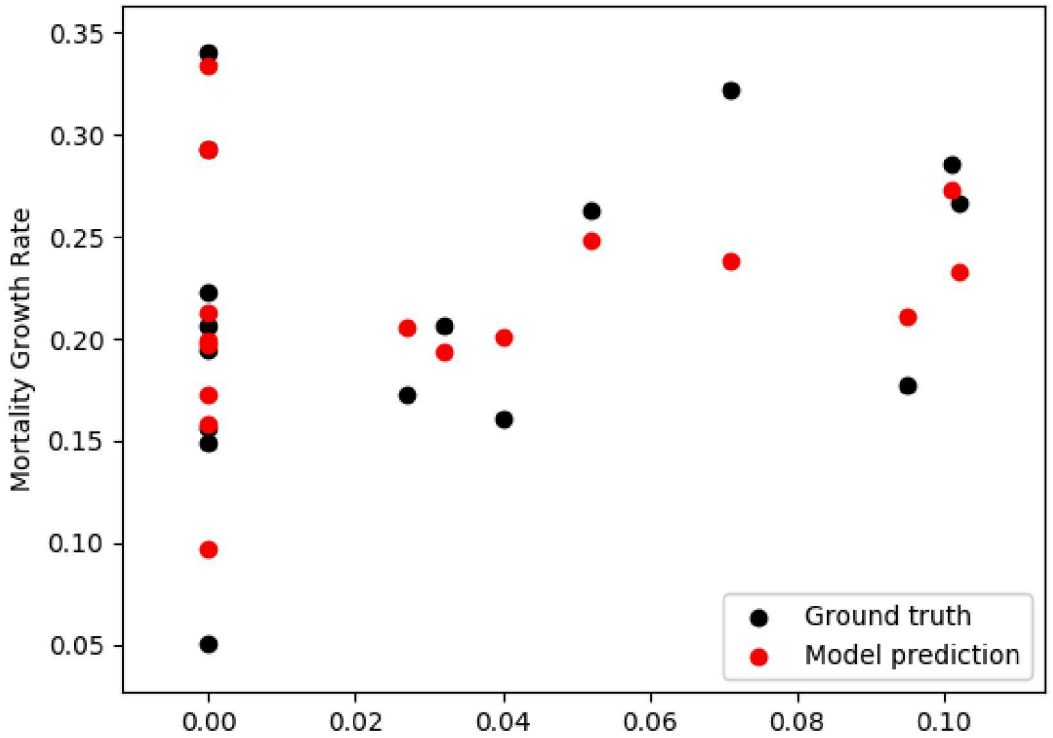
Result: Small but positive effect on predicting individual county's exponential rate of positive tests.



Two More Example Results: Predicting County Rate of Mortality Increase

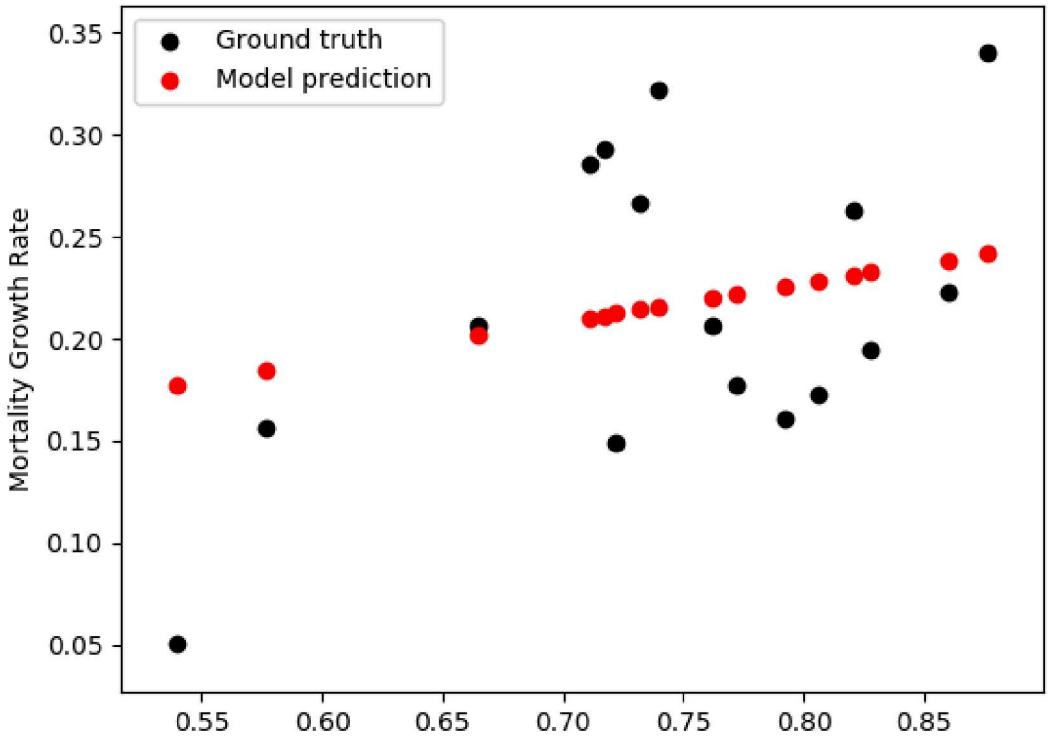
Over 130 Analyses Performed – Two more examples with strong effects (red slope)

Model Predictions vs. Black With CV Disease



% of African Americans who have heart disease

Model Predictions vs. Elderly With *No* CV



% of Elderly who *do not* have heart disease

Counterintuitive results like the second figure require further inspection and interpretation.

Next Steps

Deeper Examination of Demographic and Comorbidity Effects by County

- Fit to better outcome values (linear slopes are noisy and not informative)
- Isolate most predictive features
- Perform full joint prediction runs

UNM Data

- Evaluate large-scale models for outcome prediction based on demographic and comorbidity data

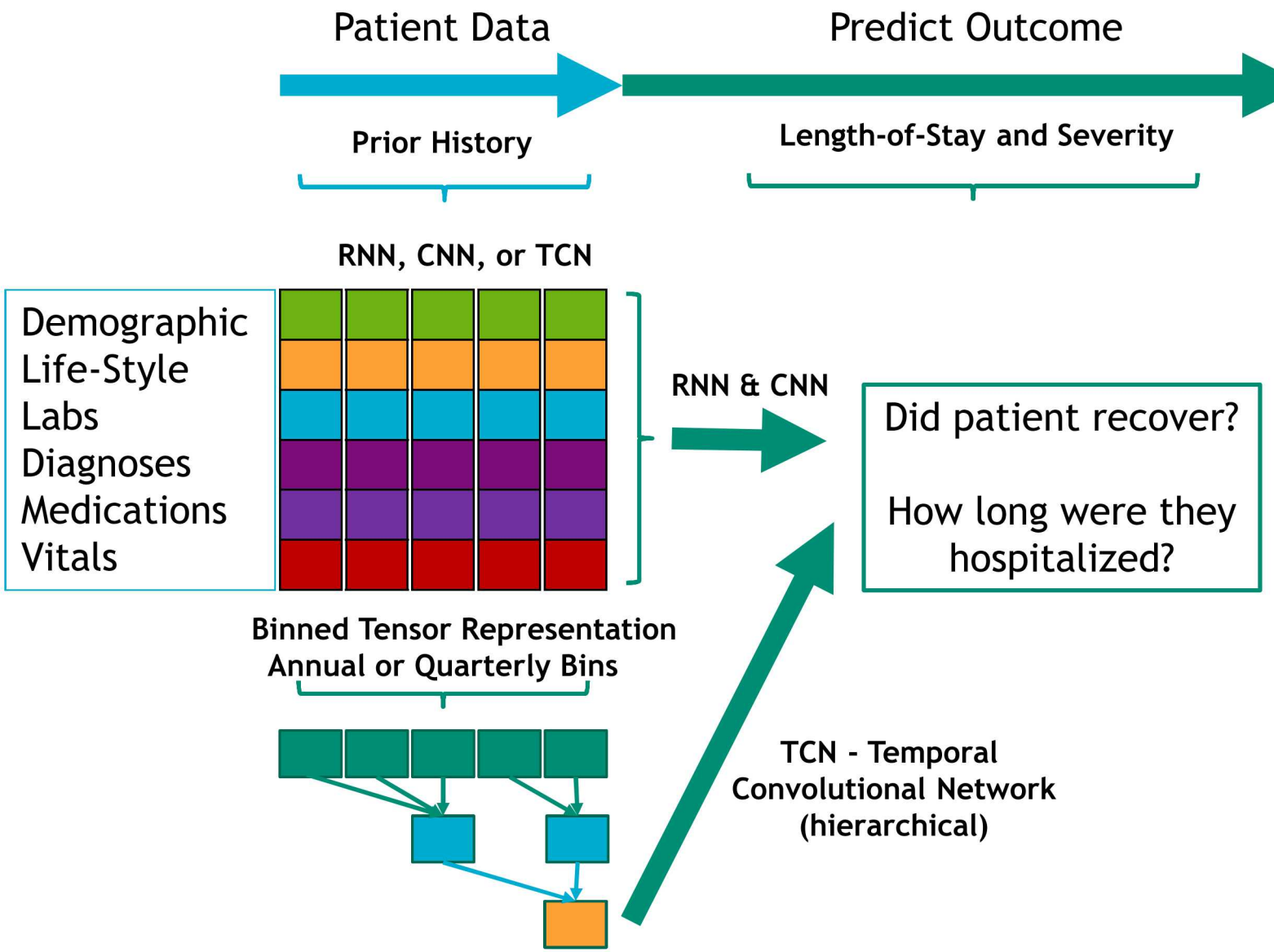
VA Data

- Apply models within VA enclave one IRB approvals are in place
- Direct examination of COVID-positive patients.

Longitudinal DNN EHR Analysis

- Deep neural network learning applied to longitudinal EHR data
- Collaboration with teams from LANL, PNNL, and Argonne.

VA Data: Longitudinal Risk Prediction Using RNNs and TCNs



Project Plan

1) Select patients who have COVID-related coding from the EHR DB.
2) Bin patient history by year or quarter. Normalize values.
3) Train model on patient data.
4) Predict probability of event on held out subsample.
5) Evaluate results in terms of accuracy, AUC and Average Precision. Compare to baseline models.

Stage-Two Deliverables

Effects of individual comorbidities

- County-level model fits reveal the effects of individual demographic and comorbidity features.

Improved patient risk prediction

- Longitudinal EHR analysis will train an improved model to more accurately predict infection severity based on a patient's medical history.
- We will produce an interpretable model based on our findings to better integrate with medical professionals.

Improved risk prognostics for all US counties

- County-level models will be extensible to all US counties as broad demographic and comorbidity prevalence data is available.
- Nuanced effects can then be incorporated into our epidemiological models.