COVID-19 Data Repository and County-level Death Count Prediction in the US

Bin Yu
UC Berkeley Statistics, EECS, CCB

github.com/Yu-Group/covid19-severity-prediction

Website: covidseverity.com

IAS Virtual Event Series
June 25, 2020
PI: Bin Yu

N. Altieri  R. Barter  J. Duncan  R. Dwivedi  K. Kumbier  X. Li  R. Netzorg

B. Park  C. Singh (Student Lead)  Y. Tan  T. Tang  Y. Wang  A. Agarwal  M. Shen  C. Zhang

Many others at UC Berkeley, UCSF, Stanford, Northeastern, Univ. of Chicago, UW-Madison, ...
On March 22, we responded to a call for data science expertise by Response4Life...

Initial Goal: Help Aid Resource Allocation
County-level 7-day severity prediction
hospital demand prediction

Data Curation
- Hospital data
- County data

Modeling
- County-level 7-day severity prediction
- Hospital demand prediction

Evaluation / Visualization
- Identify hotspots and risk factors via news articles
- Visualization
- Validate forecasts
Overview: Current Data Repository & Prediction Pipeline (Open Source)

COVID-19 Data Repository
COVID-19 Cases/Deaths + County-level Data + Hospital-level Data

Multiple county-level predictors → CLEP Ensemble + MEPI intervals + Visualizations
Curating a COVID-19 Data Repository
Data curation: scraped from a variety of sources

**COVID-19 Cases/Deaths**
- USA FACTS
- The New York Times

**County-level Data**
- (Risk Factors, Demographics, Social Mobility)
  - CDC
    - Centers for Disease Control and Prevention
    - Division for Heart Disease and Stroke Prevention
  - esri
    - COVID-19 GIS Hub
  - County Health Rankings & Roadmaps
    - Building a Culture of Health, County by County
  - CMS.gov
    - Centers for Medicare & Medicaid Services
  - United States Census Bureau
  - STREETLIGHT
  - cuebiq
  - Safe Graph
  - Social Distancing Scoreboard
  - kinsa
  - Maps
    - Mobility Trends Reports
  - Google
    - COVID-19 Community Mobility Reports

**Hospital-level Data**
- (e.g., ICU beds, staff)
  - HRSA
    - Health Resources & Services Administration
  - ArcGIS Hub
  - SAFE GRAPH
  - SAPE
  - ENCE
  -糯sa
  - NH T
  - KHN
    - Kaiser Health News

**Samuel Scarpino**
A bird’s-eye view of the hospital-level & county-level data

- ~7000 hospitals in US
- ~200 features:
  - Geographical identifiers: address, lat/long, county
  - Type of facility (e.g., short term acute care, critical access)
  - Urban/rural
  - # total beds, # Med-Surg beds, # ICU beds
  - ICU Occupancy rate
  - #Employees, #RNs
  - Total discharges, average length of stay, average daily census
  - Hospital overall rating

- COVID-19 cases and deaths (NYT and USAFacts)
- Demographics
  - Population, population density, age structure
- Health risk factors
  - Heart disease, stroke, respiratory disease, smoking, diabetes, overall mortality
- Socioeconomic risk factors
  - Social vulnerability index, unemployment, poverty, education, severe housing
- Social distancing and mobility
  - County-to-county work commute, change in distance traveled, government orders
- Other relevant data
  - Sample of flight itineraries in 2019, Kinsa temperature data, voting data
Forecasting county death counts
Curses and blessings

- Very dynamic data
- Long-term predictions have to deal with feedback
- We want to predict for all 7000 counties in the US because of R4L

- Everyday, we get new observed data to measure our predictions against -- great reality check and keeps one honest
- For PPE supplies, one week prediction is adequate (we can actually do 14 day reasonably well)
Calculate a weighted average of the predictions: higher weight to the models with better (recent) historical performance\(^1\)

Combined Linear and Exponential Predictors (CLEP)

Calculate a weighted average of the predictions: higher weight to the models with better (recent) historical performance\[^1\]

\[
w_t^m \propto \exp \left( -c(1 - \mu) \sum_{i=t_0}^{t-1} \mu^{t-i} \ell(\hat{y}_i^m, y_i) \right)
\]

Without \( \mu \), the weights are well motivated through Rissanen’s predictive MDL (Minimum Description Length) principle, and \( \mu \in (0,1) \) allows adaptation to changing dynamics.

Combined Linear and Exponential **Predictor** (CLEP)

A combination of two predictors performs well

Calculate a **weighted average of the predictions**: higher weight to the models with better historical performance

\[
E[\text{deaths}_t | t] = \exp \left( \beta_0 + \beta_1 \log(\text{deaths}_{t-1} + 1) + \beta_2 \log(\text{cases}_{t-k} + 1) \right. \\
+ \left. \beta_3 \log(\text{neigh\_deaths}_{t-k} + 1) + \beta_4 \log(\text{neigh\_cases}_{t-k} + 1) \right)
\]

Prediction Intervals based on conformal prediction[2]

Previous 5-day-ahead rel. prediction errors (%)

<table>
<thead>
<tr>
<th>Date</th>
<th>Error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apr 16</td>
<td>3.3%</td>
</tr>
<tr>
<td>Apr 17</td>
<td>6.5%</td>
</tr>
<tr>
<td>Apr 18</td>
<td>9.6%</td>
</tr>
<tr>
<td>Apr 19</td>
<td>12.6%</td>
</tr>
<tr>
<td>Apr 20</td>
<td>5.5%</td>
</tr>
<tr>
<td>Apr 25</td>
<td>?</td>
</tr>
</tbody>
</table>

Prediction Intervals:

Predicted range of error
Apr 25  [-12.6%, 12.6%]

Actual error:
Apr 25  8.8%
Data and code at covidseverity.com (searchable by county)
Covidseverity.com is an automated AI system

1. Data (daily county case and death numbers) from USAFacts is scrapped automatically to our AWS instance
2. Our CLEP prediction algorithm runs on updated data on AWS automatically (Thanks to AWS and NSF)
3. Predictions, prediction intervals, plots, and maps are generated and displayed automatically

This AI system could not spot that “1525” on May 21 for King County, WA was an error. Humans in the loop would be better.

Future of AI should be human-machine collaboration
7-day prediction: LA county (new county search function)

Thanks to Google
CLEP works also for predicting hospitalization for UC hospitals
Empirical performance of MEPI for death counts

Evaluation period: March 28--April 27. Only include days since the county has 10 deaths. Having a normalized length of 0.8 means the PI is roughly \((0.6 \, \hat{y}_{t+k}, 1.4 \, \hat{y}_{t+k})\).
7-day prediction: Mercer county (Princeton), NJ
7-day prediction: King county (Seattle), WA
High case growth
Anderson County in TX

High death growth
Lee County in FL
Covidseverity.com is an automated AI system

1. Data (daily county case and death numbers) from USAFacts is scraped automatically to our AWS instance
2. Our CLEP prediction algorithm runs on updated data on AWS automatically (Thanks to AWS and NSF)
3. Predictions, prediction intervals, plots, and maps are generated and displayed automatically

This AI system could not spot that “1525” on May 21 for King County, WA was an error. Humans in the loop would be better.

Future of AI should be human-machine collaboration

Image credit: trademed.com.
Severity Index at covidseverity.com

A score* for each hospital based on:

1. Predicted cumulative deaths
2. Predicted daily deaths

* county level predicted deaths are distributed to hospitals proportional to #employees
5000 Face Shields arrived at Temple Univ Hospital on May 8

Don Landwirth, R4L
Impacts through Response4life

- Santa Clara + Temple University Med Center in Philadelphia
- in collaboration with GetUsPPE, AeroBridge, Maker Nexus, Synergy Mill maker space,
- +65k to 25 recipients in 15 states

R4L is building a salesforce logistics system for supply chain that uses our severity index
Data and code at covidseverity.com (searchable by county)
Curating a COVID-19 data repository and forecasting county-level death counts in the United States

Nick Altieri¹ †, Rebecca L Barter¹, James Duncan⁶, Raaz Dwivedi², Karl Kumbier³, Xiao Li¹, Robert Netzorg², Briton Park¹, Chandan Singh*², Yan Shuo Tan⁴, Tiffany Tang¹, Yu Wang¹, Bin Yu*¹, ², ⁴, ⁵, ⁶

¹Department of Statistics, University of California, Berkeley
²Department of EECS, University of California, Berkeley
³Department of Pharmaceutical Chemistry, University of California, San Francisco
⁴Chan Zuckerberg Biohub, San Francisco
⁵Center for Computational Biology, University of California, Berkeley
⁶Division of Biostatistics, University of California, Berkeley

May 19, 2020

†Authors ordered alphabetically. All authors contributed significantly to this work.
*Corresponding authors

This project was initiated on March 21, 2020, with the goal of helping aid the allocation of supplies to different hospitals in the U.S., in partnership with the non-profit Response4Life.
Current directions

- Data repository a popular resource for other covid-19 activities
  In a period of two weeks, 12K visits with 1.1K unique visitors; 108 clones with 53 unique cloners

- Continued support to Response4Life

- Results and blog on CSDS atlas at Univ of Chicago

- **Hospitalization prediction** in collaboration with google (and possible collaboration with California Department of Public Health and Microsoft)

- **Causal investigation** (e.g. impact of social distancing; matching of counties) (beginning)
Thank you!

Any questions?

Please visit covidseverity.com
COVID-19 Data Repository and County Death Count Prediction

Bin Yu
UC Berkeley Statistics, EECS, CCB

github.com/Yu-Group/covid19-severity-prediction

Website: covidseverity.com

Incremental Causal Effects

Dominik Rothenhaeusler
Stanford Statistics

ONR PI Meeting
June 24, 2020
PI: Bin Yu

N. Altieri
R. Barter
J. Duncan
R. Dwivedi
K. Kumbier
X. Li
R. Netzorg
B. Park
C. Singh (Student Lead)
Y. Tan
T. Tang
Y. Wang
A. Agarwal
M. Shenl
C. Zhang

Many others at UC Berkeley, UCSF, Stanford, Northeastern, Univ. of Chicago, UW-Madison, ...
Curating a COVID-19 data repository and forecasting county-level death counts in the United States

Nick Altieri†, Rebecca L Barter, James Duncan, Raza Dwivedi, Karl Kumbier, Xiao Li, Robert Netzorg, Briton Park, Chandan Singh*, Yan Shuo Tan, Tiffany Tang, Yu Wang, Bin Yu*, 1, 2, 4, 5, 6

1Department of Statistics, University of California, Berkeley
2Department of EECS, University of California, Berkeley
3Department of Pharmaceutical Chemistry, University of California, San Francisco
4Chan Zuckerberg Biohub, San Francisco
5Center for Computational Biology, University of California, Berkeley
6Division of Biostatistics, University of California, Berkeley

May 19, 2020

†Authors ordered alphabetically. All authors contributed significantly to this work.
*Corresponding authors

This project was initiated on March 21, 2020, with the goal of helping aid the allocation of supplies to different hospitals in the U.S., in partnership with the non-profit Response4Life.
Data and code at covidseverity.com (searchable by county)
On March 22, we responded to a call for data science expertise by Response4Life...

Initial Goal: Help Aid Resource Allocation
Overview: Current Data Repository & Prediction Pipeline
(Open Source)

COVID-19 Data Repository
COVID-19 Cases/Deaths + County-level Data + Hospital-level Data

Multiple county-level predictors → CLEP Ensemble + MEPI intervals + Visualizations

[Graphs and visualizations showing data trends and predictions]
Curating a COVID-19 Data Repository
Data curation: scraped from a variety of sources

COVID-19
Cases/Deaths

County-level Data
(Risk Factors, Demographics, Social Mobility)

Hospital-level Data
(e.g., ICU beds, staff)

USA FACTS

The New York Times

CDC
Centers for Disease Control and Prevention
CDC 24/7: Saving Lives, Protecting People

Division for Heart Disease and Stroke Prevention

esri
COVID-19 GIS Hub

County Health Rankings & Roadmaps
Building a Culture of Health, County by County

USDSS
United States Diabetes Surveillance System
Division of Diabetes Translation, CDC

CMS.gov
Centers for Medicare & Medicaid Services

Census Bureau

Kinsa

COVID-19 Community Mobility Reports

Google

Maps

Mobility Trends Reports

SAFE GRAPH

ArcGIS Hub

SAM Scarpino
A bird’s-eye view of the hospital-level & county-level data

- ~7000 hospitals in US
- ~200 features:
  - Geographical identifiers: address, lat/long, county
  - Type of facility (e.g., short term acute care, critical access)
  - Urban/rural
  - # total beds, # Med-Surg beds, # ICU beds
  - ICU Occupancy rate
  - #Employees, #RNs
  - Total discharges, average length of stay, average daily census
  - Hospital overall rating

- COVID-19 cases and deaths (NYT and USAFacts)
- Demographics
  - Population, population density, age structure
- Health risk factors
  - Heart disease, stroke, respiratory disease, smoking, diabetes, overall mortality
- Socioeconomic risk factors
  - Social vulnerability index, unemployment, poverty, education, severe housing
- Social distancing and mobility
  - County-to-county work commute, change in distance traveled, government orders
- Other relevant data
  - Sample of flight itineraries in 2019, Kinsa temperature data, voting data
Forecasting county death counts
Curses and blessings

- Very dynamic data
- Long-term predictions have to deal with feedback
- We want to predict for all 7000 counties in the US because of R4L

- Everyday, we get new observed data to measure our predictions against -- great reality check and keeps one honest
- For PPE supplies, one week prediction is adequate (we can actually do 14 day reasonably well)
Combined Linear and Exponential Predictors (CLEP)

1. Separate-county exponential predictor
2. Separate-county linear predictor
3. Shared-county exponential predictor
4. Shared-county exponential predictor + demographics
5. Expanded Shared-county exponential predictor

Calculate a **weighted average of the predictions**: higher weight to the models with better (recent) historical performance[^1^]

Combined Linear and Exponential Predictors (CLEP)

Calculate a weighted average of the predictions: higher weight to the models with better (recent) historical performance

\[ w_t^m \propto \exp \left( -c(1 - \mu) \sum_{i=t_0}^{t-1} \mu^{t-i} \ell(\hat{y}_i^m, y_i) \right) \]

Without \( \mu \), the weights are well motivated through Rissanen’s predictive MDL (Minimum Description Length) principle, and \( \mu \) in (0,1) allows adaptation to changing dynamics.

Combined Linear and Exponential **Predictor (CLEP)**

A combination of two predictors performs well

![Separate-county linear predictor](image1)

![Expanded Shared-county exponential predictor](image2)

\[
E[\text{deaths}_t | t] = \exp \left( \beta_0 + \beta_1 \log(\text{deaths}_{t-1} + 1) + \beta_2 \log(\text{cases}_{t-k} + 1) + \beta_3 \log(\text{neigh_deaths}_{t-k} + 1) + \beta_4 \log(\text{neigh_cases}_{t-k} + 1) \right)
\]

Calculate a weighted average of the predictions: higher weight to the models with better historical performance

---

**Prediction Intervals** based on conformal prediction[2]

Previous 5-day-ahead prediction errors (%)

- Apr 16: 3.3%
- Apr 17: 6.5%
- Apr 18: 9.6%
- Apr 19: 12.6%
- Apr 20: 5.5%
- Apr 25: ?

Prediction Intervals:

Predicted range of error
Apr 25  [-12.6%, 12.6%]

Actual error:
Apr 25  8.8%
Maximum (absolute) error prediction intervals (MEPI)

**Step 1**
Find normalized error of our predictor in the past.

\[ \Delta_T := \frac{|y_T - \hat{y}_T|}{|\hat{y}_T|}. \]

**Step 2**
Find maximum error of past 5 days.

\[ \Delta_{\text{max}} := \max_{0 \leq j \leq 4} \Delta_{t-j}. \]

**Step 3**

\[ \widehat{\text{PI}}_{t+k} := \left[ \max \left\{ \hat{y}_{t+k}(1 - \Delta_{\text{max}}), y_t \right\}, \hat{y}_{t+k}(1 + \Delta_{\text{max}}) \right] \]

Can be applied to any ML model, and it works well under exchangeability condition on the errors.
7-day prediction: LA county (new at covidseverity.com)
CLEP works also for predicting hospitalization for UC hospitals.
High case growth
Anderson County in TX

High death growth
Lee County in FL
Empirical performance of MEPI for death counts

Evaluation period: March 28--April 27. Only include days since the county has 10 deaths. Having a normalized length of 0.8 means the PI is roughly $(0.6 \tilde{y}_{t+k}, 1.4 \tilde{y}_{t+k})$. 
A score* for each hospital based on:

1. Predicted cumulative deaths
2. Predicted daily deaths

* county level predicted deaths are distributed to hospitals proportional to #employees
5000 Face Shields arrived at Temple Univ Hospital on May 8

Don Landwirth, R4L
Impacts through Response4life

- Santa Clara + Temple University Med Center in Philadelphia
- in collaboration with GetUsPPE, AeroBridge, Maker Nexus, Synergy Mill maker space,
- +65k to 25 recipients in 15 states

R4L is building a salesforce logistics system for supply chain that uses our severity index
Impact of our work beyond R4L

- Data repository a popular resource for other covid-19 activities

In a period of two weeks, 12K visits with 1.1K unique visitors; 108 clones with 53 unique cloners

- Results and blog on CSDS atlas at Univ of Chicago
- Final project option for DS 100 at UC Berkeley (> 1000 students) and Stat 542 at University of Illinois Urbana-Champaign (graduate stat-ml course)
- **Hospitalization prediction** in collaboration with google (and possible collaboration with California Department of Public Health and Microsoft)
- **Causal investigation** *(e.g. impact of social distancing; matching of counties)* (beginning)
Data and code at covidseverity.com (searchable by county)
Incremental Causal Effects

Dominik Rothenhäusler$^1$ and Bin Yu$^2$

$^1$Department of Statistics, Stanford University
$^2$Department of Electrical Engineering and Computer Science, and Department of Statistics, University of California, Berkeley

https://arxiv.org/abs/1907.13258

Supported by ONR Grant
Incremental causal effects (Rothenhaeusler and Yu, 2019)

Causal inference from observational data is challenging

Problems with confounding, overlap, weak instruments,...
Incremental causal effects (Rothenhaeusler and Yu, 2020)

Causal inference from observational data is challenging

Problems with confounding, overlap, weak instruments,....

An important motivation for causal inference is evidence to act. Action decision might need weaker evidence than a positive average treatment effect (ATE) (e.g. whether to increase exercise time).

Moving the goalpost from ATE to other estimands can help:

- Local Average Treatment Effects (Imbens and Angrist, 1994)
- Weighted ATEs (Crump et al., 2006)
- Incremental propensity score interventions (Kennedy, 2019)
- ...
Incremental causal effects: looking for gradient effect

For a continuous treatment $T$ and smooth potential outcomes $Y(t)$ define the incremental causal effect

$$
\tau_{\text{incr}} = \mathbb{E}[\partial_t Y(T)]
$$

This corresponds to the average change in outcome if slightly increasing the treatment for every unit in the population.

It is often estimated via the average derivative $\mathbb{E}[\partial_t \mathbb{E}[Y|X, T]]$ under appropriate assumptions. Such estimands have appeared in the econ literature (Powell et al., 1989, Newey & Stoker 2003, Banerjee, 2007,...) but have received relatively little attention.
Incremental causal effects - our contributions

- **Incremental causal effects are identified under weaker assumptions** (a local ignorability and local overlap assumption)

  Conditionally on covariates, units only have to be comparable locally at current treatment $t$, not necessarily globally across all $t$

- **Incremental causal effects can be estimated with lower or equal variance** than ATE $E[Y(t+1)] - E[Y(t)]$ if the treatment distribution is Gaussian

- In high-dimensional settings, we use orthogonalization to **transform** the problem of estimation and inference of incremental effects to estimation and inference of a coefficient in a **standard regression model**

  We can use the desparsified Lasso for estimation and inference of incremental causal effects

Future work on “weak causality”

- So far: change type of intervention or target population
- Interpolate between effects that are easy to estimate and the ATE. What’s the right way to interpolate?
- Aggregate weak causal evidence across data sets
- Investigate “relaxed causal invariance constraints”
Thank you!

covidseverity.com
COVID-19 Data Repository and Severity Prediction

Yu Group
UC Berkeley Statistics, EECS, CCB

github.com/Yu-Group/covid19-severity-prediction

Website: covidseverity.com

IAS Virtual Event Series
June 25, 2020
COVID-19 Data Repository and Severity Prediction

Yu Group
UC Berkeley Statistics, EECS, CCB

- Curated data repository
- Developed ensemble prediction algorithm at county level for death counts, 7-days ahead
- Designed covid severity index at hospital level for a Salesforce logistics system by R4L

Website: covidseverity.com

5000 Face Shields arrived at Temple Univ Hospital on May 8

Don Landwirth, R4L
Initial Goal: Help Aid Resource Allocation
PI: Bin Yu

N. Altieri  
R. Barter  
J. Duncan  
R. Dwivedi  
K. Kumbier  
X. Li  
R. Netzorg  

B. Park  
C. Singh  
(Y. Tan)  
T. Tang  
Y. Wang  

Many others at UC Berkeley, UCSF, Stanford, Northeastern, Univ. of Chicago, UW-Madison, ...
Our team
from UC Berkeley Statistics/EECS and UCSF

N. Altieri  R. Barter  J. Duncan  R. Dwivedi  K. Kumbier  X. Li  R. Netzorg
B. Park  C. Singh (Student Lead)  Y. Tan  T. Tang  Y. Wang

Many others at UC Berkeley, UCSF, Stanford, Northeastern, Univ. of Chicago, UW-Madison, ...
Data Curation
- Hospital data
- County data

Modeling
- County-level 7-day severity prediction
- Hospital demand prediction

Evaluation / Visualization
- Identify hotspots and risk factors via news articles
- Visualization
- Validate forecasts
Curating a COVID-19 Data Repository
Data curation: scraped from a variety of sources

**COVID-19 Cases/Deaths**

- USAFACTS
- The New York Times

**County-level Data**

- (Risk Factors, Demographics, Social Mobility)
  - CDC: Centers for Disease Control and Prevention
  - Division for Heart Disease and Stroke Prevention
  - esri: COVID-19 GIS Hub
  - County Health Rankings & Roadmaps
  - USDSS: United States Diabetes Surveillance System
  - Division of Diabetes Translation, CDC
  - CMS.gov: Centers for Medicare & Medicaid Services
  - United States Census Bureau
  - Streetlight
  - Cuebiq
  - Introducing the Unacast Social Distancing Scoreboard
  - kinsa
  - SAFE GRAPH
  - Apple Maps: Mobility Trends Reports
  - COVID-19 Community Mobility Reports
  - Johns Hopkins University
  - Google

**Hospital-level Data**

- (e.g., ICU beds, staff)
  - HRSA: Health Resources & Services Administration
  - ArcGIS Hub
  - Health Resources & Services Administration

**Authors**

- Samuel Scarpino
A bird’s-eye view of the **hospital-level & county-level data**

- ~7000 hospitals in US
- ~200 features:
  - Geographical identifiers: address, lat/long, county
  - Type of facility (e.g., short term acute care, critical access)
  - Urban/rural
  - # total beds, # Med-Surg beds, # ICU beds
  - ICU Occupancy rate
  - #Employees, #RNs
  - Total discharges, average length of stay, average daily census
  - Hospital overall rating
- COVID-19 cases and deaths (NYT and USAFacts)
- Demographics
  - Population, population density, age structure
- Health risk factors
  - Heart disease, stroke, respiratory disease, smoking, diabetes, overall mortality
- Socioeconomic risk factors
  - Social vulnerability index, unemployment, poverty, education, severe housing
- Social distancing and mobility
  - County-to-county work commute, change in distance traveled, government orders
- Other relevant data
  - Sample of flight itineraries in 2019, Kinsa temperature data, voting data
Data Repository Traffic & Users (Last 2 weeks)

Estimated total views: ~18K
Forecasting county death counts
Combined Linear and Exponential **Predictor** (CLEP)

A combination of two models performs well


Calculate a **weighted average of the predictions**: higher weight to the models with better historical performance\(^1\)

Combined Linear and Exponential Predictors (CLEP)

Calculate a **weighted average of the predictions**: higher weight to the models with better historical performance[^2].

\[ w_t^m \propto \exp \left( -c(1 - \mu) \sum_{i=t_0}^{t-1} \mu^{t-i} \ell(\hat{y}_i^m, y_i) \right) \]

Our county-level 7-day predictive performance

Selected CA counties
Our county-level 7-day predictive performance

Rapidly Growing Counties
A score* for each hospital based on:

1. Predicted cumulative deaths
2. Predicted daily deaths

* county level predicted deaths are distributed to hospitals proportional to #employees
Mapping Deaths and the Hospital Severity Index Over Time

Predicted New Deaths for 2020-05-10

Los Angeles

Bay Area
(Interactive) map visualizations

County-level predicted cumulative # of deaths*

Hospital severity index*

*Maps for 04/15

Collaborating with the Center for Spatial Data Science (CSDS) at University of Chicago to add our predictions and severity index to the U.S. COVID-19 Atlas.
5000 Face Shields arrived at Temple Univ Hospital on May 8

Don Landwirth, R4L
5000 Face Shields arrived at Temple Univ Hospital on May 8

Don Landwirth, R4L
Impacts through Response4life

500k face shields in US by the end of may

- Santa Clara + Temple University Med Center in Philadelphia
- in collaboration with GetUsPPE, AeroBridge, Maker Nexus, Synergy Mill maker space,
- +65k to 25 recipients in 15 states in 2 weeks
- +500k outside the US

R4L is building a salesforce logistics system for supply chain that uses our severity index
Impact of our work beyond R4L

- Data repository a popular resource for other covid-19 activities
  
  In last two weeks, 12K visits with 1.1K unique visitors; 108 clones with 53 unique cloners

- Results on CSDS atlas at Univ of Chicago
- Final project option for DS 100 at UC Berkeley (> 1000 students) and Stat 542 at University of Illinois Urbana-Champaign (graduate stat-ml course)
- Possible collaboration with California Department of Public Health
- Possible causal inference through matching of counties
Curating a COVID-19 data repository and forecasting county-level death counts in the United States

Nick Altieri¹, †, Rebecca Barter³, James Duncan⁶, Razz Dwivedi², Karl Kumbier³, Xiao Li¹, Robert Netzorg², Briton Park¹, Chandan Singh*², Yan Shuo Tan¹, Tiffany Tang¹, Yu Wang¹, Bin Yu*¹, ², ⁴, ⁵, ⁶

¹Department of Statistics, University of California, Berkeley
²Department of EECS, University of California, Berkeley
³Department of Pharmaceutical Chemistry, University of California, San Francisco
⁴Chan Zuckerberg Biohub, San Francisco
⁵Center for Computational Biology, University of California, Berkeley
⁶Division of Biostatistics, University of California, Berkeley

April 29, 2020

†Authors ordered alphabetically. All authors contributed significantly to this work.
*Corresponding authors
This project was initiated on March 21, 2020, with the goal of helping aid the allocation of supplies to different hospitals in the U.S., in partnership with the non-profit Response4Life.
COVID-19 Data Repository and County Death Count Prediction

Yu Group
UC Berkeley Statistics, EECS, CCB

github.com/Yu-Group/covid19-severity-prediction
Website: covidseverity.com
PI: Bin Yu

N. Altieri  R. Barter  J. Duncan  R. Dwivedi  K. Kumbier  X. Li  R. Netzorg

B. Park  C. Singh (Student Lead)  Y. Tan  T. Tang  Y. Wang

Many others at UC Berkeley, UCSF, Stanford, Northeastern, Univ. of Chicago, UW-Madison, ...
Overview: Current Data Repository & Prediction Pipeline

COVID-19 Data Repository
COVID-19 Cases/Deaths + County-level Data + Hospital-level Data

Multiple county-level predictors → CLEP Ensemble + MEPI intervals + Visualizations
Combined Linear and Exponential **Predictor** (CLEP)

A combination of two models performs well

1. Separate-county linear predictor
2. Expanded Shared-county exponential predictor

Calculate a **weighted average of the predictions**: higher weight to the models with better historical performance¹

Death count prediction results: 4/20-5/10

Selected CA counties
Data Repository

Overview of sources (county/hospital)
- pipelines/processes
- Current users
- Current efforts
**Data:** scraped from a variety of sources

- **COVID-19 Cases/Deaths**
  - USAFACTS
  - The New York Times

- **County-level Data**
  - Risk Factors, Demographics, Social Mobility
  - CDC
    - Centers for Disease Control and Prevention
  - Division for Heart Disease and Stroke Prevention
  - esri
    - COVID-19 GIS Hub
  - County Health Rankings & Roadmaps
    - Building a Culture of Health, County by County
  - CDC
    - Introducing the Unacast Social Distancing Scoreboard
  - USDSS
    - United States Diabetes Surveillance System
    - Division of Diabetes Translation, CDC
  - Johns Hopkins University
  - CMS.gov
  - United States Census Bureau
  - Streetlight
  - KHN
  - SAFEGRAPH
  - The Center for Spatial Data Science

- **Hospital-level Data**
  - (e.g., ICU beds, staff)
  - HRSA
    - Health Resources & Services Administration
  - ArcGIS Hub

---

Samuel Scarpino
A bird’s-eye view of the hospital-level & county-level data

- ~7000 hospitals in US
- ~200 features:
  - Geographical identifiers: address, lat/long, county
  - Type of facility (e.g., short term acute care, critical access)
  - Urban/rural
  - # total beds, # Med-Surg beds, # ICU beds
  - ICU Occupancy rate
  - #Employees, #RNs
  - Total discharges, average length of stay, average daily census
  - Hospital overall rating
- COVID-19 cases and deaths (NYT and USAFacts)
- Demographics
  - Population, population density, age structure
- Health risk factors
  - Heart disease, stroke, respiratory disease, smoking, diabetes, overall mortality
- Socioeconomic risk factors
  - Social vulnerability index, unemployment, poverty, education, severe housing
- Social distancing and mobility
  - County-to-county work commute, change in distance traveled, government orders
- Other relevant data
  - Sample of flight itineraries in 2019, Kinsa temperature data, voting data
Data Repository Traffic & Users (Last 2 weeks)

Estimated total views: ~18K
Impact: 5000 Face Shields arrived at Temple Univ Hospital on May 8
Other Impacts of Our Data Repository

- Data repository a popular resource for other COVID-19 activities:
  - In last two weeks, 2.9K visits with 394 unique visitors;
  - 153 clones with 102 unique cloners
- Results on CSDS atlas at University of Chicago
- Final project option for DS100 at UC Berkeley (> 1000 students) and Stat542 at University of Illinois Urbana-Champaign (graduate stat-ml course)
- Collaboration with Google OpenSource, Microsoft’s AI for Good, on hospitalization need prediction (ongoing)
- Possible collaboration with California Department of Public Health
- Exploratory causal inference through matching of counties (ongoing)
Combined Linear and Exponential **Predictor** (CLEP)

A combination of two models performs well

Calculate a **weighted average of the predictions**: higher weight to the models with better historical performance

Current Data Repository & Pipeline (alternative to page 2)

COVID-19 Data Repository
COVID-19 Cases/Deaths + County-level Data + Hospital-level Data

Multiple county-level predictors → CLEP Ensemble + MEPI intervals + Visualizations
Our county-level 7-day predictive performance

Worst Affected Counties

- Kings County, NY
- Queens County, NY
- Bronx County, NY
Impacts through Response4life

500k face shields in US by the end of may

- Santa Clara + Temple University Med Center in Philadelphia
- in collaboration with GetUsPPE, AeroBridge, Maker Nexus, Synergy Mill maker space,
- +65k to 25 recipients in 15 states in 2 weeks
- +500k outside the US

Curating data repository

<table>
<thead>
<tr>
<th>Data variable</th>
<th>Description</th>
<th>Source data set</th>
</tr>
</thead>
<tbody>
<tr>
<td>countyFIPS</td>
<td>state-county FIPS Code</td>
<td>county_fips</td>
</tr>
<tr>
<td>STATEFP</td>
<td>state FIPS Code</td>
<td>county_popcenters</td>
</tr>
<tr>
<td>COUNTYFP</td>
<td>county FIPS Code</td>
<td>county_popcenters</td>
</tr>
<tr>
<td>CountyName</td>
<td>county name</td>
<td>county_fips</td>
</tr>
<tr>
<td>StateName</td>
<td>state abbreviation</td>
<td>county_fips</td>
</tr>
<tr>
<td>State</td>
<td>state name</td>
<td>county_latlong</td>
</tr>
</tbody>
</table>

Visualizations
Most recent 20 days zoom in

Selected CA counties
Our county-level 7-day predictive performance

Worst Affected Counties
Our county-level 7-day predictive performance

Rapidly Growing Counties
COVID-19 Data Repository and Severity Prediction

Yu Group
UC Berkeley Statistics, EECS, CCB

github.com/Yu-Group/covid19-severity-prediction

Website: covidseverity.com

California Department of Public Health, Modeling Group
May 11, 2020
Goal: Help Aid Resource Allocation
Our team
from UC Berkeley Statistics/EECS and UCSF

N. Altieri  R. Barter  J. Duncan  R. Dwivedi  K. Kumbier  X. Li  R. Netzorg

B. Park  C. Singh (Student Lead)  Y. Tan  T. Tang  Y. Wang

Many others at UC Berkeley, UCSF, Stanford, Northeastern, Univ. of Chicago, UW-Madison, ...
County-level 7-day severity prediction

Hospital demand prediction

Modeling
- County-level 7-day severity prediction
- Hospital demand prediction

Data Curation
- Hospital data
- County data

Evaluation / Visualization
- Identify hotspots and risk factors via news articles
- Visualization
- Validate forecasts

Data Curation

Modeling

Evaluation / Visualization
Curating a COVID-19 Data Repository
Data Processing Pipeline

Data Scraping
- Collect 1M records from 10+ data sources
- Monitor data changes 24/7 powered by AWS

Data Cleaning
- Handling missing and erratic entries
- Automated python script

Data Validity
- Compare data across different sources to ensure data validity
- Search for emerging data sources

For almost a month, 2 full-time students, and on-going with 1 full-time student

Data and code available: [https://github.com/Yu-Group/covid19-severity-prediction](https://github.com/Yu-Group/covid19-severity-prediction)
★ Being used by multiple research groups across the country
Data: scraped from a variety of sources

COVID-19 Cases/Deaths

USA Facts
The New York Times

County-level Data
(Risk Factors, Demographics, Social Mobility)

CDC
Centers for Disease Control and Prevention
CDC 24/7: Saving Lives, Protecting People™
Division for Heart Disease and Stroke Prevention

esri
COVID-19 GIS Hub

County Health Rankings & Roadmaps
Building a Culture of Health, County by County

USDSS
United States Diabetes Surveillance System
Division of Diabetes Translation, CDC

JOHNS HOPKINS UNIVERSITY
CMS.gov
Centers for Medicare & Medicaid Services

Census Bureau

Social Distancing Scoreboard

United States Department of Transportation

Kaiser Health News

SAFE GRAPH

Hospital-level Data
(e.g., ICU beds, staff)

HRSA
Health Resources & Services Administration

ArcGIS Hub

Samuel Scarpino
Forecasting county death counts
Calculate a **weighted average of the predictions**: higher weight to the models with better historical performance\(^2\)

---

Combined Linear and Exponential Predictors (CLEP)

Calculate a **weighted average of the predictions**: higher weight to the models with better historical performance\(^2\)

\[
\omega_t^m \propto \exp \left(-c(1-\mu) \sum_{i=t_0}^{t-1} \mu^{t-i} l(\hat{y}_i^m, y_i) \right)
\]

Combined Linear and Exponential Predictors (CLEP)

A smaller combination performed better

Separate-county linear predictor + Expanded Shared-county exponential predictor

Calculate a **weighted average of the predictions**: higher weight to the models with better historical performance[2]

Our county-level 7-day predictive performance

Selected CA counties
Most recent 20 days zoom in

Santa Clara County, CA

Los Angeles County, CA

Monterey County, CA

Sonoma County, CA

Selected CA counties
Our county-level 7-day predictive performance

Rapidly Growing Counties
Our county-level 7-day predictive performance

Rapidly Growing Counties
Prediction Intervals:

Previous 5-day-ahead prediction errors (%)

<table>
<thead>
<tr>
<th>Date</th>
<th>Error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apr 16</td>
<td>3.3%</td>
</tr>
<tr>
<td>Apr 17</td>
<td>6.5%</td>
</tr>
<tr>
<td>Apr 18</td>
<td>9.6%</td>
</tr>
<tr>
<td>Apr 19</td>
<td>12.6%</td>
</tr>
<tr>
<td>Apr 20</td>
<td>5.5%</td>
</tr>
<tr>
<td>Apr 25</td>
<td>?</td>
</tr>
</tbody>
</table>

Take the max
Prediction Intervals:

Predicted range of error
Apr 25  [-12.6%, 12.6%]

Actual error:
Apr 25  8.8%
Maximum (absolute) error prediction intervals (MEPI)

**Step 1**
Find normalized error of our predictor in the past.

\[ \Delta_\tau := \frac{|y_\tau - \hat{y}_\tau|}{|\hat{y}_\tau|}. \]

**Step 2**
Find maximum error of past 5 days.

\[ \Delta_{\text{max}} := \max_{0 \leq j \leq 4} \Delta_{\tau-j}. \]

**Step 3**

\[ \hat{\text{PI}}_{t+k} := [\max \{ \hat{y}_{t+k}(1 - \Delta_{\text{max}}), y_t \}, \hat{y}_{t+k}(1 + \Delta_{\text{max}})] \]

Can be applied to any ML model, and it works well under **exchangeability** condition on the errors.
Empirical performance of MEPI

Evaluation period: March 28--April 27. Only include days since the county has 10 deaths. Having a normalized length of 0.8 means the PI is roughly \((0.6 \, \hat{y}_{t+k}, 1.4 \, \hat{y}_{t+k})\).
A score* for each hospital based on:

1. Predicted cumulative deaths
2. Predicted daily deaths

* county level predicted deaths are distributed to hospitals proportional to #employees
Mapping Deaths and the Hospital Severity Index Over Time

Predicted New Deaths for 2020-05-10

Los Angeles

Bay Area
Collaborating with the Center for Spatial Data Science (CSDS) at University of Chicago to add our predictions and severity index to the U.S. COVID-19 Atlas.
5000 Face Shields arrived at Temple Univ Hospital on May 8

Don Landwirth, R4L
Impacts through Response4life

500k face shields in US by the end of May

- Santa Clara + Temple University Med Center in Philadelphia
- in collaboration with GetUsPPE, AeroBridge, Maker Nexus, Synergy Mill maker space,
- +65k to 25 recipients in 15 states in 2 weeks
- +500k outside the US

R4L is building a salesforce logistics system for supply chain that uses our severity index
Impact of our work beyond R4L

- Data repository a popular resource for other covid-19 activities
  In last two weeks, 12K visits with 1.1K unique visitors; 108 clones with 53 unique cloners
- Results on CSDS atlas at Univ of Chicago
- Final project option for DS 100 at Berkeley (> 1000 students) and Stat 542 at University of Illinois Urbana-Champaign (graduate stat-ml course)
- Possible causal inference through matching of counties
- Possible collaboration with California Department of Public Health (?)
Curating a COVID-19 data repository and forecasting county-level death counts in the United States

Nick Altieri¹,†, Rebecca Barter¹, James Duncan⁶, Raaz Dwivedi², Karl Kumbier³, Xiao Li¹, Robert Netzorg⁷, Briton Park¹, Chandan Singh*², Yan Shuo Tan¹, Tiffany Tang¹, Yu Wang¹, Bin Yu*¹, 2, 4, 5, 6

¹Department of Statistics, University of California, Berkeley
²Department of EECS, University of California, Berkeley
³Department of Pharmaceutical Chemistry, University of California, San Francisco
⁴Chan Zuckerberg Biohub, San Francisco
⁵Center for Computational Biology, University of California, Berkeley
⁶Division of Biostatistics, University of California, Berkeley

April 29, 2020

†Authors ordered alphabetically. All authors contributed significantly to this work.
*Corresponding authors

This project was initiated on March 21, 2020, with the goal of helping aid the allocation of supplies to different hospitals in the U.S., in partnership with the non-profit Response4Life.
**Impacts**

500k face shields in US by mid-May

- Santa Clara + Temple University Med Center in Philadelphia
- in collaboration with GetUsPPE, AeroBridge, Maker Nexus, Synergy Mill maker space, R4L
- *+65k to 25 recipients in 15 states in 2 weeks*
- *+500k outside the US*

Salesforce system

**Curating data repository**

<table>
<thead>
<tr>
<th>Data variable</th>
<th>Description</th>
<th>Source data set</th>
</tr>
</thead>
<tbody>
<tr>
<td>countyFIPS</td>
<td>state-county FIPS Code</td>
<td>county_fips</td>
</tr>
<tr>
<td>STATEFP</td>
<td>state FIPS Code</td>
<td>county_popcenters</td>
</tr>
<tr>
<td>COUNTYFP</td>
<td>county FIPS Code</td>
<td>county_popcenters</td>
</tr>
<tr>
<td>CountyName</td>
<td>county name</td>
<td>county_fips</td>
</tr>
<tr>
<td>StateName</td>
<td>state abbreviation</td>
<td>county_fips</td>
</tr>
<tr>
<td>State</td>
<td>state name</td>
<td>county_latlong</td>
</tr>
</tbody>
</table>

**Visualizations**
Impacts through Response4life

500k face shields in US by the end of May

- Santa Clara + Temple University Med Center in Philadelphia
- in collaboration with GetUsPPE, AeroBridge, Maker Nexus, Synergy Mill maker space,
- +65k to 25 recipients in 15 states in 2 weeks
- +500k outside the US

R4L is building a salesforce logistics system for supply chain that uses our severity index
Our county-level 7-day predictive performance

Focusing on 6 of the worst-affected counties

*Based on 4/8 data
Goal: Help Aid Resource Allocation
Our team
from UC Berkeley Statistics/EECS and UCSF

N. Altieri  R. Barter  J. Duncan  R. Dwivedi  K. Kumbier  X. Li  R. Netzorg
B. Park  C. Singh (Student Lead)  Y. Tan  T. Tang  Y. Wang

Many others at UC Berkeley, UCSF, Stanford, Northeastern, Univ. of Chicago, UW-Madison, ...
County-level 7-day severity prediction

Hospital demand prediction

Modeling

Hospital data

County data

Data Curation

County-level 7-day severity prediction

hospital demand prediction

Evaluation / Visualization

Identify hotspots and risk factors via news articles

Visualization

Validate forecasts
Impacts

500k face shields in US by mid-may

- Santa Clara + Temple University Med Center in Philadelphia
- in collaboration with GetUsPPE, AeroBridge, Maker Nexus, Synergy Mill maker space, R4L
- +65k to 25 recipients in 15 states in 2 weeks
- +500k outside the US

Salesforce system

Curating data repository

<table>
<thead>
<tr>
<th>Data variable</th>
<th>Description</th>
<th>Source data set</th>
</tr>
</thead>
<tbody>
<tr>
<td>countyFIPS</td>
<td>state-county FIPS Code</td>
<td>county_fips</td>
</tr>
<tr>
<td>STATEFP</td>
<td>state FIPS Code</td>
<td>county_popcenters</td>
</tr>
<tr>
<td>COUNTYFP</td>
<td>county FIPS Code</td>
<td>county_popcenters</td>
</tr>
<tr>
<td>CountyName</td>
<td>county name</td>
<td>county_fips</td>
</tr>
<tr>
<td>StateName</td>
<td>state abbreviation</td>
<td>county_fips</td>
</tr>
<tr>
<td>State</td>
<td>state name</td>
<td>county_fips</td>
</tr>
</tbody>
</table>

Visualizations
Part I: Curating a COVID-19 Data Repository
Outline of Part I: Data Curation

● Our data processing pipeline
● Overview of the data
● Frequently overlooked aspects and challenges
● Some useful tools

Data Processing Pipeline

**Data Scraping**
- Collect 1M records from 10+ data sources
- Monitor data changes 24/7 powered by AWS

**Data Cleaning**
- Handling missing and erratic entries
- Automated python script

**Data Validity**
- Compare data across different sources to ensure data validity
- Search for emerging data sources

For almost a month, 2 full-time students, and on-going with 1 full-time student

Data and code available: [https://github.com/Yu-Group/covid19-severity-prediction](https://github.com/Yu-Group/covid19-severity-prediction)
★ Being used by multiple research groups across the country
Data: scraped from a variety of sources

COVID-19 Cases/Deaths
- USA FACTS
- The New York Times

County-level Data
(Risk Factors, Demographics, Social Mobility)
- CDC
- Centers for Disease Control and Prevention
- Division for Heart Disease and Stroke Prevention
- esri
- COVID-19 GIS Hub
- County Health Rankings & Roadmaps
- Building a Culture of Health, County by County
- GHDx
- Introducing the Unacast Social Distancing Scoreboard
- USDSS
- United States Diabetes Surveillance System
- Division of Diabetes Translation, CDC
- Johns Hopkins University
- Centers for Medicare & Medicaid Services
- CMS.gov
- United States Census Bureau
- Streetlight
- KHN
- Kaiser Health News
- Safe Graph
- Cuebiq
- Kinsa

Hospital-level Data
(e.g., #ICU beds, staff)
- HRSA
- Health Resources & Services Administration
- ArcGIS Hub
- Samuel Scarpino
A bird’s-eye view of the hospital-level data

- ~7000 hospitals in US
- ~200 features:
  - Geographical identifiers: address, lat/long, county
  - Type of facility (e.g., short term acute care, critical access)
  - Urban/rural
  - # total beds, # Med-Surg beds, # ICU beds
  - ICU Occupancy rate
  - #Employees, #RNs
  - Total discharges, average length of stay, average daily census
  - Hospital overall rating
A bird’s-eye view of the county-level data

- COVID-19 cases and deaths (NYT and USAFacts)
- Demographics
  - Population, population density, age structure
- Health risk factors
  - Heart disease, stroke, respiratory disease, smoking, diabetes, overall mortality
- Socioeconomic risk factors
  - Social vulnerability index, unemployment, poverty, education, severe housing
- Social distancing and mobility
  - County-to-county work commute, change in distance traveled, government orders
- Other relevant data
  - Sample of flight itineraries in 2019, Kinsa temperature data, voting data
Our **data repository** can be found at the following link:

https://github.com/Yu-Group/covid19-severity-prediction
Now a little journey through cleaning the USAFacts COVID-19 cases/deaths data...

No data left behind!
A journey through cleaning the USAFacts COVID-19 cases/deaths data

Got the data from website

Some counties are duplicated.

Some cases/deaths cannot be allocated to a county.

Cumulative deaths counts sometimes decrease.

Are we done? 😐
Things got interesting when multiple data sources are available.

Got the data from nytimes!

different counties?

USAFacts do not cover all the counties.

Some counties changed their countyFIPS code.

NYTimes aggregated some counties together.
A journey through cleaning the USAFacts COVID-19 cases/deaths data

Things got interesting when multiple data sources are available.

- Got the data from nytimes!
- Different counties?
- Different death counts?
- Different time cut-off: CDC as of 4 PM ET
- How to deal with probable deaths?
- How to count one person who lives at A, tested at B, go to hospital at C? Flight travels?

Multiple data sources give us insights into the caveats of the data.
Additional challenges in data cleaning

- What is a “primary key”?
  - Use primary key to merge different sources of data together.
  - Ideally, key should be stable over time and no duplicates.

For county-level data
- County FIPS

⚠️ County FIPS can change over time (though this is rare)

For hospital-level data
- CMS Certification #

⚠️ Not all the hospitals have this number (e.g., Indian reservation hospitals)

For commute and county adjacency data
- Home County
- Work County
Additional challenges in data cleaning

- Missing data entries
  - Encoded as NAs, -1, 999, and more...
  - Meaning can depend on the data set

Diagram:

- **Missing Entry:** (NA, -1, 999)
  - County Health Rankings Data
    - SafeGraph mobility
      - No data
    - Mortality data
      - Suppressed due to privacy issues
Frequently overlooked questions

- Who is the audience or end user?
  - How to present the data to make it easily accessible by our modeling team, visualization team, and other researchers in the broader community
    - Clear documentation
    - Abridged version and unabridged version of the county-level data

Image Source: https://medium.com/@manjunath2137/how-to-create-the-right-target-audience-for-your-facebook-ad-9b5ed562b35f
Frequently overlooked questions

- **Who is the audience or end user?**
  - How to present the data to make it easily accessible by our modeling team, visualization team, and other researchers in the broader community
    - Clear documentation
    - Abridged version and unabridged version of the county-level data

---

Interactive Atlas of Heart Disease and Stroke - All Strokes (2014-2016)

- **Data source:** [https://www.cdc.gov/dhdsp/maps/atlas/index.htm](https://www.cdc.gov/dhdsp/maps/atlas/index.htm)
- **Last downloaded:** 04/02/2020
- **Data description:** County-level estimates of mortality rates per 100,000 (all ages, all races/ethnicities, both genders, 2014-2016) from all strokes (ICD10 codes: I60-I69)
- **Known data quality issues:** Data values within the table of "1" or "9999" indicate "Insufficient Data."
- **Short list of data columns:**
  - countyFIPS: county FIPS
  - StrokeMortality: estimate of mortality rate per 100,000 (all ages, all races/ethnicities, both genders, 2014-2016) from all strokes (ICD10 codes: I60-I69)
- **Notes:**
  - Data downloaded from the Interactive Atlas of Heart Disease and Stroke, a website developed by the Centers for Disease Control and Prevention, Division for Heart Disease and Stroke Prevention. [http://nccd.cdc.gov/DHDPAtlas](http://nccd.cdc.gov/DHDPAtlas)

---

List of columns - county level

**Identifying variables**

<table>
<thead>
<tr>
<th>Data variable</th>
<th>Description</th>
<th>Source data set</th>
</tr>
</thead>
<tbody>
<tr>
<td>countyFIPS</td>
<td>state-county FIPS Code</td>
<td>county_fips</td>
</tr>
<tr>
<td>STATEFP</td>
<td>state FIPS Code</td>
<td>county_popcenters</td>
</tr>
<tr>
<td>COUNTYFP</td>
<td>county FIPS Code</td>
<td>county_popcenters</td>
</tr>
<tr>
<td>CountyName</td>
<td>county name</td>
<td>county_fips</td>
</tr>
<tr>
<td>StateName</td>
<td>state abbreviation</td>
<td>county_fips</td>
</tr>
<tr>
<td>State</td>
<td>state name</td>
<td>county_latitude</td>
</tr>
</tbody>
</table>

**Data variables**

**Geographical identifiers**
Frequently overlooked questions

• What are the naming conventions and organization structure for data storage and preprocessing?
  ○ Improves accessibility for end users
  ○ Necessary to quickly integrate new members and volunteers
  ○ Best to set standards at the beginning
  ○ But this is very challenging because:
    ■ A good convention depends on the data we collect but we don’t know what data will be there.
    ■ Some data sets might change over time
Frequently overlooked questions

- Which file format?
  - txt, csv, pickle, xlsx
  - compressed versions
Frequently overlooked questions

- Which file format?
  - txt, csv, pickle, xlsx
  - compressed versions

- How to store the data?
  - Locally
  - GitHub
  - AWS
  - Google drive
The data team is at its best when working closely alongside everyone on the team

- In particular, modeling team depends on data team AND data team depends on modeling team
  - Determine what are relevant data sets
  - Iterative process between two teams to figure out how to clean the data
Overview of some useful tools

- Git commands: pull, push, merge conflicts
- Linux commands
  - shell commands
  - wget
    - Can easily download data from online source (including google drive)
  - cron jobs
    - To automatically update data, predictions, and visualizations daily
- AWS package
  - S3 buckets
- Google cloud package (update google sheet)
Summary: Data and code

Data repository:
https://github.com/Yu-Group/covid19-severity-prediction
Curating a COVID-19 data repository and forecasting county-level death counts in the United States

Nick Altieri¹⁺, Rebecca Barter¹, James Duncan⁶, Raaz Dwivedi², Karl Kumbier³, Xiao Li¹, Robert Netzorg², Briton Park¹, Chandan Singh⁺², Yan Shuo Tan¹, Tiffany Tang¹, Yu Wang¹, Bin Yu*¹, ², ⁴, ⁵, ⁶

¹Department of Statistics, University of California, Berkeley
²Department of EECS, University of California, Berkeley
³Department of Pharmaceutical Chemistry, University of California, San Francisco
⁴Chan Zuckerberg Biohub, San Francisco
⁵Center for Computational Biology, University of California, Berkeley
⁶Division of Biostatistics, University of California, Berkeley

April 29, 2020

⁺Authors ordered alphabetically. All authors contributed significantly to this work.
*Corresponding authors
This project was initiated on March 21, 2020, with the goal of helping aid the allocation of supplies to different hospitals in the U.S., in partnership with the non-profit Response4Life.
COVID-19 Data Repository and Severity Prediction

Yu Group
UC Berkeley Statistics, EECS, CCB

github.com/Yu-Group/covid19-severity-prediction

Website: covidseverity.com
Our team
from UC Berkeley Statistics/EECS and UCSF

N. Altieri
R. Barter
J. Duncan
R. Dwivedi
K. Kumbier
X. Li
R. Netzorg
B. Park
C. Singh
Y. Tan
T. Tang
Y. Wang

Many others at UC Berkeley, UCSF, Stanford, Northeastern, Univ. of Chicago, UW-Madison, ...
500k face shields in US by mid-May

- Santa Clara + Temple University Med
- in collaboration with GetUsPPE, AeroBridge, Maker Nexus, Synergy Mill maker space, R4L
- +65k to 25 recipients in 15 states in 2 weeks
- many more expected

Impact

Santa Clara + Temple University Med
in collaboration with GetUsPPE, AeroBridge, Maker Nexus, Synergy Mill maker space, R4L
+65k to 25 recipients in 15 states in 2 weeks
many more expected

Data Repository and Code Base

<table>
<thead>
<tr>
<th>Data variable</th>
<th>Description</th>
<th>Source data set</th>
</tr>
</thead>
<tbody>
<tr>
<td>countyFIPS</td>
<td>state-county FIPS Code</td>
<td>county_fips</td>
</tr>
<tr>
<td>STATEFP</td>
<td>state FIPS Code</td>
<td>county_popcenters</td>
</tr>
<tr>
<td>COUNTYFP</td>
<td>county FIPS Code</td>
<td>county_popcenters</td>
</tr>
<tr>
<td>CountyName</td>
<td>county name</td>
<td>county_fips</td>
</tr>
<tr>
<td>StateName</td>
<td>state abbreviation</td>
<td>county_fips</td>
</tr>
<tr>
<td>State</td>
<td>state name</td>
<td>county_idlongitude</td>
</tr>
</tbody>
</table>

Salesforce system

Impact

Santa Clara + Temple University Med
in collaboration with GetUsPPE, AeroBridge, Maker Nexus, Synergy Mill maker space, R4L
+65k to 25 recipients in 15 states in 2 weeks
many more expected

Data Repository and Code Base

<table>
<thead>
<tr>
<th>Data variable</th>
<th>Description</th>
<th>Source data set</th>
</tr>
</thead>
<tbody>
<tr>
<td>countyFIPS</td>
<td>state-county FIPS Code</td>
<td>county_fips</td>
</tr>
<tr>
<td>STATEFP</td>
<td>state FIPS Code</td>
<td>county_popcenters</td>
</tr>
<tr>
<td>COUNTYFP</td>
<td>county FIPS Code</td>
<td>county_popcenters</td>
</tr>
<tr>
<td>CountyName</td>
<td>county name</td>
<td>county_fips</td>
</tr>
<tr>
<td>StateName</td>
<td>state abbreviation</td>
<td>county_fips</td>
</tr>
<tr>
<td>State</td>
<td>state name</td>
<td>county_idlongitude</td>
</tr>
</tbody>
</table>

Salesforce system

Impact

Santa Clara + Temple University Med
in collaboration with GetUsPPE, AeroBridge, Maker Nexus, Synergy Mill maker space, R4L
+65k to 25 recipients in 15 states in 2 weeks
many more expected

Data Repository and Code Base

<table>
<thead>
<tr>
<th>Data variable</th>
<th>Description</th>
<th>Source data set</th>
</tr>
</thead>
<tbody>
<tr>
<td>countyFIPS</td>
<td>state-county FIPS Code</td>
<td>county_fips</td>
</tr>
<tr>
<td>STATEFP</td>
<td>state FIPS Code</td>
<td>county_popcenters</td>
</tr>
<tr>
<td>COUNTYFP</td>
<td>county FIPS Code</td>
<td>county_popcenters</td>
</tr>
<tr>
<td>CountyName</td>
<td>county name</td>
<td>county_fips</td>
</tr>
<tr>
<td>StateName</td>
<td>state abbreviation</td>
<td>county_fips</td>
</tr>
<tr>
<td>State</td>
<td>state name</td>
<td>county_idlongitude</td>
</tr>
</tbody>
</table>

Salesforce system

Impact

Santa Clara + Temple University Med
in collaboration with GetUsPPE, AeroBridge, Maker Nexus, Synergy Mill maker space, R4L
+65k to 25 recipients in 15 states in 2 weeks
many more expected

Data Repository and Code Base

<table>
<thead>
<tr>
<th>Data variable</th>
<th>Description</th>
<th>Source data set</th>
</tr>
</thead>
<tbody>
<tr>
<td>countyFIPS</td>
<td>state-county FIPS Code</td>
<td>county_fips</td>
</tr>
<tr>
<td>STATEFP</td>
<td>state FIPS Code</td>
<td>county_popcenters</td>
</tr>
<tr>
<td>COUNTYFP</td>
<td>county FIPS Code</td>
<td>county_popcenters</td>
</tr>
<tr>
<td>CountyName</td>
<td>county name</td>
<td>county_fips</td>
</tr>
<tr>
<td>StateName</td>
<td>state abbreviation</td>
<td>county_fips</td>
</tr>
<tr>
<td>State</td>
<td>state name</td>
<td>county_idlongitude</td>
</tr>
</tbody>
</table>

Salesforce system
Last Week: Curating a COVID-19 Data Repository

Covid 19 Cases/Deaths

USA FACTS

Risk Factors, Demographics, County-level Data

CDC
Centers for Disease Control and Prevention
CDC 24/7: Saving Lives, Protecting People™
Division for Heart Disease and Stroke Prevention

esri
COVID-19 GIS Hub

USDSS
UNITED STATES DIABETES SURVEILLANCE SYSTEM
Division of Diabetes Translation, CDC

County Health Rankings & Roadmaps
Building a Culture of Health, County by County

GHDx
Introducing the Unacast Social Distancing Scoreboard

United States Census Bureau
CMS.gov
Centers for Medicare & Medicaid Services

Hospital-level Data
(e.g., ICU beds, staff)

HRSA
Health Resources & Services Administration

ArcGIS Hub

Samuel Scarpino

The New York Times
This Week: Forecasting death counts
Ensemble different predictors

We combined many different prediction approaches
Ensemble predictors

1. Separate-county exponential model[1]

Ensemble predictors

We combined many different model approaches

1. Separate-county exponential model[1]

\[ E(\text{deaths}_t \mid t) = e^{\beta_0 + \beta_1 t} \]

Ensemble predictors

We combined many different model approaches

1. Separate-county **exponential** model\(^1\)

2. Separate-county **linear** model

Ensemble predictors

We combined many different model approaches

2. Separate-county **linear** model

\[
E[\text{deaths}_t | t] = \beta_0 + \beta_1 t
\]

Ensemble predictors
We combined many different prediction approaches

3. **Shared**-county exponential model
Ensemble predictors

We combined many different prediction approaches

3. Shared-county exponential model

\[
E(\text{deaths}_t \mid t) = e^{\beta_0 + \beta_1 \log(\text{deaths}_{t-1} + 1)}
\]
Ensemble predictors

We combined many different prediction approaches

3. **Shared**-county exponential model

4. **Shared**-county exponential + demographics model

+ Age
+ ICU Beds
+ # Hospitals
+ ....
Ensemble predictors

We combined many different prediction approaches

4. **Shared**-county exponential + **demographics** model

- County density and size
- County healthcare resources
- Demographic information

Deaths

Days

+ Age
+ ICU Beds
+ # Hospitals
+ ....
Ensemble predictors

We combined many different prediction approaches

5. Expanded Shared-county exponential model

- log(Cases)
- log(Cases in Neighboring counties)
- log(Deaths in neighboring counties)

+ Cases
+ Neighboring cases
+ Neighboring deaths
Combined Linear and Exponential Predictors (CLEP)

Calculate a **weighted average of the predictions**: higher weight to the models with better historical performance[2]

Combined Linear and Exponential Predictors (CLEP)

Calculate a **weighted average of the predictions**: higher weight to the models with better historical performance\(^2\)

\[
    w_t^m \propto \exp \left( -c \sum_{i=t_0}^{t-1} \ell(\hat{y}_i^m, y_i) \right)
\]

Combined Linear and Exponential Predictors (CLEP)

Calculate a **weighted average of the predictions**: higher weight to the models with better historical performance\(^2\)

\[
w^m_t \propto \exp \left( -c(1 - \mu) \sum_{i=t_0}^{t-1} \mu^{t-i} \ell(\hat{y}^m_i, y_i) \right)
\]

Combined Linear and Exponential Predictors (CLEP)

A smaller combination performed better in practice:

![Separate-county linear predictor]

![Expanded Shared-county exponential predictor]

Calculate a **weighted average of the predictions**: higher weight to the models with better historical performance\(^2\)

Our county-level 7-day predictive performance

Focusing on 6 of the worst-affected counties

*Based on 4/8 data
Our county-level 7-day predictive performance

Takeaway:
The 7-day forecasted predictions are fairly accurate

"Actual deaths": recorded deaths by a given day
Ongoing Work:

County Matching
Matching Counties:

Find similar counties and use these to predict trajectory
Prediction Intervals:

How confident should we be about our predictions?
Prediction Intervals:

How confident are we with the prediction for April 25?
Prediction Intervals:

How confident are we with this prediction for April 25?

Use *past experience* to determine confidence in new predictions.
Prediction Intervals:

- Recorded deaths
- Prediction for Apr 25
- Previous predictions

Previous 5-day-ahead prediction errors (%)
- Apr 16: 3.3%
- Apr 17: 6.5%
- Apr 18: 9.6%
- Apr 19: 12.6%
- Apr 20: 5.5%
- Apr 25: ?

Take the max
Prediction Intervals:

Predicted range of error
Apr 25  [-12.6%, 12.6%]

Recorded deaths
Prediction interval for Apr 25
Santa Clara, CA
Prediction Intervals:

Predicted range of error
Apr 25  [-12.6%, 12.6%]

Actual error:
Apr 25  8.8%
Maximum (absolute) error prediction intervals (MEPI)

Step 1: Find normalized error of our predictor in the past.

$$\Delta_T := \frac{|y_T - \hat{y}_T|}{|\hat{y}_T|}.$$ 

Step 2: Find maximum error of past 5 days.

$$\Delta_{\max} := \max_{0 \leq j \leq 4} \Delta_{t-j}.$$ 

Step 3: Define prediction intervals.

$$\hat{\text{PI}}_{t+k} := \left[ \max \{ \hat{y}_{t+k} (1 - \Delta_{\max}), y_t \}, \hat{y}_{t+k} (1 + \Delta_{\max}) \right]$$

Can be applied to any ML model!
Connection to conformal inference\cite{1,2}

General conformal inference recipe:  
95% percentile of all past errors 
MEPI:  
max of past 5 errors

Connection to conformal inference\cite{Shafer2008, Vovk2005}

General conformal inference recipe: 95% percentile of all past errors
MEPI: \text{max} of past 5 errors

If the errors \( \{\Delta_{t+k}, \Delta_t, \Delta_{t-1}, \Delta_{t-2}, \Delta_{t-3}, \Delta_{t-4}\} \) are exchangeable, then

\[
\mathbb{P}\left(y_{t+k} \in \hat{\Pi}_{t+k}\right) = \mathbb{P}(\Delta_{t+k} < \Delta_{\text{max}}) = 1 - \mathbb{P}(\Delta_{t+k} = \Delta_{\text{max}}) = \frac{5}{6} \approx 0.83.
\]

\cite{Shafer2008, Vovk2005}


Empirical performance of MEPI

Evaluation period: March 28--April 27. Only include days since the county has 10 deaths. Having a normalized length of 0.8 means the PI is roughly \((0.6\ \hat{y}_{t+k}, 1.4\ \hat{y}_{t+k})\).
Severity Index

A score* for each hospital based on:

1. Predicted cumulative deaths
2. Predicted daily deaths

* county level predicted deaths are distributed to hospitals proportional to #employees
Collaborating with the Center for Spatial Data Science (CSDS) at University of Chicago to add our predictions and severity index to the U.S. COVID-19 Atlas.
Curating a COVID-19 data repository and forecasting county-level death counts in the United States

Nick Altieri¹,‡, Rebecca Barter¹, James Duncan⁶, Raaz Dwivedi², Karl Kumbier³, Xiao Li¹, Robert Netzorg², Briton Park¹, Chandan Singh*², Yan Shuo Tan¹, Tiffany Tang¹, Yu Wang¹, Bin Yu*¹, 2, 4, 5, 6

¹Department of Statistics, University of California, Berkeley
²Department of EECS, University of California, Berkeley
³Department of Pharmaceutical Chemistry, University of California, San Francisco
⁴Chan Zuckerberg Biohub, San Francisco
⁵Center for Computational Biology, University of California, Berkeley
⁶Division of Biostatistics, University of California, Berkeley

April 29, 2020

‡Authors ordered alphabetically. All authors contributed significantly to this work.
*Corresponding authors

This project was initiated on March 21, 2020, with the goal of helping aid the allocation of supplies to different hospitals in the U.S., in partnership with the non-profit Response4Life.
Thank you!
Assign severity index to hospital based on predicted cumulative deaths
Surge Index

A score for each hospital based on:

(Estimated # ICU beds needed*) - (# ICU beds available)

*2x predicted cumulative number of deaths
Assign surge index based on #ICU beds
Volunteer Team: Local News and Emerging Hotspots

10-12 volunteers find local news and gather hard to find on-the-ground data

<table>
<thead>
<tr>
<th>Hospital</th>
<th>Severity</th>
<th>Deaths as of April 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beaumont Health (Ohio County) Michigan</td>
<td>3</td>
<td>328</td>
</tr>
</tbody>
</table>

Compare collected data against predicted severity
Other works -- at state or country level

Curve fitting epi. Modeling (e.g. IHME -- dominant in the US)

Compartment epi. modeling (e.g. ICL -- dominant in UK and Europe)

Both have parameters that are tuned based on data mostly from other countries

No comparisons yet on prediction with US data ...
Current & Future Directions

- Continue to update predictors
- Look at long-term trajectories
- Incorporate epidemiology models
- Concentrate on rural areas
Current situation: Exponential growth of COVID-19

Cumulative number of cases by county

- 12,800 Queens County, NY
- 10,200 Kings County, NY
- 6,900 Bronx County, NY
- 6,100 New York County, NY
- 3,200 Wayne County, MI
- 2,300 King County, WA
- 1,500 Orleans County, LA

Focusing on 6 of the worst-affected counties

*Cases* and "deaths" in this presentation are recorded cases and deaths. Data source: https://usafacts.org.

*Based on 3/30 data*
Current situation: Exponential growth of COVID-19

Cumulative number of deaths by county

Focusing on 6 of the worst-affected counties

*Based on 3/27 data
Goal: **Predict** COVID-19 at the county level
Our goal: **predict** COVID-19 at the county level

Cumulative number of **deaths** by county

*Based on 3/27 data*
Goal: **Predict** COVID-19 at the county level

Cumulative number of **deaths** by county

*Based on 3/27 data*
Goal: **Predict** COVID-19 at the county level

Cumulative number of **deaths** by county

Predicted 3-day death trends

*Based on 3/27 data*
Goal: **Predict** COVID-19 at the county level

Cumulative number of **deaths** by county

*Based on 3/30 data*