Estimation of COVID-19 prevalence rates using EM algorithm

June 2025

Jungsik Noh

UTSouthwestern
Medical Center

Outline

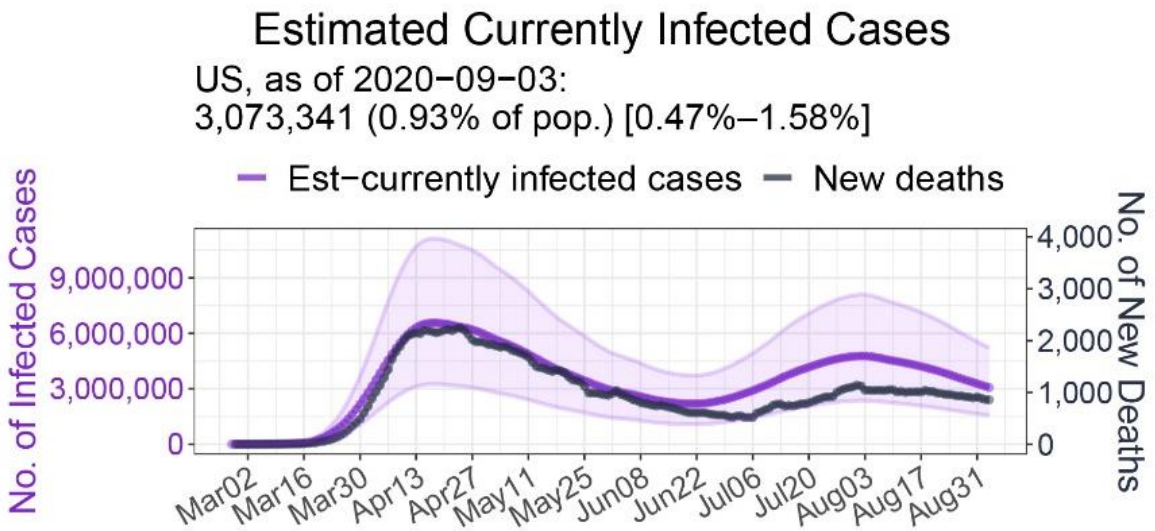
- COVID-19 data in 2020
- Data science perspective
- Expectation-maximization algorithm

PLOS ONE

RESEARCH ARTICLE

Estimation of the fraction of COVID-19 infected people in U.S. states and countries worldwide

Jungsik Noh*, Gaudenz Danuser



Back to dismal 2020…

Mar 31, 2020

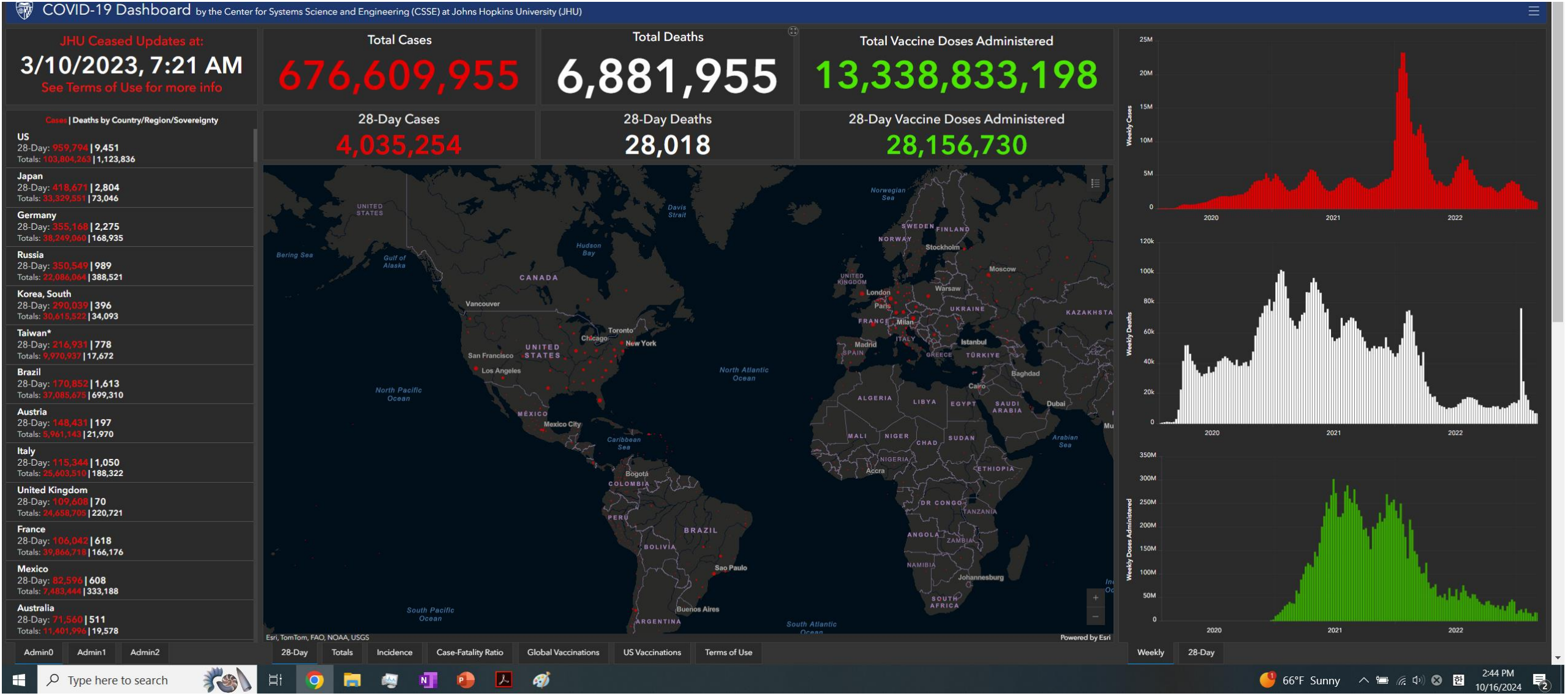
The New York Times

N.Y.C.'s 911 System Is Overwhelmed. 'I'm Terrified,' a Paramedic Says. - The New York Times

Visit ›

Data Science for the Pandemic (Data collection)

- Center for Systems Science and Engineering (CSSE) at Johns Hopkins University



Data Science for the Pandemic (Prediction)

May 14, 2020

- INSTITUTE FOR HEALTH METRICS AND EVALUATION (IHME)

Worldwide COVID-19 (12-12-2019 ~ 8-13-2020)

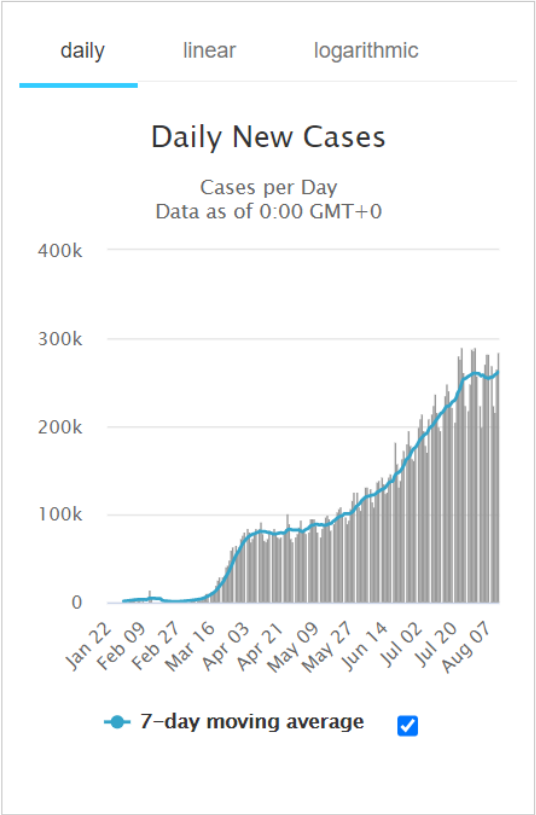
Coronavirus Cases:

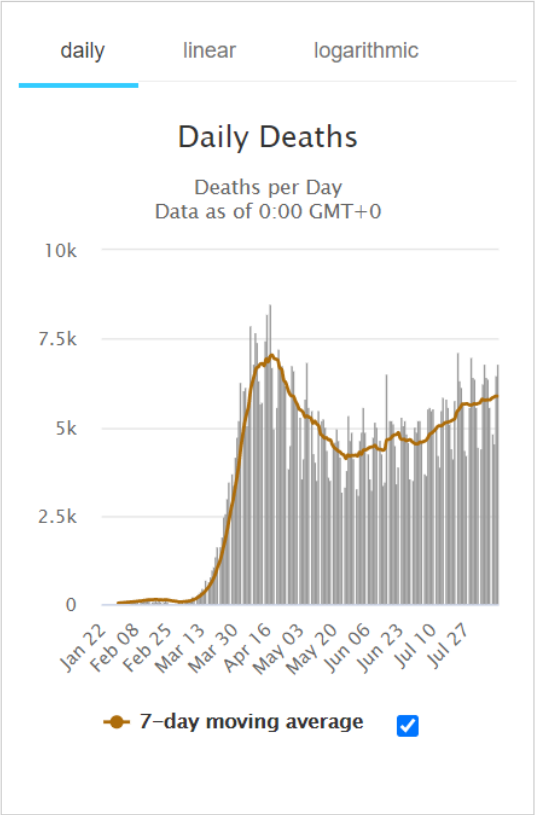
20,963,071

view by country

Deaths:

750,212





What is known (data)

- Daily (reported) new infected cases
- Daily (reported) deaths
- Daily test positivity rates
- Daily hospitalization rates
- …

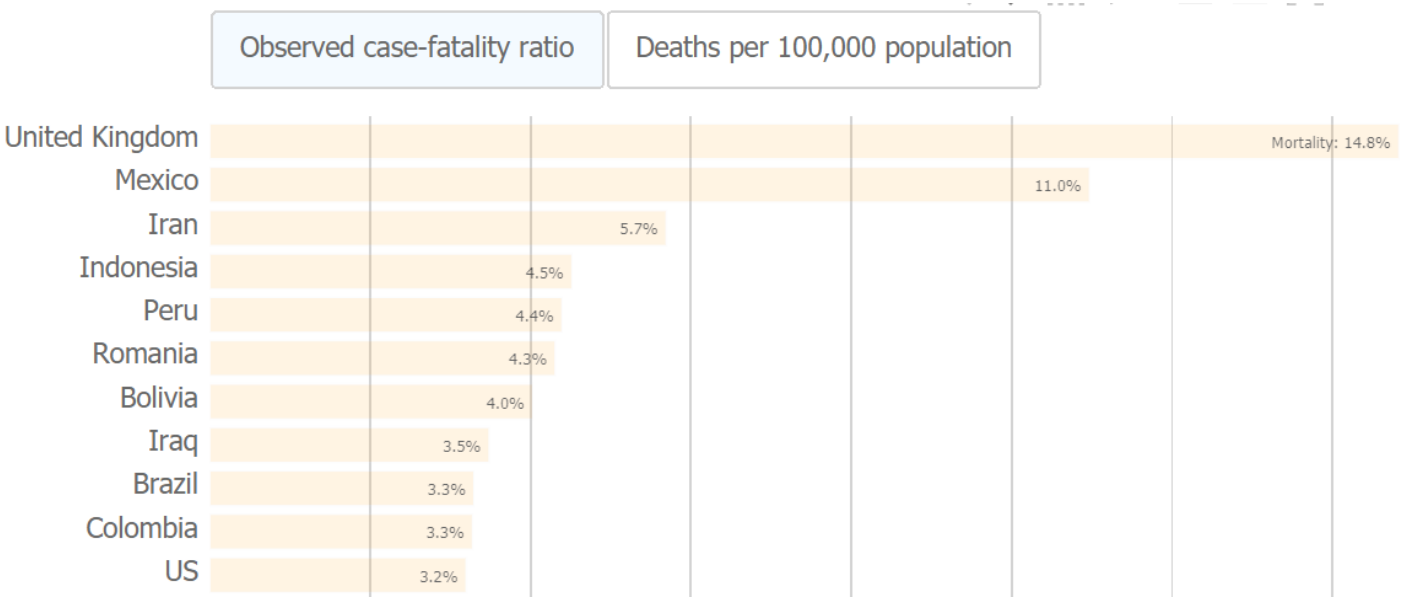
What is unknown (parameters)

- Key pandemic parameters are unknown
  - Infection-Fatality-Rate (IFR)
  - Time period between:
    - Infection to symptom-onset
    - Infectious period
    - Infection to death
    - Infection to recovery
  - …
  - *(Personally)* ***# of currently infected people in Collin county***

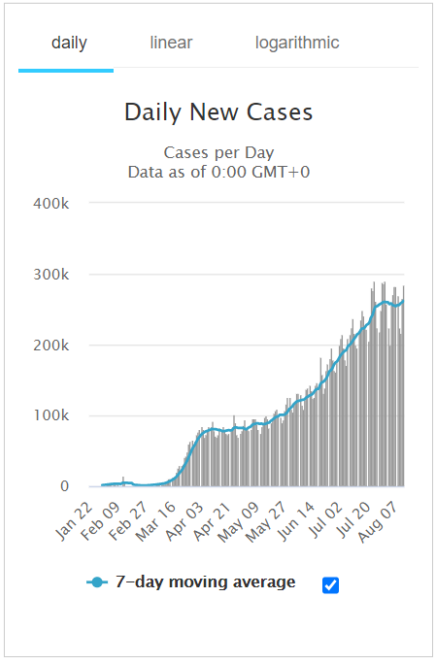
*https://www.worldometers.info/coronavirus/, CSSE

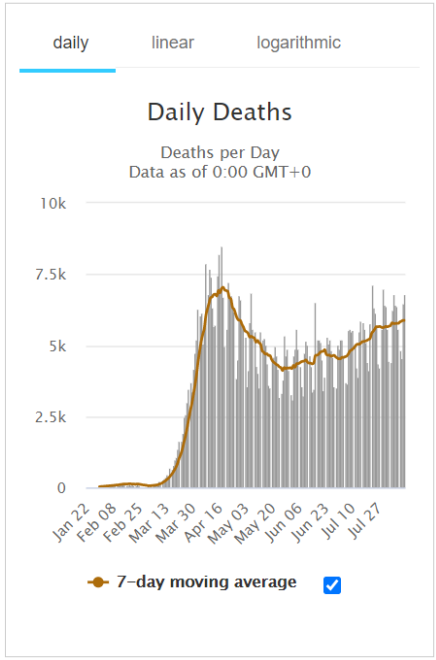
Numbers did not make sense: Case-Fatality-Rate (as of Aug 2020)

Spatial variation



Temporal variation





Under-ascertainment due to Low Testing Capacity

*https://coronavirus.jhu.edu, worldometers.info

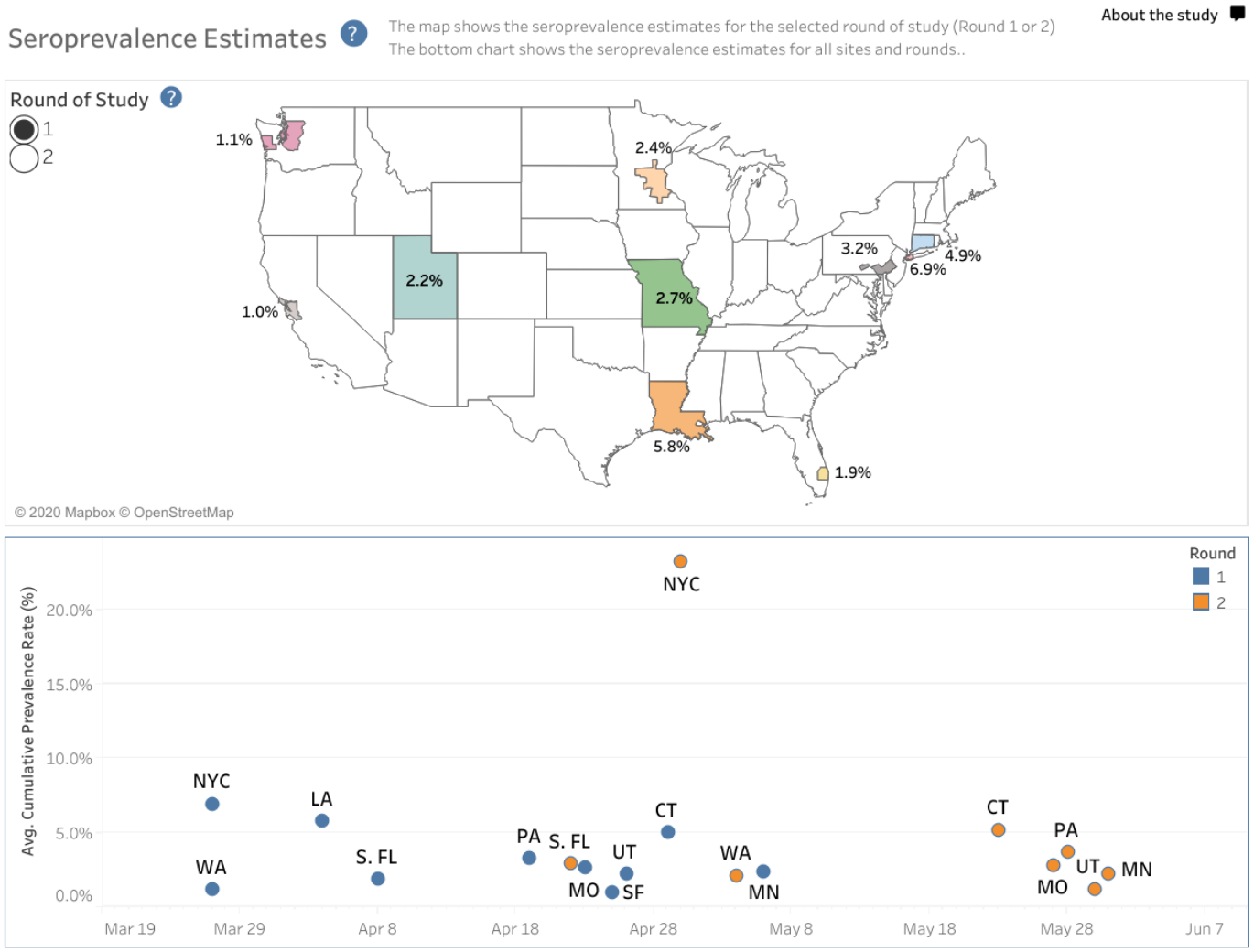
Early evidences of substantial undocumented infections

- **Ascertainment rate**

= (# of confirmed cases) /
(# of actual cases)

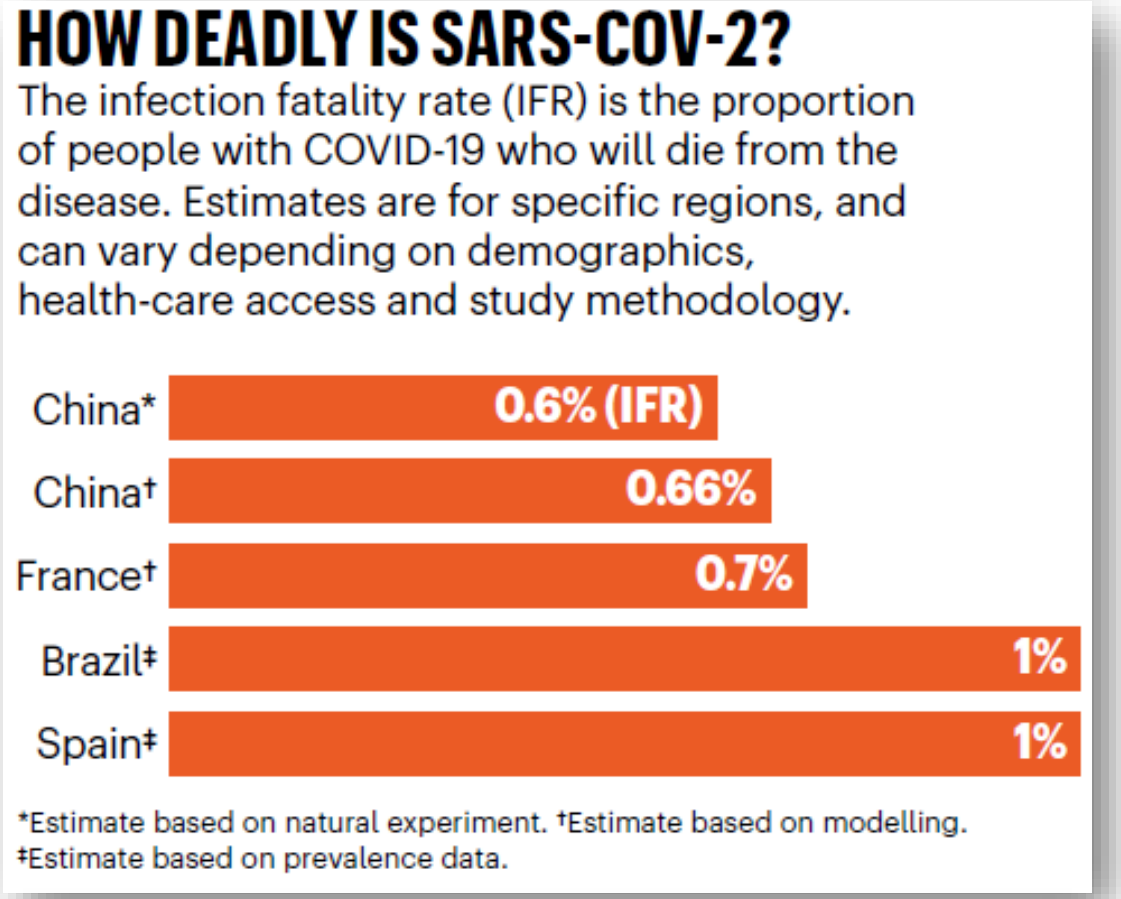
- Using computational modeling, Li et al. (Science, 2020) estimated that only 14% of all infections in China were detected or laboratory-confirmed in January 2020.

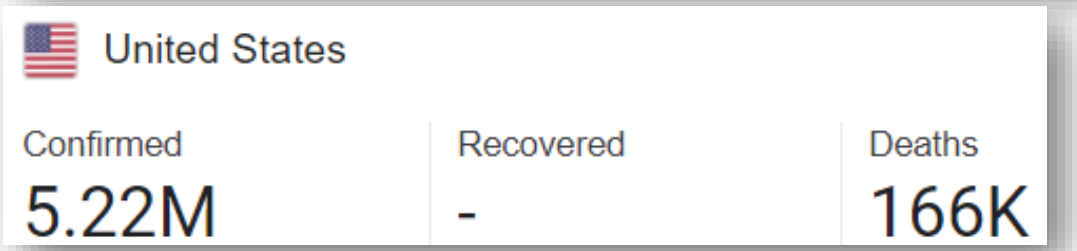
- From seroprevalence studies, CDC reported that the ascertainment rates were:
  - 4.2% in MO
  - 8.9% in NY
  - 16.7% in CT until March or April 2020



Limitations: A full list of limitations interpreting this data can be found at https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/commercial-lab-surveys.html#interpreting-serology-results

Under-reported Infections and Infection-Fatality-Rate (IFR)





- Verity et al. (Lancet, March 2020):
  - key pandemic parameters based on early pandemic data

Infection prevalence estimated from international Wuhan residents who were repatriated to their home countries

+

Age-stratified confirmed cases (>40K) in China during January

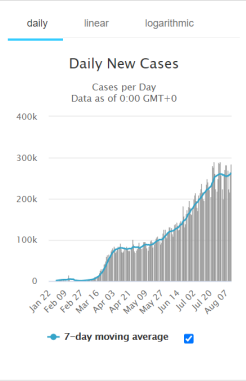
→ *"Our estimated overall infection fatality ratio for China was **0.66% (0.39–1.33)**"*

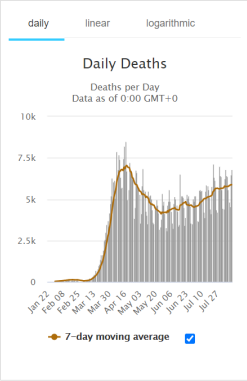
- Infection to death: 18 days
- Infection to recovery: 25 days

*Nature News in Focus, June, 2020

Formulate a data-science problem

Datasets





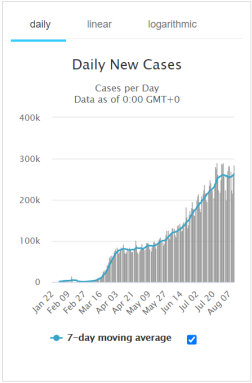
- **IFR: 0.66% (*0.39–1.33*)**
- **Infection to death: 18 days**
- **Infection to recovery: 25 days**

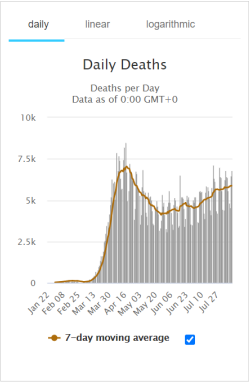
Goal of a DS project

# of currently infected people in Collin county

Formulate a data-science problem

Datasets

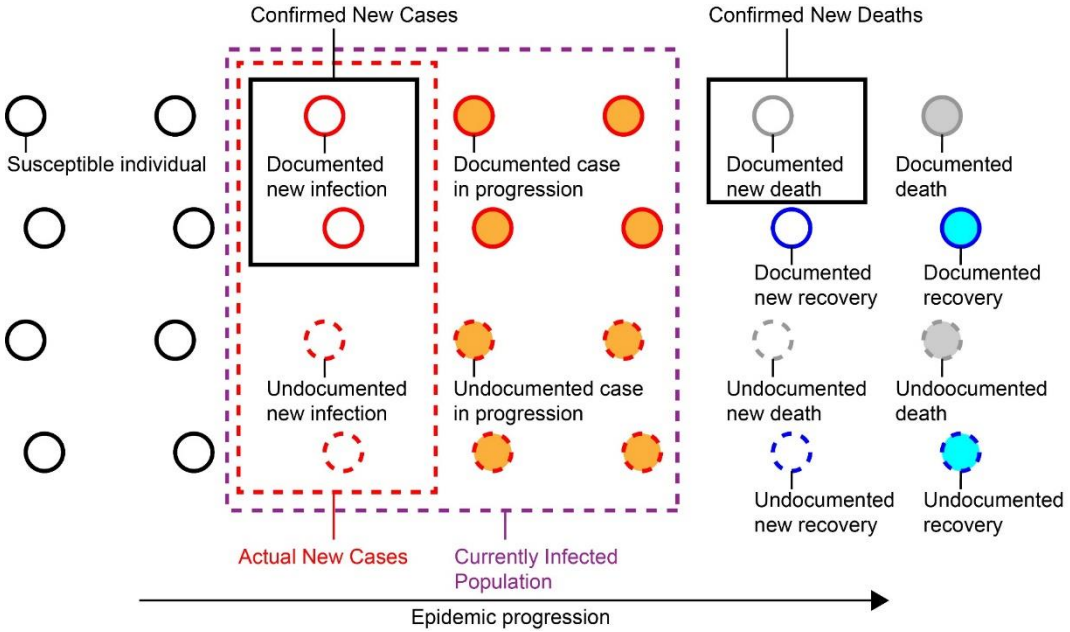




- IFR: 0.66% (*0.39–1.33*)
- Infection to death: 18 days
- Infection to recovery: 25 days

Basic equation in epidemiology

Susceptible(t) → Infected(t) → Recovered(t) or Death(t)



Goal of a DS project

# of currently infected people in Collin county

Total Infected(t)
= Currently Infected(t)
+ Total Death(t)
+ Total Recovered(t)

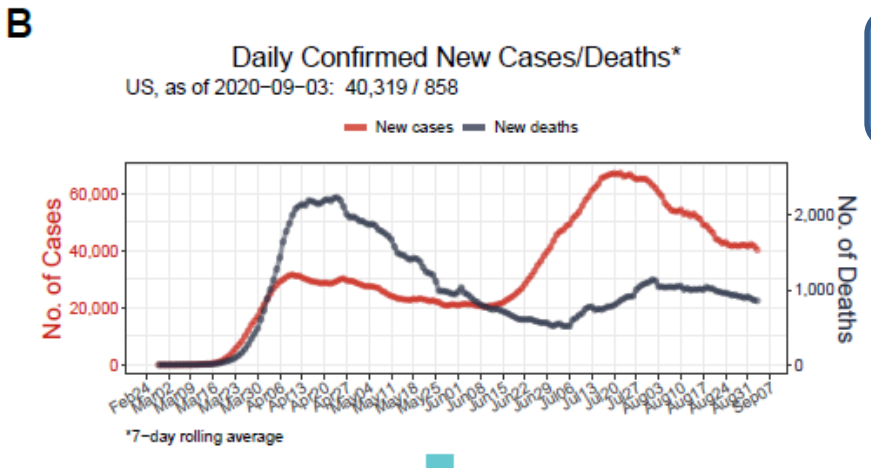
Assumption

(A1) Assume that # of deaths is accurate.
(A2) The wide uncertainty of the IFR estimate is expected to cover the true IFRs of many countries and U.S. states.

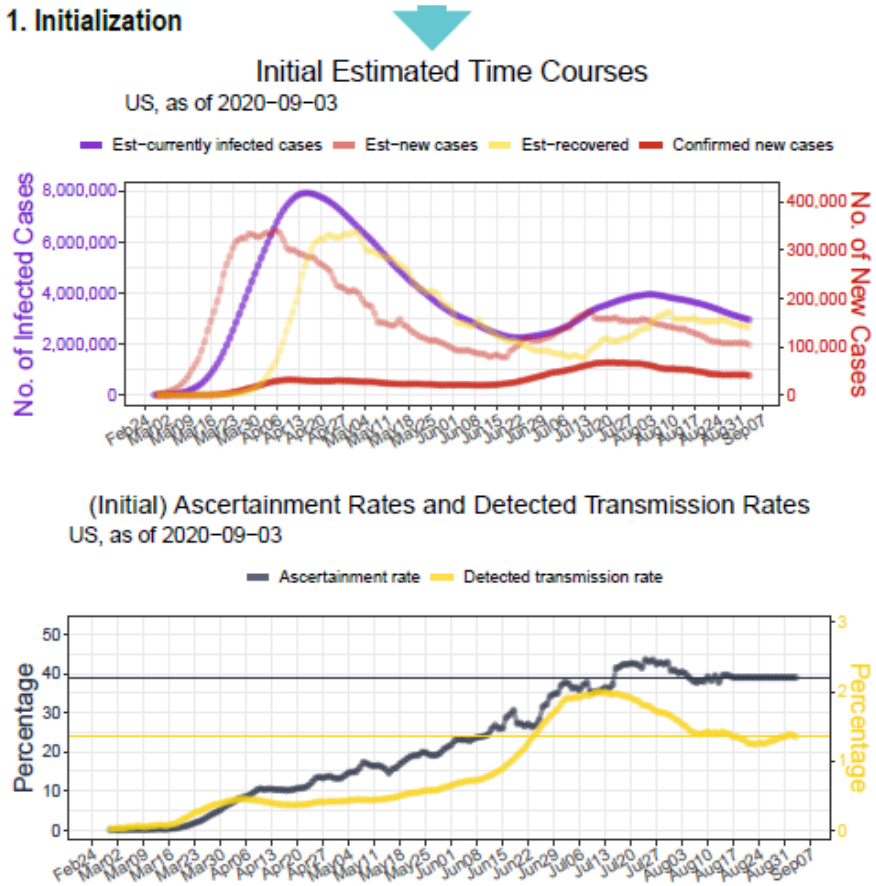
Daily cases and deaths

1. **Actual cases(t – 18) = 1/0.66 * Death(t)**
2. **Recovered(t+7) = (1/0.66 – 1) * Death(t)**
3. **Currently Infected(t)**
   = Total actual cases(t) – Total Death(t) – Total Recovered(t)

Initial latent time series



IFR: 0.66% (*0.39–1.33*)
Infection to death: 18 days
Infection to recovery: 25 days

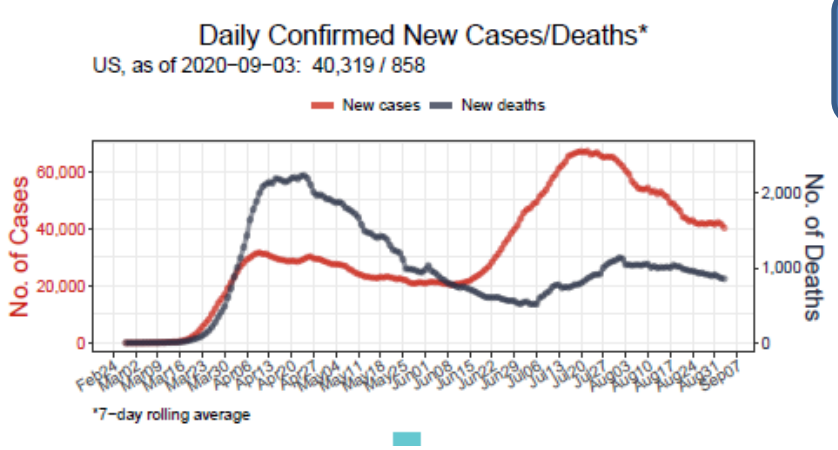


Initial estimates

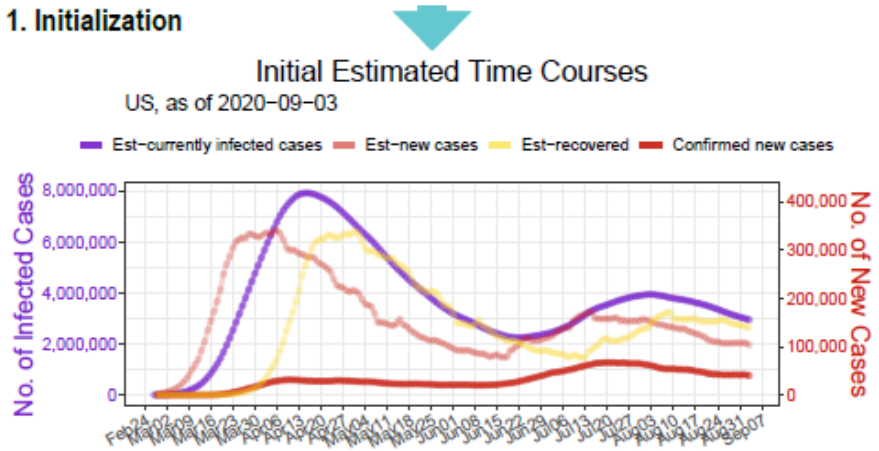
1. **Actual cases(t – 18) = 1/0.66 * Death(t)**
2. **Recovered(t+7) = (1/0.66 – 1) * Death(t)**
3. **Currently Infected(t) = Total actual cases(t) – Total Death(t) – Total Recovered(t)**
4. **Daily ascertainment rate(t) = Confirmed cases(t) / Actual cases(t)**

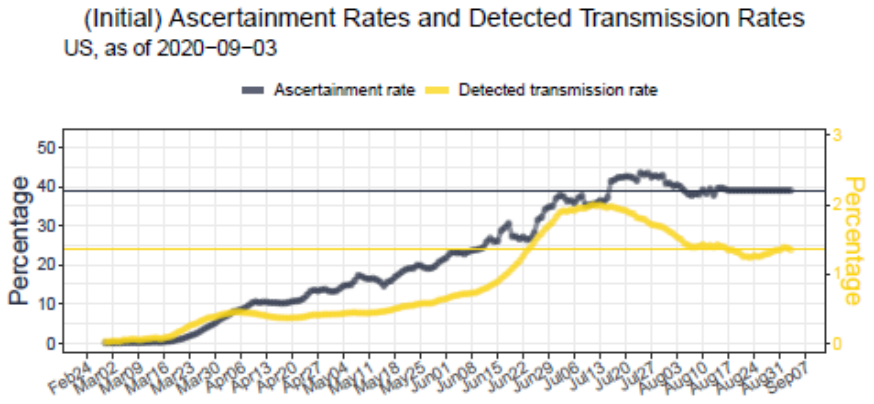
Initial latent time series

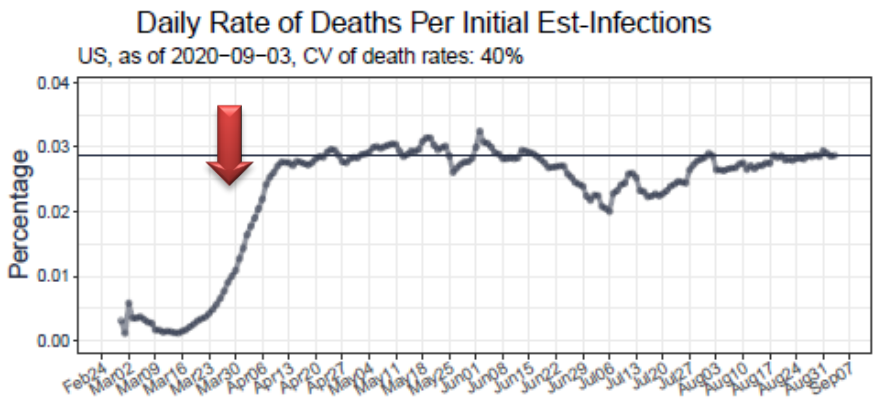
B



IFR: 0.66% (*0.39–1.33*)
Infection to death: 18 days
Infection to recovery: 25 days





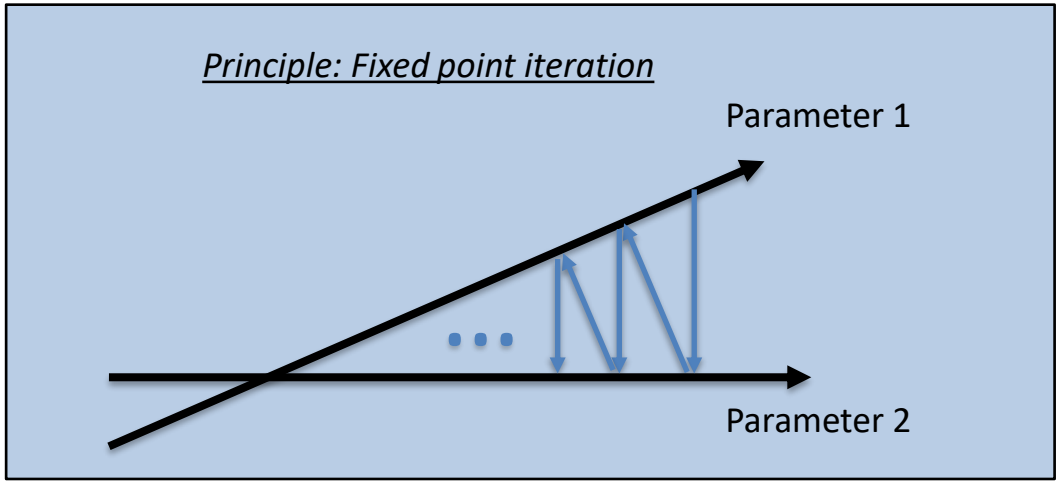


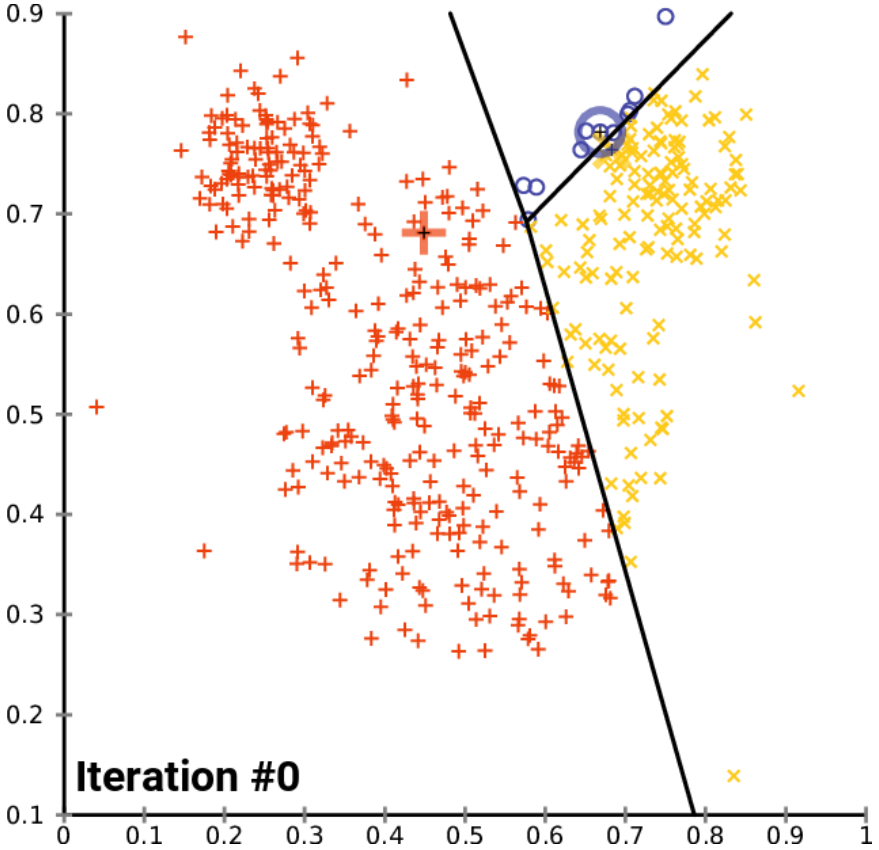
Initial estimates

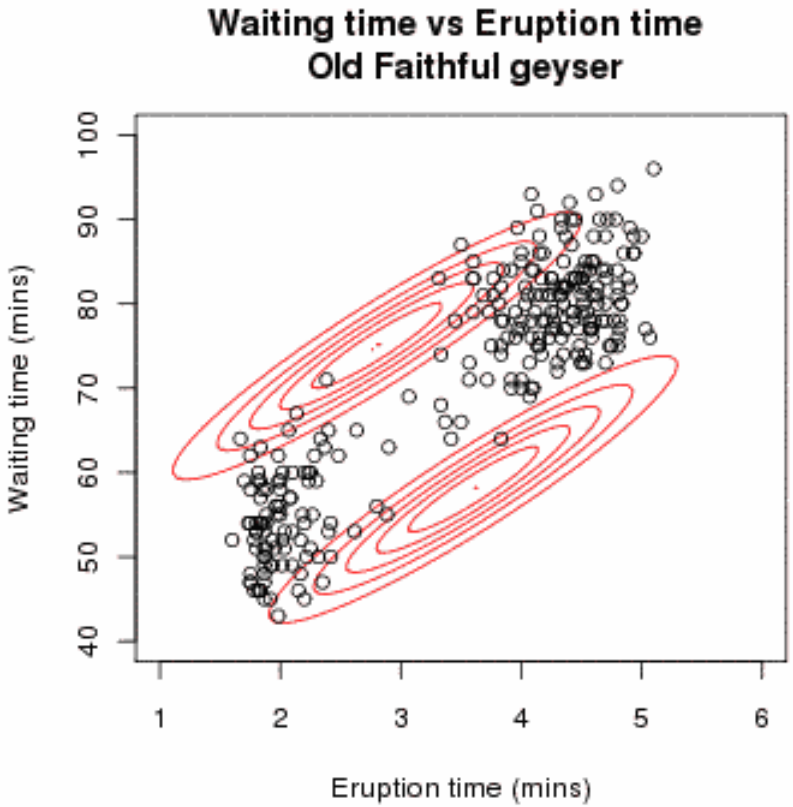
1. Actual cases(t – 18) = 1/0.66 * Death(t)
2. Recovered(t+7) = (1/0.66 – 1) * Death(t)
3. Currently Infected(t) = Total actual cases(t) – Total Death(t) – Total Recovered(t)
4. Daily ascertainment rate(t)
   = Confirmed cases(t) / Actual cases(t)
5. Detected transmission rate(t)
   = Confirmed cases(t) / Currently Infected(t-1)
6. Rate of Deaths per Currently Infected(t)
   = Deaths(t) / Currently Infected(t-1)

K-means and EM for GMM

- Both are techniques to find clusters in data points.
- K-means is a hard-thresholding.
- EM for GMM is a soft-thresholding.







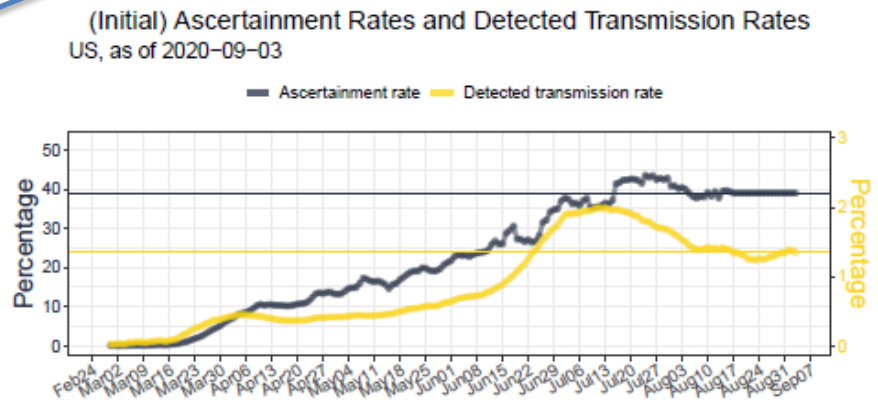
*By Chire, https://commons.wikimedia.org/w/index.php?curid=6319946

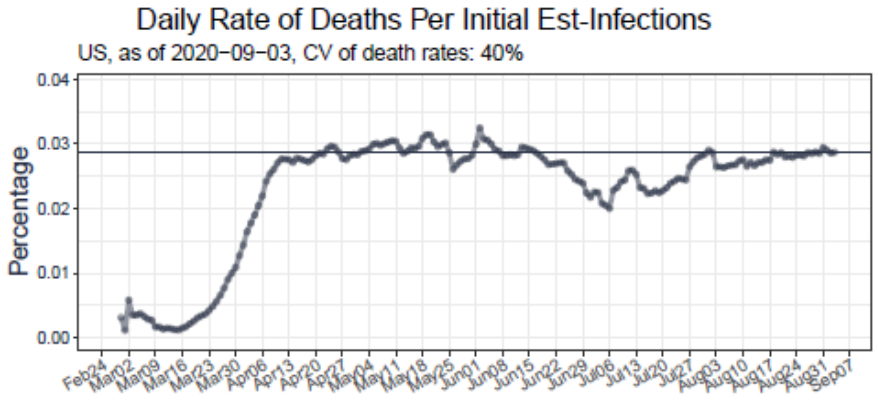
EM iterations with COVID time series

A

Currently Infected Cases: $I^{(k)}$ → Detected Transmission Rates: $DT^{(k)}$ → Spline-smoothed Detected Transmission Rates ($DT^{(k)}(h)$) and Ascertainment Rates ($A^{(k)}(h)$) → Fit a regression function from $DT^{(k)}(h)$ to $A^{(k)}(h)$ → Ascertainment Rates: $A^{(k+1)}$ → Actual Daily New Cases: $AC^{(k+1)}$ → Currently Infected Cases: $I^{(k+1)}$

$$A_t^{(0)}(h) = \beta_0 + \beta_1 DT_t^{(0)}(h) + \beta_2 \left\{DT_t^{(0)}(h)\right\}^2 + \epsilon_t, \text{ for } t = 1, 2, \ldots, T$$

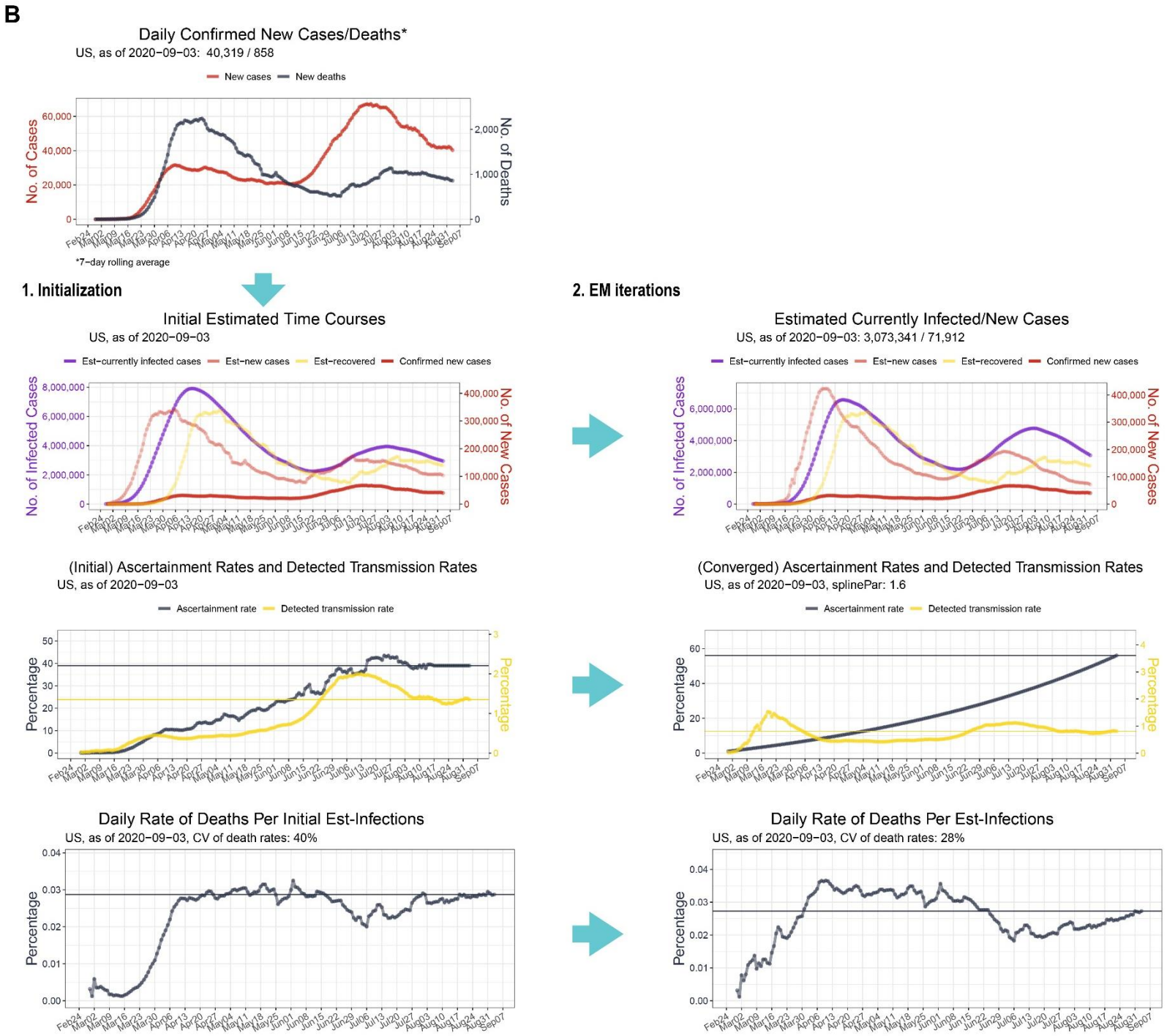




Initial estimates

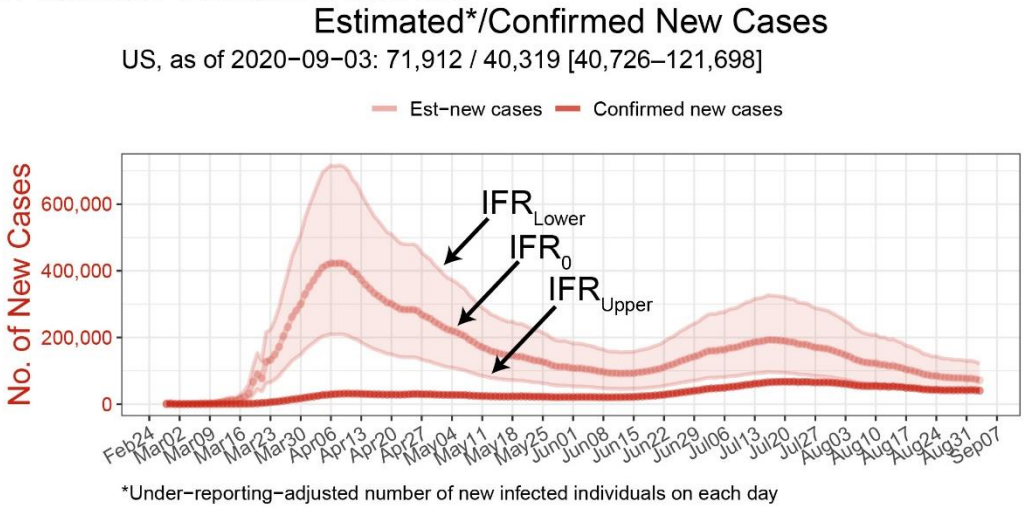
1. **Actual cases(t – 18) = 1/0.66 * Death(t)**
2. **Recovered(t+7) = (1/0.66 – 1) * Death(t)**
3. **Currently Infected(t) = Total actual cases(t) – Total Death(t) – Total Recovered(t)**
4. **Daily ascertainment rate(t)**
   A_t **= Confirmed cases(t) / Actual cases(t)**
5. **Detected transmission rate(t)**
   DT_t **= Confirmed cases(t) / Currently Infected(t-1)**
6. **Rate of Deaths per Currently Infected(t)**
   = Deaths(t) / Currently Infected(t-1)

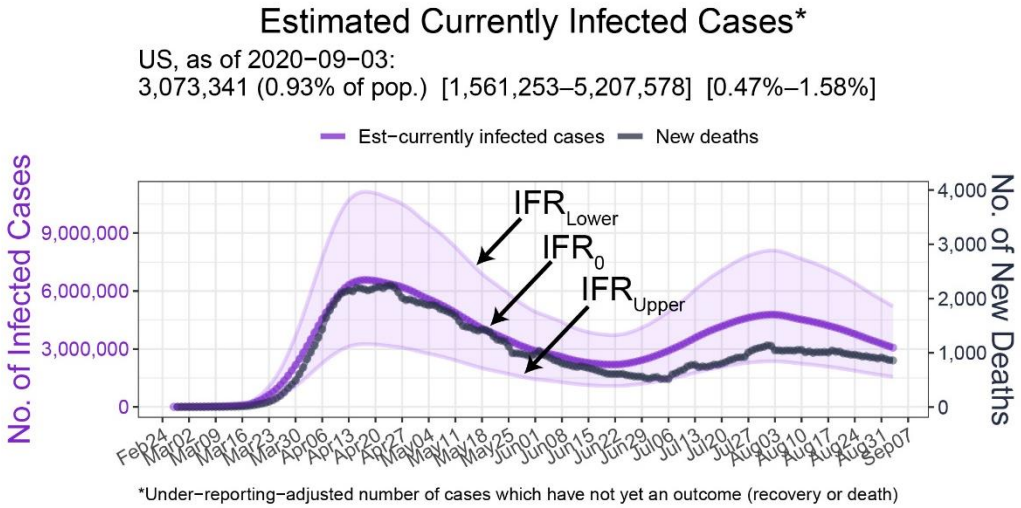
Converged estimates



Confidence intervals

3. Calculate Confidence Intervals

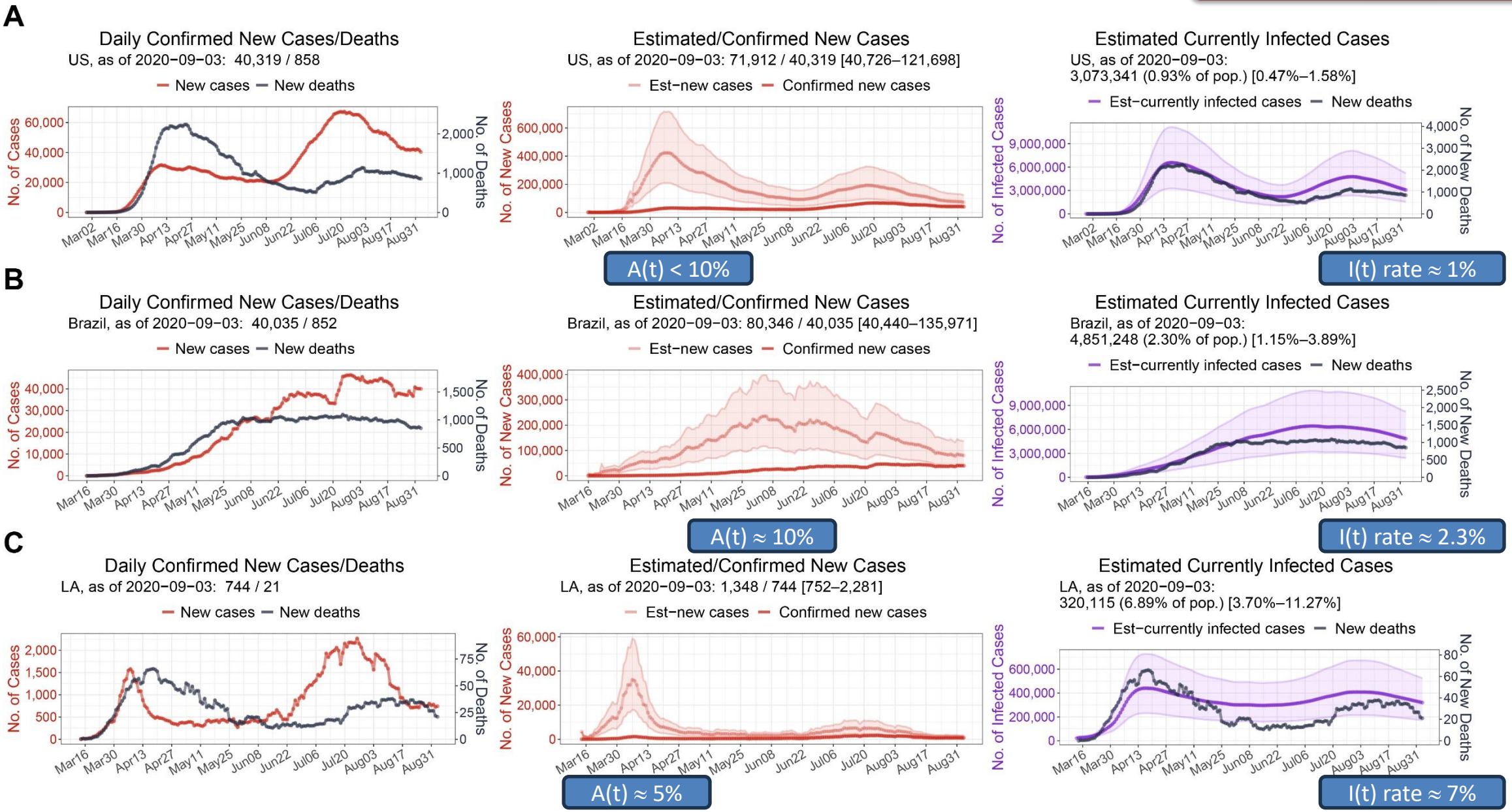




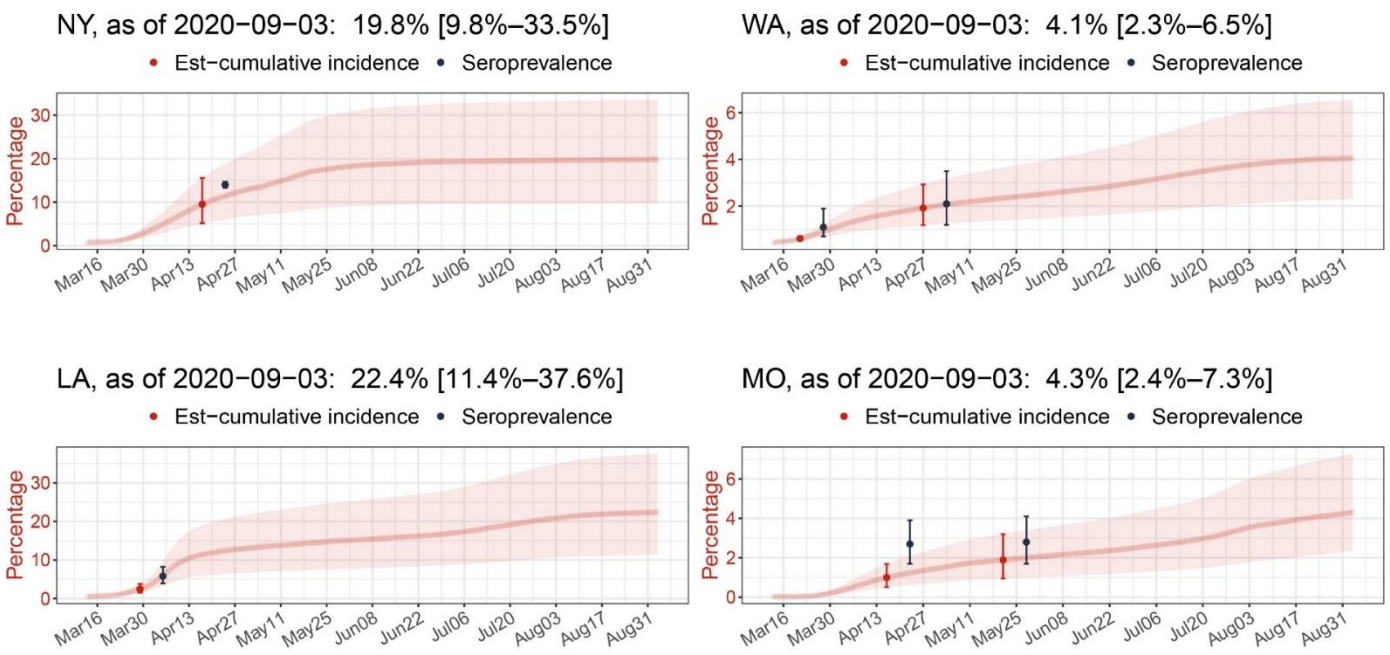
- **IFR: 0.66% (*0.39–1.33*)**
- **Infection to death: 18 days**
- **Infection to recovery: 25 days**

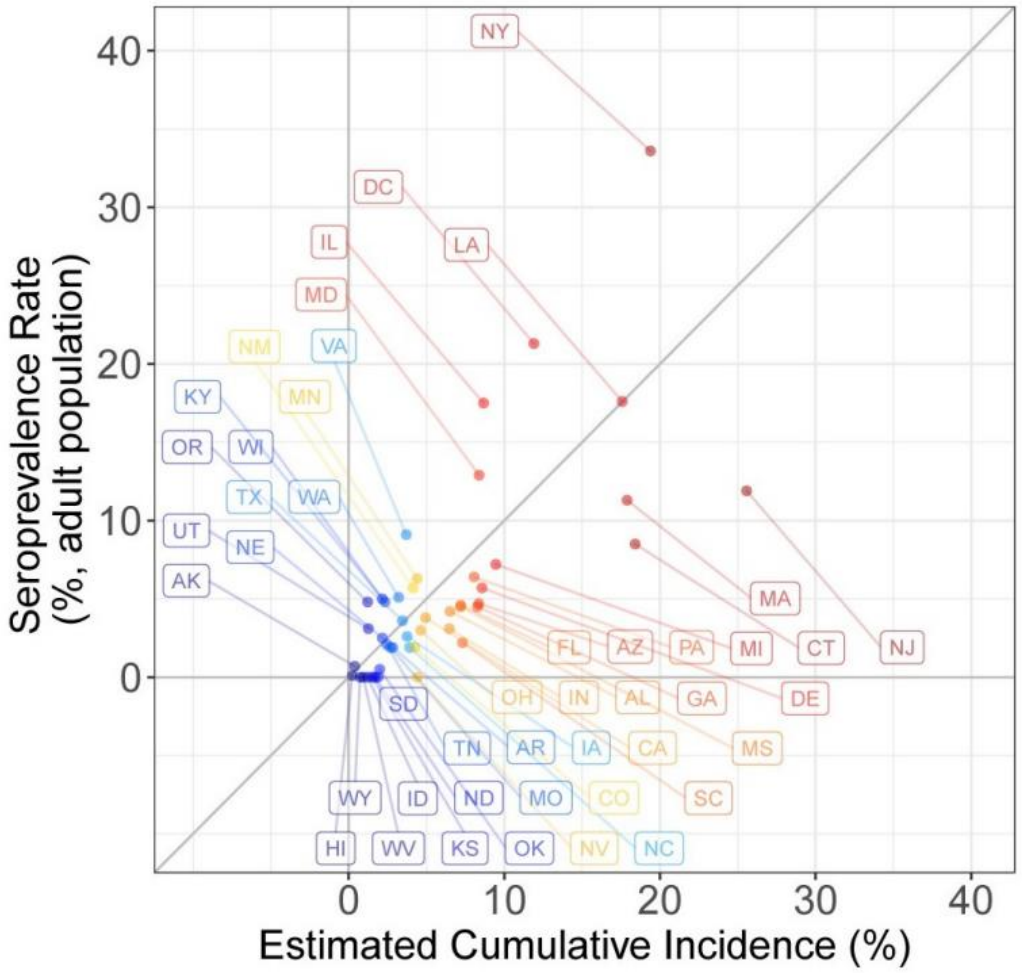
Actual cases in US, Brazil and Louisiana

Goal of a DS project



Validation using seroprevalence survey data





CDC survey

Nationwide blood tests with dialysis patients (n = 28,503) during July 2020

Summary

1. Goal of a DS project

2.