Coping with Heterogeneity and Uncertainty of COVID-19 Datasets
Issues, Approaches, and Consequences of the COVID-19 Crisis

Ambuj K Singh
Yu-Xiang Wang

University of California
Santa Barbara

April 28, 2020
Previous presentations

- (April 14) SARS-CoV2: The Virus and the Disease (Carolina Gonzalez and Lynn Fitzgibbons)
- (April 21) Current Epidemiological Models: Scientific Basis and Evaluation (Francesco Bullo)

Thanks to doctors, nurses, first responders, and everyone in the hospital/healthcare/biomedical ecosystem. And thanks to everyone else for adapting to the new lifestyle.
Lessons from history

- (1854) Cholera outbreak in London
- (1918–19) Spanish Flu

How have data and models informed decision making in past epidemics?
In August of 1854, Soho, a suburb of London, was hit hard by a terrible outbreak of cholera.

Dr. John Snow had long believed that water contaminated by sewage was the cause of cholera. A challenge to the theories of disease transmission then.

Through ‘contact-tracing’ of water (homes, restaurants, coffee shops, breweries), he established that the source was the pump on Broad Street. Shown is a memorial in his honor.
The most severe pandemic in recent history.

About 500 million people or one-third of the world’s population became infected with this virus. The number of deaths was estimated to be at least 50 million worldwide with about 675,000 in the US.

Hatchett et al analyzed Non-Pharmaceutical Interventions (NPI) in 17 US cities\(^1\).

Findings support the hypothesis that rapid implementation of multiple NPIs can significantly reduce disease transmission, but that spread will be renewed upon relaxation of such measures.

CDC model predicted 500K–1.4 million infections. A year later there were about 25K Ebola cases, including 10.3K deaths. (Scientific American, Dec 8 2014)

Prediction was more than 50 times worse. Why?
- Errors in deterministic modeling
- Difficult to model behavioral change.
- Of course, there were many good predictions post-hoc.
  “It is difficult to make predictions, especially about the future.”

“Ebola defied the prophets of doom. It never went airborne, and its economic effects were less painful than expected. Being wrong rarely feels this good. But it will be harder to catch the world’s attention next time.” Economist, Feb 2015.

---

Talk outline

- Historical context of models and interventions
- COVID-19 data
  - Heterogeneous
  - Uncertain
  - Dynamic
- Measuring the effect of NPIs (Non-Pharmaceutical Interventions)
- How do current models use available data?
- Contact tracing
- Uncertainty and bias in data and models
- Case study: fitting SIR model to Santa Barbara Data
- Challenges and opportunities
COVID-19 data is heterogeneous, uncertain, and dynamic

- **Heterogeneous:** Risk factors vary at the individual level (presence of comorbidity). Different countries (and different regions within a country) have different outcomes based on geography, culture, and the human element. Many confounding variables.

- **Uncertain:** Testing error of biomedical tests, uncertainty of transmission fraction, contact rate, recovery rate, ‘asymptomatic’ vs ‘presymptomatic’ diagnosis, biased sampling

- **Dynamic:** Tightening and loosening of interventions, behavioral changes, disease dynamics at the individual and population level

Observed *symptomatic* individuals in a homogeneous population of 450K (SIR model), initially with 200 symptomatic and 300 recovered, with $\gamma = 0.2$.

More in the presentation by Yu-Xiang
US is heterogeneous with respect to the disease.

Disease state can be measured in multiple ways: percent change in infections, death rate, percent change in positive test rate, percent change in immunity, ..
Available data and models

Observe and control disease dynamics

Biomedical
1. Transmission fraction
2. Incubation period
3. Recovery rate
4. Death rate
5. Asymptomatic fraction
6. Immunity to future infections
7. Age effects
8. Seasonal effects

Population data
1. Population density
2. Age distribution
3. # Hospital admissions
4. # ICU admissions
5. # Infections
6. # Recovery
7. # Deaths
8. # Immune
9. # Doubling rate

Contact data
1. Stratified contact rates
2. Human mobility
3. Highway traffic
4. Air traffic
5. Super-spreaders
6. Business closures
7. School closures

Ambuj Singh, Yu-Xiang Wang (UCSB) COVID-19 seminar April 28, 2020 10 / 22
Sources of data and knowledge

  - John Hopkins CSSE curates a worldwide dataset with daily numbers down to the county level\(^3\)
  - More in the presentation by Yu-Xiang

- Since policy-making depends so much on models, we should also make them public.
  - If more data improve models, so does allowing people to look under their bonnets... As well as allowing for expert critique, it is a valuable way of building up public trust. (Economist, April 4, 2020).

- Peer review system is under strain
  - 2319 articles (1823 medRxiv, 496 bioRxiv) since January 19; about 24 per day
  - Challenge of getting right information to public quickly

\(^3\)https://github.com/CSSEGISandData/COVID-19
A number of datasets can be used for measuring the effect of NPIs:

- Facebook social connectedness index
- Google COVID19 community mobility reports
- Streetlightdata (shown above)
Challenge of dealing with uncertainty

- After hearing an economist talk about his forecast’s uncertainties and why a range of estimates was needed, President Lyndon B. Johnson reportedly said: “Ranges are for cattle, give me a number.”

- “Epidemiology is a science of possibilities and persuasion, not of certainties or hard proof” (New Yorker, April 26, 2020)

- Models typically rely on multiple uncertain inputs whose interactions are difficult to analyze.

- Behavioral changes are especially hard to anticipate (earlier Ebola example).
How do current models use data?

- CHIME model
- Institute for Health Metrics and Evaluation model (IHME)
- Imperial College model
- Multi-level models

---


5 Christopher Murray. “Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months”. In: medRxiv (2020). DOI: 10.1101/2020.03.27.20043752. eprint: https://www.medrxiv.org/content/early/2020/03/30/2020.03.27.20043752.full.pdf. URL: https://www.medrxiv.org/content/early/2020/03/30/2020.03.27.20043752.


CHIME model

- One of two models endorsed to predict hospital capacity in California.
- Based on SIR model
  - Number of current hospitalizations used to predict the number (and rate) of current infections
  - This and observed doubling rate is used to predict the infection rate.
  - Use empirical estimate of recovery rate
- Predict rate of hospital admits, rate of ICU admits, ventilator need
- Can change the dynamics by “social distancing,” realized by a percent reduction in infection rate
- Assumes homogeneity
- Deterministic model

\[
\begin{align*}
\frac{dS}{dt} &= -\beta SI \\
\frac{dI}{dt} &= \beta SI - \gamma I \\
\frac{dR}{dt} &= \gamma I
\end{align*}
\]
Imperial College model

- Influenced the policy of US and UK governments
- Stochastic simulation based model based on SIR. Parameters learnt from observed data.
  - Age specific variations
  - A fraction of asymptomatic individuals
- Distribute population into age groups
- Contact rates: home, work, school, random
  - Each NPI affects these rates in different ways
- Scalability is an issue
- Can one learn directly from data?
Institute for Health Metrics and Evaluation (IHME) model

- Parameterized matching of time series from Wuhan (and other locations that have recovered) based on the number of reported deaths and the level of NPI.
  - Matching adjusted for age distribution
  - Considers four NPIs: school closures, non-essential business closures, stay-at-home recommendations, and travel restrictions.
  - Three parameters are learned: maximum death rate, inflection point, and a growth parameter.

- Spatial resolution does not allow city-level predictions.

- Sensitive to reported deaths: can be quite different from actual deaths\(^8\), and can also be affected by the quality of care.

- Difficult to model new interventions.

- Hard to change predictions quickly (over a few days) and accuracy of predictions may not improve as the forecast horizon decreases\(^9\).

----


Multi-scale mobility models

- Hierarchical model most useful at the level of a country or a large region\(^\text{10}\).
- Lowest level dynamics based on SIR (or SEIR with an intermediate ‘Exposed’ state).
- Middle level dynamics based on traffic between cities: learnt based on actual data or gravitational model.
- Top level dynamics based on air traffic.
- Successfully modeled the effect of air travel on the transmission from Wuhan\(^\text{11}\).
- Homogeneous model at the lowest level.

---


Contact tracing

- A requirement by the CDC for opening up the country\textsuperscript{12}.
- Manual implementation: in-depth interviews with those who may be infected/exposed. Number of public health employees needed: 100,000–300,000.
- Different combinations of manual and phone-based systems deployed in New Zealand, Taiwan, Singapore (TraceTogether), Australia (COVIDSafe), India (Aarogya Setu), and South Korea.
  - The number of new cases in New Zealand has dropped to single digits.
- Care19 mobile app in North Dakota, South Dakota, Utah.
- Plans for extensive phone-based tracing (Google and Apple).
  - Should contact trace include GPS?
  - Missing data and false positives
  - Privacy

Can NPIs work on campus?

- Only a personal opinion but do see
  - the opinion piece by Christina Paxson: “College Campuses Must Reopen in the Fall. Here’s How We Do It” (NY Times, April 26, 2020).

- Automated contact tracing and GPS data by opt-in; manual otherwise

- Prior estimate of individual’s disease state based on
  - mobility pattern
  - disease state of contacted individuals
  - disease state of visited places

- Build a contact network of around 35K nodes with attributes ‘at-risk’ and ‘disease state’

- Test individuals on a continuous basis with priority to ‘at-risk’ and ‘hub nodes’ in the network
  - ensure reproduction number < 1 at each node

- Build capacity for
  - healthcare
  - alternate instruction plan

- Ensure privacy of collected information: More in the presentation by Yu-Xiang.
Remainder of the talk

• A closer look at the different types of “uncertainty”

• Case study: fitting SIR model to Santa Barbara data.

• A path towards “re-opening UCSB”
Measuring uncertainty with confidence intervals, in the IMHE model, a.k.a., “Murray model”
Explanation given for the shaded-region

Uncertainty is the range of values that is likely to include the correct projected estimate for a given data category. Larger uncertainty intervals can result from limited data availability, small studies, and conflicting data, while smaller uncertainty intervals can result from extensive data availability, large studies, and data that are consistent across sources.

The model presented in this tool has a 95% uncertainty interval and is represented by the shaded area(s) on each chart.

All deaths specific to COVID-19 patients.

Shaded area indicates uncertainty

Date

Mar 01 Apr 01 May 01 Jun 01 Jul 01

Deaths per day Deaths projected
Where are these uncertainties coming from?

- Uncertainty of estimated model parameters
- Uncertainty in the natural random process
- Bias in the model itself
- Bias from sample selection
- Changes in the distribution
- Effects of new interventions

Uncertainty that vanish as datasets get larger

Uncertainty that we could measure by model-checking techniques.

Usually can only gauge in the hindsight
The need for localized models

• Different demographics, culture, social norms

• Different industry composition

• Different availability of healthcare

• But we often do not have enough local data
  • 1% of the cities have 99% percent of the data.
The histogram of counties in the US by number of cases (as of Apr 25)

Santa Barbara: 7 deaths, 469 cases

New York City: 17,126 deaths, 155,113 cases
The number of days since first confirmed case

Santa Barbara:
41 days
Recall the basic SIR model (three parameters)

\[
\begin{align*}
\frac{dS}{dt} & = -\frac{\beta IS}{N}, \quad + \quad Z_1 \\
\frac{dI}{dt} & = \frac{\beta IS}{N} - \gamma I, \quad - \left( Z_1 + Z_2 \right) \\
\frac{dR}{dt} & = \gamma I, \quad + \quad Z_2
\end{align*}
\]

We added \( Z_1, Z_2 \) to model the uncertainty not captured by the model.

\( \beta \)  
Expected number of people one patient can affect per day.

\( \gamma \)  
Percentage of infectious people recovering (or dying) per day.

\( N \)  
Size of the susceptible population.
A (very rough) sketch calculation of the uncertainty... assume the model is correct.

• Borrow $\gamma$ parameter from Italy model.

• Fitting Santa Barbara data to the deterministic / stochastic SIR model:

  • Least square fit $\beta \in [0.276, 0.299]$  $N \in [467, 499]$

  • Maximum Likelihood $\beta \in [0.076, 0.375]$
Predictions of SB cases with fitted SIR model

- Uncertainty bounds of the least square fit of deterministic SIR model is narrow, but wrong.

- Confidence bounds of the MLE of stochastic SIR model are more conservative, but too wide to be useful.

- Neither model is correct.
The parameters of SIR model should change over time as a response to the interventions
Timeline of events in Santa Barbara County and UCSB

- **Mar 10:** UCSB transitions to remote instruction
- **Mar 16:** First case reported in SB County
- **Mar 19:** State-wide “Stay-at-home” order
- **Mar 31:** Two cases confirmed within UCSB Community.
- Many more events before / after / in between...
The intervention is complex and multi-dimensional...

- “Social distancing recommendation”
- “Lock down order”
- CDC Recommends “Don’t wear masks”
- CDC Recommends “Wear masks”
- The evolution of “Fear factor”
- How much are people following these orders?
- Types of businesses to resume
Receptions of the interventions can be objectively measured using phone-tracking data!
Google’s COVID-19 Community Mobility Report: Santa Barbara County

Santa Barbara County

- **Retail & recreation**
  - -49% compared to baseline
- **Grocery & pharmacy**
  - -12% compared to baseline
- **Parks**
  - -33% compared to baseline
- **Transit stations**
  - -49% compared to baseline
- **Workplace**
  - -36% compared to baseline
- **Residential**
  - +14% compared to baseline
We have to transfer knowledge from elsewhere, but make local adjustments...

• Describing interventions by “features”

• Describe local communities by “features”

• Measure effects of interventions objectively.

• Model the data-observation process so we can infer the unseen.
A potential path towards “reopening UCSB”?

- Build a refined model on a local level.

- Describe individuals by “features”
  - Class schedule, mobility patterns on campus, social activities
  - Contact tracing

- Targeted testing and re-testing of “high risk” individuals.

- Implement measures that are known to work, and learn from feedback.
Privacy concerns creep in even as we publish only aggregate statistics....

“The movements of soldiers within Bagram air base - the largest US military facility in Afghanistan”
“All models are wrong, but some are useful.”

• How useful a model is can often be tested

• Quantify uncertainty with “ranges”

• Be transparent on the assumptions and potential pitfalls
NSF Funded project on COVID-19

In collaboration with:
Dr. Richard Beswick, Dr. Lynn Fitzgibbons
Ms. Jeanne Li

• “Interventional COVID-19 Response Forecasting in Local Communities Using Neural Domain Adaptation Models”

• We cannot do it alone!