Lessons Learned from Modeling COVID-19: Steps to Take at the Start of the Next Pandemic

Christin Glorioso, Filippo Castiglione, Kayode Oshinubi, Aviral Chharia, Jacob Barhak

Presented by Jacob Barhak

Sole Proprietor, Computational Disease Modeler
Software Developer

MIDAS Network Webinar

2025 Mar 28th
Virtual

https://www.clinicalunitmapping.com/show/Lessons_Learned_COVID19_Latest.pdf

Conflict of Interest and Disclaimer

- Payment/services info: Dr. Barhak reports non-financial support and other from Rescale, and MIDAS Network, other from Amazon AWS, Microsoft Azure, MIDAS network, other from The COVID tracking project at the Atlantic, other from John Rice and Jered Hodges,
- Financial relationships: Jacob Barhak declare(s) employment from MacroFab, United Solutions, B. Well Connected health, Pronto Telecommunications. The author had a contract with U.S. Bank / Apexon, MacroFab, United Solutions, and B. Well during the work. However, none of these companies had influence on the work reported here.
- Jacob Barhak declare(s) employment and Technical Support from Anaconda. The author contracted with Anaconda in the past and uses their free open source software tools. Also the Author received free support from Anaconda Holoviz team and Dask teams.
- Intellectual property info: Dr. Barhak has a patent US Patent 9,858,390 - Reference model for disease progression issued to Jacob Barhak, and a patent US patent Utility application #15466535 - Analysis and Verification of Models Derived from Clinical Trials Data Extracted from a Database.
- Other relationships: personal fees from United Solutions, personal fees from B. Well Connected health, personal fees and non-financial support from Anaconda.
- Also, from 2006-2012 Jacob Barhak was paid by the University of Michigan to develop software licensed under GPL license.
- However, despite all support, Dr. Barhak is solely responsible for contents of this publication.

Preprints Published Paper

Citation:

Glorioso, C., Castiglione, F., Oshinubi, K., Chharia, A., & Barhak, J. (2024). Lessons Learned from Modeling COVID-19: Steps to Take at the Start of the Next Pandemic. Preprints. https://doi.org/10.20944/preprints202411.2193.v1

Transparency:

The drafting history of this paper can be accessed via this link: https://docs.google.com/document/d/1shsxwnzH5LFs3kyO_4dhR-kEh0Al_z9J4lslUpfzH_g/edit?usp=sharing

Steps to Take at Start of the Next Pandemic

Modeling

- Estimate infectiousness curves
- Estimate transmission rate per encounter
- Estimate encounter frequency per region
- Estimate the inaccuracy of numbers reported
- Estimate mortality profile
- Include baseline models

Validation

- Start simulations that attempt to match the recorded numbers

Information

- Start recording infections, hospitalizations, and deaths per region
- Make resources available and accessible
- Determine means of disseminating information to the public
- Improve measurement of hospitalization
- Expect variations

Infrastructure and education

- Computing infrastructure
- Centralized epidemiological and clinical records
- Educate and train experts on recently emerging technologies
- Educate the press before educating the public
- Research

1. Record Infections, Hospitalizations, and Deaths per Region

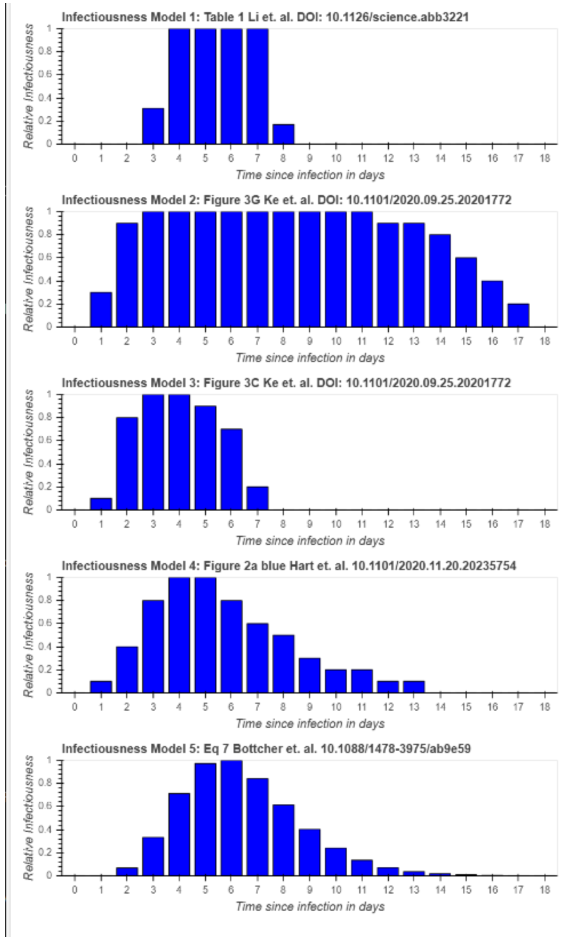
- During this last pandemic, there have been many inaccurate reports:
  - the number of cumulative hospitalizations which do not monotonically increase with time,
  - current hospitalizations higher than cumulative hospitalizations
  - mortality miscounting.
- It is important to record numbers “consistently”, especially numbers for hospitalization and deaths that may be more accurate.

1. Record Infections, Hospitalizations, and Deaths per Region

- Collected data should be made public:
  - in a trusted location
  - in multiple standardized formats such as CSV files, spreadsheets, and PDFs
- Numbers should be standardized across regions and countries with quality control
- A lack of standardization can result in misinterpretation and incorrect conclusions
  - the number of positive tests is often confused with the number of infections
  - modern methods such as wastewater metric data can help be more accurate

2. Estimate Infectiousness Curves

- In May 26, 2020, the Department of Health Services (DHS) was still asking about the average infectious period during which individuals can transmit the disease
- One approach to estimating infectiousness is to follow a patient and record the amount of virus they shed
    - Each patient may have their unique shedding curve
    - It may be highly useful to publish the overall distribution of these curves after anonymizing the patient information
- infectiousness curves can also be obtained by retrospective calculations



3. Estimate Transmission Rate per Encounter

- The probability of transmission per average encounter between two individuals sh
- This probability can be modified according to parameters such as:
  - population density
  - regional climate
  - ventilation
  - public transportation
- multiple possible formulas should be constructed

3. Estimate Transmission Rate per Encounter

- The reproduction number (R0) in SIR model depends on:
  - infectiousness
  - transmission rate
  - number of encounters
- Breaking the R0 number into its constituent factors is better for simulation
  - the compartmental SIR model is a century old
  - modern computing power allows more powerful simulation techniques

4. Estimate Encounter Frequency per Region

Need to collects parameters on interactions between individuals:

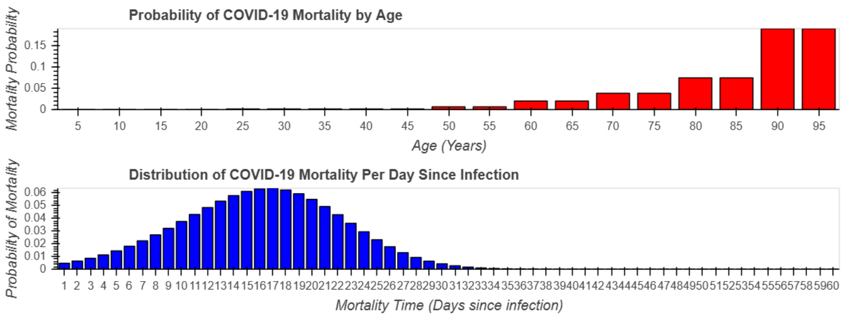
- number of encounters - average and standard deviation
- mobility data - to figure out variations
- household size - to figure out minimum interactions of sick people
- local jurisdictions restrictions such as school closures, office shutdowns, restaurant restrictions, mask mandates, mobility limitations

Such data is not available from a single source

- it can help compare which strategies seem to reduce the effect of a pandemic
- it can help determine which model best explains the observed data

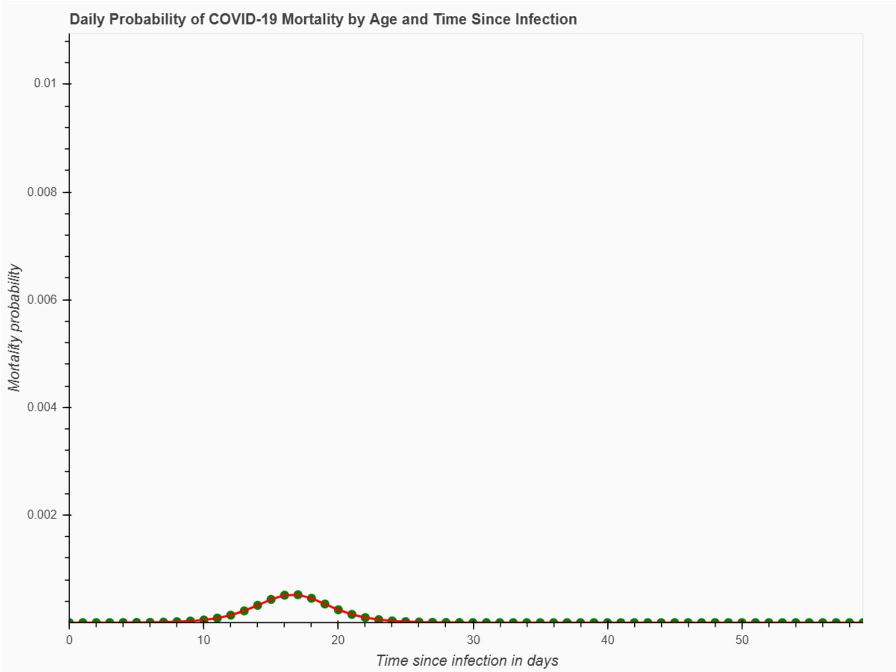
5. Estimate Mortality Profile

- The mortality table provides the death proportion of people infected
  - ideally stratified by age and/ or other factors like sex, etc.
- The mortality time provides the time of death since infection.
  - need to distinguish between:
    - time of death since infection
    - time of death since symptom onset



5. Estimate Mortality Profile

- It is possible to provide a model that provides the probability of death as a function of:
  - time since infection (day)
  - age (and possibly other factors)
- can be presented as:
  - function
  - table
  - interactive graphics



6. Estimate the Inaccuracy of Reported Numbers

- Reported numbers are inaccurate and require continuous audit because:
  - collection procedures are often not consistent
  - there are multiple types of diagnostic tests
  - procedures have different response times and change in time
- Examples:
  - # of positive tests does not scale to the # of infections in a population
  - death numbers may be inaccurate because:
    - delays in reporting
    - the cause of death not being associated directly with the pandemic

6. Estimate the Inaccuracy of Reported Numbers

- During modeling it important to distinguish between
  - modeled numbers – which represent virtual infections that should match the ground truth
  - observed numbers – which represent what we see and report
- Modelers should investigate multiple correction strategies to fix the difference between modeled numbers and observed values.
  - in some cases the accuracy of reporting is graded and this can be used in modeling

7. Start Simulations that Will Attempt to Match the Recorded Numbers

- The collected information should be verified using simulation to see how they match the observed reality.
  - it is highly unlikely that models can accurately forecast
  - the focus should not only be on pandemic forecasting
- Using ensemble modeling tools can
  - attempting to verify, compare, and construct the assumptions made
  - help pinpoint errors in comprehension
  - help improve comprehension over time

7. Start Simulations that Will Attempt to Match the Recorded Numbers

- The Reference Model for COVID-19 as an example of an ensemble model attempting to explain the pandemic
  - older findings:
    - transmission rate per encounter ~ 0.6%
    - long infectiousness model was dominant
  - recent finding simulate multiple time periods with different starts
    - transmission rate per encounter ~ 2%
    - mixture of infectiousness models
    - temperature may have an important role in transmission

7. Start Simulations that Will Attempt to Match the Recorded Numbers

The Reference Model experience shows:

- It is very useful to repeat simulations with different perspectives
- Proper explanation requires a lot of computing power
- Assumptions/models should be separated from data
- Attempt to extract invariant disease parameters that do not change with behavior

The ground hypothesis is that the phenomenon can be explained by analyzing the data:

- Many modelers believe this is the case
- it still needs to be proved so that credibility can be established

8. Make Resources Available and Accessible

All collected information should be made available to the public electronically with as few restrictions as possible.

- The Creative Commons Zero (CC0) License is recommended and makes it as close to the public domain as possible
- Seemingly permissive licenses, such as CC BY-NC 4.0 or GPL should be avoided
- Reuse of collected information should be allowed in all sorts of ways
- A competition between players should not diminish the incentives for other players
- Open-source tools and resources should be made readily available to support rapid data aggregation and subsequent modeling efforts.
- greater centralization and improved accessibility would be beneficial for the future

9. Improve Measurement of Hospitalization

- For accurate hospitalization modeling:
  - contact tracing efforts
  - techniques such as wastewater metric
  - start real-time simulations that attempt to match the recorded numbers
  - continually updating models with incoming data
  - include an iterative feedback loop

9. Improve Measurement of Hospitalization

- Simulations should incorporate:
  - hospital admission rates
  - length of stay

10. Include Baseline Models

Chharia Aviral , Jeevan Govind , Jha Rajat Aayush , Liu Meng , Berman Jonathan M. , Glorioso Christin , "Accuracy of US CDC COVID-19 forecasting models", Frontiers in Public Health, V12 (2024) . https://doi.org/10.3389/fpubh.2024.1359368

The work systematically analyzed the accuracy of US CDC COVID-19 case forecasting models

- categorizing them and then calculating their Mean Absolute Percent Error (MAPE)
- included two baseline models
- more than one-third of models fail to outperform a simple static case baseline
- two-thirds fail to outperform a simple linear trend forecast.

10.1. Recommendations for Government Officials

- a reassessment of the role of forecasting models in pandemic modeling and policy formulation
- raised concerns about “directly” hosting these models on official platforms, like the WHO, US CDC, etc.
  - those are widely trusted by the general public and press
  - the public may not be aware of significant differences / quality of models
- Current models can be used to forecast peak and decline of existing waves
- Beforehand forecast, still requires “hands-on” approach

10.2. Recommendations for Researchers

- Developing pandemic case forecasting models:
  - increased horizons such as forecasting 2-4 weeks ahead
  - A model that only forecasts one day in advance would have less utility
  - Staffing decisions for hospitals can require a lead time of 2-4 weeks to prevent over-reliance on temporary workers, or shortages.
- Focus on data collection methodologies
  - address the lack of clean, structured, and accurate datasets

11. Expect Variations

- Epidemic involve variations such as mutations that change parameters
- Mutations can change infectiousness, mortality, etc.
- mutations can be viewed as:
  - waves in the pandemic
  - mix in the population
- machine learning models will become less accurate as they are trained on datasets from previous variants
- data may be behind time in reporting by labs
- the public should be informed since previously infected or vaccinated individuals are no longer immune

12. Determine Means of Disseminating Information to the Public

- The public aware of the limitations of models and the data collected.
- Presenting a larger picture can help mitigate both public anxiety and indifference
- When reporting model results it is important to trace back all information to ensure reproducibility
- Trusting public may better follow enacted policies
- Authorities worldwide should admit errors quickly and humbly.

12.1. Computing Infrastructure

- High-Performance Computing (HPC) power is required
- Fortunately, cloud computing hardware is now readily available on demand
- Software tools need to be modernized to effectively leverage these platforms
    - older tools need modernization and replacement
    - new simulation tools should support:
        - Ensemble modeling technology
        - ability to explain and not only forecast

12.2. Centralized Epidemiological and Clinical Records

- Needed Infrastructure
  - dta lakes with public access
  - dashboards and visualization for big data
  - real time analytics
  - machine learning interfaces
  - access to electronic health record of infected individuals
- obstacles:
  - difference standards and protocols
  - geopolitical considerations

12.2. Centralized Epidemiological and Clinical Records

- Common agreement among entities wil help
- Best centralization possible within the limitations
- Attempts were made to centralize epidemiological data:
    - MIDAS network
    - DHS Master question list
- Attempts should be made to address clinical data

12.3. Educate and Train Experts on Recently Emerging Technologies

Gaps:

- The century-old SIR model is still dominant and
- Newly emerging ideas like using an Ensemble of Models are still uncommon
- Many experts in disease modeling do not have an in-depth programming background

12.3. Educate and Train Experts on Recently Emerging Technologies

Solutions:

- Train experts in:
  - version control
  - database systems
  - new modeling techniques
- Organizing modeling drills
  - Organize Hackathons for modelers and researchers based on synthetic data
  - national and local governments should collaborate to learn their roles in data collection and reporting.

12.4. Educate the Press Before Educating the Public

- Press should be aware of
  - model capabilities
  - model shortcomings
- Education channels can include: websites, articles, educational TV, interviews of modelers
- Journalists have fluent conversation and articulation skills compared to researchers
- journalists should avoid glorifying models and particularly prevent the researchers from doing so.

12.5. Research

- implementing microsimulation and agent based simulation in High-Performance Computing (HPC) infrastructures.
  - GPUs hardware
  - Recent compiler technology like Numba
  - CUDA parallelization
  - improve the assembling speed of ensemble models.
    - an ensemble model can take:
      - many versions (over 70 currently)
      - hundreds of CPU years to reach conclusions

12.5. Research

- After the COVID-19 pandemic, modelers have more datasets to work with and can effectively study how to improve things for the next pandemic:
  - data collection
  - modeling
  - reporting cycle.
- It is still an open question of how fast one can come up with a good explanation of a disease since its outbreak.
- Another unanswered question is if we can and under what conditions/constraints can apply this research to a future pandemic

Acknowledgments

- John Rice initiated the idea of this study and reviewed the first draft.
- Eric Forgoston helped substantially by providing valuable data on US states.
- Thanks to Robin Thompson, Alan Perelson, and Sen Pei for providing infectiousness models. - Lucas Böttcher helped connect authors and contributed models and other supporting information.
- Thanks to Guy Genin and Gregory Storch for providing useful information.
- The MIDAS network helped to connect authors. The MIDAS Coordination Center is supported in part by NIGMS grant 5U24GM132013 and the NIH STRIDES program.
- Thanks to Rescale, Microsoft Azure, Amazon AWS, and MIDAS for providing the computing power in previous years to support the modeling effort.