Journal of Medical Systems (2020) 44: 170
https://doi.org/10.1007/s10916-020-01645-z

EDUCATION & TRAINING

COVID-19 Prediction Models and Unexploited Data

K. C. Santosh [1]

Received: 23 June 2020 / Accepted: 11 August 2020 / Published online: 13 August 2020

Abstract

For COVID-19, predictive modeling, in the literature, uses broadly SEIR/SIR, agent-based, curve-fitting techniques/models. Besides, machine-learning models that are built on statistical tools/techniques are widely used. Predictions aim at making states and citizens aware of possible threats/consequences. However, for COVID-19 outbreak, state-of-the-art prediction models are failed to exploit crucial and unprecedented uncertainties/factors, such as a) hospital settings/capacity; b) test capacity/rate (on a daily basis); c) demographics; d) population density; e) vulnerable people; and f) income versus commodities (poverty). Depending on what factors are employed/considered in their models, predictions can be short-term and long-term. In this paper, we discuss how such continuous and unprecedented factors lead us to design complex models, rather than just relying on stochastic and/or discrete ones that are driven by randomly generated parameters. Further, it is a time to employ data-driven mathematically proved models that have the luxury to dynamically and automatically tune parameters over time.

Keywords COVID-19 · Prediction model · Data visualization · And machine learning

Background

Since December 2019, the novel Coronavirus (identified in Wuhan, China) threats globally as its spreading rate is found to be exponential. The following statement from World Health Organization (WHO) situation status report provides an idea of how sensitive the issue is:

> *Based on world health statistics, the COVI-19 pandemic is causing significant loss of life, disrupting livelihoods, and threatening the recent advances in heath and progress towards global sustainable development goals (source: report no. 114).*

Besides, they reported a clear guidance on considerations on adjusting public health and social measures. In this situation, prediction tools can help project different scenarios, such as a) number of possible confirmed (new) cases; b) number of possible hospitalized cases; and c) number of possible death cases (just to name a few). As a consequence, prediction tools are useful for several different purposes. As an example, number of possible hospitalized cases based on the severity level can help determine the need of numbers of ventilators and other sophisticated medical equipment. Further, states need to shape their health system responses in accordance with the need. For this, prediction models require to have important properties like epidemiological characteristics (of the diseases), such as incubation period, transmissibility, asymptomaticity, and severity. Other features, such as social distancing, stay-at-home orders, use of facemasks or self-quarantine, travel restriction, and contact tracing could help predict what comes next. For better understanding, prediction models are important to have better estimation about the disease and its possible threats. To be precise, according to the Centers for Disease Control and Preventions (CDC), prediction models helps respond pandemic by informing decisions about planning, resource allocation, and need the social distancing [1]. In [1], CDC prioritizes a) mortality forecast, b) hospitalization forecasts, c) COVID-19 pandemic planning scenarios, and d) COVID-19 surge.

Predictions, data simulation, and visualization

In particular, such models are crucial, where large amount of data are not possible to collect (resource constrained regions,

This article is part of the topical collection on *Education & Training*

✉ K. C. Santosh
santosh.kc@ieee.org

[1] Department of Computer Science, University of South Dakota, 414 E Clark St, Vermillion, SD 57069, USA

for instance). To amplify/visualize COVID-19 outbreak predictive analytical, it requires data visualization tools. Data visualization can help estimate the trend. Not to be confused, visualization tool cannot be considered as the prediction model. Unfortunately, in the literature, most of the prediction models are limited to data visualization. As an example, data simulations always help better understand the particular event(s). However, it must be limited to education/training. If not, subtitle (published by John Hopkins [2]), "envisions a fast-spreading coronavirus with a devastating impact" could be mistaken for newspapers headlines [3]. Similarly, the most read article in The Washington Post: "flatten the curve" with coronavirus simulator [4] helps citizens primarily aware of issues like social distancing and sanity in public health. In simple words, simulations help us build up our intuition about how diseases work in a way that words and even static charts cannot [3].

Predictive analytical results and media

In addition to unprecedented nature of the situation and many uncertainties that are related to diseases, inaccurate information can be predicted. Considering predictive analytical results and the trend of COVID-19 outbreak, recently, WHO has joined forces with the United Kingdom to run an awareness campaign named "Stop The Spread" about the risks of inaccurate and false information regarding the COVID-19 pandemic. Further, below are the few examples on how we provide information to the public. As an example, on March 31, 2020, the White House projected 100 K to 240 K Coronavirus deaths in the next two weeks [5]. Later, on April 8, 2020, we had another media statement [6] "not every model agrees: America's most influential coronavirus model just revised its estimates downward" as previous prediction was too far from actual values (84, 575 death cases in the U.S., dated: May 14, 2020).

Media did not intentionally broadcast/announce inaccurate information; instead, estimated values were based on prediction models. Not to be confused, this article is not aimed at blaming neither media nor prediction models.

Prediction models and unexploited data Correction: This must be a section like "Background" and "conclusion."

Artificial Intelligence and Augmented Intelligence play crucial roles in understanding data by using multiple different tools/techniques, such as data analytics, machine learning, and pattern recognition including anomaly detection [7, 8]. Predictive modeling requires exploiting comprehensive data. Missing one or two features/factors can deviate predictive values from actual ones.

More often, discrete models provide prediction based on their parameters, and of course input data (raw). Such input parameters are application dependent, as they need to be tuned during training validation. In case of continuous data and where there exist unavoidable uncertainties, these models behave differently. Such models do not provide coherent results, nor do they provide values close to actual data. The primary reason behind this is lack of understanding about the particular events i.e., data sentiments and additional unavoidable uncertainties/factors, such as hospital settings/capacity, number of tests on a daily basis, demographics, and population (density) and their vulnerability in that particular region. We observed that higher the population density higher the spread rate; and New York City can be considered as a real-world example (27,567 death cases and 340,661 confirmed cases, dated: May 14, 2020). This means that the exact same models with exact input parameters may not be applied for another region. Input parameters are required to be adjusted in accordance with the population density over time. Also, vulnerable citizens/individuals, which we often call "high-risk patients" cannot be just treated as healthy/normal citizens/individuals and vice-versa.

In the literature, we found COVID-19 models describe the characteristics of the disease and forecast accordingly, using mainly three different model types: a) SEIR/SIR models; b) Agent-based models; and c) Curve-fitting models [9]. Categorically, inspired from [9], let us revisit a few of them that help understand the practicalities of the models.

SEIR/SIR models Medical Research Council (MRC) Centre for Global Infectious Disease Analysis used Non-Pharmaceutical Intervention (NPI) model to reduce COVID-19 mortality and healthcare demand [10]. In their NPI model, SEIR was primarily adopted. NPI model predicted 2.2 million U.S. deaths (in an unmitigated scenario). Similarly, Columbia University used SEIR model with the name Severe COVID-19 model & Mapping Tool forecasted number of severe cases, hospitalizations, critical care, ICU use, and deaths under different social distancing scenarios, for 3-week and 6-week periods starting from April 2 [11, 12]. University of Pennsylvania named their prediction model, CHIME: COVID-19 Hospital Impact Model for Epidemics [13]. Their model allows users to vary inputs and assumptions. In their predictions, for next three months, they forecasted best- and worst-case scenarios for total number of hospitalizations, ICU bed demand, ventilator demand, and number of days these demands would exceed hospitals capacities.

Agent-based models A group of research centers and universities: Fogarty International Center, Fred Hutchison Cancer Center, Northeastern University, University of Florida and more employed agent-based COVID-19 prediction model [14]. They forecasted based on two different scenarios: a) no mitigation and b) stay-at-home. Compared to actual data, their range can be considered even though range is really wide.

Curve-fitting models Los Alamos National Laboratory employed Curve-fitting technique in their prediction model, named as Confirmed and Forecasted Cased Data Model [15]. As a fact, from their model, the best guess was for California state as of April 08, 2020, which were 4082 deaths (compared to 2974 actual deaths, dated May 14, 2020). The Institute for Health Metrics and Evaluation (IHME) [16] – an independent global health center – used curve-fitting model to project numbers of hospitalizations and deaths in the U.S. (including state-wise data) through August 2020. Their predictions vary over time as they employed curve-fitting model.

It is not a surprise that different models forecasted different results, since the exact same model with a small change in input parameters/variables (other than raw data) can significantly deviate guesses. Also, the way we employed data after social distancing (or lockdown) cannot be validated, since the collected data do not pronounce whether the lockdown was 100%. Even though their predictions are deviated from actual values, they are often limited to best guesses (for short-term prediction). Other than widely categorized (aforementioned) three different models, researchers are not limited to the use of machine learning and/or deep learning models [17, 18], where statistics and probability were taken into account. Also, they consider mathematical models for different cases using time-window: before, during, and after the lockdown. In [19], a comprehensive state-of-the-art works on forecasting COVID-19 is reported. As their models are transparent enough in terms of how they used parameters and assumptions, we still can learn and make society aware of how much worse can happen in the future. However, since they predict far from actual values, machine-learning scientists call models: "garbage-in garbage-out" [20, 21].

Therefore, it is required to check whether models take into consideration the following unprecedented factors: a) hospital settings/capacity; b) test capacity/rate (on a daily basis); c) demographics; d) population density; e) vulnerable people; and f) income/poverty. These factors are still uncovered and/or unexploited in most of the COVID-19 prediction models. These continuous and unprecedented factors lead to design complex models, not just relying on stochastic and/or discrete models. Stochastic models require fairly large amount of data to tune/stabilize their randomly generated parameters. Unlike the data-independent or discrete model, we are now required to employ data-driven models that have luxury to dynamically and automatically tune parameters over time.

While considering multiple factors that impact COVID-19 outbreak, it is a time to revisit the following items: a) is this really a complex problem? If not, why are not state-of-the-art tools accurate? We, scientists do not like to limit to win over others in terms of validation, we rather focus on developing a prediction tool, where majority of factors (mentioned earlier) are considered. In case we consider several different factors, prediction tools can be complex than expected. Within the scope, it is a time to see whether machine learning [22] deep neural networks [23] can be realized with thousands of parameters. Use of possible data analytical tools is another interest [24]. However, we must be aware of using data science and deep learning models as they require fairly large number of hyper-parameters to be tuned. This means that one must take a close look (experimentally) whether number of hyper-parameters supersedes number of input data (numeric data) for a prediction either mortality rate, death rate or recovery rate.

Conclusion

On the whole, let us quickly summarize it. In an ideal environment, prediction is somehow trivial, where the only concern is whether the data is large. However, in case of COVID-19, due to large amount of uncertainties, predictions could possibly deviate from what they should be. A few, but major uncertainties may come from multiple different sources, such as demographics, vulnerability issues that can be lung-related or heart-related diseases, hospital settings/capacity, test rate, social distancing, and income versus commodities. State-of-the-art prediction models that fall under the scope of SEIR/SIR, agent-based, and curve-fitting approaches barely include these aforementioned factors. As a consequence, their predictions are not close to actual values, nor do they produce consistent results among themselves. It is, however, not a surprise to see different results from different models, but it is unusual to have different values for the exact same objectives.

Rather than providing generic tools for predictions, it is important to focus on a few but major factors that significantly deviate the prediction values from the actual ones. Also, prediction models are expected to tune their (hyper)parameters over time. Meaning, data-driven models are expected, where (hyper)parameters need to be automatically tuned over time, with no under-fitting and over-fitting issues. If not, computer scientists call them: garbage-in garbage-out models because incorrect data can still permit statistical and probabilistic analysis [20, 21]. Besides, it is important to apply these models on different datasets and check changes in models' behavior from one region to another. As an example, the exact same model trained in New York may not be applied to South Dakota.

'What comes next' is also one of the primary issues. Along with COVID-19, lockdown in the name of social distancing impacts highly due to hunger across the world. UNICEF launches #Reimagine – a global campaign – to prevent the pandemic from becoming a lasting crisis or children [25] with the statement: As COVID-19 devastates already fragile health systems, over 440,000 additional children under five could die in the next six months in South Asia, without urgent action. As mentioned before, how can we forget income/poverty versus commodities in our prediction models?

Compliance with ethical standards

Conflict of interest Author declared no conflict of interest.

Ethical approval This article does not contain any studies with human participants performed by any of the authors.

References

1. Centers for Disease Control and Preventions (CDC). "COVID-19 Mathematical Modeling" Source: National Center for Immunization and Respiratory Diseases (NCIRD), Division of Viral Diseases (May 26, 2020) URL: https://www.cdc.gov/coronavirus/2019-ncov/covid-data/mathematical-modeling.html
2. Katie Pearce. Pandemic simulation exercise spotlights massive preparedness gap: Event 201, Johns Hopkins Center for Health Security, envisions a fast-spreading coronavirus with a devastating impact". Health Security (Nov. 06, 2019) URL: https://hub.jhu.edu/2019/11/06/event-201-health-security/
3. Amanda Makulec. Move over, data visualization. The era of 'data simulation' is here "Flatten the curve" was just the beginning. COVID-19 update, Fast Company (June 01, 2020) URL: https://www.fastcompany.com/90508780/move-over-data-visualization-the-era-of-data-simulation-is-here
4. Hary Stevens. Why outbreaks like coronavirus spread exponentially, and how to "flatten the curve" The Washington Post (March 14, 2020) URl: https://www.washingtonpost.com/graphics/2020/world/corona-simulator/
5. Fox News (by Andrew O'Reilly). "White House projects 100K to 240K coronavirus deaths as Trump tells US to prepare for 'very painful two weeks'". (Fox News: March 31, 2020) URL: https://www.foxnews.com/politics/trump-tells-americans-to-prepare-for-a-very-painful-two-weeks-as-white-house-releases-extended-coronavirus-guidelines
6. The Washington Post (by William Wan and Carolyn Y. Johnson). "America's most influential coronavirus model just revised its estimates downward. But not every model agrees.". (The Washington Post: April 08, 2020) URL: https://www.washingtonpost.com/health/2020/04/06/americas-most-influential-coronavirus-model-just-revised-its-estimates-downward-not-every-model-agrees/
7. Long, J.B., Ehrenfeld, J.M. "The Role of Augmented Intelligence (AI) in Detecting and Preventing the Spread of Novel Coronavirus." J Med Syst 44, 59 (2020). https://doi.org/10.1007/s10916-020-1536-6
8. Santosh, K.C. "AI-Driven Tools for Coronavirus Outbreak: Need of Active Learning and Cross-Population Train/Test Models on Multitudinal/Multimodal Data." *J Med Syst* **44,** 93 (2020). https://doi.org/10.1007/s10916-020-01562-1
9. Josh Michaud, Jennifer Kates, and Larry Levitt. "COVID-19 Models: Can They Tell Us What We Want to Know?" KFF (April 16, 2020) URL: https://www.kff.org/coronavirus-policy-watch/covid-19-models/
10. Ferguson et al. "Report 9 - Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand." MRC Centre for Global Infectious Disease Analysis (March 16, 2020) URL: https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/report-9-impact-of-npis-on-covid-19/
11. Charles C Branas, Andrew Rundle, Sen Pei, Wan Yang, Brendan G Carr, Sarah Sims, Alexis Zebrowski, Ronan Doorley, Neil Schluger, James W Quinn, Jeffrey Shaman. "Flattening the curve before it flattens us: hospital critical care capacity limits and mortality from novel coronavirus (SARS-CoV2) cases in US counties". MedRxiv (Apri 06, 2020) DOI: https://doi.org/10.1101/2020.04.01.20049759
12. Mapping tool: https://cuepi.shinyapps.io/COVID-19/. Columbia University (June 23, 2020, last accessed)
13. CHIME v1.1.5 (2020-04-08): https://penn-chime.phl.io. COVID-19 Hospital Impact Model for Epidemics (CHIME). University of Pennsylvania (April 2020)
14. COVID-19 Model: https://covid19.gleamproject.org/#model. Northeastern University, Fogarty International Center, Fred Hutchison Cancer Center, and University of Florida (May 15, 2020, last accessed)
15. Confirmed and Forecasted Cased Data Model: https://covid-19.bsvgateway.org. Los Alamos National Laboratory (June 20, 2020, last accessed)
16. The Institute for Health Metrics and Evaluation (IHME) COVID-19 Model: https://covid19.healthdata.org/united-states-of-america (June 20, 2020, last accessed)
17. Fong, Simon James et al. "Composite Monte Carlo decision making under high uncertainty of novel coronavirus epidemic using hybridized deep learning and fuzzy rule induction." Applied soft computing, vol. 93, 106282 (2020) https://doi.org/10.1016/j.asoc.2020.106282
18. Ermanno Cordelli et al. "Time-window SIQR analysis of COVID-19 outbreak and containment measures in Italy" IEEE 33rd International Symposium on Computer Based Medical Systems (CBMS), (2020)
19. Shinde, G.R., Kalamkar, A.B., Mahalle, P.N. et al. Forecasting Models for Coronavirus Disease (COVID-19): A Survey of the State-of-the-Art. Sn Comput. Sci. 1, 197 (2020). https://doi.org/10.1007/s42979-020-00209-9
20. "Garbage In, Garbage Out: How Anomalies Can Wreck Your Data – Heap – Mobile and Web Analytics". heapanalytics.com (May 7, 2014).
21. Steve Goldstein. "Oops — Rick Perry says broken clock is right once a day". The New York Post (Retrieved September 19, 2019)
22. Difan Zou, Lingxiao Wang, Pan Xu, Jinghui Chen, Weitong Zhang, Quanquan Gu. "Epidemic Model Guided Machine Learning for COVID-19 Forecasts in the United States." MedRxiv. DOI: https://doi.org/10.1101/2020.05.24.20111989 (2020)
23. Nancy Koleva. "When and When Not to Use Deep Learning". https://www.dataiku.com (May 1, 2020)
24. Joshi, Amit, Dey, Nilanjan, Santosh, K. C. Intelligent Systems and Methods to Combat Covid-19, SpringerBriefs in Computational Intelligence. eBook ISBN: 978-981-15-6572-4 (2020) https://doi.org/10.1007/978-981-15-6572-4
25. Anne Sophie Bonefeld. U Coronavirus – Reimagine, NICEF South Asia. https://www.unicef.org/rosa/press-releases/covid-19-devastates-already-fragile-health-systems-over-440000-additional-children (June 20, 2020, last accessed)