

Social networks, machine learning, and cladistics in the time of COVID-19

Redes sociales, aprendizaje automatizado y cladística en tiempos de COVID-19

Álvaro Díaz-Badillo^{1,2}, Carlos Ramírez-Pfeiffer², and Juan C. López-Alvarenga^{1,2*}

¹Department of Human Genetics, University of Texas Rio Grande Valley, Edinburg, Texas, USA; ²Department of Research, Universidad México-Americana del Norte, Reynosa, Tamaulipas, Mexico

*I falli commessi in un lontano passato sono molto più
facili da condannare che da correggere.*

Bocaccio

The most important cause of death throughout our history has not been wars, but infectious diseases. In 2009, H1N1 influenza alerted us about pandemics, but apparently we did not manage to learn much. The COVID-19 pandemic shows us our vulnerability once again and invites us to reflect on the tools; we have to contain similar pandemics in the future.

The virus that causes COVID-19, the infectious agent of current pandemic, belongs to the large family of coronaviruses that dwell in many species of animals, including camels, cows, cats, and bats. Sometimes, animal coronaviruses infect people, as it occurred with Middle East Respiratory Syndrome (MERS-CoV) or severe acute respiratory syndrome (SARS-CoV), and now with this new virus called SARS-CoV-2.

SARS-CoV-2 is a beta-coronavirus, like MERS-CoV and SARS-CoV, which originate in bats. The idea that this chiropter is the natural reservoir is supported by recent genetic sequences of the virus, obtained from patients of the United States and that are very similar to those obtained in China, the clinical presentation of which, COVID-19, is confused with a broad spectrum of diseases with symptoms that range from mild-to-severe. A recent publication reported severe disease in 16% of infected people in China¹, and a weekly morbidity and mortality report issued by the Centers for Disease Control

and Prevention mentions that 80% of deaths were among adults older than 65 years and that there was a higher frequency of severe cases in people older than 85 years^{2,3}. People with serious underlying medical conditions (severe heart conditions, chronic lung disease, and diabetes) also appear to be at increased risk for developing severe COVID-19.

Natural models that repeat themselves in history

The “black plague” during the middle age (1347 and 1351) had a devastating effect on a quarter part of the European population. In history, there have been several epidemics that have had a considerable emotional impact on the population, but today this emotional effect is exacerbated by the mass media, which have compared this epidemic to “the end of the world”, or that publish news about hospitals exceeded in their capacity and on the fear of food shortages⁴. Attempts to understand and predict epidemics have led to dissimilar methods for data acquisition and prediction modeling. In the 19th century, Sir Ronald Ross already attempted to generate simple linear models, and in the past century, Kermack and McKendrick⁵ used non-linear differential equations with groups of susceptible, infected and recovered individuals (SIR functions), which showed predictions that were closer to reality.

Correspondence:

*Juan C. López-Alvarenga

E-REBL 122

University Dr CP 78539

Edinburg, Texas, USA

E-mail: juan.lopezalvarenga@utrgv.edu

2444-0507/© 2020 Academia Mexicana de Cirugía. Published by Permanyer. This is an open access article under the terms of the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Date of reception: 18-04-2020

Date of acceptance: 22-04-2020

DOI: 10.24875/CIRUE.M20000146

Cir Cir (Eng). 2020;88(4):387-389

Contents available at PubMed

www.cirurgiaycirujanos.com

Experience with past epidemics, such as that of H1N1 influenza and those of SARS and MERS, can be of great help for understanding and managing COVID-19. The use of bioinformatics tools and machine learning techniques can predict COVID-19 diagnosis using age, vital signs, and clinical symptoms. Other models employ imaging and laboratory test results.

Many of the diagnostic and prognostic models that have been published are not reliable. A real problem is that some hasty models include patients who may not be representative of the population in which predicting is intended and, therefore, methodological guidelines have been developed, such as the Transparent Reporting of a Multivariate Prediction Model for Individual Prognosis (TRIPOD) initiative⁶. A meta-analysis⁷ with predictive models found that predicted mortality ranged between 8 and 59%; this considerably broad range is due to the biases that were introduced to the model. Another problem that occurs in these studies is validation of the models, since there are hardly any statistics of patients who are recovering and there are also delays in the diagnosis of confirmed cases and in the reports of fatal cases. In view of this problem, some researchers considered data obtained from other respiratory infectious diseases to validate their results. When some stochastic conditions are introduced to the model to bring it closer to reality, researchers find themselves with a lack of data for validation.

Bioinformatics as a tool for predicting infection

Hundreds of research teams all over the world are combining their efforts to collect data and develop solutions for this COVID-19 epidemic. Bioinformatics should contribute, through machine learning, to identify those who are at higher risk of contagion, to diagnose patients in a timely manner, to faster develop more efficient medicines, to predict the spread of the disease, and to define the degrees of virulence. It would also help to better understand the biology of viruses, map the sources where they come from, and predict upcoming pandemics.

As it can be observed in scientific literature, the use of bioinformatics currently offers tools that allow following the trajectory of the virus from its possible origins in the bat. Since the early stages of the epidemic in Wuhan, China, genome sequencing was carried out and seven preserved protein domains were found that gave support to the fact that this virus uses the same angiotensin-converting enzyme (ACE2) receptor⁸ as

the SARS-CoV family. In addition, sequencing in nine patients from the first cases showed similarity in 99.98%, including ACE2 receptor preservation⁹. Studies of ACE2 genetic variants in humans have found that they are associated with cardiovascular disorders, hypertension, and diabetes; however, structural changes, especially on the M2 peptidase subdomains, do not seem to produce important modifications in the structure. The rs73635825 (S19P) and rs143936283 (E329G) alleles show low affinity for SARS-CoV2 spike proteins¹⁰. This is consistent with a recent description of COVID-19 severity familial aggregation, which showed a decrease in CD4+ lymphocytes with an increase in the neutrophil/lymphocyte count ratio¹¹. Cladistics of the virus suggest that there are variants in the nitrogenous bases according to different geographical zones, and that some of these variants are found in the proteins responsible for replication. Thus, the combination of virus genomes and host variants may be functions that partly explain susceptibility.

The best approach that considers genetic variants and their interaction with the environment for the behavior of this pandemic to be understood is cladistics under the light of the theory of evolution. Cladistics are important for its possible prediction of the effects of a second wave of SARS-CoV-2 infections and even for the future infections. In addition, it classifies organisms and predicts what has not yet been observed. Cladistics is privileged, since in addition to producing hypotheses about organisms relationships, it also predicts their properties. This can be especially important for pharmaceutical companies or laboratories interested in improving diagnostic methods, isolation of agents, or development of treatments. This way, learning from current pandemic, we have a historic obligation to allocate more resources and promote research to better prepare ourselves for the next re-emergences or new infectious epidemics.

Search for new integrative models

Bernoulli initiated infectious diseases mathematical modeling in 1760, but the work by Kermack and McKendrick⁵ systematized the modeling framework.

Modeling can begin with basic questions that can be investigated: What conditions are necessary for an epidemic to occur? Which fraction of a population will become infected in an epidemic? Can partial vaccination in a population protect against the infectious agent? Validation of the model with new questions helps bringing us closer to the best prediction.

Now, we need to consider the consequences of this epidemic, integrating what we are progressively knowing in the course of research. Box's aphorism remains: "all models are wrong, but some models are useful"¹². Modeling tools should consider biology combined with social, economic, and cultural effects, which have been modified by the decisions of different governments in the world, and that have affected the flow of trading, interpersonal relationships, and ecology of the environment.

Integrative modeling should analyze the social, cultural, and economic impacts that can result in significant insured losses. Health and life insurance companies and other health-related companies should assess the risk of mortality and morbidity resulting from the spread of this type infectious diseases to determine the risk capital that is needed for extreme scenarios, given that pandemics are, so far, inevitable.

As it can be observed, from the perspective of other scenarios (e.g., H1N1 influenza in 2009), analytical and statistical models are crucial to understanding and managing the risk of infectious diseases. The 2009 pandemic was almost a failure for the world, and there is the danger of assuming that future influenza pandemics will be as "benign" as that one. The time of appearance and virulence of the next pandemic are uncertain, and there are also other emerging and re-emerging infectious diseases that are beginning to occur (measles, dengue, and hantavirus, among others), or that we have already seen in recent decades and that were previously unknown or poorly characterized, such as SARS, human immunodeficiency virus infection, and hemorrhagic fever by the Ebola virus.

Disinformation in the information century

Social networks are currently the first line of information about the COVID-19 pandemic. Although digital platforms constantly update case and fatality data, they also feed a wave of misinformation about the coronavirus, inasmuch as users focus on the popularity of what they post rather than on accuracy.

Recent studies indicate that most people in all countries of the world state that the media has helped them to understand the crisis and what they can do. However, approximately one out of every three people also state that the media has exaggerated the pandemic. Regulation of the veracity of information about the pandemic has led to the designation of sources to decide the reliability of news and event pages, in addition to the prohibition of certain advertisements during the pandemic.

It appears that the most dangerous virus of the 21st century is fake news, or pseudo-journalistic content disseminated through news portals, written press, radio, television, and social networks, whose objective is disinformation. Moreover, at the end of the day, this phenomenon can only be contained with good, real, and reliable news contrasted by integrative models.

Conclusion

In conclusion, efforts should be directed toward the search of integrative models that provide a quantitative framework for assessing risk and making "informed" decisions. To do this, we must take into account the lessons learned from the 2009 H1N1 influenza pandemic, and those that we will learn from the COVID-19 pandemic that currently affects us (April 2020). This implies improving the way, data are obtained by our population, considering the lag that exists with the confirmation of diagnoses, and making efforts to avoid the spread of misinformation on social networks. In addition, multidisciplinary forums are needed to discuss strategies for the future epidemics. If we do not have this, we will be producing a hundred picaresque tales in the style of *The Decameron*.

References

1. Wu Z, McGoogan JM. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72314 cases from the Chinese Center for Disease Control and Prevention. *JAMA*. 2020. Feb 24. doi: 10.1001/jama.2020.2648. [Epub ahead of print]
2. Centers for Disease Control and Prevention. Coronavirus Disease 2019. CDC; 2019. Disponible en: <https://www.cdc.gov/coronavirus/2019-ncov/hcp/index.html>.
3. World Health Organization. Coronavirus disease (COVID-2019) situation reports. WHO; 2020. Disponible en: <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/>.
4. Lima CKT, Carvalho PMM, Lima I, Nunes JVAO, Saraiva JS, de Souza RI, et al. The emotional impact of Coronavirus 2019-nCoV (new Coronavirus disease). *Psychiatry Res* 2020;287:112915.
5. Kermack WO, McKendrick AG. Contributions to the mathematical theory of epidemics - I. 1927. *Bull Math Biol*. 1991;53:33-55.
6. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med*. 2015; 162:55-63.
7. Wynants L, Van Calster B, Bonten MMJ, Collins GS, Debray TPA, De Vos M, et al. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. *BMJ*. 2020; 369:m1328.
8. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*. 2020;579:270-3.
9. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet*. 2020;395:565-74.
10. Hussain M, Jabeen N, Raza F, Shabbir S, Baig AA, Amanullah A, et al. Structural variations in human ACE2 may influence its binding with SARS-CoV-2 spike protein. *J Med Virol*. 2020 Apr 6. doi: 10.1002/jmv.25832. [Epub ahead of print]
11. Xia XY, Wu J, Liu HL, Xia H, Jia B, Huang WX. Epidemiological and initial clinical characteristics of patients with family aggregation of COVID-19. *J Clin Virol*. 2020;127:104360.
12. Box GEP. *Empirical model-building and response surfaces*. New York: John Wiley & Sons; 1986.