Commentaries

Accelerating the global response against the exponentially growing COVID-19 outbreak through decent data sharing

Cooper J Galvin [c,e], Luis Fernandez-Luque [b,g,h], Yu-Chuan (Jack) Li [a,d,e,f,*]

[a] International Medical Informatics Association
[b] Working Groups and SIGs, International Medical Informatics Association
[c] Biophysics program, Stanford Medical School, Stanford, CA, USA
[d] Graduate Institute of Biomedical Informatics, College of Medical Science & Technology, Taipei Medical University, Taipei, Taiwan
[e] International Center for Health Information Technology, Taipei Medical University, Taipei, Taiwan
[f] Department of Dermatology, Taipei Medical University Wan-Fang Hospital, Taipei, Taiwan
[g] Graduate Institute of Biomedical Informatics, College of Medical Science & Technology, Taipei Medical University, Taipei, Taiwan
[h] Salumedia Labs, Sevilla, Spain

ARTICLE INFO

ABSTRACT

The novel coronavirus disease 2019 (COVID-19) is a novel and exponentially growing disease, and consequently, the accelerated development of knowledge from good data is possible quickly and globally. In order to combat the global pandemic of COVID-19, all humans on earth need to make difficult strategic decisions on three very different scales, all fueled by Analytical and Artificial Intelligence-based predictive Models.

© 2020 Published by Elsevier Inc.

The novel coronavirus disease 2019 (COVID-19) is a novel and exponentially growing disease, and consequently, the accelerated development of knowledge from good data is possible quickly and globally. To create this knowledge, medical professionals and governments need to share relevant, privacy-protected, patient data with an open COVID19 registry.

In order to combat the global pandemic of COVID-19, all humans on earth need to make difficult strategic decisions on three very different scales, all fueled by Analytical and Artificial Intelligence-based predictive Models (AAIMs):

• **Personal scale**: the choices an individual makes about their lifestyle, work, movement, socializing, and other risk-associated choices — aided by the aggregated and analyzed information transferred to them through media, apps, and the environment they interact with,
• **Healthcare and pharmaceutical provider scale**: determining the best prevention, diagnosis, and treatment strategy for patients (and potential patients) given the resources available — aided by electronic health software and treatment AAIMs, all based on similar cases, and

• **Healthcare system scale**: determining the city, state, or national strategy for communication, transmission prevention, resource allocation, and tracking/restrictions to be imposed on the populace — aided by large-scale epidemiological, financial, and political AAIMs.

AAIMs, if properly designed, can be deployed rapidly to support epidemiologists and clinical, economic, and policy experts build models and perform analysis that synthesizes outcomes data from the pandemic into actionable information for the decision makers at all scales. With the support of AAIMs models and analyses, these experts are making assumptions that impact the validity of their models that inform the decisions on all scales. The assumptions of such models and its data sources should be made public, yet a bigger gap appears to be emerging on the access to that life-saving data.

The lack of access in a transparent and rigorous way to data and data models used in the current COVID-19 might be already impacting the response to the crisis and affecting millions of people. Policies have been changing constantly and varied widely for testing for and containing COVID-19 in many countries and states, thus affecting the nature of the data being collected. Further, the data that informed the early response by the United States Government and recommendation by the prime minister of the United Kingdom to let COVID-19 run its course without intervention were informed by a small sample size of 425 cases in Wuhan, potentially biased by political interest of China (Fauci

* Corresponding author. Tel.: +886-2-27361661x7600; fax: +886-2-6638-7537.
E-mail addresses: jack@tmu.edu.tw jaak88@gmail.com (Y.-C.(J.) Li).

et al., 2020; Li et al., 2020). The data that support the suggestion of "flattening the curve" of cases to spread the demand for hospital resources across the summer came from a greater number of cases reported openly from Wuhan and data on the severity of cases from Italy (Ferguson et al., 2020).Which of these models is correct depends on who has the most recent and accurate data. Are their models correct across geographies or only in those with data overrepresented in the databases?

Experts' models are only as accurate as the data that feed them, so it is in the best interest of every individual on earth to have relevant and actionable data from their location combined with large amounts of data from across the world.

Given the nature of the decisions being made, there is a dramatic need for patient-level information of geographies visited, deidentified demographic information, medical history and comorbidities, laboratory tests/image studies/medications received, and patient outcomes for all coronavirus cases witnessed worldwide. Korea, Italy, and Japan have shared more than 8000 patients' data on Kaggle, which have led to new understandings of the disease and its severity, but unfortunately, the shared data at different granularity, different data elements, and inconsistent representation of data.

There are already solutions in place to ensure a consistent representation of data in order to speed up the development of AAIMs. Universal data formats already exist for each of these categories of essential data: geospatial coordinates for geographies; age and gender (at least) for demographic information; common data model, like Observational Medical Outcomes Partnership (OMOP); and public release of diagnostic images (DIMCOM3) and laboratory testing (LOINC); and ICD-10 codes for diagnoses and outcomes. These data should be hosted publicly on

sites like Kaggle, MIMIC (medical information mart for intensive care), and RDA (research data alliance) that researchers across the globe access to provide insights. This will also facilitate a shared global effort towards the detection of data quality issues, compared with the current "black-box" approach inherent from only sharing heavy aggregated and huge-granularity data.

We implore healthcare professionals, payors, health system managers, and device makers to provide the data requested to Kaggle with all fields outlined in this letter. These data are essential for providing good information for the decisions made to combat COVID-19 and future pandemics on the personal, healthcare/pharmaceutical provider, and government scales.

Conflict of interest

None.

Funding

None.

References

Fauci AS, Lane HC, Redfield RR. Covid-19 — navigating the uncharted. N Engl J Med 2020; 382(13):1268–9.

Ferguson NM, Laydon D, Nedjati-Gilani G, et al. Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. London: Imperial College; 2020. https://doi.org/10.25561/77482.

Li Q, Guan X, Wu P, et al. Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. N Engl J Med 2020;382(13):1199–207.