Viewpoint

Data standards and standardization: The shortest plank of bucket for the COVID-19 containment

Mengchun Gong,^{a,b}* Yuanshi Jiao,^c Yang Gong,^d and Li Liu^a

^aNanfang Hospital, Southern Medical University, Guangzhou, China

^bInstitute of Health Management, Southern Medical University, Guangzhou, China

^cDigital Health China Technologies, Beijing, China

^dSchool of Biomedical Informatics, University of Texas Health Science Center at Houston, United States

Keywords: Digital Health; Data standard; Data standardization; COVID-19; IT infrastructure

In the battle against the unprecedented pandemic of COVID-19 worldwide, biomedical informatics, especially data standards and data standardization, have played significant roles in multiple aspects in containment of the pandemic, including understanding disease mechanisms,¹ improving clinical care,² triaging resource needs,³ advising policy-making,⁴ implementing public health countermeasures,⁵ enhancing technical innovation in syndromic surveillance,⁶ developing vaccines, and enabling wide coverage of vaccination.⁷

Nevertheless, the development of the standards for COVID-19 relevant data collection during the pandemic have gone through a lot of obstacles⁸ globally since the very beginning of the pandemic, which led to misleading statistics, inefficient communication, biased policy-making, and clinical risks.9 COVID-19 provided an eminent chance to test the data infrastructure in different regions and many issues and challenges have been exposed. Efforts to access and align existing healthcare data infrastructure in the context of the pandemic highlighted complicated interoperability challenges, which remain significant barriers to real-time data analytics and hurdles for improving health outcomes through datadriven responses.¹⁰ By reflecting on the COVID-19 related data standards in runological order (Figure 1), recommendations are made with the goal of promoting a globally-aligned standardization of healthcare data and the establishment of a community of common health for humankind amid the current and potentially future global public health crisis.

Recognizing the value of data standards and standardization for COVID-19 containment

It is now an era when medical practices, in both routine and emergent scenarios, are continuously recorded by digital systems, covering electronic health records and physiologic, laboratory, imaging data as well as decision-making and treatment information. Therefore, when no clinical trial data informs a rapidly evolving situation or unknown disease, the expectation would arise from the public for rapid and large-scale data collection, analysis to support strategic decision-making, and sharing of best practices.^{II} A critical component of the proposed strategy is the democratization of data: all collected information (observing necessary privacy standards) should be made publicly available immediately upon release in machine-readable formats based on open data standards and enabling data-informed decision making for all stakeholders.

Data standards empower international knowledge discovery and solution exploitation

Understanding of the clinical characteristics and responses to treatment of COVID-19 brought enormous value to clinicians when the trial-based evidence was sparse.^{I,I2} The large-scale real-world evidence generation network formed within the framework of OHDSI (Observational Health Data Sciences and Informatics)¹ has brought an innovative approach to coordinate data sources from different institutes, countries, and languages, aligned a cohort of over 4.5 million cases, and retrospectively described the unknown disease with strong representativeness on populations and regions (Europe, United States, South Korea, and China). OHDSI developed a comprehensive vocabulary system to incorporate data standards used in different countries and areas and implemented them in data processing and analytics. The high-level standardization and implementation of multiple standards enabled the OHDSI network to bring insights to clinical characteristics,¹³ treatment pathways¹⁴ and subgroup patients analysis.¹⁵



The Lancet Regional Health - Western Pacific 2022;29: 100565 Published online 11 August 2022 https://doi.org/10.1016/j. lanwpc.2022.100565

^{*}Corresponding author at: Nanfang Hospital, Southern Medical University, 1838 North Guangzhou Ave, Guangzhou, China.

E-mail address: gmc@nrdrs.org (M. Gong).

^{© 2022} The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http:// creativecommons.org/licenses/by-nc-nd/4.0/)



Figure 1. Timeline of data standards development during initial phase of COVID-19.

The network also provided important evidence on potential repurposed medications, which demonstrated an important approach to scan existing therapeutic methods in the lack of clinical trials of a new regimen.¹⁴ Last but not the least, data standardization and data sharing significantly improved the recruitment efficiency of clinical trials for new treatments and effectively monitored potential side effects of various medicinal products and the vaccines.¹⁶

The sharing of the data has been restricted to comply with related regulations. The potential of data-driven knowledge discovery and transfer has been weakened accordingly. However, in face of the high pressure, the scientific world has been robust in encouraging novel studies and data sharing without violation of data privacy. It's important to point out that the data standards and their implementation in different countries and languages have enabled multi-national studies without inflicting concerns of data governance and original data leakage. Within the coordinating mechanisms organized by OHDSI,¹ TriNetX,¹² ICODA,¹⁷ and other openscience networks, insights can be extracted, with an unprecedented scale and efficiency, from multiple independent databases around the world due to their common data model, vocabulary control, quality control, privacy protection mechanism and ethics standards.

Data standards enable data-informed decision making

Statistical analysis of the epidemiological trend required a standard nomenclature for the disease and high quality of data standardization in case reporting as well as data collection at both regional and global level.¹⁸ Inference from the epidemiological data to calculate the population size of potential contact was one of the key parameters to make policies on public health.

It is difficult to assess the accuracy of the data at the population level when the relevant data are distributed in the silos and the data owners are not willing to share it. Our experience, as illustrated in the Honghu Hybrid System (HHS),⁵ was using digital technologies to connect variable, if not all, data sources, integrated and standardized the data, and generated a near real-time surveillance system (daily) in the area with a population close to a million. Error in statistics during the emergent period of the pandemic was inevitable. A double-check mechanism, enabled by an independent channel (digital vs. manual) effectively minimized mismatched information.

Moreover, to mitigate the huge burden on medical needs and manpower shortage, many clinical decisionsupport systems (CDSS), mostly machine-learning based and data-driven, were developed and implemented in different checkpoints of the data flow¹⁹ for covering syndromic surveillance, triaging, severity classification, and outcome prediction. Although successes were reported within individual development sites, these systems could hardly be transplanted to other sites. The major reasons for such challenge include inconsistency in data standards and standardization, lack of usability for laypersons, difficulty of deployment in resource-poor settings, and potential ethical pitfalls or legal barriers.²⁰ The systems with the highest success rate of migration were the classification of chest CT images based on artificial intelligence (AI) technologies²¹ since the data in the Picture Archiving and Communication System (PACS) around the world follow the Digital Imaging and Communication in Medicine (DICOM) standard. However, the power of AI and datadriven predictive science played little role in improving the general level of clinical care for the COVID-19 patients, especially for the severe cases as the data infrastructure of standards and standardization were not ready for such challenges.

Reflection and effort on improving the level of data standardization

It is never too late to mend the fences as an old Chinese proverb said. There is an urgent need to reflect on the cause of low effectiveness of data sharing, data mining, and data science applications during the COVID-19 pandemic. The most important factor, also the shortest plank of bucket for the effort of containing the pandemic, is the lack of a widely implemented clinical data standard system and the various level of data standardization. This made the value of all the investment on hardware and software diminish. In order to quickly form an international data sharing network to generate real-world evidence and understand the disease as well as the affected populations,²² it is important to implement standards beyond the classification code (ICD). SNOMED CT (Systematized Nomenclature of Medicine - Clinical Terms), LOINC (Logical Observation Identifiers Names and Codes), and RxNorm are among the top recommended terminology systems.¹

In November 2020, the European Commission declared its commitment to the establishment of the European Health Data Space (EHDS), with the goal of facilitating access and better utilization of the European health data-eg, EHR, genomic, public health, and registry data.²³ Meanwhile, the Europe Commission announced the financial support program to member countries on implementing SNOMED CT as their core clinical vocabulary standard to enhance interoperability and increase the value of the data.²⁴ This provided a good example for the Western Pacific countries and regions to learn and build a data sharing platform for the future by clearly defining the best practices for fair benefit sharing, transparent and accountable governance of public and private sector data, true commitment to public dialogue, and global cooperation.

Recommendations for a tested preparedness

Strengthen the leadership of WHO

Reflecting on the initial phase of the COVID-19 pandemic, the identification of the pathogenic microorganism and its nomenclature, the characterization of the clinical manifestation and the definition of the diseases (from novel coronavirus pneumonia to COVID-19) have been the key steps for global coordination on research resources and implementation of public health countermeasures.10 WHO played an essential role in coordinating the expert resources, government support, and world-wide implementation, which paved the foundation for disease classification in healthcare IT systems, epidemiological statistics, and multi-center research programs. ICD has been proven efficient and cost-effective, considering the implementation in multiple languages in a short time across countries. International collaboration, under the leadership of WHO, should be strengthened to get more prepared for the future global public health emergencies. The upcoming ICD-11,25 which has been significantly modified to cope with the increasing needs in classification with more granularity, hierarchical terminology structure, coverage on clinical phenotypes, and incorporation of traditional medicine, will definitely help improve preparedness of data infrastructure in different countries.

Avoid potential bias and conflicts

Bias has been observed in the process of naming the disease. The use of the name of Wuhan city, where the world started to know about the virus, by some politicians and experts raised widespread sentimental conflicts worldwide and caused unnecessary waste of time and resources in that special period when each hour was counted for battling the disease, including taking care of patients and conducting research on understand the disease. We recommend that the bias and conflicts should be avoided, following the current naming methodology for COVID-19, to improve the implementation of the standards in all relevant countries and areas.

Equity in technology access and international collaboration

It is also recognized an unmet need to help low-to-middle income countries to accomplish standardization of the data and application of healthcare IT technologies. A regional effort to control the disease with such high transmissibility will not be successful without the involvement of all countries and regions. Training, financial support on infrastructure, free implementation of mature systems, and man-power support in data standardization and analytics are necessary and essential,⁵ especially for low-to-middle income countries and areas.²⁶

Conclusion

Healthcare IT, data sciences, and AI have failed public expectations during the COVID-19 pandemic due to the inadequate preparedness of IT infrastructure in most countries, if not all. Lack of data standards and low-tomiddle level of data standardization were part of the major causes and the shortest plank in the bucket for the containment of the pandemic. With strong coordination by WHO, a global effort to increase interoperability among the healthcare IT systems of different countries will be a fundamental step to get prepared for the next pandemic with an unknown origin.

Contributors

Dr. Gong Mengchun contributed to the conceptulisation and writing – original draft of the manuscript. Mr. Jiao Yuanshi contributed to visualisation and writing – original draft of the manuscript. Dr. Yang Gong and Dr. Liu Li contributed to writing – review & editing of this paper and provided valuable suggestions.

Declaration of interests

Mengchun Gong is an employee of Digital Health China, Co, Ltd. All other authors have no conflicts to declare.

Funding

Mengchun Gong and Li Liu were supported by the National Key Research and Development Plan of China (2020YFC2006406).

References

- I Kostka K, Duarte-Salles T, Prats-Uribe A, et al. Unraveling COVID-19: a large-scale characterization of 4.5 million COVID-19 cases using CHARYBDIS. Clin Epidemiol. 2022;14:369–384.
- 2 Xiao LS, Zhang WF, Gong MC, et al. Development and validation of the HNC-LL score for predicting the severity of coronavirus disease 2019. *EBioMedicine*. 2020;57:102880.
- 3 Wang CJ, Ng CY, Brook RH. Response to COVID-19 in Taiwan: big data analytics, new technology, and proactive testing. JAMA. 2020;323(14):1341–1342.
- 4 Flor LS, Friedman J, Spencer CN, et al. Quantifying the effects of the COVID-19 pandemic on gender equality on health, social, and economic indicators: a comprehensive review of data from March, 2020, to September, 2021. *Lancet*. 2022;399(10344):2381–2397.
- 5 Gong M, Liu L, Sun X, Yang Y, Wang S, Zhu H. Cloud-based system for effective surveillance and control of COVID-19: useful experiences from Hubei, China. J Med Internet Res. 2020;22(4): e18948.
- 6 Desjardins MR. Syndromic surveillance of COVID-19 using crowdsourced data. *Lancet Reg Health West Pac.* 2020;4:100024.
- 7 Lopez Bernal J, Andrews N, Gower C, et al. Effectiveness of the Pfizer-BioNTech and Oxford-AstraZeneca vaccines on covid-19 related symptoms, hospital admissions, and mortality in older adults in England: test negative case-control study. *BMJ*. 2021;373: nro88.

- 8 Bar-Zeev N, Shet A. Population science with individual-level data make for better policies. *Lancet Respir Med.* 2021;9(9):942–943.
- 9 Ong SWX, Young BE, Lye DC. Lack of detail in population-level data impedes analysis of SARS-CoV-2 variants of concern and clinical outcomes. *Lancet Infect Dis*. 2021;21(9):1195–1197.
- IO Gardner L, Ratcliff J, Dong E, Katz A. A need for open public data standards and sharing in light of COVID-19. *Lancet Infect Dis.* 2021;21(4):e80.
- Reeves JJ, Hollandsworth HM, Torriani FJ, et al. Rapid response to COVID-19: health informatics support for outbreak management in an academic health system. J Am Med Inform Assoc. 2020;27 (6):853–859.
- 2 Taquet M, Geddes JR, Husain M, Luciano S, Harrison PJ. 6-month neurological and psychiatric outcomes in 236 379 survivors of COVID-19: a retrospective cohort study using electronic health records. *Lancet Psychiatry*. 2021;8(5):416–427.
- 13 Burn E, You SC, Sena AG, et al. Deep phenotyping of 34,128 adult patients hospitalised with COVID-19 in an international network study. Nat Commun. 2020;11(1):5009.
- Prats-Uribe A, Sena AG, Lai LYH, et al. Use of repurposed and adjuvant drugs in hospital patients with covid-19: multinational network cohort study. *BMJ*. 2021;373.
 Duarte-Salles T, Vizcaya D, Pistillo A, et al. Thirty-day outcomes of
- 15 Duarte-Salles T, Vizcaya D, Pistillo A, et al. Thirty-day outcomes of children and adolescents with COVID-19: an international experience. *Pediatrics*. 2021;148(3):708–718.
- 16 Burn E, Li X, Kostka K, et al. Background rates of five thrombosis with thrombocytopenia syndromes of special interest for COVID-19 vaccine safety surveillance: Incidence between 2017 and 2019 and patient profiles from 38.6 million people in six European countries. *Pharmacoepidemiol Drug Saf.* 2022;31 (5):495-510.
- 177 International COVID-19 Data Alliance. https://icoda-researchorg/. Accessed I July 2022.
- 18 Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect Dis.* 2020;20(5):533–534.
 19 Chen J, See KC. Artificial intelligence for COVID-19: rapid review.
- J Med Internet Res. 2020;22(10):e21476.
 Sottile PD, Albers D, DeWitt PE, et al. Real-time electronic health record mortality prediction during the COVID-19 pandemic: a prospective cohort study. J Am Med Inform Assoc. 2021;28(11):2354-
- 2365.
 21 Li MD, Chang K, Mei X, et al. Radiology implementation considerations for Artificial Intelligence (AI) applied to COVID-19, from the AJR special series on AI applications. *AJR Am J Roentgenol.* 2022;210(11):1-0.
- 2022;219(1):1-9.
 Cosgriff CV, Ebner DK, Celi LA. Data sharing in the era of COVID-19. Lancet Digit Health. 2020;2(5):e224.
- 23 Commission and Germany's presidency of the council of the EU underline importance of the European Health Data Space. 2020. https://ec.europa.eu/commission/presscorner/detail/en/ IP_20_2049. Accessed 26 April 2022.
- 24 European Union drives use of standardized terminology in Member States with funding for SNOMED CT. https://www.snomed. org/news-and-events/articles/EU-drives-standardized-terminologyfunding-program. Accessed 26 April 2022.
- 25 WHO's new International Classification of Diseases (ICD-II) comes into effect. 2022. https://www.who.int/news/item/II-02-2022-who-s-new-international-classification-of-diseases-(icd-II)comes-into-effect. Accessed 22 April 2022.
- 26 Brotherton H, Usuf E, Nadjm B, et al. Dexamethasone for COVID-19: data needed from randomised clinical trials in Africa. Lancet Glob Health. 2020;8(9):e1125–e11e6.