https://doi.org/10.1038/s41591-022-02157-8

The case for including microbial sequences in the electronic health record

Vitali Sintchenko & Enrico Coiera

Integrating microbial sequencing data into electronic health records, while presenting privacy concerns, will improve patient care and population health and will expand the secondary uses of such data.

The growing availability of microbial genomes sequenced for health care rather than research raises the question of whether such data should be included in an individual's electronic health records (EHR). While integrating human genome data into EHR has been widely discussed[1], microbial genomic data bring unique and important challenges. One challenge is that the ownership of microbial genomic data remains ambiguous. Genomics service providers, public health agencies that fund such services, and patients consider themselves as stakeholders in genomic data governance. While human DNA defines our identity, this cannot be said about the genomes of coronaviruses, Salmonella or other microorganisms. These pathogens are temporary residents, and their genomes remain the same when sampled from different humans, especially those within a transmission chain[2].

Another challenge is that microbial genomics is a highly specialized field, with data from a growing and increasingly complex spectrum of pathogens and commensal flora that includes viruses, bacteria, medically relevant fungi and parasites. Each of these pathogens represents a different disease with unique pathology and epidemiology and comes with specific terminology standards[3].

Nevertheless, the utility of microbial genomics in controlling communicable diseases is clearly established[4], as demonstrated during the COVID-19 pandemic[5]. The inclusion of microbial genomic sequences within EHR together with comprehensive standards for data interoperability will enhance surveillance and disease prevention and can be guided by five principles (Box 1).

BOX 1

Microbial sequencing data obtained from a patient must be:

- Linked to population-level data
- Essential to the management of a patient's health problems
- Interpretable by a clinician
- Presented in a way that protects a patient's privacy when data are shared
- Integrated in a way that enables computational search, decision support or human decision making

Population level data

The first principle for including microbial sequencing data into EHR is that such data should be linked to population-level data, as this is critical for disease prevention and control analyses. This ensures that important public health and population data are available within the EHR of individual patients. Data linkage between healthcare records can improve the effectiveness of public health surveillance and interventions[4–6]. Multi-jurisdictional outbreaks and the cross-border spread of diseases have been understood and controlled through the sharing of genomic data. The similarity between the genomes of pathogens recovered from clinical cases of food poisoning and contaminated foods has been instrumental to identify sources and implement preventative public health actions[3,4].

The value of such shared data is a function of dataset size. The larger the sets of shared genome sequences, the more representative they can be of current disease activity and the easier it is to identify clusters of infections with a common source, which can then be identified and acted upon. A modelling study predicted that for each additional 1,000 genomes of foodborne bacteria added to a database, there are ~6 fewer cases per pathogen per year[7]. As genotyping of drug-resistant bacteria and pathogens with epidemic potential becomes the standard-of-care in infectious disease control, it will permit fine-grained observational analysis of treatments and outcomes, shaping best practice recommendations and identifying malpractice. Table 1 describes the main benefits and challenges of microbial genomics data linkage in health. To minimize risks such as privacy breaches, only minimal metadata (such as outbreak location) are shared along with the microbial genomes, especially in publicly accessible databases.

Patient care

The second principle for including microbial sequencing data in EHR is that microbial sequencing data must be essential to the management of a patient's health and have a specific and essential role in shaping decisions about their care. The value of microbial genome sequencing lies in the recognition and tracking of transmission events, as well as the identification of antimicrobial resistance and clinically relevant co-infections, with distinct variants of the same pathogen potentially improving the risk assessment and selection of treatment[8–10]. Three use cases in Table 2 provide examples of the added value of genomic data in patient management and infection control. In the three examples, microbial genomic data offer high-resolution diagnostic information, enable the detection of genomic markers of drug-resistance in difficult-to-treat and high-consequence diseases and allow risk stratification for healthcare-associated infections.

Genomics data are usually available as raw sequences produced by next-generation sequencing, consensus sequences inferred by bioinformatics pipelines or text files reporting laboratory interpretations. Additionally, metadata may contain information about a patient's demographics, data about specimen collection and exposure history, or the results of phenotypic tests on the sequenced pathogen. While many clinical applications would be well served by laboratory reports, some require less-processed data. Consensus sequence data may be needed to provide a high-resolution diagnosis in cases of severe disease, and raw sequence data may be expected to support the accurate detection of drug resistance in an actionable timeframe or to guide infection control measures for hospital-acquired infections (Table 2).

Table 1 | Benefits and risks of microbial genome data linkage in EHR

Benefits	More efficient clinical care through precision medicine, such as better targeted selection of therapy and case risk assessments
	More efficient public health surveillance, such as better recognition of multi-jurisdictional clusters of infections
	Enhanced value of EHR for infection control and secondary use
	Advances of knowledge and professional staff development, including research and development with artificial intelligence
Risks	Ethical concerns, such as breaches of patient privacy and confidentiality through the disclosure of transmission events without consent
	Incidental findings or invalid data interpretation, with the associated medico-legal risks, due to data users lacking context or not appreciating sequence data biases and the implications of missing data
	Limited recognition or attribution of intellectual property rights of data donors related to raw and processed sequencing data

Interpretable data

Microbial genomic data must be interpretable by a clinician. Typically, diagnostic laboratories report processed rather than raw data to requesting clinicians. Results of molecular tests such as PCR are often provided with a statement such as 'nucleic acid of a pathogen was detected' but not the primary data, like melting curves from the PCR. The specialized nature of genomic data creates significant complexity for microbial genome analysis outside of genomic laboratories. There is a risk of misinterpreting specialized genomic data or incidental findings after the initial laboratory report, which could impact patient safety and privacy, with medico-legal implications.

However, advances in genomics and data analytics challenge this approach of withholding information to avoid confusion or misinterpretation by clinicians and encourage the full sharing of sequencing data as well as the provision of support for clinicians when interpreting data. Historically, clinicians have become adept at interpreting signal data such as electrocardiograms or radiological images, either as part of comprehensive training or for specific clinical issues. There is no reason to expect that infectious disease specialists will not gradually upskill in genomics. If microbial sequence data are to be retained in EHR, competence in understanding these data should be a requirement and part of medical training, and the interpretation of genomic data should be aided by computerized decision support.

Decision making

Microbial sequence data must be presented in a way that they can be directly and meaningfully interrogated by a clinician or informatics specialist or be integrated into software that provides the requisite decision support[1]. The representation of sequence data in decision support software must permit computational access and manipulation, for example, through the creation of summary visualizations, the identification of features of interest or annotation with relevant and up-to-date external information, given that knowledge about pathogen genomics as well as therapeutic options and links to known local outbreaks evolves rapidly. Currently, the integration of structured microbial genomic information into EHR to support patient care remains limited[11]. Microbial genomics reports are generally still presented as text files, and more structured delivery will require embracing information standards. The research community needs to invest in the implementation of standards for microbial genomics metadata to ensure that these data are interoperable, findable, searchable and reusable[12].

Privacy and access

Microbial genomics data that have been generated as part of healthcare delivery have often also been stored in health databases and registries focused on notifiable conditions. The value of sharing genomics data is shaped by its timeliness. Withholding human genome sequence data is commonplace among researchers and companies, but delays in pathogen data sharing can reduce the likelihood that transmission events are recognized in an actionable timeframe, and the negative effects of data hoarding on disease control have long been recognized[13]. Microbial genomic data stored in EHR serve different purposes to the data captured in public health or open databases or biobanks. Sequencing data in EHR can support prescribing decisions, diagnostic testing and prognosis. By contrast, data in public health information systems underpin public health investigations of disease clusters and non-pharmaceutical interventions. Biobanks collate data and the associated samples from healthcare providers and researchers to facilitate reanalysis of genomic data. We would thus expect that different types of genomic data and different types of data governance arrangements would be needed to support clinical management, population health and translational research.

While open databases can support open science and crowd-sourced discovery, genomic data linked to individual and public health records can support clinical trials and can be used to measure the outcomes of treatments and population health interventions. Sharing microbial genomic data and metadata associated with diseases affecting humans and animals and involving cases across different jurisdictions requires collaboration between sectors, including human and animal health, food and environment, and relevant government, commercial and not-for-profit stakeholders[3,14].

Genomic data providers must reduce the risks to individuals of privacy-invasive or reputation-damaging inferences that could be drawn from microbial genomics data, and they must protect the legally recognized rights of individuals to access personal data and contest interpretations of their data.

Precision public health

One benefit of genomic data sharing is more efficient public health surveillance, which will enable better targeted and nuanced interventions, a model referred to as precision public health. Furthermore, local data can be compared with global context data (such as genomes reported as circulating in other countries), providing opportunities for research and development (Table 2). Indeed, the World Health Organization

Table 2 | Clinical case studies using microbial genome sequencing

Case	Scenario	Context	Added value
High-resolution diagnosis in cases of severe disease	Confirmation of diagnosis for catastrophic infection[19]	This case represents a fulminant course of fried rice–associated food poisoning, caused by pre-formed exotoxin produced by *Bacillus cereus*, in an immunocompetent person. The patient developed severe septic shock with multi-organ failure, metabolic acidosis, rhabdomyolysis and coagulopathy. Despite maximal supportive measures (continuous renal replacement therapy, plasmapheresis and broad-spectrum antimicrobials), the patient died.	Whole genome sequencing (WGS) enabled the detection of genes responsible for the production of emetic toxin (cereulide) and non-haemolytic toxin to differentiate the strains of *B. cereus* capable of causing emetic-type food poisoning from the common and less dangerous strains of this bacterium. The WGS diagnosis had important implications for diagnostic investigations and public health messaging about food preparation and storage, particularly given its occurrence during COVID-19 home isolation.
Accurate detection of drug resistance in an actionable timeframe	Resistome assessment for the treatment of extensively drug-resistant tuberculosis[20]	WGS has rapidly progressed from a research tool to a cost-effective clinical application for the diagnosis and management of tuberculosis and for public health surveillance[21,22]. Phenotypic drug susceptibility testing (DST) for *Mycobacterium tuberculosis* takes several weeks to complete, and second-line DST is often poorly reproducible, potentially leading to compromised clinical decisions. WGS can generate an in silico drug-susceptibility profile much faster and with high precision in adequately resourced settings[21].	WGS identified variants of *M. tuberculosis* associated with drug resistance two weeks before the DST report, which was issued 10 weeks after patient presentation and 8 weeks after the initial growth of *M. tuberculosis*. In the interim, the patient may have received a compromised regimen with the potential to select for further drug resistance. The in silico susceptibility profile provided comparable or superior data to the DST results for second-line drugs, in a much shorter timeframe[23].
Recognition of hospital-acquired infections	Diagnosis and risk stratification of *Clostridioides difficile* (CD) disease[24]	WGS has been instrumental for understanding different healthcare reservoirs, modes of transmission and the role of patients with asymptomatic CD colonization in transmission, as well as for detecting risk factors for CD transmission within healthcare settings.	Hospital infection control investigations based on CD genome sequencing can identify links to determine relatedness between implicated isolates in a significant proportion of cases, with a threshold of ≤2 single nucleotide polymorphisms. Independent risk factors (adjusted P <0.05) for CD acquisition, such as older age, longer inpatient duration and CD ribotype, increase the risk of onward transmission. Patients with a plausible source identified by WGS have a greater risk of recurrence and 30-day mortality. Clinical characteristics associated with increased healthcare-associated CD transmission could be used to target preventative interventions[25]. This timely recognition of hospital-acquired infections improves health care delivery and helps to stratify patient management, as acquiring CD from a recent case is associated with worse clinical outcomes. Furthermore, WGS can differentiate CD strains with high virulence and disinfectant resistance as well as a high propensity for transmission[24].

has made genomic surveillance a global health priority[15]. The onus to reduce the risks of data sharing is largely borne by data donors and relate to patient privacy, confidentiality and data security. The risk of re-identification of patients from publicly shared microbial genomic data is considered to be low, but the risk also depends on the sensitivity of the data. For example, a re-identification risk of 10% may be considered acceptable for SARS-CoV-2 sequence data, but for data associated with sexually transmitted infections, some guidelines recommend a risk threshold of 5%, as there is a greater potential harm[16].

Genomics service providers are sequencing-data donors, while clinicians and epidemiologists are data recipients, and the two have different responsibilities and expectations of this data sharing arrangement. Laboratory data donors are increasingly concerned about the legal and ethical implications of incidental findings made by data recipients, misuse of data and profiteering from re-using data with limited attribution of intellectual property, as well as the lack of accountability of data recipients due to distributed and unpredictable data re-use.

Genomic data have been treated as an emerging asset, with privately run DNA marketplaces paying individuals for their genomic and personal data with either money or in-kind payments. As personalized medicine becomes the norm, sequence data from individual people and their pathogens may become more valuable[17]. Some countries assert that microorganisms located within or isolated from their territories are their sovereign property and should be protected as elements of their natural diversity — a concept referred to as microbial sovereignty[18].

Microbial genomic testing has shifted from being a reasonable step to prevent the spread of infectious diseases to the standard of care for many of them. Microbial genomics data can improve infection control and prevent hospital-acquired infections. This genomic evidence on preventable transmission events in healthcare settings can also be used by parties outside healthcare systems for claims about malpractice. The importance of the responsible analysis and reporting of disease clusters using genomics data cannot be overstated. If accurate, such discoveries should improve health care delivery, but the economic and reputational repercussions of making invalid inferences from genomic data can be significant for healthcare systems, industries and nations.

New ways of managing microbial genomic data are required to maximize the benefits of such data for patients and society. There is value both in integrating microbial sequencing data into EHR for patient care and population health and in sharing genomics data that, as a consequence, should expand the secondary uses of data from EHR. The inclusion of microbial genomes into EHR should be based on the principle that such data are shared responsibly and in a timely manner. Data processing tools are needed to integrate and contextualize the clinical analyses of microbial sequences, which will reduce the complexity of sequencing data interpretation for healthcare providers and patients.

Vitali Sintchenko [1,2] ✉ **& Enrico Coiera** [3]
[1]Sydney Institute for Infectious Diseases, University of Sydney, Sydney, New South Wales, Australia. [2]Centre for Infectious Diseases and Microbiology – Public Health, Institute of Clinical Pathology and Medical Research, NSW Health Pathology, Sydney, New South Wales, Australia. [3]Centre for Health Informatics, Australian Institute of Health Innovation, Macquarie University, Sydney, New South Wales, Australia.
✉e-mail: vitali.sintchenko@sydney.edu.au

Published online: 16 January 2023

References

1. Marsolo, K. & Spooner, S. A. *Genet. Med.* **15**, 786–791 (2013).
2. Sintchenko, V. & Holmes, E. C. *Brit. Med. J.* **350**, h1314 (2015).
3. Lefterova, M. I., Suarez, C. J., Banaei, N. & Pinsky, B. A. *J. Mol. Diagn.* **17**, 623–634 (2015).
4. Gardy, J. L. & Loman, N. J. *Nat. Rev. Genet.* **19**, 9–20 (2018).
5. du Plessis, L. et al. *Science* **371**, 708–712 (2021).
6. Kruse, C. S., Stein, A., Thomas, H. & Kaur, H. *J. Med. Syst.* **42**, 214 (2018).
7. Brown, B., Allard, M., Bazaco, M. C., Blankenship, J. & Minor, T. *PLoS ONE* **16**, e0258262 (2021).
8. Rockett, R. J. et al. *Nat. Commun.* **13**, 2745 (2022).
9. Rockett, R. et al. *N. Engl. J. Med.* **386**, 1477–1479 (2022).
10. The CRyPTIC Consortium and the 100,000 Genomes Project. *N. Engl. J. Med.* **379**, 1403–1415 (2018).
11. Griffiths, E. J. et al. *GigaScience.* **11**, giac003 (2022).
12. Musen, M. A. *Nature* **609**, 222 (2022).
13. Marshall, E. *Science* **275**, 777–780 (1997).
14. Khoury, M. J., Armstrong, G. L., Bunnell, R. E., Cyril, J. & Iademarco, M. F. *PLoS Med.* **17**, e1003373 (2020).
15. *Global Genomic Surveillance Strategy for Pathogens with Pandemic and Epidemic Potential 2022–2032* (WHO, 2022).
16. Song, L. et al. *Front. Genet.* **12**, 716541 (2022).
17. Cartwright-Smith, L., Gray, E. & Hyatt Thorpe, J. *Vand. J. Ent. Tech. L.* **19**, 207–241 (2016).
18. Halabi, S. *Ann. Health L.* **28**, Article 3 (2019).
19. Colaco, C. M. G., Basile, K., Draper, J. & Ferguson, P. E. *BMJ Case Rep.* **14**, e238716 (2021).
20. Outhred, A. C. et al. *J. Antimicrob. Chemother.* **70**, 1198–1202 (2015).
21. Meehan, C. J. et al. *Nat. Rev. Microbiol.* **17**, 533–545 (2019).
22. Mugwagwa, T., Abubakar, I. & White, P. J. *Thorax* **76**, 281–291 (2021).
23. World Health Organization Communicable Diseases Cluster TB. *The Use of Next-Generation Sequencing Technologies for the Detection of Mutations Associated with Drug Resistance in* Mycobacterium tuberculosis *Complex: Technical Guide* (WHO, 2018).
24. Kong, L. Y. et al. *Clin. Infect. Dis.* **68**, 204–209 (2019).
25. Martin, J. S. H. et al. *Clin Infect. Dis.* **67**, 1379–1387 (2018).

Acknowledgements

This work was funded in part by the NSW Health Prevention Research Support Program and the Australian National Health and Medical Research Council Centre for Research Excellence in Digital Health.

Competing interests

The authors declare no competing interests.