Governance Framework for a FAIR and Care Health Synthetic Data Ecosystem

Helen H Chen (University of Waterloo, School of Public Health Sciences), Maura Grossman (University of Waterloo, Cheriton School of Computer Science), Shu-feng Tsao (University of Waterloo, School of Public Health Sciences)

Health data, especially electronic medical records (EMRs), are often stored in disparate systems and formats, rendering integration and standardisation difficult. Researchers and developers often depend on de-identified or aggregated data to test theories, data models, algorithms, or prototype innovations, but it often takes substantial time and resources to retrieve, aggregate, and de-identify relevant data before it can be used. One approach to solve these challenges is the creation of realistic, high-quality, synthetic health datasets that capture as many of the complexities of the original data sets but do not include any actual patient data, coined as "fully synthetic data." The fast advent of artificial intelligence (AI) techniques, especially those involving privacy preservation, makes generating high-quality health synthetic data feasible. Synthetic health data generates significant interest from the health research and technology innovation community since it would take much less time to access synthetic data. For example, the Clinical Practice Research Datalink (CPRD) published and maintained by the United Kingdom (UK) governmental agency has created synthetic datasets available for research. Several organizations create and publish synthetic data in the United States (US), such as the Agency for Healthcare Research and Quality and Centres for Medicare and Medicaid. Synthetic health data can accurately reflect the characteristics of a population of interest and serve as a valuable resource for policymakers, health researchers and health technology innovators. Synthetic health data shows great promise in protecting patient privacy, diversifying datasets, and enhancing clinical and innovative research.

However, certain challenges have emerged concerning current practices involving health synthetic data. First, data governance is lacking, given that existing health synthetic data have been created on a case-by-case basis or generated by a limited number of organizations with their data governance processes and procedures. Ideally, we would apply the Findable, Accessible, Interoperable, and Reusable (FAIR) principles when sharing synthetic health data. However, unlike the UK and the US, Canada has very limited sharable and useful, high-quality health synthetic datasets that meet FAIR standards, despite its footprint in the Common Infrastructure for National Cohort in Europe, Canada, and Africa (CINECA) projects. Moreover, the FAIR principles remain insufficient to address Indigenous data. The International Indigenous Data Sovereignty Interest Group within the Research Data Alliance has proposed principles for handling Indigenous data, including Collective Benefit, Authority to Control, Responsibility, and Ethics (CARE). Overall, data governance for synthetic health data is undeveloped, given the fragmented management of synthetic health data in Canada and elsewhere.

Another challenge is the ambiguous status of the treatment of synthetic health data in Canada's regulations, leading to uncertainty regarding patient consent and research ethics guidelines for creating, transferring, and using synthetic health data. Under the US Health Insurance Portability and Accountability Act (HIPAA), creating de-identified data is regarded as part of the healthcare operations of a covered entity. Therefore, patient consent is not required even if the de-identified data will function as a database for research. Most US organizations have adopted this position to waive the requirement for informed consent when generating synthetic health data for research or commercial purposes. In addition, approval from an institutional review board (IRB) is not required since synthetic health data is considered secondary data analysis that does not meet the definition of "human subject research." According to the European Union's (EU) General Data Protection Regulation (GDPR), synthetic health data can be treated as "pseudonymous" or "anonymous" data if such data is appropriately created, but informed consent is still required, so patients know that their data will potentially be used to create synthetic data.

Synthetic health data are not addressed in Canada's Personal Information Protection and Electronic Documents Act (PIPEDA) since PIPEDA does not distinguish between anonymous and de-identified data. In contrast, the Consumer Privacy Protection Act (CPPA) does make such a distinction. Apart from these two privacy regulations, the Government of Canada has proposed the new Artificial Intelligence and Data Act (AIDA), and its potential impact on synthetic health data and AI techniques to generate such data remain to be clarified. Although it seems likely that the generation of synthetic health data will involve ethical review and approval in Canada because original EMRs are used in this process, the research ethics board (REB) or institutional review boards (IRB) have not reached an agreement on studies involving only synthetic health data. Decisions are often made on a case-by-case basis because laws in Canada do not have an explicit definition or reference for synthetic health data. This has created ambiguity and concerns the need for informed consent and ethical approval to create and use this kind of data. Much of this confusion results from a lack of a comprehensive policy that balances the competing interests of different stakeholders.

This paper presents ethics considerations, privacy protection rules, and cost-benefit evaluation rules, which lead to a clear governance framework for creating a FAIR and CARE synthetic health data ecosystem in Canada. This data governance framework acknowledges issues relevant to Indigenous and other marginalized groups and embeds ethical, equity, diversity, and inclusion (EDI) principles. Furthermore, this framework incorporates existing processes on cost-benefit analyses recommended and employed by the Treasury Board of Canada and the Merger Enforcement Guidelines and used by the Competition Bureau of Canada. We believe such a framework should have strong policy relevance as Canada's new Artificial Intelligence and Data Act is further developed to encourage the valuation of synthetic health data to accelerate Canada's data economy and innovation.