Establishing a FAIR, CARE, and Efficient Synthetic Health Data Sharing Ecosystem for Canada

Helen Chen[1*], PhD; Maura R. Grossman[1,2], J.D., PhD; Anindya Sen[3], PhD; Shu-Feng Tsao[1], PhD

[1]School of Public Health Sciences, University of Waterloo

[2]Cheriton School of Computer Science, University of Waterloo

[3]Department of Economics, University of Waterloo

Abstract

Obtaining access to real-world health data is a significant challenge, mainly due to privacy and security implications. Consequently, researchers and technology innovators–particularly those operating in the health data science and AI technology development spaces – increasingly resort to synthetic health data to bridge the data gap. High-quality synthetic data has the potential to expedite research and development of novel technologies. However, synthetic health datasets in Canada are scarce, and no existing synthetic health datasets conform to the Findable, Accessible, Interoperable, and Reusable (FAIR) standards. Moreover, while federated machine learning offers the advantage of protecting patient privacy by not requiring the exchange of source data across nodes, it has yet to be optimized in Canada's health research environment, and there is limited use of federated learning with synthetic health data. This paper explores the ethical considerations and value proposition of generating and sharing synthetic health data. Our goal is to facilitate the development of a reliable and sustainable synthetic data infrastructure that supports the ethical, responsible, and efficient use of synthetic health data. An important contribution of this research is the establishment of a framework that balances the social benefits of innovation from data sharing with the social costs that occur when individual privacy is compromised. The use of synthetic data significantly reduces the potential for individual harm and is a cost-effective means to lower data-sharing costs. We believe that this framework will pave the way for a more robust and secure synthetic data ecosystem, enabling the generation of valuable insights that can drive positive health outcomes for Canadians.

* Corresponding author, email: helen.chen@uwaterloo.ca. We gratefully acknowledge funding support from the Office of the Privacy Commissioner of Canada Contributions Program 2023-2024. The authors declare no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper

I. Introduction

The effective management and sharing of health data, especially electronic medical records (EMRs), is often impeded by diverse information systems and data formats. This fragmentation poses significant challenges in achieving seamless integration, standardized data representation, and efficient sharing across healthcare and technology innovation ecosystems. (Kokosi et al., 2022; Kokosi & Harron, 2022). The handling of health data has been subject to stringent regulation under various laws, such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States (US), the Personal Information Protection and Electronic Documents Act (PIPEDA) in Canada, and the General Data Protection Regulation (GDPR) in the European Union (EU) (Shapiro, 2022). In their quest to test theories, models, algorithms, or prototypes, researchers and developers frequently rely on de-identified or anonymous, aggregated health data. However, it takes substantial time and resources to retrieve, aggregate, and de-identify data before it becomes accessible (Kokosi et al., 2022; Kokosi & Harron, 2022). Legal and ethical sharing of real health data remains challenging, despite initiatives like the Open Data Charter, which promotes Open Science and accessibility through data sharing (Huston et al., 2019). Nonetheless, successful examples like the Medical Information Mart for Intensive Care (MIMIC), a large, open healthcare database, have demonstrated the value of sharing real-world healthcare data for research and innovation (Johnson et al., 2016).

One potential solution to this challenge is the creation of realistic, high-quality synthetic health datasets that mimic the complexities of the original data but do not contain any real patient information (Kokosi et al., 2022). The Clinical Practice Research Datalink (CPRD) in the United Kingdom (UK) and the Agency for Healthcare Research and Quality in the US have made synthetic datasets available for research (Synthetic data, n.d.; SyH-DR, n.d). Synthetic health data can be valuable for health and data science education, ML/AI algorithm development, and health technology innovation, while safeguarding patient privacy, diversifying datasets, and enhancing health and innovative research (Gonzales et al., 2023).

In Canada, however, there is a scarcity of high-quality, sharable synthetic health datasets that adhere to Findable, Accessible, Interoperable and Reusable (FAIR) standards, despite

Canada's involvement in the Common Infrastructure for National Cohort in Europe, Canada, and Africa (CINECA) projects (CINECA, n.d.). Furthermore, there are Collective benefits, Authority to control, Responsibility, Ethics (CARE), and the First Nations principles of Ownership, Control, Access, and Possession (OCAP) pertaining to indigenous data (Carroll et al., 2021; Mecredy et al., 2018; Wilkinson et al., 2016), but applying or implementing these principles in generating and sharing synthetic health data remains both challenging and limited.

With the advent of Machine Learning (ML) and Artificial Intelligence (AI) techniques, researchers and industry have been exploring various deep learning models to generate high-quality synthetic data (Gonzales et al., 2023; Hernandez et al., 2022). Among these, generative adversarial networks (GANs) and their variants have emerged as promising synthetic data-generation approaches (Goodfellow, et al, 2014, Xu, et al, 2019, Gonzales et al., 2023; Hernandez et al., 2022; Murtaza et al., 2023). One advantage of AI techniques for generating synthetic data over conventional de-identification methods is the potential for high-fidelity data quality with a very low risk of one-to-one reverse-engineering (i.e., re-identification) back to the original health data at the population level (El Emam et al., 2020; Hernandez et al., 2022; Rajotte et al., 2022). Additionally, federated learning shows promise as another technique to safeguard data privacy and security by training AI models without centralizing datasets across multiple network nodes, therefore further reducing the potential for critical data compromises (Antunes et al., 2022; Brisimi et al., 2018). However, optimization and broader implementation remain ongoing challenges for federated learning.

Combining generative AI models and federated learning to generate synthetic health data following FAIR principles, with additional consideration of the CARE principles for Indigenous data, can create a robust and optimal health data network. Such an approach would protect sensitive patient data and accelerate health research and innovation (Antunes et al., 2022; Brisimi et al., 2018). While AI researchers are constantly seeking new ways to improve the performance of synthetic data generation, other important aspects, such as trust, transparency, and governance of synthetic health data, still require comprehensive study and consensus building. This paper presents key considerations for establishing a pan-Canadian synthetic health data ecosystem which aims to enable learning health systems and foster health technology

innovation, providing the foundation for future research and implementation in synthetic health data. An important, new contribution is the establishment of a cost-benefit framework that can be used to determine the net social benefits of the sharing of synthetic health data.

II. Synthetic Health Data Landscape

A rapid scoping review was conducted to review the literature exploring the use of synthetic health data related to health research, in general, by following the Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) and the World Health Organization's (WHO) rapid review approach (Tricco et al., 2012, 2017). Searches were completed on PubMed, Scopus, and Google Scholar using the keywords and variations listed in Table 1. We included peer-reviewed articles and grey literature written in English and published between January 1, 2012 and March 31, 2023. Meta-analyses were excluded because they yielded unrelated articles simply due to the term "synthetic" in the search queries. Methodological papers related to improvements of models or AI algorithms were also excluded due to the topic having been reviewed extensively in existing literature (Gonzales et al., 2023; Hernandez et al., 2022; Murtaza et al., 2023).

synthetic health data, synthetic data quality, synthetic data utility, synthetic data governance, synthetic data privacy, synthetic data sharing, data consents, indigenous health data, indigenous health data governance

Table 1 keywords and variants used for literature search.

Methodologies and results of this review can be found in this paper (Tsao, et. al, 2023).

Key Finding I: Data Sources Used to Generate Synthetic Health Data

Data sources used for generating synthetic health data can be broadly categorized as follows: (1) EMRs, (2) health insurance claims, (3) other administrative or health surveys, (4) bioinformatics, (5) medical images, and (6) sensor data. Depending on how synthetic health data is generated, data in these categories can be treated as longitudinal or cross-sectional in corresponding studies. Further, despite significant advancements in Natural Language Processing (NLP) techniques, particularly the recent emergence of large language models (LLMs) like

ChatGPT (Brekke et al., 2021; Tang et al., 2023), unstructured health data, such as doctors' notes in the EMRs, have been underutilized in synthetic health data generation.

Researchers often rely on open health data, such as the MIMIC-III (Johnson, et. al., 2016), in synthetic health data algorithm development. This open dataset is attractive not only because it is freely available, but also because it enables the reproducibility of research. While it is widely acknowledged that synthetic health data offers benefits such as preserving privacy, accelerating research, and driving innovations (Chen et al., 2021; Rajotte et al., 2022), some caution against its usage. Current barriers to the adoption of synthetic health data are discussed below.

Key Finding II: Concerns Regarding Research Ethics and Legal Regulations

Typically known as the Institutional Review Board (IRB) in academic health centres, or the Research Ethics Board (REB) in Canadian universities, research ethical governance bodies have universally agreed that studies involving any original or real health data require ethics approval according to operative legal regulations (Bassan & Harel, 2018; Nass et al., 2009). However, there is a lack of consensus within Canada's REBs when it comes to studies exclusively using synthetic health data. In contrast to de-identified or anonymous real patient data, the regulatory landscape governing synthetic health data and safeguarding patient privacy has remained ambiguous (Bill C-27, n.d.). This ambiguity has impacted the generation and sharing of synthetic data in Canada, with REBs granting ethics approval on a case-by-case basis for such research. Consequently, many legal and ethical questions surrounding synthetic health data in Canada remain unanswered (Arora & Arora, 2022).

In the context of the EU's GDPR, synthetic data can be categorized as pseudonymous data, anonymous data, or both, depending on the specific context of its use (López & Elbi, 2022). In the US, if synthetic data is appropriately created, it is exempt from the HIPPA regulations (Varma, 2022). Conversely, in Canada, neither PIPEDA nor the Consumer Privacy Protection Act (CPPA) explicitly addresses synthetic data, as Canada is still undergoing legal reform in this area. (Bill C-27, n.d.). Another significant concern involves patients' informed consent. In the US, HIPAA classifies the creation of de-identified data as the healthcare operations of a covered entity, thus

eliminating the requirement for informed consent from patients, even if the de-identified data will be used for research (Nass et al., 2009).

In the EU, synthetic health data can be considered as either de-identified or anonymous data under the GDPR, but regardless, informed consent is still mandated (Shapiro, 2022). In contrast, Canada's PIPEDA and CPPA generally require organizations to obtain informed consent when individuals' personal information is collected, used, or disclosed (Shapiro, 2022). The lack of clarity in PIPEDA and CPPA has made synthetic health data a grey area, resulting in potentially lengthy and inconsistent ethical reviews impeding health research and innovation.

Key Finding III: Evaluation of Synthetic Health Data

Synthetic Data generation algorithms, and the quality, utility, and privacy risks of the generated synthetic data, are interrelated. So far, there have been no universal standards to generate synthetic health data, although the use of AI/ML techniques, such as the Generative-Adversary Networks (GANs) have been gaining attention in the generation of such data (Gonzales et al., 2023; Hernandez et al., 2022; Murtaza et al., 2023). These models produce high-quality synthetic data (provided that the real-world data is also high-quality) by preserving the cohort characteristics and trends in the real-world data. However, overfitting in these models can be problematic for privacy preservation, as it can cause some synthesized records to be too similar to the real-world data (Bhanot, et al., 2021).

Given no common methods for generating synthetic health data, there are no standard evaluation metrics to assess the quality, utility, and re-identification risks of synthetic health data. Individual studies generally have their own evaluation metrics to examine how closely the synthetic health data reflect their data sources and specific use cases. Gordon et al. (2021) proposed a data-utility evaluation matrix as a framework for health data curation, which can be modified to evaluate synthetic data. Gordon's data-utility framework consists of five categories: data documentation, technical quality, coverage, access provision and value, and interest. Each category has four quality levels, ranging from bronze, the lowest quality, to platinum, the highest quality (Gordon et al., 2021). This framework can be adapted for the evaluation of synthetic health data.

Key Finding IV: Generating Synthetic Health Data in the Indigenous Data Governance Context

Currently, there are no standardized data-sharing principles for synthetic health data. However, the FAIR and CARE principles are the two main guidelines for data sharing, with the latter addressing Indigenous data (Carroll et al., 2020; Kush et al., 2020; Wilkinson et al., 2016). The FAIR principles have frequently been applied to real-world datasets, with *FAIRsharing.org* as a platform to share data that meets the FAIR principles (Sansone et al., 2019). However, implementing the FAIR principles has remained limited for synthetic health data sharing.

In the synthetic health data network, ensuring the inclusion of Indigenous people's data and upholding their right to control and access it is paramount (Carrol et al., 2021). Historical injustice and unethical treatment of Indigenous peoples have strained relationships between Indigenous peoples and the government, leading to data-sharing restrictions (Bosacrino et al., 2022). The exclusion of Indigenous datasets poses limitations in fields such as synthetic health data, which could greatly benefit from insights provided by Indigenous datasets to develop FAIR and CARE ML/AI techniques. The scarcity of representative Indigenous datasets impacts the accuracy and generalizability of generative models, inadvertently introducing biases and influencing data-driven decisions regarding Indigenous health.

To promote healthcare access and equality for Indigenous peoples, accurate Indigenous data representation is essential (Walker et al., 2017). Synthetic health data offers a promising solution to bridge this knowledge gap. However, this goal requires partnerships between the Indigenous community, AI community, and data governance agencies to address this limitation. Presently, there is no comprehensive governance framework for generating and utilizing indigenous synthetic health data. However, the FAIR and CARE principles can serve as the foundation for Indigenous synthetic data governance. The FAIR principles can act as guidelines for data producers and repositories, and the CARE principles can extend to the 'people' or 'purpose' of why that data is being used (Carrol et al., 2021). While data stakeholders inform data reuse and research reproducibility, the CARE principles can address historical inequities and provide Indigenous peoples with a platform for preserving data sovereignty.

Generative AI holds great promise for advancing health research and accelerating technology innovations. Despite its potential, generative AI's widespread adoption and utilization in the health sector are yet to be fully realized. To ensure its responsible and effective application, there is a pressing need for further research and the establishment of regulatory guidelines focusing on data governance and quality evaluation standards. In light of this, we present the following recommendations.

III. Recommendations from a FAIR and CARE Approach

Recommendation I: Establishing Synthetic Health Data Governance Guidelines

Conducting a comprehensive policy analysis concerning synthetic data governance is imperative to compare the EU's GDPR, the US's HIPAA, and Canada's PIPEDA, CPPA, and proposed Artificial Intelligence and Data Act (AIDA) (Bill C-27, n.d.). This policy analysis would aim to understand how the EU and US address synthetic health data, as compared to Canada, identify potential gaps in policies related to synthetic health data, and formulate recommendations for policy amendment or changes in Canada.

Additionally, this policy analysis would examine the methods necessary (if any) to obtain patient consent regarding the use of health data to generate synthetic health data. Following the completion of the policy analysis, a systematic three-round electronic Delphi survey (Nasa et al., 2021; Okoli & Pawlowski, 2004) will be conducted, inviting experts, data owners (i.e., patients), data custodians, and data users to co-design the ethical guidelines and synthetic health data governance framework. Key players shaping Canada's data strategy include Canadian citizens, the Canadian Institute for Health Information (CIHI), Health Canada, Statistics Canada, the First Nations Information Governance Centre (FNIGC), the ethics review boards within academic institutions, and health organizations will work together to co-design the ethical guidelines and data governance framework. Their trust and collaboration are instrumental in the successful development and deployment of synthetic health data research and data use.

Throughout the Delphi study, an extensive list of factors crucial to the governance of synthetic health data will be identified, with particular attention to factors aligned with the FAIR and CARE principles. The team will assess the level of agreement among experts and potential

users regarding the importance of these factors, aiming to bridge the differences and similarities. Should further clarification be necessary, stakeholders will be interviewed after the Delphi surveys. Furthermore, most current consent directives for EMRs do not stipulate the potential use of patients' information for producing synthetic health data to be used by hospitals and third-party entities. Policy updates need to explicitly address whether synthetic health data will be governed as human subject data or otherwise. Clear and specific policy guidelines are essential in establishing a standardized framework that governs the generation and ethical use, protection, and sharing of synthetic health data. Such guidelines will ensure transparency, accountability, and compliance with regulatory requirements, ultimately fostering trust among stakeholders and facilitating data-driven research and innovation in the health domain.

Recommendation II: Demonstrating the Responsible and Beneficial Use of Synthetic Health Data

The deep learning models for synthesizing data, the usefulness of the generated health data, and privacy concerns over the sharing of such data are interrelated. While the list of companies specializing in synthetic data generation is expanding, including Replica Analytics®, MDClone®, Mostly AI®, to name a few, it is important to note that there is no one-size-fits-all standardized algorithm or commercial tool available for generating high-quality synthetic health data. Challenges arise due to the health data's high-dimensional nature and complex interrelations. Major data custodians, such as hospitals, CIHI, and Statistics Canada, are key players in exploring the benefits and limitations of generative AI algorithms. Conducting in-depth assessments of the privacy risks and the utility associated with specific use cases of such data can offer valuable insights. By gaining deeper insights into the potential and challenges of generating and using synthetic health data, all stakeholders can make responsible and well-informed choices that align with their specific objectives and data privacy considerations. These explorations will undoubtedly pave the way for more informed decisions regarding the incorporation of synthetic health data in data-management strategies across the entire spectrum of the Canadian health research and innovation ecosystem.

Recommendation III: Establishing Standards for the Evaluation and Publication of Synthetic Health Data

The existing literature on the evaluation of synthetic health data reveals knowledge gaps that require further exploration. Because generative AI algorithms are normally tailored for specific types of data (i.e., images, tabular data, time series, or genome data) and fine-tuned for specific use cases (i.e., software testing, epidemiological study, operation optimization), it is important to assess the quality and value of synthetic health data to fulfill these diverse needs. However, there are no universal evaluation metrics to assess the performances of generative AI models and the quality of synthetic health data. Evaluation metrics will be co-developed to assess model performances and the quality of synthetic health data based on FAIR and CARE principles.

The key components of synthetic data assessment can be categorized into the following three dimensions:

Privacy Risk: The calculation of the privacy risk involves assessing the uniqueness or (re)identifiability of individuals in the dataset. Since data imbalance is common in health data, it is imperative to employ best practices to eliminate the risk of re-identification in the original data before using it to generate synthetic data. Using privacy-preserving ML/AI models to generate synthetic health data should be investigated and encouraged. When calculating the privacy risk, synthetic health data can be assessed for the risk of re-identification with a common threshold of 0.09 (El Emam et al., 2020). Additional measurements, such as cosine-similarity (NIST, n.d) between a record in the synthetic dataset and one or more records in the original dataset, can serve as valuable tools to further mitigate the risk of synthetic data records matching those in the original dataset.

Fidelity and Utility: In comparative studies of synthetic data quality, the evaluation typically assesses how closely synthetic variable distributions resemble those of the respective real-data variables. The correlation matrix of key variables serves as another common measure of gauging the fidelity between synthetic and real datasets. Additionally, the quality of synthetic data is further scrutinized by examining the performance of a machine learning model(s) built with both synthetic and real datasets (Muller et al. 2022, Foraker et al., 2020, El Emam et al.

2021). These empirical indicators provide valuable insights into the data quality of a given synthetic dataset.

However, it is important to acknowledge that such evaluations are not exhaustive and are highly dependent on specific use cases and the intended goals of the model. Nonetheless, establishing a minimal set of fidelity measurements and benchmarking machine learning models for assessing the utility of synthetic data will mark a significant step towards standardizing synthetic data quality measures. Publishing these quality indicators will bolster confidence in synthetic data generation and ensure responsible and informed usage by third parties, equipped with explicit knowledge of the limitations inherent in the synthetic dataset.

Cost-Benefit Assessment: Similar to health technology assessment (HTA), there is a need for a cost-benefit framework that quantifies the societal impacts of generating and using synthetic health data for research and commercial purposes. The first challenge in developing this cost-benefit analysis framework is accurately capturing the costs and benefits to each stakeholder. The first step involves recommending a measurement to calculate the benefits of creating and sharing synthetic health data. In HTA, the quality-adjusted life years (QALY) is used to evaluate how well medical treatments lengthen and/or improve patients' lives. In the evaluation of synthetic health data, the benefit derived from its utilization may not be directly reflected in QALY. The benefit measurement may require consideration of additional factors, including the likelihood of improved patient well-being resulting from the proposed research or commercial products. This encompasses advanced data analytics and the value of innovative technology, which could lead to better patient outcomes and optimized healthcare delivery.

The second challenge is estimating the probability of individual privacy compromise when releasing synthetic health data. Privacy breaches could arise either through membership inferencing or reverse engineering of the synthetic health data itself, or its combination with other external sources of information. Merely assessing re-identification probability is insufficient to fully evaluate the overall costs of synthetic health data release. It is important to estimate the potential harm that could be inflicted upon individuals to capture the total potential costs borne by stakeholders.

Furthermore, this framework must consider the diverse range of stakeholder groups, considering the potential systemic harmful effects experienced by certain marginalized communities. Improper handling of synthetic data generation could exacerbate systematic bias in the original data. Therefore, a comprehensive approach is necessary to ensure that equity, diversity, and inclusion (EDI) considerations are effectively integrated into the cost-benefit analysis.

As we consider the broader use of synthetic health data, we must be vigilant in avoiding the "garbage in, garbage out" problem, recognizing the significance of the quality of the original (i.e., real) data. Synthetic data presents a promising venue to mitigate privacy risks while permitting the sharing of data, thereby promoting the democratization of data for diverse applications. Such data usage will help identify potential data gaps and quality issues and incentivize data owners and custodians to prioritize the maintenance of high-quality original data. This, in turn, fosters a positive cycle of data-driven advancement in Canada's health ecosystem.

IV. Establishing Cost-Benefit Guidelines for Health Data Sharing

Building on the above discussion, this section establishes a cost-benefit framework for sharing health data, along with implications for synthetic data use. There is a general dearth of cost-benefit methods that can be used to evaluate societal net benefits from data sharing. Cost-benefit analysis in health research, including drug invention, is typically calculated using Quality Adjusted Life years (QALY), which cannot be used to directly evaluate the impact of data sharing. From a more general perspective, most cost-benefit methods are based on economic analyses that are primarily focused on net gains to different stakeholders. For example, in evaluating the societal impacts of merger decisions between firms, the Competition Bureau of Canada compares the economic costs that might occur to consumers who might pay higher prices as a result of a lessening of market competition, which are then weighed against gains to merging parties in the form of cost savings.[1] The methodology used by the Competition Bureau is known as 'Total Surplus.' The Treasury Board of Canada (2022) also recommends a similar approach in weighing

[1] Please see Competition Bureau of Canada (2011) Merger Enforcement Guidelines for further details.

quantifiable benefits to different stakeholders versus corresponding costs, to determine the desirability of competing public-sector projects. If the marginal or incremental benefits of such projects exceed the incremental costs, then the project should be undertaken. It is important to emphasize that the benefits and costs that should be measured, are those that would otherwise have not occurred but for the project.[2]

Returning to the example of merger analysis, the key economic cost is the 'deadweight loss' that occurs, if the merging firms have the market power to increase prices to consumers. Higher prices should lead to increased profits if demand for the product does not decline significantly. Part of these profits are generated at the expense of consumers who continue purchasing the goods, but at higher prices. The Competition Bureau does not count the higher prices paid by consumers as an economic cost of the merger. This is considered a transfer of economic surplus from consumers to producers. However, higher prices also imply that there will be consumers who can no longer afford the goods. The deadweight loss to society is the lost value that these consumers placed on the goods, which they can no longer afford. Firms also lose some profits if higher prices result in reduced demand. This lost 'producer surplus' is also counted as lost surplus. The total deadweight loss to consumers and producers are weighed against the gains to the merging firms in terms of 'efficiencies,' which are reductions in the incremental costs of doing business, which would not have occurred but for the merger. If these efficiency gains to the merging firms exceed the deadweight loss, then the merger will not be blocked by the Competition Bureau. This is true, even if the price increases from the merger are significant.

Hence, merger analysis is based on stakeholder analysis, where the incremental costs are deducted from the incremental benefits, to evaluate gains to society. To accomplish this, it is necessary to assume that all stakeholders value a dollar equally, which in some cases, might be a strong assumption. Further, no stakeholder receives preferential treatment, as the use of the Total Surplus approach implies that dollar gains/losses to producers and consumers are comparable. In other words, if a consumer gains a dollar at the expense of a firm or producer,

[2] For these guidelines please refer to the Treasury Board of Canada (2022) Guide to Costing

this is just a transfer of resources, and there is no net gain to society. Our construction of a cost-benefit approach to data sharing is based on similar principles.

Our objective is to understand the optimal amount of data access for society. Hence, the first step is to define the stakeholders who benefit from enhanced data sharing. Assume the existence of data (*d*), which is of interest to academic researchers who are interested in creating knowledge that will result in some benefit (*B*) to society. Further, there is a data custodian who has authority with respect to granting researchers third-party access to data. The data custodian does not gain any direct benefit from allowing access to data, but experiences costs (*C*) in establishing infrastructure that stores data and enables access by researchers. The data custodian must use a societal framework model to evaluate whether allowing data access leads to net gains for society. As noted above, releasing the data allows third parties to conduct research and extract insights which leads to a societal benefit. To simplify our model, we will assume that the data may only be used by academics for publishable research. The mandate of the data custodian is to ensure that societal benefits from knowledge creation are maximized while minimizing the probability of individual privacy being compromised. This is because the trade-off to knowledge creation from allowing access to data, is the probability that individual-specific information contained in the database may be revealed, despite the use of privacy preserving technology by the data custodian. The cost of employing these technologies is also captured by *C*.

If an individual's privacy is compromised, then we assume that they will experience some harm (*H*), which can be monetary or non-monetary. The variable H captures both the probability of being harmed as well as the actual monetary and non-monetary amount of harm. We assume that sharing more data leads to a higher probability of data being accessed by unauthorized third parties, and therefore, an increase in possible harm (*H*) to individuals, which increases at an increasing rate. Hence, if $H = H(d)$, then $\partial H/\partial d > 0$ and $\partial^2 H/\partial d^2 > 0$. In terms of other costs, the data custodian experiences economies of scale in maintaining data infrastructure and the employment of data-protection technologies. Specifically, while there are significant upfront costs in creating the infrastructure, average and marginal costs are equal and decline with each

unit of data held by the custodian. Therefore, $\partial C/\partial d < 0$ and $\partial^2 C/\partial d^2 < 0$. As noted above, the other cost in this model is the possible harm to individuals whose information is somehow retrieved by third parties who do not have permission to access the data. If societal costs (*SC*) are the sum of the operations costs of the data custodian and possible harm to individuals then the societal cost curve will be initially falling, but will start rising with increasing harm, which will occur as more data are released. We initially assume that the data custodian is not liable for the harm experienced by individuals who experience a data breach. The gap between the private cost curve of the data custodian and the societal cost curve is the amount of harm to individuals from privacy breaches.

Societal benefits from knowledge creation initially are an increasing function of data access or each unit of data. In terms of notation, this is $B = B(d)$. However, there is a plateau, beyond which no further benefits are created, even through access to more data. Hence, $\partial B/\partial d > 0$ and $\partial^2 B/\partial d^2 < 0$ up to a certain threshold, $B = \bar{B}$, after which $\partial B/\partial d = 0$. The optimal amount of data release is then defined at the point where the slopes of the marginal benefit and marginal costs (of the data custodian) are equal at zero. Figure 1 shows the basic equilibrium with the societal benefit and cost curves and the cost curve of the data custodian.

The privately optimal amount of data release, if the only costs are those experienced by the data custodian, is given by $d = \bar{d}$. Taking harm to individuals into consideration, the optimal amount of data release is given by $d = d^*$, where the slope of the societal benefit and cost curves are equal. This will be to the left of $\bar{d}$ as the SC curve will start rising before the minimum point of the custodian cost curve. This model provides a framework for estimating net societal benefits from data sharing that is analogous to approaches used by other agencies. The important lesson is that the data custodian must compare the marginal benefits of data release against the corresponding marginal social costs. These benefits and costs can be calculated by the corresponding areas beneath the curves with respect to the amount of data being released.[3] For example, suppose the data custodian receives a request for d_2-d_1 amount of data. The marginal

[3] Of course, an underlying assumption is that equations for these curves can be estimated. This is possible through econometric methods developed by economists, as long as relevant data are available.

benefit to society is given by area ABCD. The corresponding marginal social cost is CDEF. As society will receive a marginal benefit of ABCD, it makes sense for the data custodian to release the data.

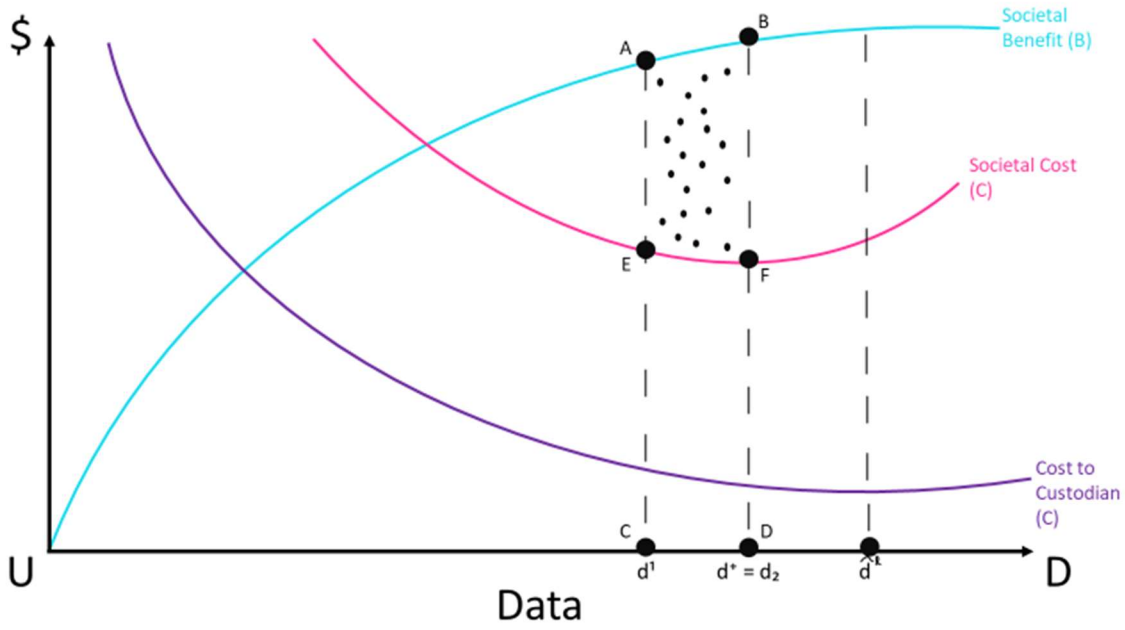

Figure 1 Basic Cost-Benefit Model

Let us further explore the implications of this model through a hypothetical example. Suppose the data custodian is responsible for individual patient records at a hospital. The custodian decides to release data in response to a request from an accredited researcher interested in developing new insights on patient treatment. Releasing the data always carries some risk of individual identification. Recent research has demonstrated that re-identifying individuals is possible even when released data are a partial sample of the entire dataset and with a limited number of variables (Rocher, et. al,2019). However, there is less consensus on the precise harm an individual may experience if they are re-identified in a dataset. In the absence of available published data for Canada, we shall assume that the individual experiences a loss of $20,000. This could occur if successful re-identification leads to information that enables a third party to launch a successful phishing or ransomware attack.[4]

[4] https://www.cbc.ca/news/canada/toronto/bmo-scam-line-of-credit-two-factor-1.6947461.

What must also be considered is the probability of a successful attack, which is different from the probability of re-identification. For example, even if an individual's information becomes available to a third party, the third party may not be able to execute a successful attack if the individual is sufficiently educated on not to respond to phishing attacks or has suitable antivirus/anti-malware software installed on their computers. In summary, the above discussion suggests that while monetary harm to an individual from a successful attack might be significant, the probability of a successful attack may be moderate. Of course, given the absence of data, it is difficult to be conclusive, and it is important to acknowledge differences across different demographic groups. For example, studies have shown the elderly to be particularly susceptible to successful cyberattacks. For simplicity, we assume that individual harm (H) from a cyberattack facilitated by the release of patient-level information for research purposes, is captured by the below function:

$$H = H(p_i, p_a, M) = p_i * p_a * M$$

Where p_i = probability of being re-identified from released data, p_a = probability of a successful cybersecurity attack, and M = monetary damage from the cybersecurity attack. Essentially, this functional form captures expected harm from data release.

With the above functional form, an increase in each of the above factors will lead to potentially greater harm to individuals who are re-identified from released data. Measuring individual harm in this manner allows us to capture how different communities may be affected by cybersecurity attacks. For example, particular groups may be more likely to be prone to successful attacks, such as seniors who are less likely to be aware of adequate cybersecurity protection measures. The harm to them is also probably going to be higher.

An Example

Assume that a group of university researchers have submitted a research proposal to access confidential, individual patient-level data and the data custodian must evaluate the net benefits to society. The objective of the proposed study is to improve decision-making processes that have the potential to improve survival rates for cancer patients. This is consistent with Kaur et al. (2022), who find that data mining and machine learning techniques are significantly

impacting hospital decision-making processes, because of access to large amounts of digital data that are typically stored in hospitals (Kaur et al. (2022). As noted by the authors, data mining and machine learning methods offer extremely promising new ways to offer effective cancer prognoses. In their words:

"*Cancer diagnosis techniques aim to detect the occurrence of cancer in a particular body organ of a person. In most cases, clinicians go for a biopsy where a tissue sample is removed from the body and analyzed for detection of cancer cells. Depending on the organ and type of cancer, different detection methods are available, and correspondingly researchers use some initial parameters, gene biomarkers, or CT images, etc. However, the diagnostic systems are not very efficient in performance. Since cancer is primarily asymptomatic in earlier stages, it is helpful to identify novel diagnostic techniques to reduce cancer mortality cases. With data mining in clinical research, various researchers have utilized different machine learning techniques to diagnose or classify cancer in other parts of the body*."

In their review of the literature, Kaur et al. (2022) note that the following patient-level information is typically used by studies: (1) Age, Gender, Marital Status; Race; (2) all cancer-related information diagnosed at the time of examination by the clinician such as Tumor size, Cancer type, Stage of cancer, Primary site; (3) Type of treatment (type of surgery, chemotherapy cycles, androgen deprivation therapy); and (4) Lifestyle attributes such as Smoking/tobacco/alcohol, and other Comorbidities. Guo et al. (2021) use such features in machine learning models to estimate the probability of a patient's individual post-surgery conditional risk of death, risk of recurrence, and risk of site-specific recurrence for cervical cancer (Guo et al., 2021). With access to data from multiple institutions, the authors are able to identify machine learning methods that offer superior predictions relative to traditional logistic or Cox regression models in estimating prognostic outcomes. Most studies on cervical cancer focus on estimating the probability of recurrence or death. In contrast, a key contribution of this study is the use of advanced models to evaluate the importance of specific recurrence sites (local or distant recurrence), which is essential for planning appropriate follow-up strategies. The authors use their research findings to construct a web-based calculator to predict postoperative survival and site-specific recurrence in cervical cancer patients, which can be employed by clinicians for

such strategies. This is important given the traditional reliance by clinicians in using their experience and knowledge to assign patients into crude categories as low- or high-risk groups to determine post-operative care, without accurately accounting for the specifics of each unique patient.

This example clearly illustrates the societal benefits resulting from access to confidential data. In terms of quantifying these benefits, we can use the value of a statistical life recommended by the Treasury Board of Canada (2022). The Value of a Statistical Life (VSL) is typically used to determine compensation for individuals involved in workplace accidents. VSL monetary amounts recommended by studies conducted by economists range from $5 million - $6 million. The Treasury Board guidelines note that with a discount rate of 5%, the $5 million value of a statistical life translates into a value of a statistical life-year of $305,000, and an estimate of $143,000 per life-year given a zero-discount rate. Suppose that improved post-operative treatment recommended by machine learning algorithms from outcomes facilitated by access to confidential data results, on average, in an increase in survival by 10 years for all cervical cancer patients treated in a hospital. If the hospital treats 20 patients on annual basis, using a conservative life-year value of $150,000, a back-of-the-envelope calculation implies that aggregate societal benefits generated for patients treated at the hospital are $30 million.

However, assume that the sharing these data has compromised the privacy of 1,000 individuals in the dataset. Further, 80% of individuals have been subjected to successful cyberattacks, with an average corresponding financial loss is $30,000. Hence, $p_i = 1$, $p_a = 0.8$, M= $30,000. Therefore, harm to each individual = $H = H(p_i, p_a, M) = p_i * p_a * M = 1 * 0.8 * 30{,}000 = \$24{,}000$. Consequently, aggregate societal harm = 1,000 * $24,000 = $24 million. If a dollar is valued equally by all stakeholders in this analysis, then sharing the data has made society better off, even though the number of people impacted by the data breach exceeded the number of patients who experienced improved survival longevity as Societal Net Benefit = Societal Marginal Benefit – Societal Social Cost = $30 m. - $24 m. = $6 million. Of course, the results are sensitive to the underlying assumptions. For example, significantly increasing the number of individuals experiencing the data breach to 5,000 implies a Societal Social Cost of $120 m. On the other hand, using a $305,000 VSL (based on a 5% discount rate) for 10 years for 50 patients, translates to a

Marginal Societal Benefit of $152,000. This methodology can also be extended to calculating societal benefits from reduced morbidity.

This discussion illustrates the sensitivity of findings to underlying assumptions. However, it also suggests that individual harm can be significantly attenuated by proper data-sharing and data-protection protocols agreed upon between the data custodian and the researchers. Further, while we cannot prove this, our belief is that the benefit of an additional life-year to an individual is extremely high, while the individual costs from data breaches, which are ubiquitous, might be more limited. We, of course, cannot say this for certain, given the absence of relevant data and research, but it is not an unreasonable assumption. Nonetheless, there is a need for more study on individual costs from unauthorized access to personal data.

Example with Liability and Synthetic Data

The above model and example do not consider the reputational harm and liability potentially experienced by the data custodian in the event of a successful data breach. The reason is that a fine paid by the data custodian is simply a monetary transfer to the impacted party, and there is no overall benefit to society. However, making the data custodian liable certainly incentivizes them to ensure adequate security protocols. The key point is that from the perspective of efficiency, the fine must be proportionate to harm experience through privacy loss. If the fine is much higher than the actual harm, then the amount of data sharing will be considerably reduced, leading to much lower net societal benefit to society.

This can be seen in Figure 2, where the data custodian's cost curve shifts up to C_2, as a result of the liability imposed on the data custodian, which results in nearly an efficient amount of data sharing, with d_1 being slightly more than d^*. Hence, the optimal fine is the monetary amount equal to the marginal damage experienced by individuals at the efficient level of data sharing, or $d = d^*$. This fine is analogous to the motivation behind carbon taxes, which theoretically is set to the amount of environmental damage at the point where societal benefit is equal to societal costs. The curve C_3 represents a scenario with a fine that exceeds true marginal harm to individuals, leading to a suboptimal amount of data release (d_2). An extremely large fine can be interpreted as being 'punitive,' and governments wanting to send a strong message to

corporations regarding the priority to be given to the protection of individual privacy. However, an alternative approach might be through the provision of synthetic data.

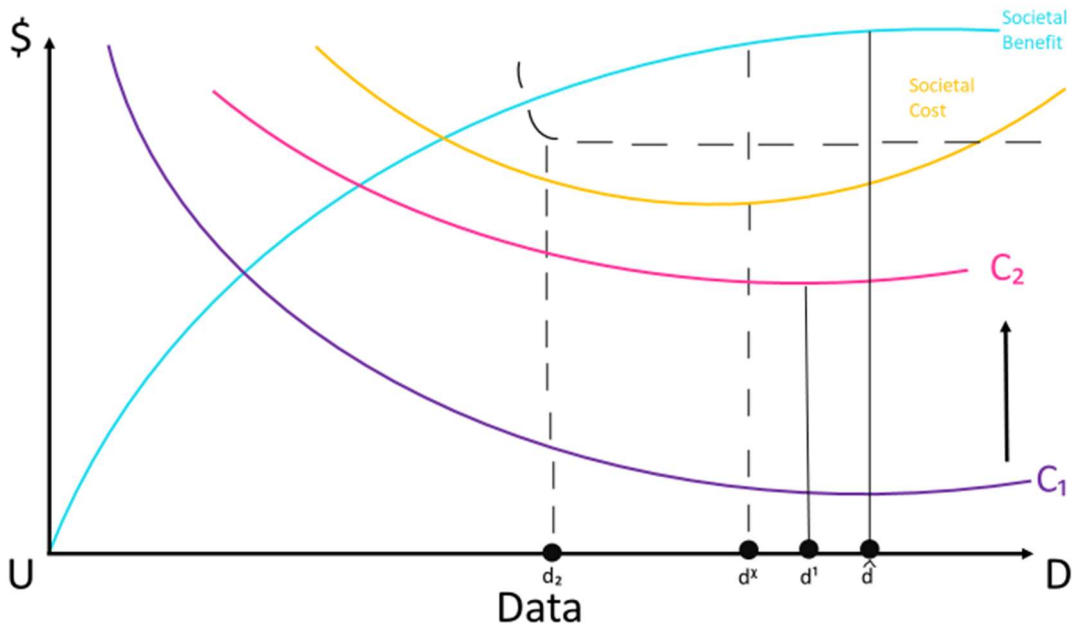

Figure 2 Cost-Benefit Analysis with Synthetic Data

Specifically, releasing synthetic data instead of real individual information, considerably reduces the degree of potential individual harm from data breaches. This could theoretically ensure that the data custodian is at the amount of efficient data sharing, $d = d^*$, with its cost curve shifting back close to C_1, assuming that the costs of producing synthetic data are low. This is an extremely important point, as it implies significantly reduced liability for firms, even in the event of a data breach. Hence, overall societal costs from data sharing will be lowered, with reduced privacy costs to individuals, and therefore, diminished liability costs to custodians as well. However, the trade-off is the possibility of reduced societal benefits, if the synthetic data are unable to produce insights that are comparable to what can be obtained from the real individual-level data. In this case, the societal benefit curve B_1 shifts down to B_2 and even releasing data $d = d^*$ leads to much lower net social benefits relative to what would occur through the release of real data. This can be illustrated with the example that we previously used.

As before, we assume that the privacy of 1,000 individuals in a dataset has been compromised. However, instead of 80%, only 10% of individuals have been subjected to successful cyberattacks, as the data accessed by cyber criminals is synthetic and not real data. As before, the financial loss to each impacted individual is $30,000. Hence, $p_i = 1$, $p_a = 0.10$, and $M = \$30{,}000$. Therefore, harm to each individual = $H = H(p_i, p_a, M) = p_i * p_a * M = 1 * 0.1 * 30{,}000 = \$3{,}000$. Therefore, aggregate societal harm = 1,000 * $3,000 = $2 million. Further assuming that the insights from the synthetic data were less useful, thus, as a result of the analysis of the data, instead of 10 years (in the previous example), cervical cancer patients only experience an increase in life expectancy of one additional year. Using the same conservative life-year expectancy value of $150,000 for 20 patients, leads to a societal benefit of $3 million. While the net social benefit is still positive at $3 million - $2 million = $1 million, it is reduced.

V. Conclusion

The movement towards producing high-quality synthetic data from machine learning and artificial intelligence methods holds significant promise in terms of producing innovative research while preserving individual privacy. What is therefore needed are data governance principles that should be followed in Canada, to ensure ethical generation and use of such data, which also respect and accommodate the concerns and needs of Indigenous peoples. This work provides recommendations for applying the FAIR and CARE principles, which are able to achieve such objectives. This paper also presents a cost-benefit framework that can be used to determine efficient levels of data sharing, aimed at balancing societal innovation against possible harm to individuals from compromises to their privacy. Our framework can be adapted to evaluate the net social benefits associated with the use of synthetic health data. Synthetic health data can greatly reduce societal costs in the form of lower risk to individual privacy from data breaches, and hence, lower liability costs for data custodians. However, the trade-off is the possibility of less useful insights from data analysis. How much less remains an area in need of further study.

References

Antunes, R. S., André da Costa, C., Küderle, A., Yari, I. A., & Eskofier, B. (2022). Federated learning for healthcare: Systematic review and architecture proposal. *ACM Transactions on Intelligent Systems and Technology*, 13(4), 1–23. https://doi.org/10.1145/3501813

Arora, A., & Arora, A. (2022). Synthetic patient data in health care: a widening legal loophole. *Lancet*, *399*(10335), 1601–1602. https://doi.org/10.1016/s0140-6736(22)00232-x

Bassan, S., & Harel, O. (2018). The ethics in synthetics: statistics in the service of ethics and law in health related research in big data from multiple sources. *Journal of Law and Health*, 31(1), 87-117.

Bhanot, K., Qi, M., Erickson, J. S., Guyon, I., & Bennett, K. P. (2021). The problem of fairness in synthetic healthcare data. *Entropy* (Basel, Switzerland), 23(9), 1165. https://doi.org/10.3390/e23091165

Bill C-27: An Act to enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act and to make consequential and related amendments to other Acts. (2022, November 4). Justice.Gc.Ca. https://www.justice.gc.ca/eng/csj-sjc/pl/charter-charte/c27_1.html

Boscarino, N., Cartwright, R. A., Fox, K., & Tsosie, K. S. (2022). Federated learning and Indigenous genomic data sovereignty. *Nature Machine Intelligence*, 1-3. https://doi.org/10.1038/s42256-022-00551-y

Brekke, P. H., Rama, T., Pilán, I., Nytrø, Ø., & Øvrelid, L. (2021). Synthetic data for annotation and extraction of family history information from clinical text. *Journal of Biomedical Semantics*, *12*(1), 11. https://doi.org/10.1186/s13326-021-00244-2

Brisimi, T. S., Chen, R., Mela, T., Olshevsky, A., Paschalidis, I. C., & Shi, W. (2018). Federated learning of predictive models from federated Electronic Health Records. *International Journal of Medical Informatics*, 112, 59–67. https://doi.org/10.1016/j.ijmedinf.2018.01.007

Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J. D., Anderson, J., & Hudson, M. (2020). The CARE principles for indigenous data governance. *Data Science Journal*, 19. https://doi.org/10.5334/dsj-2020-043

Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F. K., & Mahmood, F. (2021). Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, *5*(6), 493–497. https://doi.org/10.1038/s41551-021-00751-8

CINECA - common infrastructure for national cohorts in Europe, Canada, and Africa. (n.d.). CINECA. Retrieved December 12, 2022, from https://www.cineca-project.eu/

Competition Bureau of Canada (2011). *Merger Enforcement Guideline*. Retrieved June 1st 2023, from https://ised-isde.canada.ca/site/competition-bureau-canada/en/how-we-foster-competition/education-and-outreach/publications/merger-enforcement-guidelines#s12_0.

El Emam, K., Mosquera, L., & Bass, J. (2020). Evaluating identity disclosure risk in fully synthetic health data: Model development and validation. *Journal of Medical Internet Research*, *22*(11), e23139. https://doi.org/10.2196/23139

El Emam K, Mosquera L, Jonker E, Sood H. Evaluating the utility of synthetic COVID-19 case data. JAMIA Open. 2021 Mar 1;4(1):ooab012. doi: 10.1093/jamiaopen/ooab012. PMID: 33709065; PMCID: PMC7936723.

Foraker, RE, and others, Spot the difference: comparing results of analyses from real patient data and synthetic derivatives, *JAMIA Open*, Volume 3, Issue 4, December 2020, Pages 557–566, https://doi.org/10.1093/jamiaopen/ooaa060

Gonzales, A., Guruswamy, G., & Smith, S. R. (2023). Synthetic data in health care: A narrative review. *PLOS Digital Health*, *2*(1), e0000082. https://doi.org/10.1371/journal.pdig.0000082

Goodfellow Ian, Pouget-Abadie Jean, Mirza Mehdi, Xu Bing, Warde-Farley David, Ozair Sherjil, Courville Aaron, and Bengio Yoshua. 2014. Generative adversarial nets. In *International Conference on Advances in Neural Information Processing Systems*. 2672–2680.

Gordon, B., Barrett, J., Fennessy, C., Cake, C., Milward, A., Irwin, C., Jones, M., & Sebire, N. (2021). Development of a data utility framework to support effective health data curation. *BMJ Health & Care Informatics*, 28(1), e100303. https://doi.org/10.1136/bmjhci-2020-100303.

Guo, Chenyan, Wang, Jue, Wang, Yongming, Qu, Xinyu, Shi, Zhiwen, Meng, Yan, Qiu, Junjun, and Hua, Keqin (2021). Novel artificial intelligence machine learning approaches to precisely predict survival and site-specific recurrence in cervical cancer: A multi-institutional study. *Translational Oncology*, 14(5), 101032, ISSN 1936-5233, https://doi.org/10.1016/j.tranon.2021.101032.

Hernandez, M., Epelde, G., Alberdi, A., Cilla, R., & Rankin, D. (2022). Synthetic data generation for tabular health records: A systematic review. *Neurocomputing*, 493, 28–45. https://doi.org/10.1016/j.neucom.2022.04.053

Huston, P., Edge, V. L., & Bernier, E. (2019). Reaping the benefits of Open Data in public health. *Releve Des Maladies Transmissibles Au Canada [Canada Communicable Disease Report]*, *45*(11), 252–256. https://doi.org/10.14745/ccdr.v45i10a01

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-W. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1), 160035. https://doi.org/10.1038/sdata.2016.35

Kaur, Ishleen, Doja, M.N. Doja, & Ahmad, & Ahmad, Tanvir (2022). Data mining and machine learning in cancer survival research: An overview and future recommendations. *Journal of Biomedical Informatics*, 128(104026), ISSN 1532-0464, https://doi.org/10.1016/j.jbi.2022.104026.

Kokosi, T., De Stavola, B., Mitra, R., Frayling, L., Doherty, A., Dove, I., Sonnenberg, P., & Harron, K. (2022). An overview on synthetic administrative data for research. *International Journal for Population Data Science*, 7(1). https://doi.org/10.23889/ijpds.v7i1.1727

Kokosi, T., & Harron, K. (2022). Synthetic data in medical research. *BMJ Medicine*, 1(1), e000167. https://doi.org/10.1136/bmjmed-2022-000167

Kush, R. D., Warzel, D., Kush, M. A., Sherman, A., Navarro, E. A., Fitzmartin, R., Pétavy, F., Galvez, J., Becnel, L. B., Zhou, F. L., Harmon, N., Jauregui, B., Jackson, T., & Hudson, L. (2020). FAIR data sharing: The roles of common data elements and harmonization. *Journal of Biomedical Informatics*, 107(103421), 103421. https://doi.org/10.1016/j.jbi.2020.103421

López, C. A. F., & Elbi, A. (2022, September 22). *On synthetic data: a brief introduction for data protection law dummies*. European Law Blog. https://europeanlawblog.eu/2022/09/22/on-synthetic-data-a-brief-introduction-for-data-protection-law-dummies/

Mecredy, G., Sutherland, R., & Jones, C. (2018). First Nations data governance, privacy, and the importance of the OCAPÂ® principles. *International Journal for Population Data Science*, *3*(4). https://doi.org/10.23889/ijpds.v3i4.911

Muller, E., Zheng, X. and Hayes, J. Evaluation of the Synthetic Electronic Health Records. In Proceedings of the 1st Workshop on Scarce Data in Artificial Intelligence for Healthcare (SDAIH 2022), pages 17-22 DOI: 10.5220/0011531300003523

Murtaza, H., Ahmed, M., Khan, N. F., Murtaza, G., Zafar, S., & Bano, A. (2023). Synthetic data generation: State of the art in health care domain. *Computer Science Review*, *48*(100546), 100546. https://doi.org/10.1016/j.cosrev.2023.100546

Nasa, P., Jain, R., & Juneja, D. (2021). Delphi methodology in healthcare research: How to decide its appropriateness. *World Journal of Methodology*, *11*(4), 116–129. https://doi.org/10.5662/wjm.v11.i4.116

Nass, S. J., Levit, L. A., Gostin, L. O., & Institute of Medicine (US) Committee on Health Research and the Privacy of Health Information: The HIPAA Privacy Rule. (2009). HIPAA, the Privacy Rule, and its application to health research. *National Academies Press*. https://www.ncbi.nlm.nih.gov/books/NBK9573/

National Institute of Standards and Technology (NIST) (n.d) COSINE DISTANCE, COSINE SIMILARITY, ANGULAR COSINE DISTANCE, ANGULAR COSINE SIMILARITY (nist.gov)

Okoli, C., & Pawlowski, S. D. (2004). The Delphi method as a research tool: an example, design considerations and applications. *Information & Management*, *42*(1), 15–29. https://doi.org/10.1016/j.im.2003.11.002

Rajotte, J.-F., Bergen, R., Buckeridge, D. L., El Emam, K., Ng, R., & Strome, E. (2022). Synthetic data as an enabler for machine learning applications in medicine. *IScience*, 25(11), 105331. https://doi.org/10.1016/j.isci.2022.105331

Rocher, L., Hendrick, J.M. & de Montjoye, YA. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*, 10, 3069. https://doi.org/10.1038/s41467-019-10933-3.

Sansone, S.-A., McQuilton, P., Rocca-Serra, P., Gonzalez-Beltran, A., Izzo, M., Lister, A. L., Thurston, M., & FAIRsharing Community. (2019). FAIRsharing as a community approach to standards, repositories and policies. *Nature Biotechnology*, 37(4), 358–367. https://doi.org/10.1038/s41587-019-0080-8

Shapiro, J. (2022, May 20). Why digital privacy is so complicated. Progressive Policy Institute. https://www.progressivepolicy.org/publication/why-digital-privacy-is-so-complicated/

Synthetic data. (n.d.). Cprd.com. Retrieved December 12, 2022, from https://cprd.com/synthetic-data

Synthetic healthcare database for research (SyH-DR). (n.d.). Ahrq.gov. Retrieved December 12, 2022, from https://www.ahrq.gov/data/innovations/syh-dr.html

Tang, R., Han, X., Jiang, X., & Hu, X. (2023). *Does synthetic data generation of LLMs help clinical text mining?* https://doi.org/10.48550/ARXIV.2303.04360

Tricco, A. C., Langlois, E. V., & Straus, S. E. (2017). *Rapid reviews to strengthen health policy and systems: A Practical Guide*. World Health Organization.

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D. J., Horsley, T., Weeks, L., Hempel, S., Akl, E. A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M. G., Garritty, C., … Straus, S. E. (2018). PRISMA extension for Scoping Reviews (PRISMA-ScR): Checklist and explanation. *Annals of Internal Medicine*, 169(7), 467–473. https://doi.org/10.7326/M18-0850

Treasury Board of Canada (2023). Canada's Cost-Benefit Analysis Guide for Regulatory Proposals. Retrieved June 1st 2023, from, https://www.canada.ca/en/government/system/laws/developing-improving-federal-regulations/requirements-developing-managing-reviewing-regulations/guidelines-tools/cost-benefit-analysis-guide-regulatory-proposals.html.

Tsao SF, Sharma K, Noor H, Forster A, Chen H. Health Synthetic Data to Enable Health Learning System and Innovation: A Scoping Review. Stud Health Technol Inform. 2023 May 18;302:53-57. doi: 10.3233/SHTI230063. PMID: 37203608.

Walker, J., Lovett, R., Kukutai, T., Jones, C., & Henry, D. (2017). Indigenous health data and the path to healing. *Lancet*, 390(10107), 2022-2023. https://doi.org/10.1016/S0140-6736(17)32755-1

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1), 160018. https://doi.org/10.1038/sdata.2016.18

Xu L., Skoularidou M., Cuesta-Infante A. and Veeramachaneni K.,2019, Modeling tabular data using Conditional GAN. In *Advances in Neural Information Processing Systems*, Wallach H., Larochelle H., Beygelzimer A., d'Alché-Buc F., Fox E., and Garnett R. (Eds.), Vol. 32. Curran Associates, Inc.