MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI231079

An Approach for Generating Realistic Australian Synthetic Healthcare Data

Ibrahima DIOUF[a,1] , John GRIMES[a], Mitchell J. O'BRIEN[a], Hamed HASSANZADEH[a], Donna TRURAN[a], Hoa NGO[a], Parnesh RANIGA [a], Michael LAWLEY[a], Denis C. BAUER[a,b,c], David HANSEN[a], Sankalp KHANNA[a] and Roc REGUANT[a]
[a] *Australian e-Health Research Centre, Commonwealth Scientific and Industrial Research Organisation, Australia*
[b] *Macquarie University, Department of Biomedical Sciences, Faculty of Medicine and Health Science, Macquarie Park, Australia*
[c] *Macquarie University, Applied BioSciences, Faculty of Science and Engineering, Macquarie Park, Australia*
ORCiD ID: Ibrahima Diouf https://orcid.org/0000-0002-9672-303X, John Grimes https://orcid.org/0000-0002-9575-7641, Mitchell J. O'Brien https://orcid.org/0000-0003-0662-9101, Denis Bauer https://orcid.org/0000-0001-8033-9810, Sankalp Khanna https://orcid.org/0000-0002-9672-303X, Roc Reguant https://orcid.org/0000-0002-0350-3899

Abstract. Healthcare data is a scarce resource and access is often cumbersome. While medical software development would benefit from real datasets, the privacy of the patients is held at a higher priority. Realistic synthetic healthcare data can fill this gap by providing a dataset for quality control while at the same time preserving the patient's anonymity and privacy. Existing methods focus on American or European patient healthcare data but none is exclusively focused on the Australian population. Australia is a highly diverse country that has a unique healthcare system. To overcome this problem, we used a popular publicly available tool, Synthea, to generate disease progressions based on the Australian population. With this approach, we were able to generate 100,000 patients following Queensland (Australia) demographics.

Keywords. Synthetic healthcare data, data simulation, Synthea, Australia

1. Introduction

Healthcare data is essential for medical research, but its access remains challenging. Despite improvements in data collection and storage, access to healthcare data is still scarce [1]. Sharing patient data is one of the biggest challenges in research because researchers need to conform to regulations and obtain consent from multiple entities. Restriction of access to healthcare data is the result of preserving patient privacy [2]. At

[1]Corresponding Author: Ibrahima Diouf, The Australian e-Health Research Centre | CSIRO Health & Biosecurity, Phone: +61 3 9662 7115| Mobile:+61 467620610, email: Ibrahima.diouf@csiro.au | www.csiro.au | www.aehrc.com

the same time, data collection is costly; thus, preventing the collection of specific datasets for each project. To maintain privacy, anonymisation techniques can be used. Anonymisation is the process of sanitisation by which the individuals, whom the dataset describes, remain anonymous [3].

However, anonymised data comes at a cost. Aside from the overhead of creating a new data resource, anonymisation can cause information loss [4]. Moreover, it has been found that re-identification of individuals is possible even with anonymised data [1]. Although the risk of re-identification is low, potential exposure of sensitive data can have a high impact on the patients. Technological advances can increase the risk of re-identification. Synthetic data generation can overcome the threats to a patient's privacy and is a promising solution to minimise the risk of leaking sensitive data [5]. Synthetic data is artificially generated data that emulates existing properties of the original dataset without leaking individual details. Healthcare data generation follows two main approaches. The first is based on machine learning methods, and the second is based on statistics, generating a sample of the population compatible with observed summary statistics and disease prevalence. Machine learning methods require an initial body of training data that satisfies regulatory and consent requirements. After synthetic dataset has been generated with machine learning, researchers need to make sure that the generated data does not reproduce the original resource for any given patient [3].

Existing synthetic datasets, using machine learning or statistics, are representative of a limited set of countries, commonly north America and Europe. In both instances, their represented population, healthcare infrastructure, and environmental factors widely differ from Australia's. There is a need to bridge the gap and construct synthetic datasets unique to Australia. This paper presents a statistical data generation based on Australian aggregate statistics.

2. Methods

To generate the synthetic data we used Synthea [1]. Synthea is a patient population simulator that has been used to generate various synthetic datasets, including a synthetic representation of the population in Massachusetts (SyntheticMass). The core version of Synthea uses publicly available population aggregate statistics such as demographics, disease prevalence and incidence rates, and health reports all based in the USA. Synthea generates data based on manually curated models of clinical workflows and disease progression that cover a patient's entire life and does not use real patient data; guaranteeing a completely synthetic dataset.

Relying on curated population characteristics, the Synthea disease progression and treatment models generate the patient's data. These models are encapsulated in modules that compute individual disease groups [1-6]. Multiple records are generated by Synthea for individuals in a population (Table 1).

To generate synthetic healthcare records representative of the Australian population we curated Synthea modules using Australian data [7]. The modified inputs include demographic data of all Australian local government areas (LGAs). The updated demographics include the proportion of males and females, 5-year age groups, level of income, levels of education, names and populations of suburbs, time zones and postcodes. Additional information such as names and addresses of all Australian hospitals, primary health networks and urgent care facilities was also added.

Table 1. List of data files generated by Synthea.

Data set	Description
patients.csv	Demographics information of patients
encounters.csv	Encounter with a medical practitioner
conditions.csv	Conditions diagnosed at encounter date
allergies.csv	Presence of allergies and allergy category
imaging_studies.csv	Imaging studies such as x-ray, ultrasound…
observations.csv	Include vital signs, laboratory tests, survey…
medications.csv	Past and current medication with reason for prescription
providers.csv	Includes address and contact details of providers such as GP practices
careplans.csv	Reason for care plan, start and stop dates
procedures.csv	Assessment, screening, chemotherapy
devices.csv	Devices used for treatment for example (home nebuliser)
immunisation.csv	Immunisations history
organisations.csv	visited institutions such as hospitals, with address, phone…
payers.csv	Public and private health insurance organisations

3. Results

We developed an Australianised version of Synthea. Using this we generated 117,258 synthetic patients from Queensland. We modelled data for four common chronic conditions with their comorbidities, treatments, care plans and procedures such as MRI studies. Consistent with the expectation the prevalence of diabetes, hypertension, and chronic kidney disease increased with age, whereas the prevalence of Alzheimer's disease increases until the 80-90 age bin and then decreases (Figure 1).

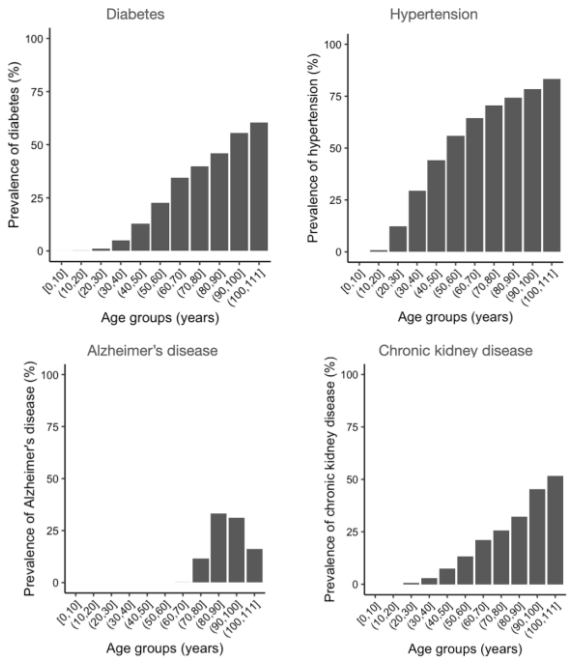

Figure 1. From top-left to right: diabetes, hypertension, Alzheimer's disease, and chronic kidney disease show the prevalence of each condition according to patients' age in 10-year bins of the Synthetic data.

Synthea also generates comorbidities associated with specific diseases. Among patients with diabetes 70% have been diagnosed with metabolic syndrome sometime during their lifetime, 63% had a diagnosis of a diabetic renal disorder, and 51% were classified as presenting microalbuminuria due to type 2 diabetes mellitus. Among patients with hypertension, the most common comorbidities were metabolic syndrome (44%), chronic kidney disease (25%) and microalbuminuria due to type 2 diabetes mellitus (21%). Among patients with Alzheimer's disease 47% have experienced metabolic syndrome, 36% have experienced chronic sinusitis and 34% have experienced chronic kidney disease. Among patients with chronic kidney disease 81% have experienced microalbuminuria due to type 2 diabetes mellitus; 55% experienced metabolic syndrome; while 51% experienced proteinuria due to type 2 diabetes mellitus.

Aside from comorbidities, Synthea also generates prescriptions. For instance, for the patients with chronic kidney disease, 67% have had a current or past prescription of Lisinopril 10mg oral tablet (used to treat hypertension); 65% had a prescription of insulin; 52% had a prescription of Hydrochlorothiazide 25mg oral tablet. Hydrochlorothiazide is used to treat hypertension and fluid retention (edema) that is caused by congestive heart failure, severe liver disease (cirrhosis), kidney disease, or treatment with a steroid or hormone medicine. Also, 48% of synthetic patients with chronic kidney disease had a 2.5mg oral prescription of amLODIPine to treat hypertension; and 5% were prescribed a 1ml Epoetin Alfa 4000unt/ml injection (Epogen) to treat anaemia caused by chronic kidney disease.

All patients with Alzheimer's disease have had a dementia management plan during the course of their disease. They also all had an assessment of health and social care needs. Synthea also generates procedures such as MRIs. Among patients with Alzheimer's, only 2.2% had a brain MRI.

4. Discussion

This manuscript showcases synthetic data generation for the Australian population. The adopted methodology follows expert-curated patterns making it realistic without requiring real data. This implies that it can be used without the standard restrictive and administrative requirements of accessing real healthcare data. This valuable resource is available and can be used by research groups to develop algorithms with limited time and cost.

Synthetic data generation requires an initial investment of time and resources after which researchers can access rapidly the desired amount of data without additional cost and/or additional regulatory requirements. Synthetic datasets have the additional advantage of enabling researchers to test their tools' scalability on datasets bigger than the real-world ones they would have access to. Furthermore, artificially generated data enables us to know -or even incorporate- patterns for the tools to detect; thus, serving as calibration method. However, synthetic data present several disadvantages: the accuracy and validity of synthetic data needs to be assessed against real-world data; It is unlikely that fully synthetic data will ever fully replicate real-world data; Synthetic data, especially those based on population summary statistics may not be able to generate outliers that are present in real-world data. Finally, real-world data may be necessary for the final validation stages of the new tools.

Synthea tries to replicate real-world patterns using disease modules. These modules are based on existing broad literature which may deviate from the Australian healthcare intricacies. For instance, Australia–via Medicare–offers universal healthcare cover available to most individuals whereas in the USA health cover is mostly privately funded. Also, the prevalence of certain conditions may diverge between the countries. Different environmental factors and lifestyle choices have large impacts on certain diseases e.g. obesity and diabetes. In future iterations of this work, we aim to customize and tailor disease modules to better reflect the patterns found in the Australian healthcare system.

The data generation method used by Synthea is a strong alternative to traditional synthetic data generation approaches as it removes the risk of individual re-identification and can be used to produce data without access to a primary source. Synthea has premade and previously tested open-source modules that anybody can use [8]. The modules are human designed following literature patterns for each disease which may miss rare or poorly understood patterns. Moreover, the architecture prevents capturing disease relationships outside each module. Future improvement of this work will include the integration of interactions of the different disease modules to capture clinical complexity.

5. Conclusions

We generated an Australian dataset and analysed on four diseases: diabetes, hypertension, Alzheimer's, and chronic kidney disease. During the process the patients were assigned realistic prescriptions, laboratory tests, comorbidities, etc. Because this approach does not require data to train the models, it can be extrapolated to any population. It is important for synthetic data to reflect the population it is being used to model. We showed that it is possible to adapt modules from a well-documented program like Synthea to generate synthetic for the Australian population.

References

[1] Walonoski J, Kramer M, Nichols J, Quina A, Moesel C, Hall D, Duffett C, Dube K, Gallagher T, McLachlan S. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. J Am Med Inform Assoc. 2018 Mar;25(3):230-8, doi: 10.1093/jamia/ocx079.

[2] Chen A. A novel graph methodology for analyzing disease risk factor distribution using synthetic patient data. Health Care Anal. 2022 Nov;2:100084, doi: 10.1016/j.health.2022.100084.

[3] Arora A, Arora A. Synthetic patient data in health care: a widening legal loophole. The Lancet. 2022 Apr;399(10335):1601-2, doi: 10.1016/S0140-6736(22)00232-X.

[4] Bhanot K, Qi M, Erickson JS, Guyon I, Bennett KP. The Problem of Fairness in Synthetic Healthcare Data. Entropy. 2021 Sep;23(9):1165, doi: 10.3390/e23091165.

[5] Ohm P. Broken promises of privacy: Responding to the surprising failure of anonymization. UCLA l Rev. 2009;57:1701.

[6] Syntheticheath [Internet]. Synthea Generic Module Framework [updated 2022 August 1; cited 2022 November 30]. Available from: https://github.com/synthetichealth/synthea/wiki/Generic-Module-Framework.

[7] Australian Bureau of Statistics [Internet]. Australian Bureau of Statistics [cited 2022 April 15, 2022]. Available from: https://www.abs.gov.au.

[8] Syntheticheath [Internet]. Synthea modules builder. [updated 2022 November 22; cited 2022 November 30]. Available from: https://syntheticheath.github.io/module-builder/.