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], 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

Ibrahima Diouf | 11/07/2023

Australia's National Science Agency

Australian e-Health Research Centre

Why generate synthetic data?

Access to data is essential for research but can be challenging

Real-world data can be costly and long to access

Synthetic data are generated to represent realistic data and/or have the same statistical properties as real data.

Synthetic data has the potential to ease data access

Synthetic data provide a greater protection of patients' sensitive information.

Different approaches for generating synthetic data

Synthetic data from real data

- Use real data to build a model that captures the distribution and structure of the real data.
- If the model is good, the synthetic data will have statistical properties similar to those of the real data.

- **approach used by Medisyn**

Synthetic data without real data

- Use existing models or background knowledge.
- Accuracy depends on the analyst’s background knowledge and the realism of the assumptions made.

- **approach used by many studies where analysts want to test scenarios for which real data are not available**

Synthetic data based on real summary statistics of a given population

- use of probability-based logic
- no risk of identification of personal information
- clinical validity may be a challenge

- **approach is used by Synthea**

Synthetic data at AEHRC

Hamed Hassanzadeh et al*:

- Historically informed data generation
  - Prospective analysis
  - Simulation of impact of variations in hospital demand and capacity on patient flow metrics

Publication: Kenny et al (2021)

Patient flow simulation using historically informed synthetic data

Digital Health Institute Summit; Brisbane.

Filip Rusak:

Synthetic MRI data

Publication: Rusak et al (2020)

3D Brain MRI GAN-Based Synthesis Conditioned on Partial Volume Maps.

Simulation and Synthesis in Medical Imaging - MICCAI, Lima, Peru (Online)

Publication: Rusak et al (2021)

Synthetic brain MRI dataset for testing of cortical thickness estimation methods v1.

CSIRO: Data Collection

John Grimes and students:

Started work on **Synthea**

Added Family history in asthma, Osteoarthritis and cystic fibrosis disease modules.

Patients' names to Australian names

Australian time zones

Australian e-Health Research Centre

The Synthea approach

Synthea mirrors the reference population:

- demographics
- disease burden
- vaccinations
- medical visits
- socio-economic factors….

Synthetic patients go through clinical journeys per disease module

A disease module simulates patients using:

- recommendations from clinical guidelines
- findings from literature
- experts’ opinions

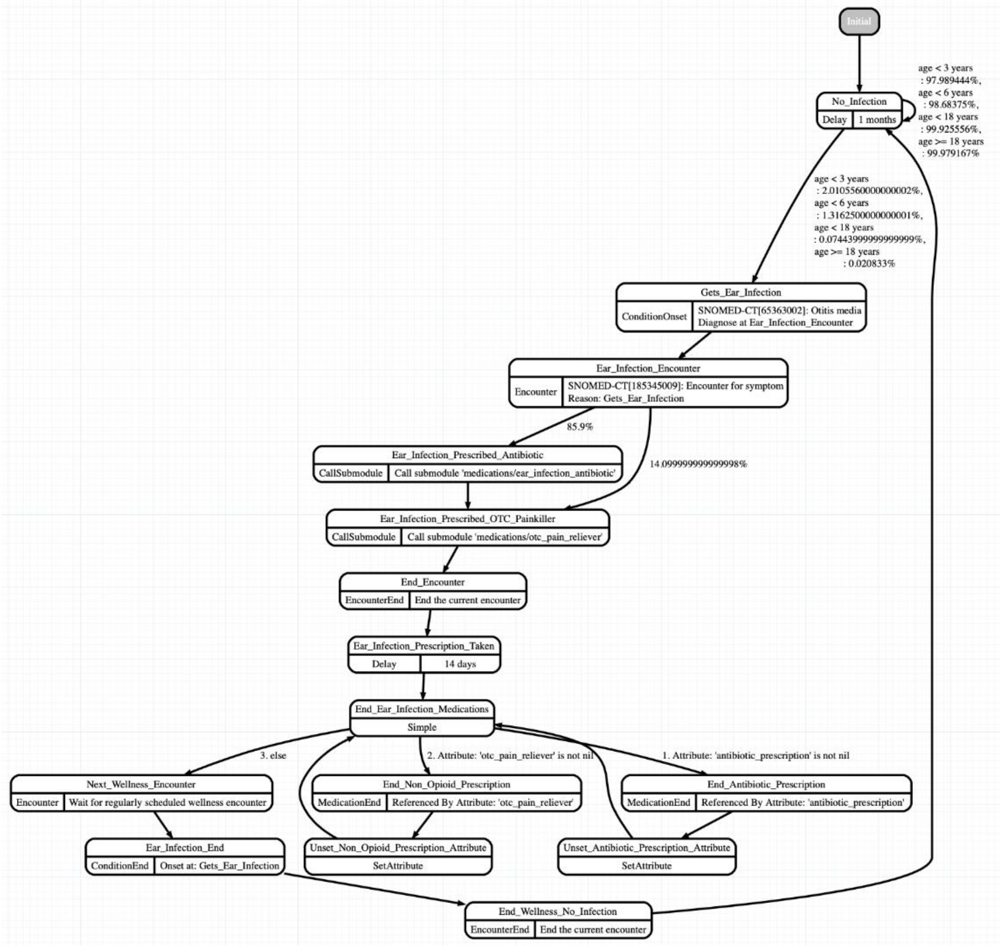
Over 60 modules

- Over >70 submodules

Synthetic data from Synthea can be exported

- standardised formats such as FHIR

Example of Synthea disease module




```json
{
  "name": "Ear Infections",
  "states": {
    "Initial": {
      "type": "Initial",
      "direct_transition": "No_Infection",
      "name": "Initial"
    },
    "No_Infection": {
      "type": "Delay",
      "exact": {
        "quantity": 1,
        "unit": "months"
      },
      "complex_transition": [
        {
          "condition": {
            "condition_type": "Age",
            "operator": "<",
            "quantity": 3,
            "unit": "years"
          },
          "distributions": [
            {
              "distribution": 0.02010556,
              "transition": "Gets_Ear_Infection",
              "remarks": [
                "72.38% of children < 3 get an ear
                 infection. This gives an incidence
                 of .7238 / (3 * 12) = 0.020105556
                 per month",
                "Source: https://www.nidcd.nih.gov
                 /health/statistics/ambulatory-care
                 -visits-diagnosis-otitis-media"
              ]
            },
            {
              "distribution": 0.97989444,
              "transition": "No_Infection"
            }
          ]
        },
```

Australian e-Health Research Centre

Data generated with Synthea

- allergies.csv
- careplans.csv
- claims_transactions.csv
- claims.csv
- conditions.csv
- devices.csv
- encounters.csv
- imaging_studies.csv
- immunizations.csv
- medications.csv
- observations.csv
- organizations.csv
- patients.csv
- payer_transitions.csv
- payers.csv
- procedures.csv
- providers.csv
- supplies.csv

Patients

Id	BIRTHDATE	DEATHDATE	DRIVERS	FIRST	LAST	MARITAL	GENDER	CITY	LAT	LON	HEALTHCARE EXPENSES	INCOME
0829aa3c	11/10/2021			Kristen940	Wolf938		F	Seisia	-10.86021218	142.3901182	7345.27	13222
b44e0d21	5/2/1977		S99978024	Valentin929	Cummings51	M	M	Morningside	-27.45419732	153.0762477	55191.9	4588
a4ffdc0	7/11/2006		S99925423	Venus149	Wuckert783		F	Kedron	-27.44963239	153.068354	68195.31	12952
aee03c13	7/9/2018			Alfred550	Graham902		M	Little Mountain	-26.81196347	153.1004452	12637.75	40971
c9068102	11/7/1962		S99946901	Shanita956	Rippin620	M	F	Elanora	-28.13850819	153.4756598	1018411.71	13670

Observations

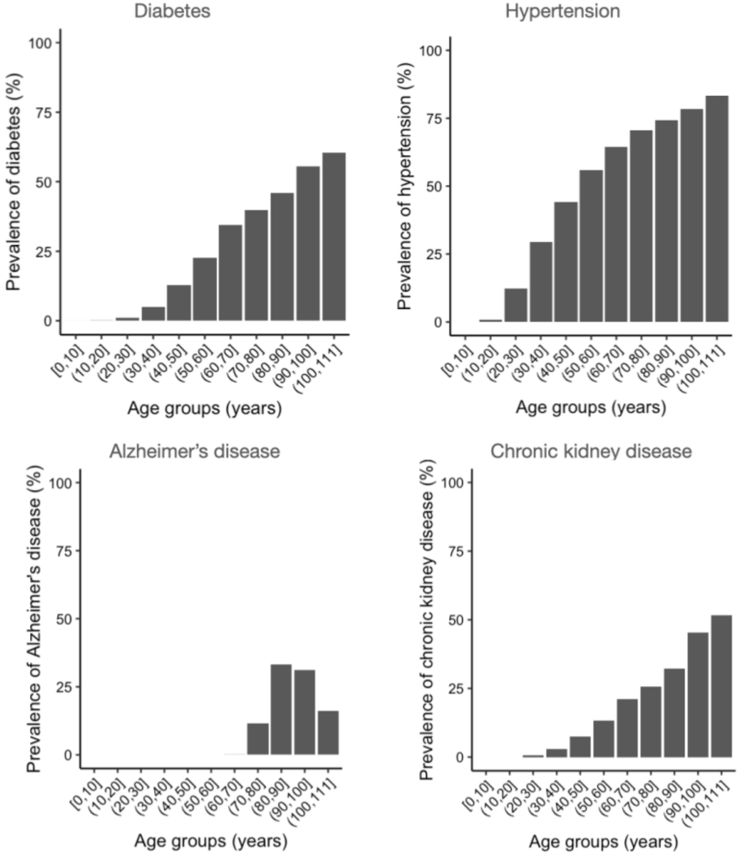
DATE	PATIENT	ENCOUNTER	CATEGORY	CODE	DESCRIPTION	VALUE	UNITS
2021-10-10T13:45:53Z	0829aa3c	a2ab0d59	vital-signs	8302-2	Body Height	53.1	cm
2021-10-10T13:45:53Z	0829aa3c	a2ab0d59	vital-signs	72514-3	Pain severity - 0-10 verbal numeric rating	3.0	{score}
2021-10-10T13:45:53Z	0829aa3c	a2ab0d59	vital-signs	29463-7	Body Weight	5.3	kg
2021-10-10T13:45:53Z	0829aa3c	a2ab0d59	vital-signs	77606-2	Weight-for-length Per age and sex	99.0	%

Medications

START	STOP	PATIENT	PAYER	ENCOUNTER	CODE	DESCRIPTION	BASE_COST	REASONDESCRIPTION
2022-09-05T13:45:53Z	2022-09-19T13:45:53Z	0829aa3f	b1c428d6	eb1bea60	198405	Ibuprofen 100 MG Oral Tablet	332.58	
2018-02-28T09:20:05Z	2018-03-16T09:20:05Z	b44e0d21	b1c428d6	6e0a0b1e	313782	Acetaminophen 325 MG Oral Tablet	212.34	Acute bronchitis (disorder)
2012-12-06T04:54:31Z	2012-12-21T04:54:31Z	a4ffdc0b	b1c428d6	e9b9e36b	198405	Ibuprofen 100 MG Oral Tablet	136.66	
2013-02-28T13:34:24Z	2013-03-10T09:34:24Z	a4ffdc0b	b1c428d6	9c905a18	834061	Penicillin V Potassium 250 MG Oral Tablet	244.33	Streptococcal sore throat
2019-11-25T06:12:31Z		a4ffdc0b	b1c428d6	f269def7	861467	Meperidine Hydrochloride 50 MG Oral Tablet	4772.04	

Generation of 117,258 synthetic patients (QLD)

Example showing four common chronic conditions



Patient characteristics within disease cohorts

Diabetes

- 70% metabolic syndrome
- 63% diabetic renal disorder
- 51% diabetes-related microalbuminuria

Chronic kidney disease

- 67% prescribed oral hypertension treatment (Lisinopril 10mg tablet)

Alzheimer's disease

- 100% dementia management plan

Australian e-Health Research Centre

Summary

Synthea's methodology follows expert-curated patterns

- realistic without requiring real data
- reduced burden of administrative requirements for access

Limitations of disease modules based on literature

- model may deviate from Australian healthcare intricacies

Limitations of current Synthea architecture

- does not capture disease relationships outside each module

Current work: machine learning-based data generation

- to capture clinical complexities present in real-world data

Australian e-Health Research Centre

Thank you

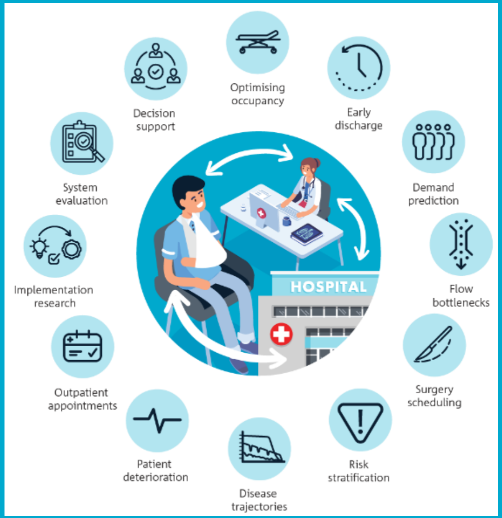
Dr. Ibrahima Diouf

Research Scientist

Ibrahima.Diouf@csiro.au

aehrc.csiro.au

Health Intelligence @ CSIRO AEHRC



Health System Productivity & Efficiency
Operational & Clinical Decision Support
Evidence-driven Policy & Healthcare Delivery