Synthetic Data: A literature review of synthetic data generation in epidemiological studies

Date: 27th June 2024
Version: v1.1

Protocol Authors and Affiliations:

Dr Caroline Morton, MRCGP, Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London UK
Dr John Tazare, PhD, Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London UK
Professor Krishnan Bhaskaran, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London UK
Professor Ruth Dobson, Wolfson Institute of Population Health, Queen Mary University London, London, UK

Protocol History

This section will detail any protocol updates. Please note that minor updates - for example to correct errors will result in an increment update only (from v1.0 to v1.1). Major protocol changes, classified as those which will impact the study conduct or analysis will result in a new protocol version (from v1.0 to v2.0).

Protocol Location/Section	Change	Rationale
v1.0	N/A	Original draft shared
v1.1	Addition of “primary care” and “secondary care” into search string; data extraction fields tightened up post-testing (see section)	Data extraction fields improved post testing; search string improved after discussion with team

Introduction

Synthetic data can be broadly defined as “microdata records created to improve accessibility whilst preventing disclosure of confidential information”[1]. When applied to electronic health record research, the types of data being generated are patient level event data, including primary and secondary care, billing or claims data, prescription data and demographic data, amongst others. These represent a synthetic patient population that can be used for a variety of purposes, including, but not limited to: training analysts, developing models, generation of without disclosure.

Synthetic data usually requires raw data or detailed information about the raw to generate a model, and then uses those models to generate new data values that have the same statistical values, and relationships as the original data[2]. The generation of synthetic data in electronic health record research has been broadly split into categories - data-driven and process-driven [3,4]. Data driven methods require the original raw data to generate the synthetic data and include methods such as imputation and taking probabilities from the summary data. Process driven methods generate synthetic data through computational or mathematical models.

The aim is to conduct a literature review to quantify the extent of and methods for generation of synthetic electronic health record data. This review will form part of a larger body of work looking at synthetic data generation in electronic health record research. For the purposes of this review, there will be a wide definition of electronic health record synthetic data. This will include data generation associated with non-research uses such as software testing, and the generation of administrative or claims records as well as strictly research datasets.

Objectives

Our primary objective is to describe the methods of generation of synthetic data in the epidemiology literature in the years 2005 to 2024. This includes whether real data is required to generate the synthetic data directly.

The secondary objectives are to describe:

- The types of synthetic data generated
- The self-described use cases of synthetic cases
- The prevalence of code sharing in synthetic data generation
- The type of license the generation of synthetic data is being used with
- Portability of synthetic data generated

Methods

Search strategy and study selection

Publication and Timeframe

We will include articles published in the electronic health record literature after 1st January 2005 and before the 30th April 2024.

Types of Studies

We will include all peer-reviewed articles of the following types:

- Original research articles
- Review articles

We will not be including narrative articles which talk about the potential of synthetic data as this systematic review is aimed at collecting current practice in generation of synthetic data. Likewise commentaries or letters to editors will be excluded. We will also exclude all papers where there is no analysis of data. Duplicate publications will be excluded if they occur.

Study identification

We will use the following search string to search PubMed:

("health data" OR "health records" OR "medical records" OR "healthcare data" OR "claims data" OR "administrative data" OR "patient records" OR "healthcare records" OR "patient data" OR "ehealth" OR "e-health" OR “ehr” OR “primary care” OR “secondary care”) AND ("dummy data" OR "mock data" OR "fake data" OR "synthetic data" OR "simulated data" OR "data generation" OR "simulate data") AND (("2005/01/01"[Date - Publication] : "2024/04/30"[Date - Publication]))

In addition to this, we will look through any reviews that describe the techniques used and manually add any cited papers.

Data Collection

Collection and Processing of Potentially Eligible Publications

The list of potentially eligible publications will be exported as a CSV and stored for record keeping. We will then screen each abstract of eligibility with a record kept for each excluded article on the reasons it was excluded. This data will be used later to give a granular breakdown of the study process.

Publication eligibility will be determined by:

1. **Abstract Screen**: Exclude non-eligible types of publications and articles where there is clearly no relevant synthetic data in the abstract.
   The screened articles will be placed into 1 of 3 categories:

   - YES include in study
   - NO exclude in study - with brief reason
   - MAYBE progress to full-text screen

2. **Full Text Screen:** Article will be read and there will be consideration whether the article involves any synthetic data generation.

   The screen articles will be placed into 1 of 2 categories:

   - YES include in study
   - NO exclude in study - with brief reason

The resulting list of included publications will be saved and the selection process displayed as a flowchart.

The eligibility screen will be done by a single reviewer as there is expected to be limited uncertainty about whether a paper should be included or not. If there is any uncertainty, papers will be discussed (CM and JT) and a joint decision taken on whether to include the article or not. All joint decisions will be documented for transparency.

As a quality control exercise, 1% of all papers at this stage will be screened by a second reviewer independently to ensure agreement. These papers will be randomly selected using a random number generator.

Data Extraction from Eligible Studies

Basic information including publication year, author list, DOI and journal will be extracted on each paper. We will use a google form to extract additional information that cannot be automatically extracted. This will include information on code sharing, linked github repo. See extraction fields section and the linked google form. The primary aim is to collect information on if synthetic data has been generated, techniques used and the use case.

Data Analysis

We will describe the prevalence of synthetic data generation overall. This will be further broken down by publication year, type of study (for example, cohort study) and use case (for example, for development of code outside a secure data environment). Where methods on how synthetic data was generated is available, we will also include if the programming code was shared, and if so how (for example, via Github). We will also include information if a readme or instructions on how to run the data generation was included.

Extraction Fields

- Journal Name
- Year of Publication
- Publication Type
    - Original Research
    - Review
    - Other (free text)
- DOI
- Title
- Authors
- Institutions
- Funders
- Type of study
- Type of data attempting to generate (admin/claims data, secondary care/primary care, other with free text).
- Does the study publish a pre-registered protocol on generating synthetic data? Yes/No (* this must go beyond a description or diagram showing implementation of an algorithm).
- Does the generation require real data as an input Yes/No?
- What is the method of synthetic data generation? (free-text)
- If yes, does the paper describe methods for dealing with privacy concerns? (free-text)
- Does the study publish the programming code for generating synthetic? Yes/No

  The following questions are only answerable if previous answer is **yes**:
    - Is the published code linked or referenced in the paper Yes/No? (*this means can click through to the code from the paper either from a reference or a link in the paper)
    - Is the published code directly accessible without needing to contact authors Yes/No?
    - Is the code based on open source and free software Yes/No? (*this means open source language such as Python, R - and not closed source such as SAS, Stata).
    - Does the code include instructions on how to run Yes/No?
    - Is the code open source Yes/No?
        - If yes, what is the license (free-text)?
    - Can the published code be directly used to generate synthetic data for another context or setting (Yes/No)

- Does the study publish synthetic data? (Yes/No)

  The following questions are only answerable if previous answer is **yes**:
    - Is the published synthetic dataset linked or referenced in the paper Yes/No?

    - Is the published synthetic dataset directly accessible without needing to contact authors Yes/No?
    - Is the synthetic dataset open source Yes/No?
        - If yes, what is the license (free-text)?

- Is the use case for the synthetic data given Yes/No?
    - If yes, what are the described use-cases (choose multiple from training models, teaching analysts, data exploration, generation of statistical code, software testing, debug statistical code or other with free text).

The Google form for collection is here. This set of extraction fields was tested on the 26th June 2024 with JT and CM with 5 papers pulled from the lists generated from the search string. JT and CM applied the abstract screening and data field extraction criteria to these papers and then met up to discuss the results.

Three papers met the criteria to be included in the study. There was agreement between JT and CM with the extraction fields but the language and meaning of certain fields was more specifically defined. These were:

1. Does the study publish a protocol on generating synthetic data? This was too broad and “pre-registered” was added to the extraction fields. This is to reduce confusion between a pre-registered protocol and a simple description or diagram of the algorithm applied.
2. The meaning of “Is the published code linked or referenced in the paper” was narrowed down to mean code can be found via a named reference in a paper or a link from the paper out to a code repository (such as Github, Gitlab) or an archive of some code.
3. “If yes, what are the described use-cases” already demonstrated a number of common use cases so a “multiple checkbox” option was added with these in rather than free-text only. Free-text is still allowed via “other”.

Descriptive figures and tables will be generated.

Data Sharing

The raw data, cleaned data and programming code will be shared on Github under a MIT license.

Table Shells

Below is a series of tables that are currently empty but will be completed by the end of the study.

References

1. What are synthetic data? :: SLS - Scottish Longitudinal Study Development & Support Unit. https://sls.lscs.ac.uk/guides-resources/synthetic-data/what-is-synthetic-data/.
2. Gonzales, A., Guruswamy, G. & Smith, S. R. Synthetic data in health care: A narrative review. *PLOS Digit. Health* **2**, e0000082 (2023).
3. Goncalves, A. *et al.* Generation and evaluation of synthetic patient data. *BMC Med. Res. Methodol.* **20**, 108 (2020).
4. Ayilara, O. F. *et al.* Generating synthetic data from administrative health records for drug safety and effectiveness studies. *Int. J. Popul. Data Sci.* **8**, (2023).