PLOS DIGITAL HEALTH

RESEARCH ARTICLE

Synthetic data in health care: A narrative review

Aldren Gonzales[1]*, **Guruprabha Guruswamy**[2], **Scott R. Smith**[1]

1 Office of the Assistant Secretary Planning and Evaluation, US Department of Health and Human Services, Washington, District of Columbia, United States of America, **2** Department of Health Administration and Policy, George Mason University, Virginia, United States of America

* aldren.gonzales@hhs.gov

OPEN ACCESS

Citation: Gonzales A, Guruswamy G, Smith SR (2023) Synthetic data in health care: A narrative review. PLOS Digit Health 2(1): e0000082. https://doi.org/10.1371/journal.pdig.0000082

Editor: Alistair Johnson, SickKids: The Hospital for Sick Children, CANADA

Received: June 29, 2022

Accepted: December 6, 2022

Published: January 6, 2023

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pdig.0000082

Data Availability Statement: All data are in the manuscript.

Funding: The authors received no specific funding for this work.

Abstract

Data are central to research, public health, and in developing health information technology (IT) systems. Nevertheless, access to most data in health care is tightly controlled, which may limit innovation, development, and efficient implementation of new research, products, services, or systems. Using synthetic data is one of the many innovative ways that can allow organizations to share datasets with broader users. However, only a limited set of literature is available that explores its potentials and applications in health care. In this review paper, we examined existing literature to bridge the gap and highlight the utility of synthetic data in health care. We searched PubMed, Scopus, and Google Scholar to identify peer-reviewed articles, conference papers, reports, and thesis/dissertations articles related to the generation and use of synthetic datasets in health care. The review identified seven use cases of synthetic data in health care: a) simulation and prediction research, b) hypothesis, methods, and algorithm testing, c) epidemiology/public health research, d) health IT development, e) education and training, f) public release of datasets, and g) linking data. The review also identified readily and publicly accessible health care datasets, databases, and sandboxes containing synthetic data with varying degrees of utility for research, education, and software development. The review provided evidence that synthetic data are helpful in different aspects of health care and research. While the original real data remains the preferred choice, synthetic data hold possibilities in bridging data access gaps in research and evidence-based policymaking.

Author summary

Synthetic data or data that are artificially generated is gaining more attention in the recent years because of its potential in making timely health care data more accessible for analysis and technology development. In this paper, we explored how synthetic data are being used by reviewing published literature and by looking at known synthetic datasets that are available to the public. Based on the available literature, it was identified that synthetic data address three challenges in making health care data accessible: it protects the privacy of individuals in datasets, it allows increased and faster access of researchers to health care

Competing interests: The authors have declared that no competing interests exist.

research data, and it addresses the lack of realistic data for software development and testing. Users should also be aware of its limitations that may include recognized risk for data leakage, dependency on imputation model, and not all synthetic data replicate precisely the content and properties of the original dataset. By explaining the utility and value of synthetic data, we hope that this review helps to improve understanding of synthetic data for different applications in research and software development.

Introduction

Data play a significant role in advancing health care delivery, public health, research, and innovations to address barriers and improve the quality of care. When researchers and innovators have timely access to real-world data, it can inform the development of new treatment, promote evidence-based policymaking, advance program evaluation, and transform outbreak responses [1–3]. However, users continue to face different challenges to accessing original data.

Most datasets containing health information are not readily available for use because they contain confidential information about individuals. Identifiable records can't also be easily shared as organization need to comply with certain regulations, such as the Health Insurance Portability and Accountability Act of 1996 (HIPAA) in the US [4]. Researchers and analysts continue to face many barriers when accessing essential datasets. Data access requirements such as the need for data use agreements, submission and approval of full protocol, completion of data request form, ethics review approval [5,6] and cost for accessing datasets that are not in the public domain remain to be a challenge.

With more people needing access to research-identifiable records, organizations are innovating ways to make data more accessible. The generation and use of synthetic datasets can potentially address many access, privacy, and confidentiality barriers [7]. In simple terms, synthetic datasets consist entirely of, or contain a subset of, not real microdata that are artificially manufactured with or without the original data. In health care, synthetic data could be an electronic health record (EHR) dataset with patient identifiable information and other sensitive information replaced with fake data to avoid reidentification. Synthetic dataset could also contain EHR records where all the original data are synthesized to produce a completely unreal record. The formal definition and types are further discussed in the succeeding section.

While synthetic data hold great potentials to advance evidence-based policymaking, research, and innovation, challenges are still present related to its development capability and confidence in using synthetic data [8]. In addition, relatively few authors have explored the topic of synthetic data, specifically its application in the health care industry and research. While some studies have explored and used synthetic data in different ways, the discussions were very focused on the project, or the specific method used.

The goal of this review article is to serve as a guide for researchers, data entrepreneurs, and innovators to improve understanding of the utility, value, and appropriateness of synthetic data for their respective applications. The paper starts by presenting the definition and types of synthetic data. Next, synthetic data generation using various software and tools are briefly discussed. The following sections summarize use cases and description of publicly available and ready-to-download synthetic datasets. Lastly, other opportunities in using synthetic data and its limitations are highlighted in the discussion section.

Methods

We conducted a narrative review of existing literature using PubMed and Scopus. The narrative review method was used to enable a thematic analysis of the different use cases [9]. The review was initially limited to peer-reviewed articles. These articles were identified by conducting an abstract/title search with the following terms: synthetic AND data OR dataset AND healthcare OR health care.

The articles were screened independently by two researchers. Articles presenting synthetic data development, use, and validation specific to health care delivery, public health, education, and research were included. After screening the 4,226 unique articles and reviewing 293 abstracts, 72 articles were included to identify the use cases. An additional targeted search was conducted using Google Scholar and Google Search engine to gather additional information from grey literature, including those related to the examples that were highlighted in this paper. Snowballing of references was also used to identify relevant articles from the articles that came out of the search.

Definition and types of synthetic data

Most literature refers to the definition of synthetic data used by the US Census Bureau. It is defined as "microdata records created by statistically modeling original data and then using those models to generate new data values that reproduce the original data's statistical properties." This definition highlights the strategic use of synthetic data because it improves data utility while preserving the privacy and confidentiality of information [10]. Depending on how it is generated, synthetic datasets can come with a reverse disclosure protection mechanism for inferences about parameters in statistical models but still with adequate variables to permit appropriate multivariate analyses [11].

The term synthetic data has been widely used to characterize datasets in various synthesized forms and levels. Some argued that the term synthetic data should only be used to refer to datasets containing purely fabricated data and without any original record [12,13]. These datasets may be developed using an original dataset as a reference or modeled using statistics. However, other literature, mainly those in the census and statistics discipline, acknowledge a more diverse sub-classification of synthetic data.

In general, synthetic data can be classified into three broad categories: fully synthetic, partially synthetic, and hybrid (published originally by Aggarwal and Chu and cited by Mohan [7]). First, Rubin proposed fully synthetic data in 1993 and further developed by Raghunathan et al. [14] in 2003. It is described as a dataset that is completely synthetic in nature and doesn't contain any real data. Because there is no reality with the dataset generated, this type is considered to have a strong privacy control but low analytic value because of data loss. Second, in partially synthetic data, select variables with sensitive values and considered to be high risk for disclosure are replaced with a synthetic version of the data. Since it contains original values, the risk for reidentification is present. The idea of 'partially synthetic' data first introduced by Little in 1993 and formally named by Reiter [15] in 2003. Lastly, hybrid synthetic data is generated using both original and synthetic data. With hybrid synthetic data, "each random record of real data, a close record in the synthetic data is chosen and then both are combined to form hybrid data." It holds privacy control characteristics with high utility in comparison to the first two categories but requires more processing time and memory [7].

A more detailed spectrum of synthetic data types is described in a working paper series by the United Kingdom's Office for National Statistics (UK's ONS). The spectrum features six levels under the synthetic and synthetically-augmented dataset types [12]. According to the UK's ONS, the synthetic structural dataset (the lowest form of synthetic data and developed using

metadata only) has no analytical value and has no disclosure risk. It can only be used for very basic code testing. On the other end of the spectrum, a replica level synthetically-augmented dataset can be used in place of the real data. This dataset has high analytic value because it preserves format, structure, joint distribution, patterns, and low-level geographies. However, since it is close to the original data, it introduces more disclosure risks.

Examples of software and tools in generating synthetic data

While published literature often refers to statistical approaches (e.g., multiple data imputation, bayesian bootstrap), the continuous development in technology has produced several tools and services to generate synthetic data programmatically. Researchers and innovators can maximize the use of software packages/libraries for R (e.g., Synthpop and Wakefield) and Python (e.g., PySynth, Scikit-learn, and Trumania) in synthesizing different types of data. Models are also available in generating synthetic images (e.g., ultrasound and computerized tomography) [16,17].

Specific to health records, there are also applications and services that users could leverage to generate synthetic data. Synthea, for example, is an open-source software package that generates high-quality, clinically realistic, synthetic patient longitudinal health records using publicly available health and census statistics, heath reports, clinical guidelines for statistical modeling [18]. MDClone's Synthetic Data Engine, on the other hand, is a commercial service that converts real EHR records into a synthetic version that is statistically comparable and maintains correlations among its variables. Health systems and universities use this synthetic data engine to accelerate data-driven medical research [19–23]. Other studies proposed the use of SynSys and Intelligent Patient Data Generator (iPDG); both are machine learning-based synthetic data generation methods for health care applications. This is not an exhaustive list of tools in generating synthetic data. As organizations explore this topic more, additional applications and services will be available.

Uses of synthetic health data

Although the use of synthetic data can be considered as a relatively new area, few peer-reviewed and grey literature have documented its value in different areas of health care research, education, data management, and health IT development. Table 1 summarizes the different use case example where synthetic data was found to be beneficial, along with an example.

Simulation studies and predictive analytics

Simulation and prediction research requires a large number of datasets to precisely predict behaviors and outcomes [31]. Real-world sources (e.g., from statistical agencies) have a significant advantage but are also most likely to be inaccessible to most researchers [32].

Synthetic data based on the real dataset can be used as a substitute or complement real data by allowing researchers to expand sample size or add variables that are not present in the original set. Synthetic data has been used in disease-specific hybrid simulation [33] and microsimulation for testing policy options [24,34] and health care financing strategies evaluation [35]. Studies also used synthetic data to validate simulation and prediction models [36] and to improve prediction accuracy [32].

Synthetic health records are also used in "in silico clinical trials" which refers to the development of "patient-specific models to form virtual cohorts for testing the safety and/or efficacy of new drugs and of new medical devices" [37]. The use of these datasets in silico clinical trials

Table 1. Summary of identified synthetic data use cases in health care and examples.

Use Case	Example
Simulation and Prediction Research	A research project used synthetic data that is close to an external benchmark to assess the impact of policy options to visit rates, prescription, and referral rates of the 65-and-over population [24].
Hypothesis, Methods, and Algorithm Testing	Experiments were conducted using publicly available and synthetic datasets to test the accuracy and robustness of the mixed-effect machine learning framework to predict longitudinal change in hemoglobin A1c among adults with type-2 diabetes [25].
Epidemiological Study/Public Health Research	A simulation study used the State of California synthetic population to test the impact of isolation, home quarantine, and other interventions in reducing the number of secondary measles cases infected by the index case and the probability of uncontrolled outbreak [26].
Health IT Development and Testing	The SMART Health IT sandbox contains synthetic clinical records that mimic a live EHR system environment that developers could use to test and demonstrate software applications in accessing clinical data using the SMART on FHIR platform [27].
Education and Training	Oregon Health and Science University used realistic synthetic clinical cardiovascular data to teach students robust risk prediction using machine learning techniques [28].
Public Release of Datasets	The Centers for Disease Control and Prevention–National Center for Health Statistics substituted select variables to prevent reidentification in the linked mortality public use files [29].
Linking data	Synthetic data were used to evaluate and test the linkage quality of an algorithm to link mothers and baby records before applying it to real-world administrative data [30].

https://doi.org/10.1371/journal.pdig.0000082.t001

can also inform the clinical trial design as well as make prediction for both the population and individual level to increase the chances of success [38].

Algorithm, hypothesis, and methods testing

Using synthetic data that reflects the content format and structure of the real data can help researchers explore variables, assess the feasibility of the dataset, and test hypotheses prior to accessing the actual dataset. Synthetic datasets can also provide another level of validation and comparison for testing methods and algorithms that would be beneficial for machine learning techniques development.

Research projects conducted experiments and have used synthetic data, public use files, and real data to verify algorithmic robustness, efficiency, and accuracy [25,39]. Another study used a publicly available synthetic claims dataset to evaluate and compare an algorithm with other models for phenotype discovery, noise analysis, scalability, and constraints analysis [40].

Epidemiological study/public health research

Datasets for epidemiology and public health studies may be limited in size, with quality concerns, challenging to obtain because of reporting procedures and privacy concerns, and expensive due to their proprietary nature [41]. More recently, the COVID-19 pandemic underscored the importance of making data accessible for public health surveillance, clinical studies (e.g., disease prognosis, drug repurposing, and new drug and vaccine development), and policy research during a health emergency. Publication of synthesized datasets can improve the timeliness of data release, support researchers in doing real-time computational epidemiology, provide a more convenient sample for sensitivity analyses, and build a more extensive test set for improving disease detection algorithms [42].

To demonstrate its utility, a study used a synthetic model with approximately eight million virtual New York City subway riders to simulate interactions and analyze the role of subway travel in spreading an influenza epidemic [43]. During the COVID-19 pandemic, several papers have documented the potentials and actual use of synthetic data for forecasting [44], improving diagnostics using synthetic images [45], and understand risk factors [46]. Other uses include epidemiological modeling [47,48], evaluation of outbreak detection algorithms [42,49], and simulation of public health events and interventions [26].

Health IT development and testing

Software testing is expensive, labor-extensive, and consumes between thirty to forty percent of the development lifecycle [50]. Because of the critical shortage of good test data, developers often create their own data or test with live data [51]. Using synthetic data can not only provide developers with a realistic dataset without privacy concerns, but it can also speed up the development lifecycle–saving cost, time, and labor.

Several tools are currently available such as the Michigan PatientGen tool that generates synthetic test records that are Fast Healthcare Interoperability Resources FHIR compatible. Another ready-to-download fully synthetic dataset under the SMART Sandbox mimics a live EHR production environment that developers could use for app testing and development [27].

Education and training

Synthetic data is useful when training students in subject areas (e.g., data science, health economics) that would require students to access a large number of realistic datasets [52]. While public use and limited use files are available, important fields for analysis (e.g., county and state, birth date) are often excluded for privacy reasons.

Oregon Health and Science University documented their use of clinical cardiovascular synthetic data to teach data science students the difficulties of working with clinical and genetic covariates for prediction analytics. Aside from citing data availability issues, they used realistic synthetic data because they need a suitable dataset for novice students to use and learn on, but realistic enough to encounter difficulties in using clinical data [28].

Public release of datasets

Releasing health datasets for public use comes with a unique challenge: preserving analytic value while ensuring the confidentiality of the records. While de-identifying microdata and data alteration can help, the probability of reidentification remains, and the alteration processes can distort the relationship of the variables [53]. Releasing partially synthesized data can mitigate disclosure risks while still allowing data users to obtain valid interferences that they could get in real data [15].

Due to high disclosure risks and to protect the confidentiality of records, the National Center for Health Statistics (NCHS) of CDC subjected linked mortality files (population survey and death certificates) to data perturbation techniques before releasing the public-use version of the dataset. Select variables that may lead to identification were replaced with synthetic values [29].

Linking data

Linking patient records with other datasets can help answer more research questions than a single source. When combining records, data processors develop algorithms and methods to automate the process and ensure accurate linkage.

Synthetic data is widely used in testing, validating, and evaluating different data linkage methods, frameworks, and algorithms either as the primary dataset or comparison dataset [30,54–57]. A research project compared the performance of different algorithms in terms of linkage accuracy and speed using nine million synthetic records [58]. While the project also used a real dataset of more than one million records, the synthetic data provided the investigators with a larger dataset to thoroughly test the capacity and efficiency of their algorithm. The use of synthetic data is also considered one of the ways to develop 'gold standard' datasets to evaluate linkage accuracy [59].

Examples of synthetic health datasets

Aside from project-specific datasets, there are publicly available and ready-to-download synthetic datasets that researchers, innovators, and data entrepreneurs can use for their purpose. The increasing number of these synthetic datasets is driven by the need for privacy-preserving datasets and the policies to make data available for public use. Below are examples of available datasets, databases, and sandboxes that contain synthetic data. This is not a comprehensive list but includes recently published datasets mentioned frequently in the literature review. These examples are also focused on US-specific datasets. Table 2 summarizes the data resources and their characteristics.

CMS 2008–2010 data entrepreneurs' synthetic public use file (DE-SynPUF)

The Centers for Medicare and Medicaid (CMS) published DE-SynPUF files to make a realistic version of Medicare claims data available to the public. The 2008–2010 synthetic files contain

Table 2. Examples of synthetic health datasets and their characteristics.

Synthetic Dataset	Data Owner/ Distributor	Type of Synthetic Dataset	Data Characteristics (type and quantity)	Use	Use Case Example
CMS 2008–2010 Data Entrepreneurs' Synthetic Public Use File (DE-SynPUF)	Centers for Medicare and Medicaid Services (Public domain)	Fully synthetic	6.8 Million beneficiary records; 112 million claims records; and 111 million prescription drug events records	Data entrepreneur analysis, software and application development, research training	Used a sub-set of the DE-SynPUF dataset to test different classification algorithms to accurately predict inpatient health care expenditure [60].
Synthea-Generated Datasets	MITRE Corporation	Fully Synthetic	One million longitudinal clinical synthetic patient records (SyntheticMass)	Innovation, development, education, and other nonclinical secondary uses	A pilot project used SyntheticMass data to assess whether data could be extracted from EHR through FHIR resources to support clinical trials [61].
US Synthetic Household Population	RTI International	Fully synthetic	Location and descriptive sociodemographic attributes of households (116 million records) and person living in those households (300 million records)	Agent-based modeling, disease outbreak simulation, distribution of resources analysis, sociodemographic pattern recognition, and disaster planning and response.	Used the dataset to simulate the impact of different influenza epidemics and the impact of utilizing pharmacies in addition to traditional (hospitals, clinic/physician offices, and urgent care centers) locations for vaccination [62].
CMS Synthetic data in Blue Button Sandbox	Centers for Medicare and Medicaid Services (inside a sandbox with access requirement)	No information	30,000 synthetic beneficiaries with claims data (Blue Button 2.0 Sandbox)	Development and testing of applications and information systems that will need to interact with CMS data systems	Blue Button 2.0 sandbox has more than 2,000 developers using the sandbox to test data exchange [63].

https://doi.org/10.1371/journal.pdig.0000082.t002

beneficiary summary, claims, and prescription data. However, DE-SynPUF contains a smaller subset of variables of the limited use files and has undergone privacy-preserving alterations. As a result, its utility to produce reliable interference about the population has weakened and made it unsuitable for analyzing the Medicare population [64,65].

The dataset maintained the data structure, format, and metadata of the CMS limited datasets. This makes it useful for training students, for researchers at the early stage of their study in designing program codes, and for health IT innovators in testing the accuracy and safety of their systems and applications [66].

One example of how DE-SynPUF was used is a project that utilized a portion of the dataset to investigate and test various classification algorithms to help address challenges in accurately predicting which beneficiaries would increase inpatient claims [60]. Another research project used DE-SynPUF to develop a framework to detect anomalous activities in specific patient groups [67]. On the technology development and data management side, the dataset has been used to test data models [68] and methodology to query data in multiple data models [69].

Synthea-generated datasets

One unique dataset that was generated using Synthea is the SyntheticMass. The dataset contains one million fictional but realistic residents of Massachusetts and mimics the geographic, disease rates, doctor's visits, vaccination, and social determinants of the real population [18,70].

Since the Synthea-generated datasets can be produced in FHIR formats, it is compatible with different programs and technologies for analysis and software development. Datasets from Synthea have been used in developing and testing health IT applications in FHIR environment [27,61], in teaching data science [28], and in modeling study [71].

More recently, the "Coherent Data Set" was produced using Synthea. This dataset combines multiple synthetic data forms together in a single package—familial genomes, magnetic resonance imaging (MRI) DICOM files, clinical notes, and physiological data [72].

US synthetic household population

The US Synthetic Household Population database contains location and sociodemographic attributes representing the entire population of the US at the household and person level. The database statistically matches the real household population and accurate spatial representation, making it a viable resource for microsimulation, planning for emergency response, simulation of disease outbreaks, and assessment of public health interventions [73,74].

Although it does not contain health information, the dataset was originally developed to support a modeling study of infectious disease agents by the National Institutes of Health [75]. The dataset was used to study different models on how infectious disease spread through social contact and applied the models to analyze how seasonal illnesses spread [76,77]. Because the dataset contains geospatial variables, a study was able to simulate the impact of influenza epidemics and how administering vaccines through pharmacies in addition to the usual locations can help increase vaccination uptake [62].

CMS synthetic data in Blue Button sandbox

CMS has sponsored different initiatives that aim to improve data access, make patient data more valuable and interoperable while minimizing the burden to health care providers [78]. Through the MyHealthEData initiative, CMS is rolling out initiatives (e.g., Blue Button 2.0) to establish standards (mostly FHIR-based) in data sharing from CMS to providers, patients, and payers [79].

CMS has published implementation guides and developed a sandbox containing 30,000 synthetic beneficiaries (in the Blue Button 2.0 sandbox) for testing purposes with over 13,000 fields from the CMS claims data warehouse mapped to FHIR [80].

Discussion

The use cases presented in this paper highlight the utility and the value of synthetic data in health care. Considering what was covered in the review, synthetic data can address three challenges in health care data. The first is protecting the privacy of individuals and ensuring the confidentiality of records. Because synthetic data can be composed purely or mixed with "fake" data, it is harder to re-identify the records [81]. Second, it improves the access of researchers and other potential users to health data. When synthesized, datasets could be made available to a wide number of users and at a faster rate because of the minimal disclosure risk [82]. Third, synthetic data address the lack of realistic data for software development and testing. Synthetic data could be cheaper for innovators for software application testing and provide them with more realistic test data for their intended test cases [83].

Given all its advantages, leveraging synthetic data can provide great opportunities in improving data infrastructure to help address some of the emerging health challenges. The data-sharing restrictions on mental health conditions such as opioid use disorder (OUD) became barriers for researchers and public health departments [84]. Generating synthetic longitudinal records of those diagnosed with OUD and those who have died because of opioid overdose can provide researchers data that could be analyzed to study patterns, identify risks, simulate policy impacts, and evaluate the effectiveness of programs. Synthesizing datasets is also useful when studying communicable diseases and stigmatized populations where there are several barriers to data sharing, for example, people diagnosed with HIV [85].

More recently, synthetic generation gained more attention as the demand for timely and accessible data increased because of the COVID-19 pandemic. One important initiative that leveraged synthetic data is the National Institutes of Health-steered National COVID Cohort Collaborative (N3C). Aside from restricted research identifiable files, N3C also generated a synthetic version of collected EHR records to make data more available to the broader research community and citizen scientists [86]. Because of the recent development, more studies are also being conducted to validate the use of synthetic data research. Recent papers on the use of synthetic data for COVID19-related clinical research have concluded that synthetic data could be used as proxy for the real dataset and that analysis of both synthetic and real datasets would yield statistically significant results–increasing the value and utility of synthetic data.

After understanding the different uses and potential applications of synthetic data, it is also important to recognize their limitations. The promise of synthetic data is to give users a dataset with artificial variables to preserve the confidentiality of the records. However, there are still risks that some quantity of the original data could be leaked. Data leakage can happen in different ways. If data contains outliers that the model captures, characteristics are reproduced in the synthetic version of the data. For example, only three individuals are diagnosed with a rare condition in a pool of synthetic records with the usual diagnoses. Because of the unique data point, that record can be easily linked to the original data. Data leakage can also happen through adversarial machine learning techniques. When an attacker has access to the synthetic data and the generative model used to create that data, records can be identified through running inversion attacks. This could be addressed through differential privacy techniques and disclosure control evaluation [12].

Not all synthetic data replicate precisely the content and properties of the original dataset–making them less useful for generating conclusions about a population. The authors agree that

the use of synthetic data in the area of clinical research is limited at the moment [18]. However, users need to understand the source and the way the synthetic dataset was generated to evaluate its appropriateness for specific types of studies or specific stages of research. The UK's ONS synthetic data high-level spectrum provides a good illustration of the quality and usability of synthetic data based on how similar it is to the original data. ONS argues that the more the synthetic data is realistic, its analytic value increases. However, the more it is close to the original dataset, the risk for disclosure also increases [12]. Taking the CMS DE-SynPUF as an example of this limitation. The synthesizing process resulted in a significant reduction in the amount of interdependence and co-variation among the variables–making it less useful for analytics [66].

Synthetic data are also dependent on the imputation model. The quality of the resulting data is highly dependent on the model used. Models can be good with identifying statistical abnormalities in the datasets but is also vulnerable to statistical noise, such as adversarial perturbation. This may cause the model to misclassify data and produce highly inaccurate outputs. Using real-world annotated data, input into the model, and testing the output for accuracy can address this issue. Models can sometimes also focus on trends and miss on distinctive cases that the real data has. This can be an issue when using the dataset for studying and generating conclusions about a certain population contained in the synthetic dataset. An example is the limited use of the synthesized version of the CanCORS data, a large-scale health survey. Upon evaluating the synthetic data which was developed using the project's model, the researchers concluded that the dataset is only useful for preliminary data analysis purposes because of the identified issues with data relationships [87].

The limitations mentioned highlight the need for validating synthetic data and the tools/models/algorithm used for production. Validation is needed to ensure that the synthetic dataset is comparable to the real-world data or useful for its intended purpose. The validation process can also be used to confirm the approach and evaluate if the model used worked as expected. Currently, there is no benchmark or standard for validating synthetic data [88]. Few studies have been conducted, and frameworks and approaches have been proposed to help in validating the realism of synthetic data. For example, the ATEN Framework for synthetic data generation also offers an approach to defining and describing the elements of realism and for validating synthetic data [88]. In another study, the authors compared the results derived from synthetic data generated by MDClone with those based on the real data of five studies on various topics. While the study showed that analyses conducted using synthetic data provide a close estimate of real data results in general, there are nuances observed in terms of accuracy (e.g., when a large number of patient records relative to the number of variables are used, it could yield to higher accuracy between the synthetic and real data) [23]. Another approach used to validate the realism of synthetic data is by looking at clinical quality measures. Using this approach, a study documented how synthetic data generated through Synthea could have some limitations in modeling heterogeneous outcomes and recognized the need to model additional factors that could influence deviation from guidelines and introduce variations in outcomes and quality [89]. The examples provided recognize the need for further exploration in the area of synthetic data validation, potentially towards a shared framework. Researchers and other data users should take into consideration the approach and results of validation studies in assessing if a specific synthetic dataset will be appropriate for their use.

Conclusion

The review showed that synthetic data has the potential to bridge data access gaps. The examples cited in this review highlighted the utility of synthesized health data in different areas of health research, software development, and training. The availability of publicly available

synthetic health datasets and off-the-shelf synthetic data generators reflects the growing interest and demand for accessible data. These tools and datasets hold great potential in increasing the access of researchers, data entrepreneurs, and health IT innovators to realistic datasets while preserving the statistical relationships and protecting the confidentiality of the records. In the future, we expect more datasets with synthetic data to be released, more tools to generate synthesized data will be developed, and more users will appreciate the utility of synthetic data. Users should evaluate synthetic datasets for quality and appropriateness for their intended use. Users should be aware and take into account the limitations when using synthetic data to maximize their potentials. Future discussions and studies should examine synthetic data's validity in different use cases, establish its utility in research, validate synthetic generation tools and techniques, and promote awareness among the research and health IT community.

Author Contributions

Conceptualization: Aldren Gonzales, Guruprabha Guruswamy, Scott R. Smith.

Data curation: Aldren Gonzales.

Formal analysis: Aldren Gonzales, Guruprabha Guruswamy.

Investigation: Aldren Gonzales, Guruprabha Guruswamy.

Methodology: Aldren Gonzales.

Project administration: Aldren Gonzales, Guruprabha Guruswamy, Scott R. Smith.

Supervision: Scott R. Smith.

Writing – original draft: Aldren Gonzales, Guruprabha Guruswamy, Scott R. Smith.

Writing – review & editing: Aldren Gonzales, Scott R. Smith.

References

1. Doshi JA, Hendrick FB, Graff JS, Stuart BC. Data, Data Everywhere, but Access Remains a Big Issue for Researchers: A Review of Access Policies for Publicly-Funded Patient-Level Health Care Data in the United States. EGEMS. 2016; 4(2):1204–. https://doi.org/10.13063/2327-9214.1204 PMID: 27141517
2. Yozwiak NL, Schaffner SF, Sabeti PC. Data sharing: Make outbreak research open access. Nature. 2015; 518(7540):477–9. https://doi.org/10.1038/518477a PMID: 25719649
3. Ho HKK, Görges M, Portales-Casamar E. Data Access and Usage Practices Across a Cohort of Researchers at a Large Tertiary Pediatric Hospital: Qualitative Survey Study. JMIR Med Inform. 2018; 6 (2):e32. https://doi.org/10.2196/medinform.8724 PMID: 29759958
4. Summary of the HIPAA privacy rule 2003 [cited 22 September 2019]. In: HHS.gov [Internet]. Available from: https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html.
5. Levenstein M, Tyler A, Bleckman J. The Researcher Passport: Improving Data Access and Confidentiality Protection. 2018 May 1 [cited 22 September 2019]. Available from: https://www.icpsr.umich.edu/files/about/researcher/ICPSR_ResearcherCredentialingWhitePaper_May2018.pdf.
6. Obtaining CMS Data for Your Research. [cited 22 September 2019]. In: National Institute on Aging [Internet] Available from: https://www.nia.nih.gov/research/dbsr/obtaining-cms-data-your-research.
7. Surendra H, Mohan H. A Review Of Synthetic Data Generation Methods For Privacy Preserving Data Publishing. J Sci Technol Res. 2017; 6(03):95–101.
8. Jarmin R. Synthetic Data: Public-Use Micro Data for a Big Data World. 2014 Oct 14 [cited 28 September 2019]. In: Census Blogs [Internet] Available from: https://www.census.gov/newsroom/blogs/research-matters/2014/10/synthetic-data-public-use-micro-data-for-a-big-data-world.html.
9. Green BN, Johnson CD, Adams A. Writing narrative literature reviews for peer-reviewed journals: secrets of the trade. J Chiropr Med. 2006; 5(3):101–17. https://doi.org/10.1016/S0899-3467(07)60142-6 PMID: 19674681
10. Philpott D. A Guide to Federal Terms and Acronyms: Bernan Press; 2017.

11. Abowd JM, Lane J. New Approaches to Confidentiality Protection: Synthetic Data, Remote Access and Research Data Centers. In: Domingo-Ferrer J, Torra V, editors. Privacy in Statistical Databases. Berlin, Heidelberg: Springer Berlin Heidelberg; 2004. pp. 282–289.

12. ONS methodology working paper series number 16—Synthetic data pilot. [cited 30 September 2019]. In: Office for National Statistics [Internet]. Available from: https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/onsmethodologyworkingpaperseriesnumber16syntheticdatapilot.

13. Siwicki B. Is synthetic data the key to healthcare clinical and business intelligence? 2020 Feb 21 [cited 11 May 2020]. In: Healthcare IT News [Internet]. Available from: https://www.healthcareitnews.com/news/synthetic-data-key-healthcare-clinical-and-business-intelligence.

14. Raghunathan T, Reiter J, Rubin D. Multiple imputation for statistical disclosure limitation. J Off Stat. 2003; 19:1–16.

15. Reiter J. Inference for partially synthetic, public use microdata sets. Surv Methodol. 2003; 29(2):181–8.

16. Cusumano D, Lenkowicz J, Votta C, Boldrini L, Placidi L, Catucci F, et al. A deep learning approach to generate synthetic CT in low field MR-guided adaptive radiotherapy for abdominal and pelvic cases. Radiother Oncol. 2020; 153:205–12. https://doi.org/10.1016/j.radonc.2020.10.018 PMID: 33075394

17. Cronin NJ, Finni T, Seynnes O. Using deep learning to generate synthetic B-mode musculoskeletal ultrasound images. Comput Methods Programs Biomed. 2020; 196:105583. https://doi.org/10.1016/j.cmpb.2020.105583 PMID: 32544777

18. Walonoski J, Kramer M, Nichols J, Quina A, Moesel C, Hall D, et al. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. J Am Med Inform Assoc. 2017. Epub 2017/10/13. https://doi.org/10.1093/jamia/ocx079 PMID: 29025144

19. MDClone Launches New Phase of Collaboration with Washington University in St. Louis. 2019 Feb 5 [cited 31 October 2019]. In: MDClone News [Internet]. Available from: https://www.mdclone.com/news-press/articles/mdclone-washington-university-collaboration.

20. Regenstrief Institute-MDClone Partnership to Accelerate Data-Driven Medical Research. 2018 Dec 20 [cited 31 October 2019]. In: Markets Insider [Internet]. Available from: https://markets.businessinsider.com/news/stocks/regenstrief-institute-mdclone-partnership-to-accelerate-data-driven-medical-research-1027826793

21. Intermountain Healthcare Collaborates with MDClone to Transform Patient Data into Actionable Insights 2019 [October 31, 2019]. Available from: https://mdclone.com/wp-content/uploads/2019/10/5c594cd44abddd8e151dcffd_Intermountain-press-release-1.pdf.

22. Marcusohn E, Epstein D, Roguin A, Zukermann R. Normal high sensitive troponin I and suspected myocardial infarction, is the rapid rule out algorythm for all? Eur Heart J. 2019; 40(Supplement_1). https://doi.org/10.1093/eurheartj/ehz748.0997

23. Reiner Benaim A, Almog R, Gorelik Y, Hochberg I, Nassar L, Mashiach T, et al. Analyzing Medical Research Results Based on Synthetic Data and Their Relation to Real Data Results: Systematic Comparison From Five Observational Studies. JMIR Med Inform. 2020; 8(2):e16492. Epub 20.2.2020. https://doi.org/10.2196/16492 PMID: 32130148

24. Davis P, Lay-Yee R, Pearson J. Using micro-simulation to create a synthesised data set and test policy options: The case of health service effects under demographic ageing. Health Policy. 2010; 97(2–3):267–74. https://doi.org/10.1016/j.healthpol.2010.05.014 PMID: 20800762

25. Ngufor C, Van Houten H, Caffo BS, Shah ND, McCoy RG. Mixed effect machine learning: A framework for predicting longitudinal change in hemoglobin A1c. J Biomed Inform. 2019; 89:56–67. Epub 2018/09/07. https://doi.org/10.1016/j.jbi.2018.09.001 PMID: 30189255

26. Enanoria WT, Liu F, Zipprich J, Harriman K, Ackley S, Blumberg S, et al. The Effect of Contact Investigations and Public Health Interventions in the Control and Prevention of Measles Transmission: A Simulation Study. PLoS One. 2016; 11(12):e0167160. Epub 2016/12/13. https://doi.org/10.1371/journal.pone.0167160 PMID: 27941976

27. SMART Health IT Sandbox. 2017 [cited 18 October 2019]. In: Smart [Internet]. Available from: https://docs.smarthealthit.org/.

28. Laderas T, Vasilevsky N, Pederson B, Haendel M, McWeeney S, Dorr DA. Teaching data science fundamentals through realistic synthetic clinical cardiovascular data. bioRxiv. 2017:232611. https://doi.org/10.1101/232611

29. Public-use Linked Mortality File. 2020 Mar [cited 7 December 2022]. Available from: https://www.cdc.gov/nchs/data/datalinkage/public-use-2015-linked-mortality-file-description.pdf.

30. Harron K, Gilbert R, Cromwell D, Van Der Meulen J. Linking data for mothers and babies in de-identified electronic health data. PLoS One. 2016; 11(10). https://doi.org/10.1371/journal.pone.0164667 PMID: 27764135

31. Ringel JS, Eibner C, Girosi F, Cordova A, McGlynn EA. Modeling health care policy alternatives. Health Serv Res. 2010; 45(5 Pt 2):1541–58. Epub 2010/08/02. https://doi.org/10.1111/j.1475-6773.2010.01146.x PMID: 21054371

32. Aljaaf AJ, Al-Jumeily D, Hussain AJ, Fergus P, Al-Jumaily M, Hamdan H. Partially synthesised dataset to improve prediction accuracy. In: Huang DS, Bevilacqua V, Premanratne P, editors. Intelligent Computing Theories and Application. Switzerland: Springer Cham; 2016. p. 855–66.

33. Amoon AT, Arah OA, Kheifets L. The sensitivity of reported effects of EMF on childhood leukemia to uncontrolled confounding by residential mobility: a hybrid simulation study and an empirical analysis using CAPS data. Cancer Causes Control. 2019; 30(8):901–8. Epub 2019/05/31. https://doi.org/10.1007/s10552-019-01189-9 PMID: 31144088

34. Symonds P, Hutchinson E, Ibbetson A, Taylor J, Milner J, Chalabi Z, et al. MicroEnv: A microsimulation model for quantifying the impacts of environmental policies on population health and health inequalities. Sci Total Environ. 2019; 697:134105. https://doi.org/10.1016/j.scitotenv.2019.134105 PMID: 32380606

35. Hennessy D. Creating a synthetic database for use in microsimulation models to investigate alternative health care financing strategies in Canada. Int J Microsimul. 2015; 8:41–74.

36. Sun Z, Wang F, Hu J. LINKAGE: An approach for comprehensive risk prediction for care management. In: Cao L, Zhang C, editors. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015 Aug 10–13. Sydney, Australia. New York: Association for Computing Machinery; 2015. p. 1145–1154.

37. Pappalardo F, Russo G, Tshinanu FM, Viceconti M. In silico clinical trials: concepts and early adoptions. Brief Bioinform. 2019; 20(5):1699–708. https://doi.org/10.1093/bib/bby043 PMID: 29868882

38. Zand R, Abedi V, Hontecillas R, Lu P, Noorbakhsh-Sabet N, Verma M, et al. Development of synthetic patient populations and in silico clinical trials. In: Bassaganya-Riera, editor. Accelerated Path to Cures. Springer Cham; 2018. p. 57–77.

39. Jayalatchumy D, Thambidurai P. Prediction of diseases using Hadoop in big data—A modified approach. In: Silhavy R, Senkerik R, Oplatkova ZK, Prokopova Z, Silhavy R. Advances in Intelligent Systems and Computing. Springer Cham; 2017. p. 229–238.

40. Chen R. Tackling chronic diseases via computational phenotyping: Algorithms, tools and applications. PhD Dissertaion. Georgia Institute of Technology; 2018. Available from: https://smartech.gatech.edu/handle/1853/60206.

41. Levin D, Finley P. Synthetic data generators for the evaluation of biosurveillance outbreak detection algorithms. Sandia National Laboratories. 2018 Oct 1 [Cited 21 October 2019]. In: OSTI.GOV [Internet]. Available from: https://www.osti.gov/servlets/purl/1481598.

42. Texier G, Jackson ML, Siwe L, Meynard J-B, Deparis X, Chaudet H. Building test data from real outbreaks for evaluating detection algorithms. PloS one; 2017. p. e0183992. https://doi.org/10.1371/journal.pone.0183992 PMID: 28863159

43. Cooley P, Brown S, Cajka J, Chasteen B, Ganapathi L, Grefenstette J, et al. The role of subway travel in an influenza epidemic: a New York City simulation. Journal of urban health: bulletin of the New York Academy of Medicine: Springer US; 2011. p. 982–95. https://doi.org/10.1007/s11524-011-9603-4.

44. Bannur N, Shah V, Raval A, White J. Synthetic Data Generation for Improved covid-19 Epidemic Forecasting. medRxiv. 2020:2020.12.04.20243956. https://doi.org/10.1101/2020.12.04.20243956

45. Karbhari Y, Basu A, Geem ZW, Han G-T, Sarkar R. Generation of Synthetic Chest X-ray Images and Detection of COVID-19: A Deep Learning Based Approach. Diagnostics. 2021; 11(5):895. https://doi.org/10.3390/diagnostics11050895 PMID: 34069841

46. Synthetic data. 2022 Nov 15 [cited 7 December 2022]. In: Clinical Practice Research Datalink (CPRD) [Internet]. Available from: https://cprd.com/content/synthetic-data.

47. Xu Z, Glass K, Lau CL, Geard N, Graves P, Clements A. A Synthetic Population for Modelling the Dynamics of Infectious Disease Transmission in American Samoa. Sci Rep. 2017; 7(1):16725. Epub 2017/12/03. https://doi.org/10.1038/s41598-017-17093-8 PMID: 29196679

48. Hashemian M, Stanley K, Osgood N. Leveraging H1N1 infection transmission modeling with proximity sensor microdata. BMC Med Inform Decis Mak. 2012; 12:35. Epub 2012/05/04. https://doi.org/10.1186/1472-6947-12-35 PMID: 22551391

49. Garcia YE, Christen JA, Capistran MA. A Bayesian Outbreak Detection Method for Influenza-Like Illness. Biomed Res Int. 2015; 2015:751738. Epub 2015/10/02. https://doi.org/10.1155/2015/751738 PMID: 26425552

50. Ariola B. DevOps and Cloud Mean the End of QA as You Know It. 2019 Jun 16 [cited 21 October 2019]. In: CIO [Internet]. Available from: https://www.cio.com/article/3409149/devops-and-cloud-mean-the-end-of-qa-as-you-know-it.html.

51. PatientGen—synthetic, realistic patient data for use in interoperability testing. [cited 18 October 2019]. In: HealthIT.gov—Interoprability Proving Ground [Internet]. Available from: https://www.healthit.gov/techlab/ipg/node/4/submission/1466.

52. Droese L, Gerbel S, Teppner S, Fiebeck J, Frömke C. Generating synthetic data for use in research and teaching. 2019 Jul [cited 7 December 2022]. In: ResearchGate [Internet]. Available from: 10.13140/RG.2.2.27534.31047.

53. Na L, Yang C, Lo C-C, Zhao F, Fukuoka Y, Aswani A. Feasibility of Reidentifying Individuals in Large National Physical Activity Data Sets From Which Protected Health Information Has Been Removed With Use of Machine Learning. JAMA Netw Open. 2018; 1(8):e186040–e. https://doi.org/10.1001/jamanetworkopen.2018.6040 PMID: 30646312

54. Pow C, Iron K, Boyd J, Brown A, Thompson S, Chong N, et al. Privacy-Preserving Record Linkage: An international collaboration between Canada, Australia and Wales. Int J Popul Data Sci. 2017;1. https://doi.org/10.23889/ijpds.v1i1.101

55. Goldstein H, Harron K, Cortina-Borja M. A scaling approach to record linkage. Stat Med. 2017; 36 (16):2514–21. Epub 2017/03/18. https://doi.org/10.1002/sim.7287 PMID: 28303597

56. Li X, Xu H, Shen C, Grannis S. Automated linkage of patient records from disparate sources. Stat Methods Med Res. 2018; 27(1):172–184. https://doi.org/10.1177/0962280215626180 PMID: 28034172

57. Boyd JH, Guiver T, Randall SM, Ferrante AM, Semmens JB, Anderson P, et al. A Simple Sampling Method for Estimating the Accuracy of Large Scale Record Linkage Projects. Methods Inf Med. 2016; 55(3):276–83. Epub 2016/04/21. https://doi.org/10.3414/ME15-01-0152 PMID: 27096424

58. Mamun AA, Mi T, Aseltine R, Rajasekaran S. Efficient sequential and parallel algorithms for record linkage. Journal of the American Medical Informatics Association. 2014; 21(2):252–62. https://doi.org/10.1136/amiajnl-2013-002034 PMID: 24154837

59. Harron KL, Doidge JC, Knight HE, Gilbert RE, Goldstein H, Cromwell DA, et al. A guide to evaluating linkage quality for the analysis of linked data. Int J Epidemiol. 2017; 46(5):1699–710. https://doi.org/10.1093/ije/dyx177 PMID: 29025131

60. Lahiri B, Agarwal N. Predicting Healthcare Expenditure Increase for an Individual from Medicare Data. 2014 [cited 22 September, 2021]. Available from: http://cci.drexel.edu/hi/hi-kdd2014/morning_5.pdf.

61. Reddy B, Zopf R, Abolafia J. Use of Fast Healthcare Interoperability Resources (FHIR) in the Generation of Real World Evidence (RWE). 2017 [cited 22 September 2021]. Available from: https://www.cdisc.org/sites/default/files/resource/Use_of_Fast_Healthcare_Interoperability_Resources_in_the_Generation_of_Real_World_Evidence.pdf.

62. Bartsch SM, Taitel MS, DePasse JV, Cox SN, Smith-Ray RL, Wedlock P, et al. Epidemiologic and economic impact of pharmacies as vaccination locations during an influenza epidemic. Vaccine. 2018; 36 (46):7054–63. https://doi.org/10.1016/j.vaccine.2018.09.040 PMID: 30340884

63. Speech: Remarks by Administrator Seema Verma at the Blue Button Developer Conference. 2019 Jul 30 [cited 25 October 2019]. In: CMS.gov Newsroom [Internet]. Available from: https://www.cms.gov/newsroom/press-releases/speech-remarks-administrator-seema-verma-blue-button-developer-conference.

64. CMS 2008–2010 Data Entrepreneurs' Synthetic Public Use File (DE-SynPUF) 2014 [October 25, 2019]. Available from: https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs/DE_Syn_PUF.html.

65. CMS 2008–2010 Data Entrepreneurs' Synthetic Public Use File (DE-SynPUF). 2022 Jun 24 [cited 7 December 2022]. In: CMS.gov [Internet]. Available from: https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs/Downloads/SynPUF_FAQ.pdf.

66. User Manual—Centers for Medicare and Medicaid Services (CMS) Linkable 2008–2010 Medicare Data Entrepreneurs' Synthetic Public Use File (DE-SynPUF). 2013 Jan 15 [cited 25 October 2019]. Available from: https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs/Downloads/SynPUF_DUG.pdf.

67. Paudel R, Eberle W, Talbert D. Detection of Anomalous Activity in Diabetic Patients Using Graph-Based Approach. In: Rus VR, Markov Z, editors. Proceedings of the Thirtieth International Florida Artificial Intelligence Research Society Conference; 2017 May 22–24; Palo Alto, California. FLAIRS; 2017. p. 423–429.

68. Lambert C, Amritansh, Kumar P. Transforming the 2.33M-patient Medicare synthetic public use files to the OMOP CDMv5: ETL-CMS software and processed data available and feature-complete. 2016 [cited 22 September 2021]. Available from: https://www.ohdsi.org/web/wiki/lib/exe/fetch.php?media=

symposium_2016:transforming_the_2.33m-patient_medicare_synthetic_public_use_files_to_the_omop_cdmv5_-_etl-cms_software_and_processed_data_available_and_feature-complete.pdf.

69. Klann JG, Phillips LC, Herrick C, Joss MAH, Wagholikar KB, Murphy SN. Web services for data warehouses: OMOP and PCORnet on i2b2. J Am Med Inform Assoc. 2018/08/08 ed2018. p. 1331–1338. https://doi.org/10.1093/jamia/ocy093 PMID: 30085008

70. SyntheticMass. [cited 31 October 2019]. In: SyntheticMass [Internet]. Available from: https://synthea.mitre.org/about.

71. Gebert T, Jiang S, Sheng J. Characterizing Allegheny County Opioid Overdoses with an Interactive Data Explorer and Synthetic Prediction Tool. arXiv:1804.08830. 2018 [cited 31 October 2019]. Available from: https://arxiv.org/abs/1804.08830.

72. Walonoski J, Hall D, Bates KM, Farris MH, Dagher J, Downs ME, et al. The "Coherent Data Set": Combining Patient Data and Imaging in a Comprehensive, Synthetic Health Record. Electronics. 2022; 11 (8):1199. https://doi.org/10.3390/electronics11081199

73. RTI U.S. Synthetic Household Population Database. [cited 25 October 2019]. Available from: https://www.rti.org/sites/default/files/brochures/rti-brochure-file-8c629303-5027-429d-86ef-d26bae408309.pdf.

74. Wheaton WD, Cajka JC, Chasteen BM, Wagener DK, Cooley PC, Ganapathi L, et al. Synthesized Population Databases: A US Geospatial Database for Agent-Based Models. Methods report (RTI Press), 2009(10), 905. https://doi.org/10.3768/rtipress.2009.mr.0010.0905 PMID: 20505787

75. Dutchen S. A New Use for Census Data: Disease Simulations. 2011 May 18 [cited 25 October 2019]. In: LiveScience [Internet]. Available from: https://www.livescience.com/14220-census-data-disease-simulations-nigms.html.

76. RTI U.S. Synthetic Household Population. [cited 25 October 2019]. In: RTI International [Internet]. Available from: https://www.rti.org/impact/rti-us-synthetic-household-population%E2%84%A2.

77. ReCONNECT to Economic Opportunity: RTI U.S. Synthetic Household Population Data. [cited 25 October 2019]. In: NC State University Institute for Emerging Issue [Internet]. Available from: https://archive.iei.ncsu.edu/reconnectnc/rti-data/.

78. CMS Advances Interoperability & Patient Access to Health Data through New Proposals. 2019 Feb 8 [cited 25 October 2019]. In: CMS.gov Newsroom [Internet]. Available from: https://www.cms.gov/newsroom/fact-sheets/cms-advances-interoperability-patient-access-health-data-through-new-proposals.

79. A 360˚ view of your patients' history. [cited 25 October 2019]. In: CMS Dta at the Point of Care [Internet]. Available from: https://dpc.cms.gov/.

80. Blue Button API Docs. [25 October 2019]. In: CMS Blue Button 2.0 [Internet]. Available from: https://bluebutton.cms.gov/developers/#sample-beneficiaries.

81. Domingo-Ferrer J, Torra V, Mateo-Sanz J, Sebe F. Re-Identification and Synthetic Data Generators: A Case Study. 2005 Nov 9 [cited 25 October 2019]. Available from: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=5bf974ebdba5df9928729845068aa5c1f860ca10

82. Abay NC, Zhou Y, Kantarcioglu M, Thuraisingham B, Sweeney L. Privacy preserving synthetic data release using deep learning. In: Bertino E, Gao W, Steffen B, Yung M, editors. Lecture Notes in Computer Science. Springer Cham; 2019. p. 510–526.

83. Rose G. When to use production vs. synthetic data for software testing. [cited 30 October 2019]. In: Software Testing News [Internet]. Available from: https://www.softwaretestingnews.co.uk/when-to-use-production-vs-synthetic-data-for-software-testing/.

84. Manatt. Overcoming Data-Sharing Challenges in the Opioid Epidemic: Integrating Substance Use Disorder Treatment in Primary Care. 2018 Jul [cited 30 October 2019]. Available from: https://www.chcf.org/wp-content/uploads/2018/07/OvercomingDataSharingChallengesOpioid.pdf.

85. Ford MA, Spicer CM, editors. Monitoring HIV care in the United States: Indicators and data systems. Washington, District of Columbia: National Academies Press; 2012. p. 1–330.

86. N3C Synthetic Data Workstream: National COVID Cohort Collaborative (N3C). [cited 6 July 2021]. In: N3C [Internet]. Available from: https://covid.cd2h.org/N3C_synthetic_data.

87. Loong B, Zaslavsky AM, He Y, Harrington DP. Disclosure control using partially synthetic data for large-scale health surveys, with applications to CanCORS. Stat Med. 2013; 32(24):4139–61. https://doi.org/10.1002/sim.5841 PMID: 23670983

88. McLachlan S, Dube K, Gallagher T, Daley B, Walonoski J. The ATEN framework for creating the realistic synthetic electronic health record. In: Zwiggelaar R, Gamboa H, Fred A, Badia SB, editors. Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies —(Volume 5); 2018 Jan 19–21; Funchal, Madeira, Portugal: SciTePress; 2018. p. 182–91.

89. Chen J, Chun D, Patel M, Chiang E, James J. The validity of synthetic clinical data: a validation study of a leading synthetic data generator (Synthea) using clinical quality measures. BMC Med Inform Decis Mak. 2019; 19(1):44. Epub 2019/03/16. https://doi.org/10.1186/s12911-019-0793-0 PMID: 30871520