Communication

Synthetic Data & the Future of Women's Health: A Synergistic Relationship

Gayathri Delanerolle , Peter Phiri * , Heitor Cavalini , David Benfield , Ashish Shetty , Yassine Bouchareb , Jian Shi , Alain Zemkoho

Posted Date: 24 May 2023

doi: 10.20944/preprints202305.1694.v1

Keywords: Womens Health; Data Science; Data Methods; Artificial Intelligence

Communication

Synthetic Data & the Future of Women's Health: A Synergistic Relationship

Gayathri Delanerolle [1,2], Peter Phiri [1,3,*,†], Heitor Cavalini [1,†], David Benfield [1,4,†], Ashish Shetty [5,6], Yassine Bouchareb [7], Jian Qing Shi [1,8,9,‡] and Alain Zemkoho [1,4,10,‡]

1 Research & Innovation Department, Southern Health NHS Foundation Trust, SO40 2RZ, Southampton, UK
2 Nuffield Department of Primary Care Health Sciences, University of Oxford, OX3 7JX, Oxford, UK
3 School of Psychology, Faculty of Environmental and Life Sciences, University of Southampton, SO17 1BJ, Southampton, UK
4 Department of Mathematics, University of Southampton, SO17 1BJ, Southampton, UK
5 Female Pelvic Medicine and Reconstructive Surgery, University College London, WC1E 6BT, London, UK
6 University College London Hospitals NHS Foundation Trust, NW1 2PG, London, UK
7 Sultan Qaboos University, College of Medicine and Health Sciences, Muscat, Oman
8 Department of Statistics and Data Science, Southern University of Science and Technology, 518055, Shenzhen, China
9 National Centre for Applied Mathematics, Shenzhen, China
10 Alan Turing Institute, 96 Euston Road, NW1 2DB, London, UK
* Correspondence: peter.phiri@southernhealth.nhs.uk
† Shared second author.
‡ Shared last author.

Abstract: Objectives: The aim of this perspective is to report the use of synthetic data as a viable method in women's health given the current challenges linked to obtaining life-course data within a short period of time and accessing electronic healthcare data. **Methods:** We used a 3-point perspective method to report an overview of data science, common applications, and ethical implications. **Results:** There are several ethical challenges linked to using real-world data, consequently, generating synthetic data provides an alternative method to conduct comprehensive research when used effectively. The use of clinical characteristics to develop synthetic data is a useful method to consider. Aligning this data as closely as possible to the clinical phenotype would enable researchers to provide data that is very similar to that of the real-world. **Discussion:** Population diversity and disease characterisation is important to optimally use data science. There are several artificial intelligence techniques that can be used to develop synthetic data. **Conclusion:** Synthetic data demonstrates promise and versatility when used efficiently aligned to clinical problems. Therefore, exploring this option as a viable method in women's health, in particular for epidemiology may be useful.

Keywords: womens health; data science; data methods; artificial intelligence

BACKGROUND

The healthcare landscape is under immense pressure to cater to a growing global population with complex clinical needs especially for women. Women have higher risks for a variety of diseases in comparison to men such as multimorbidity, cardiometabolic and neuropsychiatric diseases in addition to obstetric and gynaecological conditions [1]. Patients are more vocal about having therapies that are less invasive and more tolerable with minimal side effects. Therefore, precision and personalised medicine has become more appealing to the masses [2]. To develop such befitting research requires high-quality evidence, ideally in shorter timeframes, improved analytical methods and access to existing data within electronic healthcare records (EHRs). EHRs are a rich source of data that could enable the identification of clinical gaps and development of analytic tools [3]. In the UK, EHR systems often record if patients are taking part in a research study within their healthcare record as well as basic information linked to any intervention they may be using as at the end of a study continued standard of care may be required which can be advantageous to researchers.

 doi:10.20944/preprints202305.1694.v1

However, a significant issue with EHRs are missing values, discrepancies, inconsistencies, and limitations to access full or partial patient records [4]. This introduces several challenges to developing research associated with current knowledge and clinical practice gaps as well as evidence-based healthcare policies that would improve care. An alternative method could be to better use data science methods to support policy makers, regulators, funders and industry [5].

Data science comprises of a vast area of scientific disciplines with an array of methods, processes and algorithms to help generate and synthesise datasets in a comprehensive manner thus, it disrupts and realigns conventional science to transform healthcare practices. This interdisciplinary field can use structured and unstructured datasets either qualitative or quantitative formats. Two common conceptual methodologies of data science would be the use of statistical and artificial intelligence (AI) methods. Statistical methods have advanced rapidly over the years and heavily influenced AI methodologies. These methods can address the most common problem in research, missing data issues by way of subsidiary synthetic and augmented data solutions. Missing data issues could also influence the development of statistical theory in the context of missing data and causal inference. There are multiple specific missing data mechanisms of missing-completely-at random, missing-at-random and missing-not-at random and their variants [6]. These options provide conditions to independent confounders and measurements or treatment and selection where only part of the data is missing. These instances could contrive variables of interest and that are auxiliary to a study. Rubin et al, Heckman et al and Robins et al provide an array of thoughts on the theory of causal inference that could evaluate the existence or non-existence of causation using graphical presentation and labelling of variables [7–9].

These principles have shown synthetic data and simulators have provided early promise and can be comparatively produced in pre-annotated, less expensive and in an exhaustible manner. Synthetic data in its' essence is artificially manufactured based on real-world data and could provide a better solution for practicalities in data security, governance and privacy issues. It can also provide a solution for healthcare systems to better prepare for future population needs and report universally applicably healthcare outcomes in particular for chronic conditions where patients often present a complex disease sequalae. The National Health Service (NHS) in the UK has a dedicated lab to explore methods of developing synthetic data using SynthVAE model [10]. In addition, the Office of National Statistics (ONS) have explored the potential for using synthetic data in a pilot project to develop user guidance [11]. Synthetic data can also be used as a research enrichment method if key characteristics and variables from real-world data can be used. Research studies with frequent visits and long-term follow up can often be cumbersome as participant dropout rates are higher in these instances which impact the quality of the data. Synthetic data can therefore provide a solution to improve the data quality. This could also be considered as a data optimisation method, particularly to improve scalability required for machine learning training and testing predictive models that could assist with conditions such as endometriosis, gestational diabetes, fibroids, polycystic ovary syndrome, depression and anxiety. The ease of using synthetic data is another important feature as it removes any errors and readily convert to different formats. This is important for non-communicable disease research as it could help make-up larger sample sizes and gather long-term follow up data required to report life-course clinical epidemiology outcomes. Women's health research is often hindered by small sample sizes due to a variety of reasons including recruitment challenges, lack of representation of all ethnicities and limitations with funding. However, evidence suggests women are under-represented in clinical trials globally regardless of the explored disease area. It is further challenged for women living in rural principalities in low-middle-income countries thus, limiting the generalisability and applicability of any findings, raising confounders, biases and inequalities.

Population diversity

Population diversity is vital to develop synthetic datasets relevant to the real-world. Variables of any disease population can be modelled within a framework through random generation of known variables which is vital for women's health research due to a high prevalence of disease sequalae.

 doi:10.20944/preprints202305.1694.v1

Diffusion is another aspect of population diversity that is useful to develop blueprints of a population. Together, they can be defined as characteristics of a population where a disease population is drawn.

Ethics and data science

Ethical implications surrounding the use and share of healthcare data across different regions of the world pose a variety of governance and legal implications. Global mass immigration patterns have significantly changed influencing societal values and psychosocial behaviours impacting healthcare systems. Hence, for statisticians, synthetic data can be an attractive way to conduct an analysis whilst simultaneously designing studies taking into account all ethical considerations where healthcare benefits for users and the public outweighs any risks. These risks could be considered at the very early stages of the project development process, especially with the design of the synthetic dataset as well as availability of patient consent within the original dataset. Another facet would be to have a clear understanding of the risks and limitations associated with any technologies used, developed and the level of human oversight of these to maintain data codes of practice, fidelity, integrity and quality. Over reliance of data science models in relation to ethnicity, race and gender would further raise biases leading to ethical implications, especially in instances where tools are used as clinician decision making aids for diagnosing or treating a condition. Whilst diversity is an important aspect in research, any models developed should clearly reflect population characteristics in a realistic manner. Most healthcare organisations sharing data would also need to ensure the data used is within the legal requirements such as the Data Protection Legislation, the Human Rights Act 1998 including the general protection regulations, the statistics and registration service act 2007 and the common law duty of confidence. Another aspect to consider would be to ensure the synthetic data development methods themselves to be transparent and reproducible.

Common methods of synthetic data generation

There are a number of issues associated with synthetic data environments including real-world applicability. Compositionality representations are a hierarchical grouping method that can be used. Synthetic data is particularly developed with labelled datasets to train AI models. The data itself can be primarily categorised as fully synthetic, partially synthetic and hybrid. Fully synthetic data lacks features from the original data as it generates data based on estimated feature density to produce estimate parameters that could be deemed realistic. The real-world data is not used in fully synthetic datasets. Partially synthetic datasets retain some of the features from the real-world data as it is a useful method to address missing data by way of imputation methods. This method could also replace high-risk features within a dataset or privacy protected data features. Hybrid data is a dataset that comprises of real-world and synthetic data. The advantage of the data is that it is able to pair random records based on real-world datasets.

MACHINE LEARNING MODELS

Supervised machine learning models have been commonly used for generating synthetic data as they are able to learn to predict in associated structure and patterns. The data used is usually sequentially synthesised and can report either linear or polytomous logistic or logistic regression. Mixed-data types are reported using a variety of other models. The principle of this method comprises of spatial proximity, symmetry, causality, and functionality linked to Gestalt psychology. In the modern-day machine learning era has led to the generation of annotated compositional data with a 3D model. A 3D data model usually comprises of an upper and lower vertex with a common 2D coordinator and different z-values. The object domain of the 3D model can show fine-grained part decompositions have emerged where the findings can be mapped to manual hierarchies. Hierarchical neural nets and hierarchical convolutional networks could generate deep learning models that are aligned to demonstrate clinical feature. Based on the desired objectives and endpoints of a research project, robust data models can be developed to generate synthetic (Figure 1).

 doi:10.20944/preprints202305.1694.v1

Deep learning models

Neural networks, the basic framework for deep learning models, are proficient at learning from datasets and generalising this creating neural network architectures to generate similar data points as the original distribution. Key neural techniques for developing deep learning models include autoencoders, generative adversarial networks (GANs), neural radiance field (NeRF) and Ensembles. Autoencoders are an unsupervised neural network that learns to reconstruct data based on real-world data where the *encoder* and *decoder* compresses and decompresses the data, respectively. This method can also transform biomedical or other types of research data that are predominantly categorical and binary.

GANs are a deep learning model comprising of two types of neural networks of *generator* and *discriminator*. This method can develop synthetic data by way of an adversarial training process where the original GAN can be further improved with the addition of a variety of features. Ensembles comprise of two different types of deep learning models to generate synthetic data. NeRF data is a method that could develop new images from a 3D scene. These algorithms use the static scene as a continuous 5D function to develop a neural network to predict a new volume for each voxel to fill-in an entire picture within a scene. NeRF is useful to develop either additional or realistic images although it is cumbersome to train.

In the case of imaging applications such as scanner software to detect endometriosis lesions or tumours, synthetic data techniques, such as modified loss functions or resampling using Synthetic Minority Oversampling Technique (SMOT) whilst down- and up-sampling offer the opportunity to generate extra data to balance different classes, allowing for adequate testing and validation of machine learning and deep learning models.

Data augmentation

Data augmentation is another method used to develop new medical applications. Augmentation methods are able to increase the size of the dataset in the absence of annotated real-world data. Data augmentation is a technique to increase the sample size by way of simulating the characteristics of an already annotated real-world dataset where the statistical analyses will include mixing of real and synthetic data. Similar approaches can be used to develop synthetic data using clinical trial datasets. To ensure statistical outcomes are cohesive, often an analysis with the real-world data or clinical trial data versus the augmented, synthetic dataset can be reported. Levin and colleagues trained a neural network to diagnose a variety of ovarian cancers where the training dataset was added to the real-world dataset [12]. The model diagnosis in this instance performed equally between the trained and real-world dataset. The dataset in this instance was imaging data which is arguably more complicated than clinical data.

Classical methods for data generation

Classical approaches include statistical, probabilistic and machine learning models that are considered as baseline methods. Baseline methods are associates with anonymisation techniques and methods that do not use data modelling to generate synthetic data. These techniques can replace values, deleting sensitive attributes and inclusion of statistical noise to data. Yale and colleagues suggested that these methods could summarise and memorise any correlations identified that can then be used to generate synthetic data [13]. The statistical models are another category that can synthesise data using probabilistic and statistical models [13]. These models are able to simulate real-world data attributes where samples and conclusions could be drawn from the model to generate synthetic data that is of a categorical and numerical nature that can be tested on time-series data.

Content models

Content models are a modelling method for synthetic data generated via electronic health records (EHR) and was originally proposed by McLachlan and colleagues to develop synthetic EHRs based on publicly available health information statistics coupled with the expertise of clinicians [14].

McLachlan and colleagues further developed the Aten framework to develop synthetic labour and birth EHRs that characterised and validated the real-world applicability [14]. Walonoski and colleagues (2018) developed an open source synthetic health simulator using EHRs called *Synthea* [15]. Synthea simulates synthetic patients from birth to death in JSON format demonstrating pre-defined modules for diseases such as cancer, diabetes and infections. In particular, Synthea was used to create synthetic data linked to the COVID-19 pandemic. Chen et al validated this of Synthea to report a cohort analysis reporting a variety of clinical measures [16]. Dahmen and Cook developed a synthetic smart home senser called Synsys by way of using Hidden Markov models that can generate temporary sequences of daily activities [17]. Prophet is another method developed by Hyun et al that was validated for predicting time-series data using an additive model with a non-linear trend fit [18]. This method demonstrated to robustness to manage missing data and shifting trends. The experiment demonstrated that Prophet can express medical changes with effective prediction values based on a specific parameter [18].

SUMMARY

In summary, synthetic data shows much promise and the generation of future synthetic data vaults could generate an ecosystem where inter-disciplinary professionals could access, assess and expand using open-source tools, thus allowing transparent exchange of methods and innovative solutions for healthcare problems that are common to the global population. Underrepresented diseases can particularly benefit from this endeavour alongside of industry deploying AI technologies more responsibly.

Author contributions: This perspective article is part of the FEINMAN project. GD and PP developed the FEINMAN project as part of the ELEMI program. The mathematical intellect for the FEINMAN project is shared between GD, JQS, PP, and AZ. The first draft was written by GD. All authors critically appraised and commented on all versions of the manuscript. All authors read and approved the final manuscript.

Funding: Not applicable.

Ethics approval: Not applicable.

Consent to participate: Not applicable.

Consent for publication: All authors consented to publish this manuscript.

Acknowledgements: The authors acknowledge administrative support from Sana Sajid and Southern Health NHS Foundation Trust, Southern University of Science and Technology and University of Southampton.

Conflicts of interest: PP has received research grant from Novo Nordisk and Janssen Cilag, and other, educational, other from John Wiley & Sons, outside the submitted work. DB's work is funded by an EPSRC PhD Studentship with reference number 2612869. AZ is supported by the EPSRC grant EP/V049038/1 and the Alan Turing Institute under the EPSRC grant EP/N510129/1. All other authors report no conflict of interest. The views expressed are those of the authors and not necessarily those of the NHS, the National Institute for Health Research, the Department of Health and Social Care or the Academic institutions.

References

1. Delanerolle G, Yang X, Shetty S, Raymont V, Shetty A, Phiri P, Hapangama DK, Tempest N, Majumder K, Shi JQ. Artificial intelligence: a rapid case for advancement in the personalization of gynaecology/obstetric and mental health care. Women's Health. 2021 May;17:17455065211018111.
2. Delanerolle GK, Shetty S, Raymont V. A perspective: use of machine learning models to predict the risk of multimorbidity. LOJ Medical Sciences. 2021 Sep 14;5(5).
3. Cowie MR, Blomster JI, Curtis LH, Duclaux S, Ford I, Fritz F, Goldman S, Janmohamed S, Kreuzer J, Leenay M, Michel A. Electronic health records to facilitate clinical research. Clinical Research in Cardiology. 2017 Jan;106:1-9.
4. Gov.uk Policy paper Data saves lives: reshaping health and social care with data.[Internet]. cited on March 1st 2022. Available on: https://www.gov.uk/government/publications/data-saves-lives-reshaping-health-and-social-care-with-data/data-saves-lives-reshaping-health-and-social-care-with-data.

5. Taichman DB, Backus J, Baethge C, Bauchner H, De Leeuw PW, Drazen JM, Fletcher J, Frizelle FA, Groves T, Haileamlak A, James A. Sharing clinical trial data: a proposal from the International Committee of Medical Journal Editors. Annals of internal medicine. 2016 Apr 5;164(7):505-6.
6. Little RJ, Rubin DB. Statistical analysis with missing data. John Wiley & Sons; 2019 Apr 23.
7. Rubin DB. Causal inference using potential outcomes: Design, modeling, decisions. Journal of the American Statistical Association. 2005 Mar 1;100(469):322-31.
8. Heckman JJ. Econometric causality. International statistical review. 2008 Apr;76(1):1-27.
9. Robins JM, Hernán MA. Estimation of the causal effects of time-varying exposures in Longitudinal Analysis, Handbook of Modern Statistical Methods, Eds Fitzmaurice, G, Davidian, M., Verbeke, G., Molenberghs, G., Chapman & Hall. 2009;553:599. CRC, Bacon Raton, USA
10. NHS England. Exploring how to create mock patient data (synthetic data) from real patient data.[internet] cited on 1st March 2022. Available on https://transform.england.nhs.uk/ai-lab/explore-all-resources/develop-ai/exploring-how-to-create-mock-patient-data-synthetic-data-from-real-patient-data/
11. Office for National Statistics. ONS methodology working paper series number 16 - Synthetic data pilot. [Internet]. Cited on 1st March 2022. Available on https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/onsmethodologyworkingpaperseriesnumber16syntheticdatapilot
12. Levine AB, Peng J, Farnell D, Nursey M, Wang Y, Naso JR, Ren H, Farahani H, Chen C, Chiu D, Talhouk A. Synthesis of diagnostic quality cancer pathology images by generative adversarial networks. The Journal of pathology. 2020 Oct;252(2):178-88.
13. Yale A,J., Privacy Preserving Synthetic Health Data Generation and Evaluation, Ph.D. thesis, Rensselaer Polytechnic Institute, ISBN: 9798662575981 Publication Title: ProQuest Dissertations and Theses 27833340, 2020.
14. S. McLachlan, K. Dube, T. Gallagher, J.A. Simmonds, N. Fenton, Realistic Synthetic Data Generation: The ATEN Framework, in: A. Cliquet Jr., S. Wiebe, P. Anderson, G. Saggio, R. Zwiggelaar, H. Gamboa, A. Fred, S. Bermúdez i Badia (Eds.), Biomedical Engineering Systems and Technologies, Communications in Computer and Information Science, Springer International Publishing, 497– 523, ISBN 978-3-030-29196-9, 2019. doi:10.1007/978-3-030-29196-9_25.
15. 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. Journal of the American Medical Informatics Association. 2018 Mar 1;25(3):230-8.
16. 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 medical informatics and decision making. 2019 Dec;19(1):1-9.
17. Dahmen J, Cook D. SynSys: A synthetic data generation system for healthcare applications. Sensors. 2019 Mar 8;19(5):1181.
18. Hyun J, Lee SH, Son HM, Park JU, Chung TM. A synthetic data generation model for diabetic foot treatment. InFuture Data and Security Engineering. Big Data, Security and Privacy, Smart City and Industry 4.0 Applications: 7th International Conference, FDSE 2020, Quy Nhon, Vietnam, November 25–27, 2020, Proceedings 7 2020 (pp. 249-264). Springer Singapore.