Newcastle University

# Using Synthetic Controls to Improve Randomised Control Trials for Rare Diseases

*Author:*
Nicole Cizauskas
Student ID: 230584521
Supervisors:
Prof James Wason
Dr Svetlana Cherlin

Ph.D. Project Proposal

December, 2023

# 1 Background

## 1.1 Rare Diseases

A rare disease is defined in the EU as a disease that affects fewer than 1 person per 2,000 people [1], and defined in the US as a disease that affects fewer than 200,000 people in the USA [2]. Despite this, 6-8% of the US and EU population are affected by rare diseases [3]. Rare diseases are difficult to research, making them difficult to diagnose and treat as a result.

Treatments for rare diseases often struggle to meet the requirements for the gold standard clinical trial methodology, randomized control trials (RCTs). This is typically due to a lack of participants available for the randomization process or ethical restrictions assigning patients to control arms [4]. As a result, very few rare diseases have specific pharmacological treatments [3].

## 1.2 Randomized Control Trials

RCTs are the gold standard of clinical research due to the randomization of participants [5–7]. By randomly sorting participants into two or more groups, and assigning one for each treatment and one placebo or standard of care, a study reduces the influence of biases and possible confounding factors that come with observational clinical studies. RCTs are often double-blinded, which means that neither the participant nor the researcher knows which treatment arm is placebo. The random allocation of groups accounts for numerous demographic characteristics, such as age, gender, sex, and ethnicity, that would otherwise cause confounding. It is difficult to prove with complete certainty that a causal relationship exists, however RCTs are the best current clinical trials method to providing evidence for cause and effects in treatments.

RCTs are typically required for treatment or drug approval [6]. However, well-powered RCTs require a larger number of participants than single-arm or observational studies due to the random allocation of participants into two independent groups. RCTs can be time-consuming and costly to run compared to other trial methodologies [5]. RCTs are also difficult to perform in some circumstances, such as with rare diseases when there may be less observational evidence supporting the novel treatments being tested [8]. Participants will often not want to be a part of a control or placebo arm for trials targeting severe diseases, and so participant recruitment in these circumstances is difficult.

## 1.3 Synthetic Controls

Synthetic control arms are control groups generated based on real-world patient data with similar attributes to the experimental group [9]. They are typically designed based on previous clinical trial data, observational study data, or external data (e.g. electronic health records) [4]. Several studies had success in using synthetic control arms. For example, one phase 1-2 trial looked at establishing novel indicators of survival in acute myeloid leukaemia used a data pool of historical controls from Medidata's archive of trial data [10]; this study was able to analyse its primary outcomes of complete remission, complete remission without hematologic recovery, and overall survival using their single-arm trial combined with a synthetic control arm.

Another previous study used observational data to generate a synthetic control arm matched to a clinical trial population [11]. Propensity scores were used to match participants in the experimental arm to controls in PharMetrics, a US claims database. A sufficient number of matched participants were selected in this way.

A recent study focused on using Electronic Health Records (EHRs) to develop a synthetic control arm [12]; this study looked at the effect of a lifestyle intervention on cardiovascular health metrics, and use propensity scores to match controls from EHRs to participants in a single-arm trial. The use of EHRs allowed for a follow-up on the same control arm participants 5 years later. Overall, external data from EHRs had mixed reception, with some measurements being more useful than others in analysis.

Synthetic control arms are particularly useful in RCTs that would otherwise be restricted by participant recruitment, cost, or ethics [13, 14]. These synthetic control arms can cut the number of participants required to run an RCT by providing data to represent the control or placebo group. This also solves the problem of ethics in cases where giving participants the control or placebo could be detrimental to their health. By reducing the amount of real participants needed, the overall cost to run the RCT is reduced. The use of synthetic control arms in RCTs is gaining more legitimacy, with recent approvals from the FDA featuring synthetic controls in their analyses [13, 15, 16].

## 1.4 Generative Adversarial Networks

Generative adversarial networks (GANs) are a type of unsupervised deep learning model [17]. Originally developed as an alternative to computationally intensive Markov chains, they feature two different modules inside a single model that communicate back and forth. The first module is a generator, which creates data from an initial random

set of values. The second module is a discriminator, which compares and measures the difference between the generated data from the generator and a set of real data. The information about the difference between the generated data and the real data is backpropagated to the generator, allowing it to improve the algorithm it uses to create data. This information is also sent back into the discriminator to inform it of which data is classified as real and which is classified as generated in order to sharpen the classification skills of the discriminator. This structure can be seen in Figure 1.

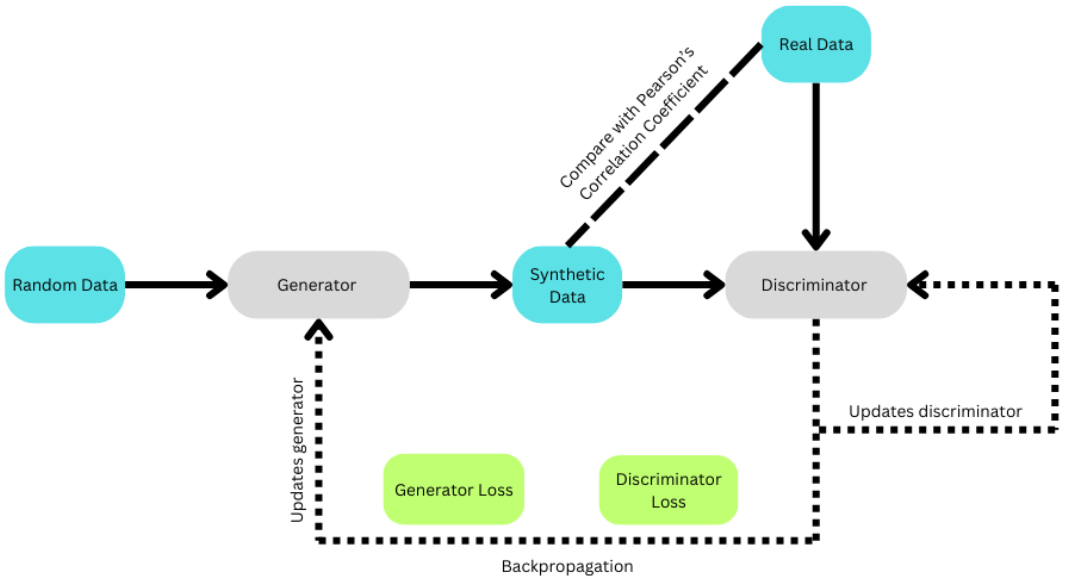



**Figure 1:** The above figure shows the structure of a Generative Adversarial Network model. The model has two main components, shown in gray: a generator, which creates data, and a discriminator, which compares the generated data to an inputted training dataset. The types of data that interact with the model are seen in blue, and include the random data that is first put into the generator, the synthetic data that comes out of the generator, and the real data that is used in the discriminator. The discriminator uses backpropagation, taking information from comparing the real data to the synthetic data and using that to instruct the generator to make more accurate synthetic data in the next cycle. It also updates the discriminator on how well it did at differentiating real data from synthetic data to improve the discrimination process. Discriminator loss and generator loss are measurements of training that show the difference between the generated data and the real data throughout the training procedure. Pearon's Correlation Coefficient is used for the final comparison of data generated by the model and original, real data used for training.

GANs are part of a recent series of generative artificial intelligence models, contributing to the development of famous uses such as GPT-4 [18]. GANs were initially suggested for image data types, and have been used to make synthetic faces that look like human faces

but do not belong to any real person [19, 20]. However, they have also been successful in generating other data types. Faces of missing people can be generated with a GAN model using text description inputs [21]. The results generated by GANs can be weighted within the model to help achieve specific training outcomes: for example, text generating models can be rewarded for "promoting diversity," or producing novel text combinations and phrases [22]. In this way, GANs are very versatile; they can receive many different data types, mixed data types, and even promote certain outcomes over others. Due to their simple two-module system, they are adaptable and accessible for novel use-cases.

### 1.5 GANs for Health Data

GANs have recently been applied to health data for a variety of methods. The most prevalent use of GANs in clinical studies is related to medical diagnostic imaging. A total of 745 publications using GANs for medical diagnostic imaging were reported between 2017 and 2022, according to a visualization analysis [23]. GANs have also been used to generate synthetic clinical data previously, particularly for the purpose of distributing medical information. Synthetic data created by GANs is not associated with any individual, and is therefore more ethical to distribute; this is why GANs are being used to anonymize electronic health records [24, 25].

A further use of GANs in health data is seen in PSSAM-GANs (Propensity Score Synthetic Augmentation Matching - Generative Adversarial Networks), which are used to provide causal propensity score matching for nonlinear treatment assignments in observational studies [26]. Using PSSAM-GAN has the added benefit of maintaining the sample size without the use of IPW (Inverse Probability Weighting). The use of GANs and other machine learning methods for synthetic control arms has been tested with data from EHRs, but has yet to be applied to RCTs [27]; as of this proposal, there is no published, accessible method for creating synthetic control arms for RCTs using GANs.

## Aim

The aim of this project is to create a machine learning model that can generate statistically indistinguishable control arms from experimental arms or previous, related data, with the ultimate goal of being applicable to rare disease RCTs.

# Objectives

The three main objectives of the project are as follows:

1. Determining which of current synthetic control arm methods are viable for RCTs
2. Creating a machine learning model that can generate control arms based on experimental arms or previous, related RCT data
3. Condensing the machine learning model and associated data cleaning processes into a library or package to make it accessible for others

# Methodology

## 1.6 Testing Current Synthetic Control Methods

There are three current methods for generating synthetic controls, each using a different type of input data: previous RCTs, observational studies, and external data (e.g. electronic health records). These three methods do not produce equal quality of synthetic controls, with previous RCTs hypothesized to produce a much higher quality [4]. However, no research has measured the difference in quality of the synthetic controls produced by these methods. It is important to know whether the difference in quality is statistically significant, and by how much, so that future models can balance the need for hard-to-get previous RCT data with an appropriate estimation of how much better it is in producing models than observational data or external data. In order to measure the difference in quality, three sets of synthetic controls will be made for the same treatment, each using one of the different methods. Synthetic controls for external and observational data will be generated using the Scpi package in R [28]. Synthetic controls based on previous RCTs will be generated using the Tidysynth package in R [29]. These will be compared to each other and the input data using Pearson's Correlation Coefficient.

## 1.7 Data Sources and Simulation

A large amount of data will be needed in order to test the differences in synthetic control methods and train the GAN model. Datasets will be requested from the Yale Open Data Access (YODA) project for the testing of synthetic control methods [30]. YODA has multiple datasets from RCTs and observational studies for the same treatments. Data

will also be simulated using the Kerus software from Exploristics [31]. This software can be used to generate statistically indistinguishable full datasets from summary level statistics [32]. Using this method, summary statistics from published articles related to treatments or diseases of interest can be utilized to generate a suitable copy of the original data. Following this, the GAN model can be provided with enough full datasets to generate a high-quality output. More data may be required for the testing of the model.

## 1.8 Building the GAN model

The Pytorch library was used to build a GAN model in Python [33]. This model was tested using a small sample of data from the Medicaldata package in R [34]; the dataset "Polyps" was chosen for testing due to its small size and simple characteristics. This dataset originated from the results of a 1993 RCT study of the nonsteroidal anti-inflammatory drug Sulindac and its impact on familial adenomatous polyposis [35]. The GAN model receives three data inputs: one column for demographic information, one column for the percentage difference between the after-treatment value and baseline value, and one column for labels of test group or control group. Data cleaning included one-hot-encoding demographic characteristics and combining the binary digits into a single column and converting all other column values to numeric values as needed. A small sample size like this will not produce robust results within the model, and this exercise was simply to build a working pipeline.
This equation describes the base loss of the GAN model:

$$L(\hat{y}, y) = [y.log\hat{y} + (1 - y).log(1 - \hat{y})]$$

$L$ represents the total loss. $y$ is the original data $\hat{y}$ is the generated data.

$$L(D(x), 1) = log(D(x))$$

In this equation, $y$ is set to 1 to represent the real data. $D$ represents the discriminator, and $G$ represents the generator. $\hat{y}$ is shown as a function of the discriminator, $D(x)$.

$$L(D(G(z)), 0) = log(1 - D(G(z)))$$

In this equation, $y$ is set to 0 to represent the fake data. $\hat{y}$ is shown as a function of the discriminator, $D(G(z))$, with $G(z)$ showing the input noise from the generator.

The final loss of the discriminator uses a maximized output of the above equations:

$$L^{(D)} = max[log(D(x)) + log(1 - D(G(z)))]$$

The final loss of the generator uses a minimized output of the above equations:

$$L^{(G)} = min[log(D(x)) + log(1 - D(G(z)))]$$

The final combined loss of the model, including both the outputs of the discriminator and generator, written for a single data point, is as follows:

$$L = \min_{G}\max_{D}[log(D(x)) + log(1 - D(G(z)))]$$

When extrapolated to the entire dataset, this is shown as follows:

$$\min_{G}\max_{D} V(D,G) = \min_{G}\max_{D}(E_{x\sim P_{data(x)}}[logD(x)] + E_{z\sim P_z(z)}[log(1 - D(G(_z)))])$$

$P_{data(x)}$ is the original data distribution. $P_z(z)$ is the input noise distribution.
The full code for the GAN model and data cleaning can be found here: https://github.com/N-cizauskas/GAN-for-Synthetic-Controls

## 1.9 Limitations

A generative synthetic model can only create data as diverse and inclusive as the datasets put into the model; with restrictions on the use of individual level patient data, this will further restrict the diversity of the data being input. Having minimal representation of minority groups (referring to broadly any group that is seen less within the dataset) in clinical data is dangerous as it may cause model overfitting. In the case of synthetic data, this would mean that certain characteristics of minority groups would be generated more often than they occur in the real world. The knock-down effect of this is that biased conclusions may be drawn as a result of the assumptions made by the synthetic data generator. To ensure minority groups are adequately represented in the final synthetic data outputs, the SMOTE (Synthetic Minority Oversampling Technique) method may be applied [36]. This method creates synthetic data of minority groups by combining them with closely related data. This will create novel combinations in data fields, unlike methods such as up-sampling or imputation which would boost the specific data fields already seen in these groups. Data closeness is determined by KNN

(K-nearest neighbour), with the default used in SMOTE being K-5 [37].

Moreover, GAN models suffer from two common problems: mode collapse and vanishing gradient [38]. Mode collapse is when the generator produces very similar values in its generated datasets and the discriminator fails to find the best strategy to determine fake vs real values [39]. As a result, the model will be stuck generating low variety datasets in a loop. Vanishing gradient occurs during the backpropagation period of training, where the network uses gradient descent to determine the minimum value of the loss function [40, 41]. Each iteration of training results in an update to its weight based on the partial derivative of the error function. If the change to the weight is too small, it will not update, and the gradient will not continue. This causes gradient-based learning models, like GANs, to be stuck without any way of continuing the training procedure. To combat these two problems, there are a variety of GAN subtypes and modifications that can be used [41–45]. There are also other machine learning models that are not affected by these limitations, such as LSTM (Long-Short Term Memory) models [46, 47], that use a memory-based infrastructure to retain information from previous iterations and compare it to current ones in order to measure weight and determine importance of change. This infrastructure could be implemented into the backpropagation phase of a GAN model if necessary.

Machine learning models are generally difficult to explain as a result of their "black box" algorithmic learning. In order to make the model more explainable, a series of tests will need to be developed. For example, changing the input data by adding or subtracting numbers within a set range and determining the range needed to significantly change the output data is one way of describing the model. More tests will need to be designed to better explain the model.

Training models on similar data to rare diseases requires a concrete understanding of similarity and a clearly understandable measurement of error. Matching of rare disease trials to other trials will be based on disease, symptoms, demographic affected, treatment, and genotype. This is not a perfect matching system, and will be limited by the availability of data. Some rare disease datasets will have closer matches than others. As a result, this is not designed to be a replacement for a randomized control trial; this model is made to be a last resort for rare diseases that are unlikely to receive funding or participation for a randomized control trial and have no other way to provide evidence for treatments.

# Ethical Issues

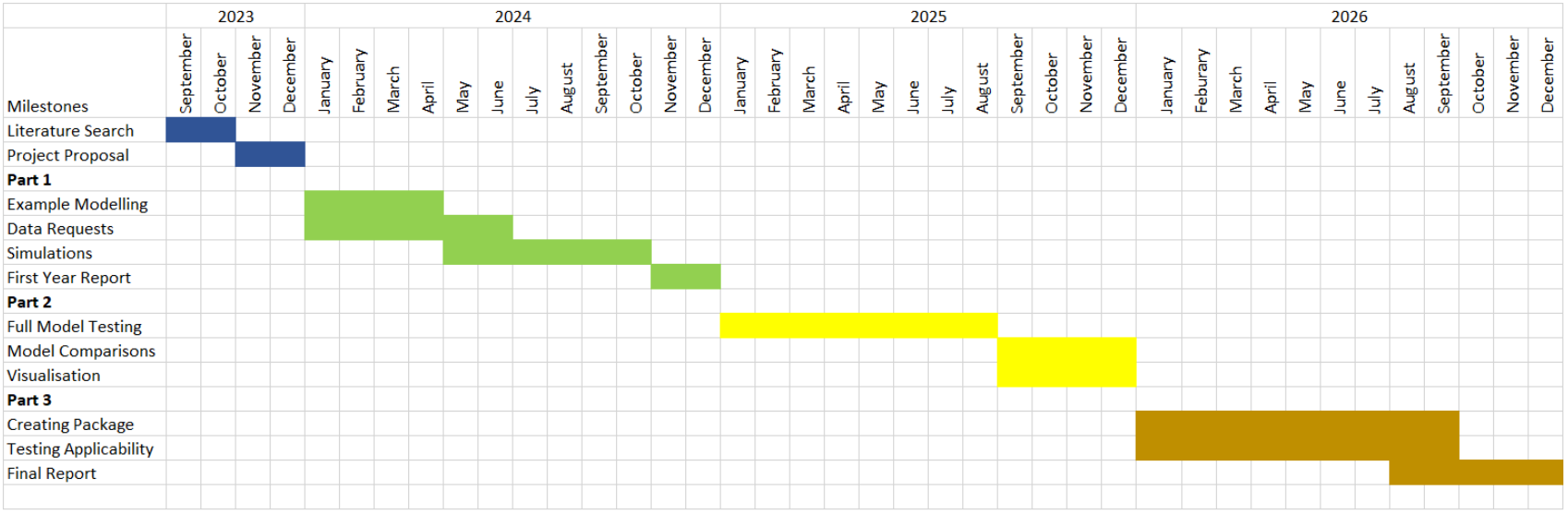
This is a primarily methodological project. Clinical datasets will be used for training and testing the models, but these will be publicly available, simulated, or anonymized by the data provider to not disclose individual level patient data. No ethical approval is required for this project.

There are ethical concerns regarding the use of full simulated datasets generated from summary level statistics: with statistically indistinguishable data being recreated, it is possible to accidentally generate an individual patient's data in full - particularly for minority patients. To prevent this, A K-anonymity of at least 3 will be required for each simulated dataset. K-anonymity is the number of individuals that have entirely unique data entries within the dataset [48, 49]. In other words, a K-anonymity of 3 would mean that there are at least 3 individuals within the dataset that have the same entry fields, for every used combination of those fields. Using a K-anonymity of 3 would ensure that no individuals would have their privacy risked during the simulation of data; the downside of this is that it may limit the specificity of the data and/or the number of individuals present from minority groups from some studies. However, as a result of recent concerns regarding ethical use of scraped training data for generative models among both researchers and the general public [50–52], this was deemed a necessary loss.

# Personal Development Plan

| Training Course | Offered By | Timeline |
|---|---|---|
| Leveraging External Information | Newcastle University | July, 2023 (taken prior to start) |
| Design and Analysis of Precision Medicine Trials | Newcastle University | September, 2023 |
| Statistical Computing, Statistical Inference | Academy for PhD Training in Statistics (APTS) | December, 2023 |
| Applied Stochastic Processes, Statistical Modelling | Academy for PhD Training in Statistics (APTS) | April, 2024 |
| Computer Intensive Statistics, High-dimensional Statistics | Academy for PhD Training in Statistics (APTS) | July, 2024 |
| Causal Inference, Statistical Machine Learning | Academy for PhD Training in Statistics (APTS) | September, 2024 |
| Building an R Package | Jumping Rivers | Unknown |
| Literature Review | Newcastle University | Unknown |

# Gantt Chart for PhD Plan



# References


[1] Commission E. EU Research on Rare Diseases;. Available from: https://research-and-innovation.ec.europa.eu/research-area/health/rare-diseases_en.

[2] Commissioner Oot. Rare Diseases at FDA; 2022. Publisher: FDA. Available from: https://www.fda.gov/patients/rare-diseases-fda.

[3] Cremers S, Aronson JK. Drugs for rare disorders. British Journal of Clinical Pharmacology. 2017 Aug;83(8):1607–1613. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5510061/.

[4] Thorlund K, Dron L, Park JJH, Mills EJ. Synthetic and External Controls in Clinical Trials – A Primer for Researchers. Clinical Epidemiology. 2020 May;12:457–467. Publisher: Dove Medical Press _eprint: https://www.tandfonline.com/doi/pdf/10.2147/CLEP.S242097. Available from: https://www.tandfonline.com/doi/abs/10.2147/CLEP.S242097.

[5] Sibbald B, Roland M. Understanding controlled trials: Why are randomised controlled trials important? BMJ. 1998 Jan;316(7126):201. Publisher: British Medical Journal Publishing Group Section: General practice. Available from: https://www.bmj.com/content/316/7126/201.

[6] Akobeng AK. Understanding randomised controlled trials. Archives of Disease in Childhood. 2005 Aug;90(8):840–844. Publisher: BMJ Publishing Group Ltd Section: Community child health, public health, and epidemiology. Available from: https://adc.bmj.com/content/90/8/840.

[7] Hariton E, Locascio JJ. Randomised controlled trials—the gold standard for effectiveness research. BJOG : an international journal of obstetrics and gynaecology. 2018 Dec;125(13):1716. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6235704/.

[8] Djurisic S, Rath A, Gaber S, Garattini S, Bertele V, Ngwabyt SN, et al. Barriers to the conduct of randomised clinical trials within all disease areas. Trials. 2017 Aug;18:360. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5539637/.

[9] Bouttell J, Craig P, Lewsey J, Robinson M, Popham F. Synthetic control methodology as a tool for evaluating population-level health interventions. J Epidemiol Community

Health. 2018 Aug;72(8):673–678. Publisher: BMJ Publishing Group Ltd Section: Theory and methods. Available from: https://jech.bmj.com/content/72/8/673.

[10] Berry DA, Elashoff M, Blotner S, Davi R, Beineke P, Chandler M, et al. Creating a synthetic control arm from previous clinical trials: Application to establishing early end points as indicators of overall survival in acute myeloid leukemia (AML). Journal of Clinical Oncology. 2017 May;35(15_suppl):7021–7021. Publisher: Wolters Kluwer. Available from: https://ascopubs.org/doi/abs/10.1200/JCO.2017.35.15_suppl.7021.

[11] Blondeau K, Schneider A, Ngwa I. A synthetic control arm from observational data to estimate the background incidence rate of an adverse event in patients with Alzheimer's disease matched to a clinical trial population. Alzheimer's & Dementia. 2020;16(S10):e043657. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/alz.043657. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/alz.043657.

[12] Ko YA, Chen Z, Liu C, Hu Y, Quyyumi AA, Waller LA, et al. Developing a synthetic control group using electronic health records: Application to a single-arm lifestyle intervention study. Preventive Medicine Reports. 2021 Dec;24:101572. Available from: https://www.sciencedirect.com/science/article/pii/S221133552100262X.

[13] Goldsack J. Synthetic control arms can save time and money in clinical trials; 2019. Available from: https://www.statnews.com/2019/02/05/synthetic-control-arms-clinical-trials/.

[14] Lyman JP, Doucette A, Zheng-Lin B, Cabanski CR, Maloy MA, Bayless NL, et al. Feasibility and utility of synthetic control arms derived from real-world data to support clinical development. Journal of Clinical Oncology. 2022 Feb;40(4_suppl):528–528. Publisher: Wolters Kluwer. Available from: https://ascopubs.org/doi/abs/10.1200/JCO.2022.40.4_suppl.528.

[15] Commissioner Oot. Statement from FDA Commissioner Scott Gottlieb, M.D., on FDA's new strategic framework to advance use of real-world evidence to support development of drugs and biologics; 2020. Publisher: FDA. Available from: https://www.fda.gov/news-events/press-announcements/statement-fda-commissioner-scott-gottlieb-md-fdas-new-strategic-framework-advanc

[16] Gretton C. Synthetic Control Arms: A Broader Clinical reach; 2021. Available from: https://www.pharmavoice.com/news/2021-10-synthetic-control-arms-a-broader-clinical-reach/612010/.

[17] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative Adversarial Nets. In: Advances in Neural Information Processing Systems. vol. 27. Curran Associates, Inc.; 2014. Available from: https://proceedings.neurips.cc/paper_files/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html.

[18] Islam A. A History of Generative AI: From GAN to GPT-4 - MarkTechPost; 2023. Available from: https://www.marktechpost.com/2023/03/21/a-history-of-generative-ai-from-gan-to-gpt-4/.

[19] Karras T, Aila T, Laine S, Lehtinen J. Progressive Growing of GANs for Improved Quality, Stability, and Variation. arXiv; 2018. ArXiv:1710.10196 [cs, stat]. Available from: http://arxiv.org/abs/1710.10196.

[20] Mahiuddin M, Khaliluzzaman M, Chowdhury MSA, Arefin MN. Fake Face Generator: Generating Fake Human Faces using GAN. International Journal of Advanced Computer Science and Applications. 2022;13(7). Available from: http://thesai.org/Publications/ViewPaper?Volume=13&Issue=7&Code=IJACSA&SerialNo=21.

[21] Shinde S, Pradhan T, Ghorpade A, Tale M. Face Generation from Textual Features using Conditionally Trained Inputs to Generative Adversarial Networks. 2023;.

[22] Xu J, Ren X, Lin J, Sun X. Diversity-Promoting GAN: A Cross-Entropy Based Generative Adversarial Network for Diversified Text Generation. In: Riloff E, Chiang D, Hockenmaier J, Tsujii J, editors. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics; 2018. p. 3940–3949. Available from: https://aclanthology.org/D18-1428.

[23] Zhang F, Wang L, Zhao J, Zhang X. Medical applications of generative adversarial network: a visualization analysis. Acta Radiologica. 2023 Oct;64(10):2757–2767. Publisher: SAGE Publications. Available from: https://doi.org/10.1177/02841851231189035.

[24] Arora A, Arora A. Generative adversarial networks and synthetic patient data: current challenges and future perspectives. Future Healthc J. 2022 Jul;9(2):190–193. Publisher: Royal College of Physicians Section: Opinion. Available from: https://www.rcpjournals.org/content/futurehosp/9/2/190.

[25] Venugopal R, Shafqat N, Venugopal I, Tillbury BMJ, Stafford HD, Bourazeri A. Privacy preserving Generative Adversarial Networks to model Electronic Health Records. Neural Networks. 2022 Sep;153:339–348. Available from: https://www.sciencedirect.com/science/article/pii/S0893608022002374.

[26] Ghosh S, Boucher C, Bian J, Prosperi M. Propensity score synthetic augmentation matching using generative adversarial networks (PSSAM-GAN). Computer Methods and Programs in Biomedicine Update. 2021;1:100020.

[27] Karletsos D, Wilson A, Krikov S, Vanderpuye-Orgle J. Generative Adversarial Networks to Generate Synthetic Controls: a Real-World Data Example;. Available from: https://www.ispor.org/docs/default-source/euro2021/fsgsisporpostersubmit-pdf.pdf?sfvrsn=f7ce5b85_0.

[28] Cattaneo MD, Feng Y, Palomba F, Titiunik R. scpi: Uncertainty Quantification for Synthetic Control Methods. arXiv; 2022. ArXiv:2202.05984 [econ, stat]. Available from: http://arxiv.org/abs/2202.05984.

[29] Dunford E. tidysynth: A Tidy Implementation of the Synthetic Control Method; 2023. Available from: https://cran.r-project.org/web/packages/tidysynth/index.html.

[30] The YODA Project; 2023. Available from: https://yoda.yale.edu/.

[31] Exploristics; 2023. Available from: https://exploristics.com/.

[32] McCartan S, Inshaw J, Young L, Nickless A, Kearney A, Warnock L, et al.. A comparison of individual participant data (IPD) and summary data to inform clinical trial simulation; 2021.

[33] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al.. PyTorch: An Imperative Style, High-Performance Deep Learning Library. arXiv; 2019. ArXiv:1912.01703 [cs, stat]. Available from: http://arxiv.org/abs/1912.01703.

[34] Higgins P. medicaldata: Data Package for Medical Datasets; 2021. Available from: https://cran.r-project.org/web/packages/medicaldata/index.html.

[35] Giardiello FM, Hamilton SR, Krush AJ, Piantadosi S, Hylind LM, Celano P, et al. Treatment of colonic and rectal adenomas with sulindac in familial adenomatous polyposis. The New England Journal of Medicine. 1993 May;328(18):1313–1316.

[36] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research. 2002 Jun;16:321–357. Available from: https://www.jair.org/index.php/jair/article/view/10302.

[37] Viadinugroho RAA. Imbalanced Classification in Python: SMOTE-ENN Method; 2021. Available from: https://towardsdatascience.com/imbalanced-classification-in-python-smote-enn-method-db5db06b8d50.

[38] Ding Z, Jiang S, Zhao J. Take a close look at mode collapse and vanishing gradient in GAN. In: 2022 IEEE 2nd International Conference on Electronic Technology, Communication and Information (ICETCI); 2022. p. 597–602. Available from: https://ieeexplore.ieee.org/abstract/document/9832406.

[39] Thanh-Tung H, Truyen T. Catastrophic forgetting and mode collapse in GANs | IEEE Conference Publication | IEEE Xplore. 2020 International Joint Conference on Neural Networks (IJCNN). 2020 Sep;Available from: https://ieeexplore.ieee.org/abstract/document/9207181.

[40] Hochreiter S. Recurrent Neural Net Learning and Vanishing Gradient. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. 1998;.

[41] Hu Z, Zhang J. Handling Vanishing Gradient Problem Using Artificial Derivative | IEEE Journals & Magazine | IEEE Xplore. Institute of Electrical and Electronics Engineers. 2021 Jan;9. Available from: https://ieeexplore.ieee.org/abstract/document/9336631.

[42] Su J. GAN-QP: A Novel GAN Framework without Gradient Vanishing and Lipschitz Constraint. arXiv; 2018. ArXiv:1811.07296 [cs, stat]. Available from: http://arxiv.org/abs/1811.07296.

[43] Weng L. From GAN to WGAN. arXiv; 2019. ArXiv:1904.08994 [cs, stat]. Available from: http://arxiv.org/abs/1904.08994.

[44] Durall R, Chatzimichailidis A, Labus P, Keuper J. Combating Mode Collapse in GAN training: An Empirical Analysis using Hessian Eigenvalues. arXiv; 2020. ArXiv:2012.09673 [cs]. Available from: http://arxiv.org/abs/2012.09673.

[45] Karthika S, Durgadevi M. Generative Adversarial Network (GAN): a general review on different variants of GAN and applications. International Conference on Communication and Electronics Systems (ICCES). 2021 Jun;Available from: https://ieeexplore.ieee.org/abstract/document/9489160.

[46] Hu Y, Huber A, Anumula J, Liu SC. Overcoming the vanishing gradient problem in plain recurrent networks. arXiv; 2019. ArXiv:1801.06105 [cs]. Available from: http://arxiv.org/abs/1801.06105.

[47] Staudemeyer RC, Morris ER. Understanding LSTM – a tutorial into Long Short-Term Memory Recurrent Neural Networks. arXiv; 2019. ArXiv:1909.09586 [cs]. Available from: http://arxiv.org/abs/1909.09586.

[48] Sweeney L. k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. 2002 Oct;10(05):557–570. Publisher: World Scientific Publishing Co. Available from: https://www.worldscientific.com/doi/abs/10.1142/S0218488502001648.

[49] El Emam K, Dankar FK. Protecting Privacy Using k-Anonymity. Journal of the American Medical Informatics Association : JAMIA. 2008;15(5):627–637. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2528029/.

[50] Falconer S. Privacy in the age of generative AI - Stack Overflow; 2023. Available from: https://stackoverflow.blog/2023/10/23/privacy-in-the-age-of-generative-ai/.

[51] Ferrara E. GenAI Against Humanity: Nefarious Applications of Generative Artificial Intelligence and Large Language Models. arXiv; 2023. ArXiv:2310.00737 [cs]. Available from: http://arxiv.org/abs/2310.00737.

[52] Marwala T, Fournier-Tombs E, Stinckwich S. The Use of Synthetic Data to Train AI Models: Opportunities and Risks for Sustainable Development. arXiv; 2023. ArXiv:2309.00652 [cs]. Available from: http://arxiv.org/abs/2309.00652.