electronics

MDPI

Review

Bias Mitigation via Synthetic Data Generation: A Review

Mohamed Ashik Shahul Hameed , Asifa Mehmood Qureshi and Abhishek Kaushik *

Dundalk Institute of Technology, A91 K584 Dundalk, Ireland
* Correspondence: abhishek.kaushik@dkit.ie

Abstract: Artificial intelligence (AI) is widely used in healthcare applications to perform various tasks. Although these models have great potential to improve the healthcare system, they have also raised significant ethical concerns, including biases that increase the risk of health disparities in medical applications. The under-representation of a specific group can lead to bias in the datasets that are being replicated in the AI models. These disadvantaged groups are disproportionately affected by bias because they may have less accurate algorithmic forecasts or underestimate the need for treatment. One solution to eliminate bias is to use synthetic samples or artificially generated data to balance datasets. Therefore, the purpose of this study is to review and evaluate how synthetic data can be generated and used to mitigate biases, specifically focusing on the medical domain. We explored high-quality peer-reviewed articles that were focused on synthetic data generation to eliminate bias. These studies were selected based on our defined inclusion criteria and exclusion criteria and the quality of the content. The findings reveal that generated synthetic data can help improve accuracy, precision, and fairness. However, the effectiveness of synthetic data is closely dependent on the quality of the data generation process and the initial datasets used. The study also highlights the need for continuous improvement in synthetic data generation techniques and the importance of evaluation metrics for fairness in AI models.

Keywords: synthetic data; artificial data; bias; fairness; AI; data generation

check for updates

Citation: Shahul Hameed, M.A.; Qureshi, A.M.; Kaushik, A. Bias Mitigation via Synthetic Data Generation: A Review. *Electronics* **2024**, *13*, 3909. https://doi.org/10.3390/electronics13193909

Academic Editor: Grzegorz Dudek

Received: 29 August 2024
Revised: 23 September 2024
Accepted: 24 September 2024
Published: 2 October 2024

Correction Statement: This article has been republished with a minor change. The change does not affect the scientific content of the article and further details are available within the backmatter of the website version of this article.

1. Introduction

Bias in AI models can be defined as systematic errors that affect the algorithm and unfairly favor certain outcomes over others [1]. These biases can originate from various sources, such as using imbalanced or underrepresented datasets to train the models, the algorithms themselves, and the ways AI models are being deployed by humans. In healthcare, these biases can lead to significant consequences of disparities in treatment outcomes for specific racial, gender, or demographic groups, resulting in unequal treatment that may lead to life-threatening incidences. They have a direct impact on patient medical outcomes. Biased models can lead to misdiagnosis, inappropriate treatment plans, and unequal access to medical resources [2]. For instance, a used model to estimate breast cancer density is trained on data that are under-inclusive of African-American women and will generate recommendations that are not well-suited for that population [3]. Moreover, Kiyasseh et al. [4] deployed a surgical AI model to assess the skill level of robotic surgeries in completing different surgical activities. However, it was found that the model exhibited an underskilling or overskilling bias. Underskilling occurs when an AI model incorrectly predicts that a certain surgical skill is of a lower grade than it actually is, hence reducing surgical performance. In contrast, overskilling occurred when the AI model incorrectly improved surgical performance by estimating that a certain skill would be a higher caliber than it was.

Therefore, the broad aim behind ensuring fairness, accuracy, and equality through bias mitigation in AI models within the medical domain is to create fair and equal outcomes for a diverse set of populations.

There are several techniques to eliminate biases from datasets. These techniques include the collection of diverse and representative data [5], handling at the algorithmic level [6], pre-processing of datasets and post-processing of the model output to reduce bias [7], and synthetic data generation [8].

Generating data is one of the several strategies to mitigate bias. It can be defined as artificial data that mimics the statistical properties and patterns of real-world information [9]. While other techniques tend to reduce or process the datasets to ensure fairness, which may result in information loss, synthetic data generation helps to preserve the data distribution and add statistically similar data samples to reduce bias. In healthcare settings, synthetic data can be utilized as a substitute for real patient data while avoiding privacy issues, which allows for the simulation and sharing of sensitive patient data. There exist multiple methods, including Generative Adversarial Networks (GANs), Variational Auto-Encoders (VAEs), the Synthetic Minority Over-sampling Technique (SMOTE), Bayesian networks, and statistical sampling methods, to generate synthetic data. These methods have been used vastly to generate new artificial datasets that capture the structure and distribution of actual datasets. These synthetic datasets can help to tackle situations where there is bias or no additional data to work with while trying to improve the fairness and accuracy of AI algorithms [4]. Using synthetic data in healthcare is also a solution to patient data privacy. Moreover, synthetic data allows you to create big, diverse datasets to train models.

This article reviews research studies that are focused on bias mitigation via synthetic data generation techniques.

Motivation

AI systems can produce biased results that both reflect and amplify human prejudices within a community, including historical and contemporary social injustice. Bias may exist in the algorithm, the original training set, or the predictions the algorithm makes. It can be passed on through racial, religious, and gender stereotypes [10]. Synthetic data can be a viable solution to preserve patient data privacy and mitigate bias in the data. Synthetic data methodologies tend to capture the structure and distributions found in real datasets while minimizing bias and protecting individually identifiable [11]. Moreover, synthetic data allows us to create big, diverse datasets to train AI models. Therefore, this paper explores the literature to highlight the application of synthetic data generation to mitigate bias. The study provides a comprehensive review of the methods and techniques used to generate synthetic data. Also, the limitations of these models are discussed. The hypothesis and formulated research questions are as follows:

Hypothesis: *Synthetic data can be used to mitigate bias in medical data effectively, resulting in unbiased, accurate, and equitable healthcare datasets.*

RQ1. What are the most common methods or techniques to generate synthetic data to handle biases in the dataset?

RQ2. What are the limitations of the existing techniques used for synthetic data generation?

The rest of the article is structured as follows:

Section 2 explains the search methodology. Section 3 gives an overview of the bias mitigation techniques. Section 4 discusses the hypothesis and research questions. Section 5 concludes the discussion.

2. Search Methodology

Figure 1 shows the detailed search methodology used to conduct this review. Several major databases, including Google Scholar, IEEE Xplore, Pubmed, ScienceDirect, and ACM Digital Library, were explored. Google Scholar was used for its broad and inclusive nature. It covers many topics and has many articles from interdisciplinary fields, so it is a great resource for literature reviews. PubMed was chosen for its focus on biomedical and life science literature. Since the focus is on AI applications in healthcare, PubMed has high-quality reviews and survey papers to obtain peer-reviewed medical research articles on the impact and implementation of AI in clinical settings. The ACM Digital Library was included for its focus on computing and information technology research. It is great for finding research on algorithm development, computational methods, and technical advancements in synthetic data generation and bias mitigation in AI. IEEE Xplore was chosen for its large collection of technology research and application implementation papers, especially in AI, ML, and synthetic data. It has technical papers and theory conference articles on the latest AI technologies and their applications in healthcare.

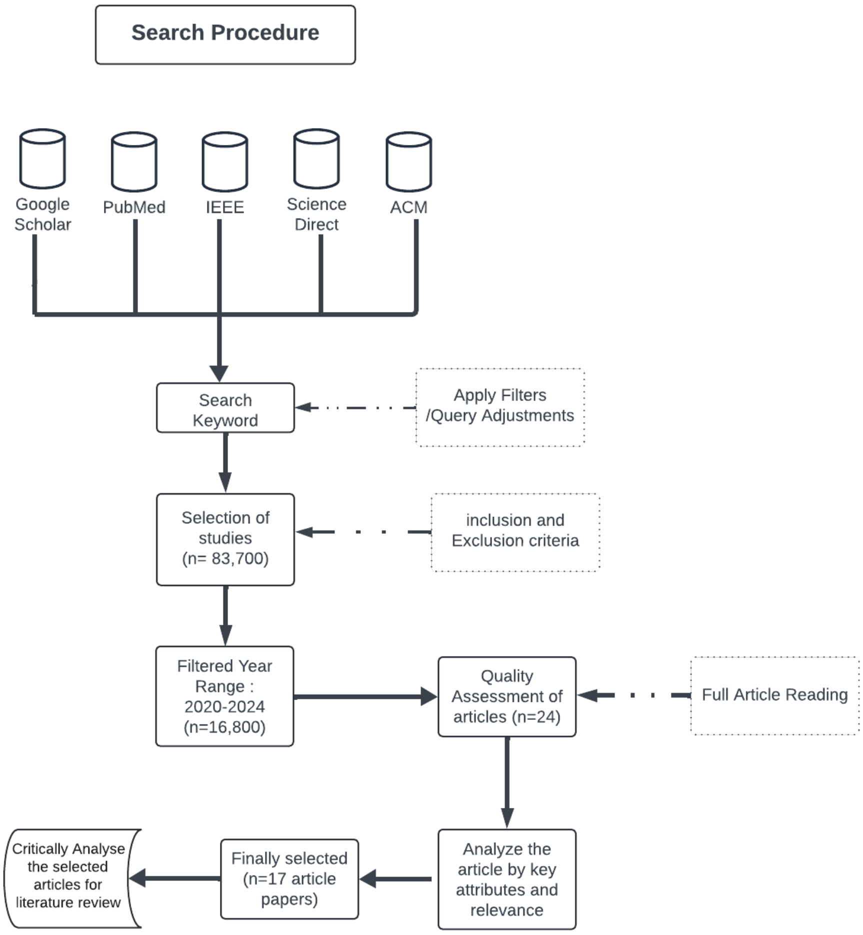
Search Procedure
Google Scholar
PubMed
IEEE
Science Direct
ACM
Search Keyword
Apply Filters /Query Adjustments
Selection of studies (n= 83,700)
inclusion and Exclusion criteria
Filtered Year Range : 2020-2024 (n=16,800)
Quality Assessment of articles (n=24)
Full Article Reading
Analyze the article by key attributes and relevance
Finally selected (n=17 article papers)
Critically Analyse the selected articles for literature review

Figure 1. Search methodology workflow.

The selections are based on the scope, relevance, and credibility of the databases and article papers in the fields of AI, Machine Learning (ML), synthetic data, and healthcare. A search query was formulated to find research articles on bias in AI, ML, Deep Learning (DL), and synthetic data generation in healthcare:

"((Bias) AND (Artificial Intelligence OR Machine Learning OR Deep Learning) AND (Synthetic Data) AND (text OR tabular dataset))"

Slight adjustments were made to the query to refine the results further. These adjustments included using synonyms of these terms, including fairness, biases, data bias, artificial data, generated data, and medical data.

To obtain the most relevant and recent studies, inclusion and exclusion criteria were devised as follows:

2.1. Inclusion Criteria

- Research articles that are focused on bias and bias handling in AI models using synthetic data generation were included.

- Research articles published between 2020 and 2024 were included to review the relevant, latest techniques.
- Research studies that were published in conferences, journals, book chapters, or proceedings were included.
- Research articles only written in the English language were selected.

2.2. Exclusion Criteria

- Research articles that did not align with our research questions were excluded.
- Non-peer-reviewed articles were not considered.

About 83,700 papers were found in this first search using the Google Scholar database platform. By applying inclusion and exclusion criteria and restricting the year of publication between 2020 and 2024, the total was dropped to about 16,800 papers. The dynamic of scientific studies in this field from 2020 to 2023 was approximately recorded as 13,200, 17,400, 21,500, and 28,000. Therefore, the average annual growth rate of articles in this field was approximately 29%.

Then, a quality assessment was performed to ensure the relevance and credibility of the papers, and the selection was narrowed down to 50 papers. The shortlisted articles were examined further for eligibility on the basis of title relevancy and abstracts; a total of 26 articles were removed, and the remaining articles were assessed by a full article reading. After a critical analysis, we finally selected 17 papers for the literature review. Table 1 summarizes the number of articles after each step.

Table 1. Number of articles selected at each step.

Stages	Number of Articles
Initial results (using query)	83,700
After applying inclusion and exclusion criteria	16,800
After quality assessment	50
Based on the abstract and conclusion	24
Full article reading (Finally selected articles)	17

The proportion of selected research studies from each database is shown in Figure 2.

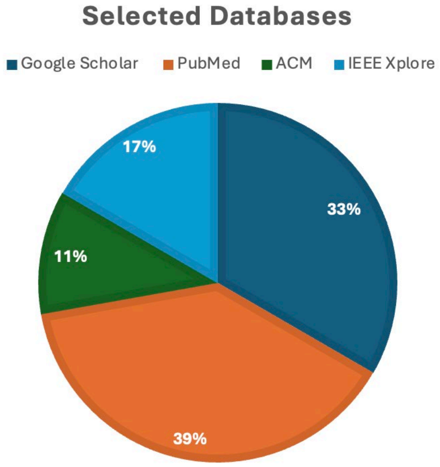

Figure 2. Proportion of the selected articles from each database.

Figure 3 shows the year-wise distribution of the selected articles. The year 2021 constituted the highest number of selected articles, preceded by the years 2022 and 2023. Moreover, the category of the selected articles is given in Table 2. A total of 12 journals, 4 conferences, and 1 doctoral dissertation were reviewed.

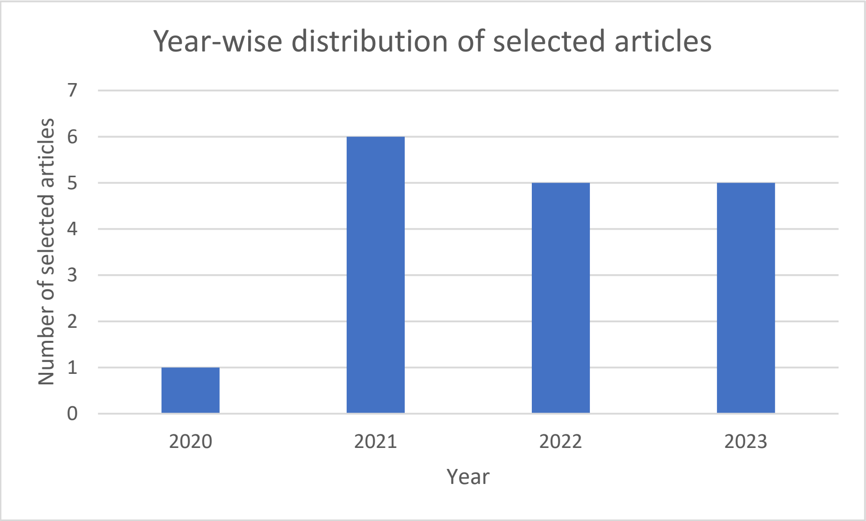

Figure 3. Year-wise distribution of the selected articles.

Table 2. Overview of selected article categories and count.

Category	Number of Articles
Academic Journal paper	12
Conference paper	4
Doctoral dissertation	1

3. Bias Mitigation via Synthetic Data Generation

Several techniques have been proposed in the literature to mitigate bias using synthetic sample generation. Table 3 summarizes the studies and the methodologies along with the dataset, modality, strengths, and weaknesses of each reviewed article.

Table 3. Overview of the methodology, dataset, modality, evaluation metric, strengths, and limitations of the reviewed research articles.

Refs.	Year	Methodology	Dataset	Modality	Evaluation Metric	Strengths	Limitations
[12]	2023	Analyses explainable artificial intelligence (XAI) methods and suggests metrics to determine the technical characteristics of the methods. Focuses on XAI methods such as LIME and SHAP, applied to machine learning models using synthetic data.	A large and heterogeneous corpus of one million Dutch EHR notes.	text data	Balanced accuracy and fairness scores to compare real and synthetic models, time-series metrics, and disparate impact metrics for representational bias and evaluating resemblance.	Auditing pipelines for robust evaluation under varying data, combines GANs, fairness metrics, and bias mitigation algorithms.	Focuses only on synthetic data generation and fairness, lack of datasets, and complexity in adversarial auditing methods.
[13]	2023	Focuses on generating synthetic data strategies for fairness assessment. Explores techniques for creating synthetic datasets that can be used to stress-test machine learning models and assess bias mitigation algorithms.	Healthcare data	Temporal and non-temporal datasets. Subgroup-level analysis of protected attributes such as gender and race	Metrics include faithfulness (correlation between feature and model predictions), monotonicity (correctness of feature), and incompleteness (effect of noise in feature).	Addresses the critical area of explainable AI, essential for user trust, provides insights into the evaluation of XAI methods, and guides future research.	The imperfection of existing XAI methods undermines user trust, does not delve into specific datasets, and lacks diverse data.

Table 3. *Cont.*

Refs.	Year	Methodology	Dataset	Modality	Evaluation Metric	Strengths	Limitations
[14]	2022	Utilized GANs. Interactive GUI tool to generate synthetic data. Integrated bias detection. Evaluated with LFR. Uses statistical fairness metrics including Statistical Parity Difference (SPD).	German Credit dataset, adult dataset	Tabular data	VP model-based clustering approach compared to clustering based solely on original patient data reduced biases by data from different hospitals with ARDS.	VP modeling approach mitigates biases by heterogeneous datasets and improves cluster discovery.	Depends on the availability of relevant observational data, and the complexity of use.
[15]	2022	Employs mechanistic virtual patient (VP) modeling to capture specific features of patients' states and dynamics while reducing biases introduced by heterogeneous datasets.	Observational data of mixed origin, including data pooled from different hospitals. ICU datasets like MIMIC or HiRID	text and tabular data	LFR improved fairness by 62% and 17.5%. Reduced SPD by 93%.	Interactive GUI, effective bias mitigation, improved fairness, comprehensive bias mitigation.	Specific to used datasets, generalizability issues, rely on specific metrics, limited to specific algorithms.
[16]	2021	DECAF framework. Structural causal model. Biased edge removal. Evaluated fairness. SCM focus on understanding causal relationships within data.	Tabular data	Tabular data	AUC, ROC, and Precision–Recall curves improved the representation of under-represented groups	Privacy preservation, effective bias identification, maintaining high data quality, and versatility for various medical datasets.	Dependence on initial dataset quality, methodological complexity, and parameter sensitivity may need adjustments for different dataset
[17]	2021	Data size reduction. Simulation of data biases. Bayes Boost approach for bias handling. Probabilistic models and synthetic data generation. Comparison with SMOTE and ADASYN for generating synthetic datasets	CPRD-based synthetic datasets (Synthetic CVD dataset) with 499,344 patients and 21 variables.	Tabular data	High-quality synthetic data, maintained utility of real data, fairness evaluated by demographic parity, equal opportunity.	Compatible with multiple fairness definitions, high-quality data, effective debiasing, and theoretical guarantees.	Requires causal relationship understanding, definition-specific results, focused on tabular data, and expert knowledge needed.
[18]	2020	GANs for generating synthetic biomedical signals using LSTM generator and CNN discriminator. Comparison with real signals to assess quality. Training using signal augmentation techniques.	Biomedical signal datasets like ECG, EEG, EMG, and PPG (17 types of ECG signals).	Time-series data (signal data)	Signal fidelity, noise levels, and usability. Signal similarity: high accuracy approx. 92%.	High-quality synthetic signals, reduce data scarcity issues, high accuracy, and versatility.	Computationally intensive, requires high-quality initial datasets, and model complexity.
[19]	2023	Review of GAN applications in medicine. Discussion on ethical security privacy conservation. Comparative analysis of different GAN architecture techniques with visual review.	Various medical datasets such as MRI, CT scans, and retinal images.	Structured and unstructured data	Accuracy, precision, recall, F1-score, and visual quality assessments. GAN metrics vary in each use case.	GANs generate synthetic medical images for datasets and new data patterns	Focus on theoretical aspects, limited empirical data, and ethical concerns.

Table 3. *Cont.*

Refs.	Year	Methodology	Dataset	Modality	Evaluation Metric	Strengths	Limitations
[20]	2022	Review of clinical papers from PubMed about AI-assessed disparities in dataset country sources, and clinical demographics (nationality, sex, expertise). Manually tagged a subsample of articles to train a model using transfer-learning techniques to predict. Studied transfer learning with the BioBERT model and automated tools like Entrez Direct and Gendarize.	--	--	Country metrics: U.S. 40.8%, Chinese 13.7%. Clinical Specialties: Radiology: 40.4%, Pathology: 9.1%.	A comprehensive analysis of 300,000 articles, highlights disparities in data, transfer-learning and automated tools to analyze the dataset.	Focuses on U.S. and China data only.
[21]	2021	Discusses the critical issues of fairness, bias, and the use of AI and ML in global health, specifically in Low- and Middle-Income Countries. It proposes a framework for appropriateness, bias, and fairness	Diagnosis clinical records of 200 consecutive patients at a clinic.	Text data	Disparate impact scores, fairness prediction from different groups, measure fairness in true positive rates across groups.	Addresses AI biases in underrepresented Indigenous populations, employing fairness metrics enhances transparency	Data specific to New Zealand, dependent on available health data.
[22]	2022	Analyzes data and algorithmic bias related to data collection and model development, training, and testing using health data collected in New Zealand. Measures fairness using DI scores, impact scores, equal opportunities and equalized tabular data.	Health data collected in New Zealand, including Maori populations. NZ-GP Harms and NZ-GDF Diabetes Mellitus (kaggle), PIMA, SACardio, MNCD-RED	Tabular data (NZ-GP Harms), Free-text	Accuracy was approximately 89.2% for both genders, models showed varying ROC for different groups.	A robust framework assessing fairness, bias, and appropriateness in AI/ML applications, offers clear guidelines for AI/ML fairness and reducing bias.	Heavily dependent on the quality and diversity of the training data, practical challenges in deploying AI/ML, and bias inherent in biological differences.
[23]	2023	Deep reinforcement learning framework for bias mitigation. Adversarial training to reduce biases during model development. Application on COVID-19 screening and patient discharge prediction	eICU Collaborative Clinical data Research Database	Text and structured data with multiple attributes such as patient demographics, diagnoses, treatments, and outcomes	Improved fairness and reduced bias in clinical AUC-ROC score of 0.818 to 0.875 using the XGBoost model and Random Forest (RF) model	Effective bias mitigation, application on real-world datasets, enhances clinical decision support.	Requires extensive resources, dependent on the quality of initial data. Limited dataset information
[24]	2022	Synthetic data generation to address the issues of small and imbalanced medical datasets. Algorithms containing tabular data of different sizes to test data, balancing using SMOTE, and ADASYN and data augmentation via Gaussian Copulas, and CTGANs	Eight medical datasets including MNCD, Bangladesh Diabetes	Tabular data	Synthetic data effectively maintained, improved synthetic data, and enhanced machine learning model training without the original dataset. Metrics: MNCD, PCD, KLD, MMD for statistical similarity and F1-score.	Evaluated multiple datasets, advanced synthetic data generation techniques such as CTGAN, and the effectiveness of Gaussian Copula in preserving data structure.	It can be time-consuming, errors encountered depend on the quality of original datasets, and challenges with CTGANs in maintaining balance and performance.

Table 3. *Cont.*

Refs.	Year	Methodology	Dataset	Modality	Evaluation Metric	Strengths	Limitations
[25]	2021	Neural language models (LSTM and GPT-2) for generating synthetic EHR text. Joint generation of synthetic text and annotations for NER with in-text PHI tag. User study for privacy assessment. Privacy was evaluated using ROUGE n-gram and BM25 scoring. Combining real and synthetic data to improve recall without manual annotation	A large and heterogeneous corpus of one million Dutch EHR notes.	Text-based (EHR data)	The accuracy of de-identification is 95%, The LSTM method produces synthetic text with higher utility compared to GPT-2.	Addresses privacy concerns, reduces manual annotation efforts and compares the utility of synthetic data to real data.	Privacy risks with synthetic text, challenges in evaluating privacy-preservation. The study focuses on Dutch EHR data.
[26]	2021	Discussed various AI models and algorithms for clinical outcome prediction. These models include decision trees, linear models, regression models, ensemble learning and neural networks.	Multiple clinical datasets include patient records (EHR), imaging, and genetic data.	Structured and unstructured data (mixed: text, tabular, images)	--	High predictive accuracy, robust model validation, personalized medicine, and effective data integration.	Model interpretability, and generalizability to diverse populations, require large datasets and lack privacy concerns.
[27]	2023	Synthetic data generation with controlled bias. Qualitative analysis of bias impact. Open-source toolkit. Controlled scenarios for bias studies.	Synthetic datasets with predefined bias	Synthetic data generation (open-source toolkit)	Qualitative analysis of bias impact.	Precise control over bias types, open source availability, flexibility to model various biases, contribution to fairness research	May not capture real-world complexities, generalization issues, requires expertise, scope of bias types might be limited.
[11]	2021	Synthetic data generation using various techniques. Comparative study. Analysis of bias reduction	Anonymized CPRD data. Large-scale datasets including real and synthetic data. Request-based access.	Tabular EHR data	Bias reduction by 15–20%, accuracy improved by 10–12%, privacy with <5% re-identification risk, and data utility retained by 90–95%.	Effective bias mitigation ensures privacy, scalable, and versatile.	Varying results and quality depend on models, resource-intensive, and metric reliance.

Detailed Analysis of the Reviewed Articles

This section provides a detailed analysis of the reviewed articles:

Hazra et al. [18] focus on creating synthetic biomedical signals by utilizing (GANs) to improve information accessibility of training data for medical students and machine learning models. The strategy includes an LSTM generator and Convolutional Neural Network (CNN) discriminator, prepared by utilizing penalty-based strategies assessed against real signals for quality. Datasets incorporate Electrocardiogram (ECG), electroencephalogram (EEG), electromyography (EMG), and Photoplethysmography (PPG) signals. The ponder reports tall-flag constancy and similitude with accuracy signals, accomplishing roughly 92% accuracy. In spite of its potential to moderate an information shortage, the approach is computationally complex and depends on the quality of initial datasets. Moreover, another approach named BayesBoost is proposed by [17]. It centers on taking care of the simulation of data biases through synthetic data generation. It employs probabilistic models and synthetic dataset generators like Bayesian systems and compares them with strategies like SMOTE and AdaSyn. The method employs Clinical Practice Research Datalink (CPRD)-based synthetic datasets, particularly the Synthetic Cardiovascular Disease (CVD) dataset that constitutes 499,344 patient records having 21 features. The results show an improvement in the Area Under Curve (AUC), Receiver Operating Characteristic (ROC), and Precision–Recall curves. The proposed model is effective in privacy preservation and bias identification. However, the approach is complex and relies on the initial dataset quality, requiring alterations for distinctive datasets.

Also, Breugel et al. [16] investigate the DEbiasing CAusal Fairness (DECAF) system that utilizes the Structural Causal model and one-sided edge expulsion to produce fair synthetic data. The consideration centers on data generation, guaranteeing high-quality synthetic data. Fairness is evaluated by demographic parity and equal opportunity. Decency is assessed by utilizing statistical equality and rise-to-opportunity measurements. The system is congruous with numerous fairness definitions, but it requires an understanding of causal connections and centers essentially on tabular data, thus requiring expert knowledge for further studies.

Gujar et al. [14] proposed a framework named GenEthos that utilizes GANs and Interactive Graphical User Interface (GUI) tools to produce synthetic data with integrated bias detection. The tool is assessed by using datasets, including the German Credit and Adult datasets. The framework is evaluated by factual fairness measurements like Statistical Parity Difference (SPD) and Disparate Impact (DI). Learning Fair Representation (LFR) has improved fairness by 62%. Reduced SPD by 93%. Despite its compelling Interactive GUI, effective bias mitigation, improved fairness moderation, and comprehensive assessment, the approach is particular to the utilized datasets and faces generalizability and specific algorithm issues.

Sharafutdinov et al. [15] utilize mechanistic virtual patient (VP) modeling to decrease biases in heterogeneous ICU datasets to identify acute respiratory distress syndrome (ARDS) using machine learning techniques. Utilizing observational information with mixed origin data from different hospitals, the VP model-based clustering approach, compared to clustering based solely on original patient data, makes strides in clustering approaches and mitigates bias. The approach depends on the accessibility of heterogeneous datasets, improved cluster discovery information, and relevant observational data, making it complex to use.

Paladugu et al. [19] audit the application of GANs in restorative imaging and AI preparation, emphasizing the generation of synthetic data information era and privacy conservation. It compares different GAN models utilizing restorative datasets like Magnetic Resonance Imaging (MRI), Computed Tomography (CT) scans, and retinal images. Measurements like precision, accuracy, review, visual quality assessments, and F1-score are utilized to assess the information that has been created. GAN metrics vary in each use case, whereas GANs appear guaranteed in making high-quality synthetic data generation and design patterns. However, this approach centers on hypothetical perspectives with restricted experimental information and limited empirical data. Moreover, Celi et al. [20] explore how bias in AI emerges from clinical medication information, centering on the populace and data-source disparities. Utilizing PubMed clinical papers from 2019, it surveys aberrations about AI-assessed disparities in dataset country sources and clinical demographics (nationality, sex, expertise). They manually tagged a subsample of articles to train a model using transfer-learning techniques with the Bidirectional Encoder Representations from Transformers for Biomedical Text Mining (BioBERT) model to make predictions. The result of the survey highlights critical aberrations, especially in the US (40.8%) and Chinese (13.7%) information. Clinical specialties: Radiology: 40.4% utilize transfer learning with BioBERT, whereas a comprehensive analysis of 300,000 articles highlights the disparities in data. Transfer-learning and automated tools to analyze the dataset are mostly trained on US and Chinese datasets.

Also, Fletcher et al. [21] investigate a system for surveying fairness, bias, and the use of AI/ML in worldwide global healthcare, specifically in low and middle-income nations. Utilizing clinical records from 200 patients to train logistic regression models accomplished around 89.2% accuracy for both genders; the ROC may vary for different groups, whereas the system depends on the quality and diversity of the training data. Additionally, Yogarajan et al. [22] analyze information and algorithmic bias in data collection and model development, training, and testing by utilizing healthcare records collected in New Zealand, centering on the Maori populations. Fairness measures, including impact scores, equal opportunities, and equalized odds, were utilized to assess bias and disparities among them. The results show evidence of bias when changes were made to algorithmic designs. However, employing fairness metrics enhances transparency to reduce bias in underrepresented populations.

Yang et al. [23] present a Deep Reinforcement Learning (DRL) system for bias mitigation that was acquired in the data collection process. The model is evaluated for COVID-19 predictions by eliminating any hospital and ethnicity-based biases. The technique includes Adversarial training to decrease biases during model development. The framework is also evaluated on the Electronic Intensive Care Unit (eICU) Collaborative Inquire Database to predict patient discharge status. The result reports an AUC-ROC score of 0.818 to 0.875 using the XGBoost model and Random Forest (RF) model.

Libbi et al. [25] explore the utilization of synthetic data information in healthcare, particularly for preserving privacy for individuals, by creating Electronic Health Record (EHR) data. Utilizing neural language models like LSTM and General-Purpose Transformers-2 (GPT-2), it aims to generate synthetic EHR content that incorporates in-text Protected Health Information (PHI) labels with annotations for Named-Entity Recognition (NER). Privacy was evaluated using a recall-oriented understudy for the Gisting Evaluation (ROUGE) n-gram and Best Match (BM25) scoring. The method utilized a dataset of one million Dutch EHR notes to train these models. The result illustrates that combining real and synthetic data information has accomplished a 95% accuracy accuracy in de-identification, especially favoring the LSTM strategy for higher utility compared to GPT-2. However, the study discoveries are specific to Dutch EHR datasets.

Pettit et al. [26] discuss the application of AI, ML, and DL in healthcare, with a specific focus on the data processing model, clinical outcome prediction, utilization of data, and AI fairness. To reduce bias in the dataset, different AI models and algorithms have been discussed. These methods include linear and regression models, decision trees, ensemble learning, and neural networks, whereas the paper underscores accuracy, precision, recall, and the F1-score to measure the performance of the AI models.

Rodriguez et al. [24] examine synthetic data generation to address the issues of small and imbalanced medical datasets containing tabular data in different sizes. The proposed method utilizes techniques like Gaussian Copulas, Conditional Tabular Generative Adversarial Networks (CTGAN), and Synthetic Data Vault (SDV). The models are evaluated on eight different medical datasets, including MNCD, MNCD-RED, BANG, EarlyDM, HeartDis, Kidney, PIMA, and SACardio. The result shows that synthetic data effectively improved and enhanced machine learning model training without the original dataset. In spite of its evaluation on multiple datasets, advanced synthetic data generation techniques such as CTGAN and SDV are dependent on the quality of original datasets and are not time efficient.

Baumann et al. [27] present a strategy to generate synthetic data with a controlled bias to analyze the impact of bias on model performance. They proposed an open-source toolkit to generate synthetic data with different types of biases. The produced synthetic datasets with predefined bias are assessed to increase awareness of bias in AI and how it affects individuals and society. Even though the toolkit offers control over biases that need to be introduced, it may not capture real-world complexities. Moreover, Draghi et al. [11] compare different synthetic data generation methods to reduce bias in healthcare data. Utilizing anonymized CPRD informative large-scale datasets comprised of both real and synthetic data, the result illustrates a reduction in bias by 15–20% and precision accuracy changes by 10–12%, with privacy metric underneath 5% of re-identification. Overall, data utility was retained up to 90–95%. The approach is effective in bias mitigation and ensures privacy. However, the proposed approach is resource-intensive.

The next section analyses all these studies given our hypothesis and research questions.

4. Discussion

Based on our investigation, synthetic data generation provides opportunities to minimize biases in the data by generating artificial samples that are statistically similar to the real data. Therefore, we fail to reject our hypothesis, stating that synthetic data can be used to mitigate bias in medical datasets effectively. The previously formulated research questions are answered below:

RQ1. What are the methods or techniques to generate synthetic data to handle biases in datasets?

- Generative Adversarial Network (GAN): GANs use two neural networks, the generator and the discriminator, which compete to produce realistic synthetic data. This technique compares actual real and synthetic data. The generator creates data, while the discriminator evaluates it against real data, improving the generator's output iteratively. For example, SynSigGAN uses an LSTM as a generator and a CNN as a discriminator to obtain biomedical signal datasets to produce high-quality synthetic biomedical signals such as ECG, EEG, and EMG in time-series data signal data [18].
- Bayesian network: The Bayesian networks use probabilistic models to simulate data with controlled biases. They can model complex dependencies between variables, providing a framework for generating synthetic datasets. For example, the BayesBoost method employs Bayesian networks to generate synthetic datasets to address underrepresented group samples in healthcare data, showing improvements in AUC and ROC curve metrics [17].
- Structural Causal Models (SCM): SCM focuses on understanding causal relationships within data. They remove biased edges in the causal graph to generate fair synthetic data, ensuring that the generated data adheres to fairness criteria like demographics and equal opportunity. For example, DECAF is a framework that uses SCMs to create fair synthetic data, maintain high-quality data utility, and achieve fairness through causally aware generative networks [16].
- Synthetic Minority Over-sampling (SMOTE): SMOTE is an oversampling technique designed to identify data imbalance among the datasets. SMOTE is a kind of data augmentation approach that creates additional data points by interpolating between the minority class samples and one of its k-nearest neighbors. This method helps the model to make the data more balanced, which enhances the minority class's representation and reduces the bias in classification tasks. For example, if a dataset has 100 instances of class A and 10 instances of class B, SMOTE would generate additional synthetic data instances of class B to balance the dataset [28].
- Gaussian Copulas: Copula-based techniques simulate synthetic data by modeling the samples from different populations with similar marginal dependencies structure between variables separately from their marginal distributions. Gaussian copulas use the multivariate uniform distributions method to examine and compare the dependence between variables to represent the relation between various features and mitigate bias among them. This technique is especially helpful for producing high-dimensional data with complex dependencies, such as medical records with multiple variables like age, blood pressure, and cholesterol levels [24].

RQ2. What are the limitations of the existing techniques used for synthetic data generation?

Despite the vast application of synthetic generation models, there are still some limitations that need to be addressed. These limitations include

- Techniques like GANs, CT-GAN, and adversarial training require significant computational resources and high-quality initial datasets with meaningful data, which can be a barrier to the widespread adoption and implementation of this method [18].
- In Bayesian networks, the generation and effectiveness of synthetic data heavily depend on the quality of the original datasets. Poor-quality or biased input data can lead to synthetic data that are also biased [17].

- Structural Causal models require a deep understanding of causal relationships between variables and expert knowledge to be implemented effectively and to understand the definition of specific results. This complexity can lead to significant changes in the inferred causal relationships among data; also, this method focuses on tabular structured data [16].
- Techniques like SMOTE can overlap samples while interpolating between instance variables. SMOTE only focuses on tabular structured data. In high-dimensional datasets, the interpolation process to generate synthetic samples that lack accuracy among instances may lead to bias in the dataset [17].
- Copula-based methods, particularly Gaussian copulas, assume a specific dependency structure between variables captured by the copula remains consistent across the dataset. These methods often require a substantial large amount of data to model the dataset accurately without bias. These methods also require a good understanding of the relationships between variables, and they might go to dependency, which makes it complex to generate synthetic data without bias [24].

5. Conclusions

This article reviewed recent research studies to investigate the application of synthetic data to mitigate bias in healthcare datasets. For this purpose, different databases, including Google Scholar, PubMed, ACM, and IEEEXplore, were explored to find relevant research studies. A total of 17 articles were selected and reviewed, including conferences, journals, and a doctoral dissertation. For each article, the methodologies, dataset, modality, evaluation metric, strengths, and weaknesses were analyzed. The overall findings show that synthetic data generation can enhance the fairness and accuracy of datasets used to train AI models by artificially generating samples representing diverse patient groups. These models, when trained on unbiased datasets, can produce fair outcomes. Different techniques, including Gaussian Copula, SMOTE, the Structural Causal method, and GANs, are discussed in this review. Each data generation method has its own limitations. However, the efficacy of synthetic data is contingent upon the quality of the generation process and the synthetic datasets. Challenges such as ensuring high-quality initial datasets, managing computational complexity, ethical considerations, and privacy concerns need to be addressed to fully realize the potential of synthetic data generation. While the study provides useful insights into the application of synthetic data to mitigate bias, it is essential to highlight certain limitations that may have influenced the outcome. One significant limitation is the selection of databases; we have not included Scopus and Web of Science databases, which limits the scope of our study. These platforms provide a greater selection of research studies that could have enhanced the findings. Therefore, in the future, wider databases will be utilized to provide a more comprehensive analysis. Despite this, the current study lays out a strong foundation for future investigation in the area of artificial data generation to handle biases.

Author Contributions: Conceptualization: M.A.S.H., A.M.Q. and A.K. methodology: M.A.S.H., A.M.Q. and A.K.; writing—review and editing A.M.Q. and A.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was managed by the CREATE-DkIT project, supported by the HEA's TU-Rise programme and co-financed by the Government of Ireland and the European Union through the ERDF Southern, Eastern & Midland Regional Programme 2021–27 and the Northern & Western Regional Programme 2021–27.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Tavares, S.; Ferrara, E. Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. *Sci* **2024**, *6*, 3. [CrossRef]
2. Jain, A.; Brooks, J.R.; Alford, C.C.; Chang, C.S.; Mueller, N.M.; Umscheid, C.A.; Bierman, A.S. Awareness of racial and ethnic bias and potential solutions to address bias with use of health care algorithms. *Proc. JAMA Health Forum. Am. Med. Assoc.* **2023**, *4*, e231197. [CrossRef] [PubMed]
3. Babic, B.; Gerke, S.; Evgeniou, T.; Glenn Cohen, I. Algorithms on Regulatory Lockdown in Medicine. *Science (1979)* **2019**, *366*, 1202–1204. [CrossRef]
4. Kiyasseh, D.; Laca, J.; Haque, T.F.; Miles, B.J.; Wagner, C.; Donoho, D.A.; Anandkumar, A.; Hung, A.J. A Multi-Institutional Study Using Artificial Intelligence to Provide Reliable and Fair Feedback to Surgeons. *Commun. Med.* **2023**, *3*, 42. [CrossRef] [PubMed]
5. Mandal, A.; Leavy, S.; Little, S. Dataset Diversity: Measuring and Mitigating Geographical Bias in Image Search and Retrieval. In Proceedings of the 1st International Workshop on Trustworthy AI for Multimedia Computing, Co-Located with ACM MM 2021, Virtual, 20–24 October 2021; Volume 21, pp. 19–25. [CrossRef]
6. Kordzadeh, N.; Ghasemaghaei, M.; Mikalef, P.; Popovic, A.; Lundström, J.E.; Conboy, K. Algorithmic Bias: Review, Synthesis, and Future Research Directions. *Eur. J. Inf. Syst.* **2022**, *31*, 388–409. [CrossRef]
7. Suresh, H.; Guttag, J. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, Virtually, 5–9 October 2021. [CrossRef]
8. Naresh Mandhala, V.; Bhattacharyya, D.; Midhunchakkaravarthy, D.; Kim, H.-J. Detecting and Mitigating Bias in Data Using Machine Learning with Pre-Training Metrics. *Ingénierie Syst. d'Inf.* **2022**, *27*, 119–125. [CrossRef]
9. Raghunathan, T.E. Synthetic Data. *Annu. Rev. Stat. Appl.* **2021**, *8*, 129–140. [CrossRef]
10. Kandpal, N.; Deng, H.; Roberts, A.; Wallace, E.; Raffel, C. Large Language Models Struggle to Learn Long-Tail Knowledge. In Proceedings of the International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023; pp. 15696–15707.
11. Draghi, B.; Wang, Z.; Myles, P.; Tucker, A. Identifying and Handling Data Bias within Primary Healthcare Data Using Synthetic Data Generators. *Heliyon* **2024**, *10*, e24164. [CrossRef] [PubMed]
12. Oblizanov, A.; Shevskaya, N.; Kazak, A.; Rudenko, M.; Dorofeeva, A. Evaluation Metrics Research for Explainable Artificial Intelligence Global Methods Using Synthetic Data. *Appl. Syst. Innov.* **2023**, *6*, 26. [CrossRef]
13. Bhanot, K.; Bennett, K.P.; Hendler, J.A.; Zaki, M.J.; Guyon, I.; Baldini, I. Synthetic Data Generation and Evaluation for Fairness. Doctoral Dissertation, Rensselaer Polytechnic Institute, Troy, NY, USA, 2023.
14. Gujar, S.; Shah, T.; Honawale, D.; Bhosale, V.; Khan, F.; Verma, D.; Ranjan, R. GenEthos: A Synthetic Data Generation System with Bias Detection and Mitigation. In Proceedings of the International Conference on Computing, Communication, Security and Intelligent Systems, IC3SIS 2022, Kochi, India, 23–25 June 2022. [CrossRef]
15. Sharafutdinov, K.; Fritsch, S.J.; Iravani, M.; Ghalati, P.F.; Saffaran, S.; Bates, D.G.; Hardman, J.G.; Polzin, R.; Mayer, H.; Marx, G.; et al. Computational Simulation of Virtual Patients Reduces Dataset Bias and Improves Machine Learning-Based Detection of ARDS from Noisy Heterogeneous ICU Datasets. *IEEE Open J. Eng. Med. Biol.* **2023**, *5*, 611–620. [CrossRef]
16. Van Breugel, B.; Kyono, T.; Berrevoets, J.; van der Schaar, M. DECAF: Generating Fair Synthetic Data Using Causally-Aware Generative Networks. *Adv. Neural. Inf. Process Syst.* **2021**, *34*, 22221–22233.
17. Draghi, B.; Wang, Z.; Myles, P.; Tucker, A.; Moniz, N.; Branco, P.; Torgo, L.; Japkowicz, N.; Wo, M.; Wang, S. BayesBoost: Identifying and Handling Bias Using Synthetic Data Generators. In Proceedings of the Third International Workshop on Learning with Imbalanced Domains: Theory and Applications, Bilbao, Spain, 17 September 2021; Volume 154.
18. Hazra, D.; Byun, Y.C. SynSigGAN: Generative Adversarial Networks for Synthetic Biomedical Signal Generation. *Biology* **2020**, *9*, 441. [CrossRef] [PubMed]
19. Paladugu, P.S.; Ong, J.; Nelson, N.; Kamran, S.A.; Waisberg, E.; Zaman, N.; Kumar, R.; Dias, R.D.; Lee, A.G.; Tavakkoli, A. Generative Adversarial Networks in Medicine: Important Considerations for This Emerging Innovation in Artificial Intelligence. *Ann. Biomed. Eng.* **2023**, *51*, 2130–2142. [CrossRef]
20. Celi, L.A.; Cellini, J.; Charpignon, M.-L.; Dee, E.C.; Dernoncourt, F.; Eber, R.; Mitchell, W.G.; Moukheiber, L.; Schirmer, J.; Situ, J.; et al. Sources of Bias in Artificial Intelligence That Perpetuate Healthcare Disparities—A Global Review. *PLoS Digit. Health* **2022**, *1*, e0000022. [CrossRef]
21. Fletcher, R.R.; Nakeshimana, A.; Olubeko, O. Addressing Fairness, Bias, and Appropriate Use of Artificial Intelligence and Machine Learning in Global Health. *Front. Artif. Intell.* **2021**, *3*, 561802. [CrossRef]
22. Yogarajan, V.; Dobbie, G.; Leitch, S.; Keegan, T.T.; Bensemann, J.; Witbrock, M.; Asrani, V.; Reith, D. Data and Model Bias in Artificial Intelligence for Healthcare Applications in New Zealand. *Front. Comput. Sci.* **2022**, *4*, 1070493. [CrossRef]
23. Yang, J.; Soltan, A.A.S.; Eyre, D.W.; Clifton, D.A. Algorithmic Fairness and Bias Mitigation for Clinical Machine Learning with Deep Reinforcement Learning. *Nat. Mach. Intell.* **2023**, *5*, 884–894. [CrossRef] [PubMed]
24. Rodriguez-Almeida, A.J.; Fabelo, H.; Ortega, S.; Deniz, A.; Balea-Fernandez, F.J.; Quevedo, E.; Soguero-Ruiz, C.; Wagner, A.M.; Callico, G.M. Synthetic Patient Data Generation and Evaluation in Disease Prediction Using Small and Imbalanced Datasets. *IEEE J. Biomed. Health Inf.* **2023**, *27*, 2670–2680. [CrossRef] [PubMed]
25. Libbi, C.A.; Trienes, J.; Trieschnigg, D.; Seifert, C. Generating Synthetic Training Data for Supervised De-Identification of Electronic Health Records. *Future Internet* **2021**, *13*, 136. [CrossRef]

26. Pettit, R.W.; Fullem, R.; Cheng, C.; Amos, C.I. Artificial Intelligence, Machine Learning, and Deep Learning for Clinical Outcome Prediction. *Emerg. Top. Life Sci.* **2021**, *5*, 729–745. [CrossRef]
27. Baumann, J.; Castelnovo, A.; Cosentini, A.; Crupi, R.; Inverardi, N.; Regoli, D. Bias On Demand: Investigating Bias with a Synthetic Data Generator. In Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23) Demonstrations Track, Macao, China, 19–25 August 2023.
28. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-Sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.