Journal of Information Systems Engineering and Management
2023, 8(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Synthetic Data Revolutionizes Rare Disease Research: How Large Language Models and Generative AI are Overcoming Data Scarcity and Privacy Challenges

Mahesh Kumar Goyal[1] and Rahul Chaturvedi[2]
1 maheshgoyal0718@gmail.com, Google LLC, Texas, United State
2 r.chaturvedi2302@gmail.com, Gilead Sciences, North Carolina, United States

ARTICLE INFO

Received: 19 Aug 2023

Accepted: 10 Oct 2023

ABSTRACT

The scarcity of patient data has been a bottleneck for finding solutions in healthcare, and in the challenging field of rare diseases, this bottleneck is worsening due to the increasing use of data analytics. With medical research anchored problems—limited availability of data and stringent privacy regulations preventing data sharing—synthetic data is becoming a developing solution. Using advanced generative AI models, synthetic data can accurately represent the statistical properties of real-world patient data, while preserving the privacy of patients. Gan and VAEs are among other powerful aids to develop synthetic high quality dataset for a rare disease. By balancing between privacy and utility, these models helped generate data to support research and analytics without compromising patient confidentiality and without reducing analytical performance. Further protection against data breaches and re—identification risks with generated AI can be accomplished by integrating differential privacy with federated learning. Yet caution is maintained regarding bias in generative models, the ethics of using synthetic data in healthcare, as well as the tradeoff between data fidelity and privacy. In this study we discuss the use of generative AI to create synthetic data for rare disease research, potential implications for privacy preserving analytics, ethical dilemmas, and future research. In this area, a scheme is designed to effectively utilize generative AI, resulting in a need for more innovation and interdisciplinarity. Through generative AI, in the coming years, synthetic data creation will transform how data can be shared securely and ethically, and efficiently for rare disease research while accelerating the development of new treatments, ultimately improving patient outcomes.

Keywords: Rare disease, Generative AI, Synthetic data, Rare disease research, Privacy-preserving analytics, Generative adversarial networks (GANs), Variational autoencoders (VAEs)

1. INTRODUCTION

Data explosion for analysis has revolutionized the research of health care, creating possibilities unlike those in the past to understand disease mechanisms to develop new treatments and deliver improved patient care. Because these conditions are rare, fewer than 1 in 2,000 people are affected, research in these diseases is challenged by small patient populations, with limited data access and the requirement for highly specialized expertise. Data drives the effort to better diagnose, predict and manage rare genetic disorders(Schembri, 2019), or uncommon cancers, among many others. But it comes with a troubling dependence on data that has also spawned significant concerns around the privacy and security of sensitive patient information(Ambrose; Basu, 2012). The challenge today is to reconcile comprehensive analytics with the extreme restriction of private requirements that are becoming more popular in the form of regulations like GDPR, HIPAA, and CCPA(W. F. Shah, 2023), in the context of rare diseases, where each data point has extreme value. When you talk about rare diseases, the data captured through various trials and experiments are sensitive data in an

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2023, 8(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

organization that involves different risks, such as breach disclosure, the re-identification of anonymized data, unauthorized access, and the potential for stigmatization. Due to these associated risks, several versions of privacy-preserving mechanisms have lately been invented that guarantee data security with consideration for analytics. (Zhao, n.d.)Existing privacy-enhancing technologies, such as anonymization and encryption are not able to balance data utility and privacy in the context of rare diseases. These might result in reducing the data quality or usability. (Cho, 2023)The process will make it less valuable especially when dealing with already limited datasets. Therefore, there was a gap in sophisticated approaches allowing data sharing without loss of privacy or data utility, particularly for rare disease research. To estimate the prevalence of RD, a study funded by the European Commission and developed by the European Organization for Rare Diseases (Eurordis) and Orphanet (Nicholl, 2014)was carried out. While the study did provide some prevalence data (Fig-1), it also revealed significant challenges in this area, including: uncertainty, incongruities between the sources of information and the small sample size of the population, as well as the poor quality of the methods used in most epidemiological studies. Moreover, the study reported that few centres are equipped for essential biochemical/ genetic investigations.

Disease	Estimated prevalence (per 100000)
Brugada syndrome	50
Erythropoietic protoporphyria	50
Guillain-Barré syndrome	47
Familial melanoma	46
Autism, genetic types	45
Tetralogy of Fallot	45
Scleroderma	42
Great vessels transposition	32·5
Focal dystonia	30
Marfan's syndrome	30
Non-Hodgkin malignant lymphoma	30
Retinitis pigmentosa	27·5
Gelineau's disease	26
Multiple myeloma	26
α1 antitrypsin deficiency	25
Congenital diaphragmatic hernia	25
Juvenile idiopathic arthritis	25
Neurofibromatosis type 1	25
Oesophageal atresia	25
Polycythaemia vera	25

Fig-1

1.1 The Rise of Synthetic Data: A Promising Solution

Synthetic data is, therefore, beginning to offer a good solution to these challenges in the area of rare diseases studies. Even though synthetic data mimics real-world patient data to the extent that the distribution of their variables matches that of actual records, no such records are produced, thus making certain that no personal details are revealed(Gusev et al., 2022). Like this it ensures that no information leakage has occurred and significant conclusions can still be made even when the diseases are rare. The use of Synthetic data is the intermediate position that makes it possible to meet the

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2023, 8(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

privacy regulation while conducting the strong analyses and also involving the rare disease community(Appenzeller et al., 2022). The rationale for this approach is to 'amalgamate' privacy and

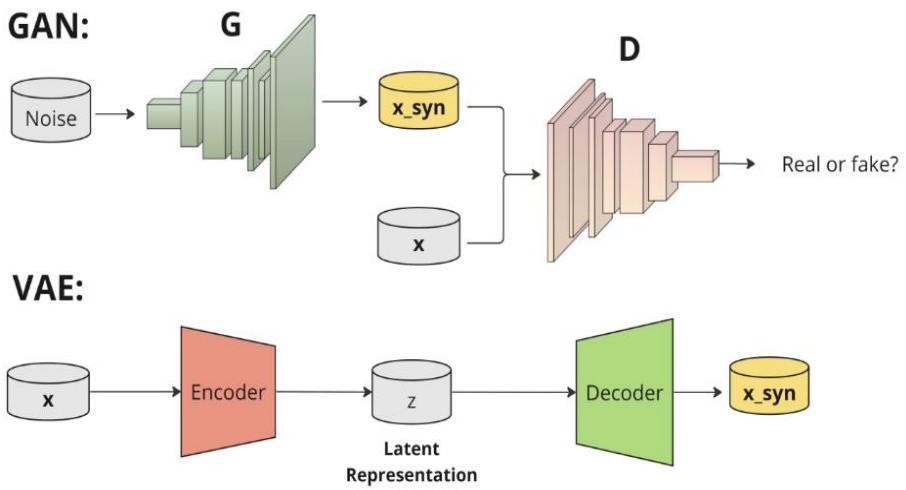

Fig-2

usefulness; and due to that, it has gained prominence in research in recent years, especially for managing the issues of limited datasets in rare disorders.

1.2 Generative AI: The Key to Creating Synthetic Data for Rare Disease

Generative AI is considered one of the first technologies capable of generating synthetic data for emulation of rare diseases. This paper identifies three enabling technologies for creating synthetic data; they include GANs, VAEs, (Bond-Taylor et al., 2022)and diffusion models. Some of these models learn some patterns or relationships for a dataset, and generate artificial data with statistical properties preserved from the original data that are not sensitive. (Appenzeller et al., 2022)This capability represents an extraordinary alternative for generative AI: privacy-preserving applications of analytics without much probability of compromising the data and even re-identification, which is especially valuable when working with low incidence diseases . These generative AI models have the following benefits in contrast to the current approaches to synthesizing data: They can also incorporate complex patterns and distributions within the data to have closer representation of the original data and even capture such detailed patterns of rare disease. Such specificity becomes possible due to the fidelity of the synthesized data that remains useful for a great number of applications, such as modeling, studying diseases, and creating applications. Second, generative AI may contain features such as differential privacy so that it will be more protectionist over some information. That concerns an extra layer of protection where, as per the synthetic data above, it is impossible to deduct one input included in an initial dataset, something crucial to patient trust in rare disease research.(Stockdale et al., 2019) Generative adversarial networks (GANs) employ adversarial training to implicitly model the real data distribution. Variational autoencoders (VAEs) utilize variational inference techniques to approximate the real distribution. Diffusion models take a different approach by gradually adding noise to the input data and then learning to reverse this process to generate new samples. The difference between two methods can be seen in Fig-2 .

1.3 Generative AI and associated challenges

These benefits are, however, associated with challenges when implementing synthetic data and

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2023, 8(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

generative AI for privacy-preserving analyses, particularly for RDFs. One of these is that generative models might be biased-often they reproduce or even amplify biases in the data set, which can result in unfair and often inaccurate representation of specific patient populations. On the ethical front, they raise issues regarding the equity of synthetic data; particularly within scarce diseases, such as rare diseases, and despite Group 1 population heterogeneity likely affecting study results. Another disadvantage is that confidentiality and accuracy are incompatible.

Where generative AI seeks to produce good synthetic data, achieving perfect balance between utility and privacy is sailing in the realm of impossibility. Fully generative models could be necessary, intricate and with a necessity for computationally expensive implementations, which cannot be inspiring for low-resource organizations or groups researching rare diseases .(Burnworth, 2015)

Another important issue regarding the formation of synthetically generated data, especially for rare disease conditions, is the question of ethical concern. As these synthetic datasets grow more realistic, one could use them to build a fake story or skew the truth that you get a wrong perception of the disease occurrence rates, nature, or treatment efficacy. Ethical protection of Synthetic Data can only come from guidelines and oversight to prevent manufacturer's foul play. Subsequently, there are still legal and regulatory perspectives about an organization that uses synthetic data. As mentioned, assuming there are several necessary regulations drafted to limit or control the use of synthetic data, they are still under development more so where relating to rare diseases.

It aims to provide an overview of research in generative AI for synthetic data generation, focusing on the question of whether and how the scope of improving privacy-preserving analyses in rare disease research can be defined. A clear picture of how these technologies can solve thorny privacy problems will emerge based on the strengths and weaknesses of various generative AI models. The following paper outlines several major ethical and practical concerns concerning appropriate synthetic data usage concerning rare diseases.

2. METHODOLOGY

2.1 Understanding Generative AI Models

Generative AI models have emerged as a pervasive foundation in artificial intelligence, ("The Transformation of Photography by Artificial Intelligence Generative AI Technology," 2023) revolutionizing how machines simulate and create data. In this respect, generative models have excited the functionality of the machines, enabling them to learn complex data distributions and generate new data that is realistic, capturing the properties of the original dataset quite closely. Generative AI started its journey with statistical methods, wherein most of the early models relied on frameworks like GMMs and HMMs to capture data distribution. Further, these approaches faced quite a few limitations concerning scalability and complexity. With the advent of methodologies about machine learning and then deep learning, this area has undergone substantial quantum leaps whereby algorithms could analyze and replicate complex patterns, structures, and dependencies within datasets. This evolution set the stage for sophisticated generative models such as GANs and VAEs, which blew the roof off data generation with their novel use of neural network architectures in concert with probabilistic techniques. These models have proven particularly valuable in generating synthetic data for rare diseases, where data scarcity is a significant challenge.

The generative models depend on a dataset's latent distribution learning to develop new and synthetic data points. (Beaulac, 2023a)The process is initiated by training the model with existing data, even if limited, to find and learn a dataset's underlying structures, relationships, and variations, including the subtle characteristics specific to rare diseases. Another essential part is the latent space, where all data goes through intelligent compression and abstraction to meaningful patterns. This is a blueprint for the model to reverse-engineer to reconstruct or generate data. In other words, while decoding the

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

latent space, the model synthesizes data that closely resembles the statistical properties in the actual dataset without leaking any information about the specific content of the instances. Concretely, this disjunction of statistical representation and individual instances is highly desirable in sensitive applications with potential privacy issues, such as in the study of rare diseases (Weyrauch & Rakov, 2013) .

Applications range from the creation of realistic images and videos to the synthesis of data sets for analytics, and now, generating synthetic patient records for rare diseases. These models create the possibility of generating synthetic patient records that maintain patient privacy while supporting research and model training in healthcare, particularly for conditions where real-world data is scarce. This could mean, in rare disease research, the creation of comprehensive datasets that reflect the diverse manifestations and progression patterns of rare diseases, enabling researchers to develop better diagnostic tools and treatment strategies. For instance, generative models in medical research can enable the creation of synthetic survey data and electronic health records in population trend studies while anonymizing it. Generative models have become essential in capturing privacy because they can produce high-quality, synthetic data representative of real-world data without exposure, even for rare and complex conditions. This capability supports compliance with stringent data privacy regulations and fosters innovation by allowing secure and ethical data sharing across industries, accelerating research into rare diseases(Bagade, 2016). As AI constantly improves, generative AI is bound to take on an even more critical role in solving the balance between privacy and utility challenges, particularly in the context of rare disease research.

2.2 How Generative AI Creates Synthetic Data

Synthetic data are therefore datasets that they are artificially created to look like real datasets but they do not contain sensitive information. It is best characterized by its ability to mimic actual datasets for training, testing, trending, and analysis models. Synthetic data provides several advantages over anonymized data (Ebrahimi Atani; Sadeghpour, 2018)because the latter is synthesized from the original data, so the probability of re-identification is almost zero .

With synthetic data, it can be compared and used in different analyses since it is not unique to a particular person and the advantages perhaps most significant in the health sector. It enables analysis that would not reveal raw patient data while it helps researchers record synthetic patient records safely. It helps organisations to meet requirements of the laws, for example, GDPR and HIPAA by providing secure means for sharing data(Demircan, 2022). This is so as to warrant that data utility is preserved as it adheres to the set legalities .

One major strength of synthetic data is that it can overcome the issue of data scarcity in areas where the real data acquisition is problematic. It enables the generation of large datasets of different contexts, which can be particularly valuable when working with subjects such as autonomous driving and healthcare since generating this data is costly, risky, or can be collected only with significant concerns to data privacy.

Synthetic data promotes creativity in a contained environment. For instance, it can be used to teach antifraud programs in financial institutions or target consumers in retail without compromising its data or the consumer's trust. This enables a company to sample new concepts so that they can be contained within the legal and security parameters.

There is for example the Theledon scooter that adds mobility to the disabled. Likewise, synthetic data opens up privacy compliant analysis in multiple verticals. With all these, it can be described as a secure, flexible replacement for real data taking organizations through privacy, regulatory and availability hurdles. Synthetic data not only serves as an answer to today's question but acts as a key that opens the door to tomorrow's dystopian ideas of data management and analytics, (Healthcare

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2023, 8(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Data Analytics and Management, 2019)with a focus on privacy.

2.3 Overview of Generative AI Techniques

Many sophisticated techniques examine real-world data imitation's statistical properties and patterns for generating synthetic data from machines, with particular relevance to creating comprehensive datasets for rare diseases. Among the widely applicable ones, two methods are salient: Generative Adversarial Networks and Variational Auto-encoders(Zhang et al., 2021). These have opened new avenues in data synthesis, making high-quality, realistic synthetic data possible that could be used for a range of analysis and research purposes, including the study of rare diseases. GANs follow the working principle based on some weird dual network architecture, one acting as a generator network and another serving discriminator purposes. The generator generates fake information, and the discriminator evaluates the reality of this phony information by testing it against actual data. This goes into an adversarial process that forms the backbone of GANs. Thus, the generator would iteratively get better at generating artificial data, narrowing the difference with accurate data with every round. At the same time, the discriminator will further develop its skills in finding the difference between them. This adversarial dynamic allows GANs to do exceptionally well in generating realistic and diverse datasets, and they are highly effective for a wide variety of applications, including the synthesis of images and videos and even the augmentation of machine learning datasets, now being applied to create synthetic patient data for rare diseases.

Variational Autoencoders represent another powerful approach to generative AI, using probabilistic models to learn the data distribution and reconstruct them. VAEs work by encoding input data into a latent space and decoding it back into the data domain. Controlled variations are introduced in the decoding phase, generating synthetic data similar to the original but with subtle differences. Unlike GANs, VAEs are designed more with interpretability and continuous data generation, making them more apt for structured data applications such as medical records, including those for rare diseases, financial transactions, and sensor data. (Bylone, 2010)Their capability of encoding and decoding with the least loss of information ensures that synthetic data retains its utility and fidelity for analytics and model training, which is crucial for accurately representing the complexities of rare diseases.

Each of those is a generative AI technique with its respective merits, which enables both the researcher and practitioner to choose based on specific use cases or characteristics of the dataset, particularly important in rare disease research where the data can be both complex and sparse. GANs work best on unstructured data where realism and diversity are critical, such as generating synthetic images for rare skin conditions. VAEs are favored for structured and tabular data when maintaining statistical integrity and interpretability, which is crucial, for instance, generating synthetic electronic health records for patients with rare genetic disorders. This direction opens up new paths toward broader applications due to its bias mitigation and increased fidelity and applicability, each more niche. Yet, it is equally enabled by more extensive and more hybrid architectures. Further development will establish generative AI as a more effective way of using traditional methods and extend to emerging innovations that will keep widening the capabilities of synthetic data generation. Furthermore, advances in the quality and practical value of synthetic datasets will be further complemented by broader challenges around bias reduction and computational efficiency, thus securing a role for generative AI as an essential enabler of privacy-preserving data analytics, especially in the under-resourced field of rare disease research (Appenzeller et al., 2022).

2.4 Generative AI's Role in Ethical and Legal Data Sharing

Generative AI has turned a new leaf in treading the ethical and legal complexities of data sharing today, where privacy and security are vital factors concerning data resources. (Kroll, 2018)This is

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2023, 8(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

particularly true in healthcare, and even more so in the context of rare diseases where patient data is both highly sensitive and incredibly valuable for research. Sharing real-world data creates formidable barriers when considering privacy-related concerns, regulatory restrictions, and ethical considerations for organizations across sectors. It can help fight such issues by enabling the creation of synthetic data possessing the same statistical properties and patterns as accurate data without sensitive or identifiable information. This would allow organizations to share and use data for research, development, and collaboration without compromising privacy, which is crucial for advancing research into rare diseases.

From an ethical perspective, generative AI decreases the chances of misuse or having negative consequences in the physical world for information. For example, in health care, one can provide synthetic data regarding the patients, especially those with rare diseases, (Reimer et al., 2018) to train predictive models or study medical conditions without causing actual harm to patient confidentiality (Sezgin et al., 2023) . That way, necessary research is done without ethical violations related to the exposure of personal health information. Thus, in rare disease research, synthetic data could enable the creation of extensive datasets that reflect the diversity of patient experiences and disease manifestations, fostering collaboration and accelerating the development of new treatments. Because that would be possible, it would also mean organizations have kept their ethical standard high while reaping some benefits from advanced data analytics, particularly in a field where every data point is precious.

On the legislative pole, synthetic data created with the help of generative AI helps organizations follow toughened laws about data privacy, like GDPR or the California Consumer Privacy Act of 201(Dhar, 2021) . All these regulations impose severe requirements on data collection, storage, and sharing regarding personal data, very often hampering using real-world datasets, especially for rare diseases. Generative AI bridges this gap by creating privacy-preserving synthetic data that meets compliance requirements while sacrificing little analytical value. This dual ability to uphold legal compliance with preserving data utility-positions generative AI as cornerstone technology for secure and responsible data sharing, which is essential for building trust within the rare disease community.

2.5 Case Studies: Generative AI in Action-Powered Synthetic Data Creation

Speeding Up Drug Development: Synthetic data coupled with rare disease research is already making big strides especially in drug development(Lacoste, 2018). Just consider a much more exciting example of Gaucher disease – when researchers had 100 patient charts to analyze, it was a mere drop in the bucket. With the help of special artificial intelligence tools called GANs, they trained 900 extra synthetic patient profiles that were indistinguishable from those found in their raw data. It allowed them to notice disease patterns they never knew existed and guided them towards new potential drug areas. The best part? The synthetic data were realistic that it could employ it for hypothesis check and application of ai models on predicted responses of patients to certain treatments.

Enhanced Disease Diagnosis through the use of AI: Developing dependable approaches to classifying rare diseases has even in the past proved difficult but not with synthetic data. A most outstanding example arises from research on Early-Onset Sarcoidosis. They all had the same problem; they lacked enough medical image data to tackle. They applied a tool called VAE to create more objective medical images that can be used in training AI to detect early signs of the ailment. Rather than waiting several years for enough raw data of real patient scans for testing how accurate the diagnostic tools would be, they could use the data to test their tools now, and at the same time protect the privacy and security of the patients.

Understanding of Disease progress : One of the most complex challenges the science of rare diseases faces is determining the disease progression. To my surprise, several researchers

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2023, 8(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

investigating Friedreich's Ataxia (Lobanova, 2018) were smart enough to fabricate patient data that might illustrate the disease in three different phases spanning several decades. This let them examine specific disease variations that might occur in only one out of every ten thousand patients. The sort of data which can otherwise take a researcher an entire lifetime of investigations to amass using conventional research techniques.

Restoring interest in Clinical Trials: For the clinical trial conducted on rare disease (Corrigendum,2024) , there arises some of the ethical issues associated with the experiments. In their study of Fabry disease, researchers made a new breakthrough by using synthetic control groups. To mimic the placebo effect that would have otherwise been obtained had the patient demographics been given actual placebos, they produced artificial data. This meant that fewer of their patients had to be given placebos while they were still able to keep their experiments academic and realistic – a fact, which was quite advantageous to all the participants involved.

Personalizing Treatment plan: A new perspective on how different patients respond to treatments also came from synthetic data treatment. Researchers who developed the disorder of Multiple System Atrophy found new patient subtypes by using fake ones in their studies. (Pellecchia et al., 2020) Out of this they began to understand that patients could be different and might require different treatments – an important realization that helped them enhance the quality of treatment offered to the patients.

3. PRIVACY IN RARE DISEASE DATA

3.1 Privacy Risks in Rare Disease Data Usage

Most have moved into the core of different industries' decision-making with data analytics, and this is particularly true in the field of rare disease research(Ghasemaghaei, 2019), where data can provide critical insights into diagnosis, treatment, and patient care. With sensitivity and personal information in data, especially in the context of rare diseases where even small amounts of information can be identified, the risks colossally grew. General sources of privacy risks usually include potential misuse, unauthorized access, and re-identification, even in anonymized or aggregated forms. Probably one of the most significant risks is the threat of re-identification attacks. Simply put, the attacked anonymized data is matched against other datasets to unmask personal information. Similarly, membership inference attacks can enable adversaries to infer whether a particular individual is part of a dataset, leading to even more privacy compromises. These risks are particularly acute in rare disease research, where the small patient population increases the likelihood of re-identification (Beaulac, 2023a). Such risks are most heightened in domains like health and financial information, where data usually contains private information. The proliferation of sophisticated analytics tools and machine learning algorithms has brought concerns about privacy to a critical level because it might unearth deep insights and correlations that could reveal personal details, especially in the context of rare diseases where specific combinations of symptoms or genetic markers can be highly identified. Therefore, balancing the accruing benefits from data analytics with the need for safeguarding privacy can be a critical issue, effectively demanding robust approaches to preserving privacy, particularly in rare disease research.

3.2 Key Privacy-Preserving Techniques

Researchers and practitioners have discussed different techniques of preserving privacy, wherein the risks above are mitigated, but sensitive information is protected and simultaneously offers data utility, especially crucial in the field of rare diseases.

3.2.1 Differential Privacy

Among various protection methods, differential privacy is one the most effective mechanisms to protect individual-level information during rare disease data analytics, and it is particularly relevant

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2023, 8(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

for rare disease research where each data point is highly valuable. It injects carefully calibrated noise into the datasets or queries for ambiguous identification of the presence or absence of specific individuals. This way, analytics results remain accurate and statistically valid while personal information cannot be disclosed. Differential privacy has emerged from government census data to machine learning because it allows model training without revealing private information, making it suitable for training models on rare disease data(Elder, 2015). In this regard, differential privacy is a cornerstone for modern privacy-preserving analytics: It quantifies these guarantees, providing a strong framework for protecting patient data in rare disease research.

3.2.2 Federated Learning

Another powerful technique for privacy preservation, especially in distributed environments, is federated learning, which is particularly useful for rare disease research that often involves collaboration between multiple institutions. Federated learning allows machine learning models to be trained directly in decentralized devices or local servers without aggregating the raw data in a centralized location(Bonawitz et al., 2021). In this technique, only model updates and not raw data are shared across the network; hence, there is a lot of reduction in the risk of a data breach. That is increasing use in applications such as personalized healthcare; patient data can rest on local devices while still contributing to learning global predictive models, enabling collaborative research on rare diseases without compromising patient privacy. Federated learning enhances privacy while reducing many logistical and regulatory burdens of centralized data collection, making it an ideal approach for multi-institutional rare disease studies .

3.2.3 Contributions of Generative AI to Data Privacy

Generative AI has created analytics that preserves privacy by generating synthetic data, which does not reveal sensitive information and represents total replicas of properties like accurate data, offering a promising solution for rare disease research. Further ahead, these are used for training models, testing, and various forms of research for which actual real-world data might be compromised, particularly important in the context of rare diseases where data is scarce and privacy concerns are paramount. These privacy risks are minimized because the generative AI generates realistic but wholly synthetic data, enabling research on rare diseases without exposing real patient data. Second, generative AI models can also be combined with other techniques for privacy enhancement, such as differential privacy, to improve their security further. Popular generative AI models include GANs and VAEs .

It is designed to explicitly prevent memorizing sensitive information from the original dataset by introducing noise or altering the training process, which is crucial for maintaining the privacy of individuals with rare diseases. Another use case of generative AI is federated learning, which synthesizes data on decentralized devices and trains models securely without aggregating data centrally, facilitating collaborative research on rare diseases across institutions. Generative AI also overcomes data shortages, improving accessibility in areas that may be sensitive regarding privacy, such as in rare disease research. For instance, artificial patient data can train diagnostic algorithms without necessarily exposing accurate patient records in healthcare, enabling the development of new diagnostic tools for rare diseases. In rare disease research, synthetic transaction data and financial flows would permit funding allocation and investment research without exposing individual or institutional confidentiality. Such applications allow generative AI to strike a good balance between privacy and utility, where each enables the other, making it an essential enabler of secure data-sharing practices in an increasingly data-driven world, with the potential to significantly advance research and treatment for rare diseases (Lochmüller et al., 2021).

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2023, 8(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

4. IMPACT OF SYNTHETIC DATA

4.1 Role of Synthetic Data in Research and Analytics

Synthetic data has already become a transformative factor in contemporary research and analytics because it resolves two significant challenges: data accessibility, privacy, and regulatory compliance, and it holds particular promise for the field of rare disease research. In many areas, such as healthcare, finance, and social sciences, access to real-world data may partly be prohibited for research or analytics purposes because of confidentiality and legal compliance concerns(Tarun Kaniganti, 2021). Synthetic data can be substituted for actual data using its statistical properties and patterns without holding sensitive or identifiable information, enabling research on rare diseases that would otherwise be impossible due to data limitations . In this fashion, accurate data need not be compromised for privacy while being studied by researchers and analysts, particularly in the context of rare diseases where patient data is both scarce and highly sensitive. For example, in health, most applications are made to patient records by synthesizing them to build predictive models, conduct drug trials, or build diagnostic tools on synthetic data without exposing patients, offering a new avenue for advancing research into rare diseases. Further, synthetic datasets representing diverse patient populations can be used to test the efficacy and safety of new treatments for rare diseases, potentially accelerating the drug development process. Therefore, synthetic data will not only innovate but also help organizations comply with various regulations on data privacy, such as the GDPR and HIPAA, particularly important for rare diseases that may be subject to additional protections. Besides, synthetic data is crucial in training artificial intelligence and machine learning models, which require more extensive and diverse datasets, and it can fill the gap between unavailable data and the growing demand for quality datasets in rare disease research. This fills the gap between unavailable data and the growing demand for quality datasets so that research and analytics can be empowered in hitherto impossible ways, potentially leading to breakthroughs in the understanding and treatment of rare diseases .

4.2 Data Utility with Privacy

One of the most critical challenges regarding synthetic data is finding the right balance between data utility and privacy: synthetic datasets have to be sufficiently realistic to serve the purpose of analytics, and this is especially important in the context of rare diseases where the nuances of the data can have significant implications for research and treatment. However, they also have to ensure that no sensitive information from the original data is leaked. Generative AI models must become sophisticated enough to learn complex relationships and patterns within datasets, all while keeping the privacy constraints in view, to provide this balance. Noise Injection into Process: Noise is injected into the generation of synthetic data using techniques (Gaborit Jaouen, 2023)like differential privacy so that no specific individual data points may be inferred, protecting the privacy of individuals with rare diseases . That said, too much intensive noise injection in the synthetic data tends to reduce its usefulness analytically. On the contrary, if too little noise is injected into the data, what comes out could be synthetic data for which information related to the underlying data is exposed unduly. Therefore, this trade-off between utility and privacy is one of the biggest challenges demanding continuous improvement or enhancement in the generative AI techniques and mechanisms of respecting the privacy of data subjects, particularly in the sensitive area of rare disease research. Each organization should critically assess the peculiar demands concerning its use cases and ensure that the generated synthetic data meets standards on confidentiality and the need for analyses. To this end, striking a proper balance would open new opportunities for synthetic data regarding their secured and practical usage across various sectors, particularly in rare disease research where the balance between utility and privacy is paramount.

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2023, 8(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

4.3 Ethical Considerations

4.3.1 Bias and Fairness within Generative Models

One of the most important ethical issues connected with synthetic data is the problem of bias and fairness in connection with generative models, particularly in the context of rare diseases. For example, in cases where biased data is used to train an AI generative model, such as race, gender, or social class, the resulting synthetic data may also be biased, even magnified. This would arguably translate into unrealistic or unfair outcomes in decision-making or predictive models once synthetic data comes into play. This is a significant concern in rare disease research, where biases in the data could lead to misdiagnosis, inadequate treatment strategies, or disparities in access to care. Biased health data, for example, will eventually drive resources into synthetic data that retains the differences in diagnosis or medical treatment options for selected demographics, potentially exacerbating existing health disparities. All this could demand careful training data curation and tap into fairness-aware algorithms in data generation, especially for rare diseases. (Rasmussen, 2020)It will be instructive on the part of the researchers and practitioners that, with openness about generative AI systems, any synthetic data generated must be representative and produced equitably, taking into account the diversity within and across rare disease populations.

4.3.2 Synthetic Data Misuse Risks

The use of synthetic data is growing year by year[Fig-3]. While synthetic data has many advantages, it is not entirely bereft of the dark side in its usage. The possibility of generating realistic datasets raises several concerns about unethical usage in creating deceptive or fraudulent datasets with vicious purposes. For example, manipulating synthetic data related to rare diseases may show wrong reports or mislead stakeholders, potentially impacting funding, research priorities, or even public perception of these conditions. Risks are also associated with using synthetic data in adversarial scenarios, including creating pseudo-profiles or simulating particular events that may affect trust and accountability in digital ecosystems. Therefore, organizations must declare unequivocal ethical rules and establish efficient monitoring systems to use synthetic data responsibly, particularly in the sensitive area of rare disease research (Chan et al., 2023). The need can also arise for the intervention of regulatory bodies to spell out standards and best practices regarding the generation and use of synthetic data, especially for rare diseases. Addressing these ethical considerations helps harness the potential of synthetic data while minimizing risks associated with misuse, ensuring that research on rare diseases is conducted ethically and responsibly(Schieppati, 2008).

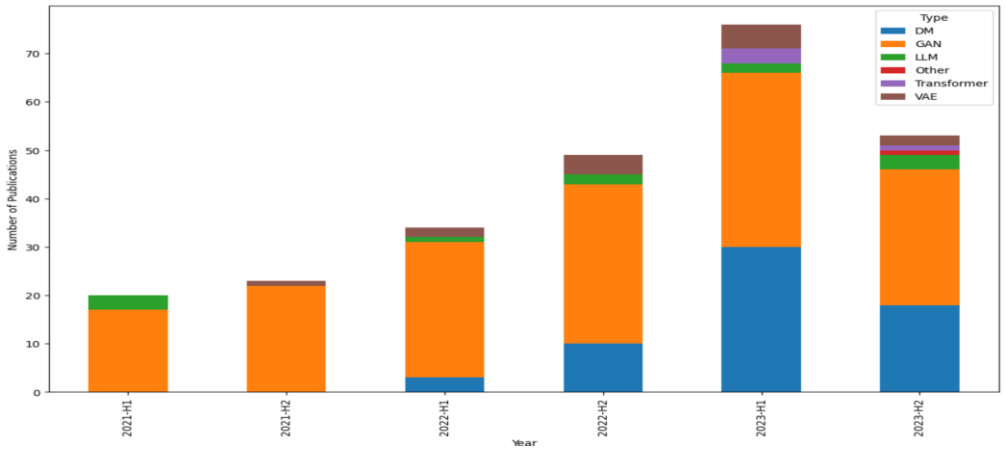

Fig-3

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2023, 8(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

4.4 Challenges and Solutions in Bias Mitigation and Computational Efficiency

4.4.1 Bias Mitigation in Generative AI

A big concern with generative AI is bias(Park, 2023), especially in the context of rare diseases. Most models tend to pick up the bias from the original data and sometimes exaggerate it. Suppose there is some bias in training data where the samples are not very well representative; then, this bias can be reflected in synthetic data, potentially leading to inaccurate or unfair representations of certain rare disease populations. This sets the scene for inequitable or wrong analytics and decision-making. How to address the challenge: The following are strategic ways to overcome these challenges, specifically tailored for rare disease research.

Bias Detection and Quantification: Generative AI systems should include a methodology for detecting and quantifying bias within the training data, with a focus on identifying biases that may be unique to or particularly impactful for rare diseases. The techniques, fairness-aware machine learning algorithms, or exploratory data analysis allows the detection of patterns of underrepresentation and overrepresentation, ensuring that the synthetic data accurately reflects the diversity of the rare disease population.

Fairness-Aware Training: one might reweight or re-sample during the training phase to deal with the imbalance. Models such as FairGAN are explicitly designed to produce synthetic data that minimizes bias while preserving utility, which is crucial for generating representative datasets for rare diseases.

Synthetic data post-processing: Post-processing approaches analyze and change generated synthetic datasets to verify whether they satisfy the predefined criteria concerning fairness. The resulting outcome can be balanced and precise with the involved groups, ensuring that the synthetic data for rare diseases does not perpetuate existing biases.

Diverse and representative training data: One of the best solutions is to ensure the original training data is diverse and representative, hence minimizing the chances of biases while synthesizing the data. This is particularly important for rare diseases, where data from diverse populations may be limited.

Equipped with these measures, organizations can generate synthetic data that is fair and useful, thereby reducing the danger of biased decision-making, especially in the context of rare diseases where accurate and equitable representation is essential.

4.4.2 Computational Efficiency in Generative AI

One of the significant barriers to adopting generative AI, especially models such as GANs and VAEs, is computational requirements, especially for resource-constrained organizations or rare disease research groups that may have limited funding. These demands arise from the high-dimensional data involved, iterative training processes, and the need for substantial memory and processing power.

Solutions to enhance computational efficiency include:

Model Optimization: Improved model architectures such as lightweight GAN and VAE (T., 2018)reduce the computation overhead without compromising performance. Besides, a lighter model may be reached for effective generation with pruning, quantization, and knowledge distillation, making it more feasible to generate synthetic data for rare diseases.

Distributed Computing and Parallel Processing: Distributed computing frameworks allow training generative models over multiple machines. It enables fast computation by parallel processing, thereby reducing the time required for training, which is beneficial for research groups working with complex rare disease data .

AI Cloud-Based Platforms: Cloud platforms, including Google Cloud, AWS and Microsoft Azure, have

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2023, 8(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

made resources scalable for AI workloads(Wankhede et al., 2020). These platforms enable organizations to access high-performance computing power without significant upfront investments in infrastructure, making it more accessible for rare disease researchers to utilize generative AI.

Federated Learning: It helps decrease the necessity of the centralized data processing system, as federated learning makes it possible to perform model training locally on decentralized devices by aggregating the updates. Hence, it minimizes computing demands while preserving privacy, which is particularly well-suited for collaborative rare disease research across multiple institutions.

Adaptive Training Techniques: Adaptive training methods, such as curriculum learning, gradually increase the difficulty of tasks while training. This optimizes the use of computational resources by focusing on more straightforward tasks first so that they can build up to more complicated patterns, which can be beneficial when working with complex and nuanced rare disease data .

With those biases that bias the mitigation or in increasing computational efficiency, improvements regarding fairness and accessibility could ensure such enablement for organizations by optimizing the benefits involved in using generative AI to generate synthetic data. These solutions have made generative AI much more viable by improving its utility and reaching a broader range of users, including those in the under-resourced field of rare disease research.

5. FUTURE DIRECTIONS

5.1 Advancements in Generative AI Models

Generative AI will continue with more sophisticated models, more efficient, synthesizing data with higher fidelity, stronger privacy guarantees, with significant implications for rare disease research. Diffusion models and hybrid architectures will be increasingly important due to its perfect trade-off between realism, diversity, and computational efficiency. For instance, diffusion models provide even more fine-grained control over the data generation process, and one could construct datasets that are not only statistically correct but also highly customizable, which is crucial for accurately representing the complexities of rare diseases. Further integration with other techniques, such as federated learning and differential privacy, continues to enhance the functionality of Generative AI for privacy-preserving data synthesis, enabling more secure and collaborative research on rare diseases (C. V. Shah, 2019).

A different line of research focuses on enhancing the scalability of the generative models, that is, their ability to deal with large, complex datasets without significant losses in performance, which is particularly important for creating comprehensive datasets for rare diseases. Improvements in explainability and interpretability may also be their fundamental features, giving users confidence that they understand what such a model does with any synthetic data created, fostering trust and transparency in rare disease research.

5.2 Wider Ramifications for Data Science and Policy

It is not only the field of data science but also well beyond it, where synthetic data and generative AI are coming of age, with profound implications for rare disease research. Allowing safe, privacy-aware data sharing can increase collaboration among organizations, researchers, and governments, accelerating the pace of discovery and innovation in rare diseases. This effect has enormous implications in domains such as health, where access to truly diverse datasets often poses confidentiality issues that make sharing impossible, (McMillen, 2004)particularly for rare diseases where collaboration is essential. On the other hand, increased integration of artificial information in everyday life also encourages demand to adapt to current legal and regulatory frameworks.

Policymakers must determine ownership, accountability, and ethical usage of synthetic data,

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

particularly in the context of rare diseases. Such as how synthetic data have the same legal protections as actual data. How will regulatory bodies ensure that synthetic data is used responsibly, especially in sensitive areas like rare disease research? Answering such questions will be foundational to creating a healthy ecosystem where generative AI technologies are adopted with ethics in mind, promoting responsible innovation in rare diseases. Ultimately, synthetic data could make access to data more accessible and enable the application of smaller organizations or researchers with restricted resources to high-quality but prohibitively costly or legally encumbering datasets, democratizing rare disease research and potentially leading to breakthroughs in treatment and care.

5.3 Research Opportunities in Privacy-Preserving Data Synthesis

Specific research opportunities exist in privacy-preserving data synthesis, many of which lie around existing challenges and limitations, with particular relevance to rare diseases. One of the most critical areas involves the development of methods that reduce possible bias in generating synthetic data. The generative AI models often inherent biases from the original dataset, which may raise fairness issues when used for decision-making or analysis, especially in rare disease research where biases can have significant consequences. In this respect, detection, quantification, and correction of these biases during training and generation lie at the heart of many researchers' focus, with a particular emphasis on addressing biases that may be unique to or exacerbated in rare diseases.

Another promising avenue is the creation of standardized benchmarks and evaluation metrics for synthetic data, particularly for rare disease applications. More importantly, a well-defined set of criteria on synthetic datasets' quality, utility, and privacy checks would enable more consistent and reliable development across the field. Interdisciplinary research involving AI, cryptography, and data privacy will also lead to the development of new techniques that can help securely scale up synthetic data systems, enabling more extensive and collaborative research on rare diseases. As the demand for privacy-preserving analytics increases, it will ensure that generative AI remains the leading driver for secure and ethical data innovation in these directions, with the potential to revolutionize research and ultimately improve the lives of individuals with rare diseases.

REFERENCES

[1] Ambrose, P., & Basu, C. (2012). Interpreting the Impact of Perceived Privacy and Security Concerns in Patients' Use of Online Health Information Systems. *Journal of Information Privacy and Security*, *8*(1), 38–50. https://doi.org/10.1080/15536548.2012.11082761

[2] Appenzeller, A., Leitner, M., Philipp, P., Krempel, E., & Beyerer, J. (2022). Privacy and Utility of Private Synthetic Data for Medical Data Analyses. *Applied Sciences*, *12*(23), 12320. https://doi.org/10.3390/app122312320

[3] Beaulac, C. (2023). A moment-matching metric for latent variable generative models. *Machine Learning*, *112*(10), 3749–3772. https://doi.org/10.1007/s10994-023-06340-x

[4] Bonawitz, K., Kairouz, P., McMahan, B., & Ramage, D. (2021). Federated Learning and Privacy. *Queue*, *19*(5), 87–114. https://doi.org/10.1145/3494834.3500240

[5] Bond-Taylor, S., Leach, A., Long, Y., & Willcocks, C. G. (2022). Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(11), 7327–7347. https://doi.org/10.1109/tpami.2021.3116668

[6] Burnworth, M. (Mindy) T. (Ed.). (2015). *Rare Diseases e-Resource Guide eReport*. American Society of Health-System Pharmacists. https://doi.org/10.37573/9781585285082

[7] Chan, C.-H., Parker, S., & Pearce, D. A. (2023). The international rare disease research consortium (IRDiRC): making rare disease research efforts more efficient and collaborative around the world. *Rare Disease and Orphan Drugs Journal*, *2*(4), 28. https://doi.org/10.20517/rdodj.2023.23

[8] Cho, J. (2023). Comparative analysis of open government data topics and usability. *Quality & Quantity*, *57*(6), 5655–5671. https://doi.org/10.1007/s11135-023-01630-x

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

[9] Dhar, T. (2021). The California Consumer Privacy Act: The ethos, similarities and differences vis-a-vis the General Data Protection Regulation and the road ahead in light of California Privacy Rights Act. *Journal of Data Protection & Privacy*, *4*(2), 170. https://doi.org/10.69554/glsa8501

[10] Elder, J. H. (2015). A new training program in data analytics & visualization. *Big Data and Information Analytics*, *1*(1). https://doi.org/10.3934/bdia.2016.1.1i

[11] Gaborit, M., & Jaouen, L. (2023). Using data-driven techniques to provide feedback during material characterisation. *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, *265*(5), 2305–2309. https://doi.org/10.3397/in_2022_0330

[12] Kroll, J. A. (2018). Data Science Data Governance [AI Ethics]. *IEEE Security & Privacy*, *16*(6), 61–70. https://doi.org/10.1109/msec.2018.2875329

[13] Lacoste, J. (2018). Research in Rare Disease: From Genomics to Proteomics. *ASSAY and Drug Development Technologies*, *16*(1), 12–14. https://doi.org/10.1089/adt.2017.828

[14] Lobanova, I. (2018). Friedreich's ataxia. *International Physical Medicine & Rehabilitation Journal*, *3*(6). https://doi.org/10.15406/ipmrj.2018.03.00156

[15] McMillen, D. (2004). Privacy, confidentiality, and data sharing: Issues and distinctions. *Government Information Quarterly*, *21*(3), 359–382. https://doi.org/10.1016/j.giq.2004.05.001

[16] Nicholl, C. (2014). Towards a European platform for Rare Diseases Registries. *Orphanet Journal of Rare Diseases*, *9*(Suppl 1), O6. https://doi.org/10.1186/1750-1172-9-s1-o6

[17] Park, Y. J. (2023). How we can create the global agreement on generative AI bias: lessons from climate justice. *AI & SOCIETY*, *39*(4), 2149–2151. https://doi.org/10.1007/s00146-023-01679-0

[18] Rasmussen, K. (2020). Knowing what to do and how to do it: High transparency and careful curation of data and metadata. *IASSIST Quarterly*, *44*(3). https://doi.org/10.29173/iq984

[19] Reimer, A., Bruckner-Tuderman, L., & Ott, H. (2018). Mapping health care of rare diseases: the example of epidermolysis bullosa in Germany. *Orphanet Journal of Rare Diseases*, *13*(1). https://doi.org/10.1186/s13023-018-0944-x

[20] Schembri, F. (2019). Artificial intelligence could diagnose rare disorders using just a photo of a face. *Science*. https://doi.org/10.1126/science.aaw5607

[21] Shah, C. V. (2019). Privacy-Preserving Digital Payments: AI and Big Data Integration for Secure Biometric Authentication. *Global Research and Development Journals*, *4*(12), 1–9. https://doi.org/10.70179/grdjev09i100014

[22] Shah, W. F. (2023). Preserving Privacy and Security: A Comparative Study of Health Data Regulations - GDPR vs. HIPAA. *International Journal for Research in Applied Science and Engineering Technology*, *11*(8), 2189–2199. https://doi.org/10.22214/ijraset.2023.55551

[23] Stockdale, J., Cassell, J., & Ford, E. (2019). "Giving something back": A systematic review and ethical enquiry into public views on the use of patient data for research in the United Kingdom and the Republic of Ireland. *Wellcome Open Research*, *3*, 6. https://doi.org/10.12688/wellcomeopenres.13531.2

[24] T., J. (2018). Comparative Study of GAN and VAE. *International Journal of Computer Applications*, *182*(22), 1–5. https://doi.org/10.5120/ijca2018918039

[25] Tang, Z. (2023). The Transformation of Photography by Artificial Intelligence Generative AI Technology. *Journal of Artificial Intelligence Practice*, *6*(8). https://doi.org/10.23977/jaip.2023.060809

[26] Wankhede, P., Talati, M., & Chinchamalatpure, R. (2020). COMPARATIVE STUDY OF CLOUD PLATFORMS -MICROSOFT AZURE, GOOGLE CLOUD PLATFORM AND AMAZON EC2. *Journal of Research in Engineering and Applied Sciences*, *05*(02), 60–64. https://doi.org/10.46565/jreas.2020.v05i02.004

[27] Weyrauch, M., & Rakov, M. V. (2013). Efficient MPS Algorithm for Periodic Boundary Conditions and Applications Section: Solid matter:https://doi.org/10.15407/ujpe58.07.0657

[28] Zhao, Y. (n.d.). *Privacy-preserving data analytics* [Doctoral dissertation, Nanyang Technological University]. https://doi.org/10.32657/10356/160032

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2023, 8(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

[29] Schieppati, Arrigo, et al. "Why rare diseases are an important medical and social issue." *The Lancet* 371.9629 (2008): 2039-2041.

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.