MedWGAN Based Synthetic Dataset Generation for Uveitis Pathology

Heithem Sliman, Imen Megdiche, Loay Ajramy, Adel Taweel, Sami Yangui, Aida Drira, Elyes Lamine

Intelligent Systems with Applications 18 (2023) 200223

Contents lists available at ScienceDirect

Intelligent Systems with Applications

journal homepage: www.journals.elsevier.com/intelligent-systems-with-applications

MedWGAN based synthetic dataset generation for Uveitis pathology

Heithem Sliman [a,*], Imen Megdiche [b,a,**], Loay Alajramy [c], Adel Taweel [c], Sami Yangui [d], Aida Drira [e], Elyes Lamine [f,a]

[a] Institut National Universitaire Champollion, ISIS, Université de Toulouse, Castres, France
[b] IRIT, Institut de Recherche en Informatique de Toulouse, Université de Toulouse, France
[c] College of Engineering and Technology, Birzeit University, Palestine
[d] LAAS-CNRS, Université de Toulouse, INSA Toulouse, 31400, France
[e] CHU de Nice, France
[f] Centre Génie Industriel, IMT Mines Albi, Université de Toulouse, Albi, France

ARTICLE INFO

Keywords:
Synthetic data [1,2]
Rare disease
GAN
Data augmentation
Open dataset
Uveitis

ABSTRACT

Clinical decision support based on artificial intelligence (AI) methods has increasingly been employed in medical applications to support medical diagnosis. Developing efficient AI methods, however, depends necessarily on the availability of sufficiently large amount of data to provide reliable results. But, in medicine, it is not always possible to find sufficient amount of real data on all pathologies, particularly, for rare diseases. This paper proposes a methodological framework for generating synthetic data using data augmentation techniques combined with epidemiological profiles. It focuses on Uveitis, a rare disease in ophthalmology, which is difficult to diagnose because of the disparity in prevalence of its etiologies. The generated synthetic data have been qualitatively validated by specialist ophthalmologists and quantitatively tested using machine learning methods. Results show that, of a randomly selected sample of the generated data, more than 55% were assessed as good or excellent, which is very promising for generating synthetic, validated as near-real, medical data for rare diseases. They also show that the proposed framework is consistent in generating synthetic data, for Uveitis pathology, of different dataset sizes, achieving more than 80% diagnosis prediction accuracy for 2000 patient records or larger.

1. Introduction

Artificial Intelligence (AI) has undergone considerable development in recent years in the field of medicine and its promising applications in several areas of medical diagnosis. However, the development of AI approaches, to provide reliable results, require the availability of sufficiently large amount of real data. Unfortunately in medicine, it is not always possible to provide so much data on all pathologies. This problem is particularly true for rare diseases. This paper focuses on Uveitis, a rare disease in ophthalmology. Uveitis corresponds to the inflammation of the intermediate tunic of the eye called uvea, as shown in Fig. 1, which is composed of the choroid extended anteriorly by the ciliary body and by the iris. Inflammatory damage to the retina, secondary to primary inflammatory damage to the uvea, is considered to be a full fledged uveitis (Haute, 2020).

Uveitis is located at the crossroads of several medical specialties and represents a real diagnostic and therapeutic challenge. It may belong to the manifestations of a general disease or may affect only the eye. Causes of Uveitis can be a combination of multiple and diverse etiologies, including purely ophthalmological diseases, infectious diseases, systemic diseases, and even drug causes. Sève et al. (2018) describe sixty possible etiologies and classify them into 5 groups, of unequal importance, which denotes the challenges of Uveitis representation.

* Corresponding author.
** Principal corresponding author.
E-mail addresses: heithem.sliman@univ-jfc.fr (H. Sliman), imen.megdiche@univ-jfc.fr (I. Megdiche), lalajramy@birzeit.edu (L. Alajramy), ataweel@birzeit.edu (A. Taweel), yangui@laas.fr (S. Yangui), aidadrira@gmail.com (A. Drira), elyes.lamine@univ-jfc.fr (E. Lamine).

https://doi.org/10.1016/j.iswa.2023.200223
Received 21 July 2022; Received in revised form 9 January 2023; Accepted 30 March 2023
Available online 12 April 2023

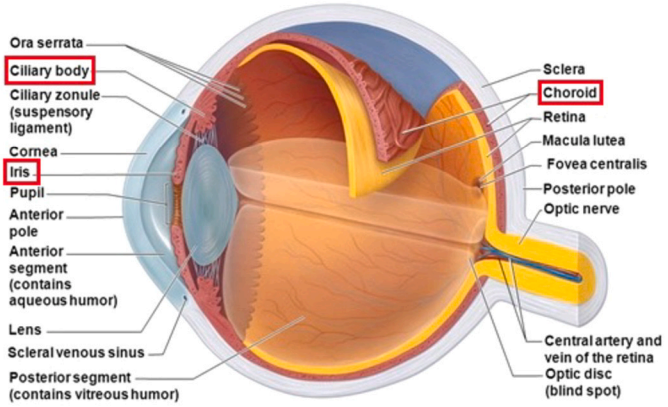


Fig. 1. Sagittal section of the human eye.

Uveitis mainly affects young adults, with 70 to 90% of patients between the ages of 20 and 60, and are responsible for 5% of legal blindness, thus ranking third in the causes of blindness worldwide (Bonnet & Brézin, 2020). Affecting mainly professionally active people, Uveitis represents a major public health problem with medico-economic consequences (Perez-Roustit, 2018). It is a relatively rare pathology, reported by studies, with an incidence of 7 to 52/100,000 people per year, and a prevalence of 38 to 284/100,000 people per year (Sève et al., 2018, Bertrand et al., 2019, Prete et al., 2016). The incidence of uveitis is estimated in the countries of the northern hemisphere at just over 50 cases per 100,000 inhabitants per year and their prevalence is at just over 100 cases per 100,000 inhabitants (Brézin, 2012). In France, although an old study, carried out in the department of "Savoie", estimated the annual incidence of Uveitis at 17 per 100,000 people per year (Vadot, 1992). However, we believe disease prevalence is similar to recent studies that report prevalence of 54 (González et al., 2018) and 60.6 per 100,000 persons (Zhang et al., 2020). Thus the need to found a dataset, for the scientific community, to improve the potential of its diagnosis.

Uveitis causal epidemiology varies according to genetic factors, environmental factors, disease definition, certain ophthalmological entities, paraclinical investigations and so forth. These cause wide heterogeneity of the disease (Bonnet & Brézin, 2020), which complicates its data model representation and generation of realistic data. These challenges and the low number of cases of Uveitis as well as the multidisciplinary management of the disease have prompted doctors to look for tools to help diagnose these pathologies, in order to shorten the delays in establishing etiological diagnosis. To improve clinical diagnosis, clinical decision-making is usually aided by Clinical Decision Support Systems (CDSS), which can utilize a knowledge-based or AI-based approach to derive its decisions. Knowledge-based (also referred to expert) systems provide decisions based on rules built using interventions of experts, while AI-based systems employ AI algorithms and techniques, e.g. machine learning, to provide their decisions using medical data.

Several CDSS have been developed to help diagnose Uveitis, since the 1990s, including the 3D shell expert system (Wiehler et al., 2006), Bayesian network for the differential diagnosis of anterior uveitis (González-López et al., 2016), and Uvemaster (Gegundez-Fernandez et al., 2017). In these systems, however, the larger the number of criteria in the rules, the more complicated it will be to implement as a system. Although these systems show modest results, in terms of accuracy, but, for example, Bayesian networks drop in performance as the number of criteria increases (Jamilloux et al., 2021), hence the increased recent interest in using AI-based or Machine Learning (ML) approaches. These approaches depend on algorithmic models, which require data to train. However, to achieve good performance, these models necessarily require large amount of medical data that represent patients, who have been diagnosed and followed for Uveitis, for example. Obtaining such data, however, can face extreme difficulties due to, on the one hand, medical data protection policies, but also, on the other hand, due to the unavailability or insufficiency of the data due to the rarity of the disease.

To address this issue, several potential solutions have been developed, such as distributed privacy preserving data mining (Scardapane et al., 2018, Ding & Sato, 2020), federated machine learning models (Yang et al., 2019), which use federated training as a means to avoid data sharing, and data anonymization techniques for privacy preserving data publishing (PPDP) (Majeed & Lee, 2020) that could allow for data pooling. Anonymization techniques aim to strike a balance in the final published data between disclosure risk and data utility, resulting in a modified version of the original data that no longer identify individual medical cases, yet the data remains vulnerable to disclosure (Hernandez et al., 2022). A potential solution, to overcome these limitations, is the generation of fully synthetic data (SD) as an alternative to real data, which is commonly known as synthetic data generation (SDG). SD is generated from a model that fits to the characteristics of a real data set. This model contains no data from the original set, but it is able to generate data similar to the original data. Although several SDG methods have been proposed, with promising results, in various application domains, such as healthcare, biometrics, and energy consumption, there is a need for more robust solutions (Hernandez et al., 2022).

The key contribution of this paper is the development of a synthetic data generation framework for Uveitis. A data generation model has been developed, based on the medWGAN approach (Baowaly et al., 2019). It uses data augmentation techniques combined with the aggregated Uveitis epidemiological profile characteristics, taking into account the distribution and the imbalance in the rarity of some Uveitis etiologies. The developed framework has been evaluated, both qualitatively, by physicians, and quantitatively, by statistical analysis and machine learning methods. Evaluation shows very promising results, with more than 55% of the generated data assessed as good or excellent, which is very satisfactory for synthetic data generation for rare diseases. It also shows the proposed approach is consistent in generating synthetic data, of different sizes, that meets Uveitis pathology and the distribution of its etiologies and profile characteristics, achieving more than 80% diagnosis prediction accuracy for datasets of 2000 records or larger. Although the developed framework is for Uveitis, however the followed approach is generic, and can be employed for other rare disease, and the number of generated synthetic samples can be increased according to the need of the employed AI algorithm.

The rest of paper is organized as follows: section 2 presents related work on synthetic data generation, section 3 describes epidemiological profile of Uveitis and the developed synthetic data generation method. In section 4 reports the evaluation of the developed method and discusses the results. Finally, section 5 concludes the work.

2. Related work

Synthetic Data (SD) are data created by a model that has been trained or built to replicate real data (RD) based on its distributions (i.e., shape and variance) and structure (i.e., correlations among attributes) (Hoptroff & El Emam, 2019). It can be employed for various applications, for example, data augmentation, which is used to balance datasets or supplement existing data before training a machine learning model, or privacy preservation, which is used to allow secure and private sharing of sensitive data. SDG has been studied in healthcare for a variety of modalities, including biological signals (Hernandez-Matamoros et al., 2020), medical pictures (Han et al., 2018), free-text content in electronic health records (EHR) (Guan et al., 2018), time series smart-home activity data (Dahmen & Cook, 2019), and EHR tabular data (Yale et al., 2020), which we focus on in this paper.

Synthetic tabular data generation approaches can be divided into three categories (Hernandez et al., 2022): (1) application-oriented approaches, (2) classical approaches, and (3) deep learning approaches. Application-oriented approaches include personalized methods, tech-

Table 1
Comparative table of GAN based approaches tested on binary data (Hernandez et al., 2022).

Publication	Data type	Num. of records	GAN based methods
Choi 2017 (Choi et al., 2017a)	Binary	1071	medGAN
Baowaly 2019 (Baowaly et al., 2018)	Binary	42214	medWGAN
Dash 2020 (Dash et al., 2020b)	Binary	-	healthGAN
Rashidian 2020 (Rashidian et al., 2020)	Binary	47412	SMOOTH-GAN
Yoon 2020 (Yoon et al., 2020)	Binary	26854	DP-GAN

niques, or frameworks that are developed to generate synthetic data for specific applications. Content Modeling for Synthetic E-Health Records (CoMSER) (McLachlan et al., 2016), Aten Framework (McLachlan et al., 2019), SynSys (Dahmen & Cook, 2019), Synthea (Walonoski et al., 2017), and Prophet (Hyun et al., 2020), are examples of these approaches. Although, these approaches have generally shown good results for the application that were specifically developed for, however, they would require major changes to modify for other applications. Additionally, to develop, these methods require significant effort and time including close collaboration of specialist physicians.

Classical approaches include baseline methods, statistical and probabilistic models, and ML models. Baseline methods, which are often used for anonymization, include techniques that simply replace values, delete sensitive attributes and add noise to the data (Nguyen, 2014). Statistical and probabilistic models synthesize data, using statistical and probabilistic techniques, that attempt to simulate real data (Tucker et al., 2020). ML models, in particular supervised ML models, however, generate data based on learned data patterns. (Rankin et al., 2020). These approaches have shown weakness in generating high quality tabular data that guarantees the privacy of the original data, as they frequently attempt to memorize real data and the correlations between attributes. However, they have, commonly, served as a benchmark to evaluate more advanced technologies (Hernandez et al., 2022).

Deep Learning (DL) approaches, on the other hand, include autoencoders, GANs and Ensembles. Autoencoders are unsupervised neural network that learn how to reconstruct data given an encoded representation of the real data (Sewak et al., 2020). Whereas GANs consists of two antagonistic neural networks: generator and discriminator, which learn to generate high quality SD by an adversarial training process (Gui et al., 2020). Ensemble methods, however, employ two different types of DL models to generate synthetic data (Dash et al., 2020a). These approaches have shown better performance in learning real data patterns and in generating more diverse data, thus were able to generate higher quality and better privacy preserving tabular data. This led to, a substantial rise in their popularity in, their use for synthetic tabular data generation and privacy preservation of real data, particularly GAN-based approaches, after their inception in 2014 (Hernandez et al., 2022, Goodfellow et al., 2014). They are considered one of the most interesting developments in AI in recent years, and have shown good results for creating synthetic data (Kavakli-Thorne et al., 2021), outperforming other approaches (Hernandez et al., 2022), particularly for binary data, thus the focus of this paper.

The GAN-based methods with the best performance are shown in Table 1, as reported by (Hernandez et al., 2022). Using a defined evaluation metric that measures resemblance between real and synthetic data, the authors report that healthGAN (Dash et al., 2020b) presented an excellent level of resemblance, med-WGAN (Baowaly et al., 2018) and SMOOTH-GAN (Rashidian et al., 2020) presented a good level of resemblance, while the resemblance for medGAN (Choi et al., 2017a) and DP-GAN (Yoon et al., 2020) was poor. However, once studied closely, healthGAN (Dash et al., 2020b) is in fact a WGAN model originally developed by (Arjovsky et al., 2017), which generally facilitates stable training but generates low quality samples or fails to converge in some settings due to the use of the weight-clipping technique (Baowaly et al., 2018). While SMOOTH-GAN, a novel model proposed by (Rashidian et al., 2020), is a conditional GAN based on WGAN-GP, adapted for healthcare data. It used gradient penalty instead of weight clipping, that resulted into a better performance than the standard WGAN (Baowaly et al., 2018). medWGAN, proposed by (Baowaly et al., 2018), is based on the medGAN (Choi et al., 2017a), however it used the WGAN-GP architecture and added autoencoder to its architecture and used the minibatch averaging technique. Utilizing the advantages of two of the most relevant models, i.e., medGAN and WGAN-GP, resulted into a significantly improved model performance. Hence, the use of medWGAN model in this work.

3. Materials and methods

3.1. Materials

It is essential to consider the epidemiology of Uveitis because the diagnostic approach will be oriented towards the search for the most frequent etiologies in the population studied, which will have important consequences on the quality of the therapeutic management. The causal epidemiology varies according to genetic factors (e.g., HLA-B27 antigen), environmental factors (e.g., outbreaks of tuberculosis), the definition of the disease (i.e. sarcoidosis), the inclusion of certain ophthalmological entities in the group of idiopathic uveitis or ophthalmological entities (i.e. pars planitis), paraclinical investigations carried out (i.e. nuclear imaging) and method of patient recruitment (i.e. tertiary centers). These account for the great heterogeneity of the disease reported in the literature (Bonnet & Brézin, 2020).

It is interesting to note that epidemiology of Uveitis changes over time in the same geographical region. Thus, in Japan, Behcet's disease, which was the first cause of Uveitis 30 years ago, now occupies sixth place, behind sarcoidosis and Vogt-Koyanagi-Harada (VKH) disease. The place of epigenetic factors is also better identified. Indeed, the risk of Uveitis associated with Behcet's disease is very high in Turkey. Surprisingly, the incidence of the disease in patients of Turkish origin migrating to Germany quickly reaches that of the German population (Scardapane et al., 2018). Thus the proposed method aims to generate a realistic dataset that could represent French patients, both treated and followed up for Uveitis, and respect the epidemiological characteristics already mentioned. It needs to draw up an epidemiological profile of Uveitis extracted from the most recent French descriptive studies. To achieve, we identified three relevant retrospective studies, on French patients, that were followed with Uveitis:

1. The first study was conducted on 121 patients treated for Uveitis in the ophthalmology department of the Croix-Rousse hospital in Lyon (Nguyen et al., 2011), from January 2002 to December 2006. Uveitis associated with the virus human immunodeficiency and post-traumatic or post-surgical endophthalmitis were excluded from the cohort. Those lost to follow-up during the 4-year of the assessment were also excluded.
2. The second study is a retrospective epidemiological study of 690 patients with a diagnosis of Uveitis, examined for the first time at the ophthalmology consultation at the Nancy regional university hospital center (Neiter et al., 2019) between January 2005 and December 2016. The patient were referred to the Regional Competence Center dedicated to systemic and autoimmune diseases for diagnostic and/or therapeutic management. The non-inclusion criteria were as follows: patients aged fewer than 18, patients with

Table 2
Extract from the resulting epidemiological profile of Uveitis.

Etiology	Percentage
idiopathic	42.521%
HLA-B27 / AS	18.181%
sarcoidosis	6.657%
Multifocal choroiditis	5.962%
Toxoplasmosis	4.888%
HSV	4.28%

a first episode of acute anterior uveitis responding well to topical treatment, patients for whom the etiological diagnosis could be made by the ophthalmologist after clinical examination without the need for initiation of systemic treatment.

3. The third study included 960 patients aged at least 18 years treated at the specialized Uveitis consultation of the Montpellier University Hospital (Perez-Roustit, 2018) between January 2003 and August 2018.

After discussion with the specialist physicians in ophthalmology, the following points were considered to generate profiles that will be used for synthetic data generation:

- Our reference profile is constructed based on the three studies described above. For each etiology, we calculated an average prevalence weighted by the size of each of the three populations following equation (1).
- All AS-type Uveitis are included in HLA-B27 uveitis, then all AS (Ankylosing spondylitis) and HLA-B27 patients are compiled under the same etiology, which we called HLA-B27/AS.
- To represent the clinical examination results for each of the identified etiologies, we based their representation, as features or columns, on the recent work by "The Standardization of Uveitis Nomenclature (SUN) Working Group", published in 2021 (The Standardization of Uveitis Nomenclature , SUN). We used the clinical description of 15 etiologies as defined by the SUN working group: HLA-B27/AS, Sarcoidosis, Multifocal choroiditis, serpiginous choroiditis, Toxoplasmosis, Herpes simplex virus (HSV), Fuchs, Birdshot, Behcet, Syphilis, Varicella zoster virus (VZV), VKH, Tuberculosis, Tubulointerstitial Nephritis and Uveitis Syndrome (TINU), and Multiple sclerosis (MS). All remaining etiologies were included in the group of the idiopathic uveitis.

$$PercEtiX_{global} = (PercEtiX_{Lyon} * 121 + PercEtiX_{Nancy} * 690 + PercEtiX_{Mont} * 960)/(121 + 690 + 960) \quad (1)$$

This resulted into the profile detailed in Table B.1 (appendix), containing all the considered etiologies, those mentioned by these three studies and taking into account the recommendations of our specialist physicians. Table 2 shows the first six etiologies from the total of 43 etiologies.

3.2. Method

In this section, we describe our methodology to generate a Synthetic Dataset for Uveitis pathology. To achieve, specialist physicians, in ophthalmology, were involved in profiling and characterizing Uveitis, based on the medical data collected from the above studies, and in validating the generated dataset. The generated and validated dataset is made available on this link.[1]

The dataset is made available for the scientific community to accelerate research and innovation on the diagnosis of Uveitis using advanced AI techniques. This dataset contains synthetic data for patients treated for Uveitis, based on the epidemiological profile, as described in Table 3.

[1] https://github.com/heithemsliman/uveitis_dataset_generation.git.

Table 3
Values distribution in original and generated datasets.

		Original dataset	Generated dataset
Age	Age < 30	25%	25.50%
	30 ≤Age < 60	62%	60.9%
	Age > 60	12%	12.75%
Gender	Female	52.5%	54.2%
	Male	47.5%	45.8%
Uveitis class	granulomatous	28%	29.35%
	non-granulomatous	72%	70.65%
Duration	chronic	40.5%	47.55%
	acute	33.5%	33.95%
	recurrent	21%	16.1%
	undetermined	5%	2.4%
Laterality	unilateral	54%	54.35%
	bilateral	36.5%	37.9%
	Alternating	9.5%	7.75%

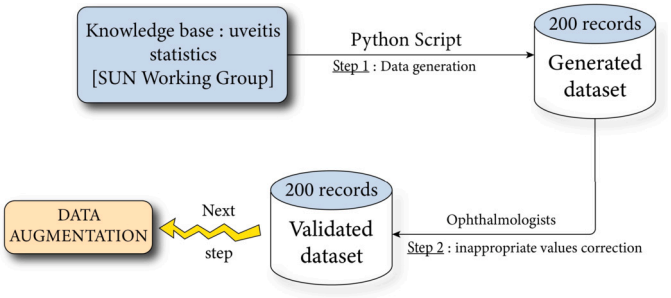


Fig. 2. Synthetic data generation methodology.

3.2.1. Data generation protocol

To generate new synthetic dataset, we undertook the following steps:

1. **Original or base dataset creation:** Since there is no real dataset of Uveitis patients available, we opted to generate an initial or base (or original) realistic dataset, which will serve as a training base for the data augmentation model (Fig. 2). To generate this dataset, we used two elements: (1) the epidemiological profile of Uveitis, partly described in Table 2, is used to determine the number of patients by etiology, and (2) the description provided by the SUN Working Group (The Standardization of Uveitis Nomenclature , SUN), to describe the results of the clinical examination within each uveitis etiology. Our generated dataset includes 200 lines, each line describes the result of the clinical examination of a patient followed for uveitis, presenting with the associated etiology.
2. **First expert validation:** The three ophthalmologists participating in this work have validated the base dataset of 200 patients, which was generated, using Python script, taking into account the clinical characteristics of each etiology. 200 patient records were deemed sufficient by medical experts for manual thorough examination as a realistic representation. Each expert has examined the document, line by line, including the values generated for each clinical observation, as described in Fig. A.1 (Appendix A). At the end of this step, we obtained a dataset of 200 synthetic patients whose etiological diagnosis is labeled by experienced doctors.
3. **Data augmentation:** In this step, we used the medWGAN model to generate a new dataset of 2000 patients, based on the learned rules from the base dataset, already validated by ophthalmologists. This enables to generate the desired number of records.

4. **Second expert validation:** The specialist physicians selected randomly 200 synthetic patients from the 2000 generated dataset. This represents 10% of the dataset, deemed to be representative, for manual validation. The objective is to assess whether a random sample of synthetic data is realistic.
5. **Data generation model validation:** To validate consistency and scalability of the model data generation, medWGAN was used to generate datasets with different sizes, ranging from 1000 to 10000 records, based on the same Uveitis etiology profiling. The generated datasets were used to assess the accuracy of diagnosis prediction of Uveitis etiologies using six different ML models. This assesses the consistency of data generation across different dataset sizes.

3.2.2. Base dataset creation algorithm

To generate our base or original realistic dataset, we initially created a new dataframe with one column for etiologies and 27 columns for examination results and other disease features. We developed an algorithm that automate the synthetic data generation. The developed algorithm is detailed in **Algorithm 1**, described below.

First, it defines the size of the dataset, which was 200 rows (line 2); this size allowed us to distribute the different etiologies on the etiology column, each according to its corresponding frequency as per the epidemiological profile already defined (lines 3 to 5) of Algorithm 1. Secondly, it processes etiologies; for all rows of each etiology, it goes through the columns, and fills each column with the corresponding values according to their frequencies, as described in the knowledge base (lines 6 to 9), while taking into account the conditions and relationships between some features or columns, as defined by the specialist physicians.

The above created the base or original dataset, which was examined and validated by the ophthalmologists for correctness.

Algorithm 1: Base dataset generation algorithm.

```
VARIABLES : df : DATAFRAME
            len_df : INTEGER
            Etiology : COLUMN
            etiology : STRING
            etiology_records : LIST
            etiology_frequency : FLOAT
            unique_value : STRING
            value_list : LIST
            value_frequency : FLOAT
INPUT     : empty dataframe with named columns
OUTPUT    : generated dataframe
1  begin
2      len_df ← 200
3      Foreach etiology do
4          etiology_records ← etiology × etiology_frequency ×len_df
5          Etiology ← etiology_records
6      Foreach etiology do
7          Foreach column ≠Etiology do
8              Foreach unique_value do
9                  value_list ←unique_value ×value_frequency ×etiology_records
10     Return df
11 end
```

3.2.3. Data augmentation algorithm

medWGAN, a GAN-based model, is used to generate the synthetic data. The original GAN is made up of two parts: a generator (G) that tries to generate realistic, but fake data, and a discriminator (D) that tries to discern the difference between the generated fake data and the real data. The generator can learn the distribution of real samples by playing an adversarial game against the discriminator if both the generator and the discriminator are sufficiently expressive (Choi et al., 2017b). The data augmentation model medWGAN is an improved version of medGAN, proposed by Choi et al. (2017a). MedGAN uses a combination of an autoencoder (Enc + Dec) and an adversarial framework to learn the distribution of discrete features, such as diagnosis. The autoencoder aids the original GAN in learning the distribution of multi-label discrete variables, in this Case 3.

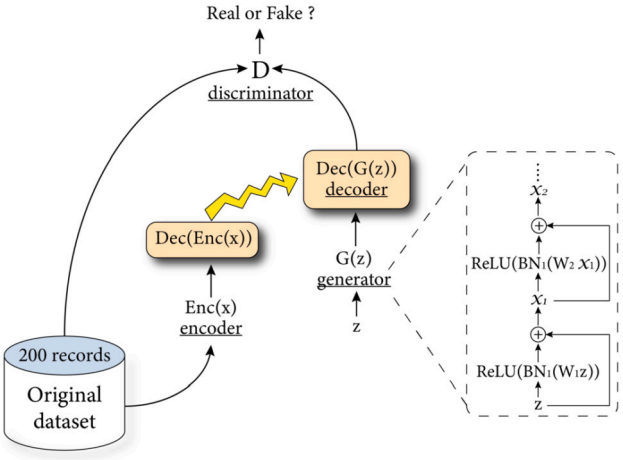


Fig. 3. The architecture of the medGAN Algorithm (Choi et al., 2017a).

Architecture of medGAN: the discrete x comes from the source data (original dataset), z is the random prior for the generator G; G is a feed-forward network with shortcut connections, as shown in the right hand side of Fig. 3; An auto-encoder (Enc and Dec) is learned from x; The same decoder Dec is used after the generator G to construct the discrete output. The discriminator D tries to differentiate real input x and discrete synthetic output Dec(G(z)).

As a contribution, medWGAN have used an improved generative network called WGAN-GP (Wasserstein GAN with gradient penalty) instead of the general GAN. It uses the same structure as that of medGAN shown in 3 (Baowaly et al., 2018). However, in medWGAN the loss function of the original GAN, which measures JS (Jensen–Shannon: a measure of similarity between two probabilities) divergence between the distributions of real and generated data, is replaced by Wasserstein Distance (Weng, 2019). Using Wasserstein Distance, which is a measure of the distance between two probability distributions, produced a much smoother value space, thus improved generated data (Weng, 2019).

4. Evaluation and results

The proposed method is evaluated in two different ways. In the first, a sample of 2000 synthetic data generated by the proposed method was evaluated and examined manually by medical experts on their correctness and validity. In the second, datasets with different sizes, generated by the proposed model, were tested using various machine learning methods on their diagnosis prediction accuracy to assess its consistency and scalability. In both ways, the aim is to evaluate the distribution of etiologies and features in the base or original dataset compared to those in the generated datasets, and their scalability consistency of generating synthetic data with different dataset sizes.

4.1. Expert evaluation of model generation

For the first generated 2000 synthetic records by the proposed model, we noticed that the error rate is higher within the class of granulomatous uveitis. This is due to the unbalanced nature of our base or original dataset, given that granulomatous uveitis accounted for a quarter of the dataset compared to non-granulomatous uveitis, which represents the remaining three quarters. To address, to balance the data, new records of granulomatous uveitis were added to the base or original dataset before training the model. These records were generated by the medWGAN network and validated by the specialist physicians, and concatenated to the base or original dataset. This created a balanced

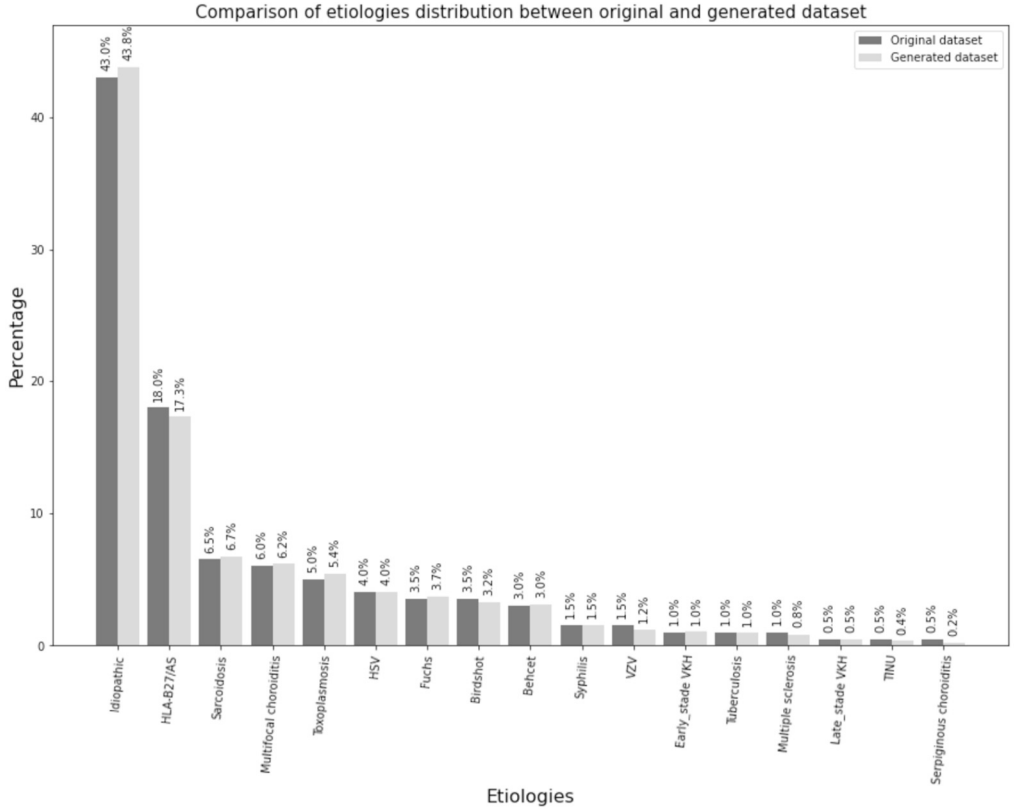


Fig. 4. Comparison of etiologies distribution between original and generated dataset.

dataset of 290 patients that we used as the base or original dataset for the initial training of our model, using 1000 epochs. To assess it correctness, a second training using the original dataset, using 500 epochs, was used to generate a new synthetic dataset. The generated dataset was assessed to have the same distribution of etiologies as that of the original dataset, and respects the defined epidemiological profile of Uveitis. The resulting distribution of etiologies is shown in Fig. 4.

Similarly, the distribution of generated values of features or columns, e.g. medical tests, was assessed. As shown in Table 3, results show that the model kept almost the same distribution in the generated dataset.

4.2. Qualitative evaluation

To assess the correctness and validity of the generated data, three ophthalmologists examined a sample of 10% randomly selected from the 2000 generated records, generated as a representative dataset for manual expert evaluation. The 2000 dataset size was deemed a suitable representative dataset, by medical experts, for the possible tedious manual expert evaluation, larger sizes may not provide additional significant evaluation gains, as shown by quantitative evaluation (see 4.3).

To ensure a representative qualitative evaluation of the proposed GAN-based work, 200 samples were randomly selected from the generated dataset, randomly shuffled the order, before manual expert evaluation by the ophthalmologists. Our specialist physicians were asked to determine how realistic those records are, using three classes of description: “Poor”, “good”, or “excellent” (excellent being most realistic). Due to time constraint, 170 records were eventually evaluated, the Results show 78 (45%) records were labeled as “poor”, 68 (40%) records were labeled as “good”, and the remaining 24 (15%) records were labeled as “excellent”, as depicted in Fig. 5.

These results are not surprising, in fact we used 27 relevant attributes that were available in the SUN article (The Standardization of Uveitis Nomenclature , SUN) to describe examination results for each

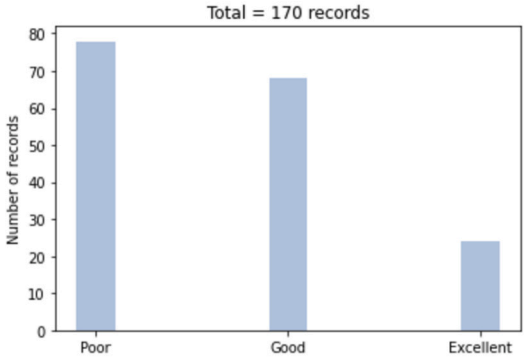


Fig. 5. Assessment results for the generated dataset’s sample.

etiology. However, our specialist physicians have noticed the absence of some important attributes that can help the algorithm to properly differentiate the etiologies, but it was difficult to include additional description, consistent with our data, from the literature. Thus the “poor” qualification is rather interpreted as a lack of information that would be necessary to integrate by additional features or columns. Additionally, 40% of records, which got a “poor” label, belonged to rare etiologies with low prevalence of less than 4%, from amongst all other rare etiologies, of which 78% were labeled as “poor” and 22% as “good”. The remaining 55%, of evaluations, were classed between “good” and excellent”. Therefore, these results are very satisfactory for, a first version of, a dataset on Uveitis, as a rare disease. As more medical information and descriptions become known and available about the disease, the quality of the generated data can, accordingly, be improved.

4.3. Quantitative evaluation

The aim for the quantitative evaluation is three folds: to test the balanced distribution of etiologies and Uveitis features in accordance

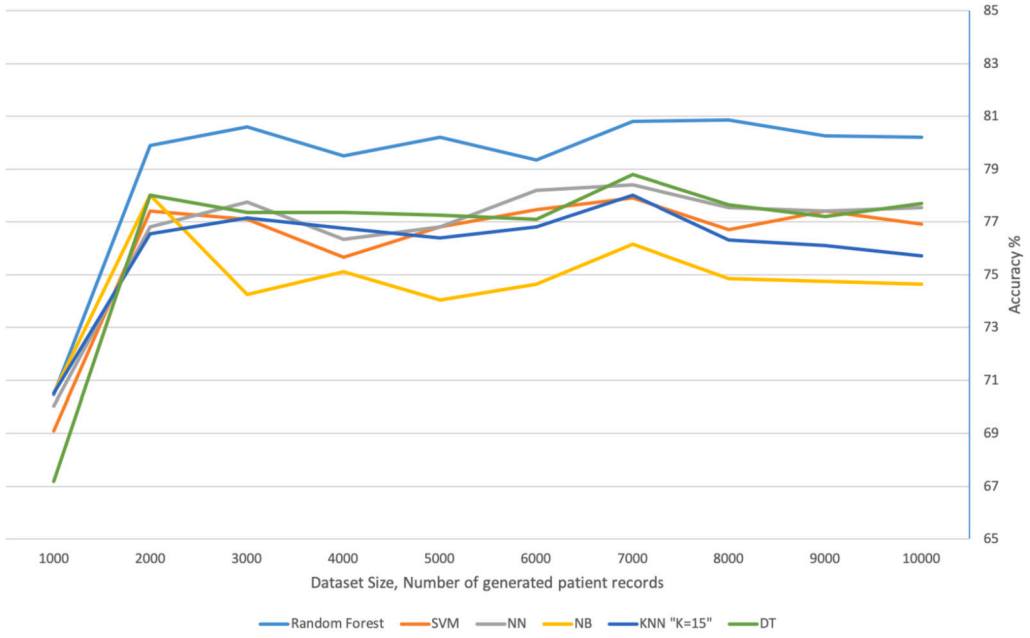


Fig. 6. Consistency and scalability evaluation of synthetic data generation method.

with its profile characteristics, to assess validity of the consistency and scalability of the proposed method and to determine the suitable dataset size that generates reasonable prediction accuracy.

For the first, on several generated datasets, statistical distribution was conducted to ensure the proposed data generation model obeys Uveitis profile characteristics as defined by the base or original dataset. Results show that our generated dataset were of good quality and draw very similar distributions, as depicted in Fig. 4 and Table 3. The proposed model kept a similar distribution of etiologies and disease features in the generated data as that in the original dataset. For Uveitis as a rare diseases, which require a specific distribution, maintaining similarity of distributions between real data and generated data is important. Results confirm that our generated data meets the requirements of a realistic dataset.

For the second, i.e. to assess the consistency and scalable data validity of the proposed synthetic data generation model, several datasets were generated with different sizes. Several machine learning methods, to draw a spectrum of behavior, were, then, applied to measure the diagnosis prediction accuracy of Uveitis etiologies. These assess how ML methods, in terms of accuracy, behave as data generation scales up or dataset size increases. For valid synthetic data generation consistency and scalability, of the proposed method, for separately generated datasets, ML methods should achieve improving prediction accuracy with increasing data sizes, until a datasize threshold. If there is inconsistency in the generated data for different sizes, diagnosis prediction accuracy for the ML methods would show random behavior. Six ML methods were selected from across a set of popular and commonly used ones for prediction to cover different types of models. The selected ML methods are Random Forest, Support Vector Machine (SVM), Neural Network (NN), Naive Bayes (NB), K-nearest neighbor (KNN) and Decision Tree (DT). To conduct the experiments, 10 datasets were generated separately, with sizes ranging from 1000 to 10000 records. Sets of these 10 datasets were generated separately, using separate runs with 500 epochs each. The six selected ML methods were then applied on each set, and the accuracy results of each for each run were averaged.

As depicted in Fig. 6, results show that, for the majority of the six tested ML methods, diagnosis prediction accuracy increases with data sizes and levels out for datasets larger than 2000 patient records (thus the selected dataset size for expert evaluation, see section 4.2), with accuracy values ranging between 70%, for smaller datasets, to more than 80%, for larger ones. Some ML methods achieve, relatively, marginal accuracy gains for larger datasets, one achieves lower accuracy with larger data sizes. The aim, in this evaluation, is not to improve prediction accuracy of ML methods, but to evaluate the consistency of the generated data validity and to assess the scalability behavior of the proposed method for different generated dataset sizes. As shown, the proposed model is consistent in generating data for synthetic electronic health records for Uveitis pathology. It achieves consistent validity across separately generated datasets, and scales well to different generation sizes.

5. Conclusion

This paper presents a methodological framework for synthetic data generation, for rare diseases, to enable accelerating the use of AI approaches to supporting their diagnosis. Our methodological framework was developed in collaboration with expert ophthalmologists on Uveitis drawing on the medical and scientific expertise of the profiles of disease. The proposed framework generates synthetic data based on the epidemiological profile representative of France population through three steps: in the first, the generation of a base or original (realistic) dataset validated by doctors, in the second, the development of Uveitis characteristic profile model, and in the third, an automatic generation of synthetic dataset by MedWGAN. The generated dataset has been evaluated by specialist physicians, making it the first publically available dataset on Uveitis. Both qualitative and quantitative evaluations show very promising results, with more than 55% of the dataset assessed by medical experts as “good” or “excellent” and achieving consistent generation of data of larger dataset sizes, with ML methods consistently obtaining more than 80% of diagnosis prediction accuracy, of Uveitis etiologies, for dataset of 2000 records or larger.

In perspectives of this work, next is to enrich medical and scientific characteristics of Uveitis, increase the size of base or original data validated by doctors to create a richer model to generate more accurate representative dataset of the rare disease. The objective is to make our work available to form a community of ophthalmologists in order to generate more accurate and reliable dataset. A second objective is to compare our sample with real dataset samples extracted from hospitals, this process is long since the prior agreement of the patient in the RGPD framework must be obtained. The third objective is to encourage the creation of a computer science community that would develop this work further using other data augmentation approaches, such as SMOTE, and develop novel AI classification uveitis approaches.

CRediT authorship contribution statement

Heithem Sliman: Conceptualization, Formal analysis, Methodology, Software, Writing – original draft. **Imen Megdiche:** Project administration, Supervision, Writing – original draft. **Loay Alajramy:** Conceptualization, Formal analysis, Methodology, Software, Writing – original draft. **Adel Taweel:** Supervision, Writing – original draft. **Sami Yangui:** Supervision, Writing – original draft. **Aida Drira:** Data curation, Resources, Validation. **Elyes Lamine:** Supervision, Writing – original draft.

Declaration of competing interest

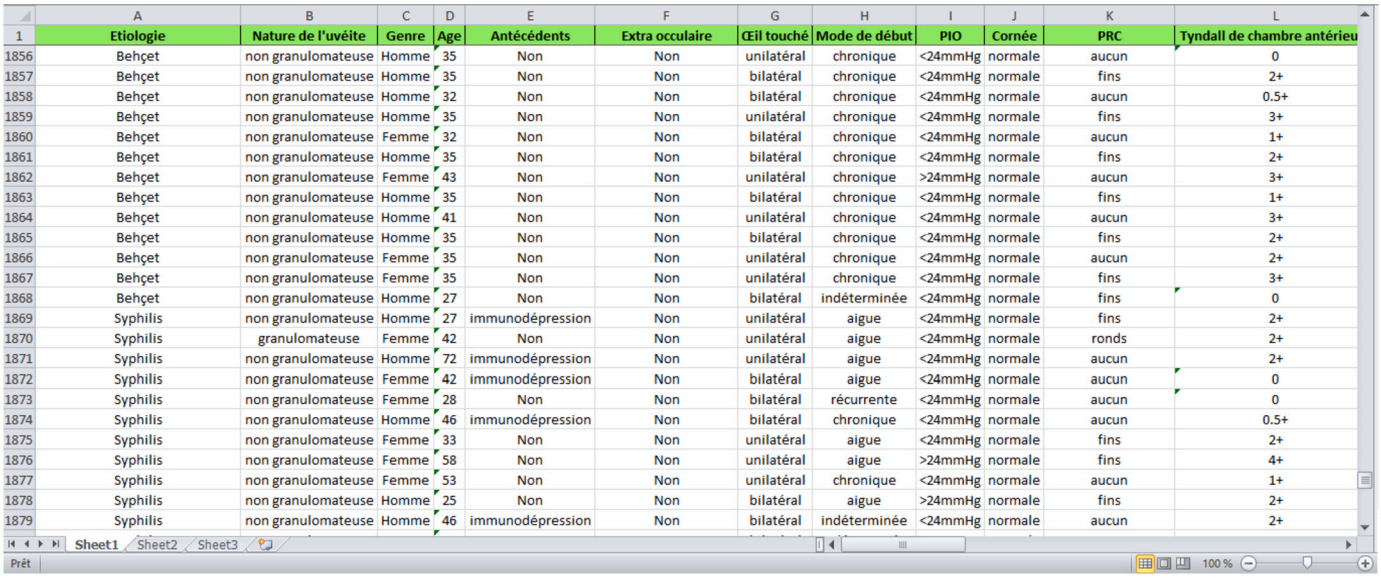
The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Imen Megdiche reports financial support was provided by Institut National Universitaire Jean Francois Champollion.

Data availability

Data will be made available on request.

Appendix A. An extract of the dataset



	A	B	C	D	E	F	G	H	I	J	K	L
1	Etiologie	Nature de l'uvéite	Genre	Age	Antécédents	Extra oculaire	Œil touché	Mode de début	PIO	Cornée	PRC	Tyndall de chambre antérieu
1856	Behçet	non granulomateuse	Homme	35	Non	Non	unilatéral	chronique	<24mmHg	normale	aucun	0
1857	Behçet	non granulomateuse	Homme	35	Non	Non	bilatéral	chronique	<24mmHg	normale	fins	2+
1858	Behçet	non granulomateuse	Homme	32	Non	Non	bilatéral	chronique	<24mmHg	normale	aucun	0.5+
1859	Behçet	non granulomateuse	Homme	35	Non	Non	unilatéral	chronique	<24mmHg	normale	fins	3+
1860	Behçet	non granulomateuse	Femme	32	Non	Non	bilatéral	chronique	<24mmHg	normale	aucun	1+
1861	Behçet	non granulomateuse	Homme	35	Non	Non	bilatéral	chronique	<24mmHg	normale	fins	2+
1862	Behçet	non granulomateuse	Femme	43	Non	Non	unilatéral	chronique	>24mmHg	normale	aucun	3+
1863	Behçet	non granulomateuse	Homme	35	Non	Non	bilatéral	chronique	<24mmHg	normale	fins	1+
1864	Behçet	non granulomateuse	Homme	41	Non	Non	unilatéral	chronique	<24mmHg	normale	aucun	3+
1865	Behçet	non granulomateuse	Homme	35	Non	Non	bilatéral	chronique	<24mmHg	normale	fins	2+
1866	Behçet	non granulomateuse	Femme	35	Non	Non	unilatéral	chronique	<24mmHg	normale	aucun	2+
1867	Behçet	non granulomateuse	Femme	35	Non	Non	unilatéral	chronique	<24mmHg	normale	fins	3+
1868	Behçet	non granulomateuse	Homme	27	Non	Non	bilatéral	indéterminée	<24mmHg	normale	fins	0
1869	Syphilis	non granulomateuse	Homme	27	immunodépression	Non	unilatéral	aigue	<24mmHg	normale	fins	2+
1870	Syphilis	granulomateuse	Femme	42	Non	Non	unilatéral	aigue	<24mmHg	normale	ronds	2+
1871	Syphilis	non granulomateuse	Homme	72	immunodépression	Non	unilatéral	aigue	<24mmHg	normale	aucun	2+
1872	Syphilis	non granulomateuse	Femme	42	immunodépression	Non	bilatéral	aigue	<24mmHg	normale	aucun	0
1873	Syphilis	non granulomateuse	Femme	28	Non	Non	bilatéral	récurrente	<24mmHg	normale	aucun	0
1874	Syphilis	non granulomateuse	Homme	46	immunodépression	Non	bilatéral	chronique	<24mmHg	normale	aucun	0.5+
1875	Syphilis	non granulomateuse	Femme	33	Non	Non	unilatéral	aigue	<24mmHg	normale	fins	2+
1876	Syphilis	non granulomateuse	Femme	58	Non	Non	unilatéral	aigue	>24mmHg	normale	fins	4+
1877	Syphilis	non granulomateuse	Femme	53	Non	Non	unilatéral	chronique	<24mmHg	normale	aucun	1+
1878	Syphilis	non granulomateuse	Homme	25	Non	Non	bilatéral	aigue	>24mmHg	normale	fins	2+
1879	Syphilis	non granulomateuse	Homme	46	immunodépression	Non	bilatéral	indéterminée	<24mmHg	normale	aucun	2+

Sheet1 Sheet2 Sheet3 Prêt 100 %

Fig. A.1. Screenshot of the generated dataset.

Appendix B. Table of etiologies

Table B.1
Resulting epidemiological profile of uveitis.

Etiology	Percentage		
idiopathic	42.521%	CMV	0.318%
HLA-B27 / AS	18.181%	Toxocariasis	0.271%
sarcoidosis	6.657%	Posner Schlossman	0.216%
Multifocal choroiditis	5.962%	APMPPE	0.216%
Toxoplasmosis	4.888%	Wegener	0.155%
HSV	4.28%	Bartonellosis	0.155%
Fuchs	3.773%	Atrophic polychondritis	0.155%
Birdshot	3.744%	Scleroderma	0.1168%
Behcet	3.421%	Drug origin	0.1168%
Syphilis	1.792%	Dental origin	0.1161%
VZV	1.584%	Juvenile chronic arthritis	0.054%
VKH	1.497%	Gougerot Sjögren	0.054%
Tuberculosis	1.246%	Sinusitis	0.054%
Oculocerebral lymphoma	0.95%	Leptospirosis	0.054%
multiple sclerosis	0.838%	Intermediate punctate choroiditis	0.054%
Lyme disease	0.779%	Trauma	0.054%
Crohn disease	0.522%	Systemic lupus erythematosus	0.038%
TINU	0.505%	Susac's syndrome	0.038%
Serpiginous choroiditis	0.488%	Fiessinger Leroy Reiter syndrome	0.038%
Inflammatory Bowel Disease (IBD)	0.48%	Eales disease	0.038%
Psoriasis	0.472%	HIV	0.038%
Paraneoplastic syndrome	0.467%		

References

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein GAN, https://doi.org/10.48550/ARXIV.1701.07875. Retrieved from https://arxiv.org/abs/1701.07875.

Baowaly, M. K., Lin, C.-C., Liu, C.-L., & Chen, K.-T. (2018). Synthesizing electronic health records using improved generative adversarial networks. *Journal of the American Medical Informatics Association*, *26*(3), 228–241. https://doi.org/10.1093/jamia/ocy142. Retrieved from arXiv:https://academic.oup.com/jamia/article-pdf/26/3/228/34151423/ocy142.pdf.

Baowaly, M. K., Lin, C.-C., Liu, C.-L., & Chen, K.-T. (2019). Synthesizing electronic health records using improved generative adversarial networks. *Journal of the American Medical Informatics Association*, *26*(3), 228–241.

Bertrand, P.-J., Jamilloux, Y., Ecochard, R., Richard-Colmant, G., Gerfaud-Valentin, M., Guillaud, M., Denis, P., Kodjikian, L., & Sève, P. (2019). Uveitis: Autoimmunity… and beyond. *Autoimmunity Reviews*, *18*(9), Article 102351. https://doi.org/10.1016/j.autrev.2019.102351. Retrieved from https://www.sciencedirect.com/science/article/pii/S156899721930151X.

Bonnet, C., & Brézin, A. (2020). Uvéites, éléments d'orientation diagnostique. *Journal Français D'ophtalmologie*, *43*(2), 145–151. https://doi.org/10.1016/j.jfo.2019.03.038. Retrieved from https://www.sciencedirect.com/science/article/pii/S0181551219304668.

Brézin, A. P. (2012). Uvéites. *La Presse Médicale*, *41*(1), 10–20. https://doi.org/10.1016/j.lpm.2011.05.011. Retrieved from https://www.sciencedirect.com/science/article/pii/S0755498211003058.

Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. F., & Sun, J. (2017a). Generating multi-label discrete patient records using generative adversarial networks. https://doi.org/10.48550/ARXIV.1703.06490. Retrieved from https://arxiv.org/abs/1703.06490.

Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. F., & Sun, J. (2017b). Generating multi-label discrete patient records using generative adversarial networks. In F. Doshi-Velez, J. Fackler, D. Kale, R. Ranganath, B. Wallace, & J. Wiens (Eds.), *Proceedings of the 2nd machine learning for healthcare conference. Proceedings of machine learning research: Vol. 68* (pp. 286–305). PMLR. Retrieved from https://proceedings.mlr.press/v68/choi17a.html.

Dahmen, J., & Cook, D. (2019). Synsys: A synthetic data generation system for healthcare applications. *Sensors*, *19*(5). https://doi.org/10.3390/s19051181. Retrieved from https://www.mdpi.com/1424-8220/19/5/1181.

Dash, S., Yale, A., Guyon, I., & Bennett, K. P. (2020a). Medical time-series data generation using generative adversarial networks. In M. Michalowski, & R. Moskovitch (Eds.), *Artificial intelligence in medicine* (pp. 382–391). Cham: Springer International Publishing.

Dash, S., Yale, A., Guyon, I., & Bennett, K. P. (2020b). Medical time-series data generation using generative adversarial networks. In M. Michalowski, & R. Moskovitch (Eds.), *AIME 2020 - international conference on artificial intelligence in medicine* (pp. 382–391). Retrieved from https://hal.inria.fr/hal-03158549.

Ding, Y., & Sato, H. (2020). Derepo: A distributed privacy-preserving data repository with decentralized access control for smart health. In *2020 7th IEEE international conference on cyber security and cloud computing (CSCloud)/2020 6th IEEE international conference on edge computing and scalable cloud (EdgeCom)* (pp. 29–35). IEEE.

Gegundez-Fernandez, J. A., Fernandez-Vigo, J. I., Diaz-Valle, D., Mendez-Fernandez, R., Cuina-Sardina, R., Santos-Bueso, E., & Benitez-del Castillo, J. M. (2017). Uvemaster: A mobile app-based decision support system for the differential diagnosis of uveitis. *Investigative Ophthalmology & Visual Science*, *58*(10), 3931–3939.

González, M. M., Solano, M. M., Porco, T. C., Oldenburg, C. E., Acharya, N. R., Lin, S. C., & Chan, M. F. (2018). Epidemiology of uveitis in a US population-based study. *Journal of ophthalmic inflammation and infection*, *8*(1), 1–8.

González-López, J., García-Aparicio, Á. M., Sánchez-Ponce, D., Muñoz-Sanz, N., Fernandez-Ledo, N., Beneyto, P., & Westcott, M. (2016). Development and validation of a Bayesian network for the differential diagnosis of anterior uveitis. *Eye*, *30*(6), 865–872.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Weinberger (Eds.), *Advances in neural information processing systems: Vol. 27* (pp. 2672–2680). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

Guan, J., Li, R., Yu, S., & Zhang, X. (2018). Generation of synthetic electronic medical record text. In *2018 IEEE international conference on bioinformatics and biomedicine (BIBM)* (pp. 374–380).

Gui, J., Sun, Z., Wen, Y., Tao, D., & Ye, J. (2020). A review on generative adversarial networks: Algorithms, theory, and applications, https://doi.org/10.48550/ARXIV.2001.06937. Retrieved from https://arxiv.org/abs/2001.06937.

Han, C., Hayashi, H., Rundo, L., Araki, R., Shimoda, W., Muramatsu, S., Furukawa, Y., Mauri, G., & Nakayama, H. (2018). Gan-based synthetic brain MR image generation. In *IEEE 15th international symposium on biomedical imaging (ISBI 2018)* (pp. 734–738).

Haute, H. (2020). *Autorité de Santé, Uvéites chroniques non infectieuses de l'enfant et de l'adulte*. Protocole National de Diagnostic et de Soins.

Hernandez, M., Epelde, G., Alberdi, A., Cilla, R., & Rankin, D. (2022). Synthetic data generation for tabular health records: A systematic review. *Neurocomputing*, *493*(C), 28–45. https://doi.org/10.1016/j.neucom.2022.04.053.

Hernandez-Matamoros, A., Fujita, H., & Perez-Meana, H. (2020). A novel approach to create synthetic biomedical signals using BiRNN. *Information Sciences*, *541*, 218–241. https://doi.org/10.1016/j.ins.2020.06.019. Retrieved from https://www.sciencedirect.com/science/article/pii/S0020025520306071.

Hoptroff, R., & El Emam, K. (2019). The synthetic data paradigm for using and sharing data. *Digital Technol.*, *19*(6). Retrieved from https://www.cutter.com/article/synthetic-data-paradigm-using-and-sharing-data-503526.

Hyun, J., Lee, S. H., Son, H. M., Park, J.-U., & Chung, T.-M. (2020). A synthetic data generation model for diabetic foot treatment. In T. K. Dang, J. Küng, M. Takizawa, & T. M. Chung (Eds.), *Future data and security engineering. Big data, security and privacy, smart city and industry 4.0 applications* (pp. 249–264). Singapore: Springer Singapore.

Jamilloux, Y., Romain-Scelle, N., Rabilloud, M., Morel, C., Kodjikian, L., Maucort-Boulch, D., Bielefeld, P., & Sève, P. (2021). Development and validation of a Bayesian network for supporting the etiological diagnosis of uveitis. *Journal of Clinical Medicine*, *10*(15). https://doi.org/10.3390/jcm10153398. Retrieved from https://www.mdpi.com/2077-0383/10/15/3398.

Kavakli-Thorne, M., Kumar, G., & Alqahtani, H. (2021). Applications of generative adversarial networks (GANs): An updated review. *Archives of Computational Methods in Engineering*, *28*, 525–552. https://doi.org/10.1007/s11831-019-09388-y. Retrieved from https://link.springer.com/article/10.1007/s11831-019-09388-y.

Majeed, A., & Lee, S. (2020). Anonymization techniques for privacy preserving data publishing: A comprehensive survey. *IEEE Access*, *9*, 8512–8545.

McLachlan, S., Dube, K., & Gallagher, T. (2016). Using the caremap with health incidents statistics for generating the realistic synthetic electronic healthcare record. In *2016 IEEE international conference on healthcare informatics (ICHI)* (pp. 439–448).

McLachlan, S., Dube, K., Gallagher, T., Simmonds, J. A., & Fenton, N. (2019). Realistic synthetic data generation: The ATEN framework. In A. Jr.Cliquet, S. Wiebe, P. Anderson, G. Saggio, R. Zwiggelaar, H. Gamboa, A. Fred, & S. Bermúdez i Badia (Eds.), *Biomedical engineering systems and technologies* (pp. 497–523). Cham: Springer International Publishing.

Neiter, E., Conart, J.-B., Baumann, C., Rousseau, H., Zuily, S., & Angioi-Duprez, K. (2019). Caractéristiques épidémiologiques et étiologiques des uvéites dans un centre hospitalier universitaire. *Journal Français D'ophtalmologie*, *42*(8), 844–851. https://doi.org/10.1016/j.jfo.2019.05.001. Retrieved from https://www.sciencedirect.com/science/article/pii/S0181551219302670.

Nguyen, A.-M., Sève, P., Le Scanff, J., Gambrelle, J., Fleury, J., Broussolle, C., Grange, J.-D., & Kodjikian, L. (2011). Aspects cliniques et étiologiques des uvéites: étude rétrospective de 121 patients adressés à un centre tertiaire d'ophtalmologie. *La Revue de Médecine Interne*, *32*(1), 9–16. https://doi.org/10.1016/j.revmed.2010.07.020. Retrieved from https://www.sciencedirect.com/science/article/pii/S024886631000874X.

Nguyen, B. (2014). Techniques d'anonymisation. *Statistique et Société*, *2*(4), 53–60. Retrieved from https://hal.archives-ouvertes.fr/hal-01113412.

Perez-Roustit, S. (2018). Epidémiologie, caractéristiques cliniques et étiologiques des uvéites prises en charge au chu de montpellier. Ph.D. thesis, Faculté de Médecine-Université de Montpellier.

Prete, M., Dammacco, R., Fatone, M. C., & Racanelli, V. (2016). Autoimmune uveitis: Clinical, pathogenetic, and therapeutic features. *Clinical and Experimental Medicine*, *16*(2), 125–136.

Rankin, D., Black, M., Bond, R., Wallace, J., Mulvenna, M., & Epelde, G. (2020). Reliability of supervised machine learning using synthetic data in health care: Model to preserve privacy for data sharing. *JMIR Medical Informatics*, *8*(7), Article e18910. https://doi.org/10.2196/18910. Retrieved from http://medinform.jmir.org/2020/7/e18910/.

Rashidian, S., Wang, F., Moffitt, R., Garcia, V., Dutt, A., Chang, W., Pandya, V., Hajagos, J., Saltz, M., & Saltz, J. (2020). SMOOTH-GAN: Towards sharp and smooth synthetic EHR data generation. In M. Michalowski, & R. Moskovitch (Eds.), *Artificial intelligence in medicine* (pp. 37–48). Cham: Springer International Publishing.

Scardapane, S., Altilio, R., Ciccarelli, V., Uncini, A., & Panella, M. (2018). Privacy-preserving data mining for distributed medical scenarios. In *Multidisciplinary approaches to neural computing* (pp. 119–128). Springer.

Sève, P., Bodaghi, B., Trad, S., Sellam, J., Bellocq, D., Bielefeld, P., Sène, D., Kaplanski, G., Monnet, D., Brézin, A., Weber, M., Saadoun, D., Cacoub, P., Chiquet, C., & Kodjikian, L. (2018). Prise en charge diagnostique des uvéites: recommandations d'un groupe d'experts. *La Revue de Médecine Interne*, *39*(9), 676–686. https://doi.org/10.1016/j.revmed.2017.09.015. Retrieved from https://www.sciencedirect.com/science/article/pii/S0248866317306719.

Sewak, M., Sahay, S., & Rathore, H. (2020). An overview of deep learning architecture of deep neural networks and autoencoders. *Journal of Computational and Theoretical Nanoscience*, *17*, 182–188. https://doi.org/10.1166/jctn.2020.8648.

The Standardization of Uveitis Nomenclature (SUN) Working Group, et al. (SUN). Development of classification criteria for the uveitis. *American Journal of Ophthalmology*, *228*, 1–280. https://doi.org/10.1016/j.ajo.2021.03.040. Retrieved from https://www.ajo.com/issue/S0002-9394(21)X0006-2.

Tucker, A., Wang, Z., Rotalinti, Y., & Myles, P. (2020). Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *npj Digital Medicine*, *3*, Article 147. https://doi.org/10.1038/s41746-020-00353-9.

Vadot, E. (1992). Epidemiology of intermediate uveitis: A prospective study in savoy. *Developments in Ophthalmology*, *23*, 33–34.

Walonoski, J., Kramer, M., Nichols, J., Quina, A., Moesel, C., Hall, D., Duffett, C., Dube, K., Gallagher, T., & McLachlan, S. (2017). Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health

care record. *Journal of the American Medical Informatics Association*, *25*(3), 230–238. https://doi.org/10.1093/jamia/ocx079. Retrieved from arXiv:https://academic.oup.com/jamia/article-pdf/25/3/230/34150150/ocx079.pdf.

Weng, L. (2019). From GAN to WGAN. https://doi.org/10.48550/ARXIV.1904.08994. Retrieved from https://arxiv.org/abs/1904.08994.

Wiehler, U., Schmidt, R., Skonetzki, S., & Becker, M. (2006). Optimierung der differenzialdiagnostischen strategie bei patienten mit sekundären uveitisformen mit einem computergestützten system. *Der Ophthalmologe*, *103*(5), 406–409.

Yale, A., Dash, S., Dutta, R., Guyon, I., Pavao, A., & Bennett, K. P. (2020). Generation and evaluation of privacy preserving synthetic health data. *Neurocomputing*, *416*, 244–255. https://doi.org/10.1016/j.neucom.2019.12.136. Retrieved from https://www.sciencedirect.com/science/article/pii/S0925231220305117.

Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *10*(2), 1–19.

Yoon, J., Drumright, L. N., & van der Schaar, M. (2020). Anonymization through data synthesis using generative adversarial networks (ADS-GAN). *IEEE Journal of Biomedical and Health Informatics*, *24*(8), 2378–2388. https://doi.org/10.1109/JBHI.2020.2980262.

Zhang, Y., Amin, S., Lung, K. I., Seabury, S., Rao, N., & Toy, B. C. (2020). Incidence, prevalence, and risk factors of infectious uveitis and scleritis in the United States: A claims-based analysis. *PLoS ONE*, *15*(8), Article e0237995.