electronics

MDPI

Article

Statistical Validation of Synthetic Data for Lung Cancer Patients Generated by Using Generative Adversarial Networks

Luis Gonzalez-Abril [1,*], Cecilio Angulo [2,3], Juan Antonio Ortega [4] and José-Luis Lopez-Guerra [5]

[1] Applied Economics I Department, Universidad de Sevilla, 41018 Sevilla, Spain
[2] Intelligent Data Science and Artificial Intelligence Research Centre, Universitat Politècnica de Catalunya, 08034 Barcelona, Spain
[3] Institut de Robòtica i Informàtica Industrial (CSIC-UPC), 08028 Barcelona, Spain
[4] Computer Science Department, Universidad de Sevilla, 41012 Sevilla, Spain
[5] Department of Radiation Oncology, University Hospital Virgen del Rocío, 41013 Sevilla, Spain
* Correspondence: luisgon@us.es

Abstract: The development of healthcare patient digital twins in combination with machine learning technologies helps doctors in therapeutic prescription and in minimally invasive intervention procedures. The confidentiality of medical records or limited data availability in many health domains are drawbacks that can be overcome with the generation of synthetic data conformed to real data. The use of generative adversarial networks (GAN) for the generation of synthetic data of lung cancer patients has been previously introduced as a tool to solve this problem in the form of anonymized synthetic patients. However, generated synthetic data are mainly validated from the machine learning domain (loss functions) or expert domain (oncologists). In this paper, we propose statistical decision making as a validation tool: Is the model good enough to be used? Does the model pass rigorous hypothesis testing criteria? We show for the case at hand how loss functions and hypothesis validation are not always well aligned.

Keywords: personalized medicine; generative adversarial network; lung cancer; validation tools

check for updates

Citation: Gonzalez-Abril, L.; Angulo, C.; Ortega, J.A.; Lopez-Guerra, J.-L. Statistical Validation of Synthetic Data for Lung Cancer Patients Generated by Using Generative Adversarial Networks. *Electronics* **2022**, *11*, 3277. https://doi.org/10.3390/electronics11203277

Academic Editors: Gorka Epelde Unanue and Darryl Charles

Received: 5 September 2022
Accepted: 8 October 2022
Published: 12 October 2022

1. Introduction

Digital twins, a concept from the industrial internet of things (IIoT), is the discipline of devising highly capable simulation models, especially those that consume data from streaming for improving performance. Devising a simulation model of clinical behavior in front of a disease is a task that is much more difficult than those for manufacturing processes, because humans are so unpredictable and engineering approaches obviously do not apply. The use of digital twins in healthcare systems is currently a hot topic under research [1]. In particular, the authors of [2] design a behavioral healthcare model for the case of lung cancer patients, which can be fed up into a decision support system [3].

Data-based solutions in the healthcare domain lead to privacy concerns [4]. Anonymization arises as a tool to mitigate risks when gathering and massively processing personal healthcare data [5]. However, data anonymization is usually seen as a major problem in data analytics because it could lead to information loss [6], reducing the knowledge contained in the dataset [7].

Nevertheless, new training procedures, such as generative adversarial networks (GANs), aim at learning representations that preserve the most relevant part of the information. As a result of the GAN-based anonymization phase, a seedbed can be obtained from the training data that allows not only to capture information from the original data avoiding privacy concerns but also to generate new synthetic information with a similar behavior to the original one [8]. Moreover, it is worth noting that obtaining clinical data has a high cost and, many times, information is very limited. By developing reliable methods for data augmentation with synthetic instances, medical professionals can benefit from this valuable information [9]. For instance, computed tomography (CT) images of the pelvis are synthetically generated in [10] for patients with cervical cancer using a conditional generative adversarial network based on a shallow U-Net (sU-Net) with an encoder/decoder depth of 2. Digital pathology and histopathological image processing are other domains where GANs are being extensively explored in medical advanced imaging [11,12]

The validation of synthetic samples, checking that they follow a distribution similar to that of real patients, is a challenge [13,14]. As a simple solution, generated synthetic data are mainly validated from the machine learning perspective using loss functions, which are those selected for model optimization. As far as the model is optimized during training, the selected loss function is minimized. Hence, the better a model is in terms of the workload of interest, the better the generated synthetic data [15]. It is well known that this implication is not always true; hence, some research is taking a different direction [16].

Performance evaluation using metrics, such as mean square error, for instance, is very complicated to be employed when information is in the form of images or graphs—in fact, in any domain that is not a metric space. For this reason, metric is usually replaced by subjective 'expert decision', oncologists in this case [17,18].

In this paper, we offer a different perspective to design an objective validation tool, which is very reasonable in decision making, applying to most of the cases in healthcare. This approach is statistical decision making—that is, to rely in hypothesis testing: is the model good enough to use? Does the model pass rigorous hypothesis testing criteria? This approach is closely related with the resemblance evaluation, especially the univariate resemblance analysis proposed in [19] when defining standardized metrics for synthetic tabular data evaluation. However, beyond the proposal of a metric, we show for the case at hand how loss minimization and hypothesis validation are not always well aligned. Hence, statistical decision making should be considered along with model optimization's losses, enforced metrics or expert assessment for validation in domains such as synthetic data generation in the healthcare domain.

The rest of this paper is structured as follows: In the next section, the available database with lung cancer patients records is introduced; the methods used, the treatment of missing values, and pre- and post-processing steps are described. Furthermore, a short introduction about generative adversarial networks is also provided. Experimentation and results are presented in Section 3, and they are validated by using the statistical criterion of goodness of fit tests. Finally, a conclusion and discussion are provided about the work developed and the results obtained.

2. Materials and Methods

In this section, the available real-world database is presented and analyzed. First, the method used to clean and pre-process this database before being used in the proposed GAN structure for synthetic data generation is described. Next, a post-processing of the synthetic data is developed in order to obtain similar data to the real data.

2.1. Database of Lung Cancer Patients

This study is carried out using data from clinical trials of diagnosed lung cancer patients once ethical approval was obtained from the associated legal body. Clinical data come from patients in the Hospital Universitario Virgen del Rocío (HUVR), which is located in Seville, Spain. This multi-center hospital complex belongs to the Andalusian Public Health System, with more than 8400 professionals under its charge and with an annual budget of more than five hundred million euros. It is a third-level hospital whose area of influence is Western Andalusia and has a staff of 1279 installed beds. The database is maintained and managed using the platform *OpenClinica* [20], which collected the information of the diagnosed lung cancer patients. In order to preserve the privacy of the patients, the dataset has been anonymized from the original database. The health team validated the anonymized dataset before any study.

Let us indicate that the considered features in the database are provided as different types: that is, nominal, ordinal and quantitative features. This fact is very important in order to elaborate an adequate data trial for the GANs because these deal with neural networks which are fed with quantitative features. Thus, all the nominal and ordinal data must be converted to numerical data in a pre-processing step. Nevertheless, nominal variables cannot be codified as quantitative features as far as no prior order is defined in them. Therefore, we hypothesize that our model will handle this type of data properly and will provide us with acceptable results [21]. The validation phase will confirm our hypothesis with respect to the nominal features.

It is worth noting that in the original dataset, outliers were previously treated for the health and statistical team in the hospital. The obtained dataset was normalized in the range from 0 to 255 in order to be converted in a 8 × 8 pixels image. This is in our interest because we want to unify all kinds of data in the form of healthcare images. In this form, data can naturally be fed to the GAN structure, as it usually works on images. Tthis form of data visualization makes the job of the medical team—understanding the information obtained from the GAN structure—easier.

The dataset under consideration, which is recorded in CSV (comma-separated values) format, contains the information about 886 lung cancer patients and 64 features. The features are related to one of the following categories: Medical record, Evolution and clinical course, Dosimetry, and Quality of life. A deep and detailed analysis of the dataset and its features can be found in [2].

An important characteristic in the dataset is missing values. Since instances in the dataset are translated into an image for each patient, these missing values are not a relevant problem. Nevertheless, it is a crucial problem when a model of machine learning is sought because the performance of these models is usually sensible to missing values.

2.2. Missing Values

The number of values of the dataset is 56,704 (886 × 64), of which 6172 are missing, that is, a percentage of 10.88%. These missing values are represented by the value ’0’ in the original dataset.

Since the main objective in this research is to replicate and validate original health data, that is, the model to be built must replicate data in a realistic way, we consider that even missing values must be replicated, because they are usually present in electronic health registers. Nevertheless, the number of missing values is too large for some instances in order to obtain a good performance from a machine learning perspective. Therefore, a study of these missing values must be carried out to eliminate those instances that are expected to downgrade the model performance, but not all of them, because we want to replicate usual registers.

In this point, two questions will be answered, and the associated decisions will be made:

1. The maximum number of missing values to be considered as valid for each instance (patient).
2. The maximum number of missing values to be considered as valid for each feature.

At this point, it is worth noting that when the medical team fills the items in the Openclinica software platform, they do not follow any standardized protocol. Hence, it can be considered that the imputation of one feature or another depends on the dedication of the doctor and the status of the patient. Therefore, if the number of missing values is high, it indicates that either the patient’s clinical history or the annotated feature is not much more relevant than the others. It is true that other motivations exist: for example, it may be that the variable has been recently considered in the software platform and not all patients have a record of it or there is new research that indicates that a feature little considered previously now is more relevant.

2.3. Pre-Processing Data

It is well known that inputs in GANs must be normalized and the range $[-1, 1]$ is recommended. Hence, a data pre-processing is required. Initially, the range of values for all the features should be $[0, 255]$ where a '0' value represents a missing value. This value is close to the value '1', which is the minimum value for all features. We have checked that, as expected, this value imputation coming from the initial dataset confuses the machine learning model when training is carried out. A solution to this problem is to define a separation space between the minimun value of the features and the '0' value. Hence, both values are moved at a distance of 90 units in the $[0, 255]$ range. For this, the transformation is performed as follows:

$$x \rightarrow \begin{cases} 0 & x = 0 \\ 90 + \dfrac{165}{254}(x-1) & x = 1, 2, \cdots, 255 \end{cases} \tag{1}$$

Next, it is scaled again to $[-1, 1]$ ($x \rightarrow \frac{x-127.5}{127.5}$) so that features take appropriate values for the machine learning model.

Once the dataset is prepared, a GAN model must be set up in order to generate synthetic data from the original ones. Thus, in the following section, GAN is briefly introduced, and the parameters required in the implementation are indicated.

2.4. Generative Adversarial Networks

Generative adversarial networks (GANs) [22] are generative models that, in short, work as follows: first, a vector noise is fed into a Generator model (usually an artificial neural network, ANN) to produce synthetic data. Next, generated data are mixed along with real data to feed a Discriminator model (again, usually an ANN), which discriminates which data come from the real dataset and which come from the synthetic data generated from the Generator.

The goal of the Generator is to fool the Discriminator and the goal of the Discriminator is not to be fooled. This confrontation leads to the Generator being increasingly capable of providing synthetic data more similar to real data. The ideal solution in the GAN model is that the percentage of success of the Discriminator for the real data and synthetic data is 50%, in both cases. The structure of the GAN model for a healthcare database can be seen in Figure 1.

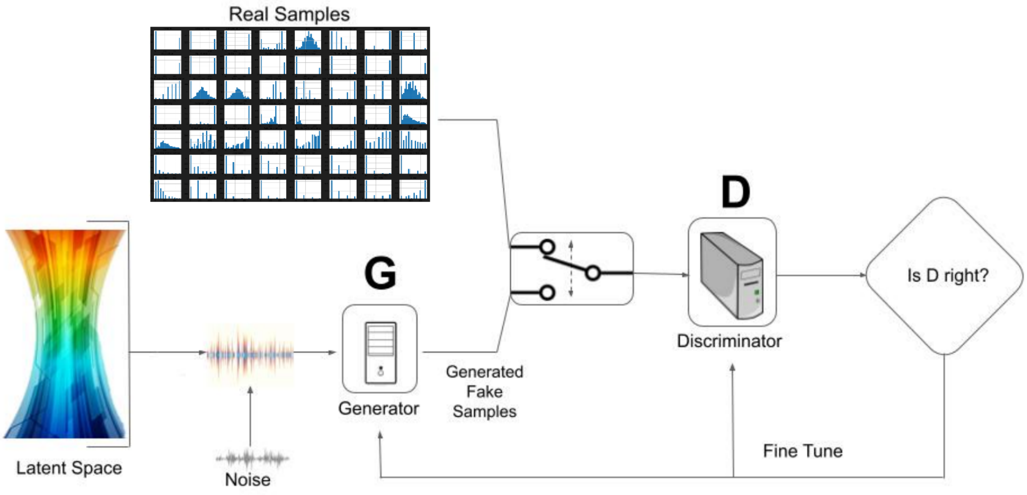

Figure 1. Structure of a Generative Adversarial Network model [2].

A brief technical introduction of the GANs model is as follows: given a real sample ($\mathbf{x}$) and some random noise vector ($\mathbf{z}$), the following terms are defined:

- $D(\mathbf{x})$ is the output of the Discriminator when a real sample $\mathbf{x}$ is processed.
- $G(\mathbf{z})$ is the output of the Generator from the noise $\mathbf{z}$, that is, the synthetic data.
- $D(G(\mathbf{z}))$ is the prediction from the Discriminator on the synthetic data.
- m is the size of samples.
- $P_{\mathbf{x}}$ and $P_{\mathbf{z}}$ are the distribution of real and noise data, respectively.
- $E_{\mathbf{x}}$ and $E_{G(\mathbf{z})}$ are the expected log likelihood from the different outputs of real and generated data.
- θ^D and θ^G are the weights of the Discriminator and Generator model, respectively.

The expression to be considered for the complete network, Discriminator and Generator, is the following, and represents a value, V,

$$V(\theta^D, \theta^G) = E_{x \sim P\mathbf{x}}[\log D(\mathbf{x})] + E_{z \sim P_{\mathbf{z}}}[\log(1 - D(G(\mathbf{z})))]. \tag{2}$$

This value function is submitted to a min–max strategy with the goal to maximize the Discriminator loss and minimize the Generator loss,

$$min_{\theta^G} \max_{\theta^D} V(\theta^D, \theta^G). \tag{3}$$

The value for the value function V is calculated as the sum of expected log likelihood for real or synthetic samples, and maximizing the resulting values leads to the optimization of the Discriminator parameters so that it learns to correctly identify both real and fake data. A database of real samples (training data) are needed so as to distinguish between real and synthetic data.

The loss function for the discriminator is the following one:

$$\nabla_{\theta^D} \frac{1}{m} \sum_{i=1}^{m} [\log D(\mathbf{x}^{(i)}) + \log(1 - D(G(\mathbf{z}^{(i)})))] \tag{4}$$

and that for the generator is,

$$\nabla_{\theta^G} \frac{1}{m} \sum_{i=1}^{m} \log(1 - D(G(\mathbf{z}^{(i)}))). \tag{5}$$

Hence, the implementation of a GAN requires an adequate setup for each employed artificial neuronal network and their training.

2.5. Post-Processing of the Output GAN

The output obtained from the GAN, that is the synthetic data, are quantitative features where values are not necessarily in the range $[-1, 1]$. Hence, a post-processing is carried out as follows:

1. Firstly, it is taken into account that (i) missing values are represented by -1, and (ii) a pre-processing step was carried out with the real dataset. Therefore, it is also necessary to consider these treatments for the synthetic dataset. For this, a threshold is considered in order to separate the output GAN in two categories: missing values and given values. The threshold is set to $-\frac{9}{17}$ in the range $[-1, 1]$ (60 in the range $[0, 255]$). On the other hand, when features are either nominal or ordinal, the first value for them in real data is $-\frac{5}{17}$ (90 in the range $[0, 255]$); hence, if the output GAN is into $[-\frac{9}{17}, -\frac{5}{17}]$, the value $-\frac{5}{17}$ is assigned. Similarly, we proceed if the output GAN is higher than 1 in the nominal and ordinal features.
   This transformation is also applied to quantitative features since, in this way, in the future, they can be converted into scaled images to $[0, 255]$, as for the initial data.

Let us name this post-processing as *first transformation*. The mathematical expression is as follows:

$$ft_1(x) = \begin{cases} -1 & \text{if} \quad x < -\frac{9}{17} \\ -\frac{5}{17} & \text{if} \quad -\frac{9}{17} \leq x \leq -\frac{5}{17} \\ 1 & \text{if} \quad x > 1 \\ x & \text{otherwise} \end{cases} \tag{6}$$

2. A *second transformation* is carried out only for nominal and ordinal features. Thus, given a synthetic value for a feature, it is transformed into the closest real value for this feature.

In order to quantify the impact of these transformations, the loss of the generator is used to quantify the loss between the real data and both the output from the GAN and the transformed output. Examples of quantifying these transformations can be seen in the next section.

3. Experimentation

In this section, the previously exposed methodology is carried out. In addition, the validation of the results obtained is analyzed.

3.1. Missing Values

As mentioned, the percentage of missing values in the data set is 10.88%. It is necessary to reduce this percentage. Thus, a study of the missing values is carried out, and by taking into account medical, statistical (cumulative histograms of the number of missing values) and machine learning criteria, the decision was to choose a maximum number of 26 missing values for each patient and 250 missing values for each feature as selection criterion.

Note that since the number of features is 64, if the threshold (26) is exceeded, it means that the patient has more than 42.18% missing values, that is, a little less than half of the values are not gathered by the medical team. After applying this filter, there are 58 patients (instances) with more than 26 missing values that are removed from the database. Therefore, the number of patients to be considered in our study is 828.

With respect to the features, since the number of patients is now 828, if the threshold (250) is exceeded, it means that the feature has more than 30.19% of missing values. After applying this filter, there exist *six* features with more than 250 missing values, which are eliminated. Hence, the number of features to be considered is 58. The features eliminated can be seen in Table 1, where the number of missing values and its percentage is also provided. A detailed description of the number of missing values by each of the 58 considered features can be observed at the end of the manuscript. It is shown that there are only 14 of 58 features with less than 10 missing values.

Table 1. The list of six features eliminated from the original dataset because the number of present missing values (# Missing Values) exceeds the threshold 250.

	Feature	# Missing Values	% Missing Values
1	PTV_Volume_cc	257	31.04
2	Studies_Level	280	33.82
3	Heart_Mean	283	34.18
4	SUV_Tumor_primary	296	35.75
5	Heart_v25	298	35.99
6	PTV_Median	318	38.41

In our previous study [2], only statistical criteria (cumulative histograms) were considered, hence only 804 patients, but 64 features composed the original database. This time, according to medical reasons, more patients were taken into consideration (from 804 to 828 patients), because they constitute a more general representation of the cohort. On the other side, some of the eight features left are important from a medical point of view; however, if they were taken into consideration, they greatly degrade the performance of the GAN model.

It can be observed in Table 2 how the original dataset has been modified. Now, the number of missing values is 2244; that is, in the dataset obtained, there are 3928 less missing values than in the original dataset.

Table 2. Summary of the obtained database after eliminating features and instances from the original dataset because of a high percentage of missing values. Symbol # stands for 'number'.

Dataset	# Features	# Patients	# Missing	% Missing
Initial	64	886	6172	10.88
Obtained	58	828	2244	4.67
Percentage (%)	92.06	93.45	36.36	—-

Hence, losing a few patients and a few features, we managed to eliminate more than 63% of the missing data. This means that despite carrying out a huge deletion of missing values, the number of them in the dataset is still relevant for validation purposes.

3.2. Pre-Processing Data

The pre-processing described in Section 2.3 is carried out on the data set for which the missing data problem has been previously treated. Now, the -1 value will denote a missing value, and $-\frac{5}{17}(=\frac{90-127.5}{127.5})$ and 1 are the minimum and the maximum values for all the features, respectively.

It is illustrated in Figure 2 how missing values are separated off the real ones for three kinds of features: Boolean (Figure 2a), ordinal (Figure 2b) and quantitative (Figure 2c). The x-axis shows the values in the new range. The y-axis represents the count of data if the variable is Boolean, categorical or ordinal and the frequencies if the variable is continuous.

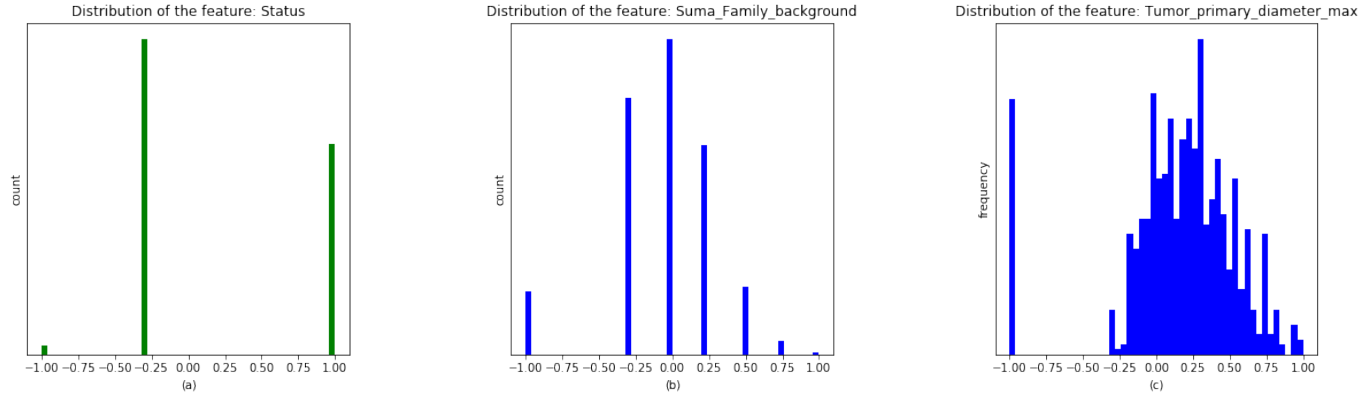

Figure 2. Bar charts of a Boolean feature (**a**) and an ordinal feature (**b**). A histogram of a quantitative feature is illustrated in (**c**).

3.3. GAN Setup

In our experimentation, the GAN used has been set up as follows: the size of the inputs in the training of the proposed GAN has been settled to *BUFFER_SIZE* = 828 (the size of the dataset), *BATCH_SIZE* = 276 (one-third of the patients (m)) and a noise vector ($\mathbf{z}$) of size 32 by following a uniform distribution in $[-1,1]$ ($\mathbf{z} \sim U(-1,1)$). The number of *EPOCHs* is 250 and the experimentation is carried out 60 times.

The architecture for the Generator $G(\mathbf{z})$ is composed by eight dense layers of size 1024, 512, 256, 128, 128, 64, 64 and 58, respectively, where the input is the noise vector of size 32 and the output is a vector of size 58, which is the same as the number of features. The architecture for the Discriminator $D(\cdot)$ is composed by five dense layers of size 512, 256, 128, 32 and 1, respectively, all of them with batch normalization and leaky (alpha = 0.01) layers. Hence, the number of trainable parameters in Generator and Discriminator (θ^G and θ^D) are 755,514 and 200,449, respectively.

The optimizer used for both, Generator and Discriminator, is the Adam optimizer with the same learning rate equal to 0.0001. The loss function for the discriminator (D) is based on the cross-entropy loss, because the discriminator performs a binary classification problem. The loss for the generator (G) is based on the mean squared error of the percentiles between real and synthetic data. The number of percentiles used is NQ = 91, which is evenly distributed from 0 to 100. This loss function is selected because by choosing the appropriate number of percentiles, we are able to empirically prove that the goodness-of-fit performance of the real and synthetic distributions can be improved. Furthermore, in order to analyze the stability of the training regime, the Fréchet Inception Distance between real and synthetic data is obtained in each iteration. A threshold value equal to 0.0001 is also set as an early stop of the code.

3.4. Model Selection

In order to run the proposed GAN, a seed is necessary to be provided for the noise vector $\mathbf{z}$. In the ideal case, the Discriminator should provide the same probabilities for the real and synthetic data, that is 50% [22]. Hence, we have defined the next function, expressed in percentage, to quantify the quality of the trained GAN structure and select the best one for experimentation:

$$quality(GAN) = 100 - (|P_R - 50| + |P_S - 50|) \tag{7}$$

where P_R and P_S are the probabilities given for the Discriminator to real $\mathbf{x}$ and synthetic G($\mathbf{z}$) data, respectively. This function is bounded in the range 0 to 100. The value 0 indicates the worst possible performance, and this is followed by a naive discriminator (if it always labels as real, then $P_R = 100\%$ and $P_S = 0\%$; and if it always labels as synthetic, then $P_R = 0\%$ and $P_R = 100\%$). On the other hand, the value 100 indicates the ideal performance when $P_R = P_S = 50\%$. Therefore, if $quality$ is close to 100, the performance of the GAN is excellent.

In the experimentation phase, many randomly chosen seeds were implemented and among all of them, the one that received a higher value in the $quality$ function was selected. The obtained highest value, $quality = 88.60\%$, indicates a very good performance of the chosen model. The result for the training of the proposed GAN on the real dataset is shown in Figure 3. Both Generator and Discriminator losses are shown in the upper part (Figure 3a,b), respectively. Both losses are stable when training is finished. The Fréchet Inception Distance, a popular metric for quantifying the distance between two distributions of images [23], is shown in Figure 3d, which is also stabilized. The probabilities given by the Discriminator to both, real (in green) and synthetic data (in blue), are shown in Figure 3c. Let us indicate that these probabilities have not completely stabilized, but this is not totally possible because of the random generation of the batch of real data in the algorithm of the GAN.

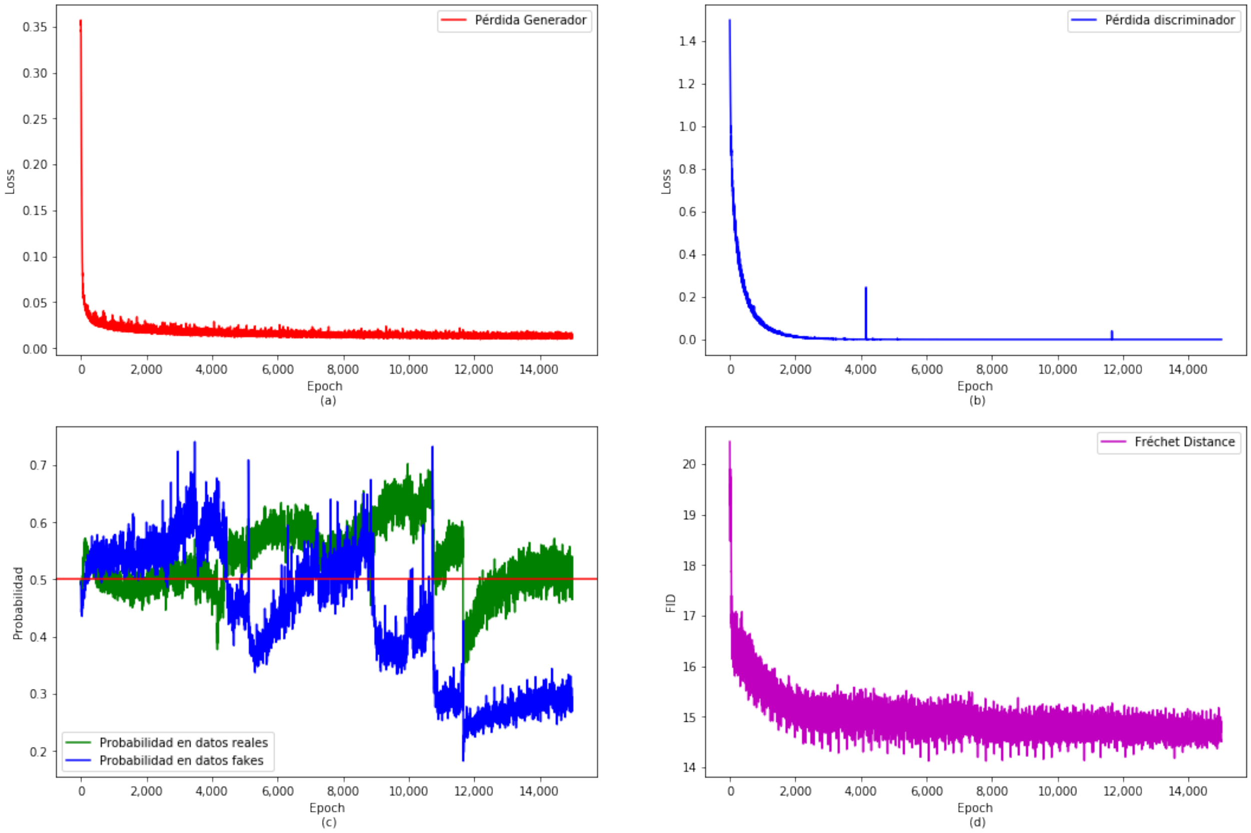

Figure 3. Training results for the implemented GAN architecture. Losses (**a**,**b**), Accuracy on real and synthetic data (**c**) and Fréchet distance (**d**) are depicted.

3.5. Generation of the Synthetic Dataset

Next, a synthetic dataset with 16, 560 instances (synthetic patients), that is, 20 times the size of the real dataset, is generated. For this synthetic dataset, the accuracy provided by the Discriminator is 50.17% for the real data, which is very close to the ideal (50%). In the case of the synthetic data, the accuracy provided by the Discriminator is 27.35%. This synthetic dataset will be used for testing and validation.

Before carrying out the proposed two post-processing transformations, let us see an example of the distribution generated by the GAN for the feature *Node*. Figure 4 shows the histogram (with 31 bins) of the generated synthetic instances/patients (upper left) and real data (lower right). It can be seen that distributions are not completely similar because the Generator is a model providing quantitative features.

Now, the post-processing presented in Section 2.5 is carried out. The post-processing is applied to the synthetic data; hence, a new dataset, called the *synthetic-patients* dataset, is generated. Before analyzing the performance of the *synthetic-patients* dataset, let us see, again for illustrative purposes, the example of the transformations on the feature *Node* (see Figure 4).

In this figure, the histogram from the raw synthetic instances is depicted on the upper left corner. On the upper right corner, the first transformation is applied, and a new histogram is obtained. Next, in the lower part left, the bar charts for the synthetic data after the second transformation can be observed. Finally, bar charts for the real data are depicted in the lower right part.

The value of the loss obtained for the feature *Node* changes when a transformation is applied. In this case, these values are 0.009208 (loss between real and synthetic data without transformation), 0.007332 (loss between real and synthetic data after the first transformation), and 0.012269 (loss between real and synthetic data after of the two transformations). The loss decreases after the first transformation, but it increases after the second transformation. The final result is worse than the one obtained initially, from a machine learning perspective (loss function), but the real distribution has a similar shape to the real data. In our experimentation, this result has always occurred: that is, the final loss value is always slightly higher than the initial loss value. However, validation using statistical tests is performing very well.

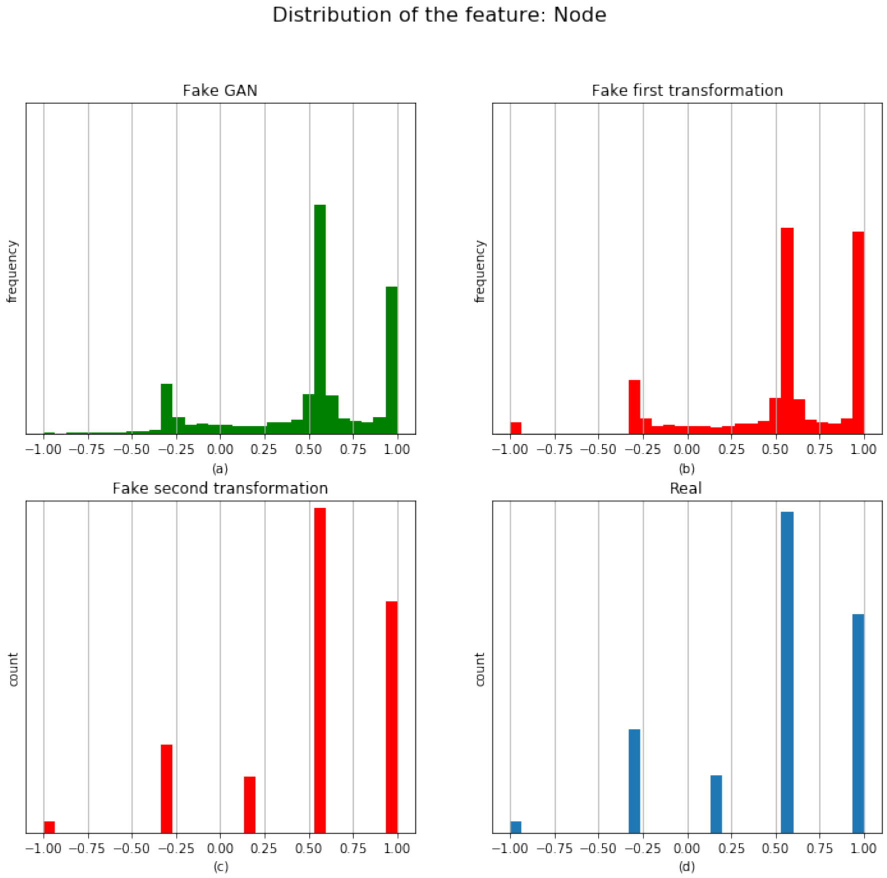

Figure 4. Distribution of the data for the feature *Node* for the synthetic database (**a**), after the first transformation (**b**) and the second transformation (**c**), and that from the real patients (**d**).

3.6. Validating the Synthetic-Patients Dataset

In order to validate whether the distribution of the *synthetic-patients* dataset can be considered similar to the real patients, goodness-of-fit tests on all the features are carried out. Thus, Pearson's chi-squared test (χ^2) is used if the feature is categorical, and the Kolmogorov–Smirnov test is used otherwise. The null hypothesis to check for each features is:

$$H_0 : \text{The synthetic-patients dataset comes from real patients distribution,}$$

and the significance level is $\alpha = 0.05 = 5\%$. Let us clarify both tests: the chi-squared test checks whether the categorical feature has the same frequencies for the synthetic patients and real patients, and the Kolmogorov–Smirnov compares the underlying continuous distributions for the synthetic patients and real patients from two independent samples.

Furthermore, in Pearson's chi-squared test, if any bin has an frequency of real patients less than 5, then these bins are combined (added) with its adjacent bins to have significance in the frequency.

The results of these tests are provided in Table 3.

This table shows the p-value and the decision with respect to H_0; that is, we can reject hypothesis H_0 if the p-value is lower than α, and we cannot reject it otherwise. Thus, for p-

value $> \alpha$, there does not exist sufficient evidence to say that the synthetic distribution of patients is different from the real distribution of patients.

Table 3. Features: number of missing, number of different values, p-value and decision on rejecting the goodness-of-fit tests between real and fake data. Symbol # stands for 'number'.

#	**Features**	**# Missing**	**# Values**	**p-Value**	**Reject H_0?**
1	Relapse_local	15	2	0.2700	False
2	Relapse_at_distance	15	2	0.1072	False
3	SocialEconomico_level	140	10	1.668×10^{-8}	**True**
4	Age	12	52	0.3036	False
5	Suma_Family_background	56	6	0.3641	False
6	Intention	8	4	0.0109	**True**
7	Smoker	0	3	0.5692	False
8	Alcoholism	3	2	0.5940	False
9	Hypertension	0	2	0.5287	False
10	Mellitus_Diabetes	0	2	0.0841	False
11	Dyslipidemia	2	2	0.5443	False
12	Heart_disease	0	2	0.1199	False
13	Thromboembolic	7	2	0.3562	False
14	Chronic obstructive pulmonary disease	2	2	0.4746	False
15	Weight_loss	2	2	0.8132	False
16	Karnofsky Performance Status	14	7	0.4717	False
17	Body mass index	118	198	0.9638	False
18	Sup_bodily	119	91	0.4481	False
19	Clinical_Stage	10	12	4.254×10^{-9}	**True**
20	Histology	5	10	1.251×10^{-15}	**True**
21	Estimated Glomerular Filtration Rate	106	4	2.171×10^{-6}	**True**
22	Anaplastic lymphoma kinase	107	4	0.6273	False
23	Tumor_primary_diameter_max	51	104	0.1895	False
24	Surgery	0	2	0.1452	False
25	Surgery_Type	0	8	1.281×10^{-6}	**True**
26	RT_Pulmonary	1	2	0.6774	False
27	Administered_Dose	26	55	4.698×10^{-23}	**True**
28	Fractionation_admin	77	20	1.143×10^{-76}	**True**
29	QT_Concomitant	18	2	0.5920	False
30	QT_indution	20	2	0.7496	False
31	Primary_overall_survival	118	218	0.4583	False
32	Primary_primary_survival	156	218	0.2205	False
33	Global health status	40	13	1.327×10^{-8}	**True**
34	Physical functioning	43	16	0.2997	False
35	Role functioning	41	7	0.8664	False
36	Emotional functioning	50	13	0.3579	False
37	Cognitive functioning	46	7	0.1723	False
38	Social functioning	46	7	0.5067	False
39	Fatigue	50	10	0.9400	False
40	Nausea and vomiting	40	7	0.8847	False
41	Pain	48	7	0.7916	False
42	Dyspnoea	42	4	0.2060	False
43	Insomnia	38	4	0.5962	False
44	Appetite loss	37	4	0.4859	False
45	Constipation	39	4	0.1960	False
46	Diarrhea	39	4	0.3015	False
47	Financial difficulties	42	4	0.3668	False
48	Dyspnoea_lung	56	10	0.1611	False
49	Coughing	46	4	0.1358	False
50	Haemoptysis	42	4	0.6604	False
51	Sore_mouth	43	4	0.9406	False
52	Dysphagia	41	4	0.7048	False
53	Peripheral_neuropathy	42	4	0.6245	False
54	Alopecia	47	4	0.3095	False
55	Tumor	27	8	9.737×10^{-23}	**True**
56	Node	13	4	0.4192	False
57	Metastasis	24	4	0.1647	False
58	Status	14	2	0.1586	False

According to the results, there are only 10 tests out of 58 in which the decision is to reject H_0, that is, the synthetic-patient dataset comes from the real patients dataset at a 0.05 level of significance in 48 of the 58 features (82.76%). Let us note that the p-value obtained in the test for the feature *Intention* is greater than the level of significance at 0.01 (1%).

For the *nine* features with a p-value less than 0.01, it is worth noting that these p-values are very close to 0; that is, the test for these features concludes that the synthetic patients dataset does not come from the real patients distribution. We think that this is the motivation because the Discriminator in the GAN provides a probability for the synthetic data of only 27.35%, which is lower than 50%. That is, the synthetic data are very similar to the real data but there are some features in which there is a significant difference between synthetic and real data.

Figure 5 shows a graphical comparative of the *nine* features with a p-value less than 0.01. It is worth noting that in some features, such as for example 'Surgery Type', which is a non-ordinal feature, from a visual viewpoint, the synthetic and real data are very similar.

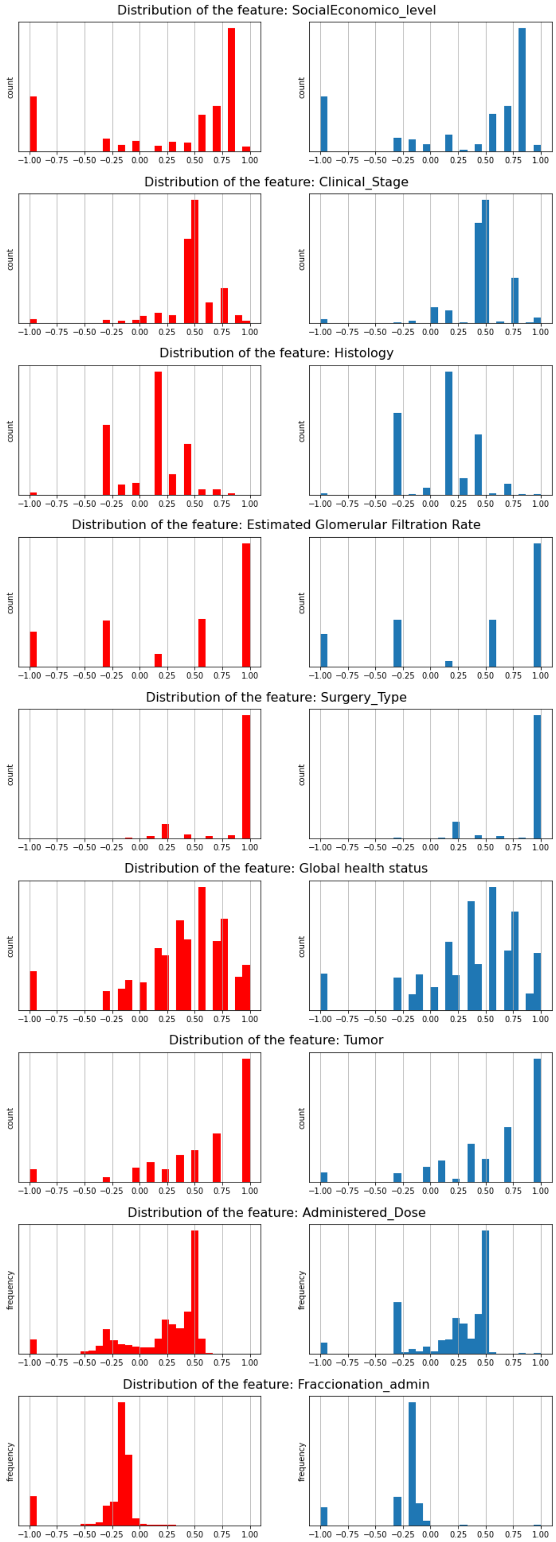

Figure 5. Distribution of the features with p-values greater than level of significance at 0.01 for synthetic and real data.

In other, as example 'Administered_Dose', the visual difference between synthetic and real data is evident.

4. Discussion and Conclusions

High-quality synthetic health data generation is a valuable resource for improving healthcare records. Generative adversarial networks can translate information from lung cancer patients in the form of images so it is possible to capture relationships across the various features in real patients.

The existence of missing values in a database is a great challenge in machine learning. The usual practice is to fill in these missing values following some data imputation procedure. However, data obtained after a medical check-up usually present many missing values. In order to replicate original records, most of these missing values must be also considered as valid data. In our approach, work is developed on a database of real patients with lung cancer where the number of missing values is large.

High-quality synthetic data are obtained using GANs in the opinion of the medical team. However, beyond this empirical evidence, some kind of objective validation tools must be provided. The presented study proposes new tools for model optimization and results validation. The GAN model is selected based on a quality index measuring the performance of the Discriminator. In the ideal case, the Discriminator's accuracy should be 50% for real and synthetic data. Hence, model optimization is working not only on the loss function but also on the reported accuracy results.

The main contribution of this paper refers to the second tool, which is associated to the statistical validation of the generated synthetic data. Transforming data into numerical features and using Pearson's chi-squared test for categorical data and the Kolmogorov–Smirnov test otherwise, a test hypothesis can be used with the null hypothesis checking whether the synthetic patients dataset distribution comes from the real patients distribution. Using this statistical test for each one of the considered features, it can be affirmed that the null hypothesis cannot be rejected for most of them, that is 48 out of 58 features (82.76%), as it is shown in Table 3. The introduced work leads to a very useful tool for validation, as it is statistically ensuring unlimited similar-to-the-original data without compromising the privacy of the original elements.

The study carried out in this paper is a novel approach for the automatic generation and validation of synthetic data in the healthcare domain. A number of limitations is still present. In particular, not all the variables are validated according to the statistical test, even though the overall model is optimized according to the loss function. How model performance varies depending on the percentage of statistically validated features and the loss function value variation is an interesting issue to be analyzed. A similar comparison could be also established with the oncologists' expert opinion.

The Kolmogorov–Smirnov test is used for continuous non-parametric one-dimension data distribution. It is one of the most used, and arguably most powerful, two-sample tests. However, as a major drawback, it is only applicable in one dimension, whereas many problems in data science cannot be compressed to one dimension without loss of information. A promising research line to be explored is the high-dimensional Kolmogorov–Smirnov distance, as introduced in [24,25]. Moreover, it can be combined with the approach in [26], where a new GAN activation function based on the Smirnov transform is used to faithfully replicate both continuous and discrete random variables.

Author Contributions: Conceptualization, L.G.-A. and C.A.; methodology, L.G.-A. and J.-L.L.-G.; software, L.G.-A., C.A. and J.A.O.; validation, L.G.-A. and J.-L.L.-G.; formal analysis, J.-L.L.-G.; investigation, L.G.-A. and C.A.; resources, J.-L.L.-G. and J.A.O.; data curation, L.G.-A. and J.-L.L.-G.; writing—original draft preparation, L.G.-A. and C.A.; writing—review and editing, C.A., J.A.O. and J.-L.L.-G.; visualization, L.G.-A.; supervision, J.-L.L.-G.; project administration, J.A.O.; funding acquisition, L.G.-A., C.A. and J.A.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partly funded by the Spanish Ministry of Science, Innovation and Universities (AEI/FEDER, UE) grant numbers PGC2018-102145-B-C21 and PGC2018-102145-B-C22 (research project EDITH).

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Ethics Committee *Comité Coordinador de Ética de la Investigación Biomédica de Andalucía* (protocol code 2282-N-20 and date 27 April 2021).

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to ethical restrictions.

Acknowledgments: We want to acknowledge the technical support given by Sara González García (Hospital Universitario Virgen del Rocío), José Mariano González Romano (Universidad de Sevilla), and Cristóbal Raya (Universitat Politècnica de Catalunya).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Elayan, H.; Aloqaily, M.; Guizani, M. Digital Twin for Intelligent Context-Aware IoT Healthcare Systems. *IEEE Internet Things J.* **2021**, *8*, 16749–16757. [CrossRef]
2. Gonzalez-Abril, L.; Angulo, C.; Ortega, J.A.; Lopez-Guerra, J.L. Generative Adversarial Networks for Anonymized Healthcare of Lung Cancer Patients. *Electronics* **2021**, *10*, 2220. [CrossRef]
3. Angulo, C.; Ortega, J.A.; Gonzalez-Abril, L. Towards a Healthcare Digital Twin. In *Frontiers in Artificial Intelligence and Applications*; Sabater-Mir, J., Torra, V., Aguiló, I., González-Hidalgo, M., Eds.; IOS Press: Oxford, UK, 2019; Volume 319, pp. 312–315.
4. Bruynseels, K.; Santoni de Sio, F.; van den Hoven, J. Digital Twins in Health Care: Ethical Implications of an Emerging Engineering Paradigm. *Front. Genet.* **2018**, *9*, 31. [CrossRef]
5. Angulo, C.; Gonzalez-Abril, L.; Raya, C.; Ortega, J.A. A Proposal to Evolving Towards Digital Twins in Healthcare. In *Proceedings of the Bioinformatics and Biomedical Engineering*; Rojas, I., Valenzuela, O., Rojas, F., Herrera, L.J., Ortuño, F., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 418–426.
6. Lee, H.; Kim, S.; Kim, J.; Chung, Y. Utility-preserving anonymization for health data publishing. *BMC Med Inform. Decis. Mak.* **2017**, *17*, 104. [CrossRef] [PubMed]
7. Gunawan, D.; Mambo, M. Set-Valued Data Anonymization Maintaining Data Utility and Data Property. In Proceedings of the 12th International Conference on Ubiquitous Information Management and Communication, Langkawi, Malaysia, 5–7 January 2018; Association for Computing Machinery: New York, NY, USA, 2018; IMCOM '18. [CrossRef]
8. Yoon, J.; Drumright, L.; Schaar, M. Anonymization Through Data Synthesis Using Generative Adversarial Networks (ADS-GAN). *IEEE J. Biomed. Health Inform.* **2020**, *24*, 2378–2388. [CrossRef] [PubMed]
9. Shin, H.C.; Tenenholtz, N.A.; Rogers, J.K.; Schwarz, C.G.; Senjem, M.L.; Gunter, J.L.; Andriole, K.P.; Michalski, M. Medical Image Synthesis for Data Augmentation and Anonymization Using Generative Adversarial Networks. In Proceedings of the Simulation and Synthesis in Medical Imaging, Granada, Spain, 16 September 2018; Gooya, A., Goksel, O., Oguz, I., Burgos, N., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 1–11.
10. Baydoun, A.; Xu, K.; Heo, J.; Yang, H.; Zhou, F.; Bethell, L.; Fredman, E.; Ellis, R.; Podder, T.; Traughber, M.; et al. Synthetic CT Generation of the Pelvis in Patients with Cervical Cancer: A Single Input Approach Using Generative Adversarial Network. *IEEE Access* **2021**, *9*, 17208–17221. [CrossRef] [PubMed]
11. Tschuchnig, M.E.; Oostingh, G.J.; Gadermayr, M. Generative Adversarial Networks in Digital Pathology: A Survey on Trends and Future Potential. *Patterns* **2020**, *1*, 100089. [CrossRef]
12. Jose, L.; Liu, S.; Russo, C.; Nadort, A.; Di Ieva, A. Generative adversarial networks in digital pathology and histopathological image processing: A review. *J. Pathol. Inform.* **2021**, *12*, 43. [CrossRef]
13. Chen, J.; Chun, D.; Patel, M.; Chiang, E.; James, J. The validity of synthetic clinical data: A validation study of a leading synthetic data generator (Synthea) using clinical quality measures. *BMC Med Inform. Decis. Mak.* **2019**, *19*, 44. [CrossRef] [PubMed]
14. Chundawat, V.; Tarun, A.; Mandal, M.; Lahoti, M.; Narang, P. TabSynDex: A Universal Metric for Robust Evaluation of Synthetic Tabular Data. *arXiv* **2022**, arXiv:2207.05295.
15. Emam, K.; Mosquera, L.; Fang, X.; El-Hussuna, A. Utility Metrics for Evaluating Synthetic Health Data Generation Methods: Validation Study. *JMIR Med. Inform.* **2022**, *10*, e35734. [CrossRef]
16. Ma, L.; Li, N.; Yu, G.; Geng, X.; Huang, M.; Wang, X. How to Simplify Search: Classification-wise Pareto Evolution for One-shot Neural Architecture Search. *arXiv* **2021**, arXiv:2109.07582.
17. Bertsimas, D.; Wiberg, H. Machine Learning in Oncology: Methods, Applications, and Challenges. *JCO Clin. Cancer Inform.* **2020**, *4*, 885–894. [CrossRef]

18. Andaur Navarro, C.L.; Damen, J.A.A.; Takada, T.; Nijman, S.W.J.; Dhiman, P.; Ma, J.; Collins, G.S.; Bajpai, R.; Riley, R.D.; Moons, K.G.M.; et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: Systematic review. *BMJ* **2021**, *375*, n2281. [CrossRef] [PubMed]
19. Hernandez, M.; Epelde, G.; Alberdi, A.; Cilla, R.; Rankin, D. Standardised Metrics and Methods for Synthetic Tabular Data Evaluation. *TechRxiv* **2021**. [CrossRef]
20. Cavelaars, M.; Rousseau, J.; Parlayan, C.; de Ridder, S.; Verburg, A.; Ross, R.; Visser, G.; Rotte, A.; Azevedo, R.; Boiten, J.; et al. OpenClinica. *J. Clin. Bioinform.* **2015**, *5*, S2. [CrossRef]
21. Piacentino, E.; Guarner, A.; Angulo, C. Generating Synthetic ECGs Using GANs for Anonymizing Healthcare Data. *Electronics* **2021**, *10*, 389. [CrossRef]
22. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In *Proceedings of the Advances in Neural Information Processing Systems*; Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2014; Volume 27.
23. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Adv. Neural Inf. Process. Syst.* **2017**, *30.* [CrossRef]
24. Hagen, A.; Strube, J.; Haide, I.; Kahn, J.; Jackson, S.; Hainje, C. A Proposed High Dimensional Kolmogorov-Smirnov Distance. In Proceedings of the Machine Learning and the Physical Sciences: Workshop at the 34th Conference on Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 6–12 December 2020.
25. Hagen, A.; Jackson, S.; Kahn, J.; Strube, J.; Haide, I.; Pazdernik, K.; Hainje, C. Accelerated Computation of a High Dimensional Kolmogorov-Smirnov Distance. *arXiv* **2021**, arXiv:2106.13706.
26. González-Prieto, Á.; Mozo, A.; Gómez-Canaval, S.; Talavera, E. Improving the quality of generative models through Smirnov transformation. *Inf. Sci.* **2022**, *609*, 1539–1566. [CrossRef]