



HAL
open science

To be or not to be, when synthetic data meet clinical pharmacology: A focused study on pharmacogenetics

Jean-baptiste Woillard, Clément Benoist, Alexandre Destere, Marc Labriffe, Giulia Marchello, Julie Josse, Pierre Marquet

► To cite this version:

Jean-baptiste Woillard, Clément Benoist, Alexandre Destere, Marc Labriffe, Giulia Marchello, et al.. To be or not to be, when synthetic data meet clinical pharmacology: A focused study on pharmacogenetics. CPT: Pharmacometrics and Systems Pharmacology, 2024, Online ahead of print. 10.1002/psp4.13240 . hal-04747078

HAL Id: hal-04747078

<https://hal.science/hal-04747078v1>

Submitted on 21 Oct 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

ARTICLE

To be or not to be, when synthetic data meet clinical pharmacology: A focused study on pharmacogenetics

Jean-Baptiste Woillard^{1,2}  | Clément Benoist^{1,2}  | Alexandre Destere^{3,4}  |
Marc Labriffe^{1,2}  | Giulia Marchello⁵ | Julie Josse⁵ | Pierre Marquet^{1,2} 

¹Pharmacology & Toxicology, Inserm, U 1248, University of Limoges, CHU Limoges, Limoges, France

²Service de Pharmacologie, Toxicologie et Pharmacovigilance, CHU Dupuytren, Limoges, France

³Department of Pharmacology and Pharmacovigilance Center, Université Côte d'Azur Medical Centre, Nice, France

⁴Inria, CNRS, Laboratoire J.A. Dieudonné, Maasai Team, Université Côte d'Azur, Nice, France

⁵Inria, PreMeDiCaL Team, University of Montpellier, Montpellier, France

Correspondence

Jean-Baptiste Woillard, INSERM U1248 P&T, University of Limoges, 2 rue du Pr Descottes, F-87000 Limoges, France.

Email: jean-baptiste.woillard@unilim.fr

Abstract

The use of synthetic data in pharmacology research has gained significant attention due to its potential to address privacy concerns and promote open science. In this study, we implemented and compared three synthetic data generation methods, CT-GAN, TVAE, and a simplified implementation of Avatar, for a previously published pharmacogenetic dataset of 253 patients with one measurement per patient (non-longitudinal). The aim of this study was to evaluate the performance of these methods in terms of data utility and privacy trade off. Our results showed that CT-GAN and Avatar used with $k = 10$ (number of patients used to create the local model of generation) had the best overall performance in terms of data utility and privacy preservation. However, the TVAE method showed a relatively lower level of performance in these aspects. In terms of Hazard ratio estimation, Avatar with $k = 10$ produced HR estimates closest to the original data, whereas CT-GAN slightly underestimated the HR and TVAE showed the most significant deviation from the original HR. We also investigated the effect of applying the algorithms multiple times to improve results stability in terms of HR estimation. Our findings suggested that this approach could be beneficial, especially in the case of small datasets, to achieve more reliable and robust results. In conclusion, our study provides valuable insights into the performance of CT-GAN, TVAE, and Avatar methods for synthetic data generation in pharmacogenetic research. The application to other type of data and analyses (data driven) used in pharmacology should be further investigated.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

Synthetic data generation has emerged as a promising approach to address privacy concerns and promote open science in pharmacological research. However, the performance of different synthetic data generation methods on “small” size datasets is not well known.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2024 The Author(s). *CPT: Pharmacometrics & Systems Pharmacology* published by Wiley Periodicals LLC on behalf of American Society for Clinical Pharmacology and Therapeutics.

WHAT THE QUESTION DID THIS STUDY ADDRESS?

This study aimed to implement and compare the performance of three synthetic data generation methods, namely CT-GAN, TVAE, and Avatar, for a non-longitudinal pharmacogenetic dataset of 253 patients previously published in terms of data utility and privacy trade off.

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

CT-GAN and Avatar had the best overall performance in terms of data utility and privacy. Data augmentation increased the number of false-positive findings. In addition, applying the algorithms multiple times improved the stability of the results.

HOW MIGHT THIS CHANGE DRUG DISCOVERY, DEVELOPMENT, OR THERAPEUTICS?

Our findings in this study can guide the selection of appropriate synthetic data generation methods for pharmacogenetic and transactional “small” datasets in pharmacology in the context of collaborative works and open science.

INTRODUCTION

Machine learning is changing the way we understand and use data in pharmacology. As the volume and complexity of biomedical data are growing, traditional analytical methods are often inadequate for capturing well the intricate relationships inherent to these datasets. Machine learning, with its ability to model complex, non-linear relationships, has emerged as a powerful tool for pharmacological research and clinical pharmacology, offering new avenues for drug discovery,¹ precision medicine,²⁻⁴ model informed precision dosing, and clinical decision-making.⁵⁻⁹

Data are very important for machine learning, but getting/producing clinical pharmacology data can be hard. Sharing data between researchers or combining data from multiple sources can significantly reduce the need for new data production. However, stringent regulations, such as the General Data Protection Regulation (GDPR), govern the sharing of medical data to protect individual privacy. Without such regulations, individuals would be highly vulnerable to privacy breaches. GDPR is, therefore, not merely a restrictive measure but a crucial framework that enables secure and respectful data sharing. It ensures that data sharing can benefit all stakeholders while safeguarding the privacy and security of individuals. Patients can be identified even when all identification variables have been suppressed from a database (called pseudonymization), in some cases.¹⁰ One way to share data without breaking these rules is to create synthetic data from real data, mimicking its initial properties. These simulated data act like real patient information but, when validated and documented, protect privacy by making it impossible to identify any

individual. In essence, synthetic data offer the benefits of real data for research without compromising patient privacy.¹¹

Synthetic data generation also allows augmenting dataset size, which can help or improve the performance of predictive ML algorithms.¹²⁻¹⁴ However, this approach is inherently constrained by the initial data subspace used as a template, with some exceptions.¹⁵

A recent method by Guillaudeux et al.¹⁶ (called Avatar) for creating synthetic data is based on reducing the dimensionality of the data, applying the nearest neighbors method in a latent space to create a local modeling area for each patient/record, and generating within this space a synthetic record at the stochastically weighted barycenter (by sampling in an exponential distribution and shuffling of the ranks of the nearest neighbors). The 2 hyperparameters in this process are the number of components and the number of nearest neighbors (k). This method has been approved by the Commission Nationale de l'Informatique et des Libertés (CNIL), the French agency that checks if data use complies with privacy law.

Deep learning techniques for synthetic data generation offer promising applications in pharmacology. Platforms such as Synthcity,¹⁷ developed by the Van der Schaar Lab (<https://www.vanderschaar-lab.com/>), provide sophisticated tools for generating synthetic tabular and survival data. Some of the methods are suitable for tabular data including tabular variational auto-encoders (VAE) and conditional tabular Generative Adversarial Networks (CT-GAN)¹⁸ as well as metrics addressing both privacy and fidelity in synthetic data generation. In brief, VAE first compresses the input data into a lower dimensional representation (encoding) and then reconstruct

the original data from this compressed form (decoding), while adding Gaussian randomness in the encoding process that ensures that the generated data points are varied yet similar to the original dataset.¹⁹ The CT-GANs consist of two parts: a generator that creates synthetic data and a discriminator that tries to distinguish between real and synthetic data. The “conditional” aspect means that the generation process can be guided by specific conditions or features, allowing for more controlled and targeted data generation.¹⁸ These methods are known to perform very well when the amount of data is rather large, but little is known about their performances in smaller datasets that are more common in clinical pharmacology or pharmacometrics.

Clinical pharmacology data predominantly exist in tabular form² or can be effectively transformed into it, encompassing a wide spectrum of goals like survival analysis, multiple linear regression with longitudinal data, and non-linear mixed-effect or non-parametric population pharmacokinetics analysis. However, most of the algorithms developed for data synthesis are meant for non-longitudinal data (with some exceptions including Avatar). Pharmacogenetics association studies are typical examples of such association between determinants (e.g., SNPs) and outcomes in a cross sectional design.

In this context, our study aimed to implement and evaluate three algorithms for synthetic longitudinal data generation in terms of data privacy and fidelity, using a previously published case study of pharmacogenetics determinants of the time-to-kidney graft loss.²⁰

MATERIALS AND METHODS

Patients and data

The dataset analyzed was from a pharmacogenetics study that linked SNPs in the 227 kidney donors with clinical outcome in the 253 respective renal transplant recipients treated with cyclosporine. Cox models were developed to evaluate the association between *ABCB1* genetic polymorphisms and the risk of graft loss.²⁰ The study was in accordance with the ethical standards of the responsible committee on human experimentation or with the Helsinki Declaration of 1975. The study showed that graft loss was significantly associated with the presence of the *ABCB1* variant haplotype 1236T/2677T/3435T in the donor (1236T/2677T/3435T vs. other haplotypes: hazard ratio (HR)=9.346; 95% confidence interval (CI) (2.278–38.461); $p=0.0019$) and with previous episodes of acute organ rejection (hazard ratio=3.077; 95% CI (1.213–7.812); $p=0.0178$). The tabular dataset was made

of one row per patient, and included continuous (recipient and donor age, cold ischemia time, and time to event) as well as categorical (donor and recipient sex, haplotype, acute rejection, donor CYP3A5 status and event) variables, without missing data. All the categorical variables had been converted into numeric ordinal variables and the haplotype effect considered additive (0, 1, or 2 risk haplotypes).

Synthetic data and data augmentation

We implemented a simplified version of the Avatar algorithm as described in Guillaudeux et al.¹⁶ in the original dataset. The number of components for the principal component analysis (PCA) step of the Avatar algorithm was set by default to the number of columns. Analyses were performed using the key parameter $k=5, 10$ and 20 (all synthetic data are stochastically generated within a surrounding of k neighbors for each patient) by creating a set of synthetic data of the same size as the real dataset. As the algorithm is stochastic, a seed was applied for reproducibility.

We evaluated the intra-dataset variability of the haplotype effect by 100 bootstrap samples of the synthetic dataset in which the final Cox model was applied and we reported the 5, 50 and 95th percentiles of the estimated HR. Then, we evaluated both the inter-dataset variability of the Avatar algorithm and the effect of changing the k parameter by changing 100 times the seed for the different numbers of k parameter (5,10 and 20). For each of them, we reported the 5, 50, and 95th percentiles of the haplotype HR after adjustment on the significant covariates.

We also investigated the effect of data augmentation for the k parameter = 5, 10, and 20 with an arbitrary increase to four times (the initial size of the original dataset) and applied the same analyses as above (intra- and inter-dataset variability and change in the k parameter).

Finally, survivalVAE and survival ctGAN were used as alternate algorithms to create synthetic and augmented data and the intra- and inter-dataset variability was evaluated as for the Avatar algorithm.

The code used to perform this study is available as an Rmd html file as supplemental data (https://github.com/jbwoillard/synthetic_data_pharmacogenetics/tree/main).

Evaluation of synthetic data

The synthetic data were evaluated comparing the distribution (median[*min*–*max*] values or $n(\%)$) and graphical matrix of distribution and covariation to those of the

original dataset. A multivariate Cox model was fitted to the synthetic and the augmented data, and the Kaplan–Meier curve and distribution of the hazard ratio adjusted on the significant covariates for each dataset (obtained after bootstrapping and inter-dataset variability) were compared to those of the original data. For each of the algorithms, we evaluated the percentage of each variable selection based on the BIC (Bayesian information criterion) across the 100 bootstraps, to see if the same variables as in the original analysis were selected (threshold of >50% of the bootstraps for selection).

Evaluation of privacy and fidelity

Utility was evaluated using the Kullback–Leibler distance and the Kolmogorov–Smirnov test. Privacy was assessed using the distance from the closest record (DCR) and the nearest neighbor distance ratio (NNDR),¹⁶ which evaluate the risk of re-identification of the original data from the synthetic data. Finally, we conducted a discrimination analysis using logistic regression by training a model to distinguish between the original and synthetic datasets.²¹ We calculated the AUC-ROC to assess the model's performance, with an AUC-ROC value close to 0.5 indicating indistinguishability between the original and synthetic data. All the analyses were performed in R version 4.2.1 except the surVAE, survival CT-GAN, and the metric calculation which were fitted using the Synthcity library¹⁷ in Python 3.9.

RESULTS

Data

Table 1 summarizes the original data and the various synthetic datasets. Significant differences were found for all variables across datasets, except the recipient sex proportion. In particular, the haplotype distributions differ from the original. As an example, **Figure 1** depicts the variation and covariation between features for the non-augmented Avatar $k=5$ (A), $k=20$ (B), CT-GAN (C) and surVAE (D) datasets in comparison to the original data (different colors). Similar figures for other synthetic datasets are available as supplemental data ([html R code file](#)). Interestingly, the Avatar synthetic data exhibit a central tendency for the continuous covariates (shrinkage towards the mean). **Figure 2** presents the Kaplan–Meier (KM) curves for the different synthetic datasets in comparison of the original one. SurVAE, with or without augmentation, produced disappointing results, generating KM curves significantly different from the original.

Conversely, the Avatar models with $k=10$ or $k=20$, and the non-augmented CT-GAN, exhibited the Kaplan–Meier curves closest to the real ones. However, they either underestimated survival without graft loss for the longest follow-up periods, with a marked decline in survival probability observed in the non-risk group (CT-GAN), or slightly overestimated it (Avatar with $k=10$ and $k=20$). Surprisingly, the $k=5$ dataset significantly overestimated the haplotype effect, as shown by the shorter time-to-event for the TTT/TTT group and the higher hazard ratio compared to the original data.

Variable selected in the final model using each algorithm

Table 2 details the variables each algorithm would have selected in the final model after bootstrapping. We observed that the augmented datasets exhibited a tendency to introduce “false-positive” associations (i.e., not significant after bootstrapping the original dataset). Only the Avatar models with values of $k=5$ and $k=20$ identified the one and only variable (haplotype) significant in the original dataset bootstrapping analysis, whereas Avatar with $k=10$ additionally identified acute rejection, which was selected in the original analysis but did not remain after bootstrapping. SurVAE did not select any variables.

Intra-dataset variability

Our analysis revealed significant intra-dataset variability for the haplotype HR. **Table 3** illustrates the quantile distribution obtained after performing 100 bootstrap samples for each algorithm. Among the tested algorithms, the non-augmented CT-GAN and augmented Avatar with $k=20$ yielded HR estimates closest to the original data. Non-augmented Avatar with $k=10$ or $k=20$ yielded slightly increased HR in comparison to the original, Avatar with $k=5$ even more so, while surVAE underestimated the effect and resulted in a non-significant result.

Inter-dataset variability

The quantile distribution of the haplotype HR after running 100 times each algorithm to create synthetic data is presented in **Table 4**. Interestingly, the Avatar algorithm with $k=5$ aligned more closely with the original results than the single run application of the algorithm (**Table 3**). Avatar with $k=10$ and $k=20$ diverged from those obtained previously. Each algorithm converged towards a similar central value—regardless of data augmentation.

TABLE 1 Data obtained with each synthetic data algorithm, as compared to the original.

Variable	Original	Augmented			Augmented			Augmented			Augmented CT-GAN	p
		k = 5	k = 10	k = 20	k = 5	k = 10	k = 20	SurvVAE	SurvVAE	CT-GAN		
N	253	253	1012	253	1012	253	1012	253	1012	253	1012	<0.001
Haplotype, n (%)												
No TTT	97 (38.3)	93 (36.8)	373 (36.9)	93 (36.8)	356 (35.2)	79 (31.2)	345 (34.1)	71 (28.1)	296 (29.2)	64 (25.3)	342 (33.8)	<0.001
TTT/other	123 (48.6)	144 (56.9)	577 (57.0)	146 (57.7)	602 (59.5)	159 (62.8)	614 (60.7)	146 (57.7)	560 (55.3)	137 (54.2)	485 (47.9)	<0.001
TTT/TTT	33 (13)	16 (6.3)	62 (6.1)	14 (5.5)	54 (5.3)	15 (5.9)	53 (5.2)	36 (14.2)	156 (15.4)	52 (20.6)	185 (18.3)	<0.001
CYP3A5 donor GG, n (%)	211 (83.4)	217 (85.8)	859 (84.9)	219 (86.6)	880 (87.0)	224 (88.5)	909 (89.8)	238 (94.1)	969 (95.8)	231 (91.3)	840 (83.0)	<0.001
Age recipient (year)	55[19–78]	55 [24, 74]	55 [23, 77]	54 [24, 73]	54 [24, 75]	55[25,74]	56[25,76]	56 [20, 78]	56 [24, 78]	68 [36, 78]	55 [19, 78]	<0.001
Sex recipient M, n (%)	156 (61.7)	163 (64.4)	656 (64.8)	165 (65.2)	666 (65.8)	178 (70.4)	700 (69.2)	163 (64.4)	647 (63.9)	170 (67.2)	650 (64.2)	0.114
Age donor (year)	40[12–73]	40[19, 68]	40 [15, 68]	40[15, 63]	40 [15, 64]	39 [18, 68]	40 [17, 68]	34 [13, 60]	34 [12, 66]	38 [19, 71]	30 [12, 58]	<0.001
Sex donor M n (%)	174 (68.8)	185 (73.1)	717 (70.8)	188 (74.3)	743 (73.4)	188 (74.3)	743 (73.4)	197 (77.9)	790 (78.1)	191 (75.5)	722 (71.3)	0.003
Acute rejection (n, %)	81 (32.0)	73 (28.9)	294 (29.1)	72 (28.5)	276 (27.3)	63 (24.9)	244 (24.1)	74 (29.2)	289 (28.6)	80 (31.6)	331 (32.7)	0.004
Cold ischemia time (h)	1153 [303, 2580]	1174 [456, 2362]	1158 [456, 2362]	1157 [570, 1987]	1141 [372,2040]	1135 [631, 2091]	1144 [577,2091]	1057 [588, 1912]	1058 [565, 2191]	1051 [597, 2115]	993 [303, 2580]	<0.001
Graft loss, n (%)	22 (8.7)	21 (8.3)	86 (8.5)	20 (7.9)	80 (7.9)	13 (5.1)	55 (5.4)	14 (5.5)	53 (5.2)	27 (10.7)	137 (13.5)	<0.001
Time to event (years)	5.34 [0.68, 15.83]	5.36 [0.97, 14.94]	5.45 [0.97, 14.94]	5.60 [0.96, 15.33]	5.31 [0.87, 15.33]	5.34 [1.20,15.10]	5.21 [1.02,15.10]	6.13 [1.07, 14.88]	6.27 [0.82, 15.70]	5.75 [0.92, 14.08]	5.88 [1.02, 15.12]	<0.001

Note: K is a hyperparameter for Avatar corresponding to the number of nearest neighbors used to stochastically generate the synthetic record.

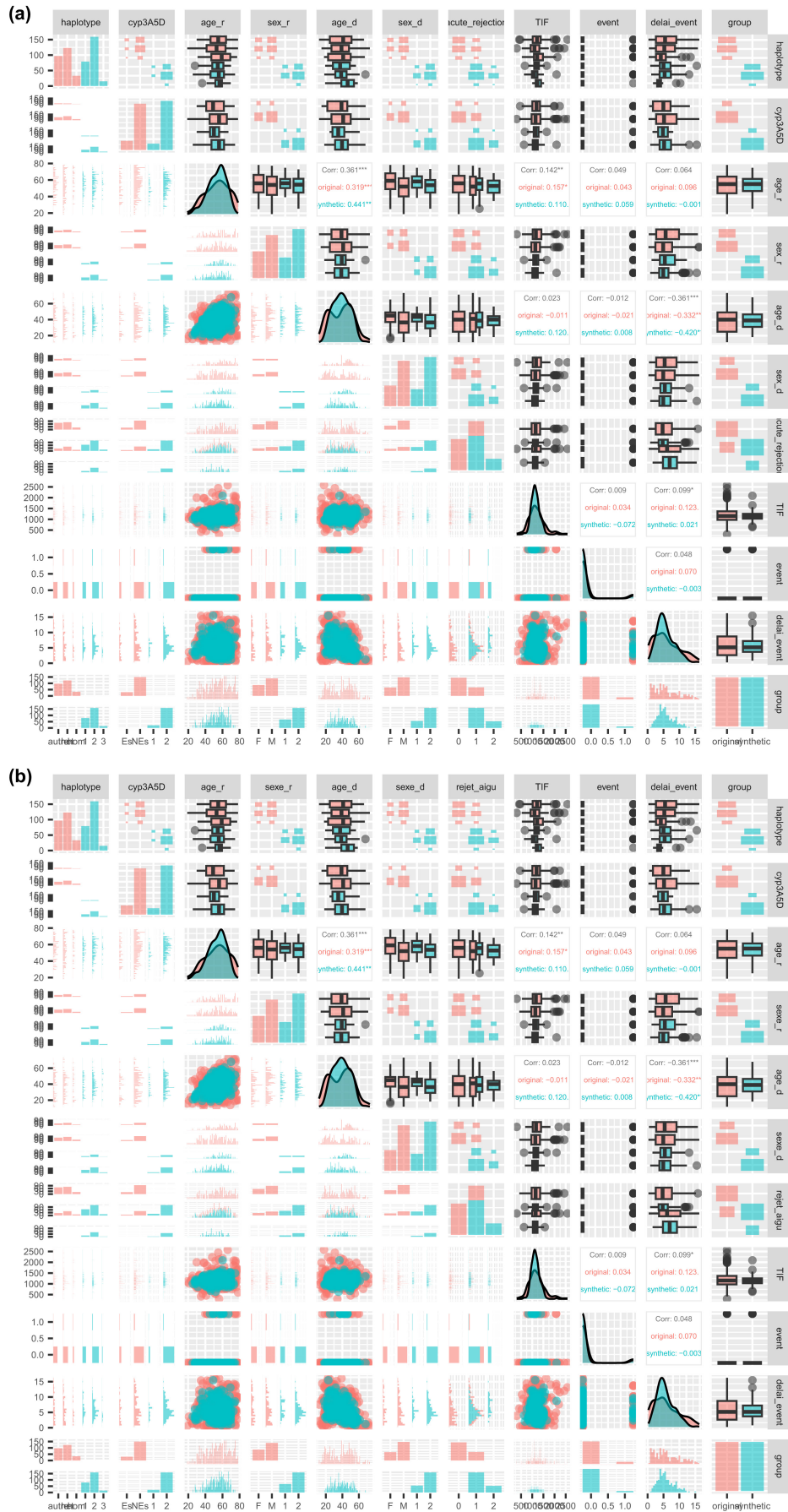


FIGURE 1 Example of variance-covariance matrix of features for the algorithms Avatar (a = with $k = 5$ and b = with $k = 20$), c = CTGAN and d = surVAE; CYP3A5D is CYP3A5 donor status, age_r and sexe_r are recipient age and sex, age_d and sexe_d are donor age and sex, reje_t_algu is acute rejection, TIF is cold ischemia time, event is graft loss, delai_event is time to event (graft loss).

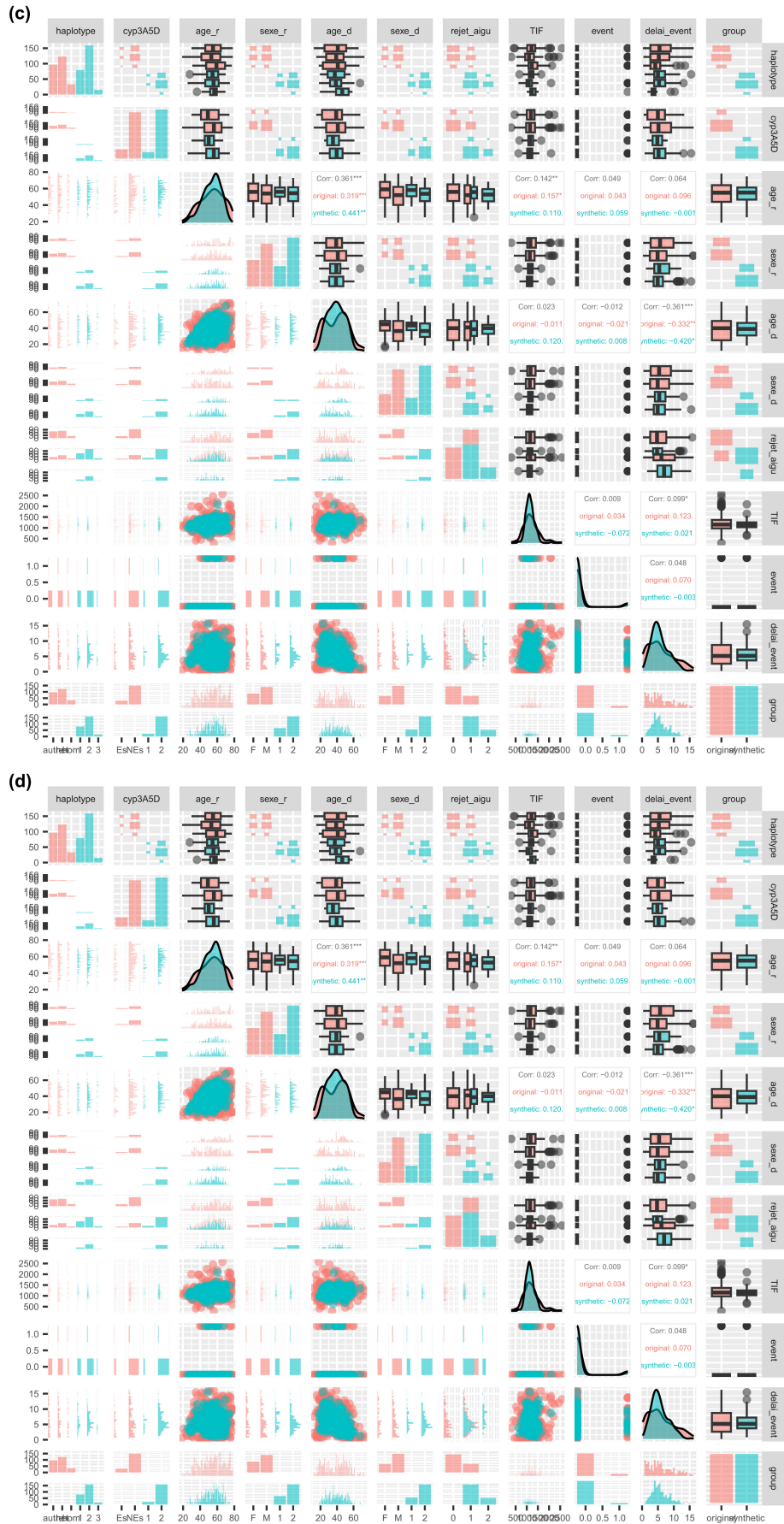


FIGURE 1 (Continued)

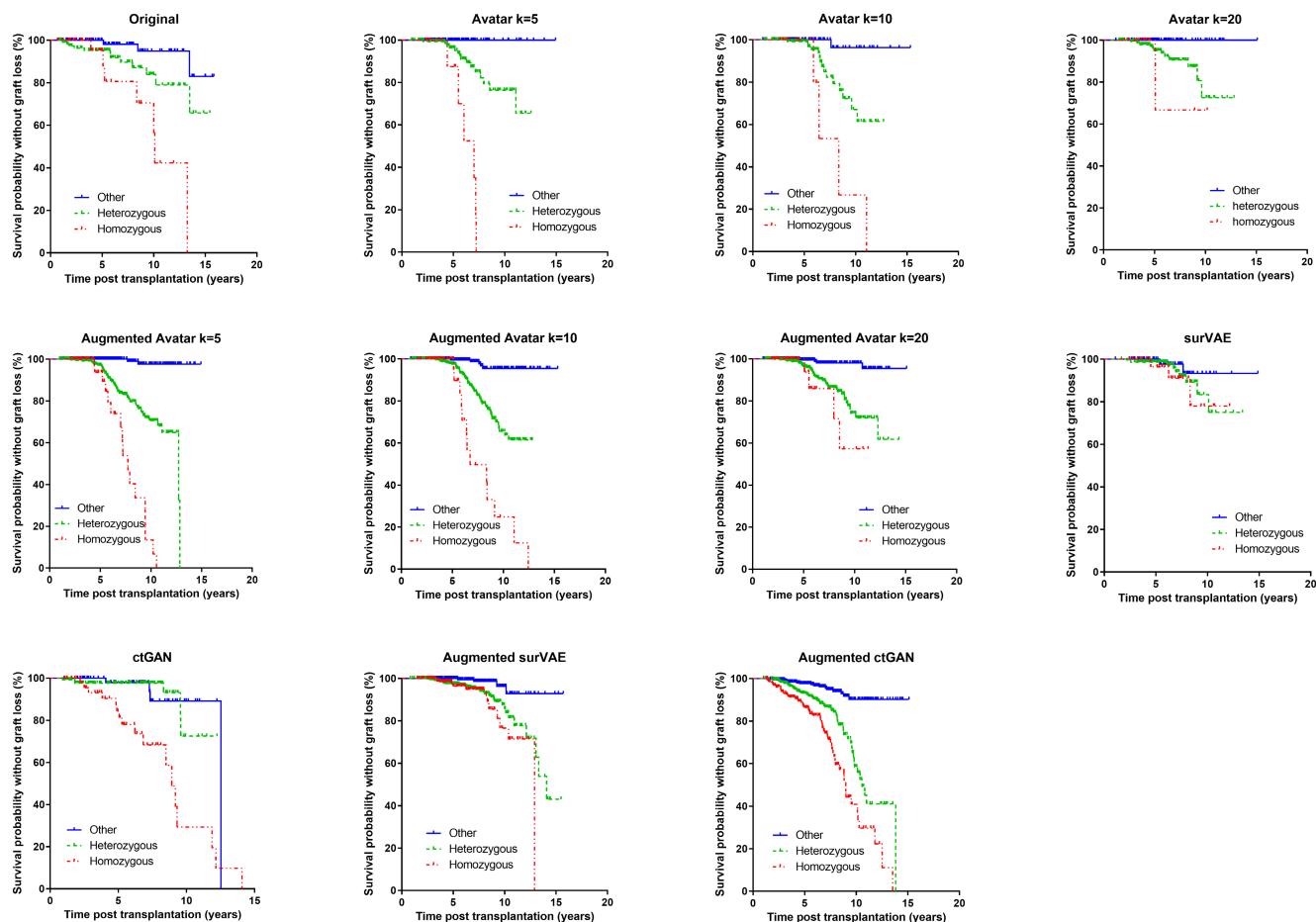


FIGURE 2 Kaplan Meier curve for the ABCB1 TTT haplotype in the original study and in the synthetic datasets obtained with Avatar $k=5$, $k=10$, and $k=20$, CT-GAN, and surVAE, augmented or not.

Additionally, the values of k do not seem to significantly affect the HR value. CT-GAN slightly underestimated the results relative to the original data, and SurVAE continued to produce disappointing results, with the 5th percentile of HR being below 1.

Privacy

The privacy and utility metrics are reported in Table 5. Briefly, the utility metrics showed that there was a good fidelity between the original data and the synthetic data whatever the method for generating synthetic data. Concerning privacy, the DCR and NNDR were the highest and close to 1 for CT-GAN and VAE, respectively, while Avatar for $k=5$ (augmented or not) yielded lower but acceptable privacy. Interestingly and as expected, when the value of k increases privacy increases, as highlighted for the Avatar $k=10$ and $k=20$ exhibiting better privacy metrics values in comparison to $k=5$. The discrimination analysis using logistic regression revealed that the AUC-ROC was approximately 0.5 for the

Avatar method. In contrast, the AUC-ROC for the CT-GAN and TVAE methods was around 0.3 (Table 5).

DISCUSSION

This study evaluates three different synthetic data generation methods for their capacity to reproduce the results obtained using a pharmacogenetic dataset characterized by a rather small sample size. We had a particular interest in the Avatar approach, recently developed by a French group, as it is certified by the French data protection authority (CNIL). This recent approach has been used successfully in some previous studies.^{22,23} The two other approaches (CT-GAN and SurVAE) have been largely used in the data augmentation/anonymization world.^{24–28} This work is also original in that it focuses on small datasets, whereas the existing methods were designed for large datasets (thousands of patients). Actually, most pharmacology studies have been performed on limited datasets (dozens or hundreds of patients).

TABLE 2 Variable selected in the final model in the 100 bootstrapping (>50% of the bootstrap) using each algorithm and comparison to original data.

Algorithm	Variable selected in the final model after bootstrapping
Original	Haplotype
$k = 5$	Haplotype
Augmented $k = 5$	Haplotype, age recipient, donor sex, acute rejection
$k = 10$	Haplotype, acute rejection
Augmented $k = 10$	Haplotype, donor age, donor sex, acute rejection
$k = 20$	Haplotype
Augmented $k = 20$	Haplotype, donor CYP3A5, donor age, acute rejection
SurvVAE	None
Augmented SurvVAE	Haplotype, donor CYP3A5
CT-GAN	Haplotype, recipient age, donor age
Augmented ctGAN	Haplotype, donor CYP3A5, donor age

Note: K is a hyperparameter for Avatar corresponding to the number of nearest neighbors used to stochastically generate the synthetic record.

While the Avatar method is founded on a machine learning approach with only a few hyperparameters to tune (such as the number of k and components for the latent projection step), CT-GAN and TVAE are based on deep learning approaches with a larger number of hyperparameters to optimize. We did not performed an exhaustive parameter search for Avatar opting to use all component to preserve as much of the original data's variability. Regarding the k parameter, we conducted experiments with different values (5, 10 and 20) to assess its impact on the synthetic data performances. This exploration can be viewed as a parameter search in a small subspace and our results indicated that $k = 10$ provided the best balance between data fidelity and privacy.

For the two deep learning methods (CT-GAN and SurvVAE), the default implementation as provided in Synthcity was used.

Our analysis revealed that the statistical descriptives of the synthetic datasets varied between the methods applied and in comparison to the original data. However, the primary goal of such analyses is not to perfectly match the statistical descriptives but to ensure that the synthetic data can select the same variables in predictive or explanatory models and conserve approximately the same effect sizes. Our results indicate that while there are differences in statistical descriptives, the synthetic data generated by the Avatar method successfully identified the key variable (haplotype).

After single-run data generation, CT-GAN achieved the best results for the ABCB1 TTT haplotype HR estimation. However, HR was slight underestimated after aggregating

TABLE 3 ABCB1 haplotype Hazard Ratio HR [95% CI] adjusted on the other significant covariates, and quantile distribution among the 100 bootstraps, for all synthetic data generation algorithms tested.

	Original	$k = 5$		$k = 10$		$k = 20$		survVAE		CT-GAN	
		$k = 5$	augmented	$k = 10$	augmented	$k = 20$	augmented	survVAE	augmented	CT-GAN	augmented
Single application	HR [95% CI]	3.22 [1.70–6.09]	10.53 [4.47–24.82]	6.45 [4.38–9.50]	5.68 [2.61–12.36]	4.44 [1.5,13.14]	3.34 [1.97,5.67]	3.34	1.75	3.20	2.42 [1.91–3.08]
Aggregation of HR after 100 bootstrap	5th percentile of the haplotype HR	2.15	7.36	5.29	3.83	2.63	2.51	1.73	0.88	2.00	1.98
	50th percentile of the haplotype HR	3.40	11.09	6.68	5.92	4.49	3.37	2.35	1.80	3.50	2.40
	95th percentile of the haplotype HR	7.06	20.47	8.68	14.68	19.49	4.82	3.25	4.23	7.77	2.91

Note: K is a hyperparameter for Avatar corresponding to the number of nearest neighbors used to stochastically generate the synthetic record.

Algorithm	5th percentile of HR	50 percentile of HR	95th percentile of HR
Original	NA	3.40	NA
$k = 5$	4.07	6.17	9.35
$k = 10$	3.98	6.43	12.84
$k = 20$	3.49	5.98	12.90
Augmented $k = 5$	4.44	5.53	7.27
Augmented $k = 10$	4.24	5.30	7.76
Augmented $k = 20$	3.58	5.30	7.25
SurvVAE	0.64	1.91	7.44
Augmented SurvVAE	0.98	1.92	3.55
CT-GAN	1.16	2.65	5.69
Augmented CT-GAN	1.46	2.50	4.45

Note: K is a hyperparameter for Avatar corresponding to the number of nearest neighbors used to stochastically generate the synthetic record.

Algorithm	KL inverse	KS test	DCR* [5th–50th–95th]	NNDR ^a [5th–50th–95th]	AUC ROC ^b
Original data	NA	NA	NA	NA	NA
Avatar $k = 5$	0.87	0.72	[0.28–0.57–1.00]	[0.14–0.33–0.75]	0.581
Avatar $k = 5$ augmented	0.93	0.92	[0.09–0.42–1.51]	[0.05–0.27–0.94]	0.534
Avatar $k = 10$	0.79	0.90	[0.50–1.00–2.14]	[0.34–0.72–0.97]	0.550
Avatar $k = 10$ augmented	0.89	0.91	[0.19–0.68–1.51]	[0.10–0.47–0.94]	0.518
Avatar $k = 20$	0.77	0.89	[0.47–1.04–1.90]	[0.34–0.74–0.98]	0.503
Avatar $k = 20$ augmented	0.76	0.88	[0.47,1.01,1.87]	[0.32–0.74–0.94]	0.474
CT-GAN	0.78	0.88	[0.83–1.78–3.39]	[0.52–0.87–0.99]	0.234
CT-GAN augmented	0.89	0.91	[0.81–1.87–3.62]	[0.52–0.87–0.99]	0.345
survVAE	0.84	0.89	[0.89–1.01–2.99]	[0.55–0.88–0.99]	0.298
survVAE augmented	0.86	0.90	[0.86–1.94–3.08]	[0.54–0.90–0.99]	0.367

Abbreviations: DCR, distance to closest record; KL, Kullback Leibler; KS, Kolmogorov Smirnov; NNDR, nearest neighbor distance ratio.

^aValue in comparison to the original data, K is a hyperparameter for Avatar corresponding to the number of nearest neighbors used to stochastically generate the synthetic record.

^bAUC-ROC for a logistic regression attempting to discriminate between original and synthetic data.

100 independent bootstrap datasets, as compared to the original effect. CT-GAN also demonstrated excellent privacy, with high DCR and NNDR values. Bootstrapping analysis of the dataset resulted in the CYP3A5 donor status, donor age, and the haplotype itself being significant risks of survival without graft loss. Even if false positive, donor age was previously reported in the literature²⁹ and donor

CYP3A5 is controversial^{30,31} and showed significance or a tendency in the original data univariate analysis.

While the Avatar approach yielded very good results compared to the original data, it produced overall an overestimation of the haplotype effect. Interestingly, in our study, increasing the value of k (up to 20, corresponding to approximately 10% of the original data size) did not

TABLE 4 Quantile distribution for haplotype HR among the 100 datasets (obtained with 100 different seeds) for each algorithm and for different k values for the avatar algorithm without adjustment on other covariates.

TABLE 5 Fidelity and privacy metrics of the synthetic data algorithms.

harm HR estimation; on the contrary, it improved it, and also enhanced the KM curve. As expected, increasing the number of K nearest neighbors improved privacy as well. Ultimately, the Avatar with $k=10$ might be the best performer, comparable to the CT-GAN.

The Avatar method resulted in the shrinkage of the synthetic data distribution compared to the original data. This is because each patient does not participate to the local space where the synthetic data will be stochastically produced. This specificity allows privacy by design as every outlier patient will be recentered automatically if they don't belong to a small cluster. While this shrinks the distribution of each variable, the impact on our estimation of the haplotype effect is unclear. We were surprised by the differences in results between a single application of Avatar with $k=20$ and multiple runs of the algorithm; the latter tended to converge towards a higher (overestimated) value. In contrast, Avatar with $k=5$ initially overestimated the results highly, but multiple applications converged towards a lower value. The variations in HR values between augmented and non-augmented datasets might be due to differences in the final model and the adjusted covariates, which result in a decreased HR value upon adjustment. Unlike the other methods, SurvVAE consistently produced disappointing results.

Based on our results and regardless of the algorithm chosen for synthetic data generation, for datasets with small sample sizes it seems very important to run the algorithm multiple times and aggregate the performance metrics. Alternatively, one could propose running the algorithm multiple times and aggregating the generated datasets before conducting survival analysis. However, this approach may lead to a decrease in the variance of effect estimation, as observed in our results for four times augmented data. In our example, CT-GAN seemed less impacted by that phenomenon because the single dataset yielded values close to the original values.

To address privacy in comparison to the original data, we used two metrics: DCR and NNDR. DCR is the Euclidean distance between a synthetic record and its closest real neighbor, with a higher distance indicating better privacy. NNDR, however, is the ratio between the distances of the closest and second-closest real neighbors for each synthetic record. A higher NNDR value (between 0 and 1) means better privacy. In this study, Avatar exhibited lower values in comparison to CT-GAN or surVAE but as discussed before, increasing the value of k increased the privacy with acceptable results for privacy starting at $k=10$. Additionally, our discrimination analysis using logistic regression showed that the AUC-ROC for the Avatar method was around 0.5, indicating a high degree of similarity between the original and synthetic data. In contrast, the AUC-ROC for CT-GAN and

TVAE was around 0.3, suggesting these methods produce synthetic data that are more easily distinguishable from the original data.

This study has a notable limitation in that its findings cannot be generalized to other types of datasets within the field of pharmacology. Originally, we intended to evaluate synthetic data generation techniques across three distinct datasets: the current pharmacogenetics dataset, a population pharmacokinetics dataset, and a longitudinal dataset. However, as the manuscript evolved, it became clear that including all three datasets would result in an overly complex and lengthy article, potentially compromising the clarity and depth of our analysis. Consequently, we chose to focus this manuscript exclusively on the pharmacogenetics dataset. This decision was made to allow for a thorough examination of the specific challenges (particularly in terms of result stability) associated with synthetic data in this context, characterized by its cross-sectional and non-longitudinal nature.

Despite this focus, there is a significant potential of synthetic data across various domains of clinical pharmacology. The analysis of population pharmacokinetics and longitudinal datasets, which present unique challenges related to data structure, variability, and privacy, is ongoing within our DIGPHAT consortium. By addressing these datasets in separate studies, we aim to provide a comprehensive and nuanced understanding of synthetic data applications across different pharmacological contexts.

In the analysis conducted here which is a specific case using survival analysis to replicate the results of a pharmacogenetic study, we observed that data augmentation increased the incidence of false positives and did not provide significant added value overall. Future research will further explore these broader applications, particularly in machine learning contexts, where synthetic data can be leveraged to augment datasets, enhance model training, and strengthen privacy protections.³²

In this work, we did not generate our data under differential privacy (DP). DP is a privacy-preserving technique that adds statistical noise to data, ensuring that the output of any analysis does not significantly differ whether an individual's data are included or not. This provides strong theoretical guarantees that individual privacy is protected and has been recommended in reports.³² While we did not generate data with DP initially, we a posteriori applied Laplace noise with an epsilon value of 0.25 (which is associated with relatively low privacy) to our dataset to evaluate the impact of DP on the utility of the synthetic data. We observed that the means of the original data with DP noise deviated significantly from the means of both the original and synthetic datasets without DP (cf. Data S1

and [code diffpriv.html](#)). This substantial loss of utility highlights the trade-off between privacy and data usefulness, particularly in the context of small and specific datasets like ours. The synthetic data generation methods we employed (CT-GAN, surVAE, and Avatar) were chosen for their ability to produce high-fidelity datasets that closely mimic the results original data. The empirical privacy metrics we used, such as DCR and NNDR, while having some pitfalls, provided a practical assessment of re-identification risk. While DP offers formal privacy guarantees, our results suggest that its implementation compromises the utility of synthetic data, making it less effective for statistical analysis and predictive modeling, which is critical in the context of our study. Therefore, the empirical privacy metrics used in our study offer a more balanced approach, ensuring adequate privacy protection without sacrificing data fidelity.

Our results might initially be seen as poor, thus limiting any future application or generalizability of the applied methods. However, although our study highlights some limitations and variabilities in the results, it also demonstrates positive findings. For instance, the haplotype variable was consistently selected by all methods except one, and similar patterns were observed in Kaplan–Meier curves for several approaches. These results, especially for the newly developed Avatar method, indicate that synthetic data are valuable for maintaining the integrity of predictive models and achieving consistent analytical outcomes. Additionally, our study addresses a critical gap in the literature by focusing on a small pharmacogenetic dataset, which is a common scenario in clinical research that often poses unique challenges.

In conclusion, no tested method allowed us to reach both perfect utility and high privacy with respect to the original dataset for the development of a survival Cox model in this pharmacogenetics case study. CT-GAN yielded good results in terms of privacy and utility, but selected some “false-positive” variables. The Avatar method recently developed is an excellent alternative for generating synthetic data, associated with a good fidelity with $k=10$ (same variables selected in comparison to the original study) and privacy in comparison to CT-GAN. On the contrary, data augmentation seems to increase the false-positive selection of variables and in the case of statistical models, the variance is reduced. surVAE yielded poor results and cannot be recommended in this context. In the case of small datasets, applying the algorithm multiple times could help to achieve more stable and reliable results.

Finally, creating synthetic data is particularly valuable because it promotes open science by providing data alongside results. Additionally, it facilitates data sharing between centers in multicenter studies as an alternative to

federated learning. This bypasses the legal complexities, costs, and time-consuming procedures often associated with traditional data sharing.

AUTHOR CONTRIBUTIONS

J-B.W., C.B., and A.D. wrote the manuscript. J-B.W., M.L., G.M., J.J., and P.M. designed the research. J-B.W., C.B., A.D., M.L., G.M., J.J., and P.M. performed the research. J-B.W., C.B., A.D., G.M., and J.J. analyzed the data.

ACKNOWLEDGMENTS

We express our sincere gratitude to Sebastien Benzekry for his invaluable advice and insightful suggestions, and to Morgan Guillaudeux for his thorough and insightful review of this paper, which significantly enhanced the quality of this work. We express our sincere gratitude to Sebastien Benzekry for his invaluable advice and insightful suggestions, and to Morgan Guillaudeux for his thorough and insightful review of this paper, which significantly enhanced the quality of this work. We also thank Christophe Battail for the insightful discussions.

FUNDING INFORMATION


This work is part of the DIGPHAT project which was supported by a grant from the French government, managed by the National Research Agency (ANR), under the France 2030 program, reference ANR-22-PESN-0017. This work was also supported by a grant from the ARS Nouvelle Aquitaine Regional APMT project.

CONFLICT OF INTEREST STATEMENT

The authors declared no competing interests for this work.

ORCID

Jean-Baptiste Woillard  <https://orcid.org/0000-0003-1695-0695>

Clément Benoist  <https://orcid.org/0009-0005-6335-7474>

Alexandre Destere  <https://orcid.org/0000-0001-6147-9201>

Marc Labriffe  <https://orcid.org/0000-0001-5840-8904>

Pierre Marquet  <https://orcid.org/0000-0001-7698-0760>

REFERENCES

1. Catacutan DB, Alexander J, Arnold A, Stokes JM. Machine learning in preclinical drug discovery. *Nat Chem Biol.* 2024;20:960-973.
2. Stankevičiūtė K, Woillard J-B, Peck RW, Marquet P, van der Schaar M. Bridging the worlds of Pharmacometrics and machine learning. *Clin Pharmacokinet.* 2023;62:1551-1565.
3. Janssen A, De Waele JJ, Elbers PWG. Towards adequate and automated antibiotic dosing. *Intensive Care Med.* 2023;49:853-856.
4. Lu J, Deng K, Zhang X, Liu G, Guan Y. Neural-ODE for pharmacokinetics modeling and its advantage to alternative machine learning models in predicting new dosing regimens. *iScience.* 2021;24:102804.

5. Poweleit EA, Vinks AA, Mizuno T. Artificial intelligence and machine learning approaches to facilitate therapeutic drug management and model-informed precision dosing. *Ther Drug Monit.* 2023;45:143-150.
6. Li Q-Y, Tang BH, Wu YE, et al. Machine learning: a new approach for dose individualization. *Clin Pharmacol Ther.* 2024;115:727-744.
7. Bica I, Alaa AM, Lambert C, van der Schaar M. From real-world patient data to individualized treatment effects using machine learning: current and future methods to address underlying challenges. *Clin Pharmacol Ther.* 2021;109:87-100.
8. Minichmayr IK, Mizuno T, Goswami S, Peck RW, Polasek TM, the American Society of Clinical Pharmacology and Therapeutics Precision Dosing Community. Recent advances addressing the challenges of precision dosing. *Clin Pharmacol Ther.* 2024;116:527-530.
9. Woillard J-B, Labriffe M, Debord J, Marquet P. Tacrolimus exposure prediction using machine learning. *Clin Pharmacol Ther.* 2021;110:361-369.
10. Culnane C, Rubinstein BIP, Teague V. Health data in an open world. *CoRR.* 2017.
11. Naik K, Goyal RK, Foschini L, et al. Current status and future directions: the application of artificial intelligence/machine learning for precision medicine. *Clin Pharmacol Ther.* 2024;115:673-686.
12. Shi G, Liu B, Walls L. Data augmentation to improve the performance of ensemble learning for system failure prediction with limited observations. In *2022 13th International Conference on Reliability, Maintainability, and Safety (ICRMS)*, 296–300. 2022. doi:10.1109/ICRMS55680.2022.9944577
13. Mumuni A, Mumuni F. Data augmentation: a comprehensive survey of modern approaches. *Array.* 2022;16:100258.
14. Iglesias G, Talavera E, González-Prieto Á, Mozo A, Gómez-Canaval S. Data augmentation techniques in time series domain: a survey and taxonomy. *Neural Comput Applic.* 2023;35:10123-10145.
15. Liu T, Qian Z, Berrevoets J, van der Schaar M. GOGGLE: generative modelling for tabular data by learning relational structure. In *The Eleventh International Conference on Learning Representations*. 2023. <https://openreview.net/forum?id=fPVRcJqspu>
16. Guillaudeux M, Rousseau O, Petot J, et al. Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis. *NPJ Digit Med.* 2023;6:37.
17. Qian Z, Cebere B-C, van der Schaar M. Synthcity: Facilitating innovative use cases of synthetic data in different data modalities. 2023.
18. Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. Modeling Tabular data using Conditional GAN. In *Advances in Neural Information Processing Systems*. 2019.
19. Mi L, Shen M, Zhang J. A probe towards understanding GAN and VAE models. *CoRR.* 2018.
20. Woillard J-B, Rerolle JP, Picard N, et al. Donor P-gp polymorphisms strongly influence renal function and graft loss in a cohort of renal transplant recipients on cyclosporine therapy in a long-term follow-up. *Clin Pharmacol Ther.* 2010;88:95-100.
21. Allen A, Siefkas A, Pellegrini E, et al. A digital twins machine learning model for forecasting disease progression in stroke patients. *Appl Sci.* 2021;11:5576.
22. Fadel M, Petot J, Gourraud P-A, Descatha A. Flexibility of a large blindly synthesized avatar database for occupational research: example from the CONSTANCES cohort for stroke and knee pain. *PLoS One.* 2024;19:e0308063.
23. Goutaudier V, Sablik M, Racapé M, et al. Design, cohort profile and comparison of the KTD-Innov study: a prospective multidimensional biomarker validation study in kidney allograft rejection. *Eur J Epidemiol.* 2024;39:549-564.
24. D'Amico S, Dall'Olio D, Sala C, et al. Synthetic data generation by artificial intelligence to accelerate research and precision medicine in hematology. *JCO Clin Cancer Inform.* 2023;7:e2300021.
25. El Kababji S, Mitsakakis N, Fang X, et al. Evaluating the utility and privacy of synthetic breast cancer clinical trial data sets. *JCO Clin Cancer Inform.* 2023;7:e2300116.
26. El Emam K, Mosquera L, Fang X, El-Hussuna A. An evaluation of the replicability of analyses using synthetic health data. *Sci Rep.* 2024;14:6978.
27. Titar RR, Ramanathan M. Variational autoencoders for generative modeling of drug dosing determinants in renal, hepatic, metabolic, and cardiac disease states. *Clin Transl Sci.* 2024;17:e13872.
28. Kikuchi T, Hanaoka S, Nakao T, et al. Synthesis of hybrid data consisting of chest radiographs and tabular clinical records using dual generative models for COVID-19 positive cases. *J Imaging Inform Med.* 2024;37:1217-1227.
29. Melk A, Sugianto RI, Zhang X, et al. Influence of donor sex and age on graft outcome in kidney transplantation. *Nephrol Dial Transplant.* 2024;39:607-617.
30. Warzyszyńska K, Zawistowski M, Karpeta E, Jałbrzykowska A, Kosieradzki M. Donor CYP3A5 expression decreases renal transplantation outcomes in white renal transplant recipients. *Ann Transplant.* 2022;27:e936276.
31. Woillard J-B, Gatault P, Picard N, Arnion H, Anglicheau D, Marquet P. A donor and recipient candidate gene association study of allograft loss in renal transplant recipients receiving a tacrolimus-based regimen. *Am J Transplant.* 2018;18:2905-2913. doi:10.1111/ajt.14894
32. Jordon J, Szpruch L, Houssiau F, et al. Synthetic data – What, why and how? 2022. <https://arxiv.org/abs/2205.03257>

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Woillard J-B, Benoist C, Destere A, et al. To be or not to be, when synthetic data meet clinical pharmacology: A focused study on pharmacogenetics. *CPT Pharmacometrics Syst Pharmacol.* 2024;00:1-13. doi:10.1002/psp4.13240