# Synthèse de données par la méthode Avatar : anonymisation et fidélité en pharmacologie de la transplantation

C. Benoist[1] & P. Marquet[1,2] & F. Stanke-Labesque[3] & J.-B. Woillard[1,2]

[1] *Service de pharmacologie, Toxicologie et pharmacovigilance, CHU de Limoges, France*

[2] *Pharmacologie & Transplantation, INSERM 1248, Université de Limoges, Limoges, France*

[3] *Univ. Grenoble Alpes, Inserm, CHU Grenoble Alpes, HP2, Grenoble 38000, France*

Auteur correspondant : Clément Benoist, Clement.BENOIST@chu-limoges.fr

**Résumé.** La confidentialité est centrale dans l'utilisation des données médicales. Néanmoins, à l'heure des entrepôts de données de santé, cette contrainte freine l'exploitation des données médicales. La question de l'anonymisation est une question difficile. Le remplacement des noms par des pseudonymes ne suffit pas : il est nécessaire qu'on ne puisse pas réidentifier les patients de la base de données. Les données médicales étant coûteuses à produire, l'idée est de produire des données synthétiques à la fois anonymisées et fidèles aux données originales. L'algorithme Avatar vise à la production de données synthétiques anonymisées ; il est reconnu conforme en termes d'anonymisation par la Commission National de l'Informatique et des Libertés, sous conditions. Nous le comparerons à l'algorithme CT-GAN (conditionnal tabular generative adversial network). Nous avons décidé de considérer des métriques de fidélité (inverse de la divergence de Kullback-Leibner, $p$-valeur du test de Kolmogorov-Smirnov) et de confidentialité (valeur de k-anonymisation). Deux jeux de données seront utilisés dans le domaine de la transplantation rénale : un jeu de données relatif à l'effet de l'inflammation sur l'exposition au tacrolimus (médicament anti-rejet, servant à prévenir la réaction immunitaire contre le greffon) et un jeu de données de pharmacocinétique de population. Nous montrerons qu'Avatar semble avoir une bonne capacité d'anonymisation sur les jeux de données de taille moyenne (inflammation du tacrolimus), sans perdre en fidélité.

**Mots-clés.** Anonymisation, confidentialité, Avatar, données génératives, médecine

**Abstract.** Confidentiality is central to the use of medical data. However, in the age of health data warehouses, this constraint is holding back the exploitation of medical data. Anonymization is a difficult issue. Replacing names with pseudonyms is not enough : patients must not be re-identified in the database. As medical data is expensive to produce, the idea is to produce synthetic data that is both anonymized and faithful to the original data. The Avatar algorithm is designed to produce anonymized synthetic data ; it is recognized as compliant in terms of anonymization by the Commission National de l'Informatique et des Libertés, subject to certain conditions. We will compare it with the CT-GAN (conditional tabular generative adversial network) algorithm. We have decided to consider fidelity metrics (inverse of the Kullback-Leibner divergence, $p$-value of the Kolmogorov-Smirnov test) and confidentiality (k-anonymization value). Two datasets will be used in the field of renal

transplantation : a dataset relating to the effect of inflammation on exposure to tacrolimus (anti-rejection drug, used to prevent the immune reaction against the graft) and a population pharmacokinetics dataset. We'll show that Avatar appears to have good anonymization capability on medium-sized datasets (tacrolimus inflammation), without losing fidelity.

**Keywords.** Anonymisation, privacy, Avatar, generative data, medicine.

# 1 Introduction

La confidentialité des données médicales est particulièrement importante dans de nombreux domaines, par exemple en médecine : par exemple, la connaissance d'un diagnostic médical pourrait mener à une augmentation des primes d'assurance ou à une discrimination à l'embauche. Une première idée pour garantir la confidentialité est de supprimer les identifiants et de les remplacer par des pseudonymes. Néanmoins certaines variables sensibles , seules ou combinées, éventuellement croisées avec d'autres jeux de données, permettent de ré-identifier le patient. Par exemple, Culnane *et al.* [2017] montrent que, sur une base de données de santé australienne en open data (MBS/MPS), 10% de la base de données étaient potentiellement ré-identifiable. Certaines ré-identifications se fondaient sur les données rares (prescriptions ou maladies rares, grossesses multiples). Les données ne sont anonymisées que si on ne peut pas réidentifier les patients (sinon, elles sont pseudonymisées). D'un point de vue légal, le **r**èglement **g**énéral sur la **p**rotection des **d**onnées (RGPD) s'applique à partir du moment où les données sont pseudonymisées.

Ces considérations légitimes de vie privée ralentissent l'échange de données sensibles entre les acteurs utilisant les données de santé. Ainsi une idée sera de créer des données synthétiques répondant à deux impératifs concurrents : la fidélité qui exige que les distributions des données originales et des données synthétiques soient identiques et l'anonymisation qui requiert qu'on ne puisse déduire des informations supplémentaires sur les patients que ces patients aient servi à générer les données synthétiques ou non (en s'inspirant de la confidentialité différentielle). Dans les applications médicales, on peut considérer la fidélité quand les jeux de données synthétiques aboutissent aux mêmes résultats dans les analyses statistiques ; on peut aussi considérer des métriques qui indiquent le niveau de fidélité. De même, en médecine, on peut considérer qu'un jeu de données est anonyme si on ne peut pas ré-identifier les patients dans le sens courant ; de même, on peut calculer des métriques pour évaluer l'anonymisation d'un jeu de données.

Il existe des méthodes de synthèse de données similaires à un jeu de données telles que les GAN (generative adversial network), la méthode Synthpop, la méthode SMOTE (**s**ynthetic **m**inority **o**versampling **te**chnique) ou Stable diffusion. Néanmoins, ces méthodes ne sont pas orientées vers la confidentialité. La méthode Avatar [Guillaudeux et al., 2023] vise à l'anonymisation des données, elle a obtenu une certification de la CNIL (**c**ommission **n**ationale de l'**i**nformatique et des **l**ibertés), autorité française indépendante chargée de la vie privée et de la protection des données, en très grande majorité dans le domaine de l'informatique. Cette certification concerne la capacité de l'algorithme à anonymiser les données.

Les algorithmes d'anonymisation, qui génèrent des données, peuvent être également uti-

lisés pour augmenter les données.

# 2 Matériels et méthodes

## 2.1 Données

Deux jeux de données sont considérés :

- un jeu de données de pharmacocinétique de population [1] pour du tacrolimus en transplantation cardiaque. Le jeu de données comporte 29 observations et 18 variables. Il y a des variables continues et des variables catégorielles. Il y a des variables démographiques (âge, sexe), des données de concentrations du tacrolimus dans le sang.
- un jeu de données sur l'effet d'inflammation sur l'exposition au tacrolimus (NCT00812786) en transplantation hépatique ; l'exposition au tacrolimus correspond à sa courbe de concentration en fonction du délai avant la prise. Le jeu de données comporte 1573 observations. Il y a 11 variables, parmi lesquelles des marqueurs d'exposition du tacrolimus et des variables démographiques (age,sexe...), biologiques. Les données combinent des données catégorielles et continues.

L'objectif est de comparer les méthodes d'anonymisation sur des jeux de données de petit taille (ce qui est fréquent en médecine) et de grand taille (dans le domaine médical : la grande taille est relative).

## 2.2 Méthodes de génération de données synthétiques

### 2.2.1 Méthode à tester : Avatar

Notre méthode reprend l'algorithme d'Avatar. En entrée, nous avons un jeu de données pseudonymisées. Les données sont projetées dans un espace latent multidimensionnel par analyse en composantes principales. En considérant uniquement les $n_d$ premières composantes principales, on cherche les $k$ plus proches voisins, pour une donnée dans l'espace latent. Pour chaque individu, un avatar est synthétisé de manière stochastique à partir de ses $k$ plus proches voisins. Un nombre quelconque d'avatars peut être généré.

Considérons les $k$ plus proches voisins d'un individu $O$ dans l'espace latent, chacun sera associé à un coefficient. Soit $D_i$ l'inverse de la distance entre $O$ et le $i$-ème plus proche voisin. Soit $R_i$ une réalisation de la loi exponentielle d'intensité $\lambda$. Pour obtenir le vecteur $(C_i)_{i=1}^k$, on permute le $k$-uplet $(1/2, (1/2)^2, \ldots, (1/2)^k)$.

[1] La pharmacocinétique mesure l'exposition d'un médicament. L'exposition peut être représentée par la concentration du médicament en fonction du temps. Pour cela, on utilise des modèles paramétriques appelé modèles pharmacocinétiques. La variabilité sur les paramètres pharmacocinétiques en fonction des individus et de leur caractéristiques associées (sexe, âge...), souvent à l'aide de modèles non linéaires à effets mixtes, est le champ d'étude de la pharmacocinétique.

Le poids $P_i$ sera calculé de manière naturelle : $P_i = D_i \times R_i \times C_i$.

Le poids sera normalisé :

$$W_i = \frac{P_i}{\sum_{j=1}^{k} P_j}$$

Ainsi, $W_i$ sera le poids associé au $i$-ème voisin. Une combinaison linéaire permet d'obtenir les avatars.

L'opération peut être répétée.

### 2.2.2 Méthode de référence : CTGAN (conditionnal tabular generative adversial network)

L'algorithme CTGAN décrit dans [Xu *et al.*, 2019] est une méthode qui, à partir d'un apprentissage, génère des données synthétiques tabulaires similaires aux données d'apprentissage. Je rappelle le détail de cet algorithme.

On considère une table où $N_c$ colonnes $C_1, \ldots, C_{N_c}$ correspondent aux réalisations de variables aléatoires continues et $N_d$ colonnes $D_1, \ldots, D_{N_d}$ correspondent aux réalisations de variables aléatoires discrètes.

La première étape est de prétraiter les colonnes associées à des variables continues grâce à une normalisation à mode spécifique. Les colonnes associées à des variables aléatoires continues notées $C_i = (c_{i,j})_{i,j}$ seront modifiées en utilisant le modèle à mélange de gaussiennes variationnel : le nombre de modes $m_i$ est estimé et le mélange de gaussiennes est ajusté pour en déterminer les paramètres. Notons ces modes $\eta_k$, $k = 1, \ldots, m_i$. Ainsi la densité de $C_i$ peut s'écrire

$$f_{C_i} = \sum_{k=1}^{m_i} \mu_{i,j,k} \mathcal{N}(c_{i,j}; \eta_{i,j,k}, \varphi_{i,j,k})$$

où $\mathcal{N}(\cdot; \tilde{\eta}, \tilde{\varphi})$ est la densité de la loi normale de moyenne $\tilde{\eta}$ et d'écart-type $\tilde{\varphi}$.

Pour chaque valeur de probabilité, on calcule les densités de probabilité des $c_{i,j}$ associées à chacun de leur mode. On a des densités $\rho_{i,j,1}, \ldots, \rho_{i,j,m_i}$. Ces densités valent $\rho_{i,j,k} = \mu_k \mathcal{N}(c_{i,j}; \eta_k, \varphi_k)$.

On cherche $\ell^*_{i,j}$ qui maximise en $\ell$ la quantité $\rho_{i,j,\ell}$ ; notons ce maximum $\rho^*_{i,j}$. Nous considérons $c_{i,j,\ell^*_{i,j}}$. Nous posons $\beta_{i,j}$ le vecteur de taille $m_i$ qui vaut 0 partout sauf dans la position d'ordre $\ell^*_{i,j}$ où la valeur vaut 1. On considère aussi le scalaire $\alpha_{i,j} = \frac{c_{i,j,\ell^*_{i,j}} - \eta_{i,j,\ell^*_{i,j}}}{4\varphi_{\ell^*_{i,j}}}$.

La ligne sera représentée par :

$$r_j = \alpha_{1,j} \oplus \beta_{1,j} \oplus \cdots \oplus \alpha_{N_c,j} \oplus \beta_{N_c,j} \oplus d_{1,j} \oplus \cdots \oplus d_{N_d,j}$$

où $\oplus$ est l'opérateur de concaténation, $N_c$ (respectivement $N_d$) est le nombre de variables aléatoires continues (respectivement discrètes) et les $d_{i,j}$ sont encodés selon l'encodage One-Hot des colonnes correspondant aux variables aléatoires discrètes.

Le GAN utilise deux réseaux de neurones : un générateur, qui génère des données synthétiques et un discriminateur, qui propose un score qui évalue la proximité statistique entre les vraies données et les données synthétiques, issues du générateur (conditionnel).

Le CTGAN utilise un générateur conditionnel. Soit $i^*$ un indice de numérotation des colonnes discrètes. Soit $R$ la variable aléatoire sous-jacente à la génération d'une ligne.

Nous considère le vecteur conditionnel *cond* relatif à la condition $(D_{i*} = k^*)$. Soit $i \in [\![1, N_d]\!]$. Soit $m_i = (m_i^{(k)})_{k=1}^{N_d}$ où $m_i^{(k)}$ vaut 0 sauf si $i = i^*$ et $k = k^*$ où il vaut 1. Ce vecteur conditionnel conditionnel, concaténé à des réalisations de la loi normale centrée réduite regroupée en vecteur sert d'entrée au générateur.

Il est fourni par le générateur conditionnel $\{\hat{d}_1, \ldots, \hat{d}_{N_d}\}$. La fonction de perte est pénalisée en ajoutant la cross-entropy entre $m_{i^*}$ et $\hat{d}_{i^*}$

L'entraînement, appelé entraînement par échantillon, s'effectue de la manière suivante :

- On pose les $N_d$ vecteurs remplis de zéros $m_i = (m_i^{(k)})_{i=1}^{|D_i}$.
- On sélectionne aléatoirement un $i^*$ selon une loi uniforme discrète sur $[\![1, N_d]\!]$, qui est associés à $D_{i^*}$.
- On élabore une fonction de pondération fondée sur la colonne $D_{i^*}$
- On tire $k^*$ de la fonction de pondération précitée
- On constitue le vecteur conditionnel correspondant.

Dans l'implémentation du CTGAN et le calcul des métriques, présentées ci-après, nous avons le package *synthcity* [Zhaozhi *et al.*, 2023].

## 2.3 Métriques

Il y a plusieurs métriques : certaines évaluent la fidélité des données, certaines évaluent la confidentialité.

La première métrique est la p-valeur dans le test de Kolmogorov-Smirnov. Elle est comprise entre 0 et 1. Plus cette métrique est proche de 1, plus les distributions considérées (en particulier, entre données réelles et synthétiques) sont proches.

Une autre mesure de fidélité est la divergence estimée de Kullback-Leibler. Plus cette mesure est proche de 0, plus les distributions (par exemple, entre données réelles et synthétiques) sont proches. Dans notre cas, nous présenterons l'inverse de la distribution de Kullback-Leibler ; dans ce cas, plus la valeur de cet inverse est proche de 0, plus les distributions (par exemple, entre données réelles et synthétiques) sont différentes.

La mesure de la $k$-anonymisation est l'effectif minimum dans les catégories, qui sont reconstituées par l'algorithme des $k$-means.

Ces mesures sont obtenues par le package *synthcity* [Qian *et al.*, 2023].

| | Avatar knn=5 | Avatar knn=10 | Avatar knn=50 | CT-GAN |
|---|---|---|---|---|
| KL-inverse | 0.756/0.972 | 0.856/0.930 | 0.775/0.972 | 0.774/0.759 |
| KS test | 0.927/0.930 | 0.903/0.909 | 0.872/0.930 | 0.846/0.853 |
| $k$-anonymisation | 24 (origine : 12)/ 91 (origin : 12) | 29/119 | 25/91 | 17/52 |

TABLE 1 : Métriques comparant le jeu de données initiales et le jeu de données synthétiques, pour le jeu de donnée sur l'inflammation par le tacrolimus. La première valeur correspond au jeu de données synthétiques de même taille que le jeu de données intiales, la seconde valeur correspond à un jeu de données synthétique qui fait 4 fois la taille des données initiales.

| | Avatar knn=5 | Avatar knn=10 | CT-GAN |
|---|---|---|---|
| KL-inverse | 0.358/0.616 | 0.226/0.350 | 0.364/0.532 |
| KS test | 0.822/0.847 | 0.78/0.802 | 0.713/0.705 |
| $k$-anonymisation | 11 (orig. : 14)/ 3 (orig. : 14) | 10/5 | 11/3 |

TABLE 2 : Métriques comparant le jeu de données initiales et le jeu de données synthétiques, pour le jeu de donnée de pharmacocinétique de population. La première valeur correspond au jeu de données synthétiques de même taille que le jeu de données intiales, la seconde valeur correspond à un jeu de données synthétique qui fait 4 fois la taille des données initiales.

### 2.4 Plan de travail

Les données synthétiques par les algorithmes de génération de données suivants :

- l'algorithme Avatar avec un nombre de plus proches voisins égal à 5
- l'algorithme Avatar avec un nombre de plus proches voisins égal à 10
- l'algorithme Avatar avec un nombre de plus proches voisins égal à 50, pour le jeu de données d'inflammation du tacrolimus uniquement
- CT-GAN

Pour chacun de ces algorithmes de génération de données, deux jeux de données seront synthétisés : un de même taille que le jeu de données original, l'autre qui fait 4 fois la taille du jeu de données original.

## 3 Résultats

Les tables 1 et 2 présentent les métriques de fidélité et de confidentialité relatives respectivement au jeu de données d'inflammation par le tacrolimus et au jeu de données de pharmacocinétique de population.

# 4 Discussion

L'utilisation de deux jeux de données permet de tester l'efficacité de la méthode de synthèse Avatar sur un petit jeu de données (pharmacocinétique) et sur un jeu de données plus conséquent (inflammation du tacrolimus).

Pour le jeu de données d'inflammation (tacrolimus), on constate une amélioration des métriques de confidentialité : la valeur de k-anonymisation est plus grande que la valeur d'origine, quel que soit l'algorithme de synthèse utilisé. Cette amélioration se renforce lorsque l'on génère plus de données. On constate que Avatar est plus performant en termes de confidentialité que le CTGAN. On constate que, sur la valeur de k-anonymisation, le nombre optimal de plus proches voisins utilisé dans Avatar est 10, ce qui montre que le nombre de plus proches voisins est optimisable.

Pour le jeu de données de pharmacocinétique, on constate une détérioration de la valeur de k-anonymisation lorsque l'on utilise Avatar, détérioration qui se confirme lorsque l'on génère plus de données. Cette détéorioration s'explique vraisemblablement par le manque de données initiales ; la dégradation s'observe aussi avec CT-GAN avec approximativement la même intensité. Le nombre optimal de plus proches voisins dans Avatar sur la valeur de k-anonymisation n'est pas clair : d'un point de vue strictement numérique, il dépend de la taille de la l'échantillon généré, mais cette différence ne semble pas pertinente.

En ce qui concerne la fidélité des données synthétiques aux données réelles, les deux métriques utilisées sont : l'inverse de la divergence de Kullback-Leibner et la p-valeur de test de Kolmogorov-Smirnov.

Pour le jeu de données d'inflammation (tacrolimus), du point de vue de la $p$-valeur du test de Kolmogorov-Smirnov, pour une génération de données synthétiques de la taille du jeu de données initiales par Avatar, lorsque le nombre de plus proches voisins diminue, la $p$-valeur augmente, ce qui indique une fidélité des données générées aux données réelles qui s'améliore. Pour une génération de données de taille quatre fois, la taille des données initiale, la $p$-valeur du test de Kolmogorov-Smirnov reste globalement constante, ce qui semble indiquer que la fidélité ne dépend pas du nombre de voisins retenu.

Si on compare, en termes de $p$-valeur du test de Kolmogorov-Smirnov, la fidélité relative aux méthodes Avatar est meilleure que celle relative aux méthodes CT-GAN, ce qui est d'autant plus vrai que la quantité de données synthétiques est grande.

Toujours pour le jeu de données d'inflammation (tacrolimus), en ce qui concerne l'inverse de la divergence de Kullback-Leibner, la méthode Avatar est au moins aussi performante que la méthode CT-GAN (sauf d'un point de vue numérique, dans un cas) ; la fidélité s'améliore, lorsque l'on augmente la taille des données générées, pour les méthodes Avatar ; cette amélioration est plus modeste pour CT-GAN.

Pour le jeu de données de pharmacocinétique, l'augmentation de la taille des données synthétiques améliore la fidélité du point de vue de l'inverse de la divergence de Kullback-Leibner et de la $p$-valeur du test de Kolmogorov-Smirnov. Dans Avatar, l'utilisation de 5 plus proches voisins aboutit à une meilleure fidélité que l'utilisation de 10 plus proches voisins ; et selon ces deux métriques, Avatar est globalement et approximativement aussi performant

que CT-GAN.

Notons qu'il y a probablement, dans le jeu de données de pharmacocinétique, une variabilité des métriques (en particulier de la p-valeur du test de Kolmogorov-Smirnov) du fait de la faible taille des données.

D'une manière générale, la capacité d'anonymisation de Avatar semble nécessiter un volume de données relativement conséquent, ce qui devra être confirmée par d'autres études. Cependant, la génération de données médicales coûte cher et les jeux de données sont par conséquent réduits. Trouver un algorithme qui génère des données synthétiques, à la fois fidèles et anonymisées, reste un défi.

Comme perspective, on peut vérifier que les données synthétiques engendrent des résultats statistiques similaires à ceux obtenus par des données réelles.

# Remerciements

Ce travail fait partie du projet DIGPHAT, qui est soutenu par le gouvernement français, dirigé par l'agence nationale de la recherche (ANR) dans le cadre du programme France 2030 (référence : ANR-22-PESN-0017).

# References

Culnane, C., Rubinstein, B. I., & Teague, V. (2017). Health data in an open world. *arXiv preprint arXiv :1712.05627*.

Guillaudeux, M., Rousseau, O., Petot, J., Bennis, Z., Dein, C. A., Goronflot, T., ... & Gourraud, P. A. (2023). Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis. *NPJ Digital Medicine*, 6(1), 37.

Qian, Z., Cebere, B. C., & van der Schaar, M. (2023). Synthcity : facilitating innovative use cases of synthetic data in different data modalities. *arXiv preprint arXiv :2301.07573*.

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional gan. *Advances in neural information processing systems*, 32.