# Génération de Données Synthétiques Équitables et Préservant la Vie Privée via Autoencodeur Variationnel Basé sur le Clustering et Réseaux Antagonistes Génératifs

Malek Adouani[1], Zaineb Chelly Dagdia[2,3]

[1] Université Paris-Saclay, UVSQ, DAVID
[2] Univ. Lille, CNRS, Inria, Centrale Lille, UMR 9189 CRIStAL
[3] Univ. Lille, Inserm, CHU Lille, U1286 – INFINITE – Institute for Translational Research in Inflammation

26 février 2026

## Résumé

*Le recours à l'apprentissage automatique en santé soulève des enjeux majeurs de biais et de confidentialité. Les modèles génératifs équitables reposent souvent sur des données étiquetées, limitant leur usage en non supervisé, tandis que les approches basées sur la confidentialité différentielle protègent les données sans éliminer nécessairement les biais latents. Les méthodes existantes n'optimisent pas conjointement équité et confidentialité sans supervision explicite. Nous proposons un cadre génératif hybride combinant un VAE avec clustering sous zCDP et un WGAN-GP. Le clustering dans l'espace latent structure les représentations et guide la génération en contexte non supervisé, tandis qu'un Fairness Critic pénalise les corrélations avec les attributs sensibles qui sont des attributs démographiques (par exemple, sexe, origine, âge). Les résultats montrent une réduction efficace des biais, une forte garantie de confidentialité et une bonne utilité des données synthétiques. Cet article correspond à résumé d'un travail publié à ECML PKDD'2025 [1]. L'article original est accessible à l'adresse suivante : https://hal.science/hal-05113907v1/document.*

## Mots-clés

*Réseaux antagonistes génératifs, Apprentissage non supervisé, Atténuation des biais, Préservation de la vie privée.*

## 1 Problématique et état de l'art

L'adoption croissante des modèles d'apprentissage automatique dans des domaines sensibles tels que la santé et la finance soulève des préoccupations majeures liées aux biais algorithmiques et à la confidentialité des données. Les modèles entraînés sur des données biaisées peuvent reproduire, voire amplifier, des inégalités existantes, tandis que la protection des informations individuelles, souvent formalisée par la confidentialité différentielle (DP) constitue une contrainte incontournable. Pourtant, concilier ces deux exigences demeure particulièrement complexe, notamment dans des contextes non supervisés où les données ne sont pas annotées. La littérature sur la génération de données synthétiques s'est principalement structurée autour de deux axes distincts. D'une part, les approches orientées équité, telles que FairGAN et TabFairGAN [4], introduisent des contraintes de parité statistique entre attributs sensibles (par exemple, sexe, origine, âge) afin de limiter les discriminations. Toutefois, ces méthodes reposent sur des annotations explicites pour guider le débiaisage, ce qui restreint leur applicabilité en absence d'étiquettes. D'autre part, les modèles axés sur la confidentialité, comme les DPGANs [3], intègrent des mécanismes de DP via l'injection de bruit dans les gradients ou les données de sorties du modèle. S'ils offrent des garanties formelles de protection de la vie privée, ils ne traitent pas explicitement les biais latents, susceptibles d'être reproduits dans les données synthétiques. Ainsi, les travaux existants répondent généralement à l'un ou l'autre de ces objectifs, mais rarement aux deux simultanément, révélant une lacune dans les cadres non supervisés. Pour y remédier, nous proposons Clust-VAE-WGAN-GP, un modèle génératif hybride combinant une structuration équitable de l'espace latent et des mécanismes formels de confidentialité, afin de produire des données synthétiques à la fois équitables et privées.

## 2 Méthode proposée

Notre modèle, Clust-VAE-WGAN-GP, est une architecture hybride en deux étapes.

**Étape 1 : Autoencodeur Variationnel (VAE) Basé sur le Clustering avec zCDP :** La première étape structure les données d'entrée et garantit la confidentialité ou un VAE encode les données dans un espace latent continu. Directement dans cet espace latent, nous appliquons K-Means pour regrouper les représentations similaires et capturer la structure inhérente des données sans supervision. Les étiquettes de clusters qui en résultent servent de signal conditionnel pour la deuxième étape. Pour la confidentialité, nous intégrons la zero-Concentrated Differential Privacy (zCDP) [2], une variante mathématiquement rigoureuse de la DP, en ajoutant un bruit gaussien calibré aux gradients lors de l'entraînement du VAE. Cette étape produit des représentations latentes structurées et privées.

**Étape 2 : WGAN-GP avec Débiaisage Adversarial** La seconde étape génère les données synthétiques finales. Nous utilisons un réseau antagoniste génératif (GAN) de Wasserstein avec Pénalité de Gradient (WGAN-GP) pour sa stabilité d'entraînement. Le générateur est conditionné par les étiquettes de clusters produites par le VAE, garantissant que les données générées s'alignent sur la structure découverte. L'innovation majeure ici est l'ajout d'un discriminateur d'équité (Fairness Critic, FC), un second discriminateur entraîné spécifiquement pour pénaliser les corrélations entre les données générées et des attributs sensibles prédéfinis. Le générateur est alors entraîné de manière adversariale non seulement pour produire des données réalistes, mais aussi pour produire des données indépendantes des attributs sensibles. Ce mécanisme de débiaisage adversarial impose l'équité de manière dynamique en absences d'annotations.

# 3 Expérimentions et Résultats

Nous avons évalué notre modèle sur six jeux de données du domaine de la santé : HIV, Stress, Obesity, Cardio, Diabetes et Pediatric, et pour chaque base on a fixé les attributs sensibles qui sont genre, origine, et âge. Les performances ont été comparées à des méthodes de référence en matière d'équité (FairGAN, TabFairGAN) et de confidentialité (DPGAN, RDP-CGAN). L'évaluation a porté sur : le réalisme des données, l'équité, et la confidentialité.

## 3.1 Réalisme et Utilité des Données

Sans contraintes de confidentialité, Clust-VAE-WGAN-GP est compétitif par rapport aux méthodes de comparaisons, il obtient les meilleures performances dans quatre des six métriques de réalisme pour les jeux de données Heart et Pediatric. Par exemple, sur la base Pediatric, notre méthode atteint les plus faibles valeurs de Discrépance moyenne maximale (MMD) (0.0010) et de distance de Wasserstein (0.238), et les scores les plus élevés de Score de probabilité par dimension (DWP) (0.5977), $\alpha$-précision (0.741) et $\beta$-rappel (0.712). Cette performance est attribuée à l'inférence variationnelle du VAE et au mécanisme de conditionnement par clustering qui guide la génération.

## 3.2 Équité et Réduction des Biais

L'évaluation de l'équité a montré l'efficacité de notre approche pour atténuer les biais de manière non supervisée. Sur le jeu de données Pediatric, notre modèle a obtenu les écarts d'information mutuelle (MI) les plus faibles entre les clusters et les attributs sensibles (MI-Genre = 0.0010, MI-Âge = 0.0006), indiquant une indépendance quasi-totale. De plus, il a atteint le meilleur Score de Silhouette (0.5297) et le plus faible Indice de Davies-Bouldin (0.5025), confirmant une structure de clusters cohérente et non biaisée. Ces résultats valident l'apport du FC pour pénaliser activement les dépendances statistiques indésirables.

## 3.3 Préservation de la Confidentialité

Le modèle offre de solides garanties de confidentialité, avec un risque d'identifiabilité (EIR) très faible (entre 0.0100 et 0.0221 sur tous les jeux de données). Nous avons également analysé le compromis entre l'équité et la confidentialité. Les résultats montrent que lorsque le budget de confidentialité augmente (moins de bruit, donc moins de protection), les métriques d'équité comme le Score de Silhouette s'améliorent, indiquant que le débiaisage est plus efficace lorsque le signal des données n'est pas excessivement masqué par le bruit de la DP.

# 4 Conclusion

Nous avons présenté Clust-VAE-WGAN-GP, une nouvelle approche générative hybride qui impose simultanément l'équité et la confidentialité pour la génération de données synthétiques, en absence de label. En combinant un VAE basé sur le clustering avec zCDP et un WGAN-GP doté d'un débiaisage adversarial, notre approche parvient à générer des données de haute fidélité, équitables et privées. Les expérimentations sur six jeux de données démontrent sa supériorité par rapport aux modèles existants. Ce travail ouvre la voie à une génération de données plus éthique et fiable dans les domaines sensibles. Comme perspectives, nous envisageons d'explorer des techniques de clustering adaptatif pour améliorer davantage l'atténuation des biais.

# Remerciements

Ce travail est financé et soutenu par : le programme Horizon Europe de l'Union européenne (MSCA, convention n° 101236749), les financements France 2030 RHU RECORDS (ANR-18-RHUS-0004), IHU SEPSIS (ANR-23-IAHU-0004), iRECORDS financé par ERA PerMed (JTC_2021), le Programme d'Investissements d'Avenir (I-SITE ULNE / ANR-16-IDEX-0004 ULNE), géré par l'Agence Nationale de la Recherche (n° I-KUL-22-005-ARCHIE-INFINITE), ainsi que d'un financement de l'Inserm et du Ministère de la Santé via l'appel MESSIDORE 2023 opéré par l'IReSP (AAP-2023-MSDR-341423).

# Références

[1] Malek Adouani and Zaineb Chelly Dagdia. Fair and Privacy-preserving Synthetic Data Generation via Clustering-based Variational Autoencoder and Adversarially Debiased Wasserstein Generative Adversarial Networks with Gradient Penalty. In *ECMLPKDD 2025*, September 2025.

[2] M. Bun and T. Steinke. Concentrated differential privacy : Simplifications, extensions, and lower bounds. *arXiv preprint*, arXiv :1605.02065, 2016.

[3] G. O. Ghosheh, J. Li, and T. Zhu. A survey of generative adversarial networks for synthesizing structured electronic health records. *ACM Computing Surveys*, 55(8), 2023.

[4] A. Rajabi and O. O. Garibay. Tabfairgan : Fair tabular data generation with generative adversarial networks. *arXiv preprint*, arXiv :2109.00666, 2021.