médecine/sciences 2024 ; 40 : 661-4

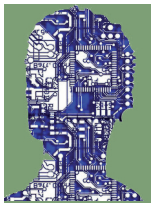
Données synthétiques en médecine : génération, évaluation et limites

Alaedine Benani[1,2], Julien Vibert[3], Stanislas Demuth[4,5]

Les récentes avancées dans le traitement et l'analyse des données ont transformé les performances des algorithmes d'aide à la décision médicale, qui jouent ainsi un rôle croissant dans la stratification des risques, la prévention, le diagnostic, le traitement et le suivi de nombreuses maladies [1].

Cependant, obtenir une base de données de qualité, complète, standardisée, volumineuse, multimodale, longitudinale, représentative, diversifiée et indemne de biais reste un défi majeur [2]. En effet, la mise en place de telles bases de données est particulièrement difficile : la multimodalité[1] demande de nombreux examens pour le même patient, le suivi longitudinal est très difficile en pratique, l'annotation[2] est coûteuse, la reproductibilité inter-observateur de la collecte est imparfaite, etc. De plus, il existe de multiples domaines où les données sont soit rares, soit difficiles à collecter. Enfin, du fait des risques notamment de réidentification des patients, les contraintes réglementaires et éthiques autour de l'utilisation des données personnelles de santé, particulièrement sensibles, compliquent encore leur accès et leur utilisation.

Dans ce contexte, les données synthétiques émergent comme une solution prometteuse [3]. Générées par diverses approches statistiques spécifiques (plutôt que collectées sur un sujet physique), elles imitent des données empiriques sans toutefois correspondre à



des individus existants et proposent donc une alternative pour surmonter les limites associées à ces dernières.

Nous explorons, ici, les enjeux, les méthodes de génération et de validation, et les applications des données synthétiques dans l'entraînement d'algorithmes d'apprentissage automatisé en médecine.

[1]Service de médecine vasculaire, hôpital européen Georges Pompidou (HEGP), AP-HP, Université Paris-Cité, Paris, France.
[2]Zoï, Paris, France.
[3]Département d'innovations thérapeutiques et essais précoces (DITEP), Inserm U981, Gustave Roussy, Villejuif, Paris, France.
[4]Inserm U1064, CR2TI — Centre de recherche en transplantation et immunologie translationnelle, Nantes Université, Nantes, France.
[5]Inserm CIC 1434, Centre d'investigation clinique, Centre hospitalier de Strasbourg, France.
alaedine.benani@aphp.fr
stanislas.demuth@inserm.fr
julien.vibert@gustaveroussy.fr

Intérêts des données synthétiques

Comment sont générées les données synthétiques ?

Les données synthétiques sont artificiellement générées, contrairement aux données empiriques issues de vrais patients. Elles imitent les propriétés statistiques des données réelles sans correspondre à des individus spécifiques. Elles sont conçues pour être indiscernables de données réelles par des algorithmes d'analyse et de traitement, pour répondre au principe d'utilité, sans divulguer des informations personnelles ou sensibles pouvant mener à l'identification d'individus pour se conformer au principe de confidentialité (*Figure 1*).

Vignette (© Lightwise/123RF).

[1] Utilisation concomitante de plusieurs types de données (images, vidéos, texte libre, texte structuré, etc.) en même temps pour l'entraînement d'un modèle d'apprentissage statistique.

[2] Le fait, pour un expert, d'assigner un label à une donnée (par exemple, définir un examen par scanner comme étant normal ou pathologique).

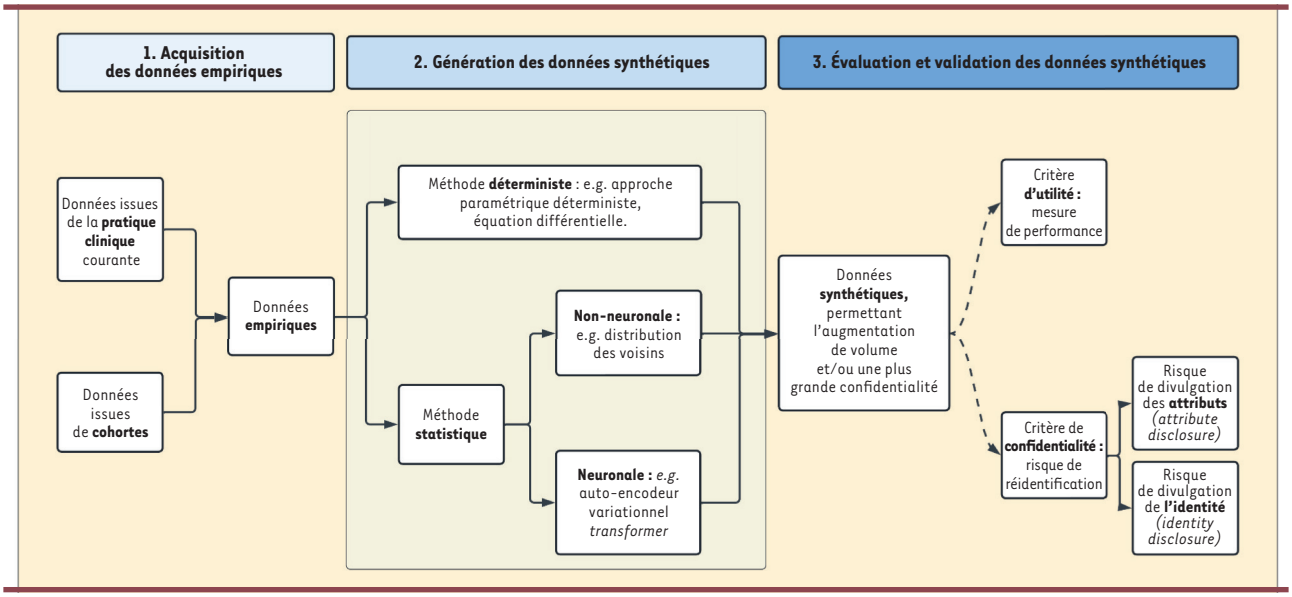


Figure 1. Acquérir les données empiriques, générer les données synthétiques puis évaluer ces dernières sont les trois grandes étapes permettant d'aboutir à un jeu de données synthétiques utilisable.

Les données synthétiques sont produites à partir de données réelles grâce à des modèles mathématiques (transformations simples, approches paramétriques, statistiques, apprentissage, réseaux de neurones, etc.). La méthode la plus couramment utilisée débute avec un jeu de données réelles utilisé pour générer un ou plusieurs autres jeux de données dans le but d'augmenter le volume de l'échantillon, de suréchantillonner des classes minoritaires, ou d'anonymiser des données [4]. La génération de données synthétiques peut se faire par des méthodes déterministes ou des méthodes statistiques.

Les méthodes déterministes utilisent des approches paramétriques pour reproduire les données. Ces modèles sont fondés sur des équations différentielles contenant une quantité finie d'inconnues. Elles ne contiennent pas de termes aléatoires, elles ne fonctionnent pas par apprentissage.

Les méthodes statistiques, quant à elles, peuvent être divisées en deux grandes catégories en fonction de leur utilisation ou non de réseaux de neurones.

Les techniques non neuronales, comme les inférences fondées sur des simulations et inférences amorties[3], les méthodes de type estimation d'atlas[4] [5], et la génération fondée sur la distribution des voisins[5], utilisent des méthodes statistiques pour synthétiser les données sans recourir à l'apprentissage profond.

L'apprentissage profond, qui se développe en parallèle de ces techniques, permet l'utilisation de technologies comme les auto-encodeurs variationnels[6], les réseaux antagonistes génératifs[7], les *transformers*[8] et la diffusion[9]. Ces modèles sont particulièrement utiles pour créer des données non structurées, telles que des images médicales.

Comment sont évaluées les données synthétiques ?

La validité des données synthétiques est généralement évaluée selon deux groupes de critères principaux : l'utilité [6] et la confidentialité [7]. Pour chacun d'eux, il existe de multiples métriques et indicateurs, dont il convient d'évaluer la pertinence en fonction du problème posé et du type d'algorithme que l'on souhaite entraîner.

L'utilité permet d'évaluer si les résultats de l'analyse de données sont similaires. Dans le cas d'un modèle d'apprentissage automatique entraîné sur un jeu augmenté par des données synthétiques, on peut même rechercher une augmentation de performances prédictives sur un jeu de validation réel. Cela implique de s'assurer que les

[3] Les inférences amorties désignent une méthode de calcul statistique utilisée pour estimer des paramètres dans des modèles complexes.

[4] Les méthodes de type estimation d'atlas désignent des techniques statistiques utilisées pour créer une référence standardisée (ou atlas) à partir de multiples jeux de données. Cet atlas représente une sorte de moyenne qui capture les caractéristiques essentielles des données d'origine.

[5] La méthode des proches voisins est une méthode d'apprentissage supervisé simple qui utilise la proximité entre les points de données pour classer ou prédire la catégorie d'un point de données.

[6] Les auto-encodeurs variationnels sont un type de réseau de neurones utilisé pour générer des données synthétiques. Le processus d'échantillonnage permet de générer de nouvelles données en créant des variations qui ressemblent aux données d'origine, mais sans correspondre exactement à des exemples spécifiques.

[7] Les réseaux antagonistes génératifs sont une classe de modèles d'apprentissage profond utilisés pour générer des données synthétiques. Ils consistent en deux réseaux de neurones qui s'entraînent ensemble dans un cadre compétitif.

[8] Les *transformers* sont une architecture de réseau de neurones introduite initialement pour le traitement du langage naturel, mais qui s'est avérée polyvalente et est utilisée dans diverses applications, y compris la génération de données synthétiques.

[9] Les modèles de diffusion génèrent des données en apprenant à inverser un processus de bruitage. Ces modèles sont particulièrement efficaces pour créer des données non structurées, comme des images médicales, en partant de bruit aléatoire et en produisant des échantillons réalistes et de haute qualité.

caractéristiques essentielles des données originales sont captées et reproduites. Les critères précis (aire sous la courbe, précision, erreur quadratique moyenne, etc.) dépendent du type d'algorithme et de la tâche souhaitée (classification ou régression).

La confidentialité est un second critère d'évaluation. Il est en effet important que les données synthétiques générées ne permettent pas de remonter aux individus à l'origine des données réelles utilisées. Les techniques de génération doivent donc être conçues pour optimiser le fait que les données synthétiques, tout en étant informatives, ne compromettent pas l'identité des personnes. Deux grands groupes de risques sont à évaluer : le risque de divulgation de l'identité (*identity disclosure*) et le risque de divulgation des attributs (*attribute disclosure*) [8].

Limites des données synthétiques

Ces dernières années, l'utilisation de données synthétiques pour des algorithmes en médecine est croissante. Elle aide à évaluer les politiques de santé publique, à améliorer l'efficacité de traitements, ou à augmenter les performances des algorithmes d'apprentissage automatique [9]. Un exemple, parmi tant d'autres [9], est la génération, par un réseau antagoniste génératif, de radiographies thoraciques de faux patients atteints de la Covid-19 (*coronavirus disease 2019*) afin d'entraîner un modèle qui détecte la maladie [10]. Cette approche présente donc un réel intérêt en termes de confidentialité et d'utilité. Cependant, elle comporte des limites.

Deux types de risques sont à distinguer : ceux inhérents aux données synthétiques et ceux liés à la propagation des défauts du jeu de données.

Pour les risques inhérents aux données synthétiques, il est important de noter que les métriques d'utilité et de confidentialité peuvent être modifiées si l'usage des données synthétiques est modifié après leur génération. Bien que ces approches soient bien étudiées pour les données tabulaires[10], elles sont encore émergentes pour les données longitudinales et non-structurées (images, vidéos, signaux). De plus, il n'existe pas de consensus clair sur les critères de confidentialité acceptables ni sur les performances minimales à atteindre. Enfin, le statut juridique, les réflexions éthiques et les sujets de propriété intellectuelle liés aux données synthétiques sont ouverts à débats.

Pour les risques liés à la propagation des défauts, les données synthétiques peuvent intégrer des biais si les modèles ne sont pas correctement ajustés ou si les données originales contiennent des biais qui n'ont pas été détectés, ce qui altère les résultats des algorithmes entraînés. Il est donc essentiel d'identifier et de mesurer ces biais (analyse de variance, tests d'ajustement, diversification des sources, rééquilibrage des classes). L'expertise clinique peut également jouer un rôle. Il existe aussi un risque de perte d'information si les modèles ne capturent pas toutes les subtilités des données réelles, ce qui peut entraîner des corrélations incorrectes. Des techniques de calibration peuvent réduire ce risque. Enfin, l'utilisation de réseaux de neurones pour générer des données synthétiques peut entraîner un surapprentissage. Dans la génération de données non structurées, comme les images, les artefacts et distorsions peuvent altérer leur utilité, même s'ils sont minimes.

Conclusion

Face aux défis imposés par la nécessité de disposer de bases de données médicales optimales, les données synthétiques émergent comme une solution prometteuse. Elles pourraient représenter une alternative intéressante aux données réelles. En plus de contourner les obstacles de la collecte de données réelles, elles peuvent également répondre aux enjeux éthiques et réglementaires. De nombreuses études ont permis de montrer l'intérêt de ces approches dans l'entraînement d'algorithmes d'apprentissage automatisé en médecine, que ce soit sur des critères d'utilité ou de confidentialité.

Les limites de ces approches tempèrent néanmoins leur utilisation. Outre les risques de perte d'information, d'introduction ou d'exacerbation de biais, ou encore de surapprentissage, se pose la question de leur intérêt comparativement à d'autres méthodes comme l'apprentissage fédéré[11].

Enfin, en médecine, la multimodalité et le suivi longitudinal sont cruciaux pour comprendre et traiter efficacement les maladies. Ces deux caractéristiques ne sont pas, pour l'instant, prises en compte efficacement par ces approches, mais la recherche progresse en ce sens. ◊

SUMMARY

Synthetic data in medicine: Generation, evaluation and limits

Recent technological advances in data science hold great promise in medicine. Large-sized high-quality datasets are essential but often difficult to obtain due to privacy, cost, and practical challenges. Here, we discuss synthetic data's generation, evaluation, and regulation, highlighting its current applications and limits. ◊

REMERCIEMENTS

Les auteurs remercient Pr Xavier Tannier, Pr Emmanuel Messas, Pr Pierre-Antoine Gourraud, Dr Pierre Bauvin, Dr Stéphane Ohayon et Dr Sylvain Bodard.

[10] Les données tabulaires sont des données organisées en tableaux structurés, sous forme de lignes et de colonnes, similaires à ceux que l'on trouve dans des feuilles de calcul. Ces données structurées sont ainsi plus faciles à manipuler.

[11] L'apprentissage fédéré consiste à répartir la tâche d'entraînement d'un algorithme sur plusieurs machines.

LIENS D'INTÉRÊT

Alaedine Benani soutient un projet de recherche financé par Zoï, Paris, France. Stanislas Demuth et Julien Vibert déclarent n'avoir aucun lien d'intérêt concernant les données publiées dans cet article.

RÉFÉRENCES

1. Wang Y, Li N, Chen L, *et al.* Guidelines, Consensus Statements, and Standards for the Use of Artificial Intelligence in Medicine: Systematic Review. *J Med Internet Res* 2023 ; 25:e46089.
2. Shilo S, Rossman H, Segal E. Axes of a revolution: challenges and promises of big data in healthcare. *Nat Med* 2020 ; 26 : 29-38.
3. Allassonnière S, Fraysse JL. *Données de santé artificielles : analyse et pistes de réflexion*. Livre Blanc, 2024
4. Shorten C, Khoshgoftaar TM. A survey on Image Data Augmentation for Deep Learning. *J Big Data* 2019 ; 6 : 60.
5. Koval I, Bône A, Louis M, *et al.* AD Course Map charts Alzheimer's disease progression. *Sci Rep* 2021 ; 11 : 8020.
6. El Emam K, Mosquera L, Fang X, *et al.* Utility Metrics for Evaluating Synthetic Health Data Generation Methods: Validation Study. *JMIR Med Inform* 2022 ; 10:e35734.
7. El Emam K, Mosquera L, Fang X. Validating a membership disclosure metric for synthetic health data. *JAMIA Open* 2022 ; 5:ooac083.
8. Goncalves A, Ray P, Soper B, et al. Generation and evaluation of synthetic patient data. *BMC Med Res Methodol* 2020 ; 20 : 108.
9. Chen RJ, Lu MY, Chen TY, *et al.* Synthetic data in machine learning for medicine and healthcare. *Nat Biomed Eng* 2021 ; 5 : 493-7.
10. Gulakala R, Markert B, Stoffel M. Generative adversarial network based data augmentation for CNN based detection of Covid-19. *Sci Rep* 2022 ; 12 : 19186.

TIRÉS À PART

A. Benani