REVUE RAPIDE SUR LA MÉTHODE AVATAR POUR L'ANONYMISATION DES DONNÉES EN CONTEXTE CLINIQUE

ÉTAT DES CONNAISSANCES

rédigé par
Marie-Pier Bouchard

en collaboration avec
Ariane Brault

Unité d'évaluation des technologies et des modes d'intervention en santé et en services sociaux, 2026

Centre intégré universitaire de santé et de services sociaux de l'Estrie – Centre hospitalier universitaire de Sherbrooke

Québec

Unité d'évaluation des technologies et
des modes d'intervention en santé
et en services sociaux

Centre intégré
universitaire de santé
et de services sociaux
de l'Estrie – Centre
hospitalier universitaire
de Sherbrooke
Québec

REVUE RAPIDE SUR LA MÉTHODE AVATAR POUR L'ANONYMISATION DES DONNÉES EN CONTEXTE CLINIQUE

ÉTAT DES CONNAISSANCES

rédigé par
Marie-Pier Bouchard

en collaboration avec
Ariane Brault

JANVIER 2026

LA MISSION

Soutenir la prise de décision des gestionnaires par diverses approches évaluatives, des technologies, des modes d'intervention, des programmes en santé, en santé publique et en services sociaux et une évaluation des interventions afin d'améliorer la santé et le bien-être de la population estrienne. L'UETMISSS fonde ses travaux sur l'évaluation rigoureuse des données scientifiques, contextuelles et des savoirs expérientiels, ces derniers provenant des usagères et usagers, de leurs proches, de la population et de l'ensemble de la communauté du CIUSSS de l'Estrie – CHUS*.

**Intervenantes et intervenants, professionnelles et professionnels, gestionnaires*

UNITÉ D'ÉVALUATION DES TECHNOLOGIES ET DES MODES D'INTERVENTION EN SANTÉ ET EN SERVICES SOCIAUX, CIUSSS DE L'ESTRIE – CHUS

Marie-Pier Bouchard, M. Sc., M.B.A.
Conseillère en évaluation à l'UETMISSS

Ariane Brault, Ph. D.
Conseillère en évaluation à l'UETMISSS

Lucien Coulibaly, Ph. D.
Conseiller en évaluation à l'UETMISSS

Cyrille Gérard Diffo, M.D., M. Sc.
Conseiller en évaluation à l'UETMISSS

Pierre Dagenais, M.D., Ph. D.
Médecin-conseil à l'UETMISSS

Marie-Andrée Roy, M. Sc.
Cheffe de service – Mobilisation des connaissances

Julien Desautels, ps.éd., Ph. D.
Coordonnateur – Enseignement et mobilisation des connaissances

Sonia Ouellet
Agente administrative cl. 1

Dépôt légal — Bibliothèque et Archives nationales du Québec, 2026
ISBN 978-2-555-03020-6 (PDF)

Pour tout renseignement sur ce document ou sur les activités de l'UETMISSS, CIUSSS de l'Estrie – CHUS, s'adresser à :
Unité d'évaluation des technologies et des modes d'intervention en santé et en services sociaux
Centre intégré universitaire de santé et de services sociaux de l'Estrie – Centre hospitalier universitaire de Sherbrooke – Hôpital et centre d'hébergement d'Youville
1036, rue Belvédère Sud
Sherbrooke (Québec) J1H 4C4
Téléphone : (819) 780-2220, poste 16648
Courriel : UETMISSS.ciussse-chus@ssss.gouv.qc.ca

Pour citer ce document : Unité d'évaluation des technologies et des modes d'intervention en santé et en services sociaux du CIUSSS de l'Estrie – CHUS (UETMISSS, CIUSSS de l'Estrie – CHUS), Revue rapide sur la méthode avatar pour l'anonymisation des données en contexte clinique : État des connaissances préparé par Marie-Pier Bouchard et Ariane Brault [En ligne], Sherbrooke, Québec (Canada), Janvier 2026, xvii, 51 p.

ÉQUIPE DE PROJET

RÉDACTION	
Marie-Pier Bouchard	Conseillère en évaluation, Direction de la coordination de la mission universitaire (DCMU), Centre intégré universitaire de santé et de services sociaux de l'Estrie – Centre hospitalier universitaire de Sherbrooke (CIUSSS de l'Estrie – CHUS)

CONTRIBUTION AU PROJET	
Ariane Brault	Conseillère en évaluation, DCMU, CIUSSS de l'Estrie – CHUS

ORIENTATION ET RÉVISION SCIENTIFIQUE	
Julien Desautels	Coordonnateur, Enseignement et Mobilisation des connaissances, DCMU, CIUSSS de l'Estrie – CHUS
Marie-Andrée Roy	Cheffe de service, Mobilisation des connaissances, DCMU, CIUSSS de l'Estrie – CHUS

APPROBATION FINALE	
Marie-Andrée Roy	Cheffe de service, Mobilisation des connaissances, DCMU, CIUSSS de l'Estrie – CHUS

AUTRES PERSONNES AYANT CONTRIBUÉ À LA CONTEXTUALISATION	
Becky Opara	Coordonnatrice du Pôle universitaire de santé numérique de l'Estrie, Faculté de médecine et des sciences de la santé - Recherche et études supérieures (FMSS), Université de Sherbrooke

REMERCIEMENTS

L'UETMISSS tient à remercier toutes les personnes ayant contribué, d'une façon ou d'une autre, à la réalisation du présent rapport. Elle remercie particulièrement M. François Rheault, professeur adjoint au département d'informatique, de l'Université de Sherbrooke pour la relecture du rapport.

DIVULGATION DE CONFLIT D'INTÉRÊTS

Aucun conflit à signaler

FINANCEMENT

Ce projet a été financé à même le budget de fonctionnement de l'UETMISSS

RAPPORT PRÉLIMINAIRE

Les résultats de ce projet ont été présentés à l'équipe demanderesse en septembre 2025.

RÉSUMÉ

Contexte – Les données de santé soutiennent l'évolution des connaissances et la prestation des soins. Leur accès est néanmoins restreint en vertu des obligations légales et éthiques visant à protéger les renseignements personnels qu'elles contiennent. En Estrie, de récents travaux de recherche ont d'ailleurs mis en lumière la contribution potentielle de la méthode avatar en tant que levier d'innovation pour l'utilisation et le partage des données de santé. Cette méthode, aussi connu sous le nom d'avatarisation des données, est une technique de génération de données synthétiques brevetée par l'entreprise *Octopize Mimethik Data*.

Puisqu'il s'agit d'une méthode récente et d'une innovation de rupture dans son domaine, peu d'études sont disponibles. Un projet pilote est prévu pour tester cette solution technologique et réaliser une évaluation organisationnelle et éthique en contexte québécois.

Objectif – Un état des connaissances préalable est requis pour soutenir l'appréciation du potentiel de valeur de la technologie d'*Octopize Mimethik Data* au CIUSSS de l'Estrie - CHUS et les évaluations prévues. Plus précisément, il vise à (a) établir le portrait de la technologie, notamment ses caractéristiques et sa performance, (b) repérer les considérations organisationnelles, contextuelles et éthiques affectant son utilisation et (c) les indicateurs pertinents pour une évaluation organisationnelle et économique de l'avatarisation des données.

Méthodologie – Une recension des écrits de type revue rapide a été réalisée pour les publications sur la méthode avatar dans le domaine de la santé au cours des 5 dernières années. Le processus de sélection des publications s'est déroulé par une double sélection indépendante des titres et des résumés ainsi qu'une lecture complète des articles. Les conflits ont été résolus par discussion jusqu'à l'atteinte d'un consensus à chaque étape du processus. Les résultats d'intérêts ont été extraits à l'aide d'une grille préétablie par la responsable de l'évaluation. En cas de doute, une seconde personne a été consultée.

Les résultats de la recension ont été présenté de manière narrative. Le présent document a fait l'objet d'une validation externe (relecture) par une personne experte externe à l'établissement.

Résultats –

Recherche documentaire

Après le retrait des doublons, ce sont 193 publications qui ont été soumises au processus de sélection. Trente-neuf ont été sélectionnées pour une lecture intégrale Au final, douze publications correspondaient aux critères de sélection et ont été inclus dans la recension des écrits.

La majorité des publications étaient d'origine françaises et quelques-unes canadiennes. Onze d'entre elles présentaient un devis évaluatif pour la méthode. La moitié des publications incluses, provenant toutes de la littérature grise, n'a pas été révisée par les pairs au moment de cet état des connaissances. Les publications reposent sur un relativement faible nombre de groupes de recherche et cinq d'entre elles impliquent du personnel d'Octopize dans les autrices et des auteurs.

Description de la méthode

La méthode avatar est un algorithme de génération de données synthétiques utilisant une approche centrée sur la personne, préservant la taille et la nature des données originales, par la simulation d'un avatar unique pour chaque individu.

Elle requiert un jeu de données tabulaires pseudoanonymisées où les rangées et les colonnes représentent respectivement les individus et les variables. Les variables employées peuvent être de nature continue, catégorielle, booléenne ou temporelle sous la forme de date.

Spécifiquement conçue pour le domaine de la santé, la méthode avatar comporte cinq étapes clés :

1) Projection des données dans un espace multidimensionnel (étape de transformation des données réduisant leur complexité)
2) Identification des k-voisins plus proches qui permet de calculer la entre les voisins et de définir une zone locale autour de chaque individu.
3) Génération pseudo-aléatoire des avatars par une simulation unique pour chaque zone locale précédemment créée.
4) Inversion de la projection pour reconvertir les avatars dans le format des données d'origine.
5) Évaluation des paramètres de protection de vie privée et d'utilité

La méthode est principalement contrôlée par les paramètres affectant la zone locale (distance mathématique utilisée, nombre de voisin (k), paramètres de la projection) ou par ceux influençant le caractère aléatoire de la génération d'un avatar. La taille de l'échantillon généré et le mode de gestion de données manquantes peuvent, quant à eux, augmenter ou compléter les données.

Performance de la méthode

Règle générale, la génération de données synthétiques par la méthode avatar tend à (a) centraliser et lisser les données, (b) retenir les associations ou les effets primaires ou les associations fortes, mais moins les effets secondaires et (c) accentuer les déséquilibres de classes.

La protection de la vie privée et l'utilité des jeux de données avatarisées sont dépendante du paramétrage de la méthode utilisée. De plus, l'utilité spécifique d'un jeu de données synthétiques est limitée aux analyses réalisées démontrant la rétention de certaines informations.

L'avatarisation peut être optimisée selon l'utilisation souhaitée par la recherche d'un équilibre entre l'utilité et la protection de la vie privée. Parmi les paramètres contrôlant la méthode, le paramètre k est le plus fréquemment employé. Un faible k préserve la précision et tend à surestimer l'effet, tandis qu'un k élevé augmente la confidentialité et tend à sous-estimer l'effet. Une faible valeur de ncp augmente l'utilité sans trop influencer la protection de la vie privée. Le poids et l'encodage différentiel de certaines variables peuvent aussi être utilisés pour optimiser le jeu de données générées. L'augmentation des données améliorent certaines mesures d'utilité mais a tendance à introduire des associations de type faux positif.

Finalement, la méthode avatar offre, avec un ajustement adéquat de ces paramètres, une performance similaire ou supérieure aux approches les plus populaires telles que CT-GAN, Synthpop et TVAE.

Considérants contextuels et organisationnels
Les considérants suivants sont importants pour encadrer l'utilisation et l'implantation de la génération de données synthétiques : (a) le cadre juridique et réglementaire des données synthétiques, (b) l'importance de l'évaluation des risques, (c) l'équilibre utilité et protection de la vie privée ainsi que (d) les perspectives et les limites concernant le partage des jeux de données synthétiques.

Limites
Les principales limites de cet état des connaissances sont l'absence de l'évaluation de la qualité des publications, la faiblesse de données économiques et organisationnelles dans la littérature, le faible nombre de groupes ayant publiés sur le sujet. La présence de personnel d'Octopize parmi les auteurs et autrices de près de la moitié des publications et l'absence d'une révision par les pairs de plusieurs publications au moment de la rédaction de cette revue sont deux autres éléments à considérer. Un élargissement de la recherche documentaire à la génération des données synthétiques dans des domaines connexes serait pertinent pour obtenir un portrait plus complet des indicateurs et des considérants.

Par ailleurs, les écrits législatifs régissant l'utilisation et la protection des données ou renseignements personnels en vigueur au Québec et au Canada étaient hors de la portée de cet état des connaissances.

Conclusion

À la condition d'un paramétrage adéquat, l'avatarisation des données semble être une alternative intéressante à l'anonymisation pour favoriser l'accès et l'utilisation des données de santé. Par contre, l'utilité des données synthétisées est dépendante de l'évaluation de ses indicateurs ce qui implique que toute autre analyse que celle pour laquelle le jeu de données a été optimisée peut conduire à des analyses erronées.

Des études supplémentaires sont nécessaires pour évaluer les dimensions éthiques, juridiques, économiques et organisationnelles de l'implantation et de l'utilisation de la technologie d'avatarisation des données par Octopize.

SUMMARY

Notice: The English translation of this summary was assisted by artificial intelligence using https://www.deepl.com/en/translator.

Context – Health data supports the advancement of knowledge and the delivery of care. However, data access is restricted by legal and ethical obligations to protect the personal information. In Estrie, recent research has highlighted the potential of the avatar method to facilitate health data sharing and utilization. This method, also known as data avatarization, is a synthetic data generation technique patented by Octopize Mimethik Data.

Since the avatar method is a recent and a disruptive innovation in its field, few studies are available. A pilot project is planned to test this technological solution and conduct an organizational and ethical assessment in Quebec.

Objective – Establishing the current state of knowledge is a prerequisite for assessing the potential value of Octopize Mimethik Data's technology for the CIUSSS de l'Estrie - CHUS and the further evaluations. More specifically, it aims to (a) present a comprehensive overview of the technology, including its characteristics and performance, (b) identify organizational, contextual, and ethical considerations regarding its use, and (c) identify relevant indicators for an organizational and economic assessment of the avatar method.

Methodology – A rapid review of the literature was conducted for publications on the avatar method in health over the past five years. The publication selection process involved an independent dual review of titles and abstracts, followed by a full reading of the articles. Conflicts were resolved through discussion. Results of interest were extracted using a grid pre-established by the main evaluator. In doubt, a second evaluator was consulted.

The results of the review were presented in narrative form and subjected to external validation (proofreading) by an expert external to the CIUSSS de l'Estrie – CHUS.

Results –

Documentary research

After removing duplicates, 193 publications were submitted to the selection process. Thirty-nine were selected for full reading. Ultimately, twelve publications met the selection criteria and were included in the literature review.

The majority of publications were French, with a few Canadian ones. Eleven had an evaluative research protocol to study the method. Half of the publications included, all from gray literature, had not been peer-reviewed at the time of this review. The publications are based on a relatively small number of research groups, and five of them included Octopize staff among the authors.

Method description

The avatar method is an algorithm for generating synthetic data using a person-centered approach, preserving the size and nature of the original data by simulating a unique avatar for each individual.

It requires a pseudo-anonymized tabular dataset where rows and columns represent individuals and variables, respectively. The variables can be continuous, categorical, Boolean, or dates.

Specifically designed for the healthcare domain, the avatar method consists of five key steps:

1) Data Projection into a multidimensional space (data transformation step reducing complexity)
2) Identification of the k-nearest neighbors, which allows the calculation of the distance between neighbors and defines a local area around each individual
3) Pseudo-random generation of local avatars
4) Inversion of the projection to return to the original encoding
5) Evaluation of privacy and utility parameters

The method is mainly controlled by parameters affecting the local area (mathematical distance, number of neighbors (k), projection parameters) or by those influencing the randomness of avatar generation. The size of the generated sample and the management of missing data can increase or supplement the data.

Method performance

The generation of synthetic data using the avatar method tends to (a) centralize and smooth data, (b) retain primary and strong associations or effects, but secondary effects to a lesser extent, and (c) accentuate class imbalances.

The privacy protection and utility of avatar datasets depend on the parameter's configuration of the method. Furthermore, the specific usefulness of a synthetic dataset is limited to analyses demonstrating the retention of certain information.

The method can be optimized to achieve the desired use by seeking a balance between utility and privacy protection. Among the parameters controlling the method, the k parameter is the most frequently used. A low k preserves accuracy and tends to overestimate the effect, while a high k increases confidentiality and tends to underestimate the effect. A low value of the number of projection components (ncp) increases utility without significantly influencing privacy protection. The weighting and differential encoding of certain variables can also be used to optimize the avatar dataset. Increasing the amount of data improves certain utility measures but tends to introduce false positive associations.

Finally, with appropriate adjustment of these parameters, the avatar method offers performance similar or superior to the most popular approaches such as conditional tabular generative adversarial network (CT-GAN), Synthpop, and tabular variational autoencoders (TVAE).

Contextual and organizational considerations

The following aspects are important to consider in order to adequately use or to implement synthetic data generation: (a) the legal and regulatory framework for synthetic data, (b) the risk assessment, (c) the balance between utility and privacy protection, and (d) the prospects and limitations of sharing synthetic datasets.

Limitations

The main limitations of this state of knowledge are the lack of evaluation of the quality of publications, the scarcity of economic and organizational data in the literature, and the small number of groups that have published on the subject. Two other factors are to be considered: (a) nearly half of the publications have Octopize staff members listed among the authors and, (b) the lack of peer-reviewing of several publications at the time of writing this review. Expanding the literature review to include synthetic data generation in related fields would be relevant to obtain a more complete picture of the indicators and considerations. Furthermore, the Quebecer and Canadian regulation and laws regarding the personal information use and protection were beyond the scope of this review.

Conclusions

Provided that it is properly configured, the avatar method appears to be an interesting alternative to anonymization for promoting health data sharing and utilization. However, the utility of synthetic data depends of its indicators' assessment, which means that any analysis other than that for which the dataset has been optimized may lead to erroneous analysis.

Further studies are needed to assess the ethical, legal, economic, and organizational dimensions for the implementation and utilization of Octopize's data avatarization technology.

TABLE DES MATIÈRES

1. INTRODUCTION 1

1.1 CONTEXTE 1
1.2 PROBLÉMATIQUE 2
1.3 ENJEUX ET BESOINS DÉCISIONNELS 2
1.4 QUESTIONS DÉCISIONNELLES 2

2. MÉTHODOLOGIE 3

2.1 QUESTIONS D'ÉVALUATION 3
2.2 MODÈLE LOGIQUE ET CADRE D'ANALYSE 3
2.2.1 Concepts clés et indicateurs 4
2.2.2 Examen des facteurs organisationnels et contextuels 5
2.3 RECENSION DES ÉCRITS 5
2.3.1 STRATÉGIE DE RECHERCHE 5
2.3.2 SÉLECTION DES ÉTUDES 6
2.4 ENJEUX D'ÉTHIQUE ET D'ÉQUITÉ 7
2.5 VALIDATION 7

3. RÉSULTATS 9

3.1 RECENSION DES ÉCRITS 9
3.1.1 Résultats de la recherche de la littérature 9
3.1.2 Description des articles inclus 9
3.1.3 Synthèse et interprétation des données probantes 10
3.2 CONSIDÉRANTS CONTEXTUELS ET ORGANISATIONNELS 17
3.3 PERSPECTIVES DES PARTIES PRENANTES 19

4. DISCUSSION 21

PERSPECTIVES D'UTILISATION 22
CONSIDÉRATIONS ÉTHIQUES À LA GÉNÉRATION ET L'UTILISATION 22
LIMITES 23

5. CONCLUSION 25

ANNEXES 27

BIBLIOGRAPHIE 49

LISTE DES TABLEAUX

Tableau I. Grille CHIP 5
Tableau II. Critères de sélection pour la recension des écrits 6
Tableau III. Résultats de protection de la vie privée de la méthode dans les publications incluses 13
Tableau IV. Résultats d'utilité de la méthode avatar dans les publications incluses 14
Tableau V Effets des paramètres de la méthode avatar sur la performance du jeu de données synthétisées 15
Tableau VI. Techniques de référence comparées à la méthode avatar dans la littérature scientifique et grise 16
Tableau VII. Tendance sur le plan de la performance comparative de la méthode avatar observée dans la littérature 17

LISTE DES FIGURES

Figure 1. Modèle logique de l'état de connaissances 3

LISTE DES ANNEXES

Annexe I. Tests et méthodes pour l'évaluation de la qualité de la génération de données synthétiques .. 27
Annexe II. Stratégie de recherche documentaire 29
Annexe III. Diagramme de flux du processus de sélection des études 39
Annexe IV. Publications exclues avec justification 41
Annexe V. Description des publications incluses présentant des données sur la méthode avatar 43
Annexe VI. Effets de la variation des paramètres de la méthode avatar sur les données synthétisées ... 45
Annexe VII. Description des paramètres de la méthode avatar et des jeux de données originaux utilisés dans les publications incluses 47

ABRÉVIATIONS

ABRÉVIATION	Définition
AIA	Attaque par inférence d'attribut
CHUS	Centre hospitalier universitaire de Sherbrooke
CIUSSS	Centre intégré universitaire de santé et de services sociaux
CRCHUS	Centre de recherche du CHUS
CT-GAN	Réseau antagoniste génératif conditionnel
DCR	Distance au plus proche voisin
DQEPP	Direction de la qualité, de l'éthique, de la performance et du partenariat
ECG	Électrocardiogramme
ECR	Essai contrôlé randomisé
HR	De l'anglais, *hidden rate*
k	Nombre de voisins
KL	Inverse de la divergence de Kullback-Leibler
KS	Test de Kolmogorov-Smirnov
LC	De l'anglais, *local cloaking*
MIA	Attaque par inférence d'appartenance
MST	Maximum Spanning Tree
N/A	Ne s'applique pas
NNDR	Ratio de distance au plus proche voisin
ncp	Nombre de dimension projetées pour le calcul des distances des voisins dans la méthode avatar
PARM	Probabilistic autoregressive model
TVAE	Auto-encodeurs variationnels tabulaires

AVANT-PROPOS

L'accès aux données de vie réelles en tant que levier d'innovation et de développement des connaissances

Les données de vie réelles (r*eal-world data*) représentent un élément clé pour soutenir l'avancement des connaissances ainsi que l'amélioration de la qualité, de l'accessibilité, de la pertinence et de la performance des soins et des services dans le domaine de la santé.

Malgré un bassin impressionnant de données recueillies par l'informatisation des dossiers médicaux et la recherche, l'accès à ces données demeure restreint en vertu des obligations légales et éthiques visant à respecter la confidentialité et à protéger les renseignements personnels des membres de la population.

Avec l'évolution des technologies, de nouvelles stratégies et approches s'ouvrent aux établissements de santé, d'enseignement et de recherche permettant d'accroitre le potentiel d'utilisation des données.

Des réflexions préalables s'imposent toutefois face à la complexité croissante des besoins ainsi qu'à la spécialisation et la multitude de solutions développées.

À l'ère des changements de paradigme et de l'évolution du système de santé au Québec, la création de valeur et l'optimisation des ressources sont au cœur des préoccupations. Ainsi, il est primordial d'intégrer une réflexion critique et éclairée balançant les besoins, les solutions et la protection des données personnels en soutien à la prise de décision concernant l'acquisition et l'implantation de ces nouvelles solutions.

C'est dans ce contexte qu'un état des connaissances a été mandatée pour soutenir les réflexions entourant l'essai d'une innovation de rupture comme solution d'anonymisation des données.

Stéphanie McMahon
Directrice
Direction de la coordination de la mission universitaire
CIUSSS de l'Estrie – CHUS

1. INTRODUCTION

1.1 CONTEXTE

Avec le développement de la santé numérique (1), la médecine s'allie de plus en plus à la technologie pour stimuler l'évolution des connaissances, transformer et améliorer la prestation des soins et des services. Pour y arriver, l'accès à des données de santé de qualité est crucial, mais demeure restreint en vertu des obligations légales et éthiques entourant la confidentialité et la protection des renseignements personnels des individus.

En effet, les données de santé (2) comprennent des données à caractère personnel ou ***sensible***, soit des informations identifiantes ou pouvant mener à l'identification d'une personne. Elles peuvent provenir de différentes sources telles les dossiers médicaux, les essais cliniques, les bases de données clinico-administratives ou encore des applications de santé mobile.

Plusieurs approches ont été développées pour permettre le partage de données de santé de manière à protéger la vie privée des individus. Parmi ces dernières, il y a l'anonymisation et plus récemment, la génération de données synthétiques.

L'anonymisation (2,3) est une technique d'épuration des données dont l'objectif est de rendre impossible la réidentification des individus par l'isolement des informations dans l'ensemble du jeu de données, le croisement de données distinctes ou la déduction d'information. Les processus d'anonymisation reposent principalement sur la permutation aléatoire (randomisation) ou la généralisation. Parmi les exemples de techniques d'anonymisation, on retrouve la **k-anonymisation** qui consiste à la suppression des dossiers hautement sensibles ainsi qu'à la généralisation pour augmenter le chevauchement et éviter des entrées uniques de données. Ces techniques ont le désavantage de réduire l'utilité des données au niveau individuel.

La génération de données synthétiques (3) est une approche de protection de la vie privée dès la conception (*privacy by design*) qui vise la création de jeux de données artificielles reproduisant les propriétés statistiques des données originales. La génération peut être réalisée au moyen de différents algorithmes utilisant des méthodes statistiques, des arbres décisionnels, des réseaux de neurones artificielles et autres. Cette approche permet en plus l'augmentation des données (3), c'est-à-dire la création des jeux de données supplémentaires ou la complétion de ceux incomplets (4). Parmi les techniques les plus récentes spécifiquement développées pour la santé, il y a l'avatarisation des données qui est la génération de données synthétiques par la **méthode avatar**. Cette méthode est commercialisée par l'entreprise française *Octopize Mimethik Data*.

En Estrie, de récents travaux de recherche ont d'ailleurs mis en lumière la contribution potentielle de la méthode avatar en tant que levier d'innovation pour l'utilisation et le partage des données de santé. Le centre de recherche du CHUS (CRCHUS) et la direction de la qualité, de l'éthique, de la performance et du partenariat (DQEPP) ont démontré l'intérêt d'acquérir les droits d'utilisation de la solution technologique brevetée d'*Octopize Mimethik Data*. Un projet pilote est en élaboration pour tester le logiciel ainsi que réaliser son évaluation organisationnelle et éthique en contexte québécois.

1.2 PROBLÉMATIQUE

Puisqu'il s'agit d'une méthode récente et d'une innovation de rupture dans son domaine, un état des connaissances est nécessaire pour comprendre la méthode, ses paramètres et l'étendue de la preuve disponible. De plus, l'état des connaissances est requis afin d'identifier les indicateurs pertinents à une future évaluation ainsi que les facteurs à considérer pour une utilisation judicieuse et adéquate de la méthode dans le domaine de la santé.

L'UETMISSS a donc été mandatée pour réaliser une revue rapide en soutien au processus d'aide à la décision concernant l'essai et l'acquisition des droits d'utilisation de la technologie d'*Octopize Mimethik Data* au CIUSSS de l'Estrie – CHUS.

1.3 ENJEUX ET BESOINS DÉCISIONNELS

Les échanges avec l'équipe demanderesse et le survol préliminaire de la littérature ont permis de cerner les principaux enjeux suivants :

- De par son caractère nouveau, il y a peu d'études sur l'avatarisation des données comparativement à d'autres méthodes de génération des données telles que les réseaux antagonistes génératifs conditionnels (CT-GAN);
- L'acceptabilité sociale des différentes parties prenantes, dont les usagères, les usagers et la population, semble peu documentée sur la génération et l'utilisation de données synthétiques à partir de leurs données cliniques.

Les besoins décisionnels sont de :

- Établir le portrait de la technologie, soit de ses caractéristiques, sa performance comme approche pour anonymiser un jeu de données et les ressources nécessaires pour la déployer;
- Repérer les considérations organisationnelles, contextuelles et éthiques affectant son utilisation;
- Dans la mesure du possible, repérer les indicateurs pertinents pour une évaluation organisationnelle et économique de l'avatarisation des données.

1.4 QUESTIONS DÉCISIONNELLES

Quel est l'état des connaissances actuelles sur la technologie d'avatarisation des données, développée par l'entreprise *Octopize Mimethik Data*?

2. MÉTHODOLOGIE

2.1 QUESTIONS D'ÉVALUATION

Cet état des connaissances sur l'avatarisation des données est réalisé de manière exploratoire en soutien à l'appréciation du potentiel de valeur de la technologie d'*Octopize Mimethik Data* au CIUSSS de l'Estrie - CHUS et aux évaluations organisationnelle et économique prévues.

Il vise plus particulièrement à répondre aux questions d'évaluation suivantes:

1) Quelles sont les caractéristiques générales et la performance de la méthode avatar en recherche et en santé?
2) Quels sont les avantages et les inconvénients de la méthode pour la protection de l'anonymat et l'utilisation des données?
3) Quels sont les facteurs organisationnels, contextuels et éthiques à considérer pour l'acquisition des droits d'utilisation et l'implantation de la méthode avatar au CIUSSS de l'Estrie - CHUS?

L'état des connaissances a pour objectif d'établir un portrait global relativement vulgarisé de la méthode avatar utilisée dans le domaine de la santé; sa description se limite donc à la présentation de ses principes généraux. Par ailleurs, l'évaluation éthique et légale, incluant la recension des écrits législatifs régissant l'utilisation et la protection des données ou renseignements personnels en vigueur au Québec et au Canada, est exclue de la portée de cet état des connaissances. Cette dernière fait l'objet d'un mandat distinct dans l'établissement.

2.2 MODÈLE LOGIQUE ET CADRE D'ANALYSE

En ce qui concerne le portrait de l'utilisation et de la performance de la méthode avatar, l'état des connaissances s'est concentré sur les concepts clés liés à l'anonymisation des données et aux indicateurs de performance fréquemment retrouvés dans la littérature.

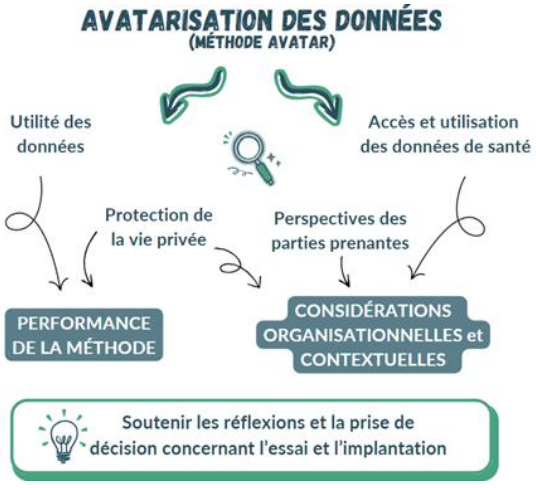


Figure 1. Modèle logique de l'état de connaissances

2.2.1 Concepts clés et indicateurs

Les indicateurs généralement employés lors de l'évaluation des données synthétiques se regroupent sous les dimensions suivantes (5,6) :

- Utilité
  Concept reposant sur l'examen de la similarité statistique entre les données originales et les données synthétiques pour un cas spécifique d'utilisation spécifique. En d'autres mots, l'utilité est la mesure dans laquelle le jeu de données synthétisées convient à l'utilisation envisagée.

  L'utilité est mesurée par un examen de la cohérence à deux niveaux (7) :
  - ⇨ Utilité générale (*broad utility*), soit la rétention d'information au niveau de la population. C'est une mesure de la fidélité. Elle peut être mesurée par la comparaison des distributions des variables (analyse univariée), des dépendances par les variables (analyse bivariée) ou celle des informations générales des données (analyse multivariée).

  - ⇨ Utilité spécifique (*Narrow or specific utility*), soit la rétention d'information au niveau individuel. C'est une mesure de la fonctionnalité, par exemple la réplicabilité des données de survie pour un groupe atteint d'un cancer. Elle peut être mesurée notamment par la performance des modèles prédictifs et des statistiques descriptives.

- Représentativité (*fairness*)
  Concept qui se rapporte à la justesse de la représentativité des données synthétiques pour les attributs d'intérêt, sur le plan de l'analyse des différents sous-groupes, comparativement au jeu de données originales (8).

- Protection de la vie privée (*privacy*)
  Concept multidimensionnel dépendant du contexte et des données d'intérêts qui est principalement mesuré par l'évaluation des risques pour le respect de la confidentialité. Parmi les indicateurs de la protection de la vie privée, il y a :

  - ⇨ Individualisation (*singling out*) (9), qui réfère au risque d'identifier une personne à partir des renseignements disponibles d'un jeu de données. C'est lorsque l'identité d'une personne peut être attribuée spécifiquement à une entrée dans le jeu de données à partir des informations qui y sont contenues telles que son diagnostic pour une maladie rare ou par la combinaison de plusieurs facteurs;

  - ⇨ Corrélation (*linkability*), qui réfère au risque «de relier des individus avec une source de données externe ayant des caractéristiques communes» (7). Par exemple, c'est d'être en mesure de déduire quel bassin ou échantillon a été utilisé pour la génération des données en croisant ces informations avec une autre source telle que les données d'admission, de remboursement par la RAMQ ou les notes au dossier médical;

  - ⇨ Inférence, qui réfère au risque que des informations sur les individus peuvent être déduites significativement (peu de doute) à partir du jeu de données anonymisées (7). Par exemple, c'est d'être en mesure de déduire une information telle que l'ethnie d'une personne à partir du fait qu'un test spécifique à ce groupe de la population a été utilisé.

Un tableau des tests, des méthodes d'évaluation et de leurs interprétations pour chaque indicateur est disponible à l'**Annexe *I*.**

2.2.2 Examen des facteurs organisationnels et contextuels

Pour l'identification des considérations organisationnelles et contextuelles, la collecte et l'analyse de la littérature scientifique et grise ont porté une attention particulière aux cinq dimensions de la valeur de l'INESSS (10) : clinique, populationnelle, organisationnelle, socioculturelle et économique.

2.3 RECENSION DES ÉCRITS

Une grille CHIP (11) et des critères de sélection ont été établis préalablement pour guider la recension des écrits.

Tableau I. Grille CHIP

CHIP	Description
Contexte	Données synthétiques anonymisées en contexte clinique (recherche ou santé)
Méthodes (**H**ow)	Méthode Avatar pour la création et l'utilisation des données
Questions (**I**ssues)	• Contexte d'utilisation (type d'étude, de données et paramètres testés, …), mesures et résultats des études ou évaluations de la méthode avatar • Avantages et inconvénients observés quant à l'emploi de la création et l'utilisation des données • Considérations organisationnelles, sociales, économiques, éthiques, juridiques à considérer
Population	Aucune précision particulière

L'anonymisation des données existe depuis bien longtemps. Elle est utilisée dans bien d'autres domaines, tels que les recensements de la population, et fait l'objet de multiples méthodologies. Toutefois, ces méthodologies alternatives et autres contextes ont été exclus de la recension. En effet, l'objectif est de soutenir la prise de décision concernant l'acquisition des droits d'utilisation de la méthode avatar, spécifiquement développée pour le domaine de la santé, dans le contexte d'un établissement de la santé. Un survol plus général de la littérature a néanmoins été utilisé pour élargir les réflexions découlant de l'analyse des résultats dans la discussion.

2.3.1 STRATÉGIE DE RECHERCHE

Un plan des concepts a été établi par la conseillère en évaluation responsable de cet état des connaissances. Un bibliothécaire (Mykola Krupko) a ensuite planifié et réalisé la recherche documentaire en trois temps.

Puisqu'il s'agit d'une méthode récente ayant peu de publication, le premier tour se concentrait les publications de tout type sur l'avatarisation des données spécifiquement.

Afin d'élargir la recherche et de compléter le portrait, le second et troisième tour visaient à identifier les synthèses récentes de connaissances, plus particulièrement les revues systématiques et les méta-analyses, respectivement sur les méthodes de génération de données synthétiques et d'anonymisation des données en santé.

Les stratégies et les historiques de recherche documentaire sont disponibles à l'***Annexe II***.

Un repérage manuel supplémentaire a été réalisé à partir des listes des références des études pertinentes, d'une liste de publication partagée par l'entreprise Octopize et de Google Scholar (notamment pour la littérature grise).

2.3.2 SÉLECTION DES ÉTUDES

Le processus de sélection des études a été réalisé sur l'ensemble des publications repérées par les trois tours de la recherche documentaire. Les critères de sélection ont été discutés et ajustés après la lecture de quelques titres et résumés entre deux membres de l'UETMISSS. La version finale des critères est présentée au tableau II.

Tableau II. Critères de sélection pour la recension des écrits

Aspect	Inclusion	Exclusion
Langue	Anglais, français	Autres langues
Date de publication	Cinq dernières années	Avant 2020
Sujet	• Avatarisation des données, méthode Avatar, algorithme avatar, données avatar, méthode développée par Octopize	• Avatarisation des patientes ou patients (*avatar patients*)
	• Données synthétiques ou génératives	• Exclusivement sur d'autres méthodes de synthèse de données
Contexte	Clinique	Tout contexte autre que clinique
Type de publication	**Avatarisation**: • Tout type d'étude ou évaluation	• Publication sans comparaison ou résultat de la méthode avatar • Lettre d'opinion • Résumé de conférence
	Anonymisation ou génération de données synthétiques: • Revue systématique ou de portée (*scoping review*), méta-synthèse ou méta-analyse	

Par la suite, une double sélection indépendante (MPB, AB) a été réalisée sur la totalité des titres et résumés avant de procéder à la lecture complète des publications retenues. Les publications dont le titre et le résumé ne permettaient pas de confirmer l'adéquation aux critères de sélection, notamment en absence d'une mention explicite concernant la méthode avatar, ont été soumises à une lecture complète.

La résolution de conflits a été réalisée lors d'une discussion entre les deux membres de l'équipe, jusqu'à l'atteinte d'un consensus. Le coefficient Kappa de Cohen a été utilisé pour déterminer la concordance d'évaluation entre les membres de l'équipe; l'atteinte d'un coefficient supérieur à 0,8 (12) correspond à un taux d'accord inter-juge élevé.

L'extraction des données (autrices ou auteurs, année de publication, pays, type de publication, éléments abordés par rapport à la méthode avatar) a été réalisée par un membre de l'UETMISSS (MPB) à partir de grilles d'extraction établies à la suite du processus de sélection.

En raison du devis sélectionné et des contraintes de temps, aucune évaluation de la qualité des études n'a été réalisée.

2.4 ENJEUX D'ÉTHIQUE ET D'ÉQUITÉ

Les quatre principes fondamentaux de la bioéthique, c'est-à-dire l'autonomie, la bienfaisance, la non-malfaisance et la justice (13), sont considérés par l'UETMISSS à toutes les étapes du processus de production d'une évaluation. Une attention particulière a été portée aux considérants en lien avec le respect et la protection de la vie privée et l'acceptabilité sociale de ce type de technologie.

Dans ce cas, la dimension éthique a été traitée transversalement notamment par le biais de la recherche et l'analyse des perspectives, des valeurs et des préférences des différentes parties prenantes dans la littérature scientifique et grise.

2.5 VALIDATION

Le rapport préliminaire a été acheminé à François Rheault, au professeur adjoint en informatique à la Faculté des sciences de l'Université de Sherbrooke à des fins de validation externe. Les objectifs de cette validation étaient :

- D'effectuer une relecture du document en portant une attention particulière à la vulgarisation des concepts et des indicateurs d'évaluation liés aux méthodes de génération de données synthétiques ainsi qu'à la cohérence entre les résultats et leurs interprétations;
- De proposer des pistes d'amélioration, s'il y a lieu.

La validation externe a été réalisée en novembre 2025 et les commentaires de Professeur Rheault ont permis de bonifier ou de préciser le contenu de ce présent document.

3. RÉSULTATS

3.1 RECENSION DES ÉCRITS

3.1.1 Résultats de la recherche de la littérature

La stratégie de recherche documentaire en trois temps a permis d'identifier 252 publications auxquelles s'ajoutent 18 publications supplémentaires repérées manuellement. Après le retrait des doublons, ce sont 193 publications qui ont été soumises au processus de sélection.

À la lecture des titres et des résumés, 39 publications ont été sélectionnées pour une lecture intégrale. Parmi celles-ci, 12 publications correspondant aux critères de sélection ont été incluses dans la recension des écrits. Parmi les publications incluses, il y a un article méthodologique avec cas d'usage (14), quatre études évaluatives (3,15–17), cinq prépublications ou articles de conférence (18–22), un mémoire de maîtrise (2) et une étude observationnelle (23).

Le diagramme de flux de la sélection et la liste des publications exclues à la lecture complète avec justification sont disponibles respectivement à l'Annexe III et l'Annexe IV.

3.1.2 Description des articles inclus

Un tableau décrivant les publications incluses est disponible à l' Annexe *V*.

L'ensemble des publications proviennent de la France et en moindre partie, du Canada.

De ce nombre, onze publications présentent un devis évaluatif pour la méthode avatar examinant particulièrement les indicateurs d'utilité et de protection de la vie privée (2,3,14–22). Parmi ces publications se retrouvent :

- La publication originale de Guillaudeux *et al.* présentant la méthode avatar pour les données tabulaires (14) ainsi que la publication portant sur son extension Chronos pour les séries temporelles (18). Ces dernières comprennent une démonstration par cas d'usage;
- Deux publications proposant une alternative (3,22) à la méthode de Guillaudeux *et al.*;
- Cinq publications évaluant l'effet de la variation de certains paramètres de la méthode avatar sur les données synthétiques (2,14,15,17,20,22);
- Six publications comparant la méthode avatar à d'autres méthodes de génération de données synthétiques ou d'anonymisation (2,3,14,15,17,21).

À noter que deux de ces publications utilisent une version simplifiée (15) ou une interface nommée AnonymHUS inspirée de la méthode avatar (2);

Quoique son sujet ne corresponde pas au CHIP, une étude observationnelle (23) a été incluse puisqu'une comparaison entre les données originales et avatar est fournie en données supplémentaires.

L'évaluation de la qualité des études n'a pas été réalisée. Toutefois, il faut noter que :

- La moitié des publications incluses, provenant toutes de la littérature grise, n'a pas été révisée par les pairs au moment de cet état des connaissances (15,18–22);
- Plusieurs des autrices et des auteurs ont contribué à plus d'une publication et par conséquent, il est fort probable que les publications reposent sur un relativement faible nombre de groupes de recherche;
- Du personnel d'Octopize est cité dans les autrices et des auteurs de 5 des 12 publications incluses (14,16,18,20,22).

3.1.3 Synthèse et interprétation des données probantes

Cette sous-section est divisée en trois volets qui présentent respectivement la méthode avatar, sa performance relative aux données originales et à d'autres méthodes de génération de données synthétiques ou d'anonymisation.

3.1.3.1 Description de la méthode avatar

La méthode avatar (14,24) est un algorithme de génération de données synthétiques utilisant une approche centrée sur la personne (patiente ou patient), préservant la taille et la nature des données originales, par la simulation d'un avatar unique pour chaque individu.

Spécifiquement conçue pour le domaine de la santé, elle requiert un jeu de données tabulaires pseudoanonymisées où les rangées et les colonnes représentent respectivement les individus et les variables. Les variables employées peuvent être de nature continue, catégorielle, booléenne ou temporelle sous la forme de date.

La méthode repose sur cinq étapes clés :

1 Projection des données

Les données sont projetées dans un espace multidimensionnel à l'aide de techniques de réduction des dimensions comme l'analyse factorielle des données mixtes (FAMD), l'analyse en composantes principales (ACP) ou l'analyse des correspondances multiples (ACM). Il s'agit en fait d'une étape de transformation des données en coordonnées numériques structurées, qui permet notamment de réduire leur complexité.

2 Identification des k – voisins plus proches

Pour chaque coordonnée (projection d'un individu à partir des données originales), les voisins les plus proches sont identifiés. La distance entre eux est calculée à l'aide d'un algorithme de K-voisins les plus proches (KNN) qui permet de définir une zone locale autour de chaque individu. Un poids est attribué à chaque voisin.

3 Génération aléatoire des avatars

Une nouvelle donnée, un avatar, est générée par une simulation unique pour chaque **zone locale** précédemment créée. Cette génération est réalisée de manière pseudo-aléatoire, c'est-à-dire par la combinaison aléatoire des informations de l'individu d'origine et de ses plus proches voisins. Une méthode empirique, reposant sur la génération des poids basée sur la distance aux voisins, une loi exponentielle et des permutations aléatoires, est utilisée.

4 Inversion de la projection

Les coordonnées avatar sont reconverties dans le format des données d'origines en opérant l'opération inverse de la projection réalisée à l'étape 1. Les données avatar, ainsi produites, ont une structure cohérente et comparable au jeu de données de départ.

5 Évaluation des paramètres de protection de la vie privée

Les mesures de protection de la vie privée sont calculées pour s'assurer que le jeu de données synthétiques respecte les critères en vigueur pour la protection de la vie privée.

Les mesures d'utilité peuvent être réalisées pour évaluer à quel degré les données avatar ont retenu les informations ou reproduit les associations, autant sur le plan individuel que populationnel.

Paramètres de la méthode avatar

La méthode avatar est contrôlée par deux principaux types de paramètres (14):

1 **Paramètres affectant l'environnement (zone) local** :

- ⇨ Distance mathématique utilisée (ex. euclidienne) ;
- ⇨ Nombre de voisins demandé (k);
- ⇨ Paramètres de la projection notamment, le nombre de dimension utilisée pour l'identification des voisins (ncp) et le poids accordé à chaque variable pour la projection.

2 **Paramètres influençant le caractère stochastique (aléatoire) de la génération d'un avatar** :

- ⇨ Loi de distribution des poids utilisée lors la génération des avatars;
- ⇨ Pourcentage de permutation appliquée à l'avatar pour chaque variable.

Le paramètre **k** est le plus fréquemment employé dans la littérature pour moduler les résultats de la méthode.

La liste des paramètres testés par publication est disponible à l'*Annexe VI*.

Dans une moindre proportion, une troisième catégorie de paramètres a aussi été testée pour tenter de pallier les enjeux fréquemment rencontrés des données de santé tels que la faible taille de l'échantillon et les données incomplètes. Il s'agit des **paramètres d'augmentation des données**, tels que la taille de l'échantillon généré (nombre d'avatar par individu) et le mode de gestion des données manquantes employé.

Variation de la méthode

Selon Guillaudeux *et al.*, la méthode avatar pourrait être adaptée à d'autres types de données avec un développement spécifique (14). Parmi les possibilités, le groupe mentionne les images, les données de grande dimension, de suivi (*tracking*) ou géospatiales et les séries temporelles.

À cet effet, Chronos (18) est une extension de la méthode Avatar pour les séries temporelles qui a été intégrée à la solution technologique d'Octopize. Son utilisation a été démontrée pour la génération de signaux d'électrocardiogramme (ECG) synthétiques. Pour le même type de signaux, une autre étude (22) propose l'emploi d'une technique alternative basée sur loi de Dirichet pour le calcul des poids dans la méthode avatar.

3.1.3.2 Performance de la méthode avatar

Les conditions de l'avatarisation des données dans les différentes publications sont disponibles à l' **Annexe *VII***. Le nombre d'individus ou de signaux, d'observations ou de variables varient grandement entre les publications. Les sources principales sont les données d'essais contrôlés randomisés, d'études observationnelles et les données clinico-administratives. Les paramètres pour l'avatarisation des données sont majoritairement une valeur k de 5, 10 ou 20 et lorsque spécifié, un ncp entre 5 et 20 ainsi qu'un poids de 10 ou 20 aux principales variables d'intérêt.

Protection de la vie privée

Pour se positionner sur le degré de protection de la vie privée des données générées par la méthode avatar, les groupes ont mesuré plus fréquemment (

Tableau III):

- La distance au proche voisin (DCR) et le ratio de DCR (NNDR);
- Le *hidden rate* (HR) et le *local cloaking* (LC), deux mesures spécifiques à la méthode avatar (25).

> Le *hidden rate* (25) est une mesure évaluant la probabilité de faire une erreur quand une attaque est faite pour associer un individu à son avatar. Il s'agit d'une mesure spécifique réalisée sur le lien, conservée temporairement, entre l'avatar et la donnée originale pendant le processus de génération.
>
> Le *local cloaking* (25) représente le nombre d'avatars qui ressemble davantage à l'individu d'origine que l'avatar qu'il a généré. Pour évaluer un jeu de données avatarisées, la médiane de la mesure pour tous les individus est utilisée.

Tel que rapporté par Demuth et *al.* (20), Octopize recommande des valeurs minimales pour DCR, NNDR et LC respectivement de 0.2, 0.3 et 3 ainsi qu'un HR de plus de 90 %. Pour leur part, le groupe juge que des valeurs légèrement plus basses sont acceptables. Bien qu'ils respectent tous les critères d'acceptabilité de Demuth et *al.*, la majorité des jeux de données avatarisées dans les publications incluses répondent aux recommandations d'Octopize.

Dans leur publication, Lebrun *et al.* réalisent des simulations d'attaque par inférence d'appartenance (MIA) ou d'attribut (AIA). La méthode avatar présente des valeurs relativement élevées, particulièrement au MIA, au suggérant une vulnérabilité du modèle. Le groupe conclue que la méthode avatar introduit donc un risque à ce niveau (3).

Tableau III. Résultats de protection de la vie privée de la méthode dans les publications incluses

Dimension	Méthode d'évaluation	Résultat♦	Interprétation	Références
Individualisation	Distance au proche voisin (DCR, médiane)*	Entre 0 et 4.39 (Une seule valeur sous 0.3)	Un risque d'individualisation est présent lorsque le DCR et le NNDR sont tous deux faibles.	(2,14,17,20)
	Ratio de DCR (NNDR, médiane)*	Entre 0.33 à 1 (Majoritairement à plus de 0.8)		(2,14,17,20)
	Hidden rate (HR)	Entre 84 et 94 %	Plus le % est élevé, plus les individus sont protégés.	(14,18,20,22)
	Local cloaking (LC, médiane)	Entre 3 et 52	Plus le nombre est élevé, plus les individus sont protégés.	(14,18,20,22)
	Taux d'individu ou information sensible	Entre 0.1 à 6.4 %	Plus le % est bas, plus les individus sont protégés.	(3,14,21)
Corrélation	Taux de protection contre les corrélations	Entre 70 et 100% (varie selon le type de corrélation)	Plus le % est élevé, plus les individus sont protégés.	(3,20)
Inférence	Risque d'attaque par inférence d'appartenance (MIA) ou d'attribut (AIA)	MIA : 0.593 AIA : 0.295	Plus la probabilité est élevée, moins les individus sont protégés.	(3)

* Sur la base d'une distance euclidienne.

Utilité

Sur le plan de l'utilité générale, les mesures les plus fréquentes sont le test de Kolmogorov-Smirnov (KS), l'inverse de la distribution de Kullback-Leibler (KL), la distance de Hellinger et les matrices de corrélation (**Tableau IV**. Résultats d'utilité de la méthode avatar dans les publications incluses). Dans l'ensemble, les résultats suggèrent des distributions similaires pour les paramètres généraux des données (utilité générale). Les examens visuels et statistiques des distributions suggèrent majoritairement une préservation de la structure et une rétention de l'information acceptables au niveau de la population (paramètres descriptifs généraux).

Pour l'utilité spécifique, les résultats primaires sont répliqués selon la majorité des publications. La réplication des résultats secondaires est davantage mitigée (17,20,23).

Tableau IV. Résultats d'utilité de la méthode avatar dans les publications incluses

Dimension	Méthode d'évaluation	Résultats ou observations	Interprétation	Références
Utilité générale	Test de Kolmogrov-Smirnov	Statistique D (2): Entre 0.157 et 0.167	Plus la valeur est élevée, plus l'échantillon est éloigné de la distribution théorique.	(2,15,17)
		Valeur de p (15,17) : Entre 0.9 et 0.927	Valeur entre 0 et 1; plus elle est proche de 1, plus les distributions considérées sont proches	
	Inverse de la divergence de Kullback-Leibler	Échantillon de grande taille : Entre 0.756 et 0.87 Échantillon de petite taille : Entre 0.226 et 0.616	Plus la valeur est proche de 0, plus les distributions sont différentes.	(20–22)
	Distance de Hellinger (médiane)	Entre 0.08 et 0.145	Score en 0 et 1 où plus la valeur s'approchant de 1, plus les distributions sont différentes	(15,17)
	Taux de différence (corrélation)	1.49 ≤ valeur > 10	Plus la valeur est basse, plus l'utilité est préservée.	(16,20)
	Statistiques descriptives	83.3 % de similitude par rapport au nombre d'individus pour les statistiques descriptives entre les données originales et avatarisées		(21)
	Examen visuel des distributions	Préservation de la structure des données		(3,14,20,21,23)
Utilité spécifique	Modèle prédictif	Justesse des prédictions entre 77 % à 95%	Plus le % est élevé, plus les prédictions sont précises.	(3,16,18)
	Mesures adaptées aux effets spécifiques	Reproduction des effets primaires, mais tendance à surévaluer leurs effets Reproduction difficile des effets secondaires		(2,3,14,16,17, 19,20,23)

D'autre part, un groupe de recherche conclut que les variables catégorielles sont mieux préservées par la méthode avatar que les variables quantitatives qui, quant à elles, ont tendance à s'agréger et à être normalisées (20). Cette tendance à la centralisation vers la moyenne est aussi observée dans les mesures d'utilité spécifique par d'autres groupes (12,14,15), notamment pour les données continues. Plusieurs groupes soulèvent une autre tendance à surestimer les effets et les associations dans les jeux de données avatarisées (23). Quelques publications rapportent des valeurs aberrantes, soit des valeurs minimales négatives biologiquement impossibles pour une variable (2), une diversité réduite de valeurs extrêmes (2,20) et une accentuation du déséquilibre des classes [augmentation des classes majoritaires et diminution des classes minoritaires] (20,23).

Quoiqu'une meilleure utilité générale n'améliore pas nécessairement l'utilité spécifique, un groupe appuie qu'une optimisation des paramètres de la méthode avatar puisse préserver les caractéristiques des variables d'intérêt pour l'utilisation future des données avatarisées (20).

Adaptation et paramétrage de la méthode

Pour de courts signaux ECG, l'utilisation de la loi de distribution de Dirichlet pour le calcul des poids présentent des résultats similaires à la méthode avatar originale pour l'utilité générale et la protection de la vie privée (22). L'adaptation de la méthode avatar aux séries temporelles, Chronos, démontre une excellente utilité spécifique aux modèles prédictifs et une bonne protection de la vie privée avec des différences visuelles pour des signaux ECG plus longs (18).

Les effets des différentes configurations des paramètres de la méthode sont présentés au **Tableau V**.

Le principal paramètre pour ajuster la méthode semble être la valeur k. Les valeurs basses de k sont associées avec une meilleure utilité, car la structure des distributions est davantage conservée (14). Au contraire, la confidentialité augmente avec des valeurs de k plus élevées, notamment parce que la structure et les frontières diminuent et que la mesure de LC augmente (14,17). Ce paramétrage réduit l'utilité du jeu de données.

Tableau V Effets des paramètres de la méthode avatar sur la performance du jeu de données synthétisées

Paramètres	Constats	Références
k	• Plus la valeur de k est faible, plus le jeu de données avatarisées est précis; • Plus la valeur de k est haute, plus le jeu de données avatarisées est confidentielle; • L'effet de taille a tendance à être surestimé par les basses valeurs de k et au contraire, sous-estimé par les plus hautes valeurs de k.	(14,17,20)
ncp	• Une petite valeur de ncp augmente l'utilité avec peu d'effet sur la protection de la vie privée.	(20)
Poids	• Le poids et l'encodage différentiel de certaines variables permettent potentiellement d'optimiser l'équilibre (juste milieu) utilité et protection de la vie privée.	(20)
Augmentation de la taille des jeux de données	• L'augmentation des données améliorent certaines mesures d'utilité comme l'inverse de la divergence de Kullback-Leibler et le test de Kolmogrov-Smirnov; • L'augmentation a tendance à introduire des associations de type faux positif.	(15,17)

Par ailleurs, les caractéristiques du jeu de données originales influencent aussi les résultats de la méthode avatar. En effet, la méthode avatar semble :

- Moins performante, notamment pour la protection de la vie privée, pour les jeux de données de départ de petite taille (15);
- Plus performante pour la génération de données multivariées, soit lorsqu'elle est utilisée pour traiter simultanément plusieurs composantes (2).

3.1.3.3 COMPARAISON DE LA MÉTHODE AVATAR

Dans la littérature, de nombreuses méthodes de génération de données synthétiques sont employées en santé (Tableau VI). Les plus fréquemment employées sont deux modèles bien documentés pour la génération de données synthétiques : CT-GAN, un algorithme d'apprentissage profond (*deep learning*) utilisant des réseaux de neurones et Synthpop, une solution technologique reposant sur des méthodes statistiques.

Voici les méthodes de référence auxquelles la méthode avatar a été comparée dans la littérature scientifique et grise :

Tableau VI. Techniques de référence comparées à la méthode avatar dans la littérature scientifique et grise

Technique	Principe ou méthode de génération de données synthétiques	Nombre d'étude (Référence)
Réseau antagoniste génératif conditionnel (CT-GAN)	Méthode d'apprentissage profond basée sur l'utilisation d'une paire de réseaux de neurones (générateur et discriminateur)	5 (2,3,14,15,17)
Synthpop	Combinaison de techniques statistiques, soit des analyses exploratoires approfondies et de modélisations statistiques	3 (2,3,14)
Auto-encodeurs variationnels tabulaires (TVAE)	Méthode d'apprentissage profond basée sur des réseaux de neurones	2 (2,17)
M-avatar	Méthode alternative à la méthode avatar qui génère les données synthétiques sur demande à partir d'un modèle global plutôt qu'une approche centrée sur l'individu.	1 (3)
Maximum Spanning Tree (MST)	Méthode de confidentialité différentielle (*differential privacy*) reposant sur une approche d'estimation marginale avant l'utilisation d'un modèle probabiliste graphique pour la génération de données synthétiques	1 (3)
PARM	Modèle Python en libre accès pour générer des données synthétiques basées sur des modèles autorégressifs probabilistes	1 (21)
SAIPH	Solution technologique qui projette le point d'origine dans un espace latent de faible dimension et reconstruit le point de données dans l'espace d'origine à partir de cette projection.	1 (3)

Sur six publications, deux publications ont comparé la performance de la méthode avatar aux méthodes d'anonymisation suivantes : la k-anonymisation (3) et l'approche de généralisation (21). Trois ont un échantillon de départ de relativement petite taille (2,17,21), trois testent plus d'une valeur de k pour les comparaisons (2,15,17). Deux comparent la méthode avatar uniquement avec une valeur k de 20 (3,14) ou une augmentation de quatre fois la taille originale (15,17).

Parmi ces publications, on observe :

- Un manque de consensus pour les méthodes d'évaluation;
- Des différences des caractéristiques des jeux de données originales;
- Une grande variation du paramétrage de la méthode.

Les tendances pour la performance comparative de la méthode avatar sont présentées au **Tableau VII**. Ce tableau présente un sommaire visuel dont l'objectif est de décrire si la méthode avatar a généralement mieux ou moins bien performé que les autres techniques dans la littérature.

Dans l'ensemble, la méthode avatar offre une performance variable comparativement aux autres méthodes de génération des données synthétiques et d'anonymisation.

Tableau VII. Tendance sur le plan de la performance comparative de la méthode avatar observée dans la littérature

Comparateur (références)	Performance comparative de la méthode avatar		
	Utilité	Protection de la vie privée	Augmentation de la taille des données
CT-GAN (2,3,14,15,17)	+/-	+/-	≈ (+/- selon le k)
Synthpop (2,3,14)	+/-	≈ ou ↘	
TVAE (2,17)	+/-	≈ ou ↗ (+/- selon le k)	≈ (si k =10 ou 20)
M-avatar (3)	↗	↘	
MST (3)	≈	↘	
PARM (21)	↗	≈ ou ↗	
SAIPH (3)	↗	↘	
K-anonymisation (21)	↗	↘	
Approches de généralisation (21)	≈	↗	
Légende pour la performance	Meilleure ↗	Similaire ≈	Moindre ↘	Variable +/-

3.2 CONSIDÉRANTS CONTEXTUELS ET ORGANISATIONNELS

Aucune publication repérée ne portait sur l'étude des facteurs contextuels et organisationnels à la génération et l'utilisation des données synthétiques en général ou spécifiquement pour la méthode avatar.

La majorité des publications mentionnent tout de même les enjeux liés à l'utilisation des données de santé ou des données synthétiques. Quelques publications élaborent, en discussion, sur les perspectives, l'encadrement et les limites de ce type de données. La majorité des publications provenant de la France se concentrent donc sur la perspective européenne.

Voici les quelques considérants pour l'utilisation et l'implantation de la génération de données synthétiques :

1 Cadre juridique et réglementaire des données synthétiques

Il s'agit de l'aspect le plus fréquemment nommé dans les publications. Il est nécessaire de considérer les lois et les règlements locaux en vigueur s'appliquant à l'utilisation de données personnelles pour la génération et de valider le statut[1] accordé aux données synthétiques par les autorités législatives (2). En effet, l'encadrement règlementaire peut différer entre la méthode de génération des données synthétiques relevant de l'intelligence artificielle et le jeu généré, un produit (2).

En France, la méthode avatar est reconnue par la Commission Nationale de l'Informatique et des Libertés (CNIL) comme une solution répondant aux exigences européennes d'anonymisation dépendamment de la configuration de ces paramètres (26).

2 Importance de l'évaluation des risques

Selon divers groupes, la démonstration de l'efficacité de l'anonymisation est nécessaire pour que les modèles de génération et leurs données synthétisées ne relèvent plus des données personnelles (2). Elle doit présenter une information suffisante sur le degré de préservation des données (utilité) et de risque pour la protection de la vie privée (21).

L'absence de consensus pour les méthodes d'évaluation ainsi que les critères de risque pour la protection de la vie privée demeure un enjeu majeur selon plus d'un groupe (2,18,19).

Sans des mesures harmonisées et des seuils d'acceptabilité, il est difficile d'évaluer si un jeu de données synthétiques ou anonymisés protège suffisamment les renseignements personnels (18).

3 Équilibre utilité et protection de la vie privée

Plus les données synthétisées sont fidèles et préservent l'information, plus le risque de réidentification augmente, diminuant ainsi la confidentialité (21). De plus, il faut garder en tête que le risque de réidentification n'est pas le même pour toutes les variables (2).

[1] Par exemple, les données synthétiques sont-elles des données personnelles ou non?

Un compromis entre l'utilité et la confidentialité des données synthétisées doit être atteint. Le degré d'altération accepté lors de la génération de données synthétiques doit être cohérent avec l'objectif prévu pour l'utilisation anticipée (ou traitement) des données (21).

Une bonne connaissance de la structure et des associations du jeu original ainsi qu'une anticipation des transformations de données nécessaires sont deux conditions gagnantes pour optimiser la génération de données synthétiques (21).

4 Perspectives et limites pour le partage des jeux de données synthétiques

L'une des principales utilisations des données synthétiques demeure le partage sécuritaire de données. Sans une connaissance de l'étendue de l'utilité du jeu de données, il est impossible de savoir si de nouvelles analyses ou l'exploration des données sont convenables (21).

Toutefois, toute information sur les paramètres, la méthode et les évaluations réalisées sur le jeu de données augmente le risque de réidentification et d'attaques. Ce dernier est intrinsèquement lié aux informations et données auxquelles l'attaquante ou l'attaquant peut avoir accès. Ainsi, la transparence et le partage d'informations doivent être balancés avec le risque auquel cela expose. (2)

Aucun élément quant aux dimensions économiques et professionnelles (ex. prix, ressources humaines, formation, entreposage des données, etc.) n'a été repéré dans la littérature.

3.3 PERSPECTIVES DES PARTIES PRENANTES

Aucune publication n'a été recensée abordant la perspective des usagères, des usagers ou de la population sur la génération ou l'utilisation de données synthétiques, toutes méthodes confondues, à partir de leurs données de santé.

Une publication aborde toutefois la perspective de différentes parties prenantes impliquées dans l'encadrement de l'utilisation des renseignements personnels telles que les autorités de contrôle (2). Selon cette publication :

- Les autorités européennes de contrôle reconnaissent généralement les données synthétiques en tant qu'une nouvelle technologie de protection améliorant la protection de la vie privée;
- Plusieurs groupes considèrent que les règlements régissant la protection des renseignements personnels, tels que le règlement général sur la protection des données (RGPD) en Europe, s'appliquent à l'utilisation des données personnelles pour la génération de données synthétiques.

4. DISCUSSION

L'utilité des jeux de données avatarisées est dépendante du paramétrage de la méthode utilisée. Règle générale, la génération de données synthétiques par la méthode avatar tend à :

- Centraliser et lisser les données (diminution de la variance);
- Retenir les associations ou les effets primaires ou les associations fortes, mais moins les effets secondaires;
- Accentuer les déséquilibres de classes.

Ces tendances suggèrent que la méthode, comme bien d'autres méthodes d'anonymisation, favorise la protection de la population majoritaire. Elle le fait en recentrant les données autour des zones à forte densité au dépend des zones davantage distinctives ou des valeurs aberrantes, qui elles sont plus faciles à ré-identifier. De plus, l'accentuation du déséquilibre de classes suggère que la méthode peut amplifier certains biais structurels des données menant à la sous-représentation de certains groupes. Quoiqu'une seule étude rapporte que la méthode a généré des données biologiquement impossibles, des recherches supplémentaires sont nécessaires pour examiner les conditions expérimentales auxquelles ses problèmes surviennent et comment la méthode peut-elle s'adapter pour les contrer.

D'un autre côté, une description des indicateurs d'utilité et de protection de la vie privée est suggérée pour définir les possibilités de réutilisation du jeu de données avatarisées. Rappelons que l'utilité spécifique d'un jeu de données synthétiques est limitée aux analyses réalisées démontrant la rétention de certaines informations. Ces propos sont appuyés par d'autres revues de la littérature sur les données synthétiques. Une des revues propose que l'implantation de telles méthodes devrait être auditée puisqu'elles sont sujettes aux erreurs (5).

Par ailleurs, plusieurs essais peuvent être nécessaires pour identifier les valeurs optimales des paramètres permettant l'atteinte des objectifs recherchés entre utilité et confidentialité pour le jeu de données synthétisées. Cette étape d'optimisation représente fort probablement un coût et un investissement de temps à considérer pour la réalisation des essais et des analyses des indicateurs.

Il est difficile de se positionner sur la performance réelle de la méthode avatar comparativement aux autres méthodes de génération des données synthétiques. D'une part, les résultats des différentes publications présentent une grande variabilité. D'autre part, l'analyse est complexifiée par le manque de consensus pour les méthodes d'évaluation, les différences des caractéristiques des jeux de données originales et la diversité des configurations respectives des paramètres de chaque méthode. Pour les indicateurs de protection de la vie privée, le manque d'harmonisation et de consensus sur leurs seuils d'acceptabilité limitent l'interprétation. Ces observations sont appuyées par plus d'une revue exploratoire sur les données synthétiques ou les indicateurs d'utilité et de protection de la vie privée (5,27). De plus, l'un de ces groupes remet en question l'utilisation d'indicateurs basés sur la similarité, tels que la DCR et le NNDR qui complexifie l'interprétation, et suggère plutôt d'utiliser des indicateurs plus modernes (5).

Néanmoins, les résultats de performance suggèrent que la méthode avatar offre, avec un ajustement adéquat de ces paramètres, une performance similaire ou supérieure aux approches les plus populaires telles que CT-GAN, Synthpop et TVAE. De plus, la variabilité observée sur le plan de la performance permet d'appuyer le potentiel et la pertinence d'optimiser les paramètres de la méthode avatar.

PERSPECTIVES D'UTILISATION

La méthode avatar est adaptée pour les données tabulaires et les séries temporelles, notamment les signaux ECG. Son utilisation a été démontrée notamment sur des jeux de données hospitaliers, d'essais cliniques ou d'études observationnelles. Les utilisations rapportées sont cohérentes avec celles suggérées par une revue de portée sur les applications de la génération de données synthétiques dans le domaine de la santé (28). Dans cette étude, Rujas et *al.*, identifie les trois grandes catégories d'utilisation :

- Développement d'IA, incluant l'entraînement et la validation de modèles ainsi que l'amélioration de la généralisabilité et de l'interprétabilité de modèles existants;
- Utilisation accrue des données secondaires, comme le partage de données, poursuite d'analyse et d'étude;
- Amélioration des connaissances cliniques, utilisation comme matériel de formation ou soutien pour le diagnostic ou le traitement.

Avant d'utiliser la méthode avatar, il est essentiel de vérifier les cadres législatifs et règlementaires associés à l'utilisation de données personnelles pour la génération ainsi que le statut accordé aux données synthétiques en vigueur au Québec et au Canada.

CONSIDÉRATIONS ÉTHIQUES À LA GÉNÉRATION ET L'UTILISATION

Aucune étude des considérants organisationnels, contextuels ou expérientiels n'a été recensée pour la génération et l'utilisation spécifiquement pour la méthode avatar. Plusieurs revues exploratoires ou systématiques, ne répondant pas tout à fait aux critères de sélection de cette revue rapide, apportent un éclairage complémentaire ou connexe.

D'un côté, il y a le statut à considérer des données synthétiques comparativement aux données dépersonnalisées. Aux États-Unis, au Canada et en Europe, la désidentification des données relève des opérations du système de santé (27). Par conséquent, aucun consentement informé des usagères et des usagers n'est nécessaire même si les données générées sont utilisées en recherche. Toutefois, le statut et la régulation des données synthétiques demeurent une zone grise et ce, peu importe l'utilisation qui en est en faite (27). Cette situation s'explique par le fait qu'elles reproduisent les caractéristiques des données réelles. Un examen éthique et légal de ces aspects est donc essentiel. Alors que certaines organisations et organismes subventionnaires québécois et canadiens commencent à se positionner sur le sujet, une vigie des cadres et des principes qui en émergent s'avère particulièrement pertinente.

D'autre part, il y a les attitudes et les perceptions des usagères et des usagers quant à l'utilisation de leurs données. Une revue systématique sur l'utilisation secondaire et le partage des données clinico-administratives et d'essais cliniques (29) révèle que les personnes consultées sont généralement à l'aise

avec l'idée que les données de santé soient utilisées pour la recherche, surtout lorsqu'elles sont dépersonnalisées ou anonymisées. Selon cette revue, obtenir un consentement préalable n'est pas un prérequis pour plusieurs personnes. D'autres revues, dont une s'intéressant au même sujet spécifiquement en contexte d'utilisation par l'intelligence artificielle, identifie néanmoins la confidentialité et les risques de sécurité comme des freins au consentement (30–32). Les manquements dans le processus de consentement éclairé et le partage non-autorisé des données en font aussi partis (30). La participation des usagères et des usagers dans le développement, l'amélioration de la protection de la vie privée à travers l'anonymisation et la promotion des standards éthiques pour l'utilisation des données sont, à l'opposé, des leviers (30,32).

Ces différents constats suggèrent que des consultations des parties prenantes ainsi qu'une démonstration des indicateurs de vie privée seraient pertinentes pour soutenir les réflexions concernant l'acquisition et l'implantation de la méthode avatar au CIUSSS de l'Estrie – CHUS.

LIMITES

Les principales limites de cette revue rapide sont liées aux publications recensées et à la méthodologie employée.

Premièrement, la portée de cet état des connaissances est limitée au domaine de la santé et de la recherche. Il faut considérer que la méthode avatar d'Octopize est récente et qu'elle a été spécifiquement développée pour le domaine de la santé. De plus, aucune publication sur la méthode avatar n'a été repérée dans un autre domaine lors de la recherche documentaire. Toutefois, un élargissement de la stratégie de recherche sur la génération de données synthétiques à des domaines connexes pourraient potentiellement enrichir le portrait des considérants organisationnels, contextuels et économiques à considérer pour l'implantation et l'utilisation de la méthode.

Deuxièmement, aucune évaluation de la qualité des études n'a été réalisée. D'autre part, la majorité des publications recensées sur la méthode avatar reposent sur un faible nombre de groupes. Parmi la liste des autrices et auteurs de plusieurs publications évaluant la méthode, il y a des membres du personnel d'Octopize ou des personnes ayant collaboré avec eux sur de précédents travaux. Le risque de conflit d'intérêt que cela représente n'est pas systématiquement adressé. Pour une bonne proportion des publications incluses, aucune révision par les pairs n'a été réalisée au moment de cette revue. Il s'agit surtout de publications de la littérature grise, soit des articles de conférence et de prépublication. Par conséquent, des études supplémentaires seront nécessaires pour valider les résultats.

Par ailleurs, les écrits législatifs régissant l'utilisation et la protection des données ou renseignements personnels en vigueur au Québec et au Canada sont hors de la portée de cet état des connaissances.

Finalement, aucune publication décrivant les dimensions économiques ou organisationnelles n'a été repérée, ce qui fait que les considérants présentés excluent leurs aspects relatifs. Compte tenu du faible nombre de publications, un élargissement de la recherche documentaire serait pertinent pour obtenir un portrait plus complet des indicateurs et des considérants. Il pourrait, par exemple, s'élargir à l'utilisation de la méthode avatar ou d'autres techniques de génération des données synthétiques dans des domaines connexes tels que les recensements gouvernementaux, les assurances et autres.

5. CONCLUSION

À la condition d'un paramétrage adéquat, l'avatarisation des données semble être une alternative intéressante à l'anonymisation pour favoriser l'accès et l'utilisation des données de santé. La génération de données synthétisées par la méthode avatar peut servir pour le partage de données, le développement ou l'entraînement d'algorithmes et l'avancement des connaissances. Toutefois, l'utilité des données synthétisées est dépendant de l'évaluation de ses indicateurs. Par conséquent, il faut considérer le risque que toute autre analyse que celle pour laquelle le jeu de données a été optimisée peut conduire à des analyses erronées.

Le partage d'information sur la génération et l'évaluation des données synthétiques favorise leurs utilisations secondaires mais accroît les risques pour la protection de la vie privée.

Des études supplémentaires sont nécessaires pour confirmer ses résultats et évaluer les dimensions éthiques, juridiques, économiques et organisationnelles de l'implantation et de l'utilisation de la technologie d'avatarisation des données par Octopize.

ANNEXES

Annexe I. Tests et méthodes pour l'évaluation de la qualité de la génération de données synthétiques

Dimension	Famille de méthode	Test ou méthode d'évaluation	Utilisation ou interprétation
Utilité générale	Analyse univariée	Distance de Hellinger	Score entre 0 et 1 où plus la valeur s'approche de 1, plus les distributions sont différentes
		Test de Kolmogorov-Smirnov (KS)	La statistique D ou la valeur de p sont utilisées. La statistique D représente la distance maximale entre la distribution de l'échantillon et de la distribution théorique; plus elle est élevée, plus l'échantillon est éloigné de la distribution théorique. La valeur de p se situe entre 0 et 1; plus elle est proche de 1, plus les distributions considérées sont proches.
	Analyse bivariée	Coefficient de corrélation ou matrice de corrélation	Comparaison entre les données originales et les données synthétiques où les probabilités (valeur de p) sont souvent utilisées pour l'interprétation.
	Analyse multi-variée	Méthode d'analyse factorielle (FAMD, PCA, MCA)	Examen des charges factorielles, des valeurs, de la variance et des graphiques pour analyser la structure et l'adéquation entre les données originales et les données synthétiques.
	Divergence de Kullback-Leibler (Peut s'appliquer à des analyses uni, bi et multivariées)		Plus la mesure est proche de 0, plus les distributions sont proches. L'inverse de la distribution de Kullback-Leibler est fréquemment employé. Dans ce cas, plus la valeur est proche de 0, plus les distributions sont différentes.
Utilité spécifique	Précision des modèles	Modèle prédictif	Utile pour vérifier ou mesurer l'acuité des prédictions du jeu de données synthétiques comparativement aux originales.
	Similarité homomor-phique	Mesure statistique (ex. Odd ratio avec IC 95%)	La comparaison des statistiques mesurant la force d'association entre une exposition et un résultat est fréquemment utilisée. Les mesures retenues dépendent du contexte et des caractéristiques du jeu de données d'origine.
Protection de la vie privée	Individualisa-tion	Distance au plus proche voisin (DCR) et ratio de distance au plus proche voisin (NNDR)	Utilise les valeurs médianes; un risque d'individualisation est présent lorsque le DCR et le NNDR sont tous deux faibles.

Dimension	Famille de méthode	Test ou méthode d'évaluation	Utilisation ou interprétation
		Hidden Rate (HR)	Mesure spécifique à la méthode avatar, soit la probabilité de faire une erreur quand une attaque est faite pour associer un individu à son avatar.
		Local cloaking (LC)	La médiane sur tous les individus est utilisée pour évaluer un jeu de données; plus le nombre est élevé et plus les individus sont protégés.
	Corrélation	Taux de protection contre les corrélations	Correspond au pourcentage d'individus qui ne seraient pas reliés avec succès à leur homologue synthétique si l'attaquante ou l'attaquant utilisait une source de données externe
	Inférence	Attaque d'inférence des attributs (AIA)	Le taux de succès de l'attaque par inférence est plus fréquemment utilisé. Il est déterminé par le calcul de la précision (proportion d'attributs correctement inférés), le rappel (*sensitivity*) et le f1-score (moyenne entre précision et rappel). Des valeurs élevées indiquent une plus grande vulnérabilité du modèle aux attaques.
		Attaque d'inférence par appartenance (MIA)	

Source : (2,5,7,15,17,18,21)

Annexe II. Stratégie de recherche documentaire

La stratégie de recherche s'est déroulée en 3 volets successifs. La méthodologie de chaque volet et les stratégies de recherche pour chaque banque de données interrogée sont présentées ci-dessous :

VOLET 1 : RECHERCHE SPÉCIFIQUE SUR L'AVATARISATION DES DONNÉES

I. Méthodologie

1. **Concepts**
- Avatarisation
- Données

2. **Bases de données et date(s) d'interrogation**
- Medline (Ovid), 2025-06-03
- Embase (Ovid), 2025-06-04
- EBM Reviews - Cochrane Database of Systematic Reviews (Ovid), 2025-06-04
- EBM Reviews - Cochrane Central Register of Controlled Trials (Ovid), 2025-06-04
- IEEE Xplore, 2025-06-04
- CINAHL, 2025-06-04

3. **Limites**
- Chronologique : 2020 à 2025
- Linguistique : français, anglais

II. Stratégie(s) de recherche

1. **Stratégie(s) de recherche pour les bases de données**

1.1 Stratégie de recherche pour Ovid MEDLINE(R) ALL
Interrogée le 2025-06-03

Ovid MEDLINE(R) <1946 to May Week 4 2025>

1 Octopize.ab,kw,ti. 1
2 (Avatar* adj3 (method* or approach* or algorithm*)).ab,kw,ti. 43
3 Avatari?ation.ab,kw,ti. 1
4 1 or 2 or 3 44
5 routinely collected health data/ 320
6 data.ab,kf,ti. 4434230
7 5 or 6 4434239
8 4 and 7 6
9 limit 8 to ((english or french) and last 5 years) 2

1.2 Stratégie de recherche pour EMBASE (Ovid)
Interrogée le 2025-06-04
1 Octopize.ab,kw,ti. 3
2 (Avatar* adj3 (method* or approach* or algorithm*)).ab,kw,ti. 119
3 Avatari?ation.ab,kw,ti. 5
4 1 or 2 or 3 124
5 exp health data/ 685233
6 data.ab,kw,ti. 7257981
7 5 or 6 7693296

8 4 and 7 34
9 limit 8 to ((english or french) and last 5 years) 14

1.3 Stratégie de recherche pour EBM Reviews - Cochrane Database of Systematic Reviews (Ovid)

Interrogée le 2025-06-04
1 Octopize.ab,kw,ti. 0
2 (Avatar* adj3 (method* or approach* or algorithm*)).ab,kw,ti. 0
3 Avatari?ation.ab,kw,ti. 0
4 1 or 2 or 3 0
5 data.ab,kw,ti. 9031
6 4 and 5 0

1.4 Stratégie de recherche pour EBM Reviews - Cochrane Central Register of Controlled Trials (Ovid)

Interrogée le 2025-06-04
1 Octopize.ab,kw,ti. 0
2 (Avatar* adj3 (method* or approach* or algorithm*)).ab,kw,ti. 19
3 Avatari?ation.ab,kw,ti. 0
4 1 or 2 or 3 19
5 Routinely Collected Health Data/ 14
6 data.ab,kw,ti. 388888
7 5 or 6 388888
8 4 and 7 7
9 limit 8 to last 5 years 0

1.5. Stratégie de recherche pour IEEE Xplore

Interrogée le 2025-06-04
11 resultsfor ((((((No Keywords Specified))) AND ((Document Title:Octopize) OR (Abstract:Octopize) OR (Author "Keywords":Octopize))) OR ((Document Title:avatarisation OR Document Title:avatarization) OR (Abstract:avatarisation OR Abstract:avatarization) OR (Author "Keywords":avatarisation OR Author "Keywords":avatarization))) OR ((Document Title:avatar* ONEAR/3 (method OR Document Title:methods OR Document Title:approach OR Document Title: approaches OR Document Title:algorithm OR Document Title:algorithms)) OR (Abstract:avatar* ONEAR/3 (method OR Abstract:methods OR Abstract:approach OR Abstract: approaches OR Abstract:algorithm OR Abstract:algorithms)) OR (Author "Keywords":avatar* ONEAR/3 (method OR Author "Keywords":methods OR Author "Keywords":approach OR Author "Keywords": approaches OR Author "Keywords":algorithm OR Author "Keywords":algorithms)))) AND ((Document Title:data) OR (Abstract:data) OR (Author "Keywords":data)) Filters Applied: 2020 – 2025

1.6. Stratégie de recherche pour CINAHL (Ebsco)

Interrogée le 2025-06-03

#	Question	Opérateurs de restriction/Opérateurs d'expansion	Résultats
S10	S8 AND S9	Opérateurs d'expansion - Appliquer des sujets équivalents Modes de recherche - Proximité	4
S9		Opérateurs de restriction - Date de publication: 20200601-20250631; Langue: English, French Opérateurs d'expansion - Appliquer des sujets équivalents Modes de recherche - Proximité	2,213,945
S8	S6 AND S7	Opérateurs d'expansion - Appliquer des sujets équivalents Modes de recherche - Proximité	12
S7	S4 OR S5	Opérateurs d'expansion - Appliquer des sujets équivalents Modes de recherche - Proximité	1,147,010
S6	S1 OR S2 OR S3	Opérateurs d'expansion - Appliquer des sujets équivalents Modes de recherche - Proximité	25
S5	TI data OR AB data	Opérateurs d'expansion - Appliquer des sujets équivalents Modes de recherche - Proximité	1,142,511
S4	(MH "Data Management+")	Opérateurs d'expansion - Appliquer des sujets équivalents Modes de recherche - Proximité	10,679
S3	TI (avatarization OR avatarisation) OR AB (avatarization OR avatarisation)	Opérateurs d'expansion - Appliquer des sujets équivalents Modes de recherche - Proximité	1
S2	TI (Avatar* N3 (method* or approach* or algorithm*)) OR AB (Avatar* N3 (method* or approach* or algorithm*))	Opérateurs d'expansion - Appliquer des sujets équivalents Modes de recherche - Proximité	24
S1	TI octopize OR AB octopize	Opérateurs d'expansion - Appliquer des sujets équivalents Modes de recherche - Proximité	0

Au total, 31 références ont été trouvées et importées dans EndNote. Il reste 27 références après la suppression des doublons.

VOLET 2 : RECENSION DES ÉCRITS SUR LA GÉNÉRATION DE DONNÉES SYNTHÉTIQUES

I. Méthodologie

1. **Concepts**
   - Donnée synthétique
   - Synthèses des connaissances

2. **Bases de données et date(s) d'interrogation**
   - Medline (Ovid), 2025-06-09
   - Embase (Ovid), 2025-06-09
   - EBM Reviews - Cochrane Database of Systematic Reviews (Ovid), 2025-06-05
   - EEE Xplore, 2025-06-09
   - CINAHL, 2025-06-09

3. **Limites**
   - Chronologique : 2020 à 2025
   - Linguistique : français, anglais

II. Stratégie(s) de recherche

1. **Stratégie(s) de recherche pour les bases de données**

1.1 Stratégie de recherche pour Ovid MEDLINE(R) ALL
Interrogée le 2025-06-09

1 (synthetic adj3 (data or dataset or datasets)).ab,kf,ti. 6024
2 exp meta-analysis/ or exp meta-analysis as topic/ or "systematic review"/ or systematic reviews as topic/ or "scoping review"/ or "scoping review as topic"/ or exp Technology Assessment, Biomedical/ 433920
3 (systematic review or meta-analysis or scoping review).pt. 388069
4 ((systematic or scoping or umbrella*) adj2 (review* or overview* or synthes*)).ti,ab,kf. 330845
5 (meta-analy* or metaanaly* or metanaly* or meta synthes* or technology assessment* or technology overview* or technology appraisal*).ti,ab,kf. 289183
6 2 or 3 or 4 or 5 504999
7 1 and 6 61
8 limit 7 to ((english or french) and last 5 years) 38

1.2 Stratégie de recherche pour EMBASE (Ovid)
Interrogée le 2025-06-09
1 (synthetic adj3 (data or dataset or datasets)).ab,kf,ti. 10049
2 exp meta analysis/ 362733
3 exp "meta analysis (topic)"/ 58643
4 "systematic review"/ 530676
5 "systematic review (topic)"/ 37468
6 "scoping review"/ 7044
7 "scoping review (topic)"/ 23
8 exp biomedical technology assessment/ 19620
9 (systematic review or meta-analysis or scoping review).pt. 0
10 ((systematic or scoping or umbrella*) adj2 (review* or overview* or synthes*)).ti,ab,kf. 513102

11 (meta-analy* or metaanaly* or metanaly* or meta synthes* or technology assessment* or technology overview* or technology appraisal*).ti,ab,kf. 460156
12 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 935185
13 1 and 12 116
14 limit 13 to ((english or french) and last 5 years) 84

1.3 Stratégie de recherche pour EBM Reviews - Cochrane Database of Systematic Reviews (Ovid)

Interrogée le 2025-06-05
1 (synthetic adj3 (data or dataset or datasets)).ti,ab,kw. 0

1.4 Stratégie de recherche pour IEEE Xplore

Interrogée le 2025-06-09
Resultsfor ((((((No Keywords Specified))) AND ((Document Title:systematic ONEAR/2 review*) OR (Abstract:systematic ONEAR/2 review*) OR (Author "Keywords":systematic ONEAR/2 review*))) OR ((Document Title:scoping ONEAR/3 review*) OR (Abstract:scoping ONEAR/3 review*) OR (Author "Keywords":scoping ONEAR/3 review*))) OR ((Document Title:meta-analysis OR Document Title:metaanalysis OR Document Title:metanalysis OR Document Title:meta synthesis) OR (Abstract:meta-analysis OR Abstract:metaanalysis OR Abstract:metanalysis OR Abstract:meta synthesis) OR (Author "Keywords":meta-analysis OR Author "Keywords":metaanalysis OR Author "Keywords":metanalysis OR Author "Keywords":meta synthesis))) AND ((Document Title:synthetic ONEAR/3 data) OR (Document Title:synthetic ONEAR/3 dataset) OR (Document Title:synthetic ONEAR/3 datasets))

1.5 Stratégie de recherche pour CINAHL (Ebsco)

Interrogée le 2025-06-09

#	Question	Opérateurs de restriction/Opérateurs d'expansion	Résultats
S11	S9 AND S10	Opérateurs d'expansion - Appliquer des sujets équivalents Modes de recherche - Proximité	12
S10		Opérateurs de restriction - Date de publication: 20200601-20250631; Langue: English, French Opérateurs d'expansion - Appliquer des sujets équivalents Modes de recherche - Proximité	2,218,173
S9	S1 AND S8	Opérateurs d'expansion - Appliquer des sujets équivalents Modes de recherche - Proximité	17
S8	S2 OR S3 OR S4 OR S5 OR S6 OR S7	Opérateurs d'expansion - Appliquer des sujets équivalents Modes de recherche - Proximité	313,161
S7	TI (meta-analy* or metaanaly* or metanaly* or meta synthes* or technology assessment* or technology overview* or technology appraisal*) OR AB (meta-analy* or metaanaly* or metanaly* or meta synthes* or technology assessment* or technology overview* or technology appraisal*)	Opérateurs d'expansion - Appliquer des sujets équivalents Modes de recherche - Proximité	139,911

S6	TI ((systematic OR scoping OR umbrella*) N2 (review* OR overview* OR synthes*)) OR AB ((systematic OR scoping OR umbrella*) N2 (review* OR overview* OR synthes*))	Opérateurs d'expansion - Appliquer des sujets équivalents Modes de recherche - Proximité	200,111
S5	PT systematic review OR PT meta-analysis OR PT scoping review	Opérateurs d'expansion - Appliquer des sujets équivalents Modes de recherche - Proximité	173,958
S4	(MH "Scoping Review")	Opérateurs d'expansion - Appliquer des sujets équivalents Modes de recherche - Proximité	13,103
S3	(MH "Systematic Review")	Opérateurs d'expansion - Appliquer des sujets équivalents Modes de recherche - Proximité	145,591
S2	(MH "Meta Analysis")	Opérateurs d'expansion - Appliquer des sujets équivalents Modes de recherche - Proximité	78,495
S1	TI (synthetic N3 (data or dataset or datasets)) OR AB (synthetic N3 (data or dataset or datasets))	Opérateurs d'expansion - Appliquer des sujets équivalents Modes de recherche - Proximité	498

Au total, 137 références ont été trouvées et importées dans EndNote. Il reste 93 références après la suppression des doublons.

VOLET 3 : RECENSION DES ÉCRITS SUR L'ANONYMISATION DES DONNÉES

I. Méthodologie

1. **Concepts**

- Anonymisation des données
- Confidentialité des données
- Synthèses des connaissances

2. **Bases de données et date(s) d'interrogation**

- Medline (Ovid), 2025-06-13
- Embase (Ovid), 2025-06-13
- EBM Reviews - Cochrane Database of Systematic Reviews (Ovid), 2025-06-13
- CINAHL, 2025-06-13

3. **Limites**

- Chronologique : 2020 à 2025
- Linguistique : français, anglais

II. Stratégie(s) de recherche

1. **Stratégie(s) de recherche pour les bases de données**

1.1 Stratégie de recherche pour Ovid MEDLINE(R) ALL

Interrogée le 2025-06-13

1 ((anonym* or "de-identification" or "de-identified" or deidentif*) adj3 (data or dataset or datasets)).ab,kw,ti. 8951
2 exp Privacy/ or exp Confidentiality/ 70892
3 (privacy or confidential* or "re-identification" or reidentif*).ab,kw,ti. 50267
4 2 or 3 106279
5 exp meta-analysis/ or exp meta-analysis as topic/ or "systematic review"/ or systematic reviews as topic/ or "scoping review"/ or "scoping review as topic"/ or exp Technology Assessment, Biomedical/ 454819
6 (systematic review or meta-analysis or scoping review).pt. 408734
7 ((systematic or scoping or umbrella*) adj2 (review* or overview* or synthes*)).ti,ab,kf. 432715
8 (meta-analy* or metaanaly* or metanaly* or meta synthes* or technology assessment* or technology overview* or technology appraisal*).ti,ab,kf. 363531
9 5 or 6 or 7 or 8 634324
10 1 and 4 and 9 42
11 limit 10 to ((english or french) and last 5 years) 26

1.2 Stratégie de recherche pour EMBASE (Ovid)

Interrogée le 2025-06-13

1 ((anonym* or "de-identification" or "de-identified" or deidentif*) adj3 (data or dataset or datasets)).ab,kw,ti. 19867
2 data privacy/ or confidential information/ 3006
3 (privacy or confidential* or "re-identification" or reidentif*).ab,kw,ti. 65356
4 2 or 3 66362
5 exp meta analysis/ or exp "meta analysis (topic)"/ or "systematic review"/ or "systematic review (topic)"/ or "scoping review"/ or "scoping review (topic)"/ or exp biomedical technology assessment/ 766473
6 (systematic review or meta-analysis or scoping review).pt. 0

7 ((systematic or scoping or umbrella*) adj2 (review* or overview* or synthes*)).ti,ab,kf. 514057
8 (meta-analy* or metaanaly* or metanaly* or meta synthes* or technology assessment* or technology overview* or technology appraisal*).ti,ab,kf. 460819
9 5 or 6 or 7 or 8 936620
10 1 and 4 and 9 84
11 limit 10 to ((english or french) and last 5 years) 52

1.3 Stratégie de recherche pour EBM Reviews - Cochrane Database of Systematic Reviews (Ovid)

Interrogée le 2025-06-13
1 ((anonym* or "de-identification" or "de-identified" or deidentif*) adj3 (data or dataset or datasets)).ti,ab,kw. 0

1.4 Stratégie de recherche pour CINAHL (Ebsco)

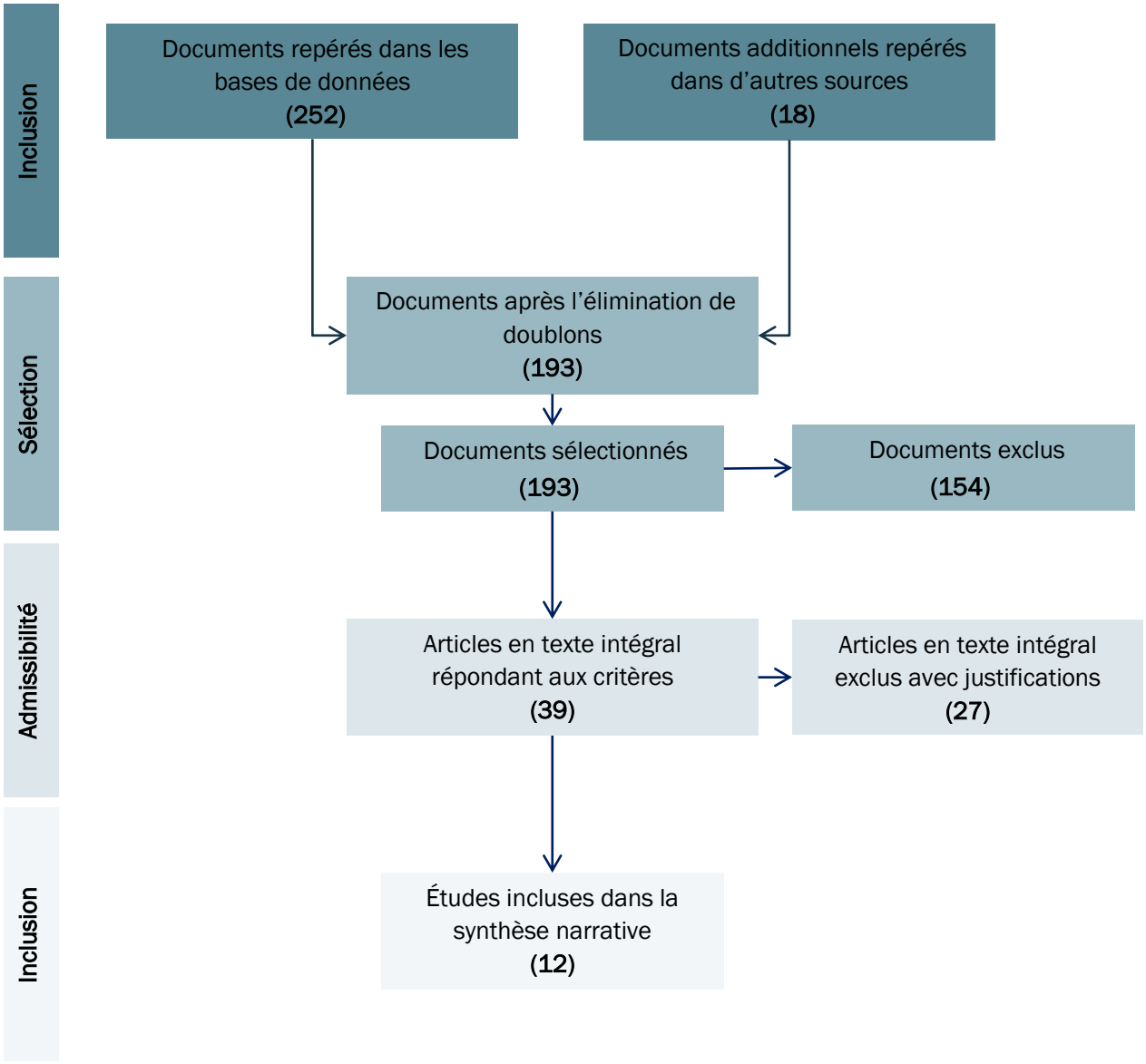
Interrogée le 2025-06-13

#	Question	Opérateurs de restriction/Opérateurs d'expansion	Résultats
S1	TI ((anonym* OR "de-identification" OR "de-identified" OR deidentif*) N3 (data OR dataset OR datasets)) OR AB ((anonym* OR "de-identification" OR "de-identified" OR deidentif*) N3 (data OR dataset OR datasets))	Opérateurs d'expansion - Appliquer des sujets équivalents Modes de recherche - Proximité	3,743
S2	(MM "Privacy and Confidentiality+")	Opérateurs d'expansion - Appliquer des sujets équivalents Modes de recherche - Proximité	10,504
S3	TI (privacy OR confidential* OR "re-identification" OR reidentification) OR AB (privacy OR confidential* OR "re-identification" OR reidentification)	Opérateurs d'expansion - Appliquer des sujets équivalents Modes de recherche - Proximité	21,673
S4	S2 OR S3	Opérateurs d'expansion - Appliquer des sujets équivalents Modes de recherche - Proximité	26,559
S5	(MH "Meta Analysis") OR (MH "Systematic Review") OR (MH "Scoping Review") OR PT systematic review OR PT meta-analysis OR PT scoping review	Opérateurs d'expansion - Appliquer des sujets équivalents Modes de recherche - Proximité	208,645
S6	TI ((systematic OR scoping OR umbrella*) N2 (review* OR overview* OR synthes*)) OR AB ((systematic OR scoping OR umbrella*) N2 (review* OR overview* OR synthes*))	Opérateurs d'expansion - Appliquer des sujets équivalents Modes de recherche - Proximité	200,537
S7	TI (meta-analy* OR metaanaly* OR metanaly* OR meta synthes* OR technology assessment* OR technology overview* OR technology	Opérateurs d'expansion - Appliquer des sujets équivalents	140,179

	appraisal*) OR AB (meta-analy* OR metaanaly* OR metanaly* OR meta synthes* OR technology assessment* OR technology overview* OR technology appraisal*)	Modes de recherche - Proximité	
S8	S5 OR S6 OR S7	Opérateurs d'expansion - Appliquer des sujets équivalents Modes de recherche - Proximité	313,674
S9	S1 AND S4 AND S8	Opérateurs d'expansion - Appliquer des sujets équivalents Modes de recherche - Proximité	10
S10		Opérateurs de restriction - Date de publication: 20200601-20250631; Langue: English, French Opérateurs d'expansion - Appliquer des sujets équivalents Modes de recherche - Proximité	2,223,245
S11	S9 AND S10	Opérateurs d'expansion - Appliquer des sujets équivalents Modes de recherche - Proximité	6

Au total, 84 références ont été trouvées et importées dans EndNote. Il reste 58 références après la suppression des doublons.

Annexe III. Diagramme de flux du processus de sélection des études



Annexe IV. Publications exclues avec justification

Auteurs(-trices), année	Titre	Raison d'exclusion
Alloza et al., 2023	A Case for Synthetic Data in Regulatory Decision-Making in Europe	SUJET : Publications portant exclusivement sur les autres méthodes de synthèse de données
Arora et al., 2025	R WE ready for reimbursement? A round up of developments in real-world evidence relating to health technology assessment: part 18	Ne corresponds pas au CHIP
Austin et al., 2024	Decades in the Making: The Evolution of Digital Health Research Infrastructure Through Synthetic Data, Common Data Models, and Federated Learning	SUJET : Publications portant exclusivement sur les autres méthodes de synthèse de données
Bachot et al., 2022	RWD106 Statistical Analysis Reproducibility Using Different Anonymization Techniques	Résumé de conférence
Benevento et al., 2023	Measuring the willingness to share personal health information: a systematic review	Ne corresponds pas au CHIP
Bouras, 2024	The Emerging Applications of Synthetic Data in Neurosurgery Research and Practice: A Systematic Review	Ne corresponds pas au CHIP
Chasseloup et al., 2023	Generation and application of avatars in pharmacometric modelling	SUJET: Avatarisation des patientes ou patients (avatar patients)
Demuth et al., 2025	Les patients virtuels pour substituer les données de référence sensibles dans une architecture de données de médecine de précision pour la sclérose en plaques	Résumé de conférence
Haber, Sax, et Prasser, 2022	Open tools for quantitative anonymization of tabular phenotype data: literature review	SUJET : Publications portant exclusivement sur les autres méthodes de synthèse de données
Hernandez et al., 2022	Synthetic data generation for tabular health records: A systematic review	SUJET : Publications portant exclusivement sur les autres méthodes de synthèse de données
Hutchings et al., 2021	A systematic literature review of attitudes towards secondary use and sharing of health administrative and clinical trial data: a focus on consent	Ne corresponds pas au CHIP
Ibrahim, et al., 2025	Generative AI for synthetic data across multiple medical modalities: A systematic review of recent developments and challenges	SUJET : Publications portant exclusivement sur les autres méthodes de synthèse de données
Le Gall et al., 2023	Génération de données synthétiques de marche: application au cas de patients atteints de sclérose en plaques	Publication sans comparaison ou résultat sur la méthode avatar, lettre d'opinion, résumé de conférence

Liu, Acharya et Tan, 2025	Preserving privacy in healthcare: A systematic review of deep learning approaches for synthetic data generation	SUJET : Publications portant exclusivement sur les autres méthodes de synthèse de données
Loni et al., 2025	A review on generative AI models for synthetic medical text, time series, and longitudinal data	SUJET : Publications portant exclusivement sur les autres méthodes de synthèse de données
Martorana et al., 2022	Aligning restricted access data with FAIR: a systematic review	Ne corresponds pas au CHIP
Moulaei, Akhlaghpour et Fatehi, 2025	Patient consent for the secondary use of health data in artificial intelligence (AI) models: A scoping review	Ne corresponds pas au CHIP
Negash et al., 2023	De-identification of free text data containing personal health information: a scoping review of reviews	Ne corresponds pas au CHIP
Pezoulas et al., 2024	Synthetic data generation methods in healthcare: A review on open-source tools and methods	SUJET : Publications portant exclusivement sur les autres méthodes de synthèse de données
Rujas et al., 2025	Synthetic data generation in healthcare: A scoping review of reviews on domains, motivations, and future applications	SUJET : Publications portant exclusivement sur les autres méthodes de synthèse de données
Sepas et al., 2022	Algorithms to anonymize structured medical and healthcare data: A systematic review	Ne corresponds pas au CHIP
Thomason, 2024	Data, digital worlds, and the avatarization of health care	Lettre d'opinion
Tsao et al., 2023	Health Synthetic Data to Enable Health Learning System and Innovation: A Scoping Review	Ne corresponds pas au CHIP
Vo et al., 2023	Multi-stakeholder preferences for the use of artificial intelligence in healthcare: A systematic review and thematic analysis	Ne corresponds pas au CHIP
Zanchi et al., 2025	Synthetic ECG signals generation: A scoping review	SUJET : Publications portant exclusivement sur les autres méthodes de synthèse de données
Laribi et al., 2024	Leveraging patients' longitudinal data to improve the Hospital One-year Mortality Risk	Publication sans comparaison ou résultat suffisant sur la méthode avatar, lettre d'opinion, résumé de conférence
Floyrac et al. 2023	Notebook pédagogique	Outil pédagogique

Annexe V. Description des publications incluses présentant des données sur la méthode avatar

Publication	Devis	Utilisation de la méthode avatar		
		Type de données	Méthode comparée	Résultat d'intérêt
Arcay, 2025 (2) France	Mémoire présentant une étude évaluative comparative avec cas pratique (Utilise AnonymHUS[2] basée sur la méthode Avatar)	Données tabulaires (pharmacologie)	• CT-GAN • Synthpop • TVAE	• Utilité générale et spécifique • Protection de la vie privée
Barreteau et al., 2023 (22) France	Article de méthodologie avec cas d'usage présentant une alternative à la méthode avatar pour le calcul des poids	Signaux ECG	Alternative pour le calcul des poids	• Utilité • Protection de la vie privée
Bennis et Gourraud, 2021 (18) France	Article de conférence pour l'ArabWIC en 2021 qui présente Chronos, une extension de la méthode avatar avec un cas d'usage **(Prépublication♦)**	Signaux ECG (séries temporelles)	• Chronos	• Utilité générale et spécifique • Protection de la vie privée
Benoist et al., 2024 (15) France	Évaluation comparative de l'effet de la taille du jeu de données sur la génération de données synthétiques (Partie du projet DIGPHAT)	Données catégorielles et continues (pharmacologie)	• CT-GAN	• Utilité générale et spécifique • Protection de la vie privée • Effet de la taille du jeu de données
Demuth et al., 2024 (20) France	Étude de la performance de la méthode avatar pour la génération de données d'ECR synthétiques **(Prépublication♦)**	ECR	N/A	• Utilité générale et spécifique • Protection de la vie privée • Variation des paramètres (k, ncp, poids, gestion des données manquantes)
Fadel et al., 2024 (16) France	Étude évaluative pour la génération de données de cohortes synthétiques	Données catégorielles et continues	N/A	• Utilité spécifique • Représentativité
Guillaudeux et al., 2023 (14) France	Article méthodologique avec cas d'usage comparatif (Publication originale de la méthode avatar)	Données tabulaires et longitudinales (ECR et séries temporelles)	• Synthpop • CT-GAN	• Utilité générale et spécifique • Protection de la vie privée • Variation des paramètres (k)

[2] Selon Arcay, AnonymHUS est une interface développée à partir de la publication originale de la méthode avatar par Guillaudeux et al. (14).

Publication	Devis	Utilisation de la méthode avatar		
		Type de données	Méthode comparée	Résultat d'intérêt
Laribi et al., 2024 (19) Canada	Article de conférence présenté au MAIS en 2024 présente l'évaluation des données synthétiques générées pour l'étude originale (33) **(Prépublication♦)**	Données tabulaires	N/A	• Utilité spécifique
Lebrun et al., 2024 (3) France, Canada	Étude évaluative de la performance de différentes méthodes de génération de données synthétiques et d'anonymisation Propose une méthode alternative, M-Avatar, à la méthode avatar.	Données tabulaires	• M-Avatar • SAIPH • Algorithme MST • Synthpop • CT-GAN • K-anonymisation	• Utilité générale et spécifique • Protection de la vie privée
Pau et al., 2023 (21) France	Étude évaluative sur la reproductivité des analyses statistiques de différentes méthodes d'anonymisation **(Prépublication♦)**	Données tabulaires	• PARM • Solution en libre service • Méthode de généralisation	• Utilité générale et spécifique • Représentativité • Protection de la vie privée
Rousseau et al., 2021 (23) France	Étude observationnelle sur les facteurs de risque pour les anévrismes dans la cohorte du projet ICAN (Comparaison des données avatar avec celles originales dans les suppléments de la publication)	Données tabulaires	N/A	• Utilité générale et spécifique
Woillard et al., 2025 (17) (France)	Étude évaluative de la performance de différentes méthodes de génération de données synthétiques (Utilise une version allégée de la méthode avatar)	Données tabulaires et catégorielles	• Synthpop • CT-GAN • TVAE	• Utilité • Protection de la vie privée • Augmentation des données

Abréviations

ArabWIC : Arab Women in Computing, CT-GAN : Réseau antagoniste génératif conditionnel, ECG : Électrocardiogramme, ECR : Essais contrôlés randomisés, k : Nombre de voisins, MAIS : *Montreal Artificial Intelligence* Symposium, MST : Maximum Spanning Tree, N/A : Ne s'applique pas, ncp : Nombre de dimensions projetées pour le calcul des distances des voisins, PARM : *Probabilistic Autoregressive Model,* TVAE : Auto-encodeurs variationnels tabulaires.

Note

♦ Prépublication : La publication n'a pas été révisée par les pairs et par conséquent, ces résultats sont à considérer et interpréter avec prudence.

Annexe VI. Effets de la variation des paramètres de la méthode avatar sur les données synthétisées

Six des douze publications incluses examinent l'effet de la variation de certains paramètres de la méthode avatar sur la génération de données. Voici les paramètres testés par chacune de ses publications.

	k	Nombre de composantes	ncp	Poids variés	Loi de distribution des poids	Augmentation du jeu de données	Mode de gestion des données manquantes
Arcay, 2025 (2)	✓	✓					
Barreteau, 2023 (22)	✓				✓		
Benoist et al., 2024 (15)	✓					✓	
Demuth et al., 2024 (20)	✓		✓	✓			✓
Guillaudeux et al., 2023 (14)	✓						
Woillard et al., 2025 (17)	✓					✓	

Abréviations

k : Nombre de voisins, ncp : Nombre de dimensions projetées pour le calcul des distances des voisins

Annexe VII. Description des paramètres de la méthode avatar et des jeux de données originaux utilisés dans les publications incluses

Publication	Jeu de données originales	Paramètres employés pour la méthode avatar
Arcay (2)	Données clinico-admistratives : 916 individus	• Avatar 1 : 20 composantes, k = 5 • Avatar 2 : 35 composantes, k=10
Barreteau et al. (22)	1000 signaux ECG de la dérivation II d'une durée de 5 secondes	• k = 10 et 20 • Loi de distribution des poids alternative : Loi de Dirichlet
Bennis et Gourraud (18)	Échantillons de 125Hz et d'une longueur de 188 provenant de 2 jeux de données de signaux ECG: • MIT-BIH Arrythmia : 18 630 signaux, 5 catégories • PTB Diagnostic ECG: 11 550 signaux, 2 catégories	• Chronos, adaptation de la méthode avatar pour les séries temporelles
Benoist et al. (15)	2 jeux de données catégorielles et continues: • Pharmacocinétique de population : 29 observations, 18 variables • Effet sur l'exposition: 1573 observations, 11 variables	• Augmentation : 1x et 4x la taille du jeu de données • k = 10 et 20
Demuth et al. (20)	Données catégorielles et quantitatives d'essais contrôlés randomisés (ECR) • ECR Clarity : 864 observations et 35 variables (7 catégorielles et 28 quantitatives) • ECR ADVANCE : 1512 observations et 25 variables (8 catégorielles et 17 quantitatives)	• Optimisation des jeux de données o k = 2, 5, 10, 15, 20, 25, 30, 40, 50, 75, 100, et 150 o ncp = 5, 10, 20, 30, 46 et la valeur maximale selon le jeu de données originales o Configuration sans poids et avec variation du poids des variables o Gestion des valeurs quantitatives manquantes non imputée par le serveur avatar, choix des traiter comme des valeurs négatives aberrantes • Examen de la performance sur les jeux de données optimisés o ECR CLARITY: k = 5, ncp = 5, poids de 20 pour la variable d'intérêt, calcul booléen des voisins o ECR ADVANCE : k = 2, ncp = 10, poids de 20, gestion des données manquantes comme valeurs aberrantes négatives

Publication	Jeu de données originales	Paramètres employés pour la méthode avatar
Fadel et al. (16)	Données catégorielles et continues de cohortes : 62 434 observations et 22 variables	• k= 20 • ncp= 20 • poids de 10 ajouté aux variables
Guillaudeux et al. (14)	• Essai clinique : 2139 individus, 26 variables • Étude observationnelle : 683 observations, 10 variables	• Démonstration de la méthode avatar : ○ 25 générations d'avatar avec K= 20 pour chaque jeu de données ○ 10 générations d'avatar avec des k de 4 à 1200 pour l'essai clinique et de 4 à 334 pour l'étude observationnelle • Comparaison: 10 échantillons de 70% des jeux de données originales par méthode
Laribi et al. (19)	Données clinico-administratives : 123 646 individus, 244 variables • Pour leur étude, deux jeux de prédicteurs ont été utilisés : un avec 12 variables uniquement et l'autre avec les 244 variables	Non spécifié
Lebrun et al. (3)	Données d'essai clinique : 2 139 individus, 26 variables	• Jeu de données séparé également en deux (moitié pour la génération des données, l'autre comme contrôle) • k= 20 • Distance de 5
Pau et al. (21)	Données d'une étude longitudinale (cohorte): 315 individus, 64 variables	Non spécifié
Rousseau et al. (23)	Données de cohorte (Les données synthétiques ne sont pas le sujet de l'étude)	Non spécifié
Woillard et al. (17)	Données longitudinales • 253 individus	• Augmentation : 1x et 4x la taille du jeu de données • k = 5, 10 et 20

Abréviations ECG : Électrocardiogramme, ECR : essai contrôlé randomisé, k : nombre de voisins, ncp : nombre de dimension utilisée pour l'identification des voisins (ncp), PTB : Physikalisch Technische Bundesanstalt, x : fois.

BIBLIOGRAPHIE

1. Université de Sherbrooke. Le numérique de la santé à l'UdeS [Internet]. [cité 11 août 2025]. Disponible sur: https://www.usherbrooke.ca/medecine/recherche/regroupements-de-recherche/psprt/numerique-de-la-sante

2. Arcay E. Gouvernance des données synthétiques obtenues par l'intelligence artificielle en santé [Internet]. Université de Strasbourg; 2025. Disponible sur: https://publication-theses.unistra.fr/public/theses_exercice/PHA/2025/2025_ARCAY_Emmanuelle.pdf

3. Lebrun T, Béziaud L, Allard T, Boutet A, Gambs S, Maouche M. Synthetic Data: Generate Avatar Data on Demand. In Doha, Qatar: Springer-Verlag; 2024. p. 193-203. Disponible sur: https://doi.org/10.1007/978-981-96-0576-7_15

4. Qu'est-ce que l'augmentation des données ? | IBM [Internet]. 2024 [cité 11 août 2025]. Disponible sur: https://www.ibm.com/fr-fr/think/topics/data-augmentation

5. Kaabachi B, Despraz J, Meurers T, Otte K, Halilovic M, Kulynych B, et al. A Scoping Review of Privacy and Utility Metrics in Medical Synthetic Data. medRxiv [Internet]. 2023;28. Disponible sur: http://linksolver.ovid.com/OpenUrl/LinkSolver?sid=OVID:embase&id=pmid:&id=doi:10.1101%2F2023.11.28.23299124&issn=&isbn=&volume=&issue=&spage=&pages=&date=2023&title=medRxiv&atitle=A+Scoping+Review+of+Privacy+and+Utility+Metrics+in+Medical+Synthetic+Data&aulast=Kaabachi

6. Pilgram L, El Emam K. Applications pratiques de la génération de données synthétiques. Futur Stat Off. 8 sept 2025;1-13.

7. Petot J, Barreteau AF. Octpize. 2023 [cité 21 août 2025]. Comment évaluer l'utilité des données synthétiques ? - Blog - Octopize. Disponible sur: https://www.octopize.io/fr/blog-posts/comment-evaluer-lutilite-des-donnees-synthetiques

8. Bhanot K, Qi M, Erickson JS, Guyon I, Bennett KP. The Problem of Fairness in Synthetic Healthcare Data. Entropy. 4 sept 2021;23(9):1165.

9. Finck M, Pallas F. They who must not be identified—distinguishing personal from non-personal data under the GDPR. Int Data Priv Law. 1 févr 2020;10(1):11-36.

10. Institut national d'excellence en santé et services sociaux (INESSS). Guide de soutien à l'appréciation de la valeur. 2024.

11. Forrester MA, Forrester MA, éditeurs. Doing qualitative research in psychology: a practical guide. Repr. Los Angeles: Sage; 2010. 262 p.

12. McHugh ML. Interrater reliability: the kappa statistic. Biochem Medica. 2012;22(3):276-82.

13. Stone EG. Evidence-Based Medicine and Bioethics: Implications for Health Care Organizations, Clinicians, and Patients. Perm J. déc 2018;22(4):18-030.

14. Guillaudeux M, Rousseau O, Petot J, Bennis Z, Dein CA, Goronflot T, et al. Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis. NPJ Digit Med. 20230310e éd. 10 mars 2023;6(1):37.

15. Benoist c, Marquet P, Stanke-Labesque f, Woillard JB. Synthèse de données par la méthode Avatar : anonymisation et fidélité en pharmacologie de la transplantation [Internet]. [cité 2 juin 2025]. Disponible sur: https://jds2024.sciencesconf.org/530702/document

16. Fadel M, Petot J, Gourraud PA, Descatha A. Flexibility of a large blindly synthetized avatar database for occupational research: Example from the CONSTANCES cohort for stroke and knee pain. PLoS One. 20240731e éd. 2024;19(7):e0308063.

17. Woillard JB, Benoist C, Destere A, Labriffe M, Marchello G, Josse J, et al. To be or not to be, when synthetic data meet clinical pharmacology: A focused study on pharmacogenetics. CPT Pharmacomet Syst Pharmacol. 20241016e éd. janv 2025;14(1):82-94.

18. Bennis Z, Gourraud PA. Application of a novel Anonymization Method for Electrocardiogram data. In: The 7th Annual International Conference on Arab Women in Computing in Conjunction with the 2nd Forum of Women in Research [Internet]. Sharjah United Arab Emirates: ACM; 2021 [cité 25 août 2025]. p. 1-5. Disponible sur: https://dl.acm.org/doi/10.1145/3485557.3485581

19. Laribi H, Raymond N, Taseen R, Poenaru D, Vallières M. Synthetic Data for Accessible Learning in Healthcare: Improving Mortality Prediction with Longitudinal Data [Internet]. 2024 [cité 25 août 2025]. Disponible sur: https://www.researchsquare.com/article/rs-5363467/v1

20. Demuth S, Rousseau O, Faddeenkov I, Paris J, Sèze J, Baciotti B, et al. Privacy-by-design generation of two virtual clinical trials in multiple sclerosis and their release as open datasets. 2024;

21. Pau D, Bachot C, Monteil C, Vinet L, Boucher M, Planchet E, et al. Comparison of Anonymization Techniques Regarding Statistical Reproducibility [Internet]. 2023 [cité 25 août 2025]. Disponible sur: https://www.ssrn.com/abstract=4516197

22. Barreteau AF, Regnier-Coudert O, Le Carpentier E, Moussaoui S. Génération de signaux anonymes à partir de données non anonymes par modèle de mélange linéaire local. In 2023. p. 1169-72. Disponible sur: https://hal.science/hal-00667856/

23. Rousseau O, Karakachoff M, Gaignard A, Bellanger L, Bijlenga P, Constant Dit Beaufils P, et al. Location of intracranial aneurysms is the main factor associated with rupture in the ICAN population. J Neurol Neurosurg Psychiatry. févr 2021;92(2):122-8.

24. Russeil G, Barreteau AF. Octopize. 2023 [cité 26 août 2025]. Comprendre le coeur de la méthode avatar - Blog Octopize. Disponible sur: https://www.octopize.io/fr/blog-posts/comprendre-le-coeur-de-la-methode-avatar

25. Crasset T, Regnier-Coudert O. Octopize. 2023 [cité 13 nov 2025]. Évaluation de la confidentialité d'un jeu de données - Blog - Octopize. Disponible sur: https://www.octopize.io/fr/blog-posts/evaluation-de-la-confidentialite-dun-jeu-de-donnees

26. Données de santé artificielles : Analyse et pistes de réflexion [Internet]. 91 p. Disponible sur: https://static.botdesign.net/docs/VF_Livre_blanc_Donn%C3%A9es_de_sant%C3%A9_artificielles-250424.pdf

27. Tsao SF, Sharma K, Noor H, Forster A, Chen H. Health Synthetic Data to Enable Health Learning System and Innovation: A Scoping Review. Stud Health Technol Inform. 2023;302:53-7.

28. Rujas M, Martín Gómez Del Moral Herranz R, Fico G, Merino-Barbancho B. Synthetic data generation in healthcare: A scoping review of reviews on domains, motivations, and future applications. Int J Med Inf. 20241217e éd. mars 2025;195:105763.

29. Hutchings E, Loomes M, Butow P, Boyle FM. A systematic literature review of attitudes towards secondary use and sharing of health administrative and clinical trial data: a focus on consent. Syst Rev [Internet]. déc 2021;10(1) (no pagination). Disponible sur: http://linksolver.ovid.com/OpenUrl/LinkSolver?sid=OVID:embase&id=pmid:33941282&id=doi:10.1186%2Fs13643-021-01663-z&issn=2046-4053&isbn=&volume=10&issue=1&spage=132&pages=&date=2021&title=Systematic+Reviews&atitle=A+systematic+literature+review+of+attitudes+towards+secondary+use+and+sharing+of+health+administrative+and+clinical+trial+data%3A+a+focus+on+consent&aulast=Hutchings

30. Moulaei K, Akhlaghpour S, Fatehi F. Patient consent for the secondary use of health data in artificial intelligence (AI) models: A scoping review. Int J Med Inf [Internet]. juin 2025;198(no pagination). Disponible sur: http://linksolver.ovid.com/OpenUrl/LinkSolver?sid=OVID:embase&id=pmid:40107041&id=doi:10.1016%2Fj.ijmedinf.2025.105872&issn=1386-5056&isbn=&volume=198&issue=&spage=105872&pages=&date=2025&title=International+Journal+of+Medical+Informatics&atitle=Patient+consent+for+the+secondary+use+of+health+data+in+artificial+intelligence+%28AI%29+models%3A+A+scoping+review&aulast=Moulaei

31. Benevento M, Mandarelli G, Carravetta F, Ferorelli D, Caterino C, Nicoli S, et al. Measuring the willingness to share personal health information: a systematic review. Front Public Health. 2023;11:1213615.

32. Vo V, Chen G, Aquino YSJ, Carter SM, Do QN, Woode ME. Multi-stakeholder preferences for the use of artificial intelligence in healthcare: A systematic review and thematic analysis. Soc Sci Med [Internet]. déc 2023;338(no pagination). Disponible sur: http://linksolver.ovid.com/OpenUrl/LinkSolver?sid=OVID:embase&id=pmid:37949020&id=doi:10.1016%2Fj.socscimed.2023.116357&issn=0277-9536&isbn=&volume=338&issue=&spage=116357&pages=&date=2023&title=Social+Science+and+Medicine&atitle=Multi-stakeholder+preferences+for+the+use+of+artificial+intelligence+in+healthcare%3A+A+systematic+review+and+thematic+analysis&aulast=Vo

33. Laribi H, Raymond N, Taseen R, Poenaru D, Vallières M. Leveraging patients' longitudinal data to improve the Hospital One-year Mortality Risk. Health Inf Sci Syst. 4 mars 2025;13(1):23.

Centre intégré
universitaire de santé
et de services sociaux
de l'Estrie – Centre
hospitalier universitaire
de Sherbrooke

Québec