Nº 11-522-X au catalogue
ISSN 1709-822X

Recueil du Symposium de 2024 de Statistique Canada : Le futur des statistiques officielles

Applications pratiques de la génération de données synthétiques

par Lisa Pilgram et Khaled El Emam

Date de diffusion : le 8 septembre 2025

Statistique Canada Statistics Canada

Canada

Recueil du Symposium de 2024 de Statistique Canada
Le futur des statistiques officielles

Applications pratiques de la génération de données synthétiques

Lisa Pilgram et Khaled El Emam[1]

Résumé

Les différents secteurs utilisent de plus en plus la génération de données synthétiques (GDS) pour partager et augmenter des données tout en protégeant la vie privée ainsi que pour en éliminer les biais. Chaque fois qu'on utilise cet outil, on doit avoir recours à un ensemble distinct de mesures d'évaluation pour tenir compte de la stochasticité du processus de GDS : la vulnérabilité liée à la divulgation des membres et des attributs est essentielle pour protéger la vie privée; la fidélité et l'utilité des tâches en aval s'appliquent de manière plus générale; l'équité et la diversité sont des facteurs pertinents pour éliminer les biais et augmenter les données, respectivement. En présentant les données probantes accumulées et en examinant des études de cas modèles, il est montré que la GDS peut offrir un bon rendement dans bon nombre de ces cas d'utilisation. Dans le présent document les principaux enseignements tirés des expériences en matière de données synthétiques sur la santé sont également partagés.

Mots clés : Données synthétiques; protection de la vie privée; partage des données; biais.

1. Introduction

Les défis liés à l'accès aux données se sont accentués avec l'accroissement de la demande en matière de données, particulièrement en ce qui concerne les données comprenant des renseignements d'identification personnelle. Pour être en mesure d'élargir l'accès à ce type de données, il importe de répondre à certaines préoccupations en matière de protection de la vie privée. Divers secteurs optent de plus en plus pour la génération de données synthétiques (GDS) comme méthode de partage de données dans le respect de la vie privée (El Emam et Hoptroff, 2020; van Breugel et coll., 2024). Les données synthétiques désignent de fausses données qui reflètent des distributions conjointes réelles, mais qui, si elles sont bien générées, ne contiennent pas de données personnelles réelles.

Même s'il existe encore certaines incertitudes concernant la réglementation contextuelle des données synthétiques, des organismes statistiques nationaux envisagent d'utiliser des données synthétiques pour partager leurs produits de données (Commission économique des Nations Unies pour l'Europe, 2022). Par ailleurs, certains grands administrateurs de données ont partagé publiquement des données synthétiques. Notons, par exemple, les fichiers Data Entrepreneur's Synthetic Public Use des Centers for Medicare & Medicaid Services (CMS) (U.S. Centers for Medicare & Medicaid Services, 2022), les données sur le cancer de Public Health England (National Disease Registration Service, 2018), les variantes synthétiques de l'ensemble de données sur les hôpitaux et les réclamations du système de santé publique français (Health Data Hub France, 2021), et les microdonnées synthétiques du registre national des naissances vivantes d'Israël (Hod et Canetti, 2024).

Les données synthétiques peuvent être utiles dans trois principaux cas d'utilisation : 1) le partage de données dans le respect de la vie privée, 2) l'élimination des biais et 3) l'augmentation des données (Jordon et coll., 2022). L'élimination des biais désigne un processus de génération de données synthétiques visant à contrebalancer les biais de représentation (tels que les biais raciaux ou de genre dans les données). L'augmentation des données est un outil utile pour s'acquitter de tâches d'analyse lorsque la taille d'échantillon des ensembles de données existants est limitée, ce qui est souvent le cas dans le secteur des soins de santé et le secteur humanitaire. L'utilisation de données synthétiques pour protéger la vie privée est le cas le plus courant, tel que le démontre la littérature portant sur les soins de santé (Kaabachi et coll., 2023).

[1]Faculté de médecine, Université d'Ottawa et Institut de recherche du CHEO, Ottawa (Ontario); lpilgram@ehealthinformation.ca et kelemam@ehealthinformatio.ca

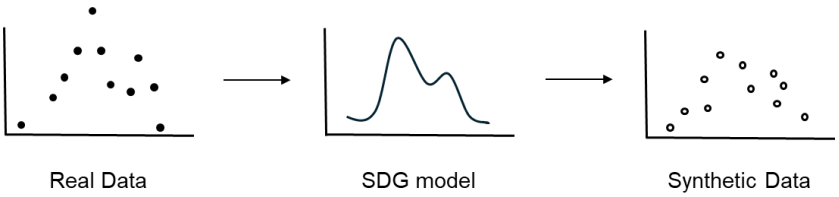
Dans le cadre du présent article, nous donnerons un aperçu des pratiques actuelles en matière de GDS et présenterons les enseignements qui ont été tirés de l'élaboration et de l'utilisation de données synthétiques pour protéger la vie privée et éliminer des biais.

1.1 Générateurs de données synthétiques

On désigne par le terme général « données synthétiques » un large éventail de méthodes, selon la façon dont les données sont générées et de ce qu'elles cherchent à imiter. De façon générale, le processus de GDS ne nécessite pas nécessairement de données réelles. Il peut s'appuyer sur les distributions connues a priori et être éclairé par des connaissances préalables, des statistiques sommaires publiées et des calculateurs de risques publiés (Templ et coll., 2017; Walonoski et coll., 2018; Jeanson et coll., 2024; Al-Dhamari et coll., 2024). Nous nous intéressons cependant dans le présent article au processus de GDS qui comprend l'entraînement à l'aide de données individuelles réelles pour en tirer la distribution conjointe qui est ensuite utilisée pour la génération.

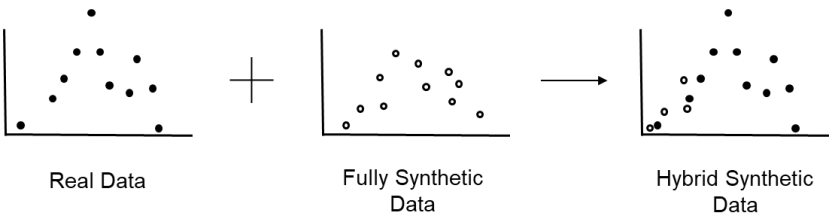
Les méthodes fondées sur les données individuelles utilisent des méthodes d'apprentissage automatique ou d'apprentissage profond, notamment des arbres de décision séquentiels (Drechsler, 2011; Nowok, 2015), ainsi que des réseaux antagonistes génératifs, qui sont l'un des types de modèles génératifs les plus utilisés dans la recherche et la pratique (Hernandez et coll., 2022; Ghosheh et coll., 2024). Elles ont souvent été appliquées pour la synthèse des données sur la santé. La figure 1.1-1 illustre de manière simplifiée le processus de la GDS.

Figure 1.1-1
Conceptualisation de la génération de données synthétiques



Les données générées peuvent être entièrement synthétiques, partiellement synthétiques et synthétiques hybrides (Little, 1993; Rubin, 1993; Surendra et MohanH, 2017). Les données entièrement synthétiques désignent un ensemble de données entièrement synthétiques, les données partiellement synthétiques un ensemble de données dont certaines colonnes sont résumées, et les données synthétiques hybrides sont des données qui contiennent à la fois des enregistrements réels et synthétiques. Les idées et données probantes que nous présentons ici portent sur les données entièrement synthétiques (c'est-à-dire le cas d'utilisation pour la protection de la vie privée) et les données hybrides (c'est-à-dire le cas d'utilisation de l'élimination des biais). La figure 1.1-2 illustre le concept de données hybrides. Sauf indication contraire, nous utiliserons le terme « données synthétiques » pour désigner uniquement des données tabulaires structurées entièrement synthétiques.

Figure 1.1-2
Conceptualisation des données synthétiques hybrides pour l'élimination des biais



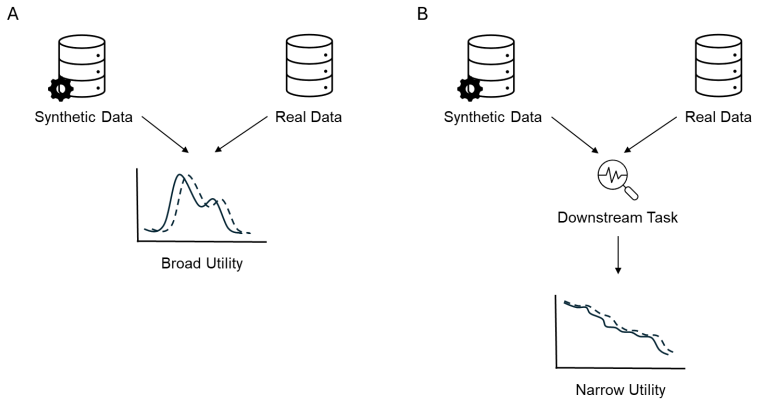
Les données synthétiques sont évaluées à l'aide de mesures d'utilité et de protection de la vie privée. Le cas d'utilisation des données synthétiques est ce qui détermine en fin de compte la manière dont elles sont évaluées. Alors que l'utilité est pertinente pour tous les cas d'utilisation, la protection de la vie privée n'est pas pertinente pour l'élimination des biais ou l'augmentation des données.

1.2 Évaluer l'utilité des données synthétiques

L'utilité peut être définie comme la mesure dans laquelle les données synthétiques conviennent à un cas d'utilisation spécifique. Par conséquent, l'évaluation de l'utilité ne comprend pas nécessairement les mêmes mesures dans tous les cas d'utilisation; il existe plusieurs façons de mesurer l'utilité (El Emam, 2020; El Emam et coll., 2022b; Kaabachi et coll., 2023; Vallevik et coll., 2024; Budu et coll., 2024).

Dans le cas de la protection de la vie privée, les mesures peuvent être réparties en deux catégories d'utilité, soit générale et restreinte. Les mesures d'utilité générale évaluent la ressemblance ou la similitude des données synthétiques avec les données réelles (c'est-à-dire la fidélité) et les mesures d'utilité restreinte en évaluent la fonctionnalité (c'est-à-dire l'utilité en aval). Il n'est pas nécessaire de savoir comment les données synthétiques seront utilisées pour en évaluer la fidélité. L'évaluation de l'utilité restreinte est toutefois liée à une tâche spécifique en aval, par exemple la réplique d'une analyse des données de survie pour les patients atteints de cancer. C'est ce qu'illustre la figure 1.2-1.

Figure 1.2-1
Utilité générale (A) et restreinte (B) des données synthétiques



L'utilité générale peut être mesurée à un niveau univarié ou multivarié. Les mesures de fidélité univariées ne fournissent toutefois pas suffisamment de renseignements, car elles ne tiennent pas compte des relations multivariées. Les mesures multivariées, en revanche, permettent d'évaluer simultanément les distributions conjointes et les liens de dépendance, ce qui est nécessaire pour tirer des conclusions sur tout l'ensemble de données. Il est souvent souhaitable que les mesures d'utilité générale reflètent ou prédisent l'utilité restreinte, ce qui signifie qu'elles peuvent servir d'indicateur pour la tâche en aval. Il semble que l'appariement des scores de propension, l'analyse par grappes et la distance multivariée de Hellinger soient des mesures particulièrement efficaces pour saisir les propriétés des données qui sont pertinentes, par exemple, dans les tâches de classification en aval (El Emam et coll., 2022b). La mesure directe de l'utilité restreinte ne constitue souvent pas une approche pratique, car les tâches en aval réelles ou représentatives peuvent ne pas être connues *a priori* ou peuvent être trop exigeantes en termes de calcul dans les processus d'optimisation des modèles génératifs.

L'utilité restreinte demeure pourtant une mesure type dans la recherche méthodologique. Elle fournit de précieuses indications sur la fonctionnalité et l'efficacité des données synthétiques dans la pratique. En fin de compte, c'est l'utilité restreinte plutôt que l'utilité générale qui importe le plus. Lorsque la tâche en aval comprend l'inférence, la question est de savoir s'il est possible de tirer les mêmes idées et conclusions des données synthétiques que des données réelles. Les mesures de reproductibilité sont pertinentes pour évaluer cela. Le tableau 1.2-1 présente les quatre mesures qui ont été utilisées pour brosser un tableau complet de la reproductibilité (El Emam et coll., 2024).

Tableau 1.2-1

Définitions de la reproductibilité des tâches inférentiellesMesure	**Explication**
Accord de décision	On examine si les estimations ont la même orientation et la même signification statistique que les données réelles.*
Accord d'estimation	On examine si les estimations synthétiques se situent dans l'intervalle de confiance (IC) à 95 % des données réelles.
Écart normalisé	On examine si l'écart entre les estimations corrobore l'hypothèse nulle d'absence d'écart.
Chevauchement des IC	La proportion du chevauchement entre l'IC à 95 % des données synthétiques et des données réelles.

*non applicable aux tâches descriptives en aval

Lorsque la tâche en aval en est une tâche de prédiction plutôt que d'inférence, l'approche « Train on synthetic, test on real » (TSTR) qui consiste à utiliser des données synthétiques pour l'entraînement du modèle et des données réelles pour le mettre à l'essai, fournit une mesure de l'utilité restreinte (Hyland et coll., 2017). L'approche TSTR emploie un ensemble de données test réelles pour évaluer, par exemple, la performance de prévision d'un modèle d'apprentissage automatique (AA) entraîné à l'aide de données synthétiques. Une faible performance indique alors que le modèle ne se généralise pas bien aux données réelles non vues. Par conséquent, l'interprétation de l'approche TSTR ne nécessite pas nécessairement une comparaison avec la performance lorsque le modèle est entraîné à l'aide de données réelles (Train on real, test on real, ou TRTR). Une telle comparaison permet toutefois de comprendre les raisons sous-jacentes d'une faible performance de l'approche TSTR. Si l'approche TRTR offre une bonne performance, cela signifie que la GDS n'a pas permis de générer des données synthétiques de grande utilité. Si la performance est mauvaise, le problème ne se situe pas au niveau de la GDS, mais des données réelles d'entraînement qui ne conviennent pas à la tâche en aval. Et si la performance de l'approche TSTR est supérieure à celle de l'approche TRTR, cela peut indiquer que le processus de GDS améliore la diversité de l'ensemble de données et sa capacité à généraliser des données non vues.

Les évaluations de l'utilité mentionnées ci-dessus concernent principalement les données synthétiques utilisées pour la protection de la vie privée, l'objectif étant de disposer de données de substitution à des données réelles. L'augmentation des données et l'élimination des biais nécessitent d'effectuer d'autres évaluations.

En ce qui concerne l'augmentation des données, l'objectif consiste à générer des enregistrements plus diversifiés qui peuvent ensuite être ajoutés aux données réelles. Des indicateurs de diversité peuvent être appliqués pour évaluer la mesure dans laquelle la GDS renforce la diversité (Sajjadi et coll., 2018; Alaa et coll., 2022). On s'attend à ce que l'ensemble de données plus diversifié soit utilisé pour entraîner des modèles offrant une meilleure performance en matière de pronostic pour des données non vues (c'est-à-dire qu'ils présentent des erreurs de généralisation plus faibles). Ces mesures peuvent être considérées comme des mesures générales pour le cas d'utilisation de l'augmentation des données. Des mesures d'utilité restreinte aux fins d'augmentation peuvent être définies semblablement à l'approche TSTR en ce qui concerne les tâches de pronostic.

Dans le cas de l'élimination des biais, l'équité doit être évaluée (Dwork et coll., 2012; Hardt et coll., 2016; Yan et coll., 2020; Juwara et coll., 2024; Vallevik et coll., 2024). Le biais peut être mesuré, par exemple, en utilisant la différence de parité statistique (DPS). La DPS compare la probabilité du résultat pour deux groupes susceptibles d'être touchés par un biais. La différence d'égalité des chances (DEC) compare leur taux de vrais positifs, et la différence moyenne des cotes (DMC) augmente la DEC avec le taux de faux positifs. (Dwork et coll., 2012; Hardt et coll. 2016; Yan et coll. 2020). L'utilisation de ces mesures en tant que mesures autonomes peut toutefois induire des erreurs, car on suppose que la disparité de la réalité de terrain est nulle. Dans la recherche sur les soins de santé, par exemple, si le résultat est un diagnostic, on s'attend à une certaine disparité entre les sexes pour un grand nombre de maladies. Cela signifie que l'analyste de données doit avoir une certaine connaissance de la disparité de la réalité de terrain pour pouvoir interpréter les biais présents dans ses données et, par conséquent, les biais présents dans les données dont on a éliminé les biais après la GDS.

1.3 Confidentialité des données synthétiques

Si, pendant la GDS, le modèle génératif est bien entraîné, c'est-à-dire qu'il ne fait pas de surajustement ou de mémorisation, il n'y aura pas de mise en correspondance biunivoque entre les données synthétiques et les données réelles. Cependant, si le modèle génératif surajuste ou mémorise (une partie) des données réelles, il peut y avoir divulgation de renseignements personnels dans les données synthétiques générées. Cela rend l'évaluation de la protection de la vie privée obligatoire dans tous les cas d'utilisation de protection de la vie privée, mais pas dans les cas d'utilisation d'augmentation ou d'élimination des biais.

La plupart des publications sur la protection de la vie privée portent sur trois concepts de divulgation de renseignements personnels : la divulgation de l'identité, la divulgation de l'appartenance et la divulgation d'attributs (Kaabachi et coll., 2023; Boudewijn et coll., 2023; Vallevik et coll., 2024; Budu et coll., 2024). On parle de divulgation d'identité lorsque l'identité d'une personne peut être attribuée à un enregistrement; de divulgation d'appartenance lorsque l'appartenance d'une personne à l'ensemble de données peut être déduite; et de divulgation d'attributs lorsque les renseignements sensibles d'une personne peuvent être déduits des attributs d'un ensemble de données.

Étant donné que le couplage avec un enregistrement réel n'est pas préservé dans le cadre de la GDS, la divulgation de l'identité devrait par définition être protégée. Il est de plus difficile d'adapter les mesures de la divulgation de l'identité aux données synthétiques pour tenir compte de ce processus inhérent. Par conséquent, la divulgation de l'appartenance et la divulgation d'attributs sont considérées comme les principales vulnérabilités à évaluer dans les données synthétiques. Il s'agit dans les deux cas d'inférences. Les inférences correctes ne portent toutefois pas atteinte à la vie privée en soi, mais peuvent se produire indépendamment de la divulgation des données d'un individu. Par exemple, un modèle peut être entraîné à l'aide de données synthétiques qui prédisent la survie des patients atteints de cancer. Si l'enregistrement d'un individu ne figure pas dans les données d'entraînement, la prédiction de survie est indépendante de cet individu et peut être considérée comme une connaissance générale plutôt que comme une atteinte à la vie privée de cet individu.

L'évaluation de la vulnérabilité de la divulgation de l'appartenance dans les données synthétiques consiste généralement à imiter une attaque adverse et nécessite un examen approfondi des hypothèses adverses. Par exemple, si l'on suppose que l'adversaire vise des cibles au hasard dans la même population que celle dans laquelle les données d'entraînement ont été prélevées, l'échantillon de cibles imitées utilisé dans l'évaluation doit correspondre à la proportion de membres et de non-membres dans la population (El Emam et coll., 2022a).
La divulgation d'attributs est une tâche de prédiction dans laquelle un modèle est entraîné pour prédire les renseignements sensibles d'une cible en fonction des données synthétiques. Par exemple, si les renseignements sensibles sont catégoriques, le modèle de classification est entraîné à l'aide de données synthétiques. Sa performance de prévision est évaluée pour les données d'entraînement de la GDS afin de déterminer si la prédiction est significative pour les personnes susceptibles de faire l'objet d'une atteinte à leur vie privée du fait de leur appartenance aux données d'entraînement de la GDS. Une aire sous la courbe ROC (de l'anglais « receiver operating characteristic », caractéristique de fonctionnement du récepteur) de 0,2, par exemple, ne permettrait pas à un adversaire de déduire des renseignements corrects par inférence. De façon générale, on considère qu'une aire sous la courbe ROC inférieure ou égale à 0,7 ou à 0,6 indique un faible degré de précision (Hond et coll., 2022). Une fois qu'on a confirmé la bonne performance de prévision, il faut la comparer à l'inférence indépendante de la divulgation (c'est-à-dire la génération de connaissances) afin d'évaluer la vulnérabilité de l'ensemble de données synthétiques. Une telle base de référence peut être établie à l'aide d'un ensemble de données test (Giomi et coll., 2023; Francis et Wagner, 2024).

Cette évaluation de la protection de la vie privée reste nécessaire même lorsque la GDS différentiellement confidentielle est appliquée. Cela s'explique par le fait que la vulnérabilité est fonction du budget de protection de la vie privée dans le cadre de la confidentialité différentielle, mais l'ampleur demeure incertaine, sauf pour les budgets qui s'approchent de 0 (Dwork et coll., 2019). Dans la pratique, les budgets consacrés à la protection de la vie privée sont généralement plus importants. Par exemple, le Bureau du recensement des États-Unis a mis en œuvre la confidentialité différentielle avec un budget relativement élevé de 18,19 (Abowd et coll., 2022).

1.4 La stochasticité de la GDS

Si l'on considère la GDS comme une forme d'imputation multiple, les mêmes considérations en termes de caractère aléatoire s'appliquent. Par conséquent, l'utilisation de plusieurs ensembles de données synthétiques et des règles de combinaison correspondantes peut accroître la robustesse des résultats. Ceci est reconnu par diverses mesures de confidentialité pour les données synthétiques (Elliot, 2015; Taub et coll., 2018; Stadler et coll., 2020), et a fait l'objet d'une évaluation complète de la reproductibilité inférentielle, un plateau ayant été atteint après 10 itérations (El Emam et coll., 2024). Ceci est particulièrement important dans la cadre de la recherche méthodologique où il est impératif de mener une estimation solide de la confidentialité et de l'utilité du modèle de GDS, par exemple, afin d'établir des données repères pour plusieurs modèles. Dans le cas d'un scénario de diffusion de données, il est plus pertinent de mesurer le respect de la vie privée pour un ou des ensembles de données synthétiques spécifiques.

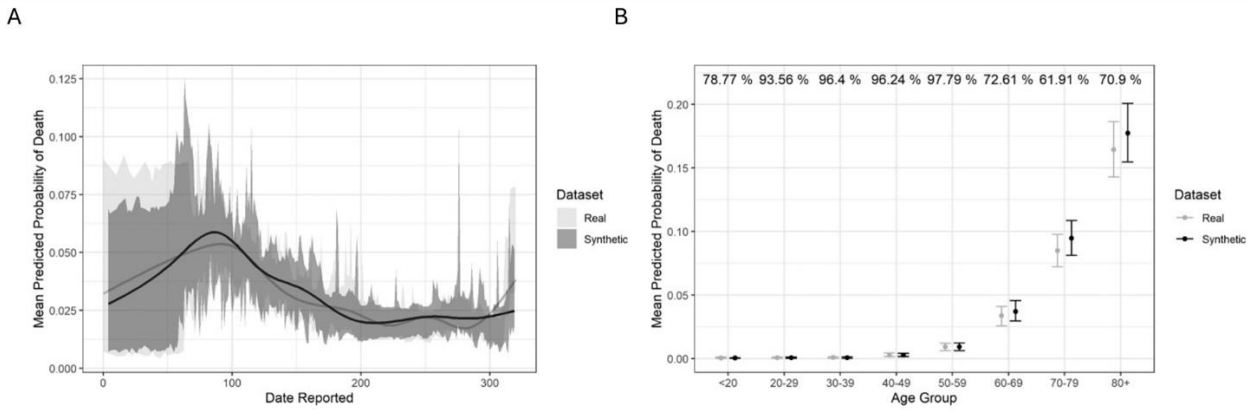
2. Génération de données synthétiques pour protéger la vie privée

Dans cette section, nous présentons deux exemples représentatifs où des données synthétiques ont été générées dans des cas d'utilisation de la GDS pour protéger la vie privée.

2.1 Données synthétiques sur les cas de COVID-19

La pandémie de COVID-19 a mis en évidence l'importance pour les chercheurs de partager des données, et a mis en lumière des obstacles importants à l'accès aux données. Ceux-ci sont principalement liés à des préoccupations en matière de protection de la vie privée. La GDS peut permettre de partager des données dans une situation telle qu'une urgence sanitaire mondiale. Dans El Emam et coll. (2021), on a évalué le potentiel qu'offrait la GDS pour partager des données sur la COVID-19 sans porter atteinte à la vie privée et en maintenant l'utilité des données. L'étude portait sur 90 514 cas de COVID-19 en Ontario. Des données synthétiques ont été générées à l'aide d'arbres de classification et de régression séquentiels. On a évalué la vulnérabilité liée à la divulgation des membres et des attributs, et l'utilité en aval a été définie comme la prédiction de la mortalité. Les résultats d'un modèle de classification binaire utilisant des arbres de décision optimisés par le gradient ont démontré que la performance de prévision ne différait que légèrement de l'ensemble de données réelles (aire sous la courbe ROC de 0,945 en utilisant des données réelles comparativement à une aire sous la courbe ROC de 0,940 en utilisant des données synthétiques). Le modèle reposant sur des données synthétiques comportait les mêmes variables explicatives de grande importance (c'est-à-dire l'âge et la date) que le modèle reposant sur des données réelles (voir la figure 2.1-1). Parallèlement, la vulnérabilité en matière de protection de la vie privée a été jugée faible : les données synthétiques constituaient de bonnes données de substitution dans le respect de la vie privée aux données réelles.

Figure 2.1-1
Probabilité de décès selon la date (A) et l'âge (B) tirée des données synthétiques (El Emam et coll., 2021).



2.2 Données synthétiques d'essais cliniques sur le cancer

Dans le secteur biopharmaceutique, le partage des données issues des essais cliniques vise à garantir la transparence et la reproductibilité. Les politiques de transparence de Santé Canada exigent par exemple une telle pratique dans le cadre de l'autorisation de mise sur le marché. Le partage des données permet également de mener des recherches à des fins secondaires et d'obtenir, par exemple, des renseignements précieux sur l'innocuité des médicaments, tout en allégeant le fardeau des patients. La GDS est donc un outil qui offre beaucoup de potentiel pour cette industrie. Dans El Kababji et coll. (2023) on a évalué des ensembles de données synthétiques se basant sur huit essais cliniques réalisés chez des patients atteints du cancer du sein afin de déterminer leur capacité à reproduire l'analyse de données réelles. On a comparé trois modèles différents de GDS : la synthèse séquentielle en arborescence, les réseaux antagonistes génératifs (RAG) et les autoencodeurs variationnels. Même si l'utilité des modèles de GDS variait en fonction de l'essai clinique, la synthèse séquentielle en arborescence a permis d'obtenir, par exemple, des estimations et des décisions très concordantes. La vulnérabilité en matière de protection de la vie privée était faible pour tous les ensembles de données synthétiques. Cette étude a révélé que le choix du modèle de GDS est important et qu'il ne sera pas cohérent pour ce qui est de différents ensembles de données. Elle a également contribué à l'accumulation des données démontrant que les données synthétiques peuvent répondre tant aux considérations de protection de la vie privée que d'utilité.

3. Génération de données synthétiques afin d'atténuer les biais

Le concept d'atténuation des biais grâce à la GDS consiste à générer des enregistrements supplémentaires correspondant au groupe sous-représenté afin d'en augmenter la représentation. Cet objectif peut être atteint grâce à la génération conditionnelle ou à l'échantillonnage de rejet après coup. Dans le domaine de la recherche biomédicale, les biais de représentation constituent un problème courant et grave. Ils se traduisent par un ensemble de données qui ne reflète pas bien la répartition de la population de patients et, au bout du compte, par des analyses qui manquent de validité externe. Avec Juwara et coll. (2024) , on a utilisé la GDS comme méthode d'élimination des biais et l'ont testé sur quatre ensembles de données relatives aux soins de santé. Dans un premier temps, les différents types de biais ont été introduits, puis l'ensemble de données dont les biais avaient été éliminés a été évalué par rapport à l'ensemble de données initial (c'est-à-dire la réalité de terrain) tel qu'il est décrit ci-dessus. L'analyse a permis de conclure que l'augmentation des données par GDS permettait de maintenir la performance de prévision pour les groupes minoritaires dans les situations où les niveaux de biais sont faibles à moyens, mais pas de manière cohérente dans les cas où le niveau de biais est élevé.

4. Observations et limites principales

De plus en plus de données probantes révèlent que la GDS permet de saisir des tendances pertinentes et assez complexes dans les données réelles. Cet outil peut donc fournir de bonnes données de substitution aux données réelles dans un cas d'utilisation pour protéger la vie privée, ou servir de complément dans un cas d'utilisation pour éliminer des biais. Nous pouvons résumer comme suit ce que nous avons appris, principalement avec des données tabulaires sur les soins de santé :

(1) L'entraînement d'un modèle génératif dépend de la taille (nombre d'observations et de variables) de l'ensemble de données source. Un petit nombre d'observations présente un risque de surajustement, mais ce risque est atténué s'il y a peu de variables. Le surajustement peut porter atteinte à la vie privée, mais peut être bénéfique sur le plan de l'utilité (par exemple, au niveau de la fidélité et de la reproductibilité). Il n'existe pas de calculateur de taille d'échantillon pour les modèles génératifs, mais de bons résultats en matière de protection de la vie privée et d'utilité ont été obtenus avec de petits ensembles de données (par exemple, dans le cas de petits essais cliniques).
(2) Il n'est pas nécessaire de savoir comment les données synthétiques seront analysées pour entraîner les modèles génératifs. Selon certaines données probantes, les modèles génératifs saisissent plusieurs tendances pertinentes dans les données sources pour des tâches d'analyse typiques.

(3) Le traitement au préalable des données réelles est une étape essentielle et peut avoir un impact non négligeable sur le rendement des modèles génératifs. Cela signifie que la mise en œuvre d'un modèle génératif a de l'importance et que les mises en œuvre ne sont pas toutes aussi efficaces, même pour un même type de modèle génératif.
(4) Il n'existe pas de type de modèle génératif supérieur aux autres pour tous les ensembles de données, mais plutôt une supériorité spécifique à chaque ensemble de données.
(5) Lorsque la charge de travail analytique consiste à estimer un paramètre et son intervalle de confiance, il faut utiliser les règles de combinaison pour calculer la valeur à partir de 10 données synthétiques afin d'éviter des estimations et des inférences non valides.
(6) La communauté a mis au point des mesures pour évaluer les différents types d'utilité (fidélité et reproductibilité) et de protection de la vie privée (divulgation de l'appartenance et divulgation d'attributs). L'élaboration de points de référence faisant consensus quant à ce qui est acceptable demeure en cours.
(7) Les données synthétiques peuvent contribuer à améliorer l'équité et la précision des modèles de pronostic et d'inférence lorsque le niveau de biais de représentation dans les données d'observation est faible ou moyen. Il est plus difficile d'atténuer de manière cohérente les effets des biais de représentation dans les ensembles de données réelles lorsque le biais est élevé.

Bibliographie

Abowd, J.M., Ashmead, R., Cumings-Menon, R., Garfinkel, S., Heineck, M., Heiss, C., Johns, R., Kifer, D., Leclerc, P., Machanavajjhala, A., Moran, B., Sexton, W., Spence, M., et Zhuravlev, P., (2022), "The 2020 Census Disclosure Avoidance System TopDown Algorithm", *Harvard Data Science Review*. Article accessible à l'adresse https://doi.org/10.1162/99608f92.529e3cb9.

Alaa, A., Breugel, B.V., Saveliev, E.S., Schaar, et M. van der (2022), "How Faithful is your Synthetic Data? Sample-level Metrics for Evaluating and Auditing Generative Models", *Proceedings of the 39th International Conference on Machine Learning*, PMLR, p 290–306.

Al-Dhamari, I., Abu Attieh, H., et Prasser, F. (2024), "Synthetic datasets for open software development in rare disease research", *Orphanet J Rare Dis* 19:265. Article accessible à l'adresse https://doi.org/10.1186/s13023-024-03254-2.

Boudewijn, A., Ferraris, A.F., Panfilo, D., Cocca, V., Zinutti, S., De Schepper, K., et Chauvenet, C.R. (2023), "Privacy Measurement in Tabular Synthetic Data: State of the Art and Future Research Directions", arXiv:2311.17453, https://doi.org/10.48550/arXiv.2311.17453.

Budu, E., Etminani, K., Soliman, A., et Rögnvaldsson, T. (2024), "Evaluation of synthetic electronic health records: A systematic review and experimental assessment", *Neurocomputing* 128253. Article accessible à l'adresse https://doi.org/10.1016/j.neucom.2024.128253.

Commission économique pour l'Europe des Nations unies (2022), "Données synthétiques pour les statistiques officielles - Guide de démarrage", https://unece.org/statistics/publications/synthetic-data-official-statistics-starter-guide. Consulté : 2023-01-24.

Drechsler, J. (2011), *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation*, Springer-Verlag, New York.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., et Zemel, R. (2012), "Fairness through awareness", Dans : *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. Association for Computing Machinery, New York, NY, USA, p. 214–226.

Dwork, C., Kohli, N., et Mulligan, D. (2019), "Differential Privacy in Practice: Expose your Epsilons!", *Journal of Privacy and Confidentiality* 9:. Article accessible à l'adresse https://doi.org/10.29012/jpc.689.

El Emam, K. (2020), "Seven Ways to Evaluate the Utility of Synthetic Data", *IEEE Security Privacy* 18:56–59. Article disponible à l'adresse https://doi.org/10.1109/MSEC.2020.2992821.

El Emam, K., et Hoptroff, R. (2020), *Practical Synthetic Data Generation*, O'Reilly, Sebastopol, CA.

El Emam, K., Mosquera, L., et Fang, X. (2022a), "Validating A Membership Disclosure Metric For Synthetic Health Data", *JAMIA* Open 5:ooac083.

El Emam, K., Mosquera, L., Fang, X., et El-Hussuna, A. (2022b), "Utility Metrics for Evaluating Synthetic Health Data Generation Methods: Validation Study", *JMIR Med Inform* 10:e35734. Article accessible à l'adresse https://doi.org/10.2196/35734.

El Emam, K., Mosquera, L., Fang, X., et El-Hussuna, A. (2024), "An evaluation of the replicability of analyses using synthetic health data", *Sci Rep* 14:6978. Article accessible à l'adresse https://doi.org/10.1038/s41598-024-57207-7.

El Emam, K, Mosquera, L., Jonker, E., et Sood, H. (2021). "Evaluating the utility of synthetic COVID-19 case data", *JAMIA* Open 4:ooab012. Article accessible à l'adresse https://doi.org/10.1093/jamiaopen/ooab012.

El Kababji, S., Mitsakakis, N., Fang, X., Beltran-Bless, A., Pond, G., Vandermeer, L., Radhakrishnan, D., Mosquera, L. Paterson, A., Shepherd, L,. Chen, B., Barlow, W.E., Gralow, J., Savard, M.-F., Clemons, M. et El Emam, K. et al (2023), "Evaluating the Utility and Privacy of Synthetic Breast Cancer Clinical Trial Data Sets", *JCO Clin Cancer Inform* e2300116. Article accessible à l'adresse https://doi.org/10.1200/CCI.23.00116.

Elliot, M. (2015), Final Report on the Disclosure Risk Associated with the Synthetic Data Produced by the SYLLS Team, Manchester University.

Francis, P., et Wagner, D. (2024), "Towards more accurate and useful data anonymity vulnerability measures", arXiv:2403.06595 [cs.CR].

Ghosheh, G.O., Li, J., et Zhu, T. (2024), "A Survey of Generative Adversarial Networks for Synthesizing Structured Electronic Health Records", *ACM Comput Surv* 56:147:1-147:34. Article accessible à l'adresse https://doi.org/10.1145/3636424.

Giomi, M., Boenisch, F., Wehmeyer, C., et Tasnádi, B. (2023), "A Unified Framework for Quantifying Privacy Risk in Synthetic Data", *Proceedings on Privacy Enhancing Technologies*.

Hardt, M., Price, E., et Srebro, N. (2016), "Equality of Opportunity in Supervised Learning", arXiv:1610.02413 [cs.LG].

Health Data Hub France (2021), "SNDS synthétiques", Dans : *Système national des données de santé*. https://documentation-snds.health-data-hub.fr/formation_snds/donnees_synthetiques/. Consulté : 2022-01-22.

Hernandez, M., Epelde, G., Alberdi, A., Cilla, R., et Rankin, D. (2022), "Synthetic data generation for tabular health records: A systematic review", *Neurocomputing*, 493:28–45. DOI:10.1016/j.neucom.2022.04.053.

Hod, S., et Canetti, R. (2024), "Differentially Private Release of Israel's National Registry of Live Births",Dans : arXiv.org. Article accessible à l'adresse https://arxiv.org/abs/2405.00267v1. Consulté : 2024-09-28.

Hond, A.A.H. de, Steyerberg, E.W., et Calster, B. van (2022), "Interpreting area under the receiver operating characteristic curve", *The Lancet Digital Health* 4:e853–e855. Article accessible à l'adresse https://doi.org/10.1016/S2589-7500(22)00188-1.

Hyland, S.L., Esteban, C., et Rätsch, G. (2017), "Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs", arXiv:170602633 [cs, stat]

Jeanson, F., Farkouh, M.E., Godoy, L.C., Minha, S., Tzuman, O., et Marcus, G. (2024), "Medical calculators derived synthetic cohorts: a novel method for generating synthetic patient data", *Sci Rep* 14:11437. Article accessible à l'adresse https://doi.org/10.1038/s41598-024-61721-z.

Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S.N., et Weller, A. (2022), "Synthetic Data -- what, why and how?", arXiv:2205.03257 [cs.LG].

Juwara, L., El-Hussuna, A., et El Emam, K. (2024), "An evaluation of synthetic data augmentation for mitigating covariate bias in health data", *Patterns*, Article accessible à l'adresse https://doi.org/10.1016/j.patter.2024.100946.

Kaabachi, B., Despraz, J., Meurers, T., Otte, K., Halilovic, M., Prasser, F., et Raisaro, J.-L. (2023), "Can We Trust Synthetic Data in Medicine? A Scoping Review of Privacy and Utility", *Metrics*, 2023.11.28.23299124.

Little, R. (1993), "Statistical Analysis of Masked Data", *Journal of Official Statistics*, 9:407–426.

National Disease Registration Service (2018), The Simulacrum. Dans : The Simulacrum. https://simulacrum.healthdatainsight.org.uk/. Consulté : 2021-11-27

Nowok, B. (2015), *Utility of synthetic microdata generated using tree-based methods*, Helsinki.

Rubin, D. (1993), "Discussion: Statistical Disclosure Limitation", *Journal of Official Statistics*, 9:462–468.

Sajjadi, M.S.M., Bachem, O., Lucic, M., Bousquet, O., et Gelly, S. (2018), "Assessing Generative Models via Precision and Recall", arXiv:180600035 [cs, stat].

Stadler, T., Oprisanu, B., et Troncoso, C. (2020), "Synthetic Data -- A Privacy Mirage", arXiv:201107018 [cs].

Surendra, H., et Mohan, H. S., (2017), "A Review Of Synthetic Data Generation Methods For Privacy Preserving Data Publishing", *International Journal of Scientific & Technology Research*

Taub, J., Elliot, M., Pampaka, M., et Smith, D. (2018), "Differential Correct Attribution Probability for Synthetic Data: An Exploration", In: Domingo-Ferrer J, Montes F (eds) *Privacy in Statistical Databases*, Springer International Publishing, Cham, p 122–137.

Templ, M., Meindl, B., Kowarik, A., et Dupriez, O. (2017), "Simulation of Synthetic Complex Data: The R Package simPop", *Journal of Statistical Software*, 79:1–38. Article accessible à l'adresse https://doi.org/10.18637/jss.v079.i10.

U.S. Centers for Medicare & Medicaid Services (2022), CMS 2008-2010 Data Entrepreneurs' Synthetic Public Use File (DE-SynPUF), https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs/DE_Syn_PUF. Consulté 2022-07-17.

Vallevik, V.B., Babic, A., Marshall, S.E., Elvatun, S., Brøgger, H.M.B., Alagaratnam, S., Edwin, B., Veeraragavan, N.R., Befring, A.K., et Nygård, J.F. (2024), "Can I trust my fake data – A comprehensive quality assessment framework for synthetic tabular data in healthcare", I*nternational Journal of Medical Informatics*, 185:105413. Article accessible à l'adresse https://doi.org/10.1016/j.ijmedinf.2024.105413.

van Breugel, B., Liu, T., Oglic, D., et van der Schaar, M. (2024), "Synthetic data in biomedicine via generative artificial intelligence", *Nature Reviews Bioengineering*, 1–14.

Walonoski, J., Kramer, M., Nichols, J., Quina, A., Moesel, C., Hall, D., Duffett, C., Dube, K., Gallagher, T., et McLachlan, S. (2018), "Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record", *J Am Med Inform Assoc*, 25:230–238. Article accessible à l'adresse https://doi.org/10.1093/jamia/ocx079.

Yan, S., Kao, H., et Ferrara, E. (2020), "Fair Class Balancing: Enhancing Model Fairness without Observing Sensitive Attributes", *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, Association for Computing Machinery, New York, NY, USA, p. 1715–1724.