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

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

# Évaluation du risque de divulgation de données synthétiques : une revue de littérature

par Zhe Si Yu

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

# Évaluation du risque de divulgation de données synthétiques : une revue de littérature

Zhe Si Yu[1]

## Résumé

L'adoption de la production de données synthétiques comme mesure de confidentialité augmente dans les organismes statistiques du monde entier, y compris à Statistique Canada. Cette approche offre une solution de rechange à la diffusion classique de fichiers publics anonymisés de microdonnées, répondant ainsi aux préoccupations en matière de protection de la vie privée et d'utilité des données. La création de données synthétiques présente toutefois des défis au niveau de l'évaluation et de l'atténuation du risque de divulgation. Le présent article passe en revue les différents types de risques de divulgation, soit la divulgation d'attribut, la divulgation de données d'adhésion et la divulgation de l'identité, et présente certaines des méthodes connexes pour mesurer le risque. L'article présente les grandes mesures d'évaluation des risques et examine les méthodes pratiques de contrôle de la divulgation dans le cadre de la synthèse des données. Les méthodes d'évaluation des risques de divulgation produisent habituellement une mesure qui peut servir à jauger le risque, mais il y a peu de consensus sur les valeurs limites de ces mesures. Il importe également de souligner l'importance de trouver le juste équilibre entre l'utilité et la confidentialité, ce qui nécessite une discussion plus approfondie dans le contexte de ces méthodes. L'article se termine en offrant des idées et des recommandations sur la gestion du risque de divulgation lors de la création de données synthétiques. Il offre également quelques perspectives sur les orientations futures de la recherche et les répercussions pratiques de la gestion des risques de divulgation par l'entremise des données synthétiques.

Mots-clés : Données synthétiques; évaluation du risque de divulgation; confidentialité.

## 1. Introduction

Les données synthétiques ont fait l'objet d'une attention particulière dans le domaine de la confidentialité des données, en raison de leur potentiel à offrir des données très utiles tout en préservant la confidentialité. Dans le contexte du présent article, les données synthétiques désignent tout ensemble de données généré par un modèle de synthèse entraîné à partir de données réelles qui répliquent statistiquement l'ensemble de données réelles. Un bon ensemble de données synthétiques doit générer des conclusions statistiques similaires à celles des données réelles. Les organismes nationaux de statistique (ONS), à l'instar de Statistique Canada, se tournent de plus en plus vers les données synthétiques comme solution de rechange à l'anonymisation classique, (Gauvin, 2021; Sallier, 2021) p. ex. la suppression. En ce moment, Statistique Canada utilise les données synthétiques principalement à des fins de formation, d'éducation et de simulation, mais jamais pour produire des statistiques officielles. Statistique Canada a déjà contribué à l'élaboration du guide officiel de la Commission économique des Nations Unies pour l'Europe (CEE-ONU) sur la génération de données synthétiques (CEE-ONU, 2022), lequel détaille l'évaluation des risques de divulgation des données synthétiques. Le présent article s'inscrit dans le prolongement des travaux que nous avons menés par le passé visant à étudier l'état actuel de la recherche en matière d'évaluation des risques de divulgation des données synthétiques.

Dans le contexte du présent article, les données entièrement et partiellement synthétiques sont définies : les ensembles de données entièrement synthétiques remplacent tous les enregistrements réels par des données générées de façon synthétique et les ensembles de données partiellement synthétiques conservent généralement les enregistrements réels, en remplaçant uniquement les attributs de nature délicate par des valeurs synthétisées. Le présent article se concentre toutefois sur les données entièrement synthétiques.

[1]Zhe Si Yu, Statistique Canada, 150, promenade Tunney's Pasture, Ottawa (Ontario), Canada, K1A 0T6, zhesi.yu@statcan.gc.ca
[2]Le contenu de ce document reflète les positions de l'auteur et pas nécessairement celles de Statistique Canada.

Malgré les avantages que procurent les données synthétiques, celles-ci ne sont pas exemptes de tout risque de divulgation. Il importe de disposer de paramètres clairs pour mesurer les risques liés aux données synthétiques, surtout si nous souhaitons que ces données soient davantage utilisées dans la pratique. La présente analyse documentaire explore les trois principaux risques de divulgation, soit la divulgation de l'identité, la divulgation d'attribut et la divulgation d'appartenance. Nous passons de plus en revue les techniques qui permettent d'évaluer ces risques, notamment en soupesant les compromis entre l'utilité des données et le risque de divulgation[2].

# 2. Types de risque de divulgation

## 2.1. Divulgation de l'identité

Il y a risque de divulgation de l'identité lorsqu'un intrus associe une personne connue à un enregistrement de données diffusé (Reiter, 2005). Il est largement admis dans la littérature que les données entièrement synthétiques présentent un faible risque de divulgation de l'identité, étant donné qu'il n'y a pas de correspondance univoque entre les enregistrements de données entièrement synthétiques et ceux de l'ensemble de données réelles (Hu, 2021; Reiter, 2005; Reiter et Mitra, 2009). Intuitivement, comme tous les enregistrements d'un ensemble de données entièrement synthétiques sont générés, il n'y a pas de couplage significatif avec l'ensemble de données réelles.

## 2.2. Divulgation d'attribut

Le risque de divulgation d'attribut dans le cadre de données synthétiques est mesuré en estimant les probabilités d'exposition de la valeur cible réelle des enregistrements dans les données d'origine (Hittmeir et coll., 2020; Hornby et Hu, 2021). Ce scénario suppose que l'intrus connaît déjà quelque peu l'ensemble de données originales (réelles) et/ou une connaissance du synthétiseur. La divulgation d'attribut couvre un éventail de préoccupations, car elle dépend du caractère sensible ou non de l'attribut lui-même (et de la nécessité de le protéger). Cette décision dépend souvent de l'ensemble de données.

De façon générale, la plupart des techniques d'évaluation du risque de divulgation d'attribut comportent un processus de calcul d'une cote basée sur la probabilité qu'un agresseur puisse deviner avec précision la valeur de l'attribut d'un enregistrement en se basant sur les connaissances qu'il possède déjà. Les modèles génératifs utilisés pour les données synthétiques reproduisent les corrélations entre les attributs en élaborant un processus génératif conditionnel. Par exemple, la méthode de spécification entièrement conditionnelle utilisée par Synthpop dans R (Nowok et coll., 2016) génère chaque attribut (sauf le premier) de manière séquentielle par conditionnement sur les attributs générés au préalable. En raison de la nature conditionnelle de ces processus, la plupart des chercheurs adoptent des méthodes d'estimation bayésiennes (Hornby et Hu, 2021; Hu et coll., 2014; Hu, 2021; Reiter et coll., 2014) pour quantifier le risque.

## 2.3 Divulgation d'appartenance

Le risque de divulgation d'appartenance est particulièrement important dans le cas des ensembles de données générés de façon synthétique. Il y a divulgation d'appartenance lorsqu'un enregistrement de la population est reconnu comme appartenant au réel ensemble de données lors de la diffusion des données synthétiques et/ou du modèle d'entraînement. Nous observons que, pour divulguer l'appartenance, l'approche adoptée est différente entre les données synthétiques et le sujet connexe bien étudié des modèles génératifs. Les modèles génératifs font partie de la catégorie plus générale dans laquelle s'inscrivent les modèles synthétiques, laquelle comprend tous les modèles statistiques qui prennent des données en entrée et génèrent des données souhaitées en sortie. La plupart des techniques explorées pour les modèles génératifs supposent un modèle de type « boîte noire » qui est néanmoins disponible pour les agresseurs, alors que le modèle n'est généralement pas disponible pour les données synthétiques (Zhao et coll., 2019). Étant donné que les hypothèses de connaissances préalables sont différentes entre le modèle génératif et la divulgation de données synthétiques, nous ne pouvons pas appliquer directement aux données synthétiques les recherches qui ont été menées

par le passé sur les modèles génératifs. Il convient d'utiliser des méthodes propres à l'évaluation de la divulgation des données synthétiques (Zhao et coll., 2019).

## 2.4 Risque perçu

L'utilisation de données synthétiques vise à représenter fidèlement les données originales d'un point de vue statistique. Cependant, l'authenticité d'une bonne réplique semblera inévitablement suspecte, ressemblant souvent à une divulgation de l'identité, créant ainsi un problème unique appelé « risque perçu ». Alors que tous les enregistrements sont fictifs dans un ensemble de données entièrement synthétiques, ils peuvent, par coïncidence, être générés de manière à être très similaires à un enregistrement réel, suffisamment pour donner l'impression que des renseignements réels ont été divulgués. Même si le risque de divulgation n'est pas réel, c'est un problème qu'il convient d'aborder.

# 3. Méthodes d'évaluation du risque de divulgation

Des méthodes spécifiques d'évaluation des risques de divulgation d'attribut et d'appartenance feront ci-dessous l'objet d'un examen. Tel qu'il est indiqué ci-dessus, le risque de divulgation de l'identité lié à des données entièrement synthétiques est faible et le risque perçu n'est pas un risque réel de divulgation. C'est la raison pour laquelle nous n'aborderons dans le présent article que les méthodes d'évaluation du risque de divulgation d'attribut et d'appartenance.

## 3.1. Divulgation d'attribut

Les attaques qui visent à divulguer des attributs sont appelées des *problèmes de classification de l'agresseur* (Hittmeir et coll., 2020; Hornby et Hu, 2021). L'agresseur, doté de certaines connaissances de base, du ou des ensembles de données synthétiques et des valeurs des attributs clés d'un certain enregistrement dans l'ensemble de données originales, cherche à prédire la valeur cible d'un certain enregistrement.

Tel qu'il est décrit dans l'article de Reiter et coll. (Reiter et coll., 2014), considérons un ensemble de données originales composé de microdonnées avec des enregistrements $n$ et des attributs $m$, où $D = \{(x_i, y_i): i = 1, \dots, n\}$ est l'ensemble de données, $x_i$ est le vecteur des valeurs des attributs non sensibles de l'enregistrement $i$, et $y_i$ est le vecteur des attributs sensibles de l'enregistrement $i$ qui font l'objet d'une synthèse. Veuillez prendre note que $X = \{x_i, i = 1, \dots, n\}$ est vide pour des données entièrement synthétiques, alors que toutes les variables seront synthétisées. Nous supposons également que les fournisseurs de données diffusent $m$ copies d'ensembles de données synthétiques, $Z = \{Z^{(1)}, \dots, Z^{(m)}\}$. Nous supposons que l'agresseur souhaite connaître l'attribut $y_i$ d'un enregistrement réel $i$, où $Y = \{y_i, i = 1, \dots n\}$. Désignons $B = \{A, S\}$ comme les connaissances préalables de l'agresseur, où $A$ est l'information connue sur les données originales et S représente les connaissances sur le synthétiseur (Reiter et coll., 2014). Ainsi, pour une supposition $y^*$ de $y_i$, l'agresseur récupère la distribution bayésienne a posteriori,

$$P(Y_i = y^* | Z, X, A, S) = \frac{P(Z | Y_i = y^*, X, A, S) P(Y_i = y^* | X, A, S)}{\sum_y P(Z | Y_i = y, X, A, S) P(Y_i = y | X, A, S)}. \quad (1)$$

Le fournisseur de données peut alors utiliser la valeur obtenue pour calculer plusieurs mesures de risque pour l'ensemble ou les ensembles de données synthétiques diffusés. Une méthode consiste à calculer

$$R_i = \left[ argmax_{y^*} P(Y_i = y^* | Z, X, A, S) = y_i \right], \quad (2)$$

puis à décider si $R = \sum_{i=1}^{n} R_i / n$ est suffisamment bas (Reiter et coll., 2014). Intuitivement, l'agresseur souhaite obtenir la probabilité maximale, parmi tous les $y_i$; ainsi, sa supposition, $y^*$, serait égale à $y_i$, d'après des renseignements préalables sur les données, les enregistrements et le modèle. Il est recommandé de maximiser $A$, c.-à-

d. de supposer que l'agresseur connaît tous les $y_i$ sauf celui visé par l'attaque. On a relevé que cette hypothèse s'avérait trop conservatrice, car dans de nombreux cas, ce niveau de connaissance ne serait pas réaliste (Reiter et coll., 2014). Néanmoins, Reiter et coll. ont montré que leur méthode peut être simulée et testée avec différentes hypothèses (Reiter et coll., 2014), ce qui signifie que nous pouvons comparer les changements dans les valeurs de R.

Si l'estimation bayésienne reste le choix le plus répandu pour évaluer la divulgation d'attribut (Hornby et Hu, 2021; Reiter et coll., 2014), il existe d'autres méthodes. En ce qui concerne le concept de *probabilité d'attribution correcte* (en anglais Correct attribution probability ou CAP) et ses dérivés (Hittmeir et coll., 2020), on suppose également que l'agresseur connaît les valeurs d'un ensemble d'attributs clés pour un individu dans l'ensemble de données originales, et qu'il souhaite connaître un attribut cible inconnu. Cette méthode ne suppose pas que l'agresseur ait accès au modèle synthétique lui-même ou qu'il le connaisse, ce qui la différencie des techniques d'estimation bayésienne mentionnées ci-dessus. La formule de l'indice de CAP est

$$CAP_{\blacksquare,j} := \mathrm{P}_{\blacksquare}\left(\mathrm{T_{o,j}} \,\middle|\, \mathrm{K_{o,j}}\right) = \frac{\sum_{i=1}^{n}[\mathrm{T}_{\blacksquare,i} = \mathrm{T_{o,i}} \wedge \mathrm{K}_{\blacksquare,i} = \mathrm{K_{o,j}}]}{\sum_{i=1}^{n}[\mathrm{K}_{\blacksquare,i} = \mathrm{K_{o,j}}]}, \quad (3)$$

Où $j \in \{1, \ldots, n\}$, $K_{\blacksquare,j}$ est un vecteur représentant les valeurs des attributs clés de l'enregistrement $j$ dans la base de données originales, $T_{\blacksquare,j}$ est la valeur correspondante de l'attribut cible, et $\blacksquare = \{o, s\}$ représente la donnée originale et la donnée synthétique, respectivement (Hittmeir et coll., 2020). L'idée à la base de cette attaque est que l'intrus peut rechercher dans tous les enregistrements des données synthétiques qui correspondent aux valeurs de l'attribut clé qu'il connaît. Ensuite, dans ces enregistrements, l'agresseur peut calculer la distribution des valeurs de l'attribut cible. La valeur moyenne de $CAP_s$ peut alors être comparée à $CAP_o$.

Étant donné que les hypothèses des techniques bayésiennes sont plus conservatrices à certains égards, bien que la *CAP* ne soit pas associée à des hypothèses, nous pouvons utiliser les deux catégories de techniques pour fixer une limite supérieure et une limite inférieure du risque d'attribut. Dans une situation plus extrême, nous pouvons utiliser la *CAP* pour estimer le risque de divulgation d'attribut des données originales, qui peut alors être défini comme la limite supérieure du risque.

## 3.2. Divulgation d'appartenance

Une des méthodes d'attaque est la méthode de partitionnement : un agresseur dispose d'un ensemble de données d'attaque, lequel est un échantillon de la même population que l'ensemble de données réelles qui est censé servir à entraîner le modèle de synthèse (Zhang et coll., 2022). L'agresseur apparie les enregistrements de son échantillon avec les données synthétiques. Il y a divulgation d'appartenance lorsqu'un enregistrement apparié se trouve également dans l'ensemble d'entraînement (réel). L'évaluation du risque de divulgation d'appartenance vise donc à quantifier la différence entre les données synthétiques et les données réelles. Nous définissons quelques termes statistiques pertinents qui sont couramment utilisés dans le cadre d'une évaluation des risques. La précision est définie comme la proportion de vrais positifs par rapport à tous les positifs, et le rappel est la proportion de vrais positifs par rapport aux vrais positifs et aux faux négatifs.

Afin d'estimer le risque sous-jacent à une telle attaque, l'administrateur des données peut séparer les données réelles en données d'entraînement et en données test. Nous supposons que l'agresseur dispose de renseignements complets sur les enregistrements $m$, où $mt$ sont tirés de l'échantillon d'entraînement et $m(1-t)$ sont tirés de l'échantillon test. $t \in (0,1)$ est une proportion choisie d'enregistrements sélectionnés pour faire partie de l'échantillon d'entraînement. La valeur par défaut $t$ sélectionnée dans le document est 0,5 (El Kababji et coll., 2023). Le modèle de synthèse est ensuite entraîné à partir des données d'entraînement, produisant des données synthétiques qui peuvent être comparées aux données réelles (tant les données d'entraînement et que les données test). Nous pouvons alors calculer la distance minimale entre chaque enregistrement des données réelles et des données synthétiques. Désignons $L$ comme la distance de Hamming et $y$ comme un enregistrement d'attaque, $y'$ comme un enregistrement synthétique et $h$ comme un seuil prédéfini. On considère qu'il y a eu appariement si $\min_{y'} L(y, y') \leq h$.

Alors, la cote $F1$ peut être calculée comme,

$$F1 = \frac{2 \times précision \times rappel}{précision + rappel} \quad (4)$$

La précision et le rappel sont tous deux calculés d'après le nombre d'enregistrements appariés qui se trouvent dans l'ensemble d'entraînement. La cote $F1$ est la moyenne harmonique de la précision et du rappel, représentant les deux de manière égale. La cote $F1$ indique le niveau de réussite des attaques. Le seuil établi dans un document de suivi de la cote $F1$ est 0,2 (El Kababji et coll., 2023), mais rien n'indique que ce seuil soit utilisé dans l'ensemble du domaine.

# 4. Discussion

## 4.1. Méthodologie

Si la littérature propose de nombreuses mesures pour évaluer le risque de divulgation, celles-ci ne sont pas sans poser de problèmes. L'un des problèmes les plus flagrants que posent les méthodes présentées ci-dessus est l'absence d'orientation concernant le seuil utilisé pour les mesures proposées, ou l'absence de consensus dans la littérature sur les seuils proposés. L'éq. (1) est une autre expérience de pensée intéressante, mais elle n'est pas utile dans une situation pratique telle quelle. Les auteurs ont estimé le risque en comparant le risque de diffuser l'ensemble de données synthétiques à celui de ne pas le faire. Étant donné qu'il faut créer un ensemble de données synthétiques, il n'a jamais été question de ne pas diffuser les données en premier lieu. Les techniques de divulgation d'appartenance se sont avérées davantage fondées sur des termes statistiques familiers que sur des cotes arbitraires. Les méthodologistes en tireraient profit étant donné qu'il existe déjà des valeurs acceptables pour, par exemple, les cotes F1 (même si elles s'appliquent davantage aux attributs catégoriels). À ce propos, il existe des mesures similaires à la cote F1, telles que $F_\beta$ (Goutte et Gaussier, 2005), qui peuvent être envisagées comme solution de rechange. Un autre point important à considérer est que ces mesures ne peuvent pas être comparées les unes aux autres, car, encore une fois, il n'existe pas de méthode de transposition entre les mesures.

Certaines mesures peuvent servir à comparer le risque entre deux ensembles de données diffusés, ce qui est beaucoup plus intuitif. L'équation (3) de la cote de la *CAP* constitue un bon moyen de comparer les méthodes classiques de contrôle de la divulgation, par exemple la suppression ou l'arrondissement, au recours à des données synthétiques. De cette manière, le risque peut également être calculé avec l'ensemble de données originales et l'ensemble de données synthétiques, et il est possible de comparer leurs valeurs pour considérer le risque comparatif. Bien qu'il n'y ait toujours pas de norme sur l'importance de la différence entre les deux valeurs, les comparaisons pourraient sembler moins arbitraires qu'une cote unique dérivée d'autres méthodes.

Un autre point qui mérite d'être pris en considération est que la plupart des publications considèrent les connaissances préalables comme un facteur clé dans la détermination du risque de divulgation, sans toutefois préciser le mode de quantification de ce renseignement. Les modèles de création d'ensembles de données synthétiques ne sont pas adaptés à la création d'enregistrements de données aberrantes tels que ceux que nous observons dans la vie réelle, par exemple Walmart parmi tous les magasins d'alimentation, car les modèles ne peuvent que générer des valeurs existantes ou, au mieux, être perturbés d'une manière ou d'une autre à partir de points de données existants (Elliot, 2014; Nowok et coll., 2016). Seules les techniques d'évaluation telles que l'éq. (2) qui tiennent compte de la divulgation maximale parmi tous les enregistrements peuvent prendre en considération les valeurs aberrantes. Les éq. (3) et (4) tiennent compte du risque moyen, qui, même s'il peut être biaisé par des valeurs aberrantes, sous-estime le risque de divulgation pour ces enregistrements en particulier.

## 4.2. Utilité

Dans le cadre de l'évaluation de la divulgation de tout ensemble de données, le risque de divulgation des données synthétiques ne peut être dissocié de leur utilité. Plus l'utilité attendue de l'ensemble de données est grande, plus le risque de divulgation est élevé. Certains s'aventurent même à dire que mesurer le risque d'un fichier revient à mesurer son utilité. La cote F1, telle qu'elle est présentée dans l'équation (4), est calculée directement à partir de la précision et du rappel (Goutte et Gaussier, 2005). Ceux-ci constituent deux mesures servant habituellement à déterminer l'utilité d'un modèle prédictif, ce qui signifie que la réduction du risque de divulgation entre directement en conflit avec le maintien simultané d'une utilité élevée. Les techniques bayésiennes, comme l'équation (2), estiment à quel point une supposition faite sur la base de renseignements antérieurs peut s'approcher de la réalité. Si l'utilité est maximisée, c'est-à-dire si l'on diffuse les données originales, la mesure de la cote calculée sera très élevée. Un ensemble de données synthétiques reproduisant parfaitement l'ensemble de données originales donnerait également lieu à une cote élevée. Cependant, à des fins de confidentialité, une cote plutôt faible est recommandée, ce qui signifie qu'il existe également un conflit inhérent entre la confidentialité et l'utilité pour les techniques bayésiennes.

L'une des hypothèses clés de nombreuses méthodes est la diffusion potentielle de plusieurs copies d'ensembles de données synthétiques à partir du modèle. Le principal avantage de la production de plusieurs ensembles de données est l'augmentation des données, ce qui signifie que toute déduction ou tout modèle tiré des ensembles de données synthétiques peut s'avérer plus robuste. Le risque de divulgation augmente avec le nombre de copies diffusées, de sorte que la diffusion de plusieurs copies doit être bien justifiée.

La diffusion du modèle synthétique en soi a des conséquences potentiellement divulgatrices. Bon nombre des techniques utilisées pour mesurer la divulgation font de la connaissance du modèle synthétique l'une des variables prises en compte, mais ne précisent pas comment cette notion doit être quantifiée. La diffusion du modèle ou des informations relatives au modèle présente des avantages évidents pour ses utilisateurs. Les utilisateurs peuvent créer davantage de copies des données synthétiques pour en accroître l'utilité, faire davantage confiance au modèle et à la méthodologie et contribuer à l'élaboration d'un modèle synthétique dans le cadre d'un schéma plus large. La diffusion du modèle sans faire preuve de prudence pourrait toutefois entraîner un risque élevé de divulgation.

## 4.3 Risque perçu

Outre le risque réel de divulgation, un problème plus insidieux à considérer est le risque perçu de divulgation. Même si l'on peut logiquement conclure qu'il n'y a aucune raison de croire que cet enregistrement synthétique contient des renseignements réels, le répondant réel dont les données ressemblent étrangement à l'enregistrement synthétique peut avoir l'impression que sa vie privée est menacée. Pour toute organisation statistique dont les fondements reposent sur la confiance, cette conclusion est extrêmement préjudiciable. Des enregistrements entièrement synthétiques peuvent finir par correspondre à des enregistrements originaux. Cette situation est suspecte, mais pas nécessairement problématique. La vraie question est de savoir si cela est dû à une mauvaise mise en œuvre du fichier synthétique lui-même (appariement réel) ou au caractère aléatoire de la méthode (appariement artificiel). La seconde option est, bien entendu, la plus souhaitable.

# 5. Recommandations

À la lumière de cette analyse documentaire, la méthode la plus applicable pour réduire le risque de divulgation lié aux données synthétiques consiste à diminuer les connaissances potentielles d'un agresseur d'un ensemble de données réelles en modifiant les données réelles avant d'effectuer la synthèse (Sallier, 2021). Nous recommandons de prendre les mesures suivantes pour les attributs continus afin de limiter l'étendue de la divulgation des attributs : troncation des valeurs extrêmes supérieures et inférieures, arrondissement à une position appropriée pour l'attribut, et regroupement des catégories en catégories plus larges.

Comme nous l'avons souligné précédemment, les valeurs aberrantes peuvent s'avérer problématiques; c'est pourquoi nous recommandons également de limiter le modèle utilisé dans la synthèse. Cela pourrait varier en fonction du modèle utilisé. En ce qui concerne les modèles basés sur un arbre de décision, il est possible de définir la taille minimale de la catégorie du nœud final (Gauvin, 2021). Par ailleurs, en ce qui concerne les valeurs aberrantes, on peut appliquer un lissage si les données ne sont pas trop asymétriques.

Enfin, d'après notre expérience et nos lectures, il apparaît évident que lorsque des données synthétiques sont produites à un niveau élevé, elles peuvent sembler très similaires aux données réelles. Les utilisateurs pourraient craindre qu'on divulgue certains renseignements, même lorsque les enregistrements sont tous synthétisés. Il est donc essentiel que toutes les données synthétiques soient bien étiquetées et identifiées comme telles afin d'éviter de tels problèmes. La génération de données synthétiques est encore une solution d'accès émergente. Une bonne communication sur ces produits de données est essentielle pour les utiliser de manière optimale.

## 6. Conclusion

Dans le cadre du présent article, nous avons examiné les différents types de risques de divulgation liés aux données synthétiques, ainsi que quelques exemples de techniques existantes d'évaluation des risques de divulgation. Les données synthétiques étant de plus en plus considérées comme un outil prometteur aux fins de contrôle de la divulgation, on demandera davantage à rendre ces techniques pratiques. Les mesures existantes dans la littérature doivent encore faire l'objet de recherches plus approfondies afin de mettre au point une procédure normalisée pour leur utilisation. Nous tenons également à souligner que les données synthétiques ne remplacent pas les données réelles. Il s'agit d'une étape intermédiaire qui aiderait les chercheurs à rationaliser leur analyse, étant donné qu'il peut s'avérer difficile d'accéder aux données réelles. Nous espérons néanmoins que la maturation des techniques d'évaluation des risques de divulgation pour les données synthétiques ne fera que soutenir l'utilisation pratique des données synthétiques elles-mêmes à Statistique Canada.

## Bibliographie


El Kababji, S., Mitsakakis, N., Fang, X., Beltran-Bless, A., Pond, G. R., Vandermeer, L., Radhakrishnan, D., Mosquera, L., Clemons, M. J., et El Emam, K. (2023), "Can synthetic data accurately mimic oncology clinical trials?", *Journal of Clinical Oncology*, 41(16_suppl), p. 1554. Article accessible à l'adresse https://10.1200/JCO.2023.41.16_suppl.1554.

Elliot, M. (2014), *Final Report on the Disclosure Risk Associated with the Synthetic Data*, Produced by the SYLLS Team.

Gauvin, H. (2021), "Generating smart deep files: the example of synthesizing hierarchical data" in *Proceedings of Statistics Canada Symposium 2021 Adopting Data Science in Official Statistics to Meet Society's Emerging Needs***,** Catalogue no. 11-522-x.

Goutte, C., et Gaussier, E. (2005), " A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation", In Losada, D.E., Fernández-Luna, J.M. (eds) *Advances in Information,* p. 345-359. Article accessible à l'adresse https://doi.org/10.1007/978-3-540-31865-1_25.

Hittmeir, M., Mayer, R., et Ekelhart, A. (2020), "A Baseline for Attribute Disclosure Risk in Synthetic Data", Dans *Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy*, p. 133–143. Article accessible en ligne à l'adresse https://dl.acm.org/doi/10.1145/3374664.3375722

Hornby, R., et Hu, J. (2021), *Bayesian Estimation of Attribute Disclosure Risks in Synthetic Data with the AttributeRiskCalculation R Package.* Manuscrit non publié. Accessible à l'adresse https://arxiv.org/abs/2103.09805

Hu, J. (2021), *Bayesian Estimation of Attribute and Identification Disclosure Risks in Synthetic Data.* . Ithaca: Cornell University Library, arXiv.org. https://10.48550/arxiv.1804.02784 Extrait du contenu de la base de données accessible au public à l'adresse https://search.proquest.com/docview/2071976619

Hu, J., Reiter, J. P., et Wang, Q. (2014), "Disclosure Risk Evaluation for Fully Synthetic Categorical Data", *Privacy in Statistical Databases*, p. 185–199. Springer International Publishing. Accessible à l'adresse https://10.1007/978-3-319-11257-2_15 .

Nowok, B., Raab, G. M., et Dibben, C. (2016), *synthpop: Bespoke Creation of Synthetic Data in R*. Foundation for Open Access Statistic. Accessible à l'adresse https://10.18637/jss.v074.i11.

Reiter, J. P. (2005), *Estimating Risks of Identification Disclosure in Microdata*. Informa UK Limited. Accessible à l'adresse https://10.1198/016214505000000619.

Reiter, J. P., et Mitra, R. (2009). Estimating Risks of Identification Disclosure in Partially Synthetic Data. *The Journal of Privacy and Confidentiality*, 1(1), p. 99-110. Article accessible à l'adresse https://10.29012/jpc.v1i1.567.

Reiter, J. P., Wang, Q., et Zhang, B. (2014), "Bayesian Estimation of Disclosure Risks for Multiply Imputed, Synthetic Data", *The Journal of Privacy and Confidentiality*, 6(1). Article accessible à l'adresse https://10.29012/jpc.v6i1.635

Sallier, K. (2021), "Synthetic Data for National Statistical Organizations: A Starter Guide: Methods and Recommendations", Article présenté dans le webinaire sur les données synthétiques de l'atelier annuel de l'UNECE HLG-MOS.

UNECE (2022), *Synthetic Data for Official Statistics: A Starter Guide.* Commission économique pour l'Europe des Nations unies. Accessible à l'adresse https://unece.org/sites/default/files/2022-11/ECECESSTAT20226.pdf .

Zhang, Z., Yan, C., et Malin, B. A. (2022), "Membership inference attacks against synthetic health data", *Journal of Biomedical Informatics*, 125. Article accessible à l'adresse https://10.1016/j.jbi.2021.103977.

Zhao, J., Chen, Y., et Zhang, W. (2019), "Differential Privacy Preservation in Deep Learning: Challenges, Opportunities and Solutions", *IEEE Access*, 7, p. 48901–48911. Article accessible à l'adresse https://10.1109/ACCESS.2019.2909559.