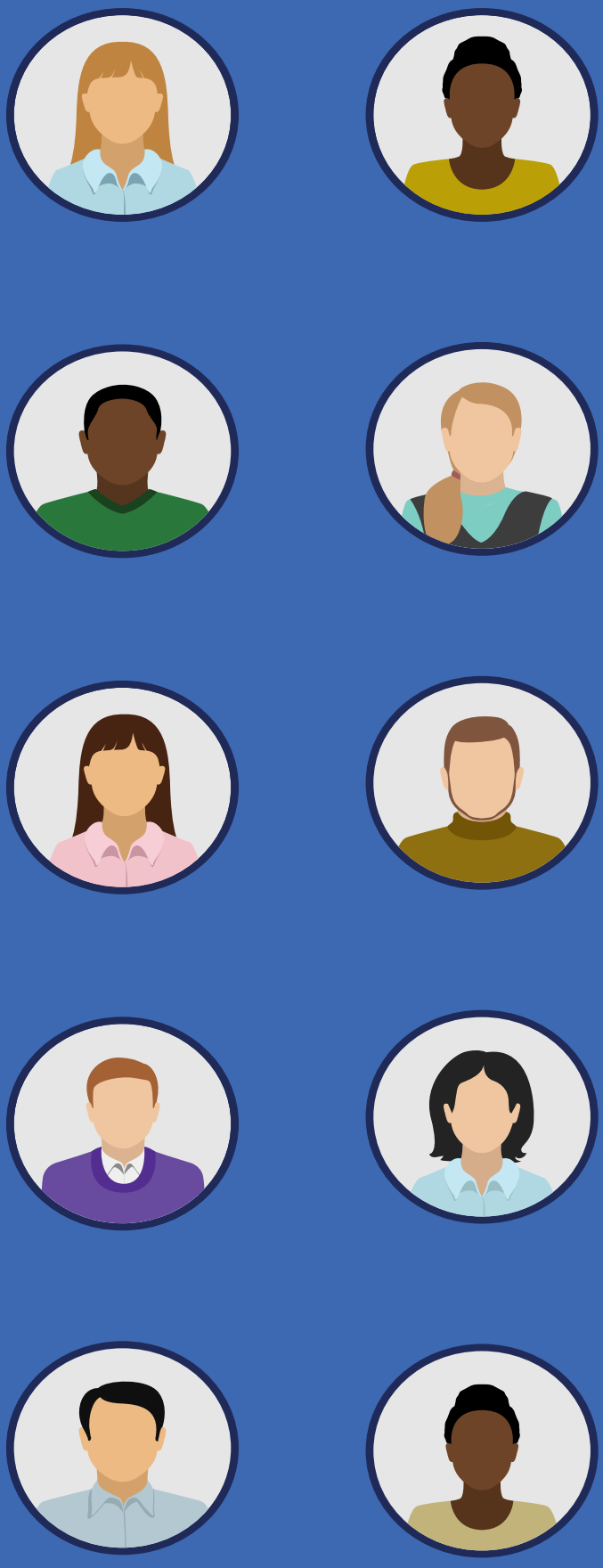


avril 2022

Dépersonnalisation des renseignements personnels pour échanger des données

Institut du savoir
sur la santé mentale et les dépendances
chez les enfants et les jeunes

Suggested citation:

L'Institut du savoir sur la santé mentale et les dépendances chez les enfants et les jeunes. *Dépersonnalisation des renseignements personnels pour échanger des données.* (2022). L'Institut du savoir sur la santé mentale et les dépendances chez les enfants et les jeunes. www.smdej.ca/equite

Pour de plus amples renseignements au sujet de ce rapport, veuillez communiquer avec Evangeline Danseco à edanseco@cymha.ca

Table des matières

Qu'est-ce que la dépersonnalisation?

La dépersonnalisation des données consiste à retirer tout renseignement personnel susceptible de permettre l'identification d'une personne. Ces renseignements comprennent notamment les renseignements personnels sur la santé.

Pourquoi dépersonnaliser les données?

Lorsque deux ou plusieurs organismes s'échangent des données contenant des renseignements personnels sur des client.e.s, il faut tenir compte de certaines exigences législatives et éthiques. La dépersonnalisation assure la protection des renseignements personnels lorsque les données sont ainsi communiquées. L'organisme qui reçoit les données dépersonnalisées peut les analyser et les utiliser à des fins stratégiques, par exemple pour appuyer les prises de décisions ou apporter des améliorations aux programmes et aux services.

Le présent document présente quelques recommandations sur la façon dont votre organisme peut recueillir, utiliser et communiquer des données.

Renseignements permettant d'identifier une personne

Les « renseignements permettant d'identifier une personne » englobent tous les renseignements qui peuvent être utilisés, soit individuellement ou en combinaison avec d'autres renseignements, pour déterminer l'identité d'une personne. Ces renseignements sont regroupés en deux catégories : 1) les renseignements directs; et 2) les renseignements indirects, ou quasi identifiants. Voici quelques exemples pour chacune des catégories.

Renseignements directs :

- Nom
- Adresse
- Numéro de téléphone
- Adresse de courriel
- Numéro de carte santé
- Numéro du dossier médical
- Adresse IP
- Numéro d'assurance sociale
- Tout autre numéro d'identification unique

Renseignements indirects ou quasi identifiants :

- Date de naissance
- Genre
- Code postal
- Dates d'événements (admission, visite à la clinique, congé)
- Origine ethnique
- Pays d'origine
- Profession
- Diagnostic

- Toute caractéristique unique ou qui n'est pas courante

Les informations tirées d'un renseignement indirect, comme le code postal, pourraient ne pas suffire pour identifier une personne. Cependant, l'utilisation de ce renseignement en combinaison avec d'autres, comme le pays d'origine, pourrait permettre de déterminer l'identité de cette même personne. Il est également important de noter que, dans certains cas, même un renseignement indirect rare pourrait faciliter cette identification. Par exemple, si dans une collectivité il n'y a qu'une famille d'immigrants provenant d'Islande, les données sur le pays d'origine pourraient permettre d'associer cette famille à un.e client.e.

Étape 1 : Sélectionner les données pertinentes à communiquer

Les données que votre organisme choisit de recevoir et d'utiliser doivent se limiter aux renseignements pertinents à l'analyse que vous entreprenez. Par exemple, si l'analyse des données de votre organisme vise uniquement à réduire les temps d'attente dans les cliniques sans rendez-vous, il ne faut pas inclure dans l'ensemble de données à analyser tout renseignement qui n'est pas lié aux délais d'attente, par exemple, le numéro de carte Santé.

Toutes les données d'identification essentielles à votre analyse doivent être classées en renseignements directs et indirects pour passer à l'étape

de la dépersonnalisation.

Étape 2 : Choisir un cadre de dépersonnalisation

Deux types de cadres peuvent être utilisés pour dépersonnaliser les données : le cadre heuristique et le cadre axé sur les risques.

- La « dépersonnalisation heuristique » applique des règles générales de dépersonnalisation aux données pour supprimer ou généraliser les renseignements.
- La « dépersonnalisation axée sur les risques » évalue le degré de risque global lié à l'échange des données en examinant le contexte et les facteurs de risque liés aux données.

L'approche axée sur les risques est recommandée par le Commissaire à l'information et à la protection de la vie privée de l'Ontario (CIPVP). Bien que ce cadre soit considéré comme la pratique exemplaire pour dépersonnaliser des données, il peut être difficile d'évaluer le niveau de risque lié à l'échange des données sans utiliser un logiciel de dépersonnalisation ou sans faire appel à des spécialistes du domaine. Des lignes directrices pour l'application de cette approche se trouvent sur le site Web du CIPVP.

Dépersonnalisation heuristique

La dépersonnalisation heuristique consiste à supprimer ou à généraliser des renseignements permettant d'identifier une personne. Cette approche élimine le besoin d'utiliser des renseignements directs pour effectuer des analyses. Ainsi, ces renseignements peuvent être retirés de l'ensemble de données.

La « généralisation » est un processus qui vise à agréger des données en regroupant les renseignements dans des catégories plus vastes, ce qui les rend moins précis. L'objectif est de réduire le nombre d'inscriptions uniques dans les données afin de limiter la probabilité qu'une personne puisse être identifiée. Si la généralisation des données ne permet pas de dépersonnaliser entièrement certains champs, ces valeurs ou inscriptions doivent être supprimées. Il faut par ailleurs porter une attention particulière à la mesure dans laquelle les données sont généralisées, car

certaines formes de dépersonnalisation peuvent limiter l'utilisation des données dans le cadre d'analyses ultérieures.

Les inscriptions de données qui comprennent les mêmes renseignements indirects sont appelées « classe d'équivalence ». Les données sont agrégées en catégories plus vastes pour augmenter la taille de chaque classe d'équivalence. La taille minimale de chaque classe d'équivalence pour assurer la protection de la vie privée est de cinq inscriptions[1].

Exemple

Examinez le tableau 1 ci-dessous en portant uniquement attention aux colonnes de l'âge et du genre. Vous pouvez constater que les personnes du même âge et du même genre forment une classe d'équivalence. Seules deux personnes présentent les mêmes valeurs d'âge (35 ans) et de genre (H). Ces deux personnes forment une classe équivalente (hommes de 35 ans) dont la taille de groupe est de 2.

S'il y avait eu une autre ligne où l'âge et le genre étaient égaux à « 35 » et « H » respectivement, alors la taille de la classe d'équivalence des hommes de 35 ans aurait été de 3. Comme toutes les autres personnes dans ce tableau n'ont pas le même âge et le même genre, chacune d'entre elles forme sa propre classe d'équivalence, dont la taille est de 1.

Le tableau 1 illustre comment les données peuvent être généralisées pour modifier la taille des classes d'équivalence. Lorsque l'âge est transformé en catégorie d'âge, la taille des classes d'équivalence change. Cependant, une inscription unique est maintenue, malgré la catégorisation de la variable de l'âge (femme de 90 ans). Dans de tels cas, le champ de l'âge de cette inscription peut être généralisé (> 50), la valeur de l'âge peut être supprimée, ou l'inscription (c'est-à-dire l'intégralité de la ligne) peut être supprimée.

Il convient de noter que l'exemple fourni vise à démontrer comment la taille des classes d'équivalence peut être modifiée. Dans cet exemple, pour que les données soient considérées comme dépersonnalisées, la taille de la plus petite classe d'équivalence devrait être de 5.

1 *Privacy Analytics*, 2020. *« The definitive guide to de-identification »* [livre blanc].

Tableau 1: Exemple de généralisation de données en catégories

Âge	Catégorie d'âge	Genre
35	De 30 à 40	H
35	De 30 à 40	H
42	De 40 à 50	F
38	De 30 à 40	H
48	De 40 à 50	F
50	De 40 à 50	F
90	> 50	F

Utilisation des variables de l'âge et du genre :

- pour les hommes de 35 ans, la taille de la classe est n=2;
- toutes les autres inscriptions sont uniques; la taille de la classe est de 1 pour chaque ligne.

Utilisation de la catégorie d'âge et des variables du genre :

- pour les hommes de 30 à 40 ans, la taille de la classe est n=3;
- pour les femmes de 40 à 50 ans, la taille de la classe est n=3;
- pour les femmes de > 50 ans, la taille de la classe est n=1.

Dépersonnalisation axée sur les risques[2]

Cette méthode évalue le degré de risque global lié à l'échange des données en examinant les facteurs de risque contextuels et associés aux données. Elle vise à déterminer le degré de dépersonnalisation requis en fonction du risque de réidentification des personnes dans l'ensemble de données. La « réidentification » est le concept voulant que des renseignements tirés d'un ensemble de données dépersonnalisé soient associés à une ou plusieurs personnes en particulier. Les étapes de la dépersonnalisation axées sur les risques sont décrites ici.

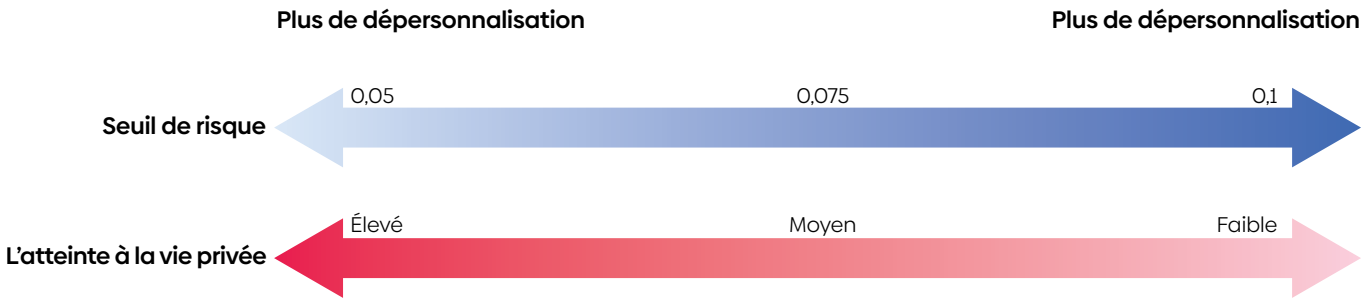
1. Déterminer les variables de renseignements directs et indirects.
2. Déterminer le seuil de risque de réidentification acceptable.
3. Mesurer le degré de risque global (risque lié aux données multiplié par le risque lié au contexte).
4. Dépersonnaliser les données, puis réévaluer le degré de risque global et s'assurer qu'il est inférieur au seuil de risque établi.

2 Cette section a été adaptée depuis le guide *De-identification guidelines for structured data* (en anglais seulement) du CIPVP.

Seuil de risque de réidentification

Avant d'effectuer une dépersonnalisation, l'organisme qui communique les données doit déterminer un seuil de risque de réidentification acceptable. Il s'agit du « degré minimal de dépersonnalisation qui doit être appliqué à un ensemble de données pour qu'il soit dépersonnalisé[3] » [traduction]. Quant au seuil de réidentification, il est le degré de risque de référence par rapport auquel le degré de risque global est mesuré.

Le CIPVP recommande d'établir le seuil de risque de réidentification en fonction d'une évaluation de l'atteinte à la vie privée[4]. Cette évaluation est « la mesure dans laquelle la divulgation de l'ensemble de données porterait atteinte à la vie privée d'une personne » [traduction] et est classée comme faible, moyenne ou élevée. La figure ci-dessous présente une ligne directrice sur les seuils de risque pour chaque niveau d'atteinte à la vie privée et le risque de réidentification correspondant. Les scénarios dans lesquels l'atteinte à la vie privée d'une personne est jugée élevée devraient viser un seuil de risque de réidentification plus faible (inférieur à 0,1). Dans un tel cas, une dépersonnalisation des données plus poussée serait nécessaire.



Risque global

Le risque global est calculé en multipliant le risque lié au contexte par le risque lié aux données. Ensuite, ce risque global est comparé au seuil de risque pour déterminer le degré de dépersonnalisation requis.

3 Commissaire à l'information et à la protection de la vie privée de l'Ontario, 2016. *« De-identification guidelines for structured data »*.

4 Veuillez consulter le guide *De-identification guidelines for structured data* du Commissaire à l'information et à la protection de la vie privée pour obtenir la liste des facteurs à prendre en compte dans l'évaluation de l'atteinte à la vie privée.

Risque lié au contexte

Le risque lié au contexte évalue la probabilité d'une attaque par réidentification portée contre les données. Une attaque par réidentification prend notamment la forme de piratage ou de cybermenaces et elle constitue une tentative visant à déterminer l'identité de personnes contenues dans un ensemble de données dépersonnalisées. Différents types d'attaques peuvent se produire. Celles-ci varient selon le type de diffusion des données6 (diffusion publique ou privée). Il peut être difficile de mesurer le degré de risque lié au contexte sans l'aide de spécialistes du domaine ou sans utiliser un logiciel de dépersonnalisation.

Pour simplifier l'évaluation du degré de risque lié au contexte expliquée dans le présent document, nous utiliserons le niveau de risque contextuel le plus élevé, dont la valeur est 1, dans un cas d'attaque par réidentification. Pour obtenir une description plus détaillée du risque lié au contexte et des types d'attaques par réidentification, veuillez consulter le guide De-identification Guidelines for Structured Data du CIPVP.

Risque lié aux données

Le risque lié aux données est déterminé en mesurant la probabilité de trouver l'identité des personnes pour chaque ligne inscrite dans les données. Le risque de réidentification pour chacune de ces lignes est calculé en déterminant la proportion inverse de la taille de la classe d'équivalence (1/taille de la classe d'équivalence). Par exemple, chaque ligne d'une classe d'équivalence dont la taille est de 10 aura une probabilité de 1 sur 10 d'être correctement identifiée. Les lignes des classes d'équivalence plus petites sont les plus susceptibles d'être réidentifiées, comparativement aux classes d'équivalence plus grandes. Le risque de réidentification qui s'y rapporte doit être appliqué à toutes les lignes inscrites dans les données au moment de déterminer le degré de risque lié aux données.

Par exemple, un ensemble de données dont les classes d'équivalence sont 5, 10 et 20 comporterait des risques de réidentification correspondants de 0,20, 0,10 et 0,05 respectivement. La classe d'équivalence d'une taille de 5 lignes comporte le plus grand risque de réidentification (0,20) comparativement aux autres classes d'équivalence. Par conséquent, le degré de risque lié aux données pour cet ensemble de données est de 0,20, puisqu'il s'agit du risque de réidentification de la plus petite classe d'équivalence.

Exemple

Deux organismes ont convenu d'échanger des données contenant des renseignements médicaux personnels et ont signé une entente d'échange de données. Comme il s'agit d'une divulgation de données privées, on peut supposer que le risque d'atteinte à la vie privée serait faible, ce qui nécessiterait un degré moindre de dépersonnalisation. L'organisme titulaire des données établit un seuil de risque de réidentification de 0,2.

- Le degré de risque global avant la dépersonnalisation a été évalué à 0,33 (risque lié aux données de 0,33 multiplié par le risque lié au contenu de 1,0).
- Le résultat du risque global devient ainsi 0,1 après la dépersonnalisation.
- Étant donné que cet indice est inférieur au seuil déterminé de 0,2, les données ont été dépersonnalisées à un niveau acceptable et peuvent être communiquées d'un organisme à l'autre.

Étape 3 : Dépersonnalisation des données

La dernière étape du processus consiste à dépersonnaliser les données. Les renseignements directs peuvent être remplacés, tandis que les renseignements indirects peuvent être remplacés et généralisés, conjointement à d'autres techniques, pour modifier la taille des classes équivalentes. De plus, des logiciels spécialisés peuvent être utilisés pour effectuer la dépersonnalisation.

Une fois ce processus terminé, le degré de risque global est réévalué et comparé au seuil de risque de réidentification initialement établi.

Conventions pour dépersonnaliser les renseignements indirects courants

La présente section établit des exemples de généralisation pour trois renseignements indirects couramment utilisés : les dates, les codes postaux et les caractéristiques non courantes.

Dates

Les dates sont habituellement présentées au format « jour, mois, année ». Les données de date généralisées peuvent être présentées sous forme de mois et d'année (12/2000), d'année (2000) ou d'intervalle (2000–2010). Par ailleurs, les dates peuvent être indiquées sous forme d'intervalles entre les dates.

Par exemple, plutôt que d'indiquer les dates d'admission et de congé, les données peuvent être présentées sous forme de durée du séjour en jours. Il faut tenir compte de l'utilité des données transformées pour s'assurer qu'elles peuvent être utilisées dans le cadre d'analyses ultérieures. Les dates de naissance peuvent également être généralisées pour être transformées en âge. Cette donnée peut ensuite être généralisée davantage en catégories d'âge.

Code postal

Les données sur les codes postaux peuvent être généralisées pour n'inclure que les trois premiers caractères alphanumériques. D'après les données du Recensement de 2016, la taille moyenne de la population contenue dans les trois premiers caractères d'un code postal canadien est d'environ 22 000 personnes. Selon la région géographique, ce nombre peut varier de 0 à plus de 100 000 personnes[5]. Si la taille de la population dans la région géographique visée est trop petite, le code postal doit être généralisé davantage à une région plus vaste, comme les deux premiers caractères ou le premier caractère seulement.

Par exemple, la taille de la population vivant dans la région couverte par le code postal P0Y est de 30 personnes. Si ce code postal est généralisé pour être porté aux deux premiers caractères (P0), la taille totale de la population dans la région géographique augmente à plus de 200 000 personnes.

Données non courantes ou uniques

Les champs qui comportent des éléments de données non courantes ou uniques doivent être modifiés et regroupés en catégories plus vastes et plus générales. Par exemple, une profession unique comme « premier ministre » devrait être généralisée et porter l'appellation « politicien ».

---

5 *Statistique Canada*, 20 février 2019. *« Chiffres de population et des logements – Faits saillants en tableaux, Recensement de 2016 »*.

D'autres exemples peuvent inclure la généralisation du pays d'origine pour indiquer plutôt la région (par exemple, remplacer « Somalie » par « Afrique subsaharienne ») et la généralisation des diagnostics en grandes catégories de diagnostic (par exemple, remplacer « trouble dépressif » par « trouble de l'humeur »).

Résumé

- Lorsque des organismes échangent des données, la dépersonnalisation – soit le retrait de renseignements personnels permettant d'identifier une personne – est nécessaire pour protéger la vie privée des individus.
- Avant de dépersonnaliser des données, les organismes doivent sélectionner les données pertinentes et choisir l'un des deux cadres de dépersonnalisation possibles : le cadre heuristique ou le cadre axé sur les risques.
- Le cadre heuristique fait appel aux règles générales de dépersonnalisation pour supprimer ou généraliser des données. La dépersonnalisation axée sur les risques, quant à elle, évalue au moyen de calculs le degré de risque global lié à l'échange des données afin de déterminer l'ampleur de la dépersonnalisation requise.
- Dans les deux cas, la dépersonnalisation consiste à remplacer et à généraliser des données afin de réduire la probabilité qu'elles puissent être réidentifiées.

Références

Cavoukian, A., & El Emam, K. (2014). *De-identification protocols: Essential for protecting privacy.* Commissaire à l'information et à la protection de la vie privée de l'Ontario.

Commissaire à l'information et à la protection de la vie privée de l'Ontario. (2016). *De-identification guidelines for structured data.*

Commissaire à l'information et à la protection de la vie privée de l'Ontario. (2018). *General data protection regulation.*

El Emam, K., Jabbouri, S., Sams, S., Drouet, Y., & Power, M. (2006). Evaluating common de-identification heuristics for personal health information. *Journal of Medical Internet Research*, 8(4), e28.

Fraser, R., & Willison, D. (2009). *Tools for de-identification of personal health information.* Electronic Health Information Laboratory.

Instituts de recherche en santé du Canada comité consultatif sur la protection de la vie privée. (2005). *Pratiques exemplaires des IRSC en matière de protection de la vie privée dans la recherche en santé.* Instituts de recherche en santé du Canada.

Nelson, G. S. (2015). *Practical implications of sharing data: A primer on data privacy, anonymization, and de-identification.* Thotwave Technologies.

Office of Civil Rights. (2012). *Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) privacy rule.*

Privacy Analytics, 2020. *« A privacy governance framework to support de-identification »* [livre blanc].

Privacy Analytics, 2020. *« De-identification 201 »* [livre blanc].

Privacy Analytics, 2020. *« De-identification 301 »* [livre blanc].

Privacy Analytics, 2020. *« The definitive guide to de-identification »* [livre blanc].

CYMHA_ON

cymhaon

cymha_on

695, avenue Industrial, Ottawa (Ontario) K1G 0Z1

— 1 613 737 2297

— 1 613 738 4894

FR — smdej.ca

EN — cymha.ca