Protection des données dans la recherche

Appariement et protection des données de santé: une contradiction?

Nicole Steck, Adrian Spoerri, Matthias Egger

Institut pour la médecine sociale et préventive (ISPM), Université de Berne

Tandis que l'appariement des données de santé permet à la recherche médicale de se pencher avec efficacité sur des questions importantes, il n'est pas rare que la protection des données dresse un obstacle infranchissable. Grâce à la méthode développée par l'Institut de médecine sociale et préventive à Berne, l'appariement de données sensibles devient possible sans l'échange d'informations identifiables.

La collecte de nouvelles données de santé pour la recherche est une activité chère et laborieuse. Pour les participants, cela signifie un investissement temporel important sur plusieurs années associé à des analyses désagréables. C'est pourquoi il serait souhaitable de faire en sorte que les données de santé existantes soient utilisées, et appariées, le mieux possible pour la recherche. Il serait ainsi possible d'en améliorer la qualité et l'exhaustivité et de se pencher sur de nouveaux sujets de recherche. Concernant les études longitudinales, il n'est pas rare qu'après quelques années seule une partie des patients initiaux soit encore sous observation, ce qui peut se traduire par une distorsion des résultats (biais de sélection). Mais il est possible d'empêcher ce biais en déterminant le statut vital de tous les patients initiaux par un appariement avec les données de mortalité. Une étude à long terme de l'efficacité des mesures préventives sur les patients âgés, réalisée dans les cabinets de médecins de famille de Soleure, a identifié de cette manière le statut vital de 98,2% des participants [1].* Autre exemple de succès pour une étude *record linkage*: l'appariement des données des recensements de 1990 et de 2000 avec celles du registre suisse du cancer de l'enfant a permis d'étudier un éventuel lien entre la proximité d'une centrale nucléaire et les cas de leucémie chez les enfants [2]. Dans le présent article, nous abordons les possibilités d'apparier les banques de données pour la recherche en Suisse et présentons une nouvelle méthode, grâce à laquelle les données de santé sensibles peuvent être croisées avec d'autres par des institutions tierces indépendantes, les centres de confiance, sans l'échange d'informations permettant d'identifier les personnes concernées.

* Les références se trouvent sous www.bullmed.ch → Numéro actuel ou → Archives → 2015 → 50/51.

Appariement des banques de données et consentement éclairé

De manière générale, il existe deux possibilités. Pour des appariements déterministes simples, il s'agit typiquement de réunir les entrées correspondant à un numéro d'identification bijectif. Dans les pays scandinaves, il est possible et autorisé d'apparier différents fichiers de données en utilisant un «numéro personnel». Une étude danoise publiée récemment s'est penchée par exemple sur la question de savoir si la prise d'antiépileptiques pendant la grossesse constituait un facteur de risque de fausse couche ou de mortinaissance [3]. Pour cela, il a été procédé à un appariement des données du registre des naissances avec celles concernant les sorties d'hôpital et les prescriptions de médicaments.

En comparaison à la Scandinavie, Statistique suisse (Office fédéral de la statistique, OFS) recueille moins de données sur la santé. Mais leurs appariements ne peuvent être réalisés que par l'Office fédéral de la statistique dans le cadre d'une convention sur la protection des données (cf. l'ordonnance sur l'appariement de données [4]). En revanche, la recherche n'a généralement pas le droit d'utiliser le numéro d'assuré social introduit en 2008 pour apparier les données.

En principe, un appariement déterministe est aussi possible avec le nom, le sexe, le code postal et la date de naissance. C'est cependant difficile parce que les données sont saisies de manière inconsistante et qu'elles font en partie défaut. C'est pourquoi, en absence de numéro d'identification explicite, on procède fréquemment à ce qu'on appelle un appariement «probabiliste». Dans ce cas, il ne s'agit pas d'apparier uniquement des données qui se correspondent exactement mais de cal-

Groupe de dialogue «Recherche prioritaire: recherche sur les soins»

Pour le corps médical, la recherche sur les soins constitue un domaine scientifique important porteur d'avenir. Dans un secteur de la santé en pleine mutation (nouveaux modèles de financement et de soins, évolutions démographiques, transferts sectoriels, etc.), il est impératif que la recherche sur les soins bénéficie d'un ancrage académique. Afin de pouvoir créer les bases scientifiques d'une recherche indépendante dénuée de tout intérêt particulier, la Fédération des médecins suisses (FMH), la Conférence des sociétés cantonales de médecine (CCM) et NewIndex soutiennent le groupe de dialogue «Recherche prioritaire: recherche sur les soins» de l'Institut de médecine sociale et préventive de l'Université de Berne.

A l'instar d'un forum, ce groupe de dialogue a pour but de solliciter l'échange d'informations: les représentants des organisations mentionnées et des groupes de recherche discutent régulièrement des travaux en cours et des projets à venir dans le domaine de la recherche sur les soins. Par ailleurs, le groupe de dialogue vise à sensibiliser le corps médical à la recherche sur les soins et à en favoriser l'acceptation tout en soulignant les avantages concrets de cette recherche pour le corps médical. Il est ouvert aux propositions en ce qui concerne les sujets à traiter, les questions, les discussions ou les demandes d'informations supplémentaires. La division Données, démographie et qualité (DDQ) de la FMH assure la coordination du groupe de dialogue et se tient à disposition pour tout complément d'information par courriel à ddq[at]fmh.ch ou par téléphone au 031 359 11 11.

culer la probabilité de réunir deux entrées d'une même personne malgré les divergences. Si par exemple dans un fichier de données, une date de naissance est indiquée par «12.02.1984», l'appariement probabiliste considéra de la même manière le «13.02.1984» ou le «02.12.1984» que le «27.09.2001». Sur la base des probabilités calculées, on peut ensuite déterminer si des entrées peuvent être attribuées à une même personne. En Suisse, les appariements probabilistes avec la date de naissance, le sexe et le lieu de domicile permettent d'obtenir de bons résultats [5]. Si en plus le nom et le prénom sont disponibles, les résultats sont même aussi bons qu'avec un numéro d'identification. Mais le nom, le prénom et la date de naissance constituent un problème en raison de la protection des données et, en général, ils ne peuvent être utilisés qu'avec un accord explicite du patient (consentement éclairé). C'est pourquoi il est recommandé, principalement pour les études de longue durée, d'ajouter une remarque en conséquence dans la déclaration de consentement. Les exceptions concernent les projets de recherche avec par exemple des données de santé anonymisées par cryptage ou collectées de manière anonyme.

Le cryptage des données de santé

Même si le patient a donné son consentement, il s'agit de favoriser la protection des données. En présence d'informations comme le nom, le prénom et la date de naissance pour apparier les données, il est préférable de veiller à ce que personne ne puisse être identifié. C'est possible en cryptant les données, c'est-à-dire en transformant les noms ou dates de naissance en série de lettres ou de chiffres non identifiables. Si le cryptage utilisé est le même sur deux fichiers de données, les séries de lettres ou de chiffres correspondants peuvent être appariées. Il est cependant indispensable que le cryptage ne soit possible qu'à sens unique. Malheureusement, les programmes de cryptage usuels ne sont pas adaptés à l'appariement des données de santé car une légère différence dans un terme à crypter débouche sur deux valeurs totalement différentes. Ainsi «Emmenegger» et «Emmeneger» ne sont plus considérés comme un seul et même nom de famille après le cryptage. Une simple faute d'orthographe peut donc avoir des conséquences considérables. A l'instar de l'appariement probabiliste, pour lequel la ressemblance des entrées prime, des méthodes de cryptage adéquates ont été recherchées, avec l'appui des fameux filtres de Bloom *(Bloom Filters)* [6]. Ces derniers permettent de calculer la similitude entre les données cryptées. L'appariement est ainsi rendu possible en dépit d'une faute d'orthographe ou autre (tab. 1).

La méthode P3RL de l'ISPM Berne

Dans ce contexte, l'Institut pour la médecine sociale et préventive (ISPM) de l'Université de Berne a développé un logiciel pour un «Privacy Preserving Probabilistic Record Linkage» (P3RL), qui permet d'apparier les données de santé personnelles stockées dans différentes banques de données tout en respectant la protection des données [7]. La méthode P3RL est pertinente lorsque les données de santé doivent être appariées par au moins deux centres différents, qui ne possèdent pas de numéros d'identification communs et dont le règlement en matière de protection des données limite l'utilisation de variables permettant d'identifier la personne comme le nom, la date de naissance, la date de décès ou l'adresse. Dans ce cas, l'appariement est réalisé par un centre de confiance indépendant. La méthode P3RL se compose de trois étapes: préparation des données, cryptage et appariement (fig. 1).

Tableau 1: Comparaison du cryptage d'un nom selon différentes méthodes. La ressemblance entre les deux noms initiaux n'est visible qu'avec le cryptage avec filtre de Bloom.

Texte	Cryptage conventionnel*	Exemple de cryptage avec filtre de Bloom
Emmenegger	078f73ae3b2852b79e143a06aa573f21	11111111111111011101100111101110010101011110010101
Emmeneger	21783f44f4696323a2267a83a2f2dd7b	11111111111111011101100111101110010101001110010101
Meier	3399f3b498509a2f63b058db71a360f3	11011010111101000001001110110111111010011111111011

* en utilisant MD5 Hash

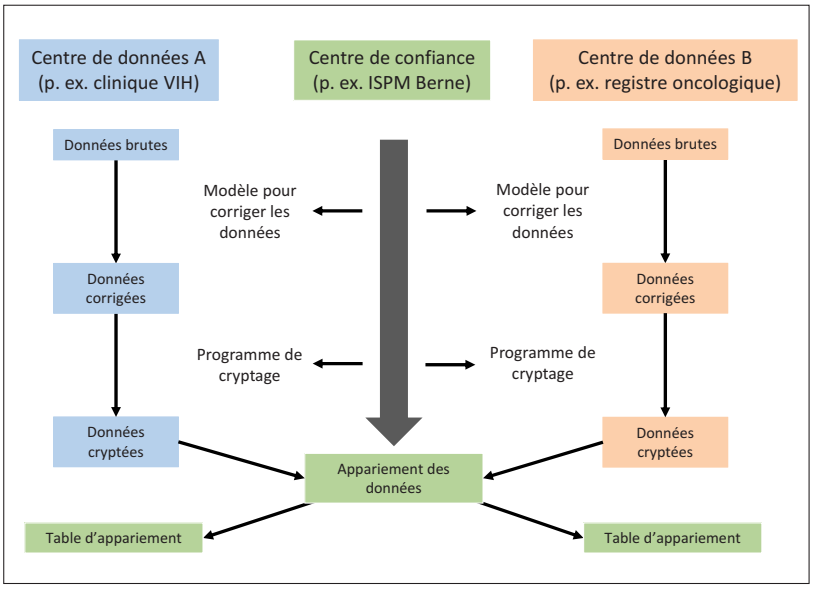

Figure 1: Déroulement et répartition des phases de travail selon la méthode P3RL à l'exemple des fichiers de données d'une clinique VIH et d'un registre du cancer qui doivent être appariés par le centre de confiance de l'ISPM à Berne.

1. Préparation des données

Dans un premier temps, les particularités et les erreurs des fichiers de données à apparier sont corrigées selon des règles uniformes. Par exemple, les caractères spéciaux ou spécifiques au langage sont uniformisés, les termes antéposés (Madame ou Dr) ou postposés (jun.) sont enlevés, une variable par prénom est créée, les dates de naissance sont présentées selon le même format, ainsi que les informations manquantes. A cet effet, le centre de confiance met à la disposition des autres centres un catalogue de règles uniformes.

2. Cryptage

Le centre de confiance leur transmet un programme de cryptage, qui repose sur les filtres de Bloom, mentionnés plus haut. Le programme permet aux administrateurs des différents centres de crypter leurs données sans devoir posséder des connaissances spécifiques. La clé du programme de cryptage est définie par les deux centres concernés sans que le centre de confiance n'en ait connaissance.

3. Appariement

Dans un dernier temps, le centre de confiance procède à l'appariement probabiliste des données. La première étape consiste à réunir toutes les entrées parfaitement concordantes. Ensuite, il établit, à l'aide des probabilités calculées pour chacun des paramètres, le meilleur appariement possible pour les entrées aux données manquantes ou légèrement divergentes. Il s'agit notamment de vérifier si le nom et le prénom ou le jour et le mois n'ont pas été interchangés lors de la saisie ou si d'autres erreurs fréquentes sont apparues. Les tableaux d'appariement et un rapport avec des informations complémentaires concernant la qualité de l'appariement sont remis aux différents centres, avant que le centre de confiance supprime toutes les données en sa possession.

Simulation

La qualité de la méthode P3RL a été testée dans le cadre d'une étude avec des données réelles et simulées. La simulation a montré que d'excellents résultats pouvaient être atteints lorsque l'appariement utilisait les noms, mais également si ces noms étaient cryptés. C'est cependant la préparation des données qui est décisive: seules les données corrigées de manière uniforme ont permis d'obtenir de bons résultats. Ainsi, l'appariement avec des noms cryptés a donné lieu à 18% de corrélations fausses si les données n'avaient pas été corrigées au préalable mais à seulement 0,7% d'erreur si les données avaient été corrigées.

Conclusion et perspective

La simulation a montré que la méthode P3RL se prête parfaitement à l'appariement de données de santé provenant de sources différentes, tout en respectant la protection des données. Même si la méthode P3RL associe un ensemble de solutions techniques complexes, elle peut être utilisée sur le terrain par des centres sans compétences techniques ni connaissances spécifiques en matière d'appariement. De plus, il est possible de tenir compte des particularités des différents centres lorsqu'il s'agit de corriger les données. En revanche, la méthode P3RL demande davantage de temps, de personnel et de moyens financiers que l'appariement de données non cryptées. Enfin, les commissions d'éthique concernées doivent évaluer à chaque projet si l'utilisation de la méthode P3RL répond à leurs exigences en matière de protection des données.

Maintenant que la méthode P3RL a été développée et testée avec succès, il est possible de l'appliquer à des projets concrets. Ce sera pour la première fois le cas lors d'une étude sur le risque de cancer des personnes atteintes du VIH en Suisse. Pour cette étude, les informations sur la population réunies par le réseau d'enregistrement du cancer (NICER) seront appariées à celles de l'étude suisse de cohorte VIH (SHCS). C'est la méthode P3RL qui a été choisie pour garantir la protection des données et obtenir un appariement de haute qualité. Le projet a été validé par la commission d'éthique du canton de Berne.

Correspondance:
Prof. Matthias Egger
ISPM
Université de Berne
Finkenhubelweg 11
CH-3012 Berne

Références

1 Stuck A, Moser A, Morf U, Wirz U, Wyser J, Gillmann G, et al. Effect of Health Risk Assessment and Counselling on Health Behaviour and Survival of Older People: Randomised Trial. PLoS Med. 12(10):e1001889. doi:10.1371/journal.pmed.1001889.

2 Spycher BD, Feller M, Zwahlen M, Roosli M, von der Weid NX, Hengartner H, et al. Childhood cancer and nuclear power plants in Switzerland: a census-based cohort study. Int J Epidemiol. 2011;40:1247–60. doi:10.1093/ije/dyr115.

3 Bech BH, Kjaersgaard MIS, Pedersen HS, Howards PP, Sørensen MJ, Olsen J, et al. Use of antiepileptic drugs during pregnancy and risk of spontaneous abortion and stillbirth: population based cohort study. BMJ. 2014;349:g5159.

4 RS 431.012.13 Ordonnance du DFI du 17 décembre 2013 concernant l'appariement des données statistiques (Ordonnance sur l'appariement de données). https://www.admin.ch/opc/fr/classified-compilation/20122208/index.html (téléchargée le 28 septembre 2015).

5 Bopp M, Spoerri A, Zwahlen M, Gutzwiller F, Paccaud F, Braun-Fahrländer C, et al. Cohort Profile: the Swiss National Cohort – a longitudinal study of 6.8 million people. Int J Epidemiol. 2009;38:379–84. doi:10.1093/ije/dyn042.

6 Schnell R, Bachteler T, Reiher J. Privacy-preserving record linkage using Bloom filters. BMC Med Inform Decis Mak. 2009;9:41. doi:10.1186/1472-6947-9-41.

7 Schmidlin K, Clough-Gorr KM, Spoerri A. Privacy preserving probabilistic record linkage (P3RL): a novel method for linking existing health-related data and maintaining participant confidentiality. BMC Med Res Methodol. 2015;15:46. doi:10.1186/s12874-015-0038-6.