Appariement entre un registre régional de pratiques en cardiologie interventionnelle et la base médico-administrative d'hospitalisation française : développement et validation d'un algorithme d'appariement déterministe

Emilie Lesaine, N. M. Belhamri, J. P. Legrand, Sandrine Domecq, P. Coste, A. Lacroix, Florence Saillour-Glenisson

Appariement entre un registre régional de pratiques en cardiologie interventionnelle et la base médico-administrative d'hospitalisation française : développement et validation d'un algorithme d'appariement déterministe

Linking Interventional Cardiology clinical registry data with French hospital administrative data: development and validation of deterministic record linkage

E. Lesaine[a,b,c]*; N-M. Belhamri[a]; J-P. Legrand[a,b,c]; S. Domecq[a,b,c]; P. Coste[d,e]; A. Lacroix[f]; F. Saillour-Glenisson[a,b,c] for the ACIRA investigators

[a] Univ. Bordeaux, ISPED, Centre INSERM U1219-Bordeaux Population Health, F-33000 Bordeaux, France
[b] CHU de Bordeaux, Pôle de santé publique, Service d'Information Médicale, F-33000 Bordeaux, France
[c] INSERM, ISPED, Centre INSERM U1219-Bordeaux Population Health, F-33000 Bordeaux, France
[d] CHU de Bordeaux Hôpital Cardiologique, Coronary Care Unit, F-33600 Pessac, France
[e] Univ. Bordeaux, Collège Sciences de la Santé, Cardiology Bordeaux, Aquitaine, F-33000 Bordeaux, France
[f] Agence Régionale de Santé Nouvelle-Aquitaine, Direction du pilotage de la stratégie et des parcours, F-33000 Bordeaux, France

*Auteur correspondant
Adresse e-mail : emilie.lesaine@u-bordeaux.fr **(**E. Lesaine)

Titre courant Algorithme d'appariement registre - PMSI

ABSTRACT

Background To recreate the in-hospital healthcare pathway for patients treated with coronary angiography or percutaneous coronary intervention, we linked the interventional cardiology registry (ACIRA) and the pseudonymized French hospital medical information system database (PMSI) in the Aquitaine region. The objective of this study was to develop and validate a deterministic merging algorithm between these exhaustive and complementary databases.

Methods After a pre-treatment phase of the databases to standardize the 11 identified linking variables, a deterministic linking algorithm was developed on ACIRA hospital stays between December 2011 and December 2014 in nine interventional cardiology centers as well as the data from the consolidated PMSI databases of the Aquitaine region from 2011 to 2014. Merging was carried out through 12 successive steps, the first consisting in strict linking of the 11 variables. The performance of the algorithm was analyzed in terms of sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV). Strategies complementary to the initial algorithm (change in the order of variables and base preprocessing) were tested. Comparative analysis of merged/unmerged patients explored potential causes of mismatch.

Results The algorithm found 97.2% of the 31,621 ACIRA stays to have sensitivity of 99.9% (95% CI [99.9; 99.9]), specificity of 97.9% (95% CI [97.7; 98.1]), PPV of 99.9% (95% CI [99.9; 99.9]) and NPV of 96.9% (95% CI [96.7; 97.1]). Complementary strategies did not yield better results. The unmerged patients were older, and hospitalized mostly in 2012 in two interventional cardiology centers.

Conclusion This study underscored the feasibility and validity of an indirect deterministic pairing to routinely link a registry of practices using hospital data to pseudonymized medico-administrative databases. This method, which can be extrapolated to other health events leading to hospitalization, renders it possible to effectively reconstruct patients' hospital healthcare pathway.

Keywords Registry; PMSI; Coronary angiography; Percutaneous coronary intervention;Deterministic linking; Validation of the linking method.

Résumé

Problématique Afin de reconstruire le parcours de soins hospitalier des patients ayant bénéficié d'un acte de coronarographie ou d'angioplastie coronaire, un appariement entre le registre de pratiques aquitain de cardiologie interventionnelle ACIRA et la base pseudonymisée du Programme de médicalisation du système d'information (PMSI) de la région ex-Aquitaine a été réalisé. L'objectif de cette étude était de développer et valider

un algorithme d'appariement déterministe entre ces bases de données exhaustives et complémentaires.

Méthodes Après une phase de prétraitement des bases pour standardiser les 11 variables d'appariement identifiées, un algorithme d'appariement déterministe a été développé sur les séjours ACIRA des patients hospitalisés entre décembre 2011 et décembre 2014 dans neuf centres de cardiologie interventionnelle et les données des bases PMSI consolidées de la région ex-Aquitaine des années 2011 à 2014. L'appariement a été effectué en 12 étapes successives, la première consistant en un appariement strict sur les 11 variables. Les performances de l'algorithme ont été analysées en termes de sensibilité, spécificité, valeur prédictive positive (VPP) et valeur prédictive négative (VPN) et testées en faisant varier l'algorithme initial selon des stratégies complémentaires (modification de l'ordre des variables et du prétraitement des bases). L'analyse comparative des patients appariés/non appariés a permis d'explorer les potentielles causes de non appariement.

Résultats L'algorithme a permis de retrouver au final 97,2 % des 31 621 séjours ACIRA avec une sensibilité de 99,9 % (IC 95% [99,9 ; 99,9]), une spécificité de 97,9 % (IC 95% [97,7 ; 98,1]), une VPP de 99,9 % (IC 95% [99,9 ; 99,9]) et une VPN de 96,9 % (IC 95% [96,7 ; 97,1]). Les stratégies complémentaires n'ont pas permis d'obtenir de meilleurs résultats. Les patients non appariés étaient plus âgés, avaient majoritairement été hospitalisés en 2012 et dans deux centres de cardiologie interventionnelle.

Conclusion Cette étude a montré la faisabilité et la validité d'un appariement indirect déterministe pour faire le lien en routine entre un registre de pratiques utilisant des données hospitalières et les bases médico-administratives pseudonymisées. Cette méthode, extrapolable à d'autres événements de santé donnant lieu à une hospitalisation, permet de façon efficace de reconstruire le parcours de soins hospitalier des patients.

Mots clés Registre ; PMSI ; Coronarographie ; Angioplastie coronaire ; Appariement déterministe ; Validation de la méthode d'appariement.

1. Introduction

Les maladies cardio-vasculaires représentaient en France en 2014 la deuxième cause de décès et un motif majeur de recours aux soins tout au long de parcours de soins complexes, incluant les hospitalisations à la phase aiguë, la réalisation d'actes de cardiologie invasive (CI - coronarographie ou angioplastie trans-luminale - ATL) et les réhospitalisations en services de médecine ou de soins de suite et de réadaptation (SSR) (1). Actuellement, les informations sur la qualité des prises en charge tout au long du parcours de soins hospitalier des patients ayant bénéficié d'un acte de CI restent parcellaires et fractionnées, par défaut d'accès à des données longitudinales les concernant. Par ailleurs, plusieurs études françaises ont montré des défauts de prise en charge des pathologies coronaires aiguës tout au long de la filière (2–4). Ces constats font émerger le besoin de gagner en connaissances sur la compréhension des parcours des patients victimes de pathologie coronaire, de leur variété, des inégalités d'accès aux soins et de leurs retentissements.

En 2010, l'Agence régionale de santé (ARS) ex-Aquitaine a décidé la mise en place du registre aquitain des actes de CI (ACIRA). Le registre de pratiques ACIRA a pour objectif de décrire les pratiques et les parcours de soins hospitaliers et post-hospitaliers des patients ayant bénéficié d'une coronarographie ou d'une angioplastie coronaire dans l'un des 11 centres de cardiologie interventionnelle (CCI) en région ex-Aquitaine, d'analyser les facteurs individuels (cliniques, sociodémographiques, économiques, géographiques), organisationnels et structurels associés à la variabilité des pratiques et des parcours de prise en charge (5). Afin d'éviter la double-saisie, de limiter les perdus de vue, d'améliorer la qualité des données et de diminuer les coûts, il a été décidé de fonder la collecte des données d'ACIRA sur des bases de données existantes, en particulier celles du Programme de médicalisation des systèmes d'information (PMSI) pour les données hospitalières de suivi.

Les registres de pratiques en cardiologie interventionnelle, tels ACIRA disposent de données cliniques précises pour décrire la prise en charge des patients avant et au décours de l'acte de CI. Le PMSI, qui couvre l'ensemble des hospitalisations dans les établissements de santé français, produit des informations médico-administratives exhaustives sur les séjours hospitaliers mais contient peu de données cliniques (6,7). Le rapprochement des données issues du registre ACIRA et des bases de données du PMSI donne ainsi accès à une analyse précise et unique du parcours de soins hospitalier des patients ayant bénéficié d'un acte de CI inclus dans le registre ACIRA.

Plusieurs cohortes de grande ampleur exploitent déjà les bases de données médico-administratives nationales, comme les enquêtes handicap-santé, les grandes cohortes épidémiologiques Constances, Elfe, le registre dijonnais des accidents vasculaires

cérébraux et le registre REIN (8–12). En parallèle, la stratégie "Intelligence Artificielle" française a créé en 2018 le "Health Data Hub" qui vise à mutualiser les ressources, enrichir les bases de données du Système national des données de santé (SNDS) avec des données cliniques de registres et à faciliter l'exploitation du SNDS pour favoriser les études, recherches ou évaluations présentant un caractère d'intérêt public « dans le respect de l'éthique et des droits fondamentaux des concitoyens » et en en accord avec les principes du Règlement général sur la protection des données (RGPD) (13). Dans la dynamique française actuelle de mutualisation des bases de données existantes, de nombreuses équipes s'interrogent sur une méthode d'appariement de qualité permettant d'enrichir leurs données à partir de celles des bases médico-administratives. La présentation détaillée d'une méthode fiable et adaptable à différents évènements de santé ayant donné lieu à une hospitalisation en France s'inscrit pleinement dans cette dynamique de partage d'expérience et de promotion de méthodes d'analyse des données médico-administratives.

Le NIR (Numéro d'inscription au répertoire national d'identification des personnes physiques) est un identifiant unique national pour chaque individu, présent dans les fichiers sociaux et médico-sociaux. L'utilisation du NIR étant très protégée par la loi, les possibilités d'appariement direct des individus par un identifiant unique, fiable et commun aux deux bases à apparier s'en trouvent limitées, ce qui a conduit les équipes de recherche à tester différentes méthodes d'appariement indirect, portant sur plusieurs variables communes aux deux bases (14,15). Les méthodes les plus fréquemment utilisées et dont la fiabilité a été prouvée à plusieurs reprises, sont les méthodes d'appariement indirects déterministe et probabiliste (16–21). De nouvelles approches basées sur l'apprentissage automatique ou « machine learning » (ML) sont plus récemment apparues (22). L'appariement déterministe consiste à rapprocher deux enregistrements d'un même individu à partir de variables d'appariement présentes dans les deux sources de données. Les enregistrements sont appariés si toutes les variables d'appariement sont strictement identiques. L'appariement déterministe est une bonne option lorsqu'il existe plusieurs variables d'appariement dont la qualité est haute : variables vérifiées, disponibles, avec un fort pouvoir discriminant (23). Ce pouvoir discriminant est favorisé par un nombre élevé de modalités de codage sur les variables d'appariement (24). L'appariement probabiliste repose sur l'évaluation de la probabilité que deux enregistrements (ou paires) correspondent au même patient au vu de leurs variables communes. Les paires dont la probabilité est supérieure à un certain seuil supérieur sont appariées, les paires dont la probabilité est inférieure à un seuil inférieur sont non-appariées et celles du milieu sont vérifiées manuellement, ce qui peut être extrêmement chronophage (24). Certaines études ont aussi mis en place des approches

probabilistes simplifiées (sans validation manuelle importante ou mesure de distance) ou combinées avec une méthode déterministe pour limiter les contraintes informatiques (24). Le Département d'information médicale du Centre hospitalier universitaire (CHU) de Dijon a développé un logiciel permettant de rendre des bases de données anonymes avant de les apparier au PMSI en utilisant un appariement probabiliste s'appuyant sur les données administratives des patients (21). Cette méthode a été utilisée avec succès pour faire le lien entre les bases du PMSI et des registres des cancers mais son adaptabilité aux autres bases de données semble complexe compte tenu de la nécessité d'avoir à disposition le logiciel ANONYMAT et des donnés identifiantes de patients pour l'appariement probabiliste (25,26). Le ML nécessite des compétences techniques très spécifiques pour la création de modèles algorithmiques complexes. Dans la plupart des cas, le ML nécessite l'utilisation au préalable de données dont le statut de couplage est connu afin de se former (22,27).

En l'absence du recueil du NIR dans ACIRA, un appariement direct avec le PMSI ex-Aquitaine était exclu pour réaliser le suivi des réhospitalisations en routine. Les bases de données ACIRA et du PMSI bénéficient de variables d'appariement discriminantes communes et de qualité. Le choix de la méthode d'appariement s'est ainsi porté sur un appariement déterministe utilisant des données non identifiantes de patients. Ce type d'appariement avec les bases de données du PMSI a été réalisé pour des registres des cancers, des accidents vasculaires cérébraux et plusieurs grandes cohortes mais il n'a, à notre connaissance, jamais été réalisé pour un registre de pratique sur la cardiologie. De par la présentation détaillée de la méthode utilisée, cette étude permet de répondre aux questions qui se posent sur la reproductibilité des algorithmes d'appariement quelle que soit la pathologie concernée. Pour répondre à notre objectif d'appariement du registre ACIRA à la base du PMSI, il a donc été nécessaire de développer l'algorithme *de novo*, de le valider afin de s'assurer de sa capacité à identifier les faux positifs et les faux négatifs, de ses performances en termes de sensibilité, de spécificité, du potentiel discriminant des variables d'appariement choisies, puis d'identifier des éléments d'amélioration pour en optimiser la performance. L'objectif de cet article est de présenter de façon détaillée le travail de développement et de validation de cet algorithme d'appariement déterministe entre les bases de données du registre de pratiques ACIRA et du PMSI en région ex-Aquitaine.

2. Méthodes

2.1. Sources de données

Le registre ACIRA se caractérise par un recueil continu, multicentrique, prospectif et exhaustif des données médicales nominatives des patients âgés de plus de 18 ans résidant en France métropolitaine pris en charge pour des actes de coronarographie et d'ATL réalisés dans les 11 CCI d'ex-Aquitaine depuis le 1er décembre 2011. Le registre ACIRA recueille pour chaque CCI, des données issues du dossier patient (informations socio-démographiques, géographiques, données cliniques, bio marqueurs cardiaques), du logiciel de cardiologie interventionnelle (indications de l'acte, techniques interventionnelles utilisées), ainsi que des variables PMSI extraites du système d'information du CCI par une requête automatique, dénommées variables « PMSI CCI » dans le reste du manuscrit (5). Les variables « PMSI CCI », codées par le Département d'information médicale du CCI, sont utilisées pour contrôler l'exhaustivité des actes codés dans le logiciel de cardiologie interventionnelle et comme variable d'appariement aux bases PMSI régionales. L'exhaustivité des actes est calculée *a posteriori* à partir de la base PMSI régionale accessible au niveau de l'ARS. Les données de suivi pendant un an (mortalité, réhospitalisations) sont issues des bases de données médio-administratives françaises.

Le PMSI est un recueil pérenne et exhaustif d'informations pseudonymisées, quantifiées et standardisées permettant de décrire et de valoriser l'activité des établissements. Les recueils PMSI des quatre champs (MCO, SSR, hospitalisation à domicile - HAD et psychiatrie) s'appliquent respectivement aux établissements publics, privés à but lucratif et non lucratif. Les bases PMSI contiennent les diagnostics médicaux, les actes réalisés au cours des séjours et les données de mouvements du patient : identification des établissements de santé, durée de séjour, mode, mois et année d'entrée et de sortie, durée des hospitalisations, délais entre les séjours, décès hospitalier. Les bases de données du PMSI sont pseudonymisées au niveau des établissements de santé par un algorithme de hachage qui transforme de façon irréversible le numéro d'identification du patient tout en permettant d'effectuer un suivi des différentes hospitalisations, car pour un patient donné, le même numéro anonyme est obtenu tout au long de ses séjours hospitaliers. Jusqu'en 2015, les ARS centralisaient, homogénéisaient et chainaient les données PMSI adressées par chaque établissement de santé de la région, aux données hospitalières pré-existantes dans la base PMSI régionale.

2.2. Etapes de construction et validation de l'algorithme d'appariement entre le registre ACIRA et le PMSI régional

L'algorithme d'appariement a été développé et validé sur les données des bases PMSI consolidées de la région ex-Aquitaine des années 2011 à 2014 et les séjours ACIRA de patients ayant bénéficié d'un acte de CI entre décembre 2011 et décembre 2014 dans

neuf CCI, deux CCI ne nous ayant pas transmis les données nécessaires au moment de la réalisation de l'appariement. L'activité de ces deux CCI représentait 22 % des actes réalisés en ex-Aquitaine. Le développement et la validation de l'algorithme d'appariement déterministe ont suivi les étapes suivantes.

Etape I : Choix des variables d'appariement

Afin de pouvoir apparier les données ACIRA extraites des CCI avec celles issues de la base PMSI régionale, 11 variables PMSI communes aux deux bases de données (variables « PMSI CCI » issues d'ACIRA et variables issues de la base PMSI régionale, dénommées variables « PMSI région ») ont été identifiées au préalable grâce à un travail préliminaire mené avec l'ARS ex-Aquitaine (Tableau 1). La première étape a consisté à identifier *a priori* les variables communes aux deux bases les plus spécifiques d'un séjour donné et présentant un nombre important de modalités différentes. La capacité de ces 11 variables à lier de façon unique un séjour ACIRA à un séjour PMSI a ensuite été testée avec succès sur un échantillon de 2477 séjours ACIRA inclus dans cinq CCI en 2012 (99,1 % des séjours appariés).

Etape II : Prétraitement pour standardiser les variables d'appariement

Certaines des 11 variables « PMSI CCI » issues d'ACIRA ont dû être corrigées à partir d'un programme automatique de façon à ce que leur format et leur type correspondent à ceux des variables « PMSI région ». Ces corrections ont été les suivantes : harmonisation du mode de calcul des durées de séjours avec celui utilisé dans la base PMSI régionale, homogénéisation du format des diagnostics principaux et diagnostics reliés, retraitement des résumés d'unité médicale (RUM) en résumé de sortie standardisé (RSS), correction des codes postaux en codes géographiques PMSI pour le lieu de résidence. La base du PMSI régional comporte des "variables contrôles" créées lors de la procédure de pseudonymisation, concernant la présence et la cohérence de certaines variables directement identifiantes, comme le numéro de sécurité sociale, le sexe, la date de naissance, le numéro administratif du séjour et des "variables contrôles" créées lors de la génération des fichiers anonymes (fusion ANO-HOSP/HOSP-PMSI, fusion ANO-PMSI/Fichier PMSI). Si le codage d'une des "variables contrôles" correspondait à une erreur, les données des séjours correspondant étaient supprimées.

Etape III : Algorithme initial d'appariement déterministe sur les séjours.

L'appariement déterministe a été réalisé en 12 étapes successives. Le tableau 1 présente les 12 étapes d'appariement avec pour chaque étape, les combinaisons de variables utilisées. La première étape consistait en un appariement déterministe sur l'ensemble des 11 variables d'appariement identifiées en phase de test alors que les 11 autres étapes avaient pour principe de lier les séjours sur une partie des 11 variables ou après modifications de certaines d'entre elles, modifications tenant compte de corrections apportées lors de l'incorporation des données « PMSI CCI » dans la base PMSI régionale. Le travail préliminaire réalisé sur les variables PMSI a aussi guidé le choix des variables à retirer de l'algorithme lors des différentes étape et ce, selon les deux principes suivants : 1) les variables étaient retirées par ordre d'étapes établi en fonction de la proportion de données manquantes ou incohérentes mesurées au cours du travail préliminaire (les variables étaient retirées d'autant plus tôt qu'elles présentaient une proportion importante de données manquantes ou incohérentes) ; 2) les suppressions ou modifications de plusieurs variables ont été regroupées dans les étapes finales. De fortes proportions de données manquantes sur la variable « nombre de DAS » (3717 données manquantes) et d'incohérences sur les variables « diagnostic relié (DR) » (108 différences), « groupage des RSS en Groupe homogène de malades (GHM RSS - 233 différences) » et « durée de séjour » (1974 différences) ont été identifiées dans la base ACIRA lors du travail préliminaire. La variable « nombre de DAS » a ainsi été enlevée des étapes 2, 4, 6, 8, 10 et 11, le « DR » des étapes 9, 10 et 11 et le « GHM RSS » de l'étape 12. Pour la « durée de séjour », l'écart d'un jour constaté entre les variables « PMSI CCI » et « PMSI région » correspondait aux séjours sans nuitées pouvant être corrigés à "0 jour" dans le PMSI régional (étapes 3, 4 et 11). Dans les étapes 5 et 6, le code géographique du lieu de résidence ACIRA a été recherché parmi l'ensemble des codes géographiques disponibles dans la base PMSI régionale pour un patient donné, afin de prendre en compte les déménagements survenus entre le moment de l'extraction des données « PMSI CCI » et le gel de la base PMSI régionale pour cette étude. Dans les étapes 7 et 8, la variable d'appariement "code géographique de résidence" a été remplacée par le "département de résidence" lorsque le code géographique était codé "XX999" dans la base PMSI régionale, où "XX" représente le numéro de département.

Que ce soit au niveau des bases du PMSI régional ou du registre ACIRA, tous les séjours d'un même patient sont liés entre eux par les traits d'identification du patient. Un patient était considéré comme apparié dès qu'un des séjours de ce

patient avait été apparié à une étape donnée, son séjour était alors supprimé des deux bases de données pour les étapes suivantes. Cependant, une étape de contrôle automatique était réalisée afin de ne pas associer à tort des patients pour les cas de sur-appariement : patient de la base du PMSI régional apparié par l'algorithme à plusieurs patients dans la base ACIRA ou inversement. Les autres séjours ACIRA d'un même patient étaient contrôlés pour vérifier que l'appariement était réalisé pour le même patient dans la base PMSI régionale.

Etape IV : Stratégies complémentaires d'appariement déterministe

Pour optimiser et valider l'algorithme initial développé à l'étape 3, huit stratégies alternatives d'appariement déterministe ont été testées, faisant varier différents paramètres de l'algorithme alternativement et de façon combinée : ordre des 12 étapes, modalités de prétraitement des bases de données et de la sélection de la population d'étude (type d'acte et âge des patients) (Tableau2) :

- Modification de l'ordre des 12 étapes (test de deux ordres d'étapes supplémentaires, stratégies 1 - ordre 1 et 2 - ordre 2) : La stratégie 1 appliquait le même principe global d'ordonnancement des étapes que la stratégie initiale et n'apportait que des modifications mineures d'ordre des étapes (2, 3, 4, 5 d'un côté et 8, 9, 10, 11, 12 de l'autre). La stratégie 2 a renversé le principe d'ordonnancement des étapes de la stratégie initiale en positionnant l'étape 11, comportant de nombreuses modifications sur les variables d'appariement, juste après l'étape d'appariement sur l'ensemble des 11 variables ;
- Action sur le travail préparatoire des bases de données :
    - Stratégie 3 : pas de prétraitement des données ACIRA ;
    - Stratégie 4 : pas de suppression des séjours selon la valeur des variables de contrôle ;
- Modification de la population d'étude :
    - Stratégie 5 : sélection des actes selon leur type (coronarographie ou ATL) ;
    - Stratégie 6 : pas de sélection des patients de plus de 17 ans ;
    - Stratégie 7 : pas de suppression des séjours selon la valeur des variables de contrôle et sélection des actes selon leur type (coronarographie ou ATL) ;
    - Stratégie 8 : sélection des actes selon leur type (coronarographie ou ATL) et pas de sélection des patients âgés de plus de 17 ans.

Etape V : Calcul des indices de performance

L'évaluation de la qualité de l'appariement est essentielle afin d'identifier des sources potentielles de biais qui pourront affecter les résultats des études basées sur ces bases de données (28).

Le registre ACIRA a été considéré comme la base de référence dans le calcul des performances de l'algorithme (29–31). Pour cette étude, les indicateurs de performance ont été redéfinis dans un contexte de rapprochement de bases de données :

- **les vrais positifs (VP)** correspondaient au nombre de séjours ACIRA associés de façon unique avec un séjour dans la base PMSI régionale (relation 1-1) ;
- **les vrais négatifs (VN)** correspondaient au nombre de séjours ACIRA non associés à un séjour dans la base PMSI régionale (relation 1-0) ;
- **les faux positifs (FP)** ou "collision" correspondaient aux séjours différents dans la base ACIRA associés à un même séjour dans la base PMSI régionale (relation n-1) ;
- **les faux négatifs (FN)** ou "doublon" correspondaient aux séjours ACIRA identiques associés à des séjours différents dans la base PMSI régionale (relation 1-n) (32).

Les indices de performance (sensibilité, spécificité, valeur prédictive positive et valeur prédictive négative) ont été calculés à l'issue de chaque stratégie d'appariement. Les séjours classés en vrais négatifs, faux positifs ou faux négatifs étaient considérés comme "non appariés". Des intervalles de confiance ont été calculés pour chacune des estimations.

Etape VI : Identification des causes de non appariement

Les patients appariés et non appariés ont été décrits et comparés (tests de Chi-deux et de Student) selon leurs caractéristiques socio-démographiques (âge, sexe des patients) et des critères techniques (année de la base PMSI régionale utilisée pour l'appariement et CCI de prise en charge qui a communiqué les variables d'appariement). Une analyse descriptive spécifique des **"**variables contrôles**"** a été réalisée à partir des patients non appariés à la stratégie initiale mais appariés à la stratégie complémentaire 4. Les données du mois de décembre 2011 ont été incluses avec l'année 2012 et celles de 2014 avec l'année 2013 pour les analyses.

2.3. Considérations éthiques et règlementaires

Le registre ACIRA a reçu les autorisations nécessaires au traitement des données de santé nominatives et à l'appariement avec les données du PMSI régional : avis favorable du Comité consultatif sur le traitement de l'information en matière de recherche dans le domaine de la santé (CCTIRS) du 10 février 2011 "n°11-060" ; autorisation de la Commission nationale de l'informatique et des libertés (CNIL) "décision DR-2011-436" du 27 octobre 2011 ; avenant spécifique pour le recueil des données de suivi à partir de la base PMSI régionale "décision DR-2012-619" du 17 décembre 2012. Les patients inclus dans ACIRA ont été individuellement informés incluant le lien avec la base PMSI régionale.

Le traitement des données et les analyses statistiques ont été réalisées avec le logiciel SAS 9.3.

3. Résultats

3.1. Méthode initiale d'appariement déterministe

Entre décembre 2011 et décembre 2014, 31 621 séjours correspondant à 26 618 patients pris en charge dans neuf CCI en région ex-Aquitaine ont été inclus dans la base de données ACIRA. A l'issue de la méthode initiale d'appariement comprenant les 12 étapes, 30 739 séjours ACIRA (97,2 %) ont été appariés aux séjours de la base PMSI régionale correspondant à 25 773 patients ACIRA (96,8 %) appariés (Tableau 3). Au cours de la première étape d'appariement, 72,4 % des séjours ont pu être appariés. Le nombre de patients et séjours appariés et les indices de performance pour chaque étape sont disponibles en matériel supplémentaire. Les indices de performance pour les appariements sur les séjours étaient les suivants : sensibilité de 99,9 % (IC 95% [99,9 ; 99,9]), spécificité de 97,9 % (IC 95% [97,7 ; 98,1]), valeur prédictive positive de 99,9 % (IC 95% [99,9 ; 99,9]) et valeur prédictive négative de 96,9 % (IC 95% [96,7 ; 97,1]). Sur les 882 séjours non appariés, 18 collisions et 27 doublons ont été constatés.

3.2. Stratégies complémentaires testées

Les stratégies 1, 2 et 6 retrouvaient des résultats de performance et un nombre de patients non appariés sensiblement identiques à ceux de la stratégie initiale (Tableau 3). A l'exception de 48 d'entre eux, la quasi-totalité des patients ont été appariés au même numéro d'étape, et ce quel que soit l'ordre des étapes. Parmi ces 48 patients, seuls 13 ont été appariés un autre patient dans la base PMSI régionale, et ce uniquement pour la stratégie 2. Les stratégies 3, 4 et 7 diminuaient le nombre de patients non appariés ainsi que les indices de performance. Les stratégies 5 et 8 amélioraient la performance de l'appariement mais augmentaient le nombre de patients non appariés. Au final, les

stratégies complémentaires explorées n'ont pas permis de conclure à de meilleurs résultats que ceux de la stratégie d'appariement initiale.

3.3. Causes de non-appariement

Les patients non appariés étaient en moyenne plus âgés (68,9 ans) que les patients appariés (67,7 ans ; p=0,006) (Tableau 4). Le genre des patients ne différait pas entre les groupes de patients appariés et non appariés. La proportion de patients non appariés était plus importante en 2012 qu'en 2013 (p<0,001). Un effet centre a été également constaté avec des proportions de patients non appariés plus importantes dans les centres I et F (10,2 et 10,9 % respectivement), alors que les sept autres centres avaient au plus 2,4 % de patients non appariés (p<0,001).

Les "variables contrôles" des patients non appariés à la stratégie initiale mais appariés à la stratégie complémentaire 4 ont été décrites pour 317 séjours. A noter que 89 séjours avaient été appariés à la stratégie initiale mais non appariés à la stratégie complémentaire 4, principalement du fait de collisions. Pour ces 317 séjours, à l'exception d'un séjour pour lequel toutes les variables contrôles étaient manquantes, les "variables contrôles" différentes de "0" concernaient le numéro de sécurité sociale. Ce numéro n'était pas renseigné pour 277 séjours et la longueur était incorrecte pour 33 séjours.

4. Discussion

4.1. Synthèse des résultats

Notre étude a montré, et la faisabilité d'un appariement entre un registre de pratiques et des bases médico-administratives françaises utilisant des variables indirectement identifiantes, et ses très bons résultats aussi bien en termes de proportion de séjours ACIRA appariés à la base PMSI que de performance. Ainsi, la méthode d'appariement déterministe entre le registre de pratiques ACIRA et la base du PMSI régional, appliquée sur les actes de coronarographie et d'angioplastie coronaire de neuf CCI entre décembre 2011 et décembre 2014, a permis, au terme de 12 étapes, de retrouver 97,2 % des 31 621 séjours ACIRA avec une sensibilité de 99,9 %, une spécificité de 97,9 %, une valeur prédictive positive de 99,9 % et une valeur prédictive négative de 96,9 %. Sur les 882 séjours non appariés, 18 collisions et 27 doublons ont été constatés. La première des 12 étapes, consistant en un appariement strict sur les 11 variables d'appariement, retrouvait 72,4 % des séjours.

Cette méthode d'appariement entre un registre de pratiques et la base médico-administratives du PMSI permet en routine de réaliser le suivi des patients et de reconstruire l'ensemble du parcours de soins hospitalier des patients ayant bénéficié d'un

acte de CI. Cet appariement est possible à grande échelle, sans nécessité de recueillir les variables directement identifiantes des patients, tout en limitant le nombre de perdus de vue et réduisant les coûts de fonctionnement (28). Nous pensons que cette méthode est adaptable à d'autres événements de santé à condition que ces événements aient donné lieu à un séjour en hospitalisation conventionnelle, afin de recueillir les variables hospitalières d'appariement au niveau de chaque établissement de santé. Il n'est pas nécessaire que ces évènements aient donné lieu à un codage d'acte au cours du séjour puisque l'appariement ne s'appuie pas sur le codage CCAM. Cette méthode a pu être utilisée sur des séjours pour accident vasculaire cérébral en 2013 avec un taux d'appariement de 98,% à la base PMSI régionale ex-Aquitaine, ce qui renforce sa validité externe. En revanche, il appartient à chaque équipe de s'assurer de la bonne performance de l'appariement, même avec les conditions pré-requises.

4.2. Comparaison aux données de la littérature

A notre connaissance, seules trois équipes françaises ont publié leurs méthodes d'appariement avec les bases de données médico-administratives pseudonymisées Françaises (10,21,33,34). Dans l'étude sur le premier certificat de naissance, les appariements déterministes utilisaient les données PMSI pseudonymisées issues directement des établissements de santé. Ces données comportaient des variables d'appariement non présentes dans la base PMSI régionale, en particulier des variables relatives à l'accouchement et aux dates d'entrée et de sortie du séjour. Les appariements déterministes retrouvaient 92,5 % des nouveau-nés et 85,5 % des séjours PMSI. L'appariement probabiliste améliorait ces proportions. L'étude sur les parcours de soins pré-dialyse des insuffisants rénaux chroniques terminaux utilisait un appariement déterministe avec les bases du Système national des données de santé (SNDS) et retrouvait 90,2 % des patients. La méthode d'appariement probabiliste utilisant les données identifiantes des patients et développée par le CHU de Dijon avait une sensibilité de 97 % et une spécificité de 93 %, mais ces résultats de performance publiés en 1998 sont probablement maintenant supérieurs grâce à l'amélioration de la qualité des données du PMSI (21). La comparaison de nos résultats avec des études internationales est rendue difficile par la nécessité d'adapter les variables et méthodes d'appariement aux bases médico-administratives des pays concernés. Néanmoins, notre méthode d'appariement a obtenu des résultats proches voir meilleurs que plusieurs études internationales, principalement américaines. Entre 88 et 97 % des patients inclus dans des registres sur les pathologies cardio-neuro-vasculaires ou rhumatologiques ont pu être appariés aux bases de données médico-administratives par un appariement déterministe (19,35–41). La littérature concernant les performances de l'appariement probabiliste est

plus riche, en particulier aux Etats-Unis, en Grande-Bretagne et en Nouvelle-Zélande. Une revue de la littérature internationale menée en 2007 sur la qualité de l'appariement probabiliste retrouvait des spécificités entre 99 et 100 %, des sensibilités entre 74 et 98 % et des VPP entre 68 et 99 % (30). Cette étude mettait en avant la difficulté de trouver un gold-standard, l'importance de la qualité et du nombre des variables d'appariement.

4.3. Forces et limites de la méthode d'appariement déterministe utilisée

Afin de valider la méthode d'appariement et d'essayer d'améliorer ses performances, huit stratégies complémentaires ont été explorées mais aucune n'a permis d'améliorer la méthode initiale. L'ordre des étapes et la sélection sur l'âge des patients n'ont eu de retentissement ni sur le nombre de patients appariés ni sur les indices de performance (stratégies 1, 2 et 6). Le choix d'une stratégie d'appariement avec suppression des patients appariés entre les étapes présentait le risque potentiel d'apparier un même patient ACIRA à un mauvais patient de la base PMSI régionale, ceci du fait de l'absence ou de modifications de certaines variables d'appariement lors des différentes étapes. Ce risque, s'il s'était confirmé, aurait dû être sensible à l'ordre des étapes. Or, notre analyse de sensibilité nous a montré que la modification de l'ordre des étapes, n'a modifié qu'à la marge le nombre de patients appariés, ainsi que les performances globales de l'appariement. De plus, dans les deux stratégies modifiant l'ordre des étapes, seuls 13 patients ACIRA ont été appariés à des patients différents dans la base PMSI régionale. Nous en concluons donc que les erreurs d'appariement ont été minimes. Les stratégies 3 et 4 d'exclusion de l'étape de prétraitement et de non prise en compte du niveau d'exactitude des données d'identification dans le PMSI régional ont présenté l'avantage d'augmenter le nombre de patients appariés mais ont eu plusieurs effets négatifs sur les performances de l'algorithme : augmentation du nombre de faux négatifs (82 doublons en plus) ayant entraîné une baisse de la VPN et de la spécificité. Ces résultats montrent en creux tout l'intérêt de ces phases de prétraitement et de prise en compte des variables contrôles qui tout en étant entièrement automatiques, donc non chronophage, permettent de diminuer le nombre de doublons et donc d'améliorer les performances. La sélection sur le type d'acte ne modifiait pas le nombre de doublons et de collisions mais réduisait le nombre de séjours appariés (stratégie 5). Nous avons choisi au final la stratégie qui offrait le meilleur compromis entre le nombre de séjours/patients appariés et la performance globale.

Les 12 étapes successives ont notamment permis de prendre en compte les éventuelles données manquantes sur les variables d'appariement (28). Deux méthodes ont été utilisées concernant celles de la variable « nombre de DAS » dans ACIRA. La première a été de combiner plusieurs variables d'appariement. Ainsi, la diminution potentielle

d'appariement entre séjours du fait des données manquantes sur cette variable était compensée par le potentiel d'appariement des 10 autres variables. La seconde a consisté à supprimer cette variable de l'algorithme du fait des incohérences qu'elle générait dans les étapes 2, 4, 6, 8, 10 et 11.

L'analyse de sensibilité réalisée par l'intermédiaire du test de huit stratégies alternatives a permis de montrer la robustesse de l'algorithme développé et le bon pouvoir discriminant des variables d'appariement utilisées. Cependant, toutes les stratégies alternatives n'ont pu être explorées, en particulier celles utilisant d'autres variables d'appariement, ce qui laisse envisager des perspectives d'améliorations. Les extractions PMSI ont été demandées aux CCI plus de deux ans après l'acte, ce qui peut expliquer certaines discordances entre les données « PMSI CCI » et « PMSI région », en particulier le code géographique de résidence. Par ailleurs, les informaticiens des CCI ont pu extraire certaines des variables demandées aussi bien à partir du logiciel administratif que du PMSI de leur établissement, ce qui pourrait expliquer certaines des divergences constatées. Une vigilance accrue a été apportée à la provenance des variables PMSI d'appariement issues des CCI. L'augmentation de la proportion de séjours appariés entre 2012 et 2014 laisse espérer une amélioration de la qualité du remplissage des données PMSI par les CCI au cours du temps. L'identification de performances moindres dans deux CCI ont permis de mettre en place des actions ciblées afin d'améliorer la qualité des variables d'appariement et des données PMSI. Les patients non appariés étaient plus âgés ce qui laisse supposer des séjours plus complexes donnant lieu à davantage de diagnostics, de RUM, de séjours, sources de "collisions" et d'erreurs potentielles. Une autre hypothèse concerne la qualité des données qui est plus susceptible d'être moins bonne chez les personnes âgées, en particulier les ayant-droits. Ces hypothèses restent cependant à vérifier pour mieux identifier les causes de non appariement et améliorer l'algorithme.

Le registre ACIRA a été considéré comme la base de référence dans le calcul des performances de l'algorithme, or des doublons, des erreurs d'enregistrement ne sont pas à exclure. Un retour aux dossiers sur un échantillon de séjours, en particulier pour les doublons et les collisions permettrait de mieux comprendre nos résultats, en particulier pour les patients âgés, et apporter ainsi des pistes d'amélioration pour l'algorithme.

5. Perspectives

Le rapprochement entre le registre de pratiques ACIRA et les bases du PMSI de la région ex-Aquitaine va permettre de reconstruire le parcours de soins hospitalier des patients ayant bénéficié d'un acte de coronarographie ou d'angioplastie coronaire et de mener des analyses approfondies sur les facteurs associés à ces réhospitalisations, ceci afin d'aider

à structurer les filières de prise en charge, déterminer la part des réhospitalisations évitables, cibler les situations et populations à risque pour faciliter la recherche d'actions d'amélioration.

La méthode déterministe a été choisie pour sa simplicité de mise en œuvre, son coût moindre en ressources informatiques et humaines ainsi que pour son applicabilité, compte tenu de la présence de nombreuses variables d'appariement de qualité communes aux deux bases (peu de données manquantes et d'erreurs) et présentant de nombreuses modalités de codage différentes (24). D'autres méthodes d'appariement auraient pu être utilisées. Nous n'avons pas choisi la méthode probabiliste qu'elle soit classique, simplifiée ou combinée, pour des raisons de complexité de mise en œuvre et de cout en ressources informatiques et humaines. La méthode de ML aurait également pu être utilisée mais la constitution d'un gold-standard est chronophage et en pratique difficile à réaliser, en particulier pour les grandes bases de données médico-administratives (42). Cette dernière approche est tout de même en cours de test pour l'appariement d'ACIRA aux bases de données nationales de mortalité.

Au vu de sa simplicité, de son faible coût de mise en œuvre, et de ses très bonnes performances, nous avons souhaité présenter en détail la méthode utilisée et les résultats obtenus pour permettre à d'autres équipes de terrain de la tester, et aux chercheurs d'y trouver les informations potentiellement utiles dans une démarche de création d'un outil générique d'appariement déterministe et ce, quelle que soit la pathologie.

L'algorithme est en cours de déploiement sur les données PMSI nationales 2012-2019, accessibles sur la plateforme sécurisée de l'ATIH (Agence technique de l'information sur l'hospitalisation). Cet algorithme sera adapté pour l'analyse de l'ensemble du parcours de soins aussi bien hospitalier qu'ambulatoire des patients via un accès aux bases médico-administratives du SNDS, en cours de demande. Le périmètre national du SNDS permettra l'appariement aux séjours hospitaliers hors région ex-Aquitaine. Le SNDS comporte les données de l'Assurance maladie, les données hospitalières du PMSI, les causes médicales de décès, les données relatives au handicap et un échantillon de données en provenance des organismes d'Assurance maladie complémentaire, ces deux dernières sources de données devant être prochainement ajoutées (43). Depuis avril 2017, l'accès aux données du SNDS est devenu possible, sur autorisation de la CNIL. Dans le but de garantir la confidentialité des données de santé, les données sont pseudonymisées, chaque individu se voit attribuer un code spécifique auquel est rattaché l'ensemble des données du SNDS le concernant. Le rapprochement entre le registre ACIRA, disposant de données directement identifiantes, et les bases du SNDS est envisagé selon un appariement indirect à partir des données d'état civil des patients

(nom, prénom, date et lieu de naissance) par l'intermédiaire d'un organisme tiers qui gère le système national de gestion des identités.

Dans la dynamique actuelle française de mutualisation des bases de données pour favoriser la recherche, à l'instar de la mise en place du "Health data hub", plusieurs équipes vont être amenées à utiliser des méthodes d'appariement déterministe ou probabiliste pour enrichir leurs données. Le traitement de ces données est encadré par le RGPD et un chapitre spécifique de la loi informatique et libertés, afin de garantir la bonne réutilisation de ces données dans le respect de l'obligation d'information des patients et de sécurité des données.

6. Conclusion

Cette étude a montré la faisabilité et les très bonnes performances d'un appariement déterministe accessible, reproductible et nécessitant des ressources humaines et informatiques limitées pour faire le lien en routine entre un registre de pratiques utilisant des données hospitalières et les bases médico-administratives pseudonymisées. Cette méthode, qui peut être adaptée à d'autres événements de santé donnant lieu à une hospitalisation, va permettre aux chercheurs de tirer profit des avantages de ces deux types de bases de données pour une analyse des parcours de soins : la finesse des données cliniques des registres de pratiques ainsi que la puissance et la richesse des données médico-administratives. Ces méthodes sont actuellement essentielles pour utiliser tout le potentiel des bases de données existantes et optimiser les prises en charge des patients.

Remerciements

Le registre ACIRA remercie tous les patients qui participent au registre, l'Agence régionale de santé Nouvelle-Aquitaine, financeur du registre, ainsi que tous les investigateurs : Dr C. Abadie, Clinique de St Augustin, Bordeaux ; Dr P. Ancelin, CH de Mont de Marsan, Mont de Marsan ; Dr S. Boutreux, CH de Périgueux, Périgueux ; Dr V. Buhaj, CH de Périgueux, Périgueux ; Dr F. Casteigt, Polyclinique Bordeaux Nord Aquitaine, Bordeaux ; Dr JM. Clerc, CH de Périgueux, Périgueux ; Pr P. Coste, CHU de Bordeaux, Bordeaux ; Dr D. Crenn, CH de Libourne, Libourne ; Dr N. Delarche, CH de Pau, Pau ; Dr V. Gilleron, CHU de Bordeaux, Bordeaux ; Dr A. Hassan, CH de Mont de Marsan, Mont de Marsan ; Dr B. Karsenty, Hôpital privé St Martin, Pessac ; Dr G. Laplace, Clinique cardiologique d'Aressy, Aressy ; Dr B. Larnaudie, Clinique de Caudéran les pins-francs, Bordeaux ; Dr JL. Leymarie, Clinique de St Augustin, Bordeaux ; Dr N. Marque, Clinique cardiologique d'Aressy, Aressy ; Dr F. Perret, Hôpital privé St Martin, Pessac ; Dr JM. Perron, CH de Libourne, Libourne ; Dr S. Debeugny, CH de Pau, Pau ;

Dr C. Robin, Clinique cardiologique d'Aressy, Aressy ; Dr MP. Benetier, ARS Nouvelle Aquitaine, Bordeaux ; Dr I. Jamet, ARS Nouvelle Aquitaine, Bordeaux.

Références

1. Hospitalisations pour motif cardio-vasculaire: motifs de recours en court séjour, mortalité à un an et comparaison avec les causes initiales de décès. DRESS, Paris; mars 2017. https://drees.solidarites-sante.gouv.fr/IMG/pdf/dd12.pdf.
2. Gabet A, De Peretti C, Iliou M-C, Nicolau J, Olié V. National trends in admission for cardiac rehabilitation after a myocardial infarction in France from 2010 to 2014. Arch Cardiovasc Dis 2018;111(11):625‑33.
3. Puymirat E, Simon T, Cayla G, Cottin Y, Elbaz M, Coste P, et al. Acute Myocardial Infarction: Changes in Patient Characteristics, Management, and 6-Month Outcomes Over a Period of 20 Years in the FAST-MI Program (French Registry of Acute ST-Elevation or Non-ST-Elevation Myocardial Infarction) 1995 to 2015. Circulation 14 nov 2017;136(20):1908‑19.
4. Bataille S, Loyeau A, Mapouata M. Indicateurs de pratiques 2003-2014 [Internet]. Service des registres de l'ARS Île-de-France e-MUST et CARDIO-ARSIF; 2016 juin [cité 9 sept 2019]. Disponible sur: http://www.cardio-arsif.org/Downloads/INDICATEURS_DE_PRATIQUE_2014.pdf
5. Lesaine E, Saillour-Glenisson F, Leymarie J-L, Jamet I, Fernandez L, Perez C, et al. The ACIRA registry: a regional tool to improve the healthcare pathway for patients undergoing percutaneous coronary interventions and coronary angiographies in the French Aquitaine region - Study design and first results. Crit Pathw Cardiol. 16 sept 2019;
6. Moulis G, Lapeyre-Mestre M, Palmaro A, Pugnet G, Montastruc J-L, Sailler L. French health insurance databases: What interest for medical research? Rev Med Interne. juin 2015;36(6):411‑7.
7. Boudemaghe T, Belhadj I. Data Resource Profile: The French National Uniform Hospital Discharge Data Set Database (PMSI). Int J Epidemiol. 1 avr 2017;46(2):392‑392d.
8. Zins M, Goldberg M. La cohorte Constances. 2011. http://www.constances.fr/medias/base-documentaire/2014/1415268206-protocole-scientifique.pdf.
9. Salines G, De Launay C. Les cohortes : intérêt, rôle et position de l'InVS.2010.http://opac.invs.sante.fr/doc_num.php?explnum_id=495.
10. Raffray M, Pladys A, F.Gao, Couchoud C, Vigneau C, Bayat S. Etude du parcours de soins pré-dialyse des insuffisants rénaux chroniques terminaux ayant démarré la dialyse en urgence. Rev DÉpidémiologie Santé Publique. mars 2019;67:S87.

11. Montaut A, Calvet L, Bouvier G, Gonzalez L. L'appariement handicap-santé et données de l'assurance maladie. Paris: Direction de la recherche, des études, de l'évaluation et des statistiques; 2013 janv p. 35. Report No.: 39.
12. Lainay C, Benzenine E, Durier J, Daubail B, Giroud M, Quantin C, et al. Hospitalization within the first year after stroke: the Dijon stroke registry. Stroke. janv 2015;46(1):190‑6.
13. Mission de préfiguration du « Health Data Hub » [Internet]. Paris, France; 2018 oct. Disponible sur: https://solidarites-sante.gouv.fr/IMG/pdf/181012_-_rapport_health_data_hub.pdf
14. Bounebache SK, Quantin C, Benzenine E, Obazinski G, Rey G. Revue Bibliographique des Méthodes de Couplage des Bases de Données : Applications et Perspectives dans le Cas des Données de Santé Publique. Journal de la Société Française de Statistique, Vol. 159 No. 3. 2018;
15. Perlbarg J, Allonier C, Boisnault P, Daniel F, Fur PL, Szidon P, et al. Faisabilité et intérêt de l'appariement de données individuelles en médecine générale et de données de remboursement appliqué au diabète et à l'hypertension artérielle. Sante Publique (Bucur). juill 2014;Vol. 26(3):355‑63.
16. Guesdon M, Benzenine E, Gadouche K, Quantin C. Securizing data linkage in french public statistics. BMC Med Inform Decis Mak [Internet]. oct 2016 [cité 14 mars 2019];16. Disponible sur: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5053094/
17. Jaro MA. Probabilistic linkage of large public health data files. Stat Med. mars 1995;14(5‑7):491‑8.
18. Tromp M, Ravelli AC, Bonsel GJ, Hasman A, Reitsma JB. Results from simulated data sets: probabilistic record linkage outperforms deterministic record linkage. J Clin Epidemiol. mai 2011;64(5):565‑72.
19. Setoguchi S, Zhu Y, Jalbert JJ, Williams LA, Chen C-Y. Validity of deterministic record linkage using multiple indirect personal identifiers: linking a large registry to claims data. Circ Cardiovasc Qual Outcomes. mai 2014;7(3):475‑80.
20. Fellegi IP, Sunter AB. A Theory for Record Linkage. J Am Stat Assoc. 1 déc 1969;64(328):1183‑210.
21. Quantin C, Bouzelat H, Allaert FA, Benhamiche AM, Faivre R, Dusserre L. Automatic record hash coding and linkage for epidemiological follow-up data confidentiality. 1998;(37):271‑7.
22. Goldstein H, Harron K, Cortina-Borja M. A scaling approach to record linkage. Stat Med. 20 juill 2017;36(16):2514‑21.

23. Zhu Y, Matsuyama Y, Ohashi Y, Setoguchi S. When to conduct probabilistic linkage vs. deterministic linkage? A simulation study. J Biomed Inform. août 2015;56:80‑6.
24. Doidge JC, Harron K. Demystifying probabilistic linkage: Common myths and misconceptions. Int J Popul Data Sci [Internet]. [cité 14 janv 2021];3(1). Disponible sur: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6281162/
25. Quantin C, Benzenine E, Hägi M, Auverlot B, Abrahamowicz M, Cottenet J, et al. Estimation of national colorectal-cancer incidence using claims databases. J Cancer Epidemiol. 2012;2012:298369.
26. Quantin C, Benzenine E, Fassa M, Hagi M, Fournier E, Gentil J, et al. Evaluation of the interest of using discharge abstract databases to estimate breast cancer incidence in two French departments. Stat J IAOS. 2012;28:73‑85.
27. Hejblum BP, Weber GM, Liao KP, Palmer NP, Churchill S, Shadick NA, et al. Probabilistic record linkage of de-identified research datasets with discrepancies using diagnosis codes. Sci Data. janv 2019;6:180298.
28. Harron KL, Doidge JC, Knight HE, Gilbert RE, Goldstein H, Cromwell DA, et al. A guide to evaluating linkage quality for the analysis of linked data. Int J Epidemiol. 01 2017;46(5):1699‑710.
29. Guesdon M, Quantin C, Benzenine E. Appariement de données pseudonymisées. In p. 20. Disponible sur: http://www.jms-insee.fr/2015/S09_4_ACTE_GUESDON_JMS2015.PDF
30. Silveira DP da, Artmann E. Accuracy of probabilistic record linkage applied to health databases: systematic review. Rev Saude Publica. oct 2009;43(5):875‑82.
31. Newgard CD. Validation of probabilistic linkage to match de-identified ambulance records to a state trauma registry. Acad Emerg Med Off J Soc Acad Emerg Med. janv 2006;13(1):69‑75.
32. Goldberg M, Quantin C, Guegen A, Zins A. Bases de données médico-administratives et épidémiologie : intérêts et limites. 2012;
33. Lebreton E, Vincelet C, Chatignoux E, Menguy C, Crenn Hebert C, Février Y-M, et al. [Record linkage of hospital discharge data and first health certificates: a test in the Val d'Oise]. Rev Epidemiol Sante Publique. août 2014;62(4):257‑66.
34. Roussot A, Benzenine E, Cottenet J, Lannelongue C, Giroud M, Quantin C. Patients fibrinolysés en bourgogne : identification et caractéristiques. J Gest Déconomie Médicales. 2013;31(7):487‑97.
35. Mao J, Etkin CD, Lewallen DG, Sedrakyan A. Creation and Validation of Linkage Between Orthopedic Registry and Administrative Data Using Indirect Identifiers. J Arthroplasty. 2 févr 2019;

36. Pasquali SK, Jacobs JP, Shook GJ, O'Brien SM, Hall M, Jacobs ML, et al. Linking clinical registry data with administrative data using indirect identifiers: implementation and validation in the congenital heart surgery population. Am Heart J. déc 2010;160(6):1099‑104.
37. Jacobs JP, Edwards FH, Shahian DM, Haan CK, Puskas JD, Morales DLS, et al. Successful linking of the Society of Thoracic Surgeons adult cardiac surgery database to Centers for Medicare and Medicaid Services Medicare data. Ann Thorac Surg. oct 2010;90(4):1150‑6; discussion 1156-1157.
38. Curtis JR, Chen L, Bharat A, Delzell E, Greenberg JD, Harrold L, et al. Linkage of a de-identified United States rheumatoid arthritis registry with administrative data to facilitate comparative effectiveness research. Arthritis Care Res. déc 2014;66(12):1790‑8.
39. Hammill BG, Hernandez AF, Peterson ED, Fonarow GC, Schulman KA, Curtis LH. Linking inpatient clinical registry data to Medicare claims data using indirect identifiers. Am Heart J. juin 2009;157(6):995‑1000.
40. Grannis SJ, Overhage JM, McDonald CJ. Analysis of identifier performance using a deterministic linkage algorithm. Proc AMIA Symp. 2002;305‑9.
41. Kim TJ, Lee JS, Kim J-W, Oh MS, Mo H, Lee C-H, et al. Building Linked Big Data for Stroke in Korea: Linkage of Stroke Registry and National Health Insurance Claims Data. J Korean Med Sci. 31 déc 2018;33(53):e343.
42. Pita R, Mendonça E, Reis S. A Machine Learning Trainable Model to Assess the Accuracy of Probabilistic Record Linkage. Big Data Anal Knowl Discov. 2017;214‑27.
43. Tuppin P, Rudant J, Constantinou P, Gastaldi-Ménager C, Rachas A, de Roquefeuil L, et al. Value of a national administrative database to guide public decisions: From the système national d'information interrégimes de l'Assurance Maladie (SNIIRAM) to the système national des données de santé (SNDS) in France. Rev Epidemiol Sante Publique. oct 2017;65 Suppl 4:S149‑67.

Tableau 1

Variables d'appariement utilisées lors des douze étapes d'appariement déterministe sur les séjours, entre les bases du registre ACIRA et du PMSI régional

	Etapes											
	1	2	3	4	5	6	7	8	9	10	11	12
FINESS	x	x	x	x	x	x	x	x	x	x	x	x
Age	x	x	x	x	x	x	x	x	x	x	x	x
Sexe	x	x	x	x	x	x	x	x	x	x	x	x
Année de sortie de l'établissement	x	x	x	x	x	x	x	x	x	x	x	x
Mois de sortie de l'établissement	x	x	x	x	x	x	x	x	x	x	x	x
Code géographique du lieu de résidence	x	x	x	x	Dem	Dem	Dpt	Dpt	x	x	x	x
GHM RSS	x	x	x	x	x	x	x	x	x	x	x	
Diagnostic principal du séjour	x	x	x	x	x	x	x	x	x	x	x	x
Nombre de diagnostics associés	x		x		x		x		x			x
Diagnostic relié	x	x	x	x	x	x	x	x				x
Durée de séjour	x	x	-1 j	-1 j	x	x	x	x	x	x	-1 j	x

-1j : prise en compte des corrections sur la "durée de séjour" dans la base PMSI régionale ; Dem : prise en compte des déménagements des patients entre l'extraction ACIRA et le gel de la base PMSI régionale ; Dpt : remplacement par la variable "département de résidence" ; GHM : groupe homogène de malades ; RSS : résumé standardisé de sortie.

Tableau 2

Stratégies alternatives à la méthode initiale d'appariement indirect.

	Stratégies								
	Initiale	1	2	3	4	5	6	7	8
Réalisation de l'appariement en 12 étapes									
Ordre initial (ordre 0)[a]	oui			oui	oui	oui	oui	oui	oui
Changement de l'ordre (ordre1)[b]		oui							
Changement de l'ordre (ordre 2)[c]			oui						
Travail préparatoire des bases de données									
Pré-traitement des données ACIRA	oui	oui	oui	non	oui	oui	oui	oui	oui
Prise en compte du niveau d'exactitude des données d'identification dans le PMSI régional[d]	oui	oui	oui	oui	non	oui	oui	non	oui
Sélection des actes d'angioplastie et de coronarographie	non	non	non	non	non	oui	non	oui	oui
Sélection sur l'âge des patients > 17 ans	oui	oui	oui	oui	oui	oui	non	oui	non

[a] : Ordre 0 : étape 1 → étape 2 → étape 3 → étape 4 → étape 5→ étape 6 → étape 7 → étape 8 → étape 9 → étape 10 → étape 11 → étape 12 ;
[b] : Ordre 1 : étape 1 → étape 3 → étape 5 → étape 2 → étape 4→ étape 6 → étape 7 → étape 9 → étape 12 → étape 8 → étape 10 → étape 11 ;
[c] : Ordre 2 : étape 1 → étape 11 → étape 10 → étape 8 → étape 12→ étape 9 → étape 7 → étape 6 → étape 4 → étape 2 → étape 5 → étape 3.
[d] : suppression des séjours dès qu'une variable de contrôle égale à 0 vs aucune suppression
DAS : diagnostic associé ; DR : diagnostic relié ; GHM : groupe homogène de malades ; PMSI : programme de médicalisation des systèmes d'information ; RSS : résumé standardisé de sortie.

Tableau 3

Nombre de patients/séjours non appariés et indices de performance sur les différentes stratégies d'appariement testées sur les séjours

	Nombre de patients non appariés (n=26 618)	Nombre de séjours non appariés (n=31 621)	Indices de performance							
			Spécificité		Sensibilité		Valeur prédictive positive		Valeur prédictive négative	
			%	IC 95 %	%	IC 95 %	%	IC 95 %	%	IC 95 %
Stratégie initiale	845	882	97,9	97,7 - 98,1	99,9	99,9 - 99,9	99,9	99,9 - 100,0	96,9	96,7 - 97,1
Stratégie 1	845	882	97,9	97,7 - 98,1	99,9	99,9 - 99,9	99,9	99,9 - 100,0	96,9	96,7 - 97,1
Stratégie 2	849	886	97,3	97,1 - 97,5	99,9	99,9 - 99,9	99,9	99,9 - 100,0	96,9	96,7 - 97,1
Stratégie 3	606	691	96,9	96,7 - 97,1	99,6	99,6 - 99,7	99,9	99,9 - 100,0	83,8	83,4 - 84,2
Stratégie 4	575	654	96,7	96,5 - 96,9	99,6	99,6 - 99,7	99,9	99,9 - 100,0	82,9	82,4 - 83,3
Stratégie 5	1014	1058	98,3	98,1 - 98,4	99,9	99,9 - 99,9	99,9	99,9 - 100,0	97,4	97,2 - 97,6
Stratégie 6	845	882	97,9	97,7 - 98,1	99,9	99,9 - 99,9	99,9	99,9 - 100,0	96,9	96,7 - 97,1
Stratégie 7	742	826	97,5	97,3 - 97,7	99,7	99,6 - 99,7	99,9	99,9 - 100,0	86,8	86,4 - 87,1
Stratégie 8	1014	1058	98,3	98,1 - 98,4	99,9	99,9 - 99,9	99,9	99,9 - 100,0	97,4	97,2 - 97,6

Patients/séjours non appariés: vrais négatifs, faux positifs et faux négatifs.

Tableau 4

Répartition des patients selon le résultat de l'appariement initial, en fonction de leurs caractéristiques socio-démographiques, de l'année de réalisation de l'acte et du centre de cardiologie interventionnelle (n=26 618).

	Patients appariés[a] (n=25 773)			Patients non appariés[b] (n=845)		
	n	%	IC 95 %	n	%	IC 95 %
Sexe masculin	17 970	69,7	69,2 - 70,3	580	68,6	65,5 - 71,8
Années						
2011-2012	11 989	46,5	45,9 - 47,1	484	57,3	53,9 - 60,6
2013-2014	13 784	53,5	52,9 - 54,1	361	42,7	39,4 - 46,1
Centres de cardiologie interventionnelle						
A	2 141	98,8	98,3 - 99,3	26	1,2	0,7 - 1,7
B	3 441	98,9	98,6 - 99,2	40	1,1	0,8 - 1,4
C	2 745	98,5	98,0 - 99,0	41	1,5	1,0 - 2,0
D	2 583	98,0	97,5 - 98,5	53	2,0	1,5 - 2,5
E	6 347	99,2	99,0 - 99,4	49	0,8	0,6 - 1,0
F	2 225	89,1	87,9 - 90,3	272	10,9	9,7 - 12,1
G	1 657	97,6	96,9 - 98,3	41	2,4	1,7 - 3,1
H	1 919	99,2	98,8 - 99,6	16	0,8	0,4 - 1,2
I	2 715	89,8	88,7 - 90,9	307	10,2	9,1 - 11,3

[a] : Patients appariés: vrais positifs; [b] Patients non appariés: vrais négatifs, faux positifs et faux négatifs.

IC: intervalle de confiance.

Tableau supplémentaire

Tableau 5

Nombre de patients/séjours non appariés et indices de performance sur les différentes étapes d'appariement testées sur les séjours

	Nombre de patients non appariés (n=26 618)	Nombre de séjours non appariés (n=31 621)	Indices de performance							
			Spécificité		Sensibilité		Valeur prédictive positive		Valeur prédictive négative	
			%	IC 95 %	%	IC 95 %	%	IC 95 %	%	IC 95 %
Etape 1	7739	8732	99,9	99,8 - 99,9	99,9	99,9 - 100,0	99,9	99,9 - 100,0	99,8	99,7 - 99,8
Etape 2	5610	6193	99,8	99,8 - 99,9	99,9	99,9 - 99,9	100,0	99,9 - 100,0	99,7	99,6 - 99,7
Etape 3	3844	4209	99,6	99,5 - 99,6	99,9	99,9 - 99,9	99,9	99,9 - 100,0	99,4	99,3 - 99,5
Etape 4	1126	1186	98,4	98,3 - 98,6	99,9	99,9 - 99,9	99,9	99,9 - 100,0	97,7	97,5 - 97,9
Etape 5	1091	1146	98,4	98,3 - 98,5	99,9	99,9 - 99,9	99,9	99,9 - 100,0	97,6	97,4 - 97,8
Etape 6	1083	1137	98,4	98,2 - 98,5	99,9	99,9 - 99,9	99,9	99,9 - 100,0	97,6	97,4 - 97,8
Etape 7	1076	1122	98,4	98,2 - 98,5	99,9	99,9 - 99,9	99,9	99,9 - 100,0	97,6	97,4 - 97,7
Etape 8	1076	1122	98,4	98,2 - 98,5	99,9	99,9 - 99,9	99,9	99,9 - 100,0	97,6	97,4 - 97,7
Etape 9	1074	1120	98,4	98,2 - 98,5	99,9	99,9 - 99,9	99,9	99,9 - 100,0	97,5	97,4 - 97,7
Etape 10	1001	1047	98,2	98,1 - 98,4	99,9	99,9 - 99,9	99,9	99,9 - 100,0	97,4	97,2 - 97,6
Etape 11	985	1031	98,2	98,1 - 98,4	99,9	99,9 - 99,9	99,9	99,9 - 100,0	97,3	97,2 - 97,5
Etape 12	845	882	97,9	97,7 - 98,1	99,9	99,9 - 99,9	99,9	99,9 - 100,0	96,9	96,7 - 97,1

Patients/séjours non appariés: vrais négatifs, faux positifs et faux négatifs.