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

Julie Perlbarg,
et al.

S.F.S.P. | *Santé Publique*

2014/3 - Vol. 26
pages 355 à 363

Article disponible en ligne à l'adresse:
ISSN 0995-3914
---
http://www.cairn.info/revue-sante-publique-2014-3-page-355.htm
---

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

Feasibility and practical value of statistical matching of a general practice database and a health insurance database applied to diabetes and hypertension

Julie Perlbarg[1, 2], Caroline Allonier[1], Philippe Boisnault[2, 3], Fabien Daniel[1], Philippe Le Fur[1, 2], Philippe Szidon[2, 3], Yann Bourgueil[1, 2]

Résumé

Objectif : Le monde de la Santé Publique en France est soucieux d'améliorer l'utilisation des bases de données nationales. L'objectif du projet était de construire un outil de recherche en soins ambulatoires en appariant des données médicales et des données de remboursement.

Méthodes : Les sources étaient la base de données du SNIIRAM et la base de données de l'Observatoire de Médecine Générale (OMG). Le SNIIRAM est une base médico-administrative nationale regroupant les données ayant servi au remboursement des soins et l'OMG est une base de données médicales en soins ambulatoires fournissant les motifs de recours aux soins appelés Résultats de Consultation (RC). À partir des données des patients ayant consulté un des 30 médecins généralistes sélectionnés en 2008, nous avons réalisé un appariement probabiliste des deux bases.

Résultats : La procédure d'appariement a permis d'apparier 89 211 séances et 29 088 patients. La comparaison des Affections de Longue Durée (ALD) et des RC a montré que 94 % des patients en ALD diabète avaient un RC diabète dans l'année. Mais seulement 65 % des patients avec un RC diabète étaient déclarés en ALD. L'appariement a permis d'identifier 12 % de patients diabétiques sans traitement antidiabétique et hors ALD qui n'étaient pas repérables dans le SNIIRAM.

Conclusion : Cette étude a décrit une méthodologie novatrice d'appariement de bases de données. Elle a également montré les apports de ce modèle de données appariées pour le ciblage de populations à risque. D'autres pistes d'exploitation sur l'analyse des comorbidités, des pratiques et des parcours de soins sont encore envisageables.

Mots-clés : Bases de données ; Systèmes d'information ; Appariement probabiliste ; Assurance maladie ; Médecine générale ; Résultats de consultation ; Épidémiologie.

Summary

Objectives: *Public Health actors in France are striving to improve the use of national databases for public health and research. The main objective of this project was to develop a research tool in ambulatory care by matching medical data and reimbursement data.*

Methods: *Data sources were the health insurance database (SNIIRAM) and the General Practice Observatory (OMG) database. The SNIIRAM is a national medical and administrative database comprising data used in healthcare reimbursement. The OMG is a medical database on ambulatory care recording presenting complaints called "Results of Consultation" (RC). Based on data for patients who consulted one of the 30 general practitioners selected in 2008, we performed a probabilistic matching of the two databases.*

Results: *The linkage procedure allowed matching of 89,211 consultations or doctor visits and 29,088 patients. Comparison of long-term diseases (ALD) and RC showed that 94% of patients with diabetes as ALD had at least one RC coded as diabetes during the year, but only 65% of patients with one RC coded as diabetes were reported as ALD for this disease. Matching of the databases identified 12% of diabetic patients without antidiabetic treatment and without ALD for this affection; these patients were therefore not identifiable in the SNIIRAM database.*

Conclusion: *This study describes an innovative database matching methodology. It also illustrates the contribution of this model of matched data in terms of targeting populations at risk. Other approaches to analysis of comorbidities, medical practices and care pathways could be proposed.*

Keywords: *Databases; Information systems; Probabilistic linkage; Health insurance; General practice; Results of consultation; Epidemiology.*

[1] IRDES (Institut de Recherche et Documentation en Économie de la Santé) – 10, rue Vauvenargues – 75018 Paris – France.
[2] Prospere (Partenariat pluridisciplinaire de recherche sur l'organisation des soins de premiers secours).
[3] SFMG (Société française de Médecine générale).

Correspondance : J. Perlbarg
julieperlbarg@gmail.com

Réception : 03/03/2014 – *Acceptation :* 28/05/2014

Introduction

Les bases de données médico-administratives nationales exhaustives sont actuellement au cœur d'interrogations et de réflexions quant à leur enrichissement et l'amélioration de leur exploitation à des fins de connaissance et de pilotage. Dans le domaine de la santé, la France est l'un des rares pays au monde à avoir investi dans un système d'information individuelle exhaustif de données médico-administratives et sociales centralisées, gérées par les organismes publics. Le Système national d'information interrégimes de l'assurance maladie (SNIIRAM) rassemble désormais l'ensemble des informations ayant donné lieu à un remboursement par l'Assurance maladie. Pour chaque assuré et pour chaque professionnel, les actes de ville, les prescriptions, les séjours hospitaliers (depuis 2009 avec le chaînage du Programme de médicalisation des systèmes d'information (PMSI) et du SNIIRAM), les indemnités journalières et les autres prestations ayant donné lieu à remboursement, sont ainsi identifiés. Le dernier rapport charges et produits 2014 publié par la Caisse nationale de l'assurance maladie des travailleurs salariés (CNAMTS) [1] témoigne de l'intérêt de ces informations à des fins d'analyse et de régulation du système de santé.

L'enrichissement des données de remboursements avec des données individuelles socio-économiques a été initié dès 1988 dans le cadre de l'enquête santé protection sociale [2] et étendu depuis aux enquêtes Handicap-santé [3] ainsi qu'à la cohorte Constance. Il permet d'augmenter considérablement les dimensions d'analyses notamment pour ce qui concerne les inégalités de recours aux soins et les effets des politiques publiques en termes de régulation. De même, l'apport d'informations individuelles sur les situations identifiées par les médecins et les décisions qu'ils prennent, paraît constituer une voie très féconde, ayant fait l'objet de plusieurs travaux prospectifs [4, 5], pour étudier les pratiques comme les parcours de soins à des fins de connaissance et d'évaluation aussi bien en termes d'efficacité que d'efficience.

Il faut cependant avoir à l'esprit que l'utilisation des bases de données nationales à des fins de recherche ou de surveillance est encore très contraignante en France et doit faire face à de nombreux obstacles, tant au niveau juridique que technique et organisationnel [6]. Les démarches d'utilisation des bases de données nationales, surtout dans un cadre de recherche impliquant de passer par un identifiant de données, sont compliquées. Le recueil du NIR (numéro d'inscription au répertoire, identifiant individuel unique utilisé par toutes les bases de données nationales) est impossible en l'absence de décret en Conseil d'État spécifique à chaque étude. De plus, l'utilisation de ces bases de données médico-administratives à des fins de recherche n'est pas une finalité des institutions détentrices, celles-ci ne bénéficient donc d'aucun moyen financier alloué à cette activité.

Certains pays ont cependant déjà réussi à mettre au service de la santé publique et de la recherche leurs systèmes d'informations médico-sociales, notamment les pays scandinaves et le Canada [7], en créant des *Population data centers*, largement accessibles aux équipes de recherche et dont les connexions entre sources de données sont croissantes. Ainsi, le *Clinical Practice ResearchDatalink* [8] en Angleterre relie désormais pour les cinq millions de personnes inscrites auprès des 650 *practices* participantes les dossiers médicaux et les actes réalisés par les médecins généralistes avec les informations des séjours hospitaliers, les données de mortalité ainsi que les données populationnelles issues des recensements. Cette ouverture très large des bases de données au service de la recherche doit cependant être particulièrement bien encadrée pour éviter toute dérive ou problème de ré-identification et de diffusion de données individuelles.

Les acteurs de la santé publique en France sont à l'heure actuelle très soucieux d'améliorer l'exploitation de ces bases de données nationales pour la santé publique et la recherche, comme le montre le rapport du Haut Conseil de la santé publique paru en mars 2012 [9], tout en assurant le bon usage de leur utilisation. Le rapport de l'Inspection générale des affaires sociales rendu par Pierre-Louis Bras en septembre 2013 fait un état des lieux de l'utilisation actuelle des bases médico-administratives et propose des principes pour sa gouvernance, son évolution et son bon usage [10]. Quelques projets de grande ampleur s'appuient déjà largement sur des appariements de bases de données nationales, comme les grandes cohortes épidémiologiques Constances et Elfe ainsi que les programmes de suivi post-professionnels Spirale et Espri. Des enquêtes en population, telle que l'enquête Handicap-santé de l'INSEE [3], ont également récemment testé la faisabilité d'un appariement avec la base de données du SNIIRAM par le NIR.

Il existe deux principaux types d'appariement concernant les bases de données médico-administratives [11]. Il est possible d'apparier ces bases de données entre elles, à l'image du CépiDc (Centre d'épidémiologie sur les causes médicales de décès) qui apparie les causes individuelles de décès aux données socioprofessionnelles de la Caisse nationale d'assurance vieillesse. Une autre possibilité est d'apparier des enquêtes, dont des suivis de cohortes,

avec ces bases de données afin d'enrichir les données individuelles.

Plusieurs méthodes d'appariement sont connues. Une des plus fréquemment utilisées et dont la fiabilité a été prouvée à plusieurs reprises est la méthode d'appariement probabiliste [12-14].

L'objectif de ce projet spécifique porté par l'équipe de recherche émergente PROSPERE était de tester la faisabilité d'un outil de recherche en soins ambulatoires en appariant des données médicales issues des médecins généralistes et des données de remboursement issues du SNIIRAM. Cet outil devrait permettre, dans le cadre du projet général de l'équipe portant sur la performance des organisations de soins primaires, de tester, d'une part des méthodes et des indicateurs d'analyse de l'organisation des soins de premier recours, et d'autre part de constituer un échantillon-témoin pour des travaux de recherche évaluative sur l'organisation des soins de premier recours.

Méthodes

Populations sources

Le travail a consisté à élaborer un nouveau modèle de données alimenté par les données de consommation inter-régime (DCIR) issues du SNIIRAM d'une part et les données issues de l'Observatoire de médecine générale (OMG) d'autre part. L'OMG a été construit et géré par la Société française de médecine générale (SFMG) et constitue à ce jour la seule base de données de médecine générale non commerciale [15]. Sur les 800 médecins généralistes appartenant à la SFMG, environ 150 étaient équipés d'un logiciel adapté et étaient en mesure de collecter et transmettre des données pour l'OMG. Le test a porté sur un échantillon de 30 médecins de l'OMG sur l'année 2008 sélectionnés pour leur ancienneté, la qualité et l'exhaustivité du codage des dossiers médicaux. Ainsi, les données concernant l'ensemble des patients ayant consulté au moins une fois dans l'année un des 30 médecins généralistes ont été récupérées.

Une des clés de l'exhaustivité du recueil d'informations dans l'OMG était le renseignement en temps réel et en continu au cours de la séance, le médecin saisissant les informations, structurées à l'aide de thésaurus pour le diagnostic et la thérapeutique médicamenteuse, utiles à sa pratique. Aucune information supplémentaire spécifique à l'OMG n'avait à être recueillie, la base de données pouvant être assimilée à un dossier médical informatisé. Les données recueillies dans l'OMG étaient organisées en domaines. Tout d'abord, le domaine « patient » contenait des données anonymes liées à l'individu, telles que la date de naissance, le sexe, le département de résidence et le type d'assurance si disponible. Ensuite, le domaine « séance » comportait les données caractérisant les contextes de prise en charge des patients tels que la date et le type de séance (consultation ou visite). Il donnait une vision synchronique du suivi des patients. Le domaine des « données médicales individuelles » comportait des éléments médicaux simples tels que le poids, la taille ou la tension artérielle. Le domaine « diagnostic » contenait les résultats de consultation, organisés en symptômes, syndromes et diagnostics, selon le dictionnaire élaboré par la SFMG et transcodables en CIM 10 (classification internationale des maladies, dixième version), correspondant à tout ce qui avait été réellement pris en charge par le médecin au cours de la séance dans le contexte d'incertitude propre à l'exercice de la médecine générale. Enfin, le domaine des « décisions » comportait les prescriptions médicamenteuses, de biologie, d'imagerie, les arrêts de travail et les adressages à d'autres praticiens. Seules les données des domaines « patient », « séance » et « diagnostic » ont été utilisées dans la phase test de 2008.

Le DCIR (SNIIRAM) regroupe des données de production, c'est-à-dire l'ensemble des données ayant servi au remboursement des patients. Ainsi, l'ensemble des prestations et leurs coûts associés sont renseignés. Ces données proviennent de sources diverses telles que les professionnels de santé, les pharmacies et les établissements de santé. Les données d'hospitalisation à partir du PMSI sont intégrées au SNIIRAM mais non disponibles pour notre test d'appariement. Seules les données du SNIIRAM concernant les bénéficiaires du régime général d'Assurance Maladie ont été récupérées pour la phase-test.

Appariement des données et construction de la base finale

La construction de ce nouveau modèle de données reposait sur le même modèle d'origine que l'OMG. Ainsi, le premier niveau d'observation était la séance (consultation ou visite) et c'est autour de ce domaine que venaient s'articuler les autres tables du SNIIRAM. La construction de cette base est passée par plusieurs étapes :

- la Société française de médecine générale, gestionnaire de l'OMG, fournissait la liste des numéros ADELI des 30 médecins généralistes investigateurs de l'OMG et volontaires pour participer au test ;

- l'ensemble des actes exécutés par chaque médecin généraliste était extrait à partir du numéro ADELI dans la base de données du SNIIRAM ;
- à partir des actes extraits, il était possible de récupérer la liste de l'ensemble des bénéficiaires et donc les files actives des 30 médecins. Un identifiant unique était alors attribué à chaque patient ;
- à partir de cette liste, il était possible d'extraire l'ensemble des prestations effectuées pour chaque bénéficiaire sur l'année 2008.

Après cette étape, étaient constituées d'une part, une base de données de l'OMG organisée au niveau séance regroupant les informations sur les séances et les patients de 30 médecins généralistes en 2008, et d'autre part, une base de données provenant du SNIIRAM contenant l'ensemble des prestations des bénéficiaires du régime général de l'Assurance Maladie ayant consulté au moins une fois un de ces 30 médecins au cours de l'année 2008.

Un préalable indispensable à tout appariement de bases de données est l'identification des sujets. Deux méthodes sont possibles. La première est celle de l'appariement direct des sujets repérés par un identifiant unique commun aux deux bases. L'OMG ayant été constitué dans un objectif d'amélioration des connaissances sur les pratiques en médecine générale, l'accès au NIR n'a pas été demandé lors de la constitution de cette base de données. Cette première méthode n'a donc pu être appliquée. La deuxième méthode est celle de l'appariement probabiliste que nous avons mis en œuvre dans ce projet. Cette méthode nécessitant la présence de variables discriminantes, la difficulté était d'en prendre suffisamment pour éviter les erreurs d'appariement, mais en nombre limité pour éviter les échecs d'appariement [16]. Les six variables suivantes ont été choisies : numéro ADELI du médecin, date de séance, type de séance (consultation ou visite), mois de naissance, année de naissance et genre du bénéficiaire. C'est la confrontation d'informations sur les dates et natures des contacts ainsi que les informations liées à l'identité du patient qui a permis de rapprocher les deux bases. Ce test d'appariement a fait l'objet d'une autorisation de la Commission nationale de l'informatique et des libertés (CNIL) et d'un avis favorable de l'Institut des données de santé (IDS). Techniquement, la SFMG a transmis à l'Institut de recherche et de documentation en économie de la santé (IRDES) par porteur un fichier crypté au format txt et les données SNIIRAM extraites par la CNAMTS (Caisse nationale de l'assurance maladie des travailleurs salariés) en format csv ont été récupérées par l'Irdes *via* une plateforme sécurisée. La procédure d'appariement a été réalisée avec le logiciel SAS 9.0.

Analyse de la base appariée : repérage des pathologies chroniques

Afin de tester l'exploitabilité et l'intérêt des données appariées, nous nous sommes intéressés, dans le prolongement d'une recherche précédente sur les seules données OMG [17], à l'apport de ces deux bases pour l'identification des patients diabétiques et des patients hypertendus. Nous avons dans un premier temps comparé les populations identifiées par l'existence d'un code affection de longue durée (ALD) diabète ou hypertension artérielle (HTA) provenant du SNIIRAM avec les populations identifiées par les résultats de consultation (RC) diabète ou HTA provenant de l'OMG. Ceci nous a permis de répondre à la question suivante dans l'exemple du diabète : « Est-ce qu'un patient déclaré en ALD « diabète de type I ou diabète de type II » a bien eu au moins une fois dans l'année un RC codé en « diabète de type I » ou en « diabète de type II » ? Puis, nous nous sommes intéressés à son corollaire : « Est-ce qu'un patient ayant eu un RC codé en « diabète de type I » ou en « diabète de type II » est déclaré en ALD « diabète de type I ou diabète de type II » ? Le même travail a été réalisé pour l'hypertension artérielle. Cependant, l'ALD « hypertension artérielle sévère » a été retirée par le Décret n° 2011-726 du 24 juin 2011. Pour notre étude concernant l'année 2008, l'ALD était toujours effective, mais ce test n'aura pas vocation à être repris sur les années suivantes en ce qui concerne cette pathologie. Nous avons ensuite renouvelé l'analyse en associant les critères des traitements remboursés dans la base SNIIRAM pour identifier les patients diabétiques (un traitement par antidiabétique oral ou par insuline remboursé au moins une fois dans l'année) et les patients hypertendus (deux traitements antihypertenseurs de classes différentes remboursés au moins une fois dans l'année). Une liste des médicaments antidiabétiques et antihypertenseurs utilisés dans les algorythmes de décision est tenue à la disposition des lecteurs.

Nous avons ensuite confronté les deux méthodes de repérage des patients diabétiques et des patients hypertendus dans le SNIIRAM et dans l'OMG afin de quantifier l'apport de la base de données appariées.

Résultats

Faisabilité de l'appariement

La base de données de l'OMG pour l'appariement était constituée de 37 992 patients représentant 126 793 séances.

La base de données du SNIIRAM était constituée de 35 730 patients représentant 117 509 séances. Cet écart était en grande partie lié à la non prise en compte des régimes autres (Mutualité sociale agricole, Régime social des indépendants, Sections locales mutualistes) que le régime général dans la base du SNIIRAM. La création d'un identifiant d'accrochage au niveau séance a permis d'identifier les séances uniques côté OMG et côté SNIIRAM. Environ 75 % des séances uniques s'appariaient d'emblée. Une séance appariée permettait d'identifier le patient correspondant dans les deux bases. Cette seule condition a été retenue comme assez discriminante pour considérer l'individu comme identifié. Cet individu était donc retiré, ainsi que ses séances, qu'elles soient appariées ou non. Dans le nouveau groupe de séances restantes, nous pouvions à nouveau isoler des séances uniques. En effet, ces séances avaient pu être présentes en faux doublons lors du premier tour d'appariement (par erreur de date par exemple), le retrait des patients considérés comme appariés a permis de retirer toutes les séances de ce patient, y compris celles contenant des erreurs sur les variables discriminantes. Les nouvelles séances uniques ont pu être appariées entre elles, ce qui permettait d'apparier de nouveaux individus. Cette séquence d'appariement a été répétée plusieurs fois jusqu'à épuisement des séances pouvant être appariées. Au total, près de 80 % des patients ont pu être appariés. La procédure d'appariement a permis d'apparier 89 211 séances et 29 088 patients. Nous avons ainsi obtenu deux bases de données, une avec comme unité la séance et l'autre avec comme unité le patient. Chaque base contenait des observations appariées, des observations provenant uniquement de la base OMG et des observations provenant uniquement de la base SNIIRAM (figure 1).

Les 89 211 séances appariées sont décrites dans le tableau I. Parmi elles, 2 608 (3 %) étaient des visites alors que le taux de visite était de près de 7 % sur l'ensemble des séances de la base SNIIRAM. Les pourcentages de visites étaient très variables d'un médecin à l'autre. En effet, un

Tableau I : Description des séances appariées (IRDES – Prospere, France, 2008)

Séances appariées (N = 89 211)	N	%
Type de séance		
Visites	2 608	3
Consultations	86 603	97
Séances auprès du médecin traitant	58 541	66
Séances avec affection de longue durée	17 579	20

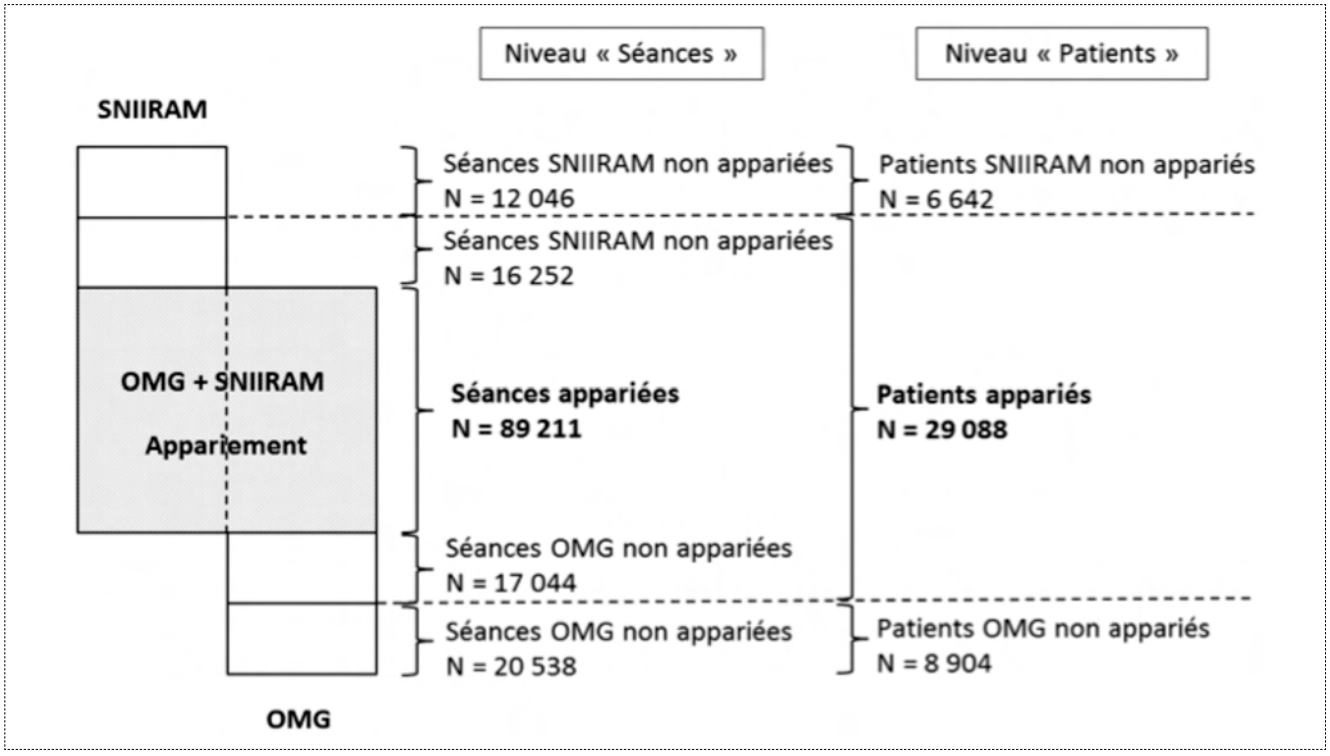

Figure 1 : Description des bases de données (IRDES – Prospere, France, 2008)

Tableau II : Description des patients appariés (IRDES – Prospere, France, 2008)

Patients appariés (N = 29 088)	N	%	Moyenne	Écart-type
Sexe masculin	13 270	46		
Âge en années			38	23,5
Couverture maladie universelle de base	615	2		
Couverture maladie universelle complémentaire	1 658	6		
Au moins une affection de longue durée (ALD)	3 718	13		
Nombre de séances auprès d'un médecin généraliste			6	4,8
Nombre de séances auprès du médecin traitant			2	2,9

tiers des médecins enregistrait moins de 1 % de visites alors qu'un quart d'entre eux en relevait plus de 5 %.

Le tableau II décrit les caractéristiques sociales et médicales des 29 088 patients appariés. Notre population d'étude, *a priori* non représentative de la population générale dans le cadre du test de faisabilité, avait en moyenne 38 ans et était constituée de 46 % d'hommes. Selon les informations issues du SNIIRAM, 2 % des patients appariés avaient été bénéficiaires de la Couverture maladie universelle (CMU) de base au moins une fois dans l'année. Six pour cent des patients avaient été bénéficiaires de la CMU complémentaire au moins une fois dans l'année. Les chiffres en population générale étaient respectivement de 3 % et 7 % [18]. Au sein de notre échantillon, 3 718 (13 %) patients présentaient une affection de Longue Durée (ALD). Les 4 869 ALD présentes dans notre population (plusieurs ALD par patient sont possibles) sont décrites dans le tableau III. En moyenne, nos patients ont eu six contacts avec un médecin généraliste dans l'année, ce qui est conforme aux données en population générale [18]. Les patients ont consulté en moyenne deux fois leur médecin traitant alors que 66 % des séances ont été réalisées auprès du médecin traitant. Cela illustre le fait que les patients consultant souvent un médecin généraliste voient régulièrement leur médecin traitant, alors que les patients qui consultent moins souvent ont plus tendance à choisir un médecin généraliste autre que leur médecin traitant.

Apport de l'appariement des données dans le repérage des pathologies chroniques

La comparaison des ALD et des RC dans le cas du diabète a montré que 94 % (521/553) des patients en ALD diabète ont été pris en charge spécifiquement pour leur diabète par un des 30 médecins généralistes au moins une fois dans l'année. Par contre, seulement 65 % (521/799) des patients ayant consulté pour diabète (au moins un RC diabète dans l'année) étaient déclarés en ALD diabète. Mais 75 % d'entre eux avaient au moins une ALD déclarée, que ce soit pour un diabète ou pour une autre pathologie. Cela nous confirme le caractère souvent polypathologique des patients diabétiques.

Dans le cas de l'hypertension artérielle, 90 % (323/360) des patients ayant une ALD pour ce motif ont consulté un des 30 médecins généralistes pour cela au moins une fois dans l'année. En revanche, seulement 11 % (323/2 979) des patients ayant eu un RC codé « hypertension artérielle » étaient déclarés en ALD spécifique. Ce chiffre passe à 36 % (1 069/2 979) lorsqu'on considère l'ensemble des ALD. Ceci est probablement lié au critère de gravité indispensable à la déclaration en ALD « hypertension artérielle » définie par une hypertension artérielle traitée par au moins deux médicaments anti-hypertenseurs.

Nous avons ensuite comparé la méthode de repérage des patients diabétiques et hypertendus avec l'information sur les traitements remboursés. Pour le diabète, 684 patients avaient été remboursés pour au moins un médicament antidiabétique au cours de l'année 2008. Parmi eux, 92 % avaient eu un RC diabète. À l'inverse, 20 % des patients ayant eu un RC diabète n'ont pas eu de traitement antidiabétique remboursé dans l'année. En croisant le repérage par l'ALD diabète et le repérage par le traitement, on s'aperçoit que 12 % des patients ayant eu un RC diabète ne sont pas repérables dans le SNIIRAM car ne sont pas déclarés en ALD et n'ont pas eu de traitement antidiabétique remboursé dans l'année.

Dans l'exemple de l'hypertension artérielle, nous avons également retrouvé 12 % de patients ayant eu un RC hypertension artérielle qui n'étaient pas directement repérables dans le SNIIRAM par l'ALD hypertension artérielle ou la présence de deux traitements antihypertenseurs de classes différentes remboursés dans l'année.

Tableau III : Répartition des affections de longue durée (N = 4 869) présentes dans la population d'étude (N = 3 718) et comparaison aux données du régime général [19] (IRDES – Prospere, France, 2008)

	Affections de longue durée	N	% dans la population d'étude	% sur la population du régime général en 2007
30	Tumeur maligne, affection maligne du tissu lymphatique ou hématopoïétique	993	26,7	20,7
8	Diabète de type 1 et diabète de type 2	762	20,5	19,7
23	Affections psychiatriques de longue durée	549	14,8	11,9
12	Hypertension artérielle sévère	495	13,3	12,8
13	Infarctus coronaire	429	11,5	10,2
5	Insuffisance cardiaque grave, troubles du rythme graves, cardiopathies valvulaires graves, cardiopathies congénitales graves	330	8,9	6,9
3	Artériopathies chroniques avec manifestations ischémiques	232	6,2	4,8
14	Insuffisance respiratoire chronique grave	177	4,8	3,7
1	Accident vasculaire cérébral invalidant	122	3,3	2,9
9	Formes graves des affections neurologiques et musculaires (dont myopathie), épilepsie grave	110	3,0	2,4
15	Maladie d'Alzheimer et autres démences	110	3,0	2,7
6	Maladies chroniques actives du foie et cirrhoses	96	2,6	2
22	Polyarthrite rhumatoïde évolutive grave	76	2,0	2
98	Pathologies hors liste	60	1,6	–
16	Maladie de Parkinson	41	1,1	1
19	Néphropathie chronique grave et syndrome néphrotique primitif	37	1,0	1,1
24	Rectocolite hémorragique et maladie de Crohn évolutives	37	1,0	1,3
7	Déficit immunitaire primitif grave nécessitant un traitement prolongé, infection par le virus de l'immuno-déficience humaine (VIH)	31	0,8	1,1
17	Maladies métaboliques héréditaires nécessitant un traitement prolongé spécialisé	31	0,8	0,5
21	Périarthrite noueuse, lupus érythémateux aigu disséminé, sclérodermie généralisée évolutive	29	0,8	0,6
27	Spondylarthrite ankylosante grave	28	0,8	0,7
25	Sclérose en plaques	18	0,5	0,8
26	Scoliose structurale évolutive (dont l'angle est égal ou supérieur à 25 degrés) jusqu'à maturation rachidienne	17	0,5	0,2
99	Polypathologie	14	0,4	–
11	Hémophilies et affections constitutionnelles de l'hémostase graves	13	0,3	0,3
10	Hémoglobinopathies, hémolyses, chroniques constitutionnelles et acquises sévères	9	0,2	0,1
20	Paraplégie	8	0,2	0,4
2	Aplasie médullaire et autres cytopénies chroniques	6	0,2	0,1
29	Tuberculose active, lèpre	6	0,2	0,1
28	Suites de transplantation d'organe	2	0,1	0,1
18	Mucoviscidose	1	0,03	0,06

Discussion

Cette étude a montré, pour la première fois en France, la faisabilité avec un niveau de qualité encourageant d'un appariement probabiliste de bases de données cliniques et médico-administratives. Par ailleurs, nous avons pu commencer à explorer l'intérêt et les apports du rapprochement de ces deux sources pour identifier des populations selon la morbidité identifiée par les médecins de premiers recours.

Ce nouveau modèle de bases de données paraît utile pour étudier l'épidémiologie des pathologies identifiées et prises en charge par le système de soins. L'information clinique permet ainsi de repérer les pathologies à un stade probablement précoce avant traitement médicamenteux ou mise sous ALD, ce que ne permettent pas les seules données du SNIIRAM. Ces patients sont probablement au début de leur maladie car non traités par des médicaments, peut-être traités par des mesures hygiéno-diététiques, et ne sont pas encore déclarés en ALD. Les actions de prévention étant surtout utiles à ce stade, il paraît important de mieux les repérer à l'échelle nationale. Un suivi de cohorte pourrait apporter de nombreux éléments sur la dynamique des maladies, notamment chroniques, et leurs prises en charge selon la complexité présentée. Une analyse complémentaire de la polypathologie par les données cliniques nous paraît également une piste à privilégier notamment en termes de parcours de soins et de polyprescription dans la suite des travaux déjà menés par Clerc *et al.* [20].

Cette étude présentait cependant plusieurs limites. Tout d'abord la phase-test initiée en 2009 a été réalisée sur l'année la plus récente et seules les données de l'année 2008 étaient exploitables. Une vision pluriannuelle pourrait permettre d'améliorer le repérage des patients et de mieux appréhender les parcours de soins des patients présentant une pathologie chronique. Par ailleurs, même si l'appariement paraît de bonne qualité, il est important de rappeler que 23 % des patients OMG n'ont pas pu être appariés comme 19 % des patients identifiés dans les bases de l'assurance maladie. Ces écarts sont probablement dûs au fait que seules les données des bénéficiaires du régime général étaient accessibles dans la base SNIIRAM pour l'année 2008. Les patients des autres régimes (MSA, RSI, SLM) n'ont donc pas pu être appariés. Cependant, pour les données de l'année 2009 et des suivantes, tous les régimes de sécurité sociale ont été intégrés au SNIIRAM, ce qui devrait améliorer l'appariement. De plus, il existe probablement un sous-codage dans l'OMG qui explique également que tous les patients et les séances OMG ne soient pas appariées. Par exemple, nous avions proportionnellement beaucoup moins de visites dans notre population que dans la population générale. Ceci est probablement dû au fait que certains médecins de l'OMG enregistrent moins souvent et de façon moins précise les actes de visites que les actes de consultations.

La comparaison entre les ALD et les RC n'est pas toujours aisée car ces deux classifications sont basées sur des logiques très différentes. Les ALD résultent d'un processus médical et administratif [21] pour permettre une prise en charge des dépenses pour des situations pathologiques définies réglementairement avec certains critères de gravité qui ont comme caractéristique commune de nécessiter une prise en charge longue, compliquée et coûteuse et après accord des médecins-conseils. La mise sous ALD peut varier selon les médecins traitants (par exemple quand un patient bénéficie de la CMU, les médecins ne font pas toujours la démarche de demande d'ALD, les soins étant déjà pris en charge par la CMU), mais également selon les pratiques des services médicaux des caisses primaires d'assurance maladie. *A contrario*, les RC regroupent des symptômes, syndromes et pathologies qui ont été réellement pris en charge par le médecin au cours de la consultation. Par ailleurs, les ALD ne regroupent qu'une petite partie de l'ensemble des pathologies possibles, même si ce sont aussi les plus compliquées à prendre en charge en termes de santé publique, minorant ainsi la polypathologie qui est un phénomène d'importance croissante mal appréhendé par les systèmes de santé.

Enfin, les données de l'OMG sont à ce jour assez restreintes. Nous n'avions en particulier pas de recueil systématique des antécédents et des facteurs de risques majeurs de certaines pathologies.

Conclusion

Cette étude nous a permis de décrire une méthode efficace et novatrice en termes d'appariement de bases de données médicales de ville et médico-administratives. Les premières exploitations réalisées sur la base-test illustrent les apports potentiels de ce modèle de données appariées en termes de ciblage de populations à risque, d'estimation de la morbidité prise en charge en médecine générale, de l'analyse du parcours de soins en fonction des pathologies et de leurs associations et des stratégies des médecins généralistes face aux patients présentant des pathologies

isolées ou multiples. La comparaison des prescriptions (médicaments et actes) effectuées par les médecins et des acquisitions et remboursements des patients, ainsi que l'analyse des tableaux de patients polypathologiques permettront également d'enrichir les méthodes d'évaluation de la qualité des interventions et des soins en analysant par exemple des écarts de pratiques par rapport aux référentiels ainsi que de nombreuses études de pharmaco-épidémiologie. Une extraction sur plusieurs années consécutives et sur l'ensemble des régimes composant le SNIIRAM améliorerait la qualité de l'appariement. Cette deuxième étape envisagée dans le cadre du projet Prospere pour les années 2009, 2010 et 2011 et pour un échantillon de 80 médecins généralistes n'a pu être menée à terme en raison de la fermeture de l'OMG fin 2011. Cette issue illustre les limites du bénévolat et la nécessité pour un projet d'envergure nationale d'un portage institutionnel et professionnel élargi qui paraît désormais envisageable avec la création récente du Collège national de la médecine générale (CMG). Ce dernier a manifesté avec le comité d'interface de la médecine générale son souhait de voir se développer un tel outil dans le cadre de la stratégie nationale de santé [22].

La constitution d'une base de données élargie, représentative des pratiques des quelques 53 000 médecins généralistes libéraux exclusifs hors exercice particulier auprès desquels sont inscrits près de 90 % des assurés en France, nécessite la participation d'un échantillon de 1 000 à 2 000 médecins. Un tel projet qui relève d'un investissement pour l'avenir serait alors, à l'image des grandes infrastructures de recherche constituées avec les cohortes en épidémiologie ou les bio-banques dans le domaine des sciences biomédicales, au service du renforcement des soins primaires et des services de santé comme domaine de recherche mais aussi plus largement au service de la recherche en santé publique.

Aucun conflit d'intérêt déclaré

Références

1. Caisse nationale de l'assurance maladie des travailleurs salariés (CNAMTS). Améliorer la qualité du système de santé et maîtriser les dépenses : propositions de l'Assurance Maladie pour 2014. Rapport Charges et Produits. Paris : CNAMTS ; 2013.
2. Allonier C, Dourgnon P, Rochereau T. Enquête sur la santé et la protection sociale 2008. Paris : l'Institut de recherche et de documentation en économie de la santé (IRDES), Rapport n° 547 ; 2010.
3. Montaut A, Calvet L, Bouvier G, Gonzalez L. L'appariement handicap-santé et données de l'assurance maladie. Série sources et méthodes. (40). Paris : Direction de la recherche, des études, de l'évaluation et des statistiques (Drees) ; 2013.
4. Chevreul K, Le Fur P, Renaud T, Sermet C. Faisabilité d'un système d'information public sur la médecine de ville. Paris : Institut de recherche et documentation en économie de la santé (IRDES), rapport n° 535 ; 2006.
5. Livre blanc 2010 du Comité d'experts : 7 propositions au service de la recherche. Paris : Institut des données de santé (IDS), rapport ; 2010.
6. Goldberg M, Quantin C, Guégen A, Zins M. Bases de données médico-administratives et épidémiologie : intérêts et limites. Courrier des statistiques (INSEE). 2012.
7. Institute for Clinical Evaluative Sciences (ICES) [Internet]. Available from <http://www.ices.on.ca/>.
8. Clinical Practice Research Datalink [Internet]. Available from <http://www.cprd.com>.
9. Pour une meilleure utilisation des bases de données nationales pour la santé publique et la recherche. Paris : Haut conseil de la santé publique (HCSP), rapport ; 2012.
10. Bras PL. Rapport sur la gouvernance et l'utilisation des données de santé. Paris : Inspection générale des affaires sociales (IGAS), rapport ; 2013.
11. Gensbittel MH, Riandey B. Appariements sécurisés et statistique (2000-2011) : une décennie d'expériences. Courrier des Statistiques (INSEE). 2011;(131).
12. Tromp M, Ravelli AC, BonselGJ, Hasman A, ReitsmaJB. Results from simulated data sets: probabilistic record linkage outperforms deterministic record linkage. J Clin Epidemiol. 2011;64(5):565-72.
13. Da Silveira DP, Artmann E. Accuracy of probabilistic record linkage applied to health databases: systematic review. Rev Saúde Pública. 2009;43(5):875-82.
14. Mason CA, Tu S. Data linkage using probabilistic decision rules: a primer. Birth Defects Res A Clin Mol Teratol. 2008;82(11):812-21.
15. Observatoire de la médecine générale (OMG) [Internet]. Available from <http://omg.sfmg.org/content/com/>.
16. Jaro MA. Probabilistic linkage of large public health data files. Stat Med. 1995;14(5-7):491-8.
17. Saint-Lary O, Boisnault P, Naiditch M, Szidon P, Duhot D, Bourgueil Y, *et al.* Performance scores in general practice: a comparison between the clinical versus medication-based approach to identify target populations. PloS One. 2012;7(4):e35721.
18. Institut de recherche et documentation en économie de la santé (IRDES) – Bases de données Eco-Santé en ligne [Internet]. Available from <http://www.ecosante.fr/>.
19. Païta M, Weill A. Les personnes en affection de longue durée au 31 décembre 2007. Points de repère (CNAMTS). 2008;(20).
20. Clerc P, Lebreton J, Mousques J, Hebbrecht G, de Pouvourville G. Étude Polychrome: construction d'une typologie des pathologies chroniques en médecine générale, pour une analyse de la polyprescription. Prat Organ Soins. 2008;39(1):43-51.
21. Haut Conseil de la santé publique. La prise en charge et la protection sociale des personnes atteintes de maladie chronique. Paris : Haut Conseil de la santé publique (HCSP), rapport. 2009.
22. Collège de médecine générale [Internet]. Available from <http://www.lecmg.fr/>.