# L'apport des bases de données médico-administratives

**Céline Ribet**
**Mireille Cœuret-Pellicer**
**Julie Gourmelen**
Inserm U1018, Plate-forme de recherche Cohortes épidémiologiques en population — Centre de recherche en épidémiologie et santé des populations, université de Versailles-Saint-Quentin, UMRS 1018

La France est l'un des rares pays dont les organismes de protection médico-sociale ou de gestion hospitalière disposent de systèmes d'information centralisés couvrant de façon exhaustive et permanente l'ensemble de la population. Les données enregistrées en routine comportent des informations sur le recours aux soins, les hospitalisations, le handicap, les prestations sociales et l'activité professionnelle. Bien que n'ayant pas à l'origine de finalité épidémiologique, ces bases offrent un intérêt potentiel majeur pour la réalisation de telles études mais sont encore très peu exploitées. On présentera ici les principaux systèmes d'information, leur exploitation potentielle en santé publique ainsi que les précautions que nécessite leur utilisation.

## Description des principales données disponibles

### Les données socioprofessionnelles

Les événements socioprofessionnels des individus sont informatisés dans les systèmes nationaux des différents régimes d'assurance vieillesse. Pour toute personne ayant appartenu au moins une fois au cours de sa vie au régime général de la Sécurité sociale, c'est la Caisse nationale d'assurance vieillesse (Cnav) qui procède à l'enregistrement des données lui permettant de garantir le droit au paiement de la retraite. Pour répondre à cet objectif, la Cnav a mis en œuvre plusieurs systèmes nationaux lui permettant de collecter et traiter les données sociales issues de différents organismes et régimes gestionnaires des prestations sociales, dont le principal est le Système national de gestion des carrières (SNGC). Cette base de données permet de retracer, pour chaque individu dès l'âge de 16 ans et jusqu'à la liquidation de ses droits à la retraite, ses différentes périodes d'activité : périodes d'activité professionnelle (par l'intermédiaire des déclarations transmises par les employeurs) ou périodes assimilées (chômage, maladie, maternité ou congés parentaux ; informations transmises respectivement par l'Assurance chômage, l'Assurance maladie, et les caisses d'allocations familiales). Le SNGC contient donc l'ensemble des données inhérentes à la carrière des assurés du régime général, y compris les données concernant d'éventuelles périodes effectuées dans d'autres régimes de base (régimes des indépendants, des agriculteurs…) ainsi que dans certains régimes particuliers ou spéciaux (SNCF, EDF…).

Un autre système d'information mis en œuvre par la Cnav est le Répertoire national inter-régimes des bénéficiaires de l'assurance maladie (RNIAM), qui permet de connaître l'organisme de rattachement de chaque bénéficiaire d'un régime d'assurance maladie par l'intermédiaire du NIR (lire encadré).

### Les données de mortalité

Le statut vital et les causes de décès des sujets d'une enquête peuvent être obtenus auprès du Centre d'épidémiologie sur les causes médicales de décès (CépiDC)

## NIR, RNIPP, SNGI

Le « numéro d'inscription au répertoire », ou NIR, est l'identifiant unique et invariable de tout individu. Ce numéro à treize caractères (plus deux pour la clé de contrôle), dont la composition est précisée par décret, est attribué à une seule et unique personne, et une personne ne possède qu'un NIR. Une fois attribué, il ne change plus.

L'attribution de ce numéro et son association aux autres éléments d'identification d'un individu (nom patronymique, prénoms, date et lieu de naissance, numéro de l'acte de naissance, sexe) se font dès la naissance sur la base des informations enregistrées par l'état civil. Au moment du décès, s'ajoutent les date et lieu de décès et le numéro de l'acte.

Pour les personnes nées en France métropolitaine ou dans les DOM, qu'elles soient françaises ou étrangères, c'est l'Insee qui a en charge cette immatriculation et qui procède à sa conservation au sein du Répertoire national d'identification des personnes physiques (RNIPP). Pour les personnes nées à l'étranger, à Mayotte et dans les TOM, c'est la Cnav qui met en œuvre d'une part l'immatriculation (uniquement lorsque l'inscription est demandée par un organisme habilité), et d'autre part la conservation au sein du Système national de gestion des identités (SNGI). Ces deux fichiers ont pour finalité de certifier l'état civil et le statut vital d'une personne auprès des organismes de sécurité sociale, de l'administration fiscale, de la Banque de France, du Système informatique pour le répertoire des entreprises et des établissements (Sirene). Leur utilisation repose sur de fortes obligations légales ; ainsi, ils ne peuvent être servir à des fins de recherche des personnes.



de l'Inserm selon la procédure décrite dans le décret n° 98-37. Cette procédure permet d'apparier des données d'état civil et de statut vital hébergées par l'Insee aux causes médicales de décès anonymes.

### Les données d'hospitalisation

Le Programme de médicalisation du système d'information des hôpitaux (PMSI) consiste en un recueil exhaustif systématique et standardisé d'informations médicales et administratives pour tout séjour d'un patient dans un établissement de soins. Il concerne aujourd'hui tous les établissements (publics et privés) et tous les types de séjours (médecine, chirurgie, obstétrique, soins de suite et de réadaptation, psychiatrie, urgences, soins à domicile). L'objectif principal du PMSI est de décrire l'activité d'un établissement à des fins d'allocation budgétaire. L'information est médicalisée et repose sur un classement des séjours en « groupes médicalement homogènes » (GHM), à partir du codage des diagnostics établis au cours d'un séjour et des principaux actes pratiqués. Ces informations sont anonymisées puis rassemblées dans une base de données nationale gérée par l'Agence technique de l'information sur l'hospitalisation (ATIH).

### Les données de l'Assurance maladie

Il existe en France un grand nombre de régimes d'assurance maladie, disposant chacun de son propre système d'information contenant les données nécessaires à la liquidation des prestations de ses assurés. Ces données comprennent des informations détaillées sur les soins présentés au remboursement (consultations, médicaments, prélèvements biologiques…), ainsi que sur les assurés, les établissements de soins et les professionnels de santé. Les services médicaux des caisses disposent de leurs propres fichiers comportant des informations médicales structurées sur les affections de longue durée (ALD), les accidents du travail et les maladies professionnelles.

La nécessité de suivre l'ensemble des dépenses tous régimes confondus a abouti en 2003 à la création du Système national d'informations inter-régimes de l'Assurance maladie (SNIIR-AM). Ces données concernent aujourd'hui tous les régimes d'assurance maladie, pour la médecine de ville comme pour l'hospitalisation. Elles sont individualisées par bénéficiaires, professionnels de santé et établissements, et médicalisées (les actes sont codés selon la Classification commune des actes médicaux et les pathologies selon la CIM10).

Grâce à un identifiant anonyme commun, les données du PMSI sont également désormais intégrées au SNIIR-AM.

## Utilité des bases dans un cadre épidémiologique

Les bases médico-administratives offrent de nombreux avantages inhérents à leur constitution.

Les données sont individuelles. Ainsi, l'accès à ces bases de données peut servir à sélectionner des individus en vue de l'inclusion dans une enquête épidémiologique à partir de critères tels qu'une pathologie, un recours à des soins spécifiques ou une profession. Un exemple récent est l'étude des effets du Médiator : il a été possible d'identifier dans le SNIIR-AM toutes les personnes ayant eu une prescription remboursée de ce médicament, et de suivre leur devenir médical, avec les résultats que l'on sait [46].

Les données sont quasi exhaustives par rapport à la population française. Elles permettent donc de disposer d'effectifs immenses pour certaines analyses. Cette exhaustivité peut aider à prendre en compte les effets de sélection à l'inclusion et au cours du suivi, qui sont une source majeure de biais dans les enquêtes épidémiologiques (lire *Aspects méthodologiques liés à l'analyse de données longitudinales et aux effets de sélection, p. 18*) :

- la constitution d'un fichier de « non-participants », pour lesquels on pourra disposer de données sur leurs consommations de soins, leurs hospitalisations et leurs caractéristiques socioprofessionnelles, permet d'étudier les facteurs liés à la non-participation ;
- le suivi de façon « passive », à travers ces bases, des personnes incluses dans des études mais qui ne répondent plus aux questionnaires permet de pallier le problème des perdus de vue.

Enfin, ces données sont parfois plus fiables que des informations obtenues par auto-questionnaire. Par exemple, les informations sur la carrière professionnelle, qui servent au calcul des retraites, sont pour des raisons évidentes particulièrement complètes et validées, toute erreur pouvant en effet avoir un impact économique sur les bénéficiaires comme sur la collectivité.

Ces avantages font que, couplées à des enquêtes auprès des personnes, ces bases de données peuvent faire l'objet d'utilisations très diversifiées dans le cadre des études épidémiologiques et peuvent apporter des solutions satisfaisantes à divers problèmes fréquemment rencontrés lors de la mise en œuvre de ces études, qu'il s'agisse de l'inclusion ou du suivi des sujets ou de l'accès à des données concernant des événements d'intérêt.

## Tenir compte des limites

Si ces bases de données constituent un intérêt certain, il faut toujours garder à l'esprit qu'elles ont été construites uniquement pour répondre aux objectifs de gestion des organismes qui les ont constituées. Leur utilisation par des épidémiologistes nécessite d'une part un important travail de réflexion concernant l'accès à ces données, leur appariement aux données d'enquêtes et les circuits de confidentialité à mettre en œuvre, et d'autre part un travail crucial de contrôle et de validation des données.

### L'accès aux données

L'identification des personnes dans les bases de données médico-administratives et sociales repose sur le « numéro d'inscription au répertoire », ou NIR, communément

*Les références entre crochets renvoient à la Bibliographie générale p. 51.*

appelé numéro Insee ou numéro de Sécurité sociale (voir encadré). Or, en dehors même des études épidémiologiques, l'utilisation directe de cet identifiant est soumise à de fortes contraintes juridiques (plusieurs lois et décrets définissent son accès, son usage et sa conservation dans les systèmes d'information). Il est possible de trouver des solutions à cette difficulté, mais elle constitue actuellement un obstacle formel pour la plupart des études en dehors d'un éventuel partenariat avec un organisme habilité à détenir ce numéro.

Reste ensuite un important travail pour définir les procédures de transmissions sécurisées entre les différents intervenants (fournisseurs de données, responsables de la gestion de l'étude, chercheurs), afin de garantir aux données à caractère personnel une confidentialité conforme aux textes.

Ainsi, l'accès et l'utilisation de ces bases de données restent complexes et nécessitent, dans des conditions compatibles avec les contraintes de qualité des études épidémiologiques, des moyens lourds et des compétences spécialisées. Il est vraisemblable que très peu d'équipes d'épidémiologie en France disposent actuellement de ces ressources.

### La validité des données

Comme déjà évoqué, l'utilisation de ces bases de données en dehors des champs pour lesquels elles ont été développées nécessite un travail complexe de contrôle et de validation, particulièrement dans le cas des études épidémiologiques où la précision des données concernant les événements de santé est cruciale.

Dans le cas précis des données de santé, aucune de ces bases prise isolément ne permet d'obtenir des informations complètes et d'une validité suffisante.

Les données de consommations de soins ne comportent pas d'information sur la nature des maladies traitées et excluent par définition l'automédication, les prestations non présentées au remboursement, et n'informent pas sur l'observance des traitements délivrés. Il est également établi que la prévalence des ALD enregistrées est systématiquement inférieure à la prévalence réelle des affections pour différentes raisons : patient atteint de l'une de ces maladies mais ne répondant pas aux critères de sévérité exigés ou ne demandant pas à bénéficier du dispositif, par exemple s'il est déjà exonéré du ticket modérateur au titre d'une autre affection.

La validité des diagnostics, que ce soit pour les causes de décès, les ALD ou le PMSI, dépend fortement de la qualité du codage à la production de l'information, celle-ci pouvant être affectée par différents problèmes (variabilité entre praticiens, biais liés aux finalités budgétaires du PMSI…). Plusieurs études ont montré que l'utilisation du PMSI ne pouvait pas se suffire du diagnostic principal, mais nécessitait des algorithmes complexes alliant les codes diagnostics aux codes actes spécifiques de la pathologie étudiée [10, 11].

Dans de nombreuses situations, il est donc nécessaire de mettre en place des procédures de validation de ces données. Les méthodes utilisées peuvent être variées : retour à des informations du dossier médical *via* les médecins traitants, confrontation avec des questionnaires remplis par les sujets, croisement avec d'autres sources (données de registre, causes de décès…). Une voie prometteuse est le développement d'algorithmes incluant des données provenant de l'appariement de l'ensemble de ces bases (remboursements de médicaments enregistrés dans le SNIIR-AM, diagnostics des ALD, actes et diagnostics du PMSI).

## Conclusion

L'utilisation des bases de données d'origine socio-médico-administrative peut grandement faciliter les travaux de recherche en santé, voire améliorer la qualité des études. La résolution des problèmes évoqués pour optimiser leur utilisation pourra contribuer au développement en France de grandes cohortes comparables à celles qui existent dans d'autres pays.