[Epidemiological studies based on medical and administrative databases : a potential strength in France].

Marcel Goldberg, Mireille Coeuret-Pellicer, Céline Ribet, Marie Zins

COHORTES EPIDEMIOLOGIQUES ET BASES DE DONNEES D'ORIGINE ADMINISTRATIVE : UN RAPPROCHEMENT POTENTIELLEMENT FRUCTUEUX

Marcel Goldberg[1,2], Mireille Coeuret-Pellicer[1,2], Céline Ribet[1,2], Marie Zins[1,2]

1- Inserm U1018, Plateforme de recherche Cohortes épidémiologiques en population - Centre de recherche en Épidémiologie et Santé des Populations

16 avenue Paul Vaillant-Couturier, F-94807, Villejuif, France

2- Université de Versailles-Saint Quentin, UMRS 1018, France

Auteur correspondant :

Marcel Goldberg

Téléphone : 33 1 77 74 74 26

Fax : 33 1 77 74 74 02

Email : marcel.goldberg@inserm.fr

RESUME

L'épidémiologie développe des cohortes de centaines de milliers de sujets qui sont suivis sur des décennies. La France dispose d'un atout potentiel, les bases de données médicosociales nationales : *Programme de Médicalisation du Système d'Information* hospitalier, *Système national d'information inter régimes de l'assurance maladie*, SNIIR-AM, système d'information de la Caisse nationale d'assurance vieillesse.

Elles offrent de nombreux avantages : exhaustivité de la population, absence de perdus de vue pendant le suivi, données souvent fiables, appariement avec des enquêtes. Des problèmes de validité des données médicales se posent cependant, et nécessitent un important travail de réflexion méthodologique, de contrôle et de validation de données. Il reste également de nombreux problèmes légaux et techniques à résoudre.

ABSTRACT

Population-based epidemiological cohorts may include nowadays hundreds of thousands of subjects, followed-up during decades. France has a major potential strength: nationwide medical and social databases set up for administrative purposes. The main databases useful for epidemiology are the social security database which contains individual medical data from different sources, and the retirement fund database on employment social benefits.

These databases have several advantages: they cover the whole French population, with no lost to follow-up, data are often of good quality and it is possible to link them with individual surveys. However medical data are not always ascertained and an important methodological and practical work has to be done, as well as some legal and practical problems have to be solved for an optimal use.

LES COHORTES EPIDEMIOLOGIQUES EN POPULATION : UN BESOIN MECONNU EN FRANCE

La cohorte épidémiologique est un type d'enquête dont le principe est le suivi longitudinal, à l'échelle individuelle, d'un groupe de sujets. Il faut distinguer les cohortes de malades souffrant d'une pathologie particulière, et les cohortes en population générale. Les premières, dont l'objectif est d'étudier l'évolution d'une maladie, incluent un nombre souvent restreint de sujets (quelques milliers pour les plus importantes) habituellement recrutés en milieu médical, et les données recueillies sont très détaillées, incluant notamment des investigations biocliniques approfondies. Les secondes, établies en population générale, sont celles qui font l'objet de cet article. Elles s'intéressent aux causes des maladies, particulièrement les maladies plurifactorielles aux déterminants environnementaux et génétiques multiples. Ces cohortes doivent inclure et suivre, souvent pendant des décennies, de très vastes échantillons pour lesquels sont recueillies de façon prospective des données personnelles, de mode de vie, sociales, professionnelles et environnementales, et qui s'accompagnent de biobanques. Ce type de cohorte, selon la définition de l'ANRS « *doit être conçu pour répondre à plusieurs questions de recherche épidémiologique, clinique, biologique ou de santé publique même si certaines ne sont pas encore formulées de façon précise au démarrage de la cohorte* ». Globalement, les études de cohorte sont celles qui permettent de proposer les meilleures conditions pour juger en termes de causalité du rôle sur la santé de facteurs de risque (ou d'interventions préventives), en permettant de prendre en compte les évolutions temporelles et les interactions entre facteurs.

Actuellement, l'épidémiologie fait face à la nécessité de développer des études de taille autrefois inimaginable. Qu'il s'agisse de mettre en évidence des risques de faible ampleur associés à l'exposition à des agents potentiellement pathogènes, d'évaluer l'efficacité d'interventions dont on attend des bénéfices d'ampleur modeste, ou de décrire la distribution et l'évolution d'événements peu fréquents, ce sont aujourd'hui des études cas-témoins en population générale de milliers ou de dizaines de milliers de sujets qui sont mises en place, ou des cohortes de centaines de milliers, voire de millions de sujets qui sont suivis de façon prospective pendant des périodes qui s'étendent sur des décennies [1].

Dans ce paysage, la France ne distingue pas particulièrement par ses grandes réalisations. À titre d'illustration, on constate que les cohortes prospectives françaises se caractérisent par leur taille relativement faible, aucune ne dépassant un petit nombre de dizaines de milliers de sujets, alors que certaines cohortes prospectives dans d'autres pays peuvent atteindre plusieurs centaines de milliers de sujets, voire plus. À titre d'illustration, on peut citer en Grande-Bretagne la *One Million Women Study* [2], le projet *UK Biobank* [3] qui a mis en place le suivi prospectif de 500 000 personnes, ou la *Norvegian Mother and Child Cohort Study* qui a inclus 100 000 femmes à la 18ème semaine de grossesse, puis leurs 100 000 nouveau-nés, ainsi que 70 000 pères, soit au total 270 000 personnes [4]. La *Nurses'Health Study* a été mise en place aux États-Unis dès 1976 et assure le suivi prospectif de près de 250 000 infirmières [5]. Actuellement se mettent en place en Europe de nouvelles très grandes cohortes, comme LifeGene en Suède (*www.lifegene.ki.se*), LifeLines aux Pays-Bas (*www.lifelines.net*), ou la Cohorte nationale allemande (www.helmholtz.de/en/research/health/the_latest_insights/insights_archive/who_stays_healthy/) qui doivent inclure et suivre plusieurs centaines de millions de sujets recrutés en population générale. On peut citer aussi l'exemple des pays scandinaves, qui disposent de multiples registres dans le domaine de la santé, de la protection sociale ou de l'activité

économique, couvrant la totalité de la population de ces pays, et qui sont largement ouverts aux chercheurs, permettant par appariement de ces bases de données de constituer des cohortes dont l'effectif se compte en millions de sujets et qui sont à l'origine d'une immense bibliographie scientifique.

La relative modestie des cohortes françaises s'explique par de nombreuses raisons. Outre le nombre notoirement trop faible des épidémiologistes [6], on se heurte aujourd'hui en France à de nombreuses difficultés d'ordre financier, organisationnel et technique.

Les coûts des cohortes sont élevés, car l'épidémiologie fait essentiellement appel à des données qui sont le plus souvent recueillies auprès des personnes elles-mêmes par des moyens divers : entretiens, autoquestionnaires, examens médicaux, collecte de matériel biologique, etc. Ces coûts restent finalement modestes si on les compare à ceux des grands instruments de physique ou à ceux de la recherche spatiale, voire au prix d'une journée d'hospitalisation dans un service de CHU, mais, ils sont largement supérieurs aux budgets qu'il est possible de demander aux organismes nationaux de financement de la recherche pour des études épidémiologiques de grande dimension. En effet, contrairement aux autres pays scientifiquement avancés, la France n'a pas mis en place un système de financement adapté, et continue *de facto* d'ignorer l'importance scientifique de telles plateformes de recherche, malgré des efforts récents (appels à projet « *Très grandes infrastructures de recherche - Cohortes* » 2009 et « *Cohortes* » 2010 des Investissements d'avenir). Cependant, les budgets qui ont été distribués sont très loin des coûts véritables, et très largement inférieurs aux financements des cohortes étrangères citées plus haut, montrant bien à quel point les besoins scientifiques sont actuellement sous-estimés par les autorités françaises de la recherche.

D'autres difficultés tiennent à la nécessité de l'implication à long terme des équipes dont la pérennité n'est souvent pas assurée. Un autre obstacle est la quasi impossibilité de disposer de personnel stable et d'un niveau de qualification suffisant, notamment du fait de l'absence de statut reconnu pour ce type d'activité dans les organismes publics de recherche, alors que la durée des projets est incompatible avec un trop fréquent renouvellement des personnels techniques qualifiés qui doivent assurer la continuité des procédures et des recueils de données.

Or, si on veut que la France se dote d'outils épidémiologiques d'envergure comparable à ce qui existe dans les pays de niveau scientifique comparable, de nouvelles cohortes prospectives sont indispensables, dont l'effectif ne se comptera plus en dizaines, mais en centaines de milliers de sujets.

LES BASES DE DONNEES MEDICO-ADMINISTRATIVES

Une grande partie des coûts des cohortes prospectives en population vient de la nécessité de « tracer » les sujets et de recueillir pour chacun des données de santé et de situation sociale. Or, de ce point de vue, notre pays dispose d'un atout potentiel d'importance. Il existe en effet en France des systèmes d'information gérés par des organismes de protection médicosociale ou de gestion hospitalière extrêmement puissants, dont peu de pays disposent à l'échelle nationale.

On utilise encore très peu en France les possibilités offertes par ces bases de données, qui offrent pourtant un intérêt potentiel majeur pour la réalisation d'études épidémiologiques. On se restreindra ici à la description des deux principaux systèmes d'information de nature

médicale et administrative qui contiennent des données individuelles d'intérêt général pour les épidémiologistes, qu'il s'agisse de l'inclusion et du suivi des sujets, ou de l'accès à des données concernant des événements d'intérêt, de santé ou de vie socioprofessionnelle.

Bases de données concernant des événements de santé

Outre les données de mortalité (statut vital et causes de décès) qui peuvent être obtenus par l'accès au Répertoire national d'identification des personnes physiques (RNIPP) et à la base de données du Centre d'épidémiologie des causes de décès de l'Inserm (CépiDc), il existe différentes bases de données réunissant des données diverses pouvant être utilisées dans des protocoles épidémiologiques.

Le PMSI (Programme de Médicalisation du Système d'Information) a pour objectif de produire des informations à contenu médical sur l'activité hospitalière. Il consiste en un recueil exhaustif d'informations administratives et médicales pour chaque séjour hospitalier (essentiellement diagnostic principal, diagnostics associés et actes pratiqués), qui sont centralisées dans une base de données nationale.

Les systèmes d'informations des différents régimes de l'Assurance maladie enregistrent des données très détaillées sur les consommations de soins remboursés (médicaments, consultations de professionnels de santé, etc.), dont l'objectif premier est la liquidation des prestations d'assurance maladie. Des informations médicales diverses sur les Affections longue durée (ALD), les Accidents du travail (AT) et les Maladies professionnelles (MP), dont l'objectif initial est le contrôle des pathologies ouvrant droit à une prestation, sont également enregistrées. L'ensemble des bases de données concernant les événements de santé est désormais réuni au sein du *Système national d'information inter régimes de l'assurance maladie* (SNIIR-AM). Les données du SNIIR-AM incluent tous les régimes de l'assurance maladie : CNAMTS, MSA, RSI et les 16 autres régimes spéciaux, et concernent aussi bien la médecine de ville que les hospitalisations. Il s'agit d'une base de données individuelles mais anonymes qui rassemble les données décrites ci-dessus (y compris le PMSI). Chaque personne est identifiée par un numéro d'anonymat permanent non réversible, qui permet de chaîner toutes les données le concernant dans les différentes sources qui alimentent le SNIIR-AM. Au total, le SNIIR-AM qui couvre la totalité de la population française, constitue la plus grande base de données de santé au monde.

Bases de données concernant des événements socioprofessionnels

La Caisse nationale d'assurance vieillesse (Cnav) a notamment pour rôle d'assurer le droit au paiement de la retraite. Pour cela, la Cnav a mis un place un système permettant de collecter et traiter les données sociales issues de différents organismes et régimes gestionnaires des prestations sociales pour chaque individu dès l'âge de 16 ans et jusqu'à la liquidation de ses droits à la retraite : périodes d'activité professionnelle ou assimilées (chômage, maladie, maternité ou congés parentaux…), incluant les employeurs et la catégorie socioprofessionnelle.

UN APPORT POTENTIEL MAJEUR POUR L'EPIDEMIOLOGIE

Quel intérêt ?

Dans un contexte épidémiologique, ces bases de données peuvent faire l'objet d'utilisations très diversifiées. En effet, si les bases de données d'origine administrative utilisées de façon isolée sont insuffisantes pour réponde à la majorité des questions posées par les

épidémiologistes, elles offrent de nombreux avantages : quasi exhaustivité de la population cible (et par conséquent absence de biais de sélection et effectifs immenses pour certaines analyses), quasi absence de perdus de vue pendant le suivi, données parfois plus fiables que celles obtenues par déclaration pour certaines informations (comme les consommations de soins ou la carrière professionnelle, par exemple). Couplées avec des enquêtes auprès des personnes, les bases de données de type administratif peuvent apporter des solutions satisfaisantes à divers problèmes courants en épidémiologie : traçage des sujets au cours du suivi de cohortes, y compris de très longue durée ; acquisition permanente de données d'intérêt, ce qui permet le suivi de nombreux problèmes ; validation de données de déclaration ; analyse des biais de participation à toutes les étapes (inclusion et suivi).

L'accès aux données individuelles des bases de données médico-administratives peut concerner divers types de procédures épidémiologiques ; on en citera quelques unes à titre d'illustration des potentialités de ces systèmes.

Constitution de cohortes ad hoc : les bases de données citées peuvent permettre de sélectionner des sujets selon des critères variés : pathologie, recours à des soins spécialisés, profession ou situation d'emploi, etc. Un exemple récent particulièrement médiatisé 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. Un autre exemple est l'étude nationale *Entred* concernant les personnes souffrant de diabète. En 2002, 10 000 patients diabétiques traités, tirés au sort dans les fichiers de soins remboursés, ont reçu un questionnaire ; après accord du patient, un questionnaire a été envoyé à son médecin traitant, et des données complémentaires ont été fournies par une requête du système d'information de l'Assurance maladie portant sur les soins remboursés et par une enquête auprès des hôpitaux [7].

Extraction de données concernant des sujets sélectionnés : il est possible de retrouver pour une personne donnée les enregistrements de données le concernant, sous réserve de disposer de certaines informations indispensables (cf. plus loin : Aspects légaux). Ainsi, on peut apparier aux bases du SNIIR-AM ou de la Cnav les sujets inclus dans une cohorte de population pour lesquels les épidémiologistes ont recueilli des données personnelles, et enrichir celles-ci sans collecte supplémentaire.

« Traçage » de sujets inclus dans des enquêtes : une des sources majeures de biais des études longitudinales est le problème des « perdus de vue », c'est-à-dire des sujets qu'on ne retrouve plus ; un des avantages du recours aux bases nationales est qu'il permet d'éviter les perdus de vue, ou du moins d'en limiter fortement le nombre, car les personnes sont toujours suivies dans les bases nationales.

Quelles limites ?

Les bases de données du PMSI et de l'Assurance maladie ne contiennent pas certaines données qui peuvent être essentielles, mais elles peuvent apporter une aide considérable à la réalisation de très nombreuses enquêtes épidémiologiques. Cependant, des problèmes de validité de ces bases de données se posent de façon parfois cruciale.

L'utilisation du PMSI comme source d'information sur les pathologies s'avère délicate et ne peut reposer uniquement sur le diagnostic principal. Il est nécessaire de développer des algorithmes plus complexes alliant les codes diagnostics aux codes actes spécifiques de la pathologie étudiée [8,9].

L'utilisation des bases de données de l'Assurance maladie dans une optique épidémiologique nécessite un important travail de réflexion méthodologique, de contrôle et de validation de données. Ainsi, les données de remboursement ne comportent pas d'information sur la nature des maladies traitées, et excluent par définition l'automédication et les prestations non présentées au remboursement. La base des ALD codées par des médecins reste une base de données à vocation médico-sociale, et ses limites sont connues : imprécision des diagnostics, absence d'exhaustivité des cas déclarés, risque de double déclaration [10].

Dans de nombreuses situations, il est donc nécessaire de mettre en place des procédures de validation des diagnostics extraits des bases de données. Celles-ci peuvent reposer sur des méthodes très variées : retour au médecin traitant, 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 d'ALD, de remboursement de médicaments, de diagnostics et d'actes enregistrés dans le SNIIR-AM et le PMSI. Ainsi un travail récent a montré qu'il est possible à partir de ce type de données d'identifier avec d'excellentes sensibilité et spécificité les patients souffrant d'une maladie de Parkinson [11].

LES PERSPECTIVES

Comme le montre l'expérience de certains pays, l'utilisation de bases de données d'origine administrative peut grandement faciliter les travaux des épidémiologistes, voire améliorer la qualité des études. Il reste cependant de nombreux problèmes à résoudre pour leur utilisation optimale.

Aspects légaux

L'identification des personnes dans les bases de données médico-administratives repose sur le « Numéro d'inscription au répertoire » (NIR, communément appelé numéro Insee ou numéro de sécurité sociale). Or la loi Informatique et libertés interdit de fait de collecter ce numéro dans le cadre d'une étude épidémiologique. Il est possible dans certaines circonstances de trouver des solutions à cette difficulté, mais elle constitue actuellement un obstacle formel pour la plupart des études. Les pouvoirs publics réfléchissent actuellement à une évolution des textes pour assouplir les conditions d'utilisation du NIR, et il faut espérer que ces efforts aboutiront prochainement.

Un très important travail est également nécessaire pour définir les procédures d'accès, de transmission sécurisée, de vérification de cohérence et de complétude, de maintien de l'intégrité des données. Les bases de données citées sont complexes, et leur utilisation dans des conditions compatibles avec les contraintes de qualité des études épidémiologiques nécessite des moyens lourds et des compétences spécialisées. Il est vraisemblable qu'aucune équipe d'épidémiologie en France ne dispose actuellement de ces ressources, et seule une structure de type « plateforme scientifique et technique » pourrait les développer et permettre à la communauté scientifique de bénéficier réellement des bases de données nationales d'origine administrative.

L'exemple d'autres pays montre que tout ceci est faisable, potentiellement très utile et pourrait contribuer au développement en France de grandes cohortes comparables à celles qui existent ailleurs.

RÉFÉRENCES

1. Thompson A. Thinking big: large-scale collaborative research in observational epidemiology. *Eur J Epidemiol*, 2009. 24: 727-31.
2. Darling GM, Davis SR, Johns JA. Hormone replacement therapy compared with simvastatin for postmenopausal women with hypercholesterolemia. *N Eng J Med* 1998; 338:64.
3. Collins, R. and UK Biobank Steering Committee. UK Biobank: Protocol for a large-scale prospective epidemiological resource. 2007, Manchester: UK Biobank Coordinating Centre.
4. Naess O et al. Cohort profile: cohort of Norway (CONOR). *Int J Epidemiol*. 2008 Jun;37(3):481-5.
5. Egan KM et al. Active and passive smoking in breast cancer: Prospective results from the Nurses' Health Study. *Epidemiology* 2002, 13, 138–145.
6. Goldberg M, Salamon R. État des forces épidémiologiques en France - L'épidémiologie humaine. In : Valleron AJ, Ed. Épidémiologie : conditions de son développement, et rôle des mathématiques. Rapport sur la Science et la Technologie n° 23, Comité RST de l'Académie des sciences. *Éditions EDP Sciences*, 2006.
7. Bulletin épidémiologique hebdomadaire. Les enquêtes Entred : des outils épidémiologiques et d'évaluation pour mieux comprendre et maîtriser le diabète. *BEH*, Numéro thématique 10 novembre 2009 / n°42-43.
8. Couris CM, Forêt Dodelin C, Rabilloud M et al. Sensibilité et spécificité de deux méthode d'identification des cancers du sein incidents dans les services spécialisés à partir des données médico-administratives. *Rev Epidemiol Sante Publique* 2004, 52, 151-60.
9. Couris CM, Colin C, Rabilloud M et al. Method of correction to assess the number of hospitalized incident breast cancer cases based on claims databases. *J Clin Epid*, 2002, 55 : 386-391.
10. Fender, P, Weill, A. Épidémiologie, santé publique et bases de données médico-tarifaires. *Rev. Epidemiol. Sante Publique*, 2004 ; 52: 113-117.
11. Moisan F, Gourlet V, Mazurie JL et al. Prediction model of Parkinson's disease based on antiparkinsonian drug claims. *Am J Epidemiol* 2011;174:354-363.