médecine/sciences 2012 ; 28 : 430-4

Cohortes épidémiologiques et bases de données d'origine administrative

Un rapprochement potentiellement fructueux

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

> Les cohortes épidémiologiques actuelles peuvent inclure des centaines de milliers de sujets qui sont suivis pendant des décennies. Pour la constitution et le suivi de ces mégacohortes, la France dispose d'un atout potentiel : les bases de données médicosociales nationales hospitalières, de l'Assurance maladie, des retraites et des causes de décès. Elles offrent de nombreux avantages : exhaustivité de la population, absence de perdus de vue pendant le suivi, données souvent fiables, appariement avec des enquêtes. Cependant, des problèmes de validité des données médicales se posent 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. <

[1] Inserm U1018, Plate-forme de recherche Cohortes épidémiologiques en population, Centre de recherche en épidémiologie et santé des populations, 16, avenue Paul Vaillant-Couturier, 94807, Villejuif, France ;
[2] Université de Versailles-Saint Quentin, UMRS 1018, France.
marcel.goldberg@inserm.fr

Les cohortes épidémiologiques en population : un besoin méconnu en France

Lorsqu'il s'agit de juger en termes de causalité du rôle sur la santé de facteurs de risque (ou d'interventions préventives), l'épidémiologie dispose de deux approches méthodologiques principales : (1) les études cas-témoins, dont le principe est de réunir un nombre adéquat de sujets souffrant de la maladie étudiée (les cas) et de sujets indemnes (les témoins), et de rechercher dans le passé l'exposition des deux groupes aux facteurs de risque d'intérêt ; (2) les études de cohorte.

Le principe des cohortes épidémiologiques

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 ou dizaines de milliers pour les plus importantes) habituellement recrutés en milieu médical. Les données recueillies sont très détaillées et incluent notamment des investigations biocliniques approfondies. Les secondes, établies en population générale, 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 (Agence nationale pour la recherche sur le sida et les hépatites virales) « *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* ».

Actuellement, l'épidémiologie fait face à la nécessité de développer des études d'une 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

m/s n° 4, vol. 28, avril 2012
DOI : 10.1051/medsci/2012284022

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].

La place des cohortes françaises dans le monde

Dans ce paysage, la France ne se distingue pas particulièrement par ses grandes réalisations. À titre d'illustration, on constate que les cohortes prospectives françaises se caractérisent par une taille relativement faible, aucune ne dépassant quelques dizaines de milliers de sujets, alors que certaines cohortes prospectives dans d'autres pays peuvent atteindre plusieurs centaines de milliers de sujets, voire plus. On peut citer en Grande-Bretagne la *Million Women Study* [2], le projet *UK Biobank* [3] qui a mis en place le suivi prospectif de 500 000 personnes, et la *Norvegian Mother and Child Cohort Study* qui a inclus 100 000 femmes à la 18e 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[1], qui doivent inclure et suivre plusieurs centaines de milliers 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. Ces registres sont largement ouverts aux chercheurs et permettent, 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 par divers moyens : 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 centre hospitalo-universitaire. Cependant, 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 projets du programme Investissements d'avenir, Très grandes infrastructures de recherche-Cohortes en 2009 et 2010). Cependant, les budgets qui ont été distribués sont très loin de répondre aux coûts véritables et très largement inférieurs aux financements des cohortes étrangères citées plus haut. Ces comparaisons montrent bien à quel point ces 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. Ainsi, 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 l'on veut que la France se dote d'outils épidémiologiques d'envergure comparables à ceux qui existent dans les pays de niveau scientifique proche, 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 données médico-administratives

Une grande partie des coûts des cohortes prospectives en population vient de la nécessité de suivre la trace des 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. En effet, il existe en France des systèmes d'information extrêmement puissants gérés par des organismes de protection médicosociale ou de gestion hospitalière 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 ont 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 obtenues par l'accès au Répertoire national d'identification des personnes

[1] www.helmholtz.de/en/research/health/the_latest_insights/insights_archive/who_stays_healthy/

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 informations diverses pouvant être utilisées dans des protocoles épidémiologiques.

Le Programme de médicalisation du système d'information (PMSI) 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 sont désormais réunies 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, Mutualité sociale agricole (MSA), Régime social des indépendants (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 la concernant dans les différentes sources qui alimentent le SNIIR-AM. Au total, le SNIIR-AM couvre la totalité de la population française et 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, etc.), incluant les employeurs et la catégorie socioprofessionnelle.

Un apport potentiel majeur pour l'épidémiologie

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 répondre à la majorité des questions posées par les épidémiologistes, elles offrent de nombreux avantages : une quasi exhaustivité de la population cible - et par conséquent une absence de biais de sélection et des effectifs immenses pour certaines analyses -, une quasi absence de perdus de vue pendant le suivi, et des données parfois plus fiables que celles obtenues par déclaration pour certaines informations (comme les consommations de soins ou la carrière professionnelle, par exemple). Couplées à 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 pendant une 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 pour illustrer les 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 concerne 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) [7]. Un autre exemple est l'étude nationale Entred qui concerne 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 au 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 [8].

Extraction de données concernant des sujets sélectionnés

Il est possible de retrouver, pour une personne sélectionnée, les enregistrements de données la concernant, sous réserve de disposer de certaines informations indispensables (voir 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 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 parfois de façon 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 [9, 10].

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 pour lesquelles un remboursement n'a pas été sollicité. La base des affections longue durée codées par des médecins reste une base de données à vocation médicosociale et ses limites sont connues : imprécision des diagnostics, absence d'exhaustivité des cas déclarés et risque de double déclaration [11].

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, etc.). Une voie prometteuse est le développement d'algorithmes incluant des données d'affections longue durée, 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 une sensibilité et une spécificité excellentes les patients souffrant d'une maladie de Parkinson [12].

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 que leur utilisation soit 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 (Institut national de la statistique et des études économiques) ou numéro de Sécurité sociale. Or la loi Informatique et libertés interdit la collecte de ce numéro dans le cadre d'une étude épidémiologique. Il est possible dans certaines circonstances de trouver des solutions à ce problème, mais il 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 travail très important est également nécessaire pour définir les procédures d'accès, de transmission sécurisée, de vérification de cohérence et de complétude, ainsi que 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. Seule une structure de type plate-forme 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. ◊

SUMMARY

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

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 and 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, and some legal and practical problems have to be solved for an optimal use. ◊

CONFLITS D'INTÉRÊTS

Les auteurs déclarent n'avoir aucun conflit d'intérêts concernant les données publiées dans cet article.

RÉFÉRENCES

1. Thompson A. Thinking big: large-scale collaborative research in observational epidemiology. *Eur J Epidemiol* 2009 ; 24 : 727-31.
2. The Million Women Study Collaborative Group. The million women study: design and characteristics of the study population [peerreviewed research]. http://breast-cancer-research.com/vol1no1/19aug99/research/1.
3. Collins R, UK Biobank Steering Committee. *UK biobank: protocol for a large-scale prospective epidemiological resource*. Manchester : UK Biobank Coordinating Centre, 2007.
4. Naess O, Sogaard AJ, Arnesen E, *et al.* Cohort profile: cohort of Norway (CONOR). *Int J Epidemiol* 2008 ; 37 : 481-5.
5. Egan KM, Stampfer MJ, Hunter D, *et al.* Active and passive smoking in breast cancer: prospective results from the Nurses' Health Study. *Epidemiology* 2002 ; 13 : 138-45.
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*. Paris : Éditions EDP Sciences, 2006.
7. Weill A, Païta M, Tuppin P, *et al.* Benfluorex and valvular heart disease: a cohort study of a million people with diabetes mellitus. *Pharmacoepidemiol Drug Saf* 2010 ; 19 : 1256-62.
8. Bulletin épidémiologique hebdomadaire. Les enquêtes Entred : des outils épidémiologiques et d'évaluation pour mieux comprendre et maîtriser le diabète. *BEH* 2009 ; n° thématique : 42-3.
9. Couris CM, Forêt Dodelin C, Rabilloud M, *et al.* Sensibilité et spécificité de deux méthodes d'identification des cancers du sein incidents dans les services spécialisés à partir des données médico-administratives. *Rev Epidemiol Sante Publ* 2004 ; 52 : 151-60.
10. 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 Epidemiol* 2002 ; 55 : 386-91.
11. Fender P, Weill A. Épidémiologie, santé publique et bases de données médico-tarifaires. *Rev Epidemiol Sante Publ* 2004 ; 52 : 113-7.
12. Moisan F, Gourlet V, Mazurie JL, *et al.* Prediction model of Parkinson's disease based on antiparkinsonian drug claims. *Am J Epidemiol* 2011 ; 174 : 354-63.

TIRÉS À PART

M. Goldberg