L'apport des bases de données d'origine administrative aux cohortes épidémiologiques : l'exemple de la cohorte Constances

Marie Zins et Marcel Goldberg
Inserm et Université de Versailles-Saint-Quentin

La France, contrairement à d'autres pays développés, ne dispose pas encore d'une cohorte de grande taille pour sa recherche épidémiologique. En revanche, notre pays dispose de bases de données administratives extrêmement riches. Exploiter ces gisements pour constituer à moindre coût de grandes cohortes est une voie de progrès prometteuse : c'est ce qui est entrepris dans la cohorte « Constances ». D'autres expériences pourront suivre, si les problèmes légaux et méthodologiques sont correctement pris en charge.

Les cohortes épidémiologiques en population : un besoin encore méconnu 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. Les cohortes en population générale s'intéressent essentiellement 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 »[1]. 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 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].

1. Une biobanque est une collection d'échantillons biologiques destinés à la recherche scientifique, en biologie (notamment en génomique) et en médecine.

Dans ce paysage, 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 en Suède, aux Pays-Bas, ou en Allemagne 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, 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, notamment du fait de 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, auto-questionnaires, examens médicaux, collecte de matériel biologique, etc. Ces coûts sont largement supérieurs aux budgets qu'il est possible de demander aux organismes nationaux de financement de la recherche 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 du fait de la quasi impossibilité de disposer de personnel stable et d'un niveau de qualification suffisant en l'absence de statut reconnu pour ce type d'activité dans les organismes publics de recherche. Pourtant la durée des projets est incompatible avec un trop fréquent renouvellement des personnels 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 d'un niveau scientifique équivalent, 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 « 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. qu'il s'agisse de l'inclusion et du suivi des sujets, ou de l'accès à des données concernant des événements de santé ou de vie socioprofessionnelle d'intérêt. On se restreindra ici à la description des deux principaux systèmes d'information de nature médicale et administrative.

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 réuni au sein du Système national d'information inter régimes de l'assurance maladie (SNIIRAM) qui concerne aussi bien la médecine de ville que les hospitalisations. 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 SNIIRAM. Au total, le SNIIRAM 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 en place un système permettant de collecter et traiter les données sociales issues de différents organismes et régimes gestionnaires des prestations sociales pour chaque individu 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 les cohortes

Dans un contexte épidémiologique, ces bases de données 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 par exemple). Couplées avec des données recueillies auprès des personnes, ces bases de données peuvent apporter des solutions satisfaisantes à

divers problèmes rencontrés par les cohortes : traçage des sujets au cours du suivi, 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), allégement des questionnaires.

… malgré certaines limites

Des problèmes de validité des données médicales se posent. Ainsi, l'utilisation du PMSI comme source d'information sur les pathologies s'avère délicate et ne peut reposer uniquement sur le diagnostic principal [5]. De plus, 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. Les ALD ont des limites connues : imprécision des diagnostics, absence d'exhaustivité des cas déclarés, risque de double déclaration [6].

Dans de nombreuses situations, il est donc nécessaire de mettre en place des procédures de validation des diagnostics extraits des bases de données : 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 SNIIRAM 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 [7].

L'exemple de la cohorte Constances www.constances.fr

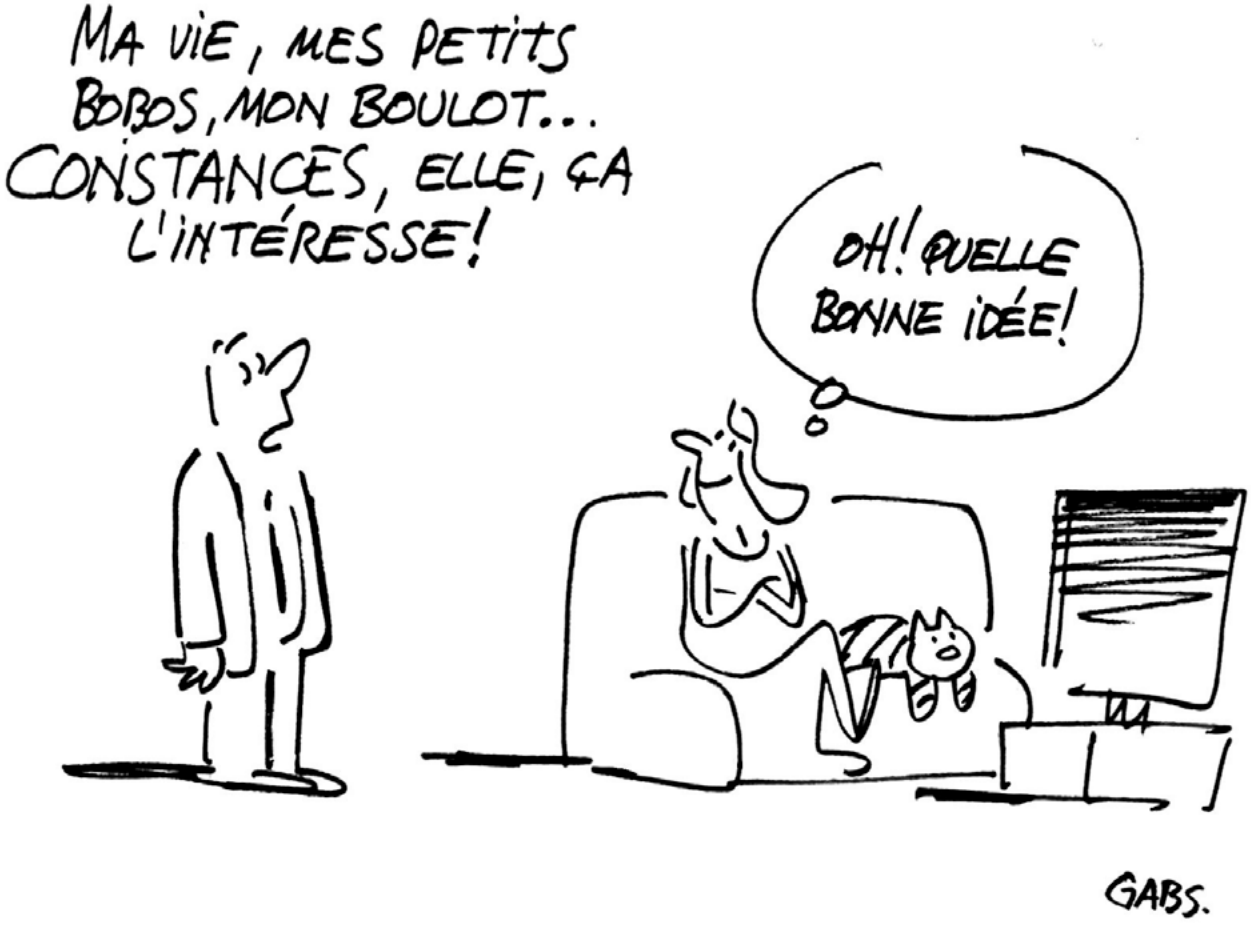
Récemment, grâce à un partenariat avec la Caisse nationale d'assurance maladie des travailleurs salariés (CNAMTS) et un important financement des Investissements d'avenir dans le cadre des Infrastructures nationales en biologie et santé, la cohorte Constances a pu être initialisée [8].

Constances est une importante cohorte épidémiologique destinée à fournir des informations à visée de santé publique et de contribuer au développement de la recherche épidémiologique en constituant une infrastructure largement accessible à la communauté scientifique.

Constances est un échantillon représentatif de la population couverte par le Régime général de Sécurité sociale (plus de 85 % de la population française) âgée de 18 à 69 ans, constitué par tirage au sort. L'effectif total prévu est de 200 000 sujets qui seront inclus sur une période de 5 ans ; le recrutement a commencé courant 2012 et actuellement (février 2014) environ 35 000 sujets sont déjà inclus et les données de 25 000 d'entre eux ont déjà été appariées avec succès avec le SNIIRAM et la Cnav. Sa structure est proportionnelle à la population-cible pour le sexe, l'âge et la catégorie sociale. Les personnes éligibles sont celles qui habitent dans 16 départements dont les Centre d'examens de santé (CES) participent à Constances. L'inclusion des participants se fait dans ces CES : les volontaires complètent un questionnaire concernant leur santé, leurs modes vie et un historique professionnel et bénéficient d'un examen de santé complet ; des prélèvements de sang et d'urine permettent de constituer une biobanque. Le suivi est « actif » : un questionnaire est complété chaque année et une invitation à revenir au CES tous les 5 ans pour un nouvel examen de santé est proposée. Il est également « passif » par appariement annuel avec les bases de données de la Cnav et du SNIIRAM ; le statut vital et les causes de décès sont également suivis dans les bases du CépiDc-Inserm. Les principales données recueillies à l'inclusion et durant le suivi concernent notamment la situation sociale et professionnelle, la santé (morbidité, capacités fonctionnelles physiques et cognitives), le recours aux soins, les comportements, l'exposition à des facteurs de risque professionnels et environnementaux.

L'apport des bases de données d'origine administrative est essentiel à toutes les étapes de la mise en place et du suivi de la cohorte. La constitution de l'échantillon repose sur les bases de données de la Cnav permettant un tirage au sort tenant compte des caractéristiques sociodémographiques et professionnelles ; la Cnav fournit également des données individuelles à l'inclusion et pendant le suivi. Le SNIIRAM fournit des données de santé et de consommation de soins exhaustives et très détaillées.

Outre l'accès à des données nombreuses, les bases de données administratives offrent également d'autres avantages. Elles garantissent la quasi-absence de perdus de vue pendant le suivi de la cohorte, même en l'absence de réponses aux questionnaires, ce qui est essentiel dans le contexte d'études longitudinales. Elles permettent également de contrôler les effets de sélection et les biais potentiels occasionnés par les effets de sélection à l'inclusion comme pendant le suivi (attrition). En effet, les personnes tirées au sort qui ne participent pas ou qui abandonnent la cohorte diffèrent des participants pour de nombreux paramètres liés à la santé et la position sociale. Pour tenir compte de ces différences une « cohorte contrôle» a été constituée, selon une procédure agréée par la CNIL, par tirage au sort d'un échantillon parmi les non-participants. Les mêmes données de la Cnav et du SNIIRAM sont extraites pour les deux cohortes (participants et non-participants), à l'inclusion et durant le suivi ; il est ainsi possible d'identifier les facteurs liés à la non-participation et de produire des estimations de prévalence de maladies et de facteurs de risque redressées pour ces facteurs par des méthodes de repondération.



Les perspectives

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 administratives repose sur le « Numéro d'inscription au répertoire » (NIR). Or la loi Informatique et libertés, qui exige un décret en Conseil d'État, rend pratiquement impossible la collecte de ce numéro dans le cadre d'une étude épidémiologique, ce qui constitue actuellement un obstacle insurmontable pour la plupart des études. Les pouvoirs publics réfléchissent actuellement à une évolution des textes pour assouplir les conditions d'utilisation du NIR.

Par ailleurs, un très important travail méthodologique et technique est nécessaire en raison de la complexité et du volume de ces bases de données. Leur utilisation dans des conditions compatibles avec les contraintes de qualité des études épidémiologiques nécessite des moyens lourds et des compétences spécialisées. 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] 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.
[6] Fender, P, Weill, A. Épidémiologie, santé publique et bases de données médico-tarifaires. Rev. Epidemiol. Sante Publique, 2004 ; 52: 113-117.
[7] 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.
[8] Zins M, Bonenfant S, Carton M, Coeuret-Pellicer M, Guéguen A, Gourmelen J, Nachtigal M, Ozguler A, Quesnot A, Ribet C, Rodrigues G, Serrano A, Sitta R, Brigand A, Henny J, Goldberg M. The CONSTANCES Cohort: an Open Epidemiological Laboratory. BMC Public Health 2010; 10:479.