Les nouvelles « méga-cohortes » en population en Europe

Marie Zins
Marcel Goldberg
Inserm U1018, plate-forme de recherche Cohortes épidémiologiques en population – Centre de recherche en épidémiologie et santé des populations, université de Versailles-Saint-Quentin

Les premières grandes cohortes épidémiologiques en population ont été mises en place après la Seconde Guerre mondiale aux États-Unis et en Grande-Bretagne. Elles ont été fondatrices de l'épidémiologie moderne, et beaucoup sont toujours actives et continuent d'apporter des données précieuses dans des domaines variés. Aux États-Unis, on peut citer la cohorte Framingham, initiée en 1949 [35], la Nurses' Health Study mise en place en 1976 chez des infirmières [19], ou en Grande-Bretagne la British Doctors' Study suivie depuis 1951 [16], les cohortes Whitehall I et Whitehall II mises en place respectivement en 1967-1969 et en 1985 [34], ou encore les cohortes de naissance inaugurées en 1946 avec la Birth Cohort 1946. En France, l'Étude prospective parisienne consacrée à l'étude des risques cardio-vasculaires a été une cohorte pionnière, initiée dès 1967 [17]. Quelques-unes de ces cohortes « historiques » sont décrites dans les articles *Les cohortes épidémiologiques dans la recherche internationale*, p. 31 et *Les cohortes « historiques » en France*, p. 37.

La nécessité de cohortes prospectives de très grande taille

Aujourd'hui, les cohortes en population sont devenues un outil scientifique relativement courant, et il est impossible de recenser toutes les cohortes existant à l'échelle internationale. Ainsi, on comptait en 2004 pas moins de 47 cohortes épidémiologiques ayant contribué de façon significative à l'établissement des risques de divers cancers dus au tabac, c'est-à-dire ayant un recul et une puissance suffisants pour mettre en évidence de tels risques, y compris pour des cancers peu fréquents.

Cependant, la nature des questions scientifiques auxquelles on demande aux cohortes d'apporter des réponses évolue. Actuellement, la recherche sur les causes des maladies de nature environnementale, professionnelle, sociale, nutritionnelle, biologique et génétique, ou en pharmacoépidémiologie, concerne de plus en plus des risques de faible ampleur, donc difficiles à mettre en évidence : effets potentiellement cancérigènes des téléphones portables, des faibles doses de rayonnements ionisants, rôle de polymorphismes génétiques vis-à-vis de maladies multifactorielles, etc.

Dans ce contexte scientifique, des cohortes de très grande envergure, avec un suivi à long terme et un phénotypage (caractérisation des maladies) de haute qualité, sont nécessaires pour assurer une puissance statistique suffisante permettant de mieux comprendre le rôle des divers facteurs personnels et environnementaux et leur interaction avec des caractères génétiques complexes. Par exemple, des associations établies entre des polymorphismes génétiques et des maladies chroniques montrent des risques relatifs typiquement compris entre 1,1 et 1,4, et la mise en évidence de façon fiable de tels effets exige de très vastes ensembles de données. Ainsi, les études cas-témoins doivent réunir des milliers de cas, même pour les situations les plus simples ; lorsque la question de recherche concerne l'étude des interactions gène-environnement et gène-gène et l'analyse approfondie de relations de cause à effet, des dizaines de milliers de cas sont souvent nécessaires. Des dizaines de milliers de sujets peuvent aussi être nécessaires pour étudier un phénotype quantitatif (pression artérielle par exemple), parce que les effets d'origine génétique peuvent être très faibles [3].

Les « méga-cohortes » en Europe

C'est dans ce contexte qu'on voit se mettre en place, depuis les années 2000, une nouvelle génération de « méga-cohortes » en population, et on estime qu'actuellement plus de 100 cohortes prospectives de population de grande dimension sont à divers stades de développement et de réalisation dans le monde, dont certaines dans des pays européens dont on présente les principales.

Certaines cohortes sont déjà en place en Europe

On peut ainsi citer en Grande-Bretagne la Million Women Study (http://www.millionwomenstudy.org/introduction/), qui a inclus entre 1996 et 2001 plus d'un million de femmes âgées de 50 ans et plus, ou le projet UK Biobank, qui a inclus, de 2006 à 2010, 500 000 personnes âgées de 40 à 69 ans (http://www.ukbiobank.ac.uk/). En Norvège (pays de 4,5 millions d'habitants, soit 13 fois moins peuplé que la France), la cohorte CONOR (Cohort of Norway) suit 200 000 adultes, et la cohorte MoBa (Norwegian Mother and Child Cohort Study) a inclus 270 000 mères, pères et leurs enfants (http://www.fhi.no/eway/?pid=238). La cohorte européenne EPIC (European Prospective Investigation into Cancer and Nutrition) réunit 520 000 participants âgés de 20 ans et plus dans 10 pays (Danemark, France, Allemagne, Grèce, Italie, Pays-Bas, Norvège, Espagne, Suède et Royaume-Uni) ; elle a été mise en place entre 1993 et 1999 avec 7 pays participants, et le Danemark, la Norvège et la Suède ont rejoint ultérieurement la cohorte (http://epic.iarc.fr/).

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

D'autres cohortes de très grande dimension sont actuellement à un stade de mise en place ou de préparation avancée

Suède : la cohorte LifeGene prévoit d'inclure 500 000 sujets âgés de 0 à 45 ans, grâce à un échantillonnage à deux niveau x : échantillon aléatoire de personnes de 18 à 45 ans, puis inclusion de leurs enfants et des membres adultes de leur famille ; les inclusions ont commencé en 2010 (http://www.lifegene.se/In-english/).

Pays-Bas : la cohorte LifeLines repose également sur un échantillonnage à deux niveaux : échantillon aléatoire de personnes âgées de 25 à 50 ans, puis inclusion des membres de leur famille (partenaires, parents et enfants). Plus de 60 000 sujets sont déjà inclus, et au total ce sont 165 000 participants qui sont attendus (http://lifelines.nl/).

France : la cohorte Constances vise à réunir un échantillon représentatif de 200 000 participants âgés de 18 à 69 ans au moment de l'inclusion qui a lieu dans des centres d'examens de santé de la Sécurité sociale ; les premières inclusions ont commencé début 2012 (http://www.constances.fr/).

Allemagne : la cohorte GeNatCo (German National Cohort) prévoit d'inclure un échantillon de 200 000 participants âgés de 20 à 70 ans ; le début des inclusions est prévu en 2013 (http://www.nationale-kohorte.de/informationen_en.html).

Malgré certaines différences, ces grandes cohortes en population présentent beaucoup de caractéristiques communes.

Elles sont centrées sur des thèmes d'intérêt général : les cohortes d'adultes s'intéressent particulièrement aux maladies chroniques et dégénératives fréquentes : cancer, maladies cardio-vasculaires et métaboliques, psychiatrie, démences, etc. Les « cohortes de naissance », qui suivent des enfants depuis la vie *in utero*, sont centrées notamment sur les facteurs liés aux différents stades du développement. Dans tous les cas, l'étude de la susceptibilité génétique à développer des maladies (ou en être protégé) est très présente (voire essentielle pour certaines cohortes), et le développement de biomarqueurs de détection précoce de pathologies est privilégié. De nombreux facteurs sont pris en compte, qu'ils soient de nature personnelle et familiale, environnementale et professionnelle, sociale, biologique et physiologique, psychologique, comportementale (alimentation, exercice physique, tabac, alcool…). L'analyse des inégalités sociales et territoriales de santé et de leurs déterminants est également présente dans plusieurs cohortes, de même que celle des consommations de soins et leur coût.

Elles incluent des recueils de données multiples reposant sur des techniques diversifiées : questionnaires, entretiens, examen médical, appariement à des bases de données nationales, et plusieurs cohortes recueillent des données de questionnaire par Internet. Certaines prévoient des investigations complémentaires spécialisées sur des sous-ensembles de participants : ainsi la cohorte allemande GeNatCo prévoit la réalisation d'IRM corps entier pour 40 000 sujets. Enfin, toutes les cohortes mettent en place des biobanques associées, destinées à stocker des échantillons biologiques divers (ADN, sérum, cellules, selles…) pendant une très longue durée pour permettre ultérieurement des analyses biologiques de marqueurs nouveaux.

Du fait de la disponibilité de données nombreuses et diversifiées sur de très importants échantillons, la plupart de ces cohortes sont gérées comme des infrastructures pratiquant une large ouverture vers la communauté de recherche, notamment sous forme d'appels publics à projets permettant ainsi à des chercheurs extérieurs de bénéficier d'un accès aux données collectées.

Enfin, certaines cohortes, comme la cohorte française Constances, sont constituées d'échantillons représentatifs de la population générale, permettant ainsi la production d'indicateurs de santé destinés aux autorités de santé publique.

La nécessité de la mise en commun de données de différentes cohortes

Lorsqu'il s'agit d'analyser les maladies complexes les plus fréquentes pour étudier des relations étiologiques complexes ou des caractères quantitatifs liés à la maladie, même les plus grandes études ne génèrent pas suffisamment de cas. Il devient alors souvent nécessaire de mettre en commun les données de plusieurs grandes cohortes, comme c'est devenu la règle pour les études de génomique dans le cadre de consortiums de recherche. La mise en commun de données à grande échelle ne concerne évidemment pas uniquement les études de génétique, mais toute l'épidémiologie est concernée pour des raisons de puissance statistique. Les études internationales comparatives sur les services de santé, les déterminants sociaux de la santé ou les habitudes alimentaires, réunissant des données de cohortes de plusieurs pays, sont également indispensables pour réduire les biais potentiels découlant de l'accès à des ensembles de données restreints et spécifiques d'une population.

Ces dernières années se sont mises en place des collaborations internationales destinées à faciliter les mises en commun de données de cohortes en population, notamment grâce à une harmonisation aussi étroite que possible des données recueillies. En effet, pour que la mutualisation des données soit possible et éviter de fastidieux et coûteux recodages des données et une perte d'information, il faut que celles-ci aient une même définition, que les variables enregistrées soient identiques, ou du moins qu'on puisse établir aisément des correspondances. Ainsi à l'échelle internationale, le consortium P^3G (Public Population Project in Genomics) promeut l'utilisation d'outils (questionnaires, échelles diverses, etc.) qu'il met à disposition de la communauté scientifique. À l'échelle européenne, le consortium BBMRI (Biobanking and Biomolecular

Resources Research Infrastructure) vise à harmoniser les méthodes de collecte et de conservation de matériel biologique, cette initiative étant relayée en France par l'Infrastructure Biobanques (lire *Coordination et partage de données de cohortes*, p. 29). Récemment s'est constitué au sein du consortium BBMRI, le projet LPC (Large Prospective Cohorts) qui associe une vingtaine de grandes cohortes en population provenant de 13 pays européens (la France y est présente par les cohortes Gazel et Constances), réunissant plus de 2,5 millions de sujets.

Ouverture à la communauté de recherche

Une autre forte tendance accompagne le développement des méga-cohortes : l'ouverture systématique des bases de données des cohortes à la communauté de recherche. En effet, les investissements nécessaires pour la construction et la maintenance des très grandes cohortes sont tels qu'il n'est plus envisageable d'en limiter l'accès aux équipes qui les mettent en œuvre. Ces méga-cohortes sont maintenant considérées comme des infrastructures de recherche qui, à l'instar d'autres très grands instruments de recherche, comme les télescopes ou les accélérateurs de particules, sont conçues pour répondre à de multiples questions provenant d'équipes diversifiées. Ainsi, en France, la cohorte Constances a été récemment reconnue comme une « *Infrastructure nationale biologie santé* » par le ministère de la Recherche et de l'Enseignement supérieur.

De nombreuses institutions de financement de la recherche en santé ont récemment exprimé leur volonté de promouvoir le partage de données, tout en prenant en compte les aspects scientifiques, éthiques et juridiques, ainsi que la nécessaire valorisation scientifique des équipes et institutions qui ont mis en place et qui gèrent les cohortes [45].