Les cohortes généralistes en population. L'exemple des cohortes Gazel et Constances

Marcel Goldberg, Marie Zins

Bull. Acad. Natle Méd., 2013, *197*, nº 2, 299-313, séance du 19 février 2013

COMMUNICATION

Les cohortes généralistes en population. L'exemple des cohortes Gazel et Constances

MOTS-CLÉS : ÉPIDÉMIOLOGIE. ÉTUDES DE COHORTE. CARACTÉRISTIQUES DE LA POPULATION

Population-based cohorts. Example of the Gazel and Constances cohorts

KEY-WORDS (Index medicus): EPIDEMIOLOGY. COHORT STUDIES. POPULATION CHARACTERISTICS

Les auteurs déclarent ne pas avoir de lien d'intérêt en relation avec le contenu de cet article.

Marcel GOLDBERG * et Marie ZINS **

RÉSUMÉ

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, des échantillons parfois très vastes, pour lesquels sont recueillies de façon prospective des données personnelles, de mode de vie, sociales, professionnelles et environnementales, et qui s'accompagnent de biobanques. Certaines sont généralistes, et concernent un vaste champ de pathologies et de facteurs de risque. Deux exemples sont présentés : la cohorte Gazel, suivie depuis près de 25 ans, constituée de 20 000 personnes âgées de 35 à 50 ans à l'inclusion, qui a déjà fait l'objet d'environ 200 publications sur des thèmes très divers ; la cohorte Constances mise en place en 2012 vise à inclure un échantillon représentatif de 200 000 adultes de 18 à 69 ans.

SUMMARY

Population-based cohorts focus on the causes of diseases, especially multifactorial diseases. Some are very large, and prospectively collect personal, lifestyle, occupational and environ-

* Inserm, Centre de recherche en Épidémiologie et Santé des Populations, U1018, Plateforme de recherche Cohortes en population, Hôpital Paul Brousse, Bât 15/16, Porte D, secteur violet — 16 avenue Paul Vaillant Couturier — 94807 Villejuif cedex ; e-mail : marcel.goldberg@inserm.fr

** Université Versailles St-Quentin en Yvelines

Tirés à part : Professeur Marcel GOLDBERG, même adresse

Article reçu le 20 janvier 2013, accepté le 18 février 2013

mental data over several decades. All include biobanks. ''Generalist'' cohorts cover a large field of diseases and risk factors. Two examples are presented here. The Gazel cohort was composed of 20 000 subjects aged 35-50 at enrolment and followed-up for 25 years, resulting in about 200 publications. The Constances cohort, created in 2012, aims to include a representative sample of 200 000 adults aged 18-69 at enrolment.

INTRODUCTION

Qu'est-ce qu'une cohorte épidémiologique ?

Le principe d'une cohorte épidémiologique est le suivi longitudinal d'un groupe de sujets. Selon les objectifs, la durée d'observation des sujets et les données individuelles recueillies de façon prospective diffèrent. Une distinction majeure doit être faite entre cohortes de malades souffrant d'une pathologie particulière, et cohortes en population générale.

Les cohortes de malades, dont **l'objectif est d'étudier l'évolution d'une maladie**, incluent un nombre souvent restreint de sujets (quelques milliers pour les plus importantes) recrutés en milieu médical, et les données recueillies sont très détaillées, incluant notamment des investigations biocliniques approfondies. Une illustration de l'apport d'un suivi longitudinal pour la connaissance de l'histoire naturelle des maladies est donnée par la Figure 1 : elle montre les principales phases de l'évolution de l'infection par le VIH et la relation entre la charge virale et le nombre de lymphocytes T_4 au cours du temps [1].

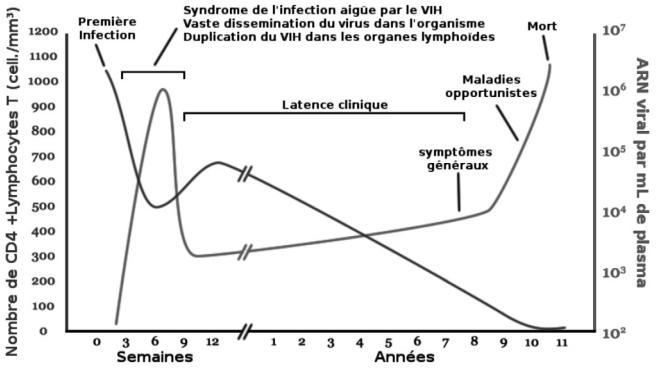

FIG. 1. — Phases de l'évolution de l'infection par le VIH

Ces cohortes sont un outil précieux, voire indispensables dans de nombreuses circonstances, pour la recherche clinique, mais elles ne prennent en compte que des personnes malades.

Les cohortes en population générale sont celles qui font l'objet de cet article. **Elles 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, des échantillons parfois très vastes, pour lesquels sont recueillies de façon prospective des données personnelles, de mode de vie, sociales, professionnelles et environnementales, et s'accompagnent de biobanques.

Le principe d'une cohorte à visée étiologique est résumé par la Figure 2.

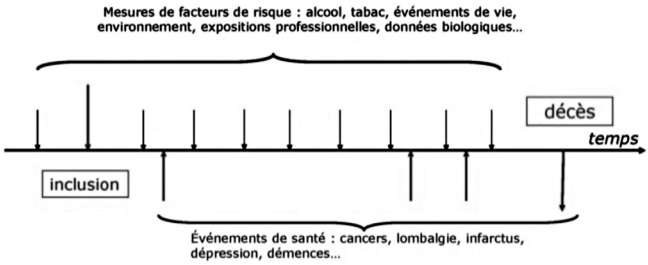

Fig. 2. — Schéma de base d'une cohorte épidémiologique

On choisit un groupe de sujets qui sont indemnes de la (des) maladie(s) étudiée(s) au début de la période d'observation. Tout au long du suivi de la cohorte, on recueille des données concernant les sujets : exposition à des facteurs de risque et incidence des maladies, et à la fin de la période d'étude, on dispose de toutes les données utiles pour calculer les risques associés aux expositions.

Certaines de ces cohortes sont « généralistes », et se caractérisent par une couverture large de problèmes de santé et de déterminants. Elles constituent alors de véritables infrastructures de recherche et de santé publique, comme le montrent les exemples des cohortes Gazel et Constances décrites plus loin.

Pourquoi des cohortes généralistes en population ?

Les études de cohorte sont celles qui permettent de proposer les meilleures conditions pour juger en termes de causalité du rôle de facteurs de risque, en permettant de prendre en compte les évolutions temporelles et les interactions entre facteurs.

Sur le plan méthodologique, les avantages principaux des cohortes sont la possibilité d'analyses longitudinales permettant de tenir compte au mieux de phénomènes liés au temps, notamment de la séquence temporelle exposition-effet. Il est ainsi possible de modéliser l'enchaînement et les interactions des différents facteurs relatifs aux conditions de vie, à l'environnement et à l'état de santé. Par ailleurs, les données d'exposition étant recueillies avant la survenue des effets, on évite certains biais

potentiels des études rétrospectives. Au total, les études de cohorte sont celles qui permettent théoriquement de proposer les meilleures conditions pour juger en termes de causalité du rôle sur la santé de facteurs de risque.

Limites et difficultés

Ainsi présentées, les cohortes en population semblent être l'instrument idéal qui répond à tous les besoins de recherche. Elles ont cependant des limites et leur mise en œuvre n'est pas sans difficultés diverses.

Puissance statistique et précision

Rappelons que pour l'estimation de la fréquence d'un phénomène (prévalence ou incidence), l'effectif de l'échantillon à observer pour une précision donnée dépend de la fréquence du phénomène dans la population. Pour l'estimation d'une mesure d'association entre exposition à un facteur de risque et une maladie, l'effectif de l'échantillon permettant de mettre en évidence une association avec une « puissance statistique » donnée dépend de l'incidence de la maladie dans la population non exposée, de la valeur supposée du risque relatif, et de la fréquence du facteur de risque dans la population étudiée. D'une façon générale, plus les phénomènes d'intérêt (maladies, expositions) sont rares, plus les associations facteur de risque/maladie sont faibles, et plus l'effectif doit être important pour une précision ou une puissance données. Dans certaines situations, il faudrait réunir des effectifs immenses pour répondre à des questions d'intérêt. Par exemple, si on s'intéresse à l'effet des pesticides sur le risque de myélome multiple, et si on fait l'hypothèse d'un risque 2 fois plus élevé chez les sujets exposés, l'effectif minimum nécessaire pour observer cette augmentation du risque après 6 ans de suivi est de 1 111 000 sujets ; 10 ans après, il est de 520 000. On voit bien que de façon réaliste les cohortes prospectives ne peuvent pas répondre à certaines questions, et que d'autres approches, notamment les études de type cas-témoins, resteront indispensables.

Effets de sélection et biais

Un biais est une erreur qui entraîne une différence systématique entre la véritable valeur d'un paramètre d'intérêt (l'incidence d'une maladie, une mesure d'association entre une maladie et un facteur de risque) et le paramètre qui est estimé par l'étude.

Une des sources majeures de biais dans les études épidémiologiques provient des effets de sélection, qui surviennent lors du recrutement ou du suivi des sujets. Or, dans la plupart des cohortes, la participation des sujets repose sur le volontariat, et il existe fréquemment des effets de sélection qui peuvent intervenir lors de la constitution de la cohorte et au long du suivi de celle-ci (attrition) [2]. Pour un objectif étiologique où on cherche à étudier les relations entre exposition à des facteurs de risque et survenue de maladies, ceci n'est généralement pas à l'origine de biais. En effet, la relation exposition — maladie n'est *a priori* pas différente entre les

sujets volontaires et ceux qui ne le sont pas. Une des raisons est qu'au moment de l'inclusion, tous sont indemnes des maladies qui seront analysées, seuls les cas incidents pendant la période de suivi étant pris en compte : des conditions très particulières seraient en effet nécessaires pour entraîner un biais dans la mise en évidence d'une relation entre une exposition et une maladie. Ainsi, pour analyser les effets du tabac sur le risque de cancer, il n'est pas nécessaire d'observer un échantillon représentatif de la population, mais de disposer d'effectifs suffisants de non-fumeurs et de fumeurs parmi lesquels le niveau d'exposition est contrasté : en effet, sur la base des connaissances actuelles, il est très vraisemblable que les mécanismes physiopathologiques et biologiques de la cancérogenèse liée au tabac sont identiques dans un échantillon de volontaires et dans l'ensemble de la population. Les effets de sélection dus au volontariat de la participation ne génèrent donc *a priori* pas de biais, ou seulement des biais minimes, lorsqu'il s'agit de comprendre comment les expositions à des facteurs de risque, les caractéristiques professionnelles et sociales, etc., influencent l'état de santé et peuvent être à l'origine de pathologies. Le problème de l'attrition au cours du suivi peut par contre être à l'origine de biais importants, car la probabilité de ne plus être suivi diffère souvent entre ceux qui sont ou ne sont pas devenus malades [3].

LA COHORTE GAZEL [4-5]

En 1989, a été mis en place un suivi épidémiologique d'une cohorte de 20 625 volontaires (15 011 hommes et 5 614 femmes) composée d'agents d'EDF-GDF âgés de 35 à 50 ans. L'objectif de Gazel était de constituer une infrastructure ouverte à la communauté scientifique, destinée à être le support d'études portant sur des thèmes diversifiés. Il s'agit donc d'une cohorte généraliste, qui n'est pas centrée sur une pathologie ou un facteur de risque spécifiques.

L'échantillon est diversifié sur le plan socio-économique et professionnel, ainsi qu'au plan géographique, les sujets vivant dans toute la France aussi bien en milieu rural qu'en milieu urbanisé, et les modes de vie ainsi que la répartition des problèmes de santé sont très voisins de ceux de la population générale française.

Les données qui font l'objet d'un recueil systématique pour toute la cohorte concernent diverses dimensions et sont recueillies auprès de différentes sources : autoquestionnaire annuel (morbidité incidente, comportements, échelles de santé mentale, variables professionnelles, personnelles et familiales, etc.) ; service du personnel d'EDF-GDF (postes de travail, situation socioéconomique) ; régime particulier de Sécurité Sociale d'EDF-GDF (absence pour raisons de santé, registres des cancers et des cardiopathies ischémiques) ; médecine du travail (conditions de travail et expositions professionnelles) ; mutuelles complémentaires d'EDF-GDF (remboursement de soins) ; bilans de santé dans les Centres d'examens de santé de la Sécurité sociale (dans ce cadre, une biobanque — sérum et ADN — a été constituée) ; causes médicales de décès.

La participation des volontaires est excellente : fin 2011 (soit après 23 ans de suivi) seuls 495 sujets (4,3 %) n'ont plus renvoyés le questionnaire annuel. Le nombre de vrais « perdus de vue » est inférieur à 1 % (personnes ayant demandé à sortir de la cohorte, ou qui ont quitté l'entreprise).

Actuellement, une cinquantaine de projets de recherche portant sur des thèmes très diversifiés ont été mis en place dans cette cohorte [6]. Des problèmes de santé aussi différents que la migraine, l'ostéoporose post-ménopausique, la pathologie cardiovasculaire ischémique, la dépression, les troubles musculo-squelettiques, l'incontinence urinaire, les accidents de circulation, les troubles cognitifs, font l'objet de projets de recherche. Des facteurs de risque comportementaux (alcool, tabac, par exemple), sociaux (support social, inégalités sociales de santé), psychologiques, professionnels (expositions chimiques, facteurs biomécaniques et psychosociaux), médicaux (consommations de médicaments et traitements) sont pris en compte. Certains de ces projets s'accompagnent du recueil de données spécifiques en complément de celles qui concernent l'ensemble de la cohorte et sont mises à disposition de chaque projet, permettant ainsi d'enrichir continuellement la base de données, grâce aux échanges de données entre équipes.

Au total, plus d'une trentaine d'équipes différentes appartenant à des structures de recherche diverses (Inserm, universités...) françaises et étrangères (Allemagne, Belgique, Canada, Danemark, Grande-Bretagne, Suède, Finlande, USA) réalisent des recherches associées à la cohorte Gazel. Celles-ci se déroulent dans le respect des règles fixées dans la charte de la cohorte : sélection des projets présentés par un Comité Scientifique, accord des instances compétentes (CNIL, CPP le cas échéant), présentation et débat pluridisciplinaire des résultats lors des journées scientifiques de la cohorte Gazel et mise en commun des données recueillies lors des enquêtes complémentaires.

Fin 2012, les sujets de Gazel étaient âgés de 59 à 74 ans : c'est pourquoi la plupart des recherches portent aujourd'hui sur divers aspects du vieillissement, prenant avantage du fait que les données sont recueillies sur les participants depuis l'âge adulte, soit beaucoup plus tôt que la plupart des cohortes de personnes âgées existantes, ce qui permet l'étude de phénomènes de vieillissement précoce.

Au total les travaux de la cohorte Gazel ont déjà été à l'origine d'environ 200 articles dans des revues internationales couvrant des domaines très variés. La production scientifique issue de Gazel augmente rapidement en raison du recul qui devient maintenant important : ceci permet aujourd'hui des analyses épidémiologiques de plus en plus puissantes, en parallèle à l'accumulation les données.

LA COHORTE CONSTANCES [7-8]

Objectifs de Constances

Malgré l'intérêt de Gazel et l'importance des travaux qui se poursuivent, cette cohorte présente diverses limites : effectif insuffisant pour l'étude de phénomènes peu fréquents, structure d'âge restreinte et absence de sujets jeunes, population issue d'une entreprise publique disposant d'un statut garantissant l'emploi, absence de sujets non français. C'est pourquoi notre équipe a entrepris la mise en place d'une nouvelle cohorte généraliste, dont les objectifs sont voisins, mais dont les caractéristiques permettent de s'affranchir de la plupart des limites de Gazel.

L'objectif du projet Constances est en effet de mettre en œuvre une très vaste cohorte épidémiologique destinée à fournir des informations à visée de santé publique et de contribuer au développement de la recherche épidémiologique. Réalisé dans le cadre d'un partenariat avec la Caisse nationale d'assurance maladie des travailleurs salariés (CNAMTS), cette cohorte a vocation à constituer une infrastructure largement accessible à la communauté de santé publique et de recherche. La mise en place d'une cohorte dont l'effectif, la qualité et la diversité des données se compareront aux plus importantes cohortes existant à l'échelle internationale, doit permettre de constituer un puissant outil pour la recherche épidémiologique en France.

Constances est une infrastructure de recherche, comme un télescope ou un accélérateur de particules, par exemple, qui ne sont pas construits pour répondre à une question de recherche spécifique, mais qui sont conçus pour aider à analyser une large gamme de problèmes scientifiques, et qui sont accessibles à la communauté des chercheurs spécialisés. Constances a été également conçue comme un outil venant en appui des objectifs de santé publique et de surveillance de l'Assurance maladie et de l'État, par le caractère particulièrement complet du dispositif de suivi et de recueil d'informations très diversifiées auprès d'un large échantillon représentatif de la population adulte.

Éléments essentiels du protocole

L'inclusion des sujets

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 de volontaires tirés au sort. L'effectif total prévu est de 200 000 sujets, et sa structure est proportionnelle à la population pour le sexe, l'âge et la catégorie sociale. Les personnes éligibles sont celles qui habitent dans les départements dont les Centre d'examens de santé de la Sécurité sociale (CES) participent à Constances. Ces CES, au nombre de 17, sont répartis dans les régions françaises, comme le montre la carte ci-dessous.

Les 17 Centres d'examens de santé Constances

La sélection des personnes éligibles est réalisée par la Caisse nationale d'assurance vieillesse (Cnav), qui tire au sort dans ses bases de données un échantillon de personnes correspondant aux critères de sexe, d'âge et de catégorie sociale définis. L'inclusion des participants est prévue sur une période de 5 ans à partir de 2012.

Les personnes tirées au sort reçoivent un courrier présentant le projet Constances et un coupon-réponse permettant de donner leur accord de principe. Les personnes ayant donné leur accord sont convoquées dans leur CES par un courrier incluant un questionnaire à compléter à domicile concernant leur santé, leurs modes vie et un historique professionnel. Les volontaires bénéficient dans leur CES d'un examen de santé et remplissent des questionnaires complémentaires (expositions professionnelles, questionnaire de santé pour les femmes).

Les principales données recueillies pendant l'examen et d'autres sources sont les suivantes :

— Données de santé : antécédents personnels et familiaux, échelles de santé et de qualité de vie, pathologies déclarées, diagnostic des affections de longue durée (ALD) et des hospitalisations, absence au travail, handicaps, limitations, incapacités et traumatismes, cause médicale de décès, comportements de santé (tabac, alcool, alimentation, activité physique, cannabis, orientation sexuelle), problèmes de santé spécifiques des femmes.
— Recours aux soins et prise en charge : professionnels de santé, médicaments, dispositifs médicaux, biologie, hospitalisations.
— Examen de santé : poids, taille, rapport taille-hanches, tension artérielle, fréquence cardiaque, vision, audition, spirométrie, biologie. Pour les personnes âgées de 45 ans et plus, tests des capacités fonctionnelles physiques et cognitives.

— Caractéristiques sociodémographiques : situation et activité professionnelle, niveau d'études, revenus, situation matrimoniale, composition du ménage, conditions de vie matérielles.
— Facteurs professionnels : histoire professionnelle, expositions professionnelles à des agents chimiques, physiques et biologiques, contraintes biomécaniques et organisationnelles, stress au travail.
— Biobanque : à l'occasion de l'examen de santé, des échantillons sanguins et d'urine sont collectés et conservés pour utilisation ultérieure.

Le suivi des participants

Il est prévu de suivre les sujets de la cohorte de deux façons complémentaires.

— Suivi « actif » : autoquestionnaire annuel pour suivre l'évolution de l'état de santé, de la situation socio-économique et professionnelle, de l'environnement familial, social et de lieu de vie, des facteurs de risque personnels et environnementaux. Une invitation à revenir au CES tous les 5 ans pour un nouvel examen de santé sera proposée.
— Suivi « passif » d'événements socioprofessionnels et de données de santé : grâce à l'appariement régulier de la cohorte avec les bases médico-administratives nationales, les principaux événements socioprofessionnels sont régulièrement extraits des bases de données de la Cnav. Des données de santé sont également extraites des bases de données de l'Assurance maladie et du PMSI, ainsi que le statut vital et les causes de décès.

Contrôle de qualité et validation des événements de santé

Un contrôle de qualité des données recueillies dans les CES a été mis en place ; il comporte notamment des visites régulières sur site d'attachés de recherche épidémiologique. Une des difficultés majeures des cohortes prospectives est l'identification de la survenue de pathologie durant le suivi. C'est pourquoi une attention particulière est portée aux diagnostics extraits des bases de données (ALD, PMSI), dont la validité doit être contrôlée ; à cet effet, des procédures de validation systématique ont été mises en place (retour au médecin, au dossier hospitalier, etc.).

Confidentialité des données

Des procédures sécurisées complexes, incluant le recours à un tiers de confiance pour gérer les coordonnées des participants, ont été élaborées pour garantir la confidentialité à toutes les étapes : recueil, transmission, stockage et utilisation des données. L'ensemble de ces procédures a été autorisé par la Commission nationale de l'informatique et des libertés.

Les recherches dans Constances

Constances est une infrastructure de recherche ouverte à la communauté scientifique. Une Charte a été établie précisant les conditions d'utilisation de la cohorte,

et un appel d'offre en direction des équipes scientifiques françaises et internationales permet aux chercheurs qui souhaitent bénéficier de la cohorte pour leurs propres travaux de proposer des projets, qui seront examinés par un Comité scientifique international.

Actuellement, plus de 40 déclarations d'intention de recherches ont été proposées par une trentaine d'équipes françaises et une dizaine d'équipes internationales, portant sur des thèmes diversifiés : pathologies spécifiques (diabète, cancer, maladie rénale chronique, pathologie respiratoire chronique, ostéo-articulaire), états de santé (vieillissement, fonctionnement physique et cognitif, troubles du sommeil, hypertension artérielle, ...), comportements (alimentation, activité physique), facteurs psychologiques, facteurs de risque professionnels et environnementaux, inégalités sociales de santé.

Collaborations scientifiques

Des collaborations ont été mises en place avec d'autres cohortes en population, tant en France (cohorte COSET-InVS sur le thème des risques professionnels, cohorte CKD-Rein concernant l'insuffisance rénale), qu'en Europe avec la Cohorte nationale allemande (200 000 adultes), et le consortium IDEAR (*Integrated Datasets across Europe for Ageing Research*) qui associe des cohortes anglaises, suédoises, allemandes et danoises.

Constances est également associée à des consortiums consacrés à l'harmonisation des données de type épidémiologique et biologique, français (Infrastructure nationale BIOBANQUES) et internationaux : consortium *Public Population Project in Genomics,* P3G et *Biobanking and Biomolecular Resources Research Infrastructure,* BBMRI.

Avancement du projet

Durant la phase de préparation, le projet de constitution de la cohorte Constances a recueilli de nombreux avis scientifiques : Conseil scientifique de la CNAMTS, Conseil scientifique de l'Institut de recherche en santé publique (IReSP), Conseil national de l'information statistique (Cnis) qui lui accordé son Label d'opportunité et de qualité statistique, Jury international des Appels à projets Cohortes et Infrastructures nationales de biologie et santé des Investissements d'avenir. Constances a également reçu le Label CQI de l'Inserm et l'autorisation de la Commission nationale de l'informatique et des libertés.

Les premières invitations à participer ont été envoyées début 2012, et la montée en charge des inclusions a commencé ; fin 2012, plus de 10 000 participants sont déjà inclus.

Le financement du projet provient de plusieurs sources. Le financeur le plus important est la CNAMTS, qui prend en charge l'essentiel des examens de santé et fournit

les données provenant du Système d'information inter-régimes de l'assurance maladie (SNIIR-AM). Pendant la phase de préparation et la réalisation des pilotes, Constances a été soutenu par la Direction générale de la santé, et le programme Très grandes infrastructures de recherche coordonné par l'IReSP. En 2012, Constances a été labellisé « Infrastructure nationale de biologie et santé » dans le cadre des Investissements d'avenir, et bénéficie à ce titre d'un important financement pour la période 2012-2019.

CONCLUSION : L'ÉMERGENCE DES « MÉGA-COHORTES »

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 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. Des dizaines de milliers de sujets peuvent être nécessaires pour étudier un phénotype quantitatif (pression artérielle par exemple), parce que les effets alléliques peuvent être aussi faibles qu'un dixième d'une déviation standard, voire moins [8].

Les « méga-cohortes » en Europe

C'est dans ce contexte qu'on voit se mettre en place une nouvelle génération de « méga-cohortes » en population. Certaines cohortes sont déjà en place en Europe. On peut ainsi citer en Grande-Bretagne la *Million Women Study* (www.millionwomenstudy.org/introduction/), qui a inclus plus d'un million de femmes âgées de 50 ans et plus, ou le projet *UK Biobank*, qui a inclus 500 000 personnes âgées de 40 à 69 ans (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 leur enfants (www.fhi.no/eway/?pid=238). La cohorte EPIC (*European Prospective Investigation into Cancer and Nutrition*) réunit 520 000 participants âgés de 20 ans et plus dans 10 pays européens (http://epic.iarc.fr/).

D'autres cohortes de très grande dimension sont actuellement à un stade de mise en place ou de préparation avancée. En Suède, la cohorte *LifeGene* prévoit d'inclure 500 000 sujets âgés de 0 à 45 ans (www.lifegene.se/In-english/). Aux Pays-Bas, la cohorte *LifeLines* prévoit de suivre 165 000 participants (http://lifelines.nl/). En Allemagne, la cohorte GeNatCo (*German National Cohort*) envisage un échantillon de 200 000 participants âgés de 20 à 70 ans (www.nationale-kohorte.de/informationen_en.html).

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

Elles concernent des thèmes d'intérêt général : les cohortes d'adultes s'intéressent particulièrement aux maladies chroniques et dégénératives fréquentes (cancers, maladies cardiovasculaires et métaboliques, maladies psychiatriques, démences, etc.). L'étude de la susceptibilité génétique 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 ; 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, examens médicaux, appariement à des bases de données nationales, et plusieurs cohortes recueillent des données de questionnaire par Internet. 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, notamment pour des nouveaux marqueurs qui n'existaient pas lors de la mise en place de la cohorte.

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 à projets permettant ainsi à des chercheurs extérieurs de bénéficier d'un accès aux données collectées.

Enfin, certaines cohortes, comme 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 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 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. Dans ce contexte, le projet LPC (*Large Prospective Cohorts*) qui associe une vingtaine de grandes cohortes en population provenant de 13 pays européens, réunissant au total plus de 2,5 million de sujets, s'est récemment constitué, la France y étant présente par les cohortes Gazel et Constances.

BIBLIOGRAPHIE

[1] Delfraissy J.F. — Mécanismes immunologiques et virologiques impliqués dans l'infection à virus de l'immunodéficience humaine : impact des traitements. *La Revue du Praticien,* 1999, *49*, 1740-1745.

[2] Goldberg M., Luce D. — Les effets de sélection dans les cohortes épidémiologiques. Nature, causes et conséquences. *Rev. Epidemiol. Santé Publique,* 2001, *49*, 477-92.

[3] Goldberg M., Chastang J.F., Zins M., Niedhammer I., Leclerc A. — Attrition during follow-up: health problems are the strongest predictors. A Study of the Gazel Cohort. *J. Clin. Epid.*, 2006, *59*, 1213-1221.

[4] Goldberg M., Leclerc A., Bonenfant S., Chastang J.F., Schmaus A., Kaniewski N., Zins M. — Cohort profile: the Gazel Cohort Study. *Int. J. Epid.*, 2007, *36*, 32-39.

[5] [En ligne] Disponible sur <www.gazel.inserm.fr> (consulté le 20 janvier 2013).

[6] Zins M., Leclerc A., Goldberg M. — The French Gazel Cohort Study: 20 years of epidemiologic research. *Advances in Life Course Research*, 2009, *14*, 135-146.

[7] Zins M., Bonenfant S., Carton M., Coeuret-Pellicer M., Guéguen A., Gourmelen J., *et al.* — The Constances Cohort: an Open Epidemiological Laboratory. *BMC Public Health*, 2010, *10*, 479.

[8] [En ligne] Disponible sur <www.constances.fr> (consulté le 20 janvier 2013)

[9] Burton P.R., *et al.* — Size matters: just how big is BIG? Quantifying realistic sample size requirements for human genome epidemiology. *Int. J. Epidemiol.,* 2009, *38*, 263-73.

DISCUSSION

M. Yvan TOUITOU

Des résultats épidémiologiques, même statistiquement significatifs, sont susceptibles d'être biaisés par des facteurs confondants, c'est-à-dire des facteurs connus ou inconnus dont il n'a

pas été tenu compte. Peut-on s'en abstraire surtout lors de résultats de faible amplitude (RR : 0,6 à 2 par exemple) ?

C'est tout l'art des épidémiologistes de prendre en compte cette difficulté. Dans Constances, nous recueillons de nombreuses données de comportement, d'environnement et de santé qui sont des facteurs de confusion potentiels pour divers problèmes de santé, ce qui permettra d'en tenir compte lors des analyses. Rappelons aussi que le jugement de causalité n'est pas uniquement « statistique » et qu'il doit prendre en compte d'autres critères comme la temporalité, l'existence d'une relation dose-effet, la réplication des résultats, la plausibilité biologique, notamment.

M. Michel HUGUIER

En augmentant le nombre de covariables prises en compte dans une étude de cohorte, n'augmente-t-on pas le risque de première espèce ?

Ce n'est pas le nombre de variables qui est véritablement en cause, mais le nombre de tests statistiques qui sont effectués. En théorie plus on teste de variables plus on augmente le risque de première espèce. Mais l'utilisation qui sera faite de Constances ne consistera pas à tester tout avec tout ! Les recherches menées consistent à vérifier des hypothèses *a priori*, avec un nombre de variables restreint, choisies de façon pertinente par rapport à ces hypothèses et au phénomène étudié : on se retrouve donc, pour une recherche donnée, dans une situation « classique » avec une hypothèse à tester et un nombre limité de variables adéquates. Cette approche comporte moins de risque de conclure à tort que les études de type « *genome-wide* » où les associations testées sont en nombre immense.

M. Georges DAVID

Quelles sont les conditions pour incorporer une nouvelle pathologie non connue au moment de la constitution de la cohorte ?

Constances est une cohorte « généraliste » qui n'est pas centrée sur une affection définie. Le dispositif mis en place doit permettre d'évoluer au cours du temps, lors du suivi des sujets et intégrer les maladies au fur et à mesure de leur occurrence. L'appariement systématique de la cohorte aux bases de données hospitalières et de l'assurance maladie permettra ainsi d'adapter le recueil de données à l'évolution des connaissances : si une nouvelle pathologie apparaît dans le futur elle fera alors l'objet de recueil de données adéquates.

M. Alain PRIVAT

Compte tenu de l'étendue de cette cohorte et de son caractère multicentrique, quelles précautions ont-elles été prises pour assurer la confidentialité des données ?

Le dispositif qui a été mis en place pour assurer la confidentialité des données est particulièrement complexe. Très schématiquement, il repose sur l'existence d'un « tiers de confiance » indépendant de l'équipe Constances. Tous les flux de données identifiantes (nom, adresse postale, etc.) passent par ce tiers de confiance seul habilité à disposer de ces données (mais qui ne dispose d'aucune autre donnée) qui sont conservées sous forme

cryptée : un système de numéros d'anonymat et de tables de correspondance permet les transferts de données individuelles de façon non identifiante. Ainsi, dans la base de données Constances, les sujets sont identifiés par un numéro non signifiant. Cette séparation physique et fonctionnelle des éléments identifiants et des données proprement dites interdit donc toute rupture de confidentialité. Ce dispositif a été mis en place en collaboration avec la Cnil, qui l'a autorisé.

M. Jean-Daniel SRAER

Est-il possible de faire des études génétiques à partir de bases de données d'une telle importance ? L'étude de Gazel n'est-elle pas biaisée par le recrutement à l'intérieur d'une entreprise EDF-GDF dans laquelle l'avenir est assuré ?

Les études génétiques seront possibles, puisqu'à l'occasion de l'examen de santé d'inclusion, du sang est recueilli et conservé dans une biobanque. Il sera donc possible d'extraire des échantillons aux fins d'analyse génétique.

Concernant Gazel, il faut considérer que cette cohorte n'a aucunement la prétention d'être représentative de la population française, ni même de celle d'EDF-GDF (les effets de sélection liés au volontariat des participants ne le permettent pas). Gazel a été conçu pour étudier des associations entre l'exposition à des facteurs de risque et le risque de développer des maladies. Or, au sein d'une cohorte dont les procédures d'inclusion ont été les mêmes pour tous les sujets (ce qui est le cas de Gazel), la relation exposition/maladie n'est *a priori* pas différente entre les sujets volontaires et ceux qui ne le sont pas. Une des raisons est qu'au moment de l'inclusion, tous sont indemnes des maladies qui seront analysées, seuls les cas incidents pendant la période de suivi étant pris en compte dans les études de cohorte : des conditions très particulières seraient en effet nécessaires pour que des effets de sélection dus au volontariat puissent entraîner un biais dans la mise en évidence ou la quantification d'une relation entre une exposition et une maladie.

Ceci n'est pas propre à Gazel, et concerne toutes les cohortes de ce type : la population de Framingham, commune prospère de la banlieue de Boston, n'est certainement pas représentative de la population américaine et encore moins de celle d'autres pays, et pourtant les résultats qui en sont issus ont valeur universelle. Cela étant, il est vrai que la composition particulière de la cohorte Gazel ne permet pas l'étude de nombreux phénomènes, du fait de la structure d'âge, de l'absence de diverses catégories de personnes, etc. Il reste néanmoins des domaines de recherche très divers où elle constitue un outil d'investigation de choix, comme le montrent les quelques 200 publications qui en sont déjà issues.