Les études de cohorte : principes et méthode

Les études de cohorte suivent un groupe important de personnes et évaluent les effets sur leur santé des facteurs de risque auxquels elles sont exposées. La fiabilité de ces études repose sur une méthodologie rigoureuse afin d'éviter tout biais, toute erreur de collecte des données ou d'interprétation des résultats.

Principe et intérêt des cohortes épidémiologiques

Marcel Goldberg
Marie Zins
Inserm U1018, plate-forme de recherche Cohortes épidémiologiques en population — Centre de recherche en épidémiologie et santé des populations, université de Versailles-Saint-Quentin, UMRS 1018

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

Qu'est-ce qu'une cohorte épidémiologique ?

La cohorte épidémiologique est un type d'enquête dont le principe est le suivi longitudinal, à l'échelle individuelle, d'un groupe de sujets. Selon les objectifs scientifiques, la durée d'observation des sujets et les données individuelles recueillies de façon prospective diffèrent. Une distinction majeure doit être faite d'emblée 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 (évolution naturelle ou sous traitement), incluent un nombre souvent restreint de sujets (quelques milliers, parfois quelques dizaines de milliers pour les plus importantes) habituellement 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, grâce à un suivi rapproché des patients, les principales phases de l'évolution de l'infection par le VIH et la relation entre la charge virale et le nombre de lymphocytes T4 au cours du temps [13].

Ces cohortes sont un outil précieux pour la recherche clinique mais, ne prenant en compte que des personnes malades, elles relèvent en fait du domaine de la recherche biomédicale « classique », avec parfois une dimension sociale (lire *Apport des sciences sociales : l'exemple de cohortes de patients infectés par le VIH*, p. 26).

Les cohortes en population générale sont celles qui font l'objet de ce dossier. 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 qui s'accompagnent de biobanques.

Le principe d'une cohorte à visée étiologique est simple, et résumé par la figure 2.

On choisit un groupe de sujets qui sont *a priori*

figure 1

Phases de l'évolution de l'infection par le VIH

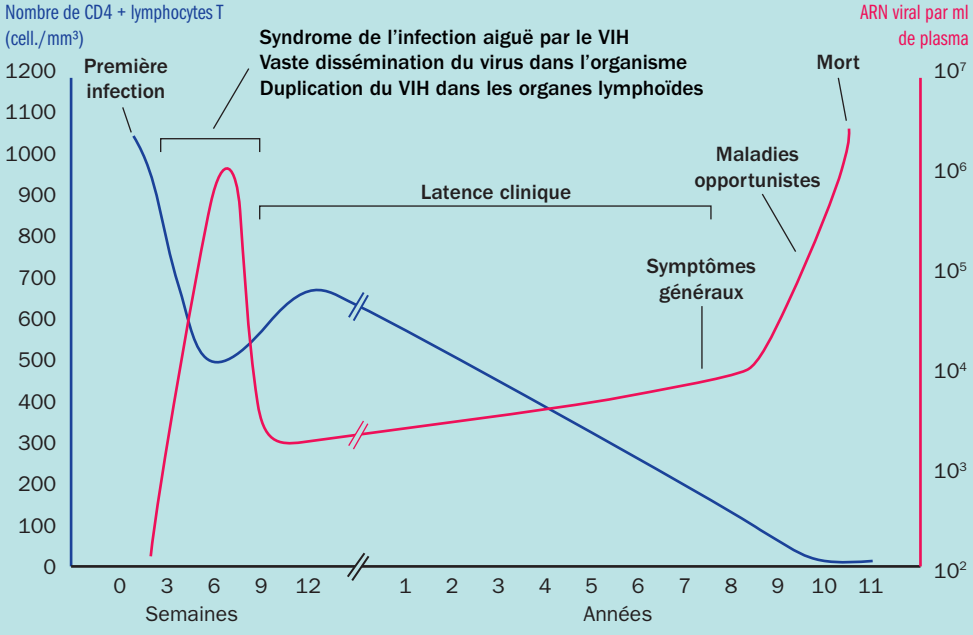
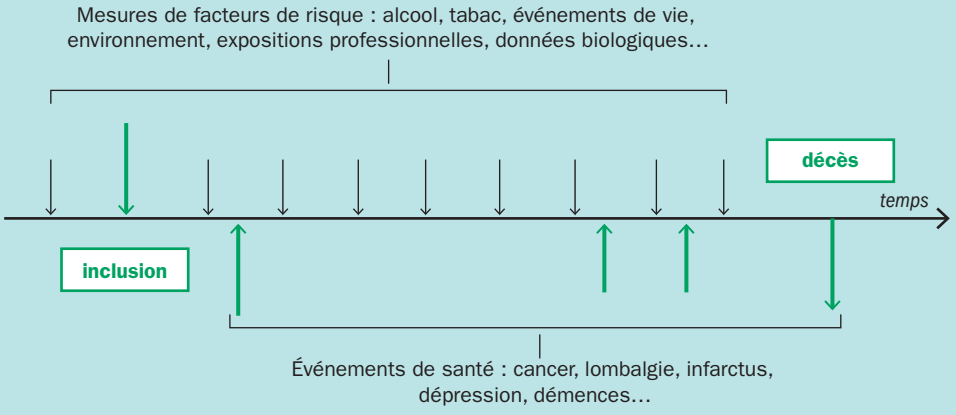


figure 2

Schéma de base d'une cohorte épidémiologique prospective



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.

Ces cohortes sont souvent « généralistes », et se caractérisent par une couverture large de problèmes de santé et de déterminants. Elles sont « *conçues 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* » selon, la définition de l'Agence nationale de recherche sur le sida, et constituent alors de véritables infrastructures de recherche et de santé publique, comme le montrent les exemples décrits dans ce numéro (lire *Les nouvelles « méga-cohortes » en population en Europe, p. 34* et *Les cohortes « historiques » en France*, p. 37).

Pourquoi des cohortes ?

Sur le plan méthodologique, les avantages principaux des cohortes sont la possibilité d'analyses épidémiologiques longitudinales permettant de tenir compte au mieux de phénomènes liés au temps, notamment de la séquence temporelle exposition (ou intervention) effet. Il est ainsi possible de modéliser l'enchaînement et les interactions des différents facteurs relatifs aux conditions de vie (alimentation, habitat, accès aux soins, réseau social…), à l'environnement (conditions de travail, expositions professionnelles et environnementales…), et à l'état de santé (états précliniques, chronologie des phénomènes pathologiques). Par ailleurs, les données d'exposition étant recueillies avant la survenue des effets analysés, 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 ou d'interventions préventives, en permettant de prendre en compte les évolutions temporelles et les interactions entre facteurs.

Les domaines d'utilisation des cohortes sont aussi diversifiés que l'épidémiologie elle-même, et concernent tous les aspects de la santé en relation avec des facteurs de risque de type varié. Outils de recherche épidémiologique, les cohortes en population peuvent également, sous certaines conditions, avoir des objectifs descriptifs et de surveillance (description, suivi de l'évolution et surveillance des pathologies et de l'exposition à des facteurs de risque), et d'évaluation de l'efficacité à court, moyen et long termes d'interventions de nature préventive ou réparatrice.

Limites et difficultés

Ainsi présentées, les cohortes longitudinales en population semblent être l'instrument idéal qui répond à tous les besoins de recherche et de santé publique. 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 à observer 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 de l'indice d'association (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 ainsi réunir des effectifs immenses pour répondre à des questions d'intérêt, ce qui constitue une des principales limites des cohortes prospectives en population. À titre d'illustration, si l'on voulait connaître la prévalence du diabète non diagnostiqué selon le sexe, l'âge et la profession et catégorie socioprofessionnelle (PCS) dans la population adulte, et sous l'hypothèse que la prévalence totale dans la population adulte serait de 1 %, on obtiendrait, dans une cohorte de 200 000 sujets, des intervalles de confiance variant entre 0,81 et 1,19, donc une précision de 1 % +- 19 %, ce qui n'est évidemment pas satisfaisant. Si l'on s'interroge sur les effets de l'exposition aux pesticides sur le risque de myélome multiple (cancer rare, dont l'incidence annuelle est d'environ 9/100 000), en retenant des hypothèses réalistes concernant la prévalence de l'exposition et l'augmentation du risque, l'effectif minimum nécessaire après six ans de suivi est de plus de 1 100 000 sujets ; dix ans après, il est d'environ 520 000 sujets. Ces exemples montrent 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, sont indispensables.

Effets de sélection, biais et représentativité

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 lorsque la population observée diffère de la population cible en raison de phénomènes liés au recrutement ou au suivi des sujets. Or, dans la plupart des cohortes épidémiologiques, 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) [25].

Lorsque l'objectif de l'étude est descriptif (estimation de la fréquence de la maladie ou de l'exposition à un facteur de risque dans la population) il faut, pour éviter les biais de sélection, que le paramètre soit estimé sur un échantillon représentatif de la population cible, c'est-à-dire en pratique tiré au sort dans une base de sondage appropriée. Le mode d'inclusion faisant appel au volontariat entraîne inévitablement des effets de sélection, même lorsqu'on procède à un tirage au sort aléatoire d'un échantillon dans une base de sondage appropriée. On rencontre en effet des non-participants à l'inclusion (personnes non retrouvées, refus, etc.), ainsi que des sujets perdus de vue en cours de suivi, qui constituent une source potentielle de biais.

Pour y remédier, on s'efforce de recueillir lors de l'inclusion un minimum de données sur les non-participants (essentiellement âge, sexe, et PCS), afin de

procéder ultérieurement à des redressements pour estimer les paramètres d'intérêt. Cette approche connaît cependant certaines limites. Ainsi, il n'est pas toujours possible de recueillir les données de redressement pour l'ensemble des sujets non participants. De plus, il n'est pas toujours facile de savoir si ces données sont suffisantes pour contrôler les biais potentiels, car on sait par exemple qu'au sein de la même catégorie socio-économique existent de larges différences à bien des égards, notamment en termes de santé, de comportements, de modes de vie, de réseaux sociaux, etc. Ainsi, la comparaison des volontaires de la cohorte Gazel aux non-participants de même catégorie socio-professionnelle, âge et genre, illustre ce point : le statut marital, les consommations d'alcool et de tabac, les comportements à risque pour la santé, l'existence de maladies psychiatriques sont fortement associés à la participation initiale à la cohorte [23].

Le même problème se pose tout au long du suivi, les non-répondants et les perdus de vue différant toujours des participants pour divers facteurs, en particulier les comportements de vie et les problèmes de santé qui jouent un rôle majeur, même à catégorie socioprofessionnelle égale, comme on a pu l'observer là aussi dans la cohorte Gazel : le risque d'attrition diffère en fonction des consommations d'alcool et de tabac, de l'état de santé perçu, de l'absentéisme médical, de la survenue de problèmes de santé mentale et de cancers notamment [24]. Or ce sont justement ce type de facteurs qui sont étudiés dans les cohortes épidémiologiques.

Finalement, on est rarement en situation de contrôler complètement les biais de sélection potentiels, car il faut pour cela disposer de données pertinentes recueillies à la fois pour les participants et l'ensemble des non-participants. Cela est parfois possible si l'on a accès à des sources de données où toute la population cible est représentée, comme les bases de données de l'Assurance maladie ou du Programme de médicalisation du système d'information des hôpitaux (PMSI) [26].

Dans un contexte où l'on cherche à étudier les relations entre exposition à des facteurs de risque et survenue de maladies (objectif étiologique), la situation est plus simple. 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 que, 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 entraîner un biais dans la mise en évidence ou la quantification 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, ou au contraire protectrices.

Le problème de l'attrition au cours du suivi peut par contre être à l'origine de biais importants si la probabilité de ne plus être suivi diffère chez les exposés et les non-exposés, et/ou chez ceux qui sont ou ne sont pas devenus malades, ce qui est souvent le cas.

Données répétées et données manquantes

Les cohortes épidémiologiques présentent deux caractéristiques particulières qui suscitent des difficultés méthodologiques : (i) les mêmes variables peuvent être recueillies à plusieurs reprises au cours du suivi pour les mêmes sujets ; (ii) ces variables peuvent être manquantes à un ou plusieurs points de mesure au cours du suivi, et cela d'autant plus fréquemment que celui-ci est de longue durée et que le recueil des données est répété.

On dispose de différentes méthodes statistiques pour traiter ces problèmes ; elles sont résumées dans l'article *Aspects méthodologiques liés à l'analyse de données longitudinales et aux effets de sélection, p. 18*.

Identification des pathologies incidentes et phénotypage

Une des difficultés majeures des cohortes de population est l'identification des pathologies incidentes parmi les sujets au cours du suivi. Les déclarations des sujets eux-mêmes sont insuffisantes : elles peuvent être imprécises, voire erronées, potentiellement entachées de biais divers, et surtout… manquantes, car une des raisons majeures de l'abandon de la participation à un suivi de cohorte est justement la survenue de pathologies [24, 25]. Par ailleurs, on ne dispose pas en France de source exhaustive et fiable d'enregistrement des pathologies incidentes à l'échelle de la population générale, sauf exceptions partielles (registres du cancer, par exemple) mais qui ne couvrent qu'un petit nombre de maladies et le plus souvent des territoires restreints. L'existence des grandes bases de données médico-administratives nationales de l'Assurance maladie et du PMSI constitue une aide potentiellement précieuse, car elles couvrent toute la population et enregistrent la quasi-totalité des consommations de soins et les problèmes de santé sérieux, et permettent donc *a priori* d'avoir connaissance de la survenue des pathologies d'intérêt.

Cependant, outre des difficultés d'accès à ces bases de données, il faut considérer que la validité des diagnostics médicaux est le plus souvent imparfaite, et que ceux-ci doivent faire l'objet de confirmations systématiques, afin d'obtenir un « phénotypage » de qualité. Cela est une activité particulièrement lourde, qui implique le retour au

médecin soignant, la recherche de documents médicaux (comptes-rendus d'anatomopathologie, imagerie, etc.), et l'examen systématique des dossiers par des « comités de validation » constitués d'experts médicaux.

Les aspects concernant l'utilisation des bases de données médico-administratives sont détaillés dans l'article *L'apport des bases de données médico-administratives, p. 21*.

Les cohortes épidémiologiques en France

Malgré les limites évoquées, on a vu se développer en France, depuis une quinzaine d'années, de nombreuses cohortes aux objectifs divers. Les cohortes françaises se caractérisent cependant par leur taille relativement faible, aucune ne dépassant un petit nombre de dizaines de milliers de sujets (lire *Les cohortes « historiques » en France*, p. 37), alors que certaines cohortes dans d'autres pays peuvent atteindre, voire dépasser, plusieurs centaines de milliers de sujets (lire *Les nouvelles « méga-cohortes » en population en Europe, p. 34*).

La relative modestie des cohortes françaises s'explique par plusieurs raisons. Outre le nombre notoirement trop faible des épidémiologistes, on se heurte aujourd'hui en France à de nombreuses difficultés d'ordre financier, organisationnel et technique. Les coûts des cohortes sont élevés, car l'épidémiologie fait essentiellement appel à des données qui sont le plus souvent recueillies auprès des personnes elles-mêmes par des moyens divers : entretiens, auto-questionnaires, examens médicaux, collecte de matériel biologique, etc. Ces coûts restent finalement modestes si on les compare à ceux des grands instruments de physique ou à ceux de la recherche spatiale, voire au prix d'une journée d'hospitalisation dans un service de CHU, mais ils sont largement supérieurs aux budgets qu'il est habituellement 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 mis en place que très récemment un système de financement spécifique, et continue *de facto* de sous-estimer l'importance scientifique de telles plates-formes de recherche, malgré des efforts récents (lire *Les grandes cohortes en santé 2008-2011, p. 39*). Cependant, les budgets qui sont distribués sont la plupart du temps très loin des coûts véritables, et d'au moins un ordre de grandeur inférieur aux financements des cohortes étrangères comparables.

D'autres difficultés tiennent à la nécessité de l'implication à long terme des équipes dont la pérennité n'est souvent pas assurée, et à la quasi-impossibilité de disposer de personnels spécialisés stables et d'un niveau de qualification suffisant, notamment du fait de l'absence de statut reconnu pour ce type d'activité dans les organismes publics de recherche, alors que la durée des projets est incompatible avec un trop fort 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 comparable à ce qui existe dans les pays de niveau scientifique comparable, de nouvelles cohortes prospectives sont indispensables, dont l'effectif ne se comptera plus en dizaines, mais en centaines de milliers de sujets.

Aspects méthodologiques liés à l'analyse de données longitudinales et aux effets de sélection

Alice Guéguen
Rémi Sitta
Inserm U1018, plate-forme de recherche Cohortes épidémiologiques en population – Centre de recherche en épidémiologie et santé des populations, université de Versailles-Saint-Quentin, UMRS 1018

Si les études de cohorte présentent de nombreux avantages méthodologiques, comme cela a été indiqué dans l'article *Principe et intérêt des cohortes épidémiologiques*, p. 14, elles n'en présentent pas moins certaines difficultés sur le plan statistique, notamment pour ce qui concerne l'analyse de données longitudinales et la prise en compte des effets de sélection.

L'analyse des données longitudinales

Le principe des études de cohorte repose sur le suivi longitudinal d'un groupe de sujets, incluant notamment le recueil répété des mêmes variables au cours du temps. Une cohorte épidémiologique est parfois le seul moyen de répondre à certaines questions de recherche, par exemple pour l'analyse de trajectoires, ou de l'incidence d'évènements irréversibles. D'autres fois, ce sera un moyen parmi d'autres, mais en général le plus efficace : en recueillant des données répétées sur les mêmes données de santé, on pourra décrire leur évolution dans le temps. Chaque sujet étant son propre « témoin », et les données mesurées sur un même sujet étant corrélées, on peut obtenir une bonne précision des estimateurs, car ces caractéristiques diminuent leur variance.

Cependant, les méthodes d'analyse classique ne sont plus utilisables, car elles fournissent des estimations dont les variances peuvent être à tort soit trop élevées soit trop faibles. Deux types de modèles ont été développés