

# 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**

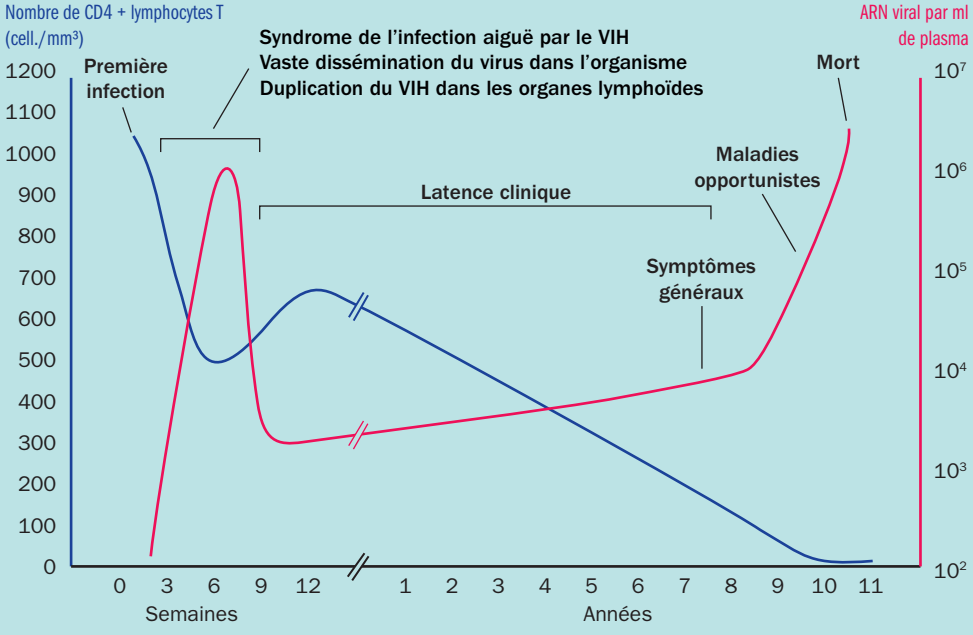
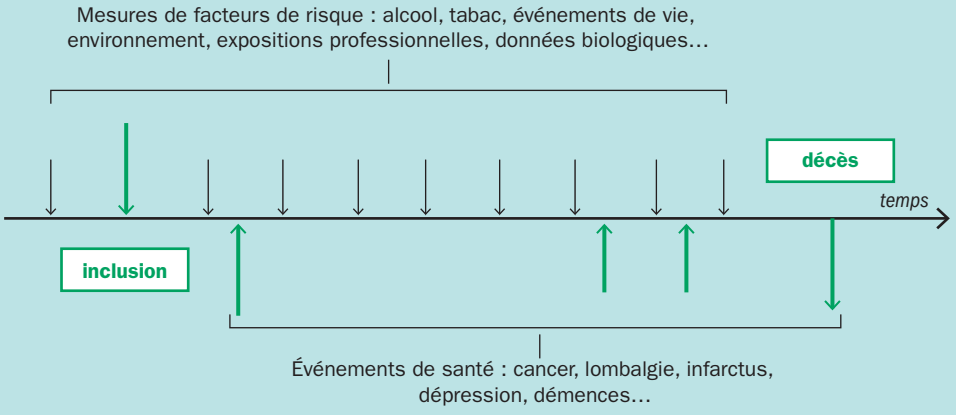



figure 2

**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

pour analyser ce type de données : les modèles mixtes et les modèles marginaux [28]. Selon la nature de la variable d'intérêt (continue et de distribution normale, binaire, à plusieurs catégories, etc.), ces méthodes sont plus ou moins faciles à mettre en œuvre.

Les modèles mixtes incluent dans la modélisation de la variable d'intérêt des effets aléatoires propres à chaque sujet. Ceux-ci se comportent comme des paramètres qui n'ont pas d'intérêt en soi, mais qui permettent de prendre en compte la corrélation entre les données. Si cette méthode se met facilement en œuvre pour des données continues et de distribution normale, il n'en va pas de même dans d'autres situations, par exemple quand la variable d'intérêt est binaire, et qu'il y a de plus peu de temps de recueil.

Les modèles marginaux ont pour objectif de modéliser directement la moyenne de la variable d'intérêt. Quand celle-ci est continue et de distribution normale, la mise en œuvre de ces modèles est facilitée grâce à l'existence de la distribution multinormale. En revanche, quand la variable d'intérêt est binaire ou a plusieurs catégories, il n'existe pas de distribution multidimensionnelle similaire. Les méthodes des GEE (*Generalized estimating equations*) ont été développées à la fin des années 80 pour pallier ce problème.

## Les effets de sélection

Les données de cohorte en population générale sont le plus souvent collectées directement auprès de sujets tirés au sort dans une population cible. Il en résulte que la population enquêtée à l'inclusion peut différer de la population cible en raison de phénomènes liés à la non-participation. Il est également possible qu'il y ait non-participation au cours du suivi. Celle-ci peut être soit définitive à partir d'un moment donné — on parle alors d'attrition –, soit intermittente (certaines personnes ne participent pas à un moment donné du suivi, puis participent de nouveau).

Les phénomènes de sélection sur la population cible, en diminuant la quantité d'information disponible, conduisent ainsi à une perte de précision dans les estimations produites à partir de la population enquêtée. Mais la conséquence la plus importante est que ces estimations peuvent être incorrectes : elles se trouvent en effet biaisées dès que certains facteurs de la participation sont liés statistiquement aux variables étudiées. Cela est particulièrement vrai dans un contexte « descriptif » où on cherche à estimer des moyennes, des fréquences, des incidences ou encore des prévalences de maladies dans une population particulière. Dans un contexte « explicatif » où l'on s'intéresse à des mesures d'association (essentiellement entre une exposition et une maladie), les biais sont en général de plus faible importance.

Habituellement, on cherche à éviter ce biais en incluant les facteurs de participation dans la modélisation à partir de données recueillies sur les seuls participants. Cette solution peut donner des résultats corrects, mais il existe des situations particulières dans lesquelles les résultats seront pourtant systématiquement biaisés [27], même lorsque tous les facteurs de participation sont connus et mesurés.

D'une manière générale, il est possible d'obtenir des estimations correctes à condition de tenir compte du mécanisme de non-participation, ce qui sous-entend qu'on le connaisse. Or ce mécanisme est inconnu, et la seule solution acceptable consistera à faire des hypothèses sur celui-ci. Les estimations produites ne seront donc valides que sous ces hypothèses. On distingue trois types de mécanismes de données manquantes.

- Données MCAR (*Missing completly at random*) : la valeur de la variable d'intérêt et la probabilité qu'elle soit manquante sont indépendantes. La plausibilité d'une telle hypothèse est quasi systématiquement remise en cause dans les enquêtes épidémiologiques, mais elle est envisageable dans d'autres études : après un prélèvement biologique, le fait qu'un tube se casse ou que l'analyseur de biologie tombe en panne conduira à des données de type MCAR. Dans la situation où les données sont MCAR, les résultats des analyses naïves effectuées sur la population enquêtée sont corrects.
- Données MAR (*Missing at random*) : après prise en compte des caractéristiques observées du sujet jusqu'à sa non-participation, la valeur de la variable d'intérêt et la probabilité qu'elle soit manquante sont indépendantes.
- Données MNAR (*Missing not at random*) : même après prise en compte des caractéristiques observées du sujet jusqu'à sa non-participation, la valeur de la variable d'intérêt et la probabilité qu'elle soit manquante sont corrélées.

La pertinence de l'hypothèse MAR ou MNAR dépend essentiellement des données dont on dispose : plus il existe de l'information potentiellement liée à la fois à la non-participation et à la variable d'intérêt, plus l'hypothèse MAR devient acceptable. Ce qui implique que si les données observées ne devaient pas être suffisantes pour la plausibilité de l'hypothèse MAR, il faudrait envisager d'enrichir les données par suffisamment d'informations supplémentaires, par exemple issues de sources extérieures à l'enquête elle-même. En tout état de cause, une bonne approche consiste à faire des analyses de sensibilité : on considère plusieurs hypothèses alternatives plausibles pour spécifier le mécanisme de non-participation, et on examine la manière dont les résultats fluctuent en fonction des hypothèses envisagées.

En résumé, l'hypothèse MCAR est rarement plausible. Sous l'hypothèse MAR, il est possible de prendre en compte le mécanisme de non-participation, mais comme cette hypothèse ne peut pas être vérifiée à partir des données observées, il est toujours préférable d'envisager l'hypothèse MNAR.

## Les méthodes

Deux méthodes ont récemment été développées pour donner des résultats sans biais sous l'hypothèse que

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



les données sont MAR : la pondération [20] et l'imputation [41]. Elles nécessitent de recueillir, pour les participants et les non-participants, des informations liées à la non-participation. Elles permettent de « reconstituer » les données manquantes des non-participants grâce aux données disponibles des participants et des non-participants. La description des méthodes se fait plus facilement dans le cas suivant : tous les sujets participent à l'inclusion, et il y a un seul temps de suivi ultérieur, où sera recueillie la variable d'intérêt. Les deux méthodes se généralisent ensuite à des situations plus complexes.

La méthode des imputations s'effectue en deux étapes : parmi les participants, on construit un modèle qui explique la variable d'intérêt par les variables observées à l'inclusion. Ce modèle est alors appliqué à chaque non-participant à partir des variables observées à l'inclusion, afin de lui prédire une valeur pour la variable d'intérêt. On ajoute souvent en pratique à la prédiction du modèle un terme reflétant la variabilité de la variable d'intérêt, et on répète cette procédure plusieurs fois pour que les données ainsi générées conservent toute la structure multidimensionnelle originelle de la population cible. Les analyses sont ensuite effectuées sur chaque jeu de données entier complété ainsi par imputation, et les résultats sont synthétisés. La généralisation au cas où il y a non-participation intermittente est plus compliquée à décrire sur le plan théorique ; elle est depuis quelques années facilement mise en œuvre grâce à l'implémentation de ces méthodes dans les logiciels statistiques.

La méthode des pondérations comprend également deux étapes : la première étape consiste à écrire un modèle de participation/non-participation qui prédit la probabilité qu'un sujet soit participant en fonction des variables observées à l'inclusion. Dans un deuxième temps, on affecte aux seuls sujets participants une pondération égale à l'inverse de ces probabilités prédites. Cette approche se justifie intuitivement ainsi : un sujet participant qui, au vu de ses caractéristiques antérieures, présente une faible probabilité de participer se verra ainsi attribuer une pondération importante, de manière à ce qu'il représente les nombreux sujets non participants ayant les mêmes caractéristiques que lui. Les estimations sont alors obtenues grâce à une analyse pondérée, effectuée sur la population des participants. Cette méthode nécessite que tous les individus de la population cible aient une probabilité de participation non nulle, car il n'y aurait sinon aucun participant pour représenter ces non-participants.

Lorsqu'il y a plusieurs temps de recueil, en cas d'attrition, la généralisation se fait simplement en modélisant la participation à chaque temps de recueil parmi les participants du temps précédent. Les probabilités modélisées sont alors multipliées entre elles, et le produit final est inversé pour fournir une pondération pour les sujets participant à tous les temps envisagés. En revanche, lorsque la non-participation est intermittente, la méthode des pondérations, en théorie possible, rend les analyses très lourdes : une solution simple, mais moins performante, consiste à considérer la non-participation comme définitive dès la première occurrence et ignorer les réponses ultérieures.

Ces deux méthodes peuvent être utilisées simultanément. Par exemple, pour la non-participation à l'inclusion, on applique quasi systématiquement une pondération, en s'appuyant sur des informations externes à l'enquête elle-même, ce qui n'empêchera pas de traiter l'attrition future soit par de l'imputation, soit par pondération (auquel cas la pondération totale sera le produit de la pondération pour non-inclusion et de celle pour attrition).

Les deux méthodes, pondérations et imputations, sont théoriquement équivalentes, mais elles ont en pratique chacune leurs avantages et leurs limites. La comparaison pondération/imputation semble indiquer une plus faible variance des estimateurs par imputation, mais parfois cela reflète uniquement la trop grande confiance implicite donnée à tort au modèle d'imputation.

### Autres aspects méthodologiques propres aux données de cohorte

Les questions méthodologiques pour les études de cohorte s'orientent dans différentes directions. Les méthodes d'analyse de données longitudinales évoquées plus haut donnent des résultats biaisés quand l'exposition varie au cours du temps et qu'il existe des variables de confusion, elles-mêmes affectées par des expositions antérieures ; les modèles marginaux structurels ont été développés à cette intention. Le décès lui-même peut être cause d'attrition, et causer des biais en particulier s'il partage des facteurs de risque avec la variable d'intérêt ; selon l'objectif, descriptif ou explicatif, l'attitude face à cette attrition est de considérer la cohorte comme mortelle ou immortelle [18]. Les cohortes épidémiologiques incluent souvent un nombre important de sujets, mais la quantité d'information recueillie par sujet est en général bien supérieure. Cela est d'autant plus vrai lorsque les cohortes intègrent des données provenant de sources externes, telles des bases de données médico-administratives nationales. Les méthodes statistiques utilisées devront alors s'adapter à ce cas particulier où le nombre de sujets est plus faible que le nombre de variables, et emprunter des méthodes issues de la fouille des données.

# L'apport des bases de données médico-administratives

**Céline Ribet**
**Mireille Cœuret-Pellicer**
**Julie Gourmelen**
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

La France est l'un des rares pays dont les organismes de protection médico-sociale ou de gestion hospitalière disposent de systèmes d'information centralisés couvrant de façon exhaustive et permanente l'ensemble de la population. Les données enregistrées en routine comportent des informations sur le recours aux soins, les hospitalisations, le handicap, les prestations sociales et l'activité professionnelle. Bien que n'ayant pas à l'origine de finalité épidémiologique, ces bases offrent un intérêt potentiel majeur pour la réalisation de telles études mais sont encore très peu exploitées. On présentera ici les principaux systèmes d'information, leur exploitation potentielle en santé publique ainsi que les précautions que nécessite leur utilisation.

## Description des principales données disponibles

### Les données socioprofessionnelles

Les événements socioprofessionnels des individus sont informatisés dans les systèmes nationaux des différents régimes d'assurance vieillesse. Pour toute personne ayant appartenu au moins une fois au cours de sa vie au régime général de la Sécurité sociale, c'est la Caisse nationale d'assurance vieillesse (Cnav) qui procède à l'enregistrement des données lui permettant de garantir le droit au paiement de la retraite. Pour répondre à cet objectif, la Cnav a mis en œuvre plusieurs systèmes nationaux lui permettant de collecter et traiter les données sociales issues de différents organismes et régimes gestionnaires des prestations sociales, dont le principal est le Système national de gestion des carrières (SNGC). Cette base de données permet de retracer, pour chaque individu dès l'âge de 16 ans et jusqu'à la liquidation de ses droits à la retraite, ses différentes périodes d'activité : périodes d'activité professionnelle (par l'intermédiaire des déclarations transmises par les employeurs) ou périodes assimilées (chômage, maladie, maternité ou congés parentaux ; informations transmises respectivement par l'Assurance chômage, l'Assurance maladie, et les caisses d'allocations familiales). Le SNGC contient donc l'ensemble des données inhérentes à la carrière des assurés du régime général, y compris les données concernant d'éventuelles périodes effectuées dans d'autres régimes de base (régimes des indépendants, des agriculteurs…) ainsi que dans certains régimes particuliers ou spéciaux (SNCF, EDF…).

Un autre système d'information mis en œuvre par la Cnav est le Répertoire national inter-régimes des bénéficiaires de l'assurance maladie (RNIAM), qui permet de connaître l'organisme de rattachement de chaque bénéficiaire d'un régime d'assurance maladie par l'intermédiaire du NIR (lire encadré).

### Les données de mortalité

Le statut vital et les causes de décès des sujets d'une enquête peuvent être obtenus auprès du Centre d'épidémiologie sur les causes médicales de décès (CépiDC)

## NIR, RNIPP, SNGI

Le « numéro d'inscription au répertoire », ou NIR, est l'identifiant unique et invariable de tout individu. Ce numéro à treize caractères (plus deux pour la clé de contrôle), dont la composition est précisée par décret, est attribué à une seule et unique personne, et une personne ne possède qu'un NIR. Une fois attribué, il ne change plus.

L'attribution de ce numéro et son association aux autres éléments d'identification d'un individu (nom patronymique, prénoms, date et lieu de naissance, numéro de l'acte de naissance, sexe) se font dès la naissance sur la base des informations enregistrées par l'état civil. Au moment du décès, s'ajoutent les date et lieu de décès et le numéro de l'acte.

Pour les personnes nées en France métropolitaine ou dans les DOM, qu'elles soient françaises ou étrangères, c'est l'Insee qui a en charge cette immatriculation et qui procède à sa conservation au sein du Répertoire national d'identification des personnes physiques (RNIPP). Pour les personnes nées à l'étranger, à Mayotte et dans les TOM, c'est la Cnav qui met en œuvre d'une part l'immatriculation (uniquement lorsque l'inscription est demandée par un organisme habilité), et d'autre part la conservation au sein du Système national de gestion des identités (SNGI). Ces deux fichiers ont pour finalité de certifier l'état civil et le statut vital d'une personne auprès des organismes de sécurité sociale, de l'administration fiscale, de la Banque de France, du Système informatique pour le répertoire des entreprises et des établissements (Sirene). Leur utilisation repose sur de fortes obligations légales ; ainsi, ils ne peuvent être servir à des fins de recherche des personnes.



de l'Inserm selon la procédure décrite dans le décret n° 98-37. Cette procédure permet d'apparier des données d'état civil et de statut vital hébergées par l'Insee aux causes médicales de décès anonymes.

### Les données d'hospitalisation

Le Programme de médicalisation du système d'information des hôpitaux (PMSI) consiste en un recueil exhaustif systématique et standardisé d'informations médicales et administratives pour tout séjour d'un patient dans un établissement de soins. Il concerne aujourd'hui tous les établissements (publics et privés) et tous les types de séjours (médecine, chirurgie, obstétrique, soins de suite et de réadaptation, psychiatrie, urgences, soins à domicile). L'objectif principal du PMSI est de décrire l'activité d'un établissement à des fins d'allocation budgétaire. L'information est médicalisée et repose sur un classement des séjours en « groupes médicalement homogènes » (GHM), à partir du codage des diagnostics établis au cours d'un séjour et des principaux actes pratiqués. Ces informations sont anonymisées puis rassemblées dans une base de données nationale gérée par l'Agence technique de l'information sur l'hospitalisation (ATIH).

### Les données de l'Assurance maladie

Il existe en France un grand nombre de régimes d'assurance maladie, disposant chacun de son propre système d'information contenant les données nécessaires à la liquidation des prestations de ses assurés. Ces données comprennent des informations détaillées sur les soins présentés au remboursement (consultations, médicaments, prélèvements biologiques…), ainsi que sur les assurés, les établissements de soins et les professionnels de santé. Les services médicaux des caisses disposent de leurs propres fichiers comportant des informations médicales structurées sur les affections de longue durée (ALD), les accidents du travail et les maladies professionnelles.

La nécessité de suivre l'ensemble des dépenses tous régimes confondus a abouti en 2003 à la création du Système national d'informations inter-régimes de l'Assurance maladie (SNIIR-AM). Ces données concernent aujourd'hui tous les régimes d'assurance maladie, pour la médecine de ville comme pour l'hospitalisation. Elles sont individualisées par bénéficiaires, professionnels de santé et établissements, et médicalisées (les actes sont codés selon la Classification commune des actes médicaux et les pathologies selon la CIM10).

Grâce à un identifiant anonyme commun, les données du PMSI sont également désormais intégrées au SNIIR-AM.

## Utilité des bases dans un cadre épidémiologique

Les bases médico-administratives offrent de nombreux avantages inhérents à leur constitution.

Les données sont individuelles. Ainsi, l'accès à ces bases de données peut servir à sélectionner des individus en vue de l'inclusion dans une enquête épidémiologique à partir de critères tels qu'une pathologie, un recours à des soins spécifiques ou une profession. Un exemple récent est l'étude des effets du Médiator : il a été possible d'identifier dans le SNIIR-AM toutes les personnes ayant eu une prescription remboursée de ce médicament, et de suivre leur devenir médical, avec les résultats que l'on sait [46].

Les données sont quasi exhaustives par rapport à la population française. Elles permettent donc de disposer d'effectifs immenses pour certaines analyses. Cette exhaustivité peut aider à prendre en compte les effets de sélection à l'inclusion et au cours du suivi, qui sont une source majeure de biais dans les enquêtes épidémiologiques (lire *Aspects méthodologiques liés à l'analyse de données longitudinales et aux effets de sélection, p. 18*) :

- la constitution d'un fichier de « non-participants », pour lesquels on pourra disposer de données sur leurs consommations de soins, leurs hospitalisations et leurs caractéristiques socioprofessionnelles, permet d'étudier les facteurs liés à la non-participation ;
- le suivi de façon « passive », à travers ces bases, des personnes incluses dans des études mais qui ne répondent plus aux questionnaires permet de pallier le problème des perdus de vue.

Enfin, ces données sont parfois plus fiables que des informations obtenues par auto-questionnaire. Par exemple, les informations sur la carrière professionnelle, qui servent au calcul des retraites, sont pour des raisons évidentes particulièrement complètes et validées, toute erreur pouvant en effet avoir un impact économique sur les bénéficiaires comme sur la collectivité.

Ces avantages font que, couplées à des enquêtes auprès des personnes, ces bases de données peuvent faire l'objet d'utilisations très diversifiées dans le cadre des études épidémiologiques et peuvent apporter des solutions satisfaisantes à divers problèmes fréquemment rencontrés lors de la mise en œuvre de ces études, qu'il s'agisse de l'inclusion ou du suivi des sujets ou de l'accès à des données concernant des événements d'intérêt.

## Tenir compte des limites

Si ces bases de données constituent un intérêt certain, il faut toujours garder à l'esprit qu'elles ont été construites uniquement pour répondre aux objectifs de gestion des organismes qui les ont constituées. Leur utilisation par des épidémiologistes nécessite d'une part un important travail de réflexion concernant l'accès à ces données, leur appariement aux données d'enquêtes et les circuits de confidentialité à mettre en œuvre, et d'autre part un travail crucial de contrôle et de validation des données.

### L'accès aux données

L'identification des personnes dans les bases de données médico-administratives et sociales repose sur le « numéro d'inscription au répertoire », ou NIR, communément

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

appelé numéro Insee ou numéro de Sécurité sociale (voir encadré). Or, en dehors même des études épidémiologiques, l'utilisation directe de cet identifiant est soumise à de fortes contraintes juridiques (plusieurs lois et décrets définissent son accès, son usage et sa conservation dans les systèmes d'information). Il est possible de trouver des solutions à cette difficulté, mais elle constitue actuellement un obstacle formel pour la plupart des études en dehors d'un éventuel partenariat avec un organisme habilité à détenir ce numéro.

Reste ensuite un important travail pour définir les procédures de transmissions sécurisées entre les différents intervenants (fournisseurs de données, responsables de la gestion de l'étude, chercheurs), afin de garantir aux données à caractère personnel une confidentialité conforme aux textes.

Ainsi, l'accès et l'utilisation de ces bases de données restent complexes et nécessitent, dans des conditions compatibles avec les contraintes de qualité des études épidémiologiques, des moyens lourds et des compétences spécialisées. Il est vraisemblable que très peu d'équipes d'épidémiologie en France disposent actuellement de ces ressources.

### La validité des données

Comme déjà évoqué, l'utilisation de ces bases de données en dehors des champs pour lesquels elles ont été développées nécessite un travail complexe de contrôle et de validation, particulièrement dans le cas des études épidémiologiques où la précision des données concernant les événements de santé est cruciale.

Dans le cas précis des données de santé, aucune de ces bases prise isolément ne permet d'obtenir des informations complètes et d'une validité suffisante.

Les données de consommations de soins ne comportent pas d'information sur la nature des maladies traitées et excluent par définition l'automédication, les prestations non présentées au remboursement, et n'informent pas sur l'observance des traitements délivrés. Il est également établi que la prévalence des ALD enregistrées est systématiquement inférieure à la prévalence réelle des affections pour différentes raisons : patient atteint de l'une de ces maladies mais ne répondant pas aux critères de sévérité exigés ou ne demandant pas à bénéficier du dispositif, par exemple s'il est déjà exonéré du ticket modérateur au titre d'une autre affection.

La validité des diagnostics, que ce soit pour les causes de décès, les ALD ou le PMSI, dépend fortement de la qualité du codage à la production de l'information, celle-ci pouvant être affectée par différents problèmes (variabilité entre praticiens, biais liés aux finalités budgétaires du PMSI…). Plusieurs études ont montré que l'utilisation du PMSI ne pouvait pas se suffire du diagnostic principal, mais nécessitait des algorithmes complexes alliant les codes diagnostics aux codes actes spécifiques de la pathologie étudiée [10, 11].

Dans de nombreuses situations, il est donc nécessaire de mettre en place des procédures de validation de ces données. Les méthodes utilisées peuvent être variées : retour à des informations du dossier médical *via* les médecins traitants, 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 provenant de l'appariement de l'ensemble de ces bases (remboursements de médicaments enregistrés dans le SNIIR-AM, diagnostics des ALD, actes et diagnostics du PMSI).

## Conclusion

L'utilisation des bases de données d'origine socio-médico-administrative peut grandement faciliter les travaux de recherche en santé, voire améliorer la qualité des études. La résolution des problèmes évoqués pour optimiser leur utilisation pourra contribuer au développement en France de grandes cohortes comparables à celles qui existent dans d'autres pays.



# Intérêt des cohortes pour la surveillance épidémiologique : exemples dans le domaine des risques professionnels

**Béatrice Geoffroy**
**Gaëlle Santin**
**Juliette Chatelot**
Institut de veille sanitaire, Département santé-travail

## Qu'est-ce que la surveillance épidémiologique ?

La surveillance épidémiologique peut être définie comme le suivi et l'analyse épidémiologique systématiques et permanents d'un problème de santé et de ses déterminants à l'échelle d'une population [22]. Si le but de la recherche épidémiologique est d'établir des relations entre des événements de santé et leurs déterminants, celui de la surveillance épidémiologique est d'éclairer la prise de décision en matière de prévention des risques pour la santé et de prise en charge. La surveillance a une approche essentiellement descriptive ; elle s'attache à connaître et à décrire les tendances concernant la fréquence des événements de santé et la distribution de leurs déterminants au sein d'une population définie, ainsi qu'à analyser leur impact sur la santé de la population d'intérêt. Les indicateurs produits sont utilisés pour identifier des groupes à risque, définir des priorités d'actions de santé publique ou évaluer l'impact de l'évolution des facteurs de risque et des actions de prévention ou de prise en charge mises en place. Les indicateurs issus de la surveillance permettent également de soulever des hypothèses et d'orienter la recherche étiologique en cas de détection de changements inexpliqués dans les caractéristiques épidémiologiques d'une maladie [2, 31].

Les exemples de surveillance basée sur des dispositifs longitudinaux développés ci-après sont issus du domaine des risques professionnels. Si les qualités des études longitudinales pour la surveillance ne sont pas spécifiques à ce domaine, il en illustre particulièrement bien les atouts. Le monde du travail se caractérise en effet par des modifications perpétuelles de l'environnement professionnel et des conditions d'emploi, liées aux changements politiques, économiques et technologiques, susceptibles d'impacter fortement les risques professionnels et par conséquent les problèmes de santé qui leur sont liés. L'étude de la santé en relation avec le travail se heurte à de nombreuses difficultés, notamment l'absence de spécificité des maladies professionnelles, le caractère multifactoriel des pathologies étudiées, qui nécessite de prendre en compte des expositions concomitantes (à la fois professionnelles et extraprofessionnelles), la survenue des pathologies souvent différée dans le temps par rapport à l'exposition professionnelle, notamment les cancers qui surviennent le plus souvent chez les personnes retraitées. Les dispositifs de surveillance longitudinaux permettent de pallier certaines de ces difficultés.

## Atouts des études de cohorte pour la surveillance épidémiologique

Plusieurs types d'études longitudinales peuvent être initiés dans le contexte de la surveillance et correspondent à des objectifs différents [8].

### Surveillance de populations spécifiques

Cette surveillance cible une population partageant des caractéristiques d'exposition communes. Ce type d'étude est généralement mené dans le but principal de détecter des événements de santé dont l'incidence serait jugée anormalement élevée par rapport à une population de référence. Dans le domaine des risques professionnels, il peut s'agir de personnes ayant la même profession, travaillant dans le même secteur d'activité, dans la même entreprise, ou exposées à une même nuisance.

L'initiation de cohortes constitue une approche intéressante pour la surveillance à l'échelle de l'entreprise. Il s'agit de reconstituer de manière rétrospective la population employée – suivie au-delà du départ à la retraite – afin de dresser un premier bilan relatif à la mortalité observée. La cohorte constituée peut ensuite servir de base à la mise en place d'une surveillance au long cours de la santé des personnels employés. Ce type d'outil permet ainsi, à l'échelon de l'entreprise, d'orienter les actions de prévention, de surveiller l'impact de l'adoption de procédés nouveaux et d'évaluer les mesures préventives mises en place. Elle facilite par ailleurs l'analyse d'éventuels signaux suspects signalés par la médecine du travail, et permet de répondre rapidement et de manière rationnelle à des préoccupations des partenaires sociaux relatives à la santé.

De même, pour documenter l'impact sur la santé de procédés nouveaux dont on suspecte le caractère nuisible pour la santé, il peut être intéressant d'initier des cohortes de travailleurs exposés et de mettre en place un suivi prospectif systématique sans *a priori* sur les conséquences de santé attendues. De telles cohortes doivent permettre de générer des hypothèses quant à la nocivité du/des produit(s). Ce type de surveillance est particulièrement pertinent dans le cas d'expositions relativement rares en population générale : c'est le cas, par exemple, de la production et l'utilisation industrielles de nanomatériaux, actuellement en plein essor, et qui font l'objet de préoccupations du point de vue de la santé des travailleurs — des risques pour la santé sont suspectés de par la taille inhabituelle de ces poussières, qui leur conférerait un potentiel de

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

nuisance spécifique. Ce type de dispositif présente en outre l'avantage d'être évolutif, le suivi de santé et le recueil de données afférentes pouvant être adaptés en fonction de la progression des connaissances. Une telle cohorte constitue une population déjà identifiée et accessible pour mener d'éventuelles études de recherche étiologique.

### Surveillance en population générale

Par ailleurs, des cohortes peuvent être initiées en population générale. Conçues comme un véritable observatoire de la santé des travailleurs au long cours, elles représentent le seul dispositif permettant de disposer d'une « *image évolutive de la réalité des risques professionnels à l'échelle de la population* » [22]. De part leur protocole, ces études sont théoriquement à même de produire une grande variété d'indicateurs propres à la population d'intérêt (fréquence des pathologies, prévalences et caractéristiques des expositions, mesures d'association entre l'exposition et la pathologie), en tenant compte de la temporalité des événements, des expositions conjointes, des temps de latence de certaines pathologies. Ce type d'étude est ainsi à même de documenter le poids des facteurs professionnels sur la santé à l'échelle populationnelle. La surveillance de ces indicateurs au fil du temps permet d'étudier les changements au regard de l'évolution des procédés et de la mise en œuvre de mesures préventives, ou d'alerter sur des modifications des caractéristiques épidémiologiques d'une maladie en relation avec les facteurs professionnels. En outre, ces cohortes en population générale sont susceptibles d'apporter rapidement des arguments en faveur d'une association entre une exposition et une pathologie, suggérée par d'autres signaux (exemple : observation de cas groupés de pathologie).

Un outil classique pour la surveillance des risques professionnels repose sur l'étude de la mortalité par cause et par profession. Les systèmes basés sur des échantillons longitudinaux de population ont pour avantage de permettre de disposer de taux de mortalité en population et de tenir compte de la carrière entière, contrairement aux systèmes classiques basés sur les seuls certificats de décès pour lesquels seule la dernière activité professionnelle est généralement renseignée [33].

En ce qui concerne les cohortes prospectives en population, elles offrent de nombreuses possibilités et modularités pour la surveillance épidémiologique. Dans le cas notamment des études par questionnaire, le recueil des données peut être planifié et couvrir des facteurs professionnels et extraprofessionnels variés, éventuellement intriqués. Dans le cadre du suivi, le recueil continu des informations sur l'état de santé et sur les facteurs de risque permet de disposer de mesures répétées dans le temps des expositions professionnelles nécessaires pour documenter des changements des conditions de travail au niveau individuel, mais également d'adapter le recueil de données selon des problématiques émergentes en termes d'exposition ou d'état de santé. En cas de détection de phénomènes de santé inexpliqués, il est enfin possible de greffer sur ce dispositif d'éventuelles études ciblées à visée analytique.

### Contraintes méthodologiques

Afin d'atteindre ces objectifs, la surveillance épidémiologique doit s'appuyer sur des indicateurs fiables, reproductibles dans le temps mais, surtout, reflétant la réalité de la situation *à l'échelle de la population d'intérêt*.

Ainsi il est nécessaire que la population d'étude soit « représentative » de la population cible. Cela signifie qu'il doit être possible, à partir des données issues du groupe de personnes suivies, d'obtenir des estimations extrapolables à la population d'intérêt. Il est donc fondamental de contrôler au mieux les effets de sélection. Pour les enquêtes en population générale, cela nécessite notamment que l'échantillon étudié soit constitué par tirage au sort dans la population cible.

L'équilibre assuré par le tirage au sort est cependant rompu dès que l'information est manquante pour certaines personnes. En effet, cette non-réponse est susceptible d'entraîner des biais de sélection, si les phénomènes étudiés sont liés à la participation. Ce problème se pose non seulement à l'inclusion, mais également au fil du suivi, et quel que soit le type de recueil de données.

Les données issues de sources externes collectées en routine (causes médicales de décès, déclarations administratives par les employeurs, consommations de soins,…), de par leur enregistrement systématique, sont moins susceptibles d'entraîner des biais de sélection. En revanche, le recueil d'information direct auprès des personnes dépend de la capacité à contacter la personne et de sa volonté et capacité à répondre, lesquelles peuvent être liées au phénomène étudié (état de santé en particulier). Dans ce cas, il est possible que les estimations obtenues sur le sous-groupe des personnes répondantes ne reflètent pas la situation de la population d'intérêt, par exemple, si les fumeurs participent plus que les non-fumeurs à une enquête cherchant à estimer la prévalence de consommation de tabac.

Il existe cependant des solutions pour corriger des biais de sélection éventuels lorsqu'on dispose chez les participants et les non-participants d'informations en lien avec le phénomène étudié (telles que les données de l'Assurance maladie). Ainsi, dans le cas d'études de cohorte sur échantillons de population, ce type d'information recueilli en continu au fil du suivi peut être utilisé pour appréhender au mieux les biais de sélection potentiels (lire *Aspects méthodologiques liés à l'analyse de données longitudinales et aux effets de sélection, p. 18*).

Par ailleurs, dans la plupart des cas, la population d'intérêt évolue au fil du temps. Pour qu'une cohorte de surveillance permette d'obtenir des estimations extrapolables à la population cible au fil du suivi, il



est indispensable de tenir compte de l'évolution de la composition de cette dernière. Certaines personnes de la cohorte initiale peuvent ne plus faire partie de la population cible à la date d'observation, tandis que de nouvelles personnes y sont entrées depuis. Afin de maintenir la capacité à décrire la population de manière transversale et prendre en compte au fil du temps l'évolution des facteurs de risque, il est nécessaire de mettre en place une cohorte dite « ouverte », c'est-à-dire avec inclusion au fil du temps des nouveaux entrants dans le champ de la population d'intérêt. Dans le cas d'étude sur échantillon de population, cela suppose de tirer au sort périodiquement et suivre de nouveaux éligibles dans la population cible. Dans le cas de la surveillance des risques professionnels, compte tenu des changements importants du tissu socio-économique, cette contrainte est fondamentale afin de tenir compte des travailleurs jeunes, des procédés nouveaux, etc.

### Conclusions

Dans le domaine de la surveillance, les études de cohorte représentent un outil majeur pour pallier la plupart des écueils des autres dispositifs classiques tels que les problèmes de temporalité, de latence ou de prise en compte de cofacteurs. Elles permettent de disposer d'une image évolutive des pathologies en lien avec les facteurs d'intérêt et de surveiller l'impact de l'évolution des risques. S'il n'est pas question dans ce contexte d'interpréter les résultats en termes de causalité étant donné l'absence d'objectif spécifique de ce type d'études, elles permettent cependant de générer des hypothèses pour la recherche. Leur protocole facilite en outre la mise en place d'études *ad hoc*. Cependant, la capacité des études à atteindre ces objectifs est totalement dépendante de la possibilité de recueillir l'information auprès d'un échantillon représentatif de la population pour laquelle on souhaite disposer d'indicateurs de santé ou d'exposition. Cette condition de représentativité doit théoriquement être réalisée à tout moment (représentativité transversale) et au long du suivi pour tous les sujets inclus. Cela suppose le plus souvent de mettre en place des dispositifs « ouverts » permettant d'inclure périodiquement de nouveaux sujets et, surtout, de mettre en œuvre tous les moyens possibles pour lutter contre la non-réponse à l'inclusion et l'attrition, et pour documenter et prendre en compte au mieux les effets de sélection afin d'obtenir des indicateurs extrapolables à la population cible surveillée.

# Apport des sciences sociales : l'exemple de cohortes de patients infectés par le VIH

**Bruno Spire**
Inserm-Sesstim,
UMR 912, Marseille

Les cohortes représentent un outil idéal pour mener des études multidisciplinaires à l'interface de l'épidémiologie médicale et des sciences sociales. L'évolution médicale d'individus concernés par un problème de santé peut être ainsi analysée de façon holistique en tenant compte du comportement et des perceptions des intéressés. Ces études se réalisent par la mise en place de questionnaires remplis par les patients régulièrement distribués au fur et à mesure du déroulement de la cohorte. Ces questionnaires sont conçus généralement en tenant compte des travaux qualitatifs menés préalablement sur des patients concernés par la pathologie d'intérêt. Nous prendrons comme exemple les cohortes de patients infectés par le VIH. Les travaux se sont principalement centrés sur l'observance au traitement, mais aussi sur la qualité de vie des patients traités. Deux cohortes ont recueilli des informations socio-comportementales, la cohorte Aproco/Copilote de la 1re génération de patients initiant une multithérapie avec antiprotéase, et la cohorte Manif2000 incluant des patients infectés par usage de drogue intraveineuse. La cohorte Aproco a inclus 1 281 patients entre 1997 et 1999 dans 47 centres en France ; la cohorte Manif2000 a inclus 467 patients entre 1995 et 1997 dans 8 centres des régions Paca et de la banlieue parisienne.

### L'importance de l'observance pour l'infection à VIH et ses particularités

Les progrès significatifs des traitements antirétroviraux hautement actifs ont relancé la problématique de l'observance. Plusieurs travaux ont mis en évidence l'observance comme facteur majeur associé au succès virologique, à la baisse de la progression clinique et de la mortalité. Le niveau d'observance requis pour assurer la meilleure réponse à long terme des multithérapies reste cependant une question ouverte. Les cohortes ont permis de suivre au cours du temps la capacité des patients dans la vie réelle à être observants et de mesurer l'impact au cours du temps de la non-observance. À partir des questionnaires administrés aux patients, des algorithmes de classification ont été établis en classant les patients comme hautement observants, modérément observants ou non observants au cours des 4 derniers jours. Ces questionnaires ont été validés en indiquant une bonne corrélation entre observance et

succès virologique et en démontrant la relation entre observance et concentrations plasmatiques d'inhibiteurs de protéase. La validité des questionnaires a été confirmée dans différentes populations très variées : les migrants d'Afrique subsaharienne vivant en France, les usagers de drogues injectables, les patients vivant au Cambodge ou au Cameroun.

### Une approche dynamique et non prédictive de l'observance

Les travaux réalisés dans ces cohortes ont permis de montrer que l'observance est un phénomène dynamique qui se modifie au cours du temps. Dans la cohorte Aproco/Copilote, seulement 26 % des patients restent hautement observants tout du long de 36 mois de suivi. 64 % ont parfois une observance élevée, et 10 % jamais.

Grâce à l'approche longitudinale, l'approche prédictive de l'observance visant à identifier *a priori* les facteurs expliquant une non-observance future a été écartée. En effet, l'analyse des déterminants de l'observance a été recherchée dans la cohorte Aproco/Copilote. Un nombre limité de caractéristiques mesurées avant traitement sont associées à la non-observance initiale. En revanche, la non-observance est mieux expliquée par les variables mesurant le vécu des patients après la mise sous traitement. Les effets secondaires perçus par le patient sont déterminants pour expliquer la non-observance, aussi bien à court terme dans son établissement qu'à plus long terme pour expliquer les ruptures d'observance. Chez les usagers de drogue de la cohorte Manif2000, les patients les moins observants sont les toxicomanes actifs ne bénéficiant pas de traitement de substitution. Ceux qui ont continué ou qui ont repris les pratiques d'injection ont plus de risque de présenter un comportement de rupture d'observance. De plus, les résultats démontrent l'impact de la précarité sociale chez les ex-usagers de drogue sur l'observance ; en revanche, chez les sujets qui restent dépendants, c'est une substitution efficace en réduisant l'injection qui est associée à une bonne observance. Ces résultats suggèrent que la prise en charge des toxicomanes séropositifs nécessite une appréhension globale de la toxicomanie, en tenant compte de l'ensemble de la problématique du patient et pas seulement du VIH.

### L'observance est encore plus capitale au début du traitement

Les données cliniques, immuno-virologiques et comportementales ont été recueillies à (quatrième mois suivant le début du suivi) M4, M12, M20, M28 et M36 après l'initiation du traitement chez les 1 281 patients de la cohorte Aproco. La suppression prolongée de la réplication virale à M28 et M36 et un gain d'au moins 200 CD4/mm$^3$ ont été utilisés comme critères de succès virologiques et immunologiques. Parmi les 582 patients suivis régulièrement jusqu'à M36, 360 patients ont des données complètes sur l'observance. Bien que 59 % soient complètement observants à M4, seulement 26 % ont maintenu un taux d'observance complète au cours des 36 mois de suivi. L'observance complète à M4 est associée à la fois à la suppression prolongée de la réplication virale et à un gain de CD4 > 200/mm$^3$ au cours des 3 années de traitement. Cependant, les patients modérément observants entre M12 et M36 ont une probabilité similaire de réponse virologique prolongée à celle des patients restés toujours complètement observants, les patients ayant présenté des épisodes de non-observance entre M12 et M36 ayant moins fréquemment une réponse prolongée. L'optimisation de l'observance semble cruciale pendant les premiers mois qui suivent l'initiation des multithérapies pour garantir l'efficacité immuno-virologique à long terme. Des déviations modérées de l'observance au cours du suivi ultérieur semblent avoir un impact moindre. Les interventions pour améliorer l'observance doivent être privilégiées au moment des premiers mois suivant la mise sous traitement.

### La dépression joue sur la progression clinique indépendamment de l'observance

La question de l'impact de la dépression sur la progression clinique avait été largement ouverte avant l'arrivée des multithérapies. La recherche d'un éventuel impact de la dépression sur la progression clinique des patients sous traitement ne peut s'étudier qu'en tenant compte de l'observance, puisque cette variable est directement associée à la non-observance. Au sein des deux cohortes Manif et Aproco/Copilote, la dépression est mesurée au décours de l'initiation du traitement par l'échelle CES-D. Cette échelle CES-D contient 20 questions qui génèrent un score hautement prédictif de la dépression. La dépression est associée à un risque accru de progression immuno-clinique et cela de façon indépendante de l'observance. Dans la cohorte Aproco/Copilote, la progression clinique était basée sur les événements cliniques classant sida, alors que dans la cohorte Manif, la progression clinique était définie par le fait d'avoir des CD4 < 200, étant donné que les patients inclus dans cette cohorte avaient des CD4 initiaux plus élevés (>35 0) et donc peu à risque de développer des événements cliniques. Ce résultat suggère que des mécanismes neuro-immunitaires pourraient jouer un rôle dans la progression de la maladie.

### Analyse des facteurs associés à l'observance au long cours

Ce type d'approche présente des difficultés d'analyse car la sélection des patients suivis et répondant régulièrement aux auto-questionnaires pose des problèmes de biais de représentativité à cause des données manquantes. Les facteurs associés à l'observance à long terme ont pu être étudiés tout en tenant compte des biais induits par les données manquantes. Celles-ci sont fréquentes dans toutes les études de cohorte car il existe une attrition naturelle (décès, perdus de vue, abandons). De plus, les données spécifiques aux auto-questionnaires sont également manquantes même si les patients sont toujours suivis dans la cohorte,



soit par refus de remplir le questionnaire ou par non-remise du questionnaire. L'analyse des données a été effectuée par une méthode statistique spécifique pour tenir compte du fait que les données manquantes ne sont pas dues au hasard (méthode d'Heckman) et corriger ainsi les biais potentiels : après correction du biais, la non-observance est indépendamment associée à l'âge jeune, un nombre élevé d'effets secondaires perçus, un traitement monoprise ou comprenant trois prises ou plus par jour, une combinaison avec antiprotéase, un score élevé de dépression et l'absence de soutien du partenaire principal. De plus, les patients nés hors Union européenne sont plus souvent retrouvés observants. Le groupe de transmission par toxicomanie et les mauvaises conditions de logement sont associés à la non-observance seulement si le biais induit par les données manquantes n'est pas corrigé.

## La qualité de vie des patients traités par antirétroviraux

### Le rôle néfaste des effets secondaires perçus

La cohorte Aproco a aussi comme objectif d'envisager l'étude de la qualité de vie des patients traités par multithérapie et de mesurer l'impact du traitement. L'échelle SF-36 qui a été choisie est une échelle générique qui explore 4 dimensions physiques et 4 dimensions mentales de qualité de vie pour laquelle il existe des valeurs de référence dans la population française en fonction de l'âge et du sexe. Une qualité de vie est considérée comme normale si les patients ont des scores de 3 échelles sur 4 physiques et 3 sur 4 mentales supérieures au 25e percentile des valeurs de la population générale. Le traitement semble montrer un impact positif sur la qualité de vie puisque la proportion de patients avec une bonne qualité de vie allait de 36 % avant traitement à 46 % après un an de traitement ($p = 0{,}001$) ; la qualité de vie est influencée à la fois par l'efficacité du traitement, mais aussi négativement par les effets secondaires perçus, en particulier par la lipodystrophie. Chez les patients infectés par voie toxicomane, la perception des effets secondaires est plus élevée, et les traitements de substitution améliorent la qualité de vie sans toutefois leur permettre de rejoindre celle des patients qui ont pu arrêter toute dépendance aux opiacés. La perception des effets secondaires est un facteur prédictif important de rupture de confiance entre le médecin et son patient. Dans une analyse menée après trois années de traitement centrée sur les scores agrégés de qualité de vie physique et qualité de vie mentale, les mêmes facteurs expliquent de bons scores de qualité de vie, mais on peut démontrer également le rôle indépendant de la relation de confiance avec le médecin prescripteur sur la qualité de vie mentale et la satisfaction des explications fournies par le médecin sur la qualité de vie physique.

### Le rôle des croyances sur la santé

Les caractéristiques psychologiques des patients avant traitement peuvent également expliquer en partie la qualité de vie après plusieurs années de traitement. Le « locus de contrôle » recueilli lors de l'inclusion dans Aproco a été mesuré. Il s'agit d'une croyance généralisée selon laquelle les événements ultérieurs dépendent soit de facteurs internes, soit de facteurs externes. Dans le cas du locus interne, l'individu établit un lien causal entre ses actions et son état de santé alors que, pour le locus externe, le patient pense soit qu'un personnage tout-puissant (le médecin) peut contrôler son état de santé soit qu'elle est sous l'influence exclusive de la chance. La mesure du « locus de contrôle » se fait grâce à trois sous-scores, un pour le locus interne, un pour le locus externe « personnage tout-puissant » et un pour le locus externe « chance ».

Les résultats montrent qu'un locus externe « personnage tout-puissant » élevé a une influence défavorable sur la qualité de vie mentale à M44, ainsi qu'un nombre élevé d'effets secondaires déclarés et le fait de ne pas avoir atteint le stade sida. En revanche, un locus interne élevé influence positivement la qualité de vie physique à M44, ainsi que le fait d'avoir un emploi et un faible nombre d'effets secondaires déclarés.

Au total, l'expérience des cohortes a montré que, une fois le traitement instauré, la qualité de vie mesurée par des échelles génériques est assez stable. La composante la plus sensible au changement de la qualité de vie dépend des effets secondaires perçus par les patients. Ces derniers semblent être l'indicateur le plus simplifié et le plus utile pour évaluer la qualité de vie au sein d'essais cliniques.

# Coordination et partage de données de cohortes

**Georges Dagher**
Infrastructure nationale Biobanques, Inserm US 13, Paris

En 2020, 75 % des décès au niveau mondial seront dus à des maladies chroniques complexes. Les progrès au niveau des technologies de haut débit, notamment en génomique, ont permis des avancées importantes, notamment en termes de définitions de sous-groupes de pathologies. De nombreuses recherches sont articulées autour d'études de caractéristiques, de différences et de singularités génétiques. Dans le champ thérapeutique, les travaux visent à développer une médecine personnalisée qui répond efficacement et rapidement aux besoins des patients. Dans ce cadre, les échantillons biologiques et leurs données associées sont essentiels pour :

- Élucider les interactions entre des facteurs génétiques et environnementaux responsables de pathologies, comprendre les mécanismes des maladies complexes et des maladies rares.
- Développer des programmes médicaux ajustés aux particularités des patients et savoir, en fonction de facteurs de risque identifiés, adapter la prévention, diagnostiquer précocement et orienter vers un traitement adapté.
- Déterminer les influences des facteurs génétiques sur les effets et les résultats des traitements, prévenir les effets délétères, fournir des médicaments sûrs, efficaces et adaptés en fonction des particularités individuelles.
- Identifier de nouvelles cibles thérapeutiques, développer de nouveaux médicaments ou améliorer ceux existants.

La recherche utilisant ces ressources, telles que la génomique, la protéomique ou encore l'imagerie, a accompli des avancées majeures dans la compréhension des facteurs physiopathologiques qui sous-tendent les maladies complexes et rares. Les études d'associations génétiques (GWAs) ont permis l'identification d'associations génétiques dans le diabète type 1 et type 2, la pathologie coronarienne, le cancer colorectal, du sein, ou de la prostate, la dégénérescence maculaire, la maladie de Crohn, l'autisme, les maladies neurodégénératives. Malgré ces avancées dans la détection des associations génétiques avec les pathologies complexes, les déterminants génétiques ne représentent qu'un des facteurs qui jouent un rôle dans leur développement. L'influence du mode de vie, des facteurs environnementaux et sociaux a été reconnue et explorée depuis bien longtemps. Cela implique que l'interaction entre le gène et l'« environnement » joue un rôle essentiel dans la chaîne causale. Il est alors important que les sciences biomédicales aient accès non seulement aux études génétiques, mais également aux données épidémiologiques qui incluent le mode de vie, et les informations socio-économiques et environnementales.

Par ailleurs, la recherche biomédicale rencontre différents obstacles qui limitent le développement de la recherche étiologique, de la recherche translationnelle ou du développement de nouvelles molécules thérapeutiques. Parmi ces obstacles figure l'accès aux échantillons biologiques provenant de cohortes d'individus bien définis sur le plan clinique et, en particulier dans le domaine de l'épidémiologie génétique, la nécessité d'accès à un grand nombre de sujets pour identifier des associations génétiques significatives sur le plan statistique dans les maladies complexes. Pour relever ces défis, il est indispensable de s'assurer de la qualité des échantillons, de la standardisation ou de l'harmonisation de leur collecte, transformation et conservation. De plus, il est nécessaire d'élargir le champ des pathologies concernées, d'aboutir à une interopérabilité des bases de données, et bien entendu de garder la confiance qu'accordent les citoyens à ces activités.

## Biobanques : un outil de mutualisation des bases de données biologiques

Le succès de cette entreprise repose principalement sur la qualité de l'échantillon et celle des informations qui lui sont associées. Il est donc important que la communauté de la recherche biomédicale coordonne mieux ses efforts pour assurer le recueil et la conservation d'échantillons biologiques dans des conditions techniques qui en garantissent la qualité et la possibilité de mise en commun pour des analyses groupées. C'est l'objectif majeur de l'infrastructure nationale Biobanques retenue dans le cadre du programme national « Investissements d'avenir ».

Biobanques[1] mobilise, au sein d'une infrastructure distribuée sur tout le territoire français, les biobanques, tumorothèques, centres de ressources microbiologiques (mBRCs) et cohortes, ainsi que des expertises associées aux collections d'échantillons biologiques. L'infrastructure s'appuie sur les acquis d'un solide réseau structuré depuis 2005, et couvre tous les champs de la recherche biomédicale. Le projet fédère 72 biobanques, dont 8 mBRCs, et implique de nombreuses équipes de recherche à travers des projets utilisant les collections d'échantillons biologiques.

La mise en place de l'infrastructure Biobanques se déroule en deux phases. Une phase de construction (2011-2016), destinée à coordonner et harmoniser les

**1.** http://www.crbfrance.fr/



différentes activités existantes ; cette phase assurera la mise en place et la préparation des services communs et des plates-formes technologiques qui seront opérationnelles progressivement au cours des prochaines années. Ultérieurement, une phase opérationnelle sera mise en œuvre (2017-2019), durant laquelle l'infrastructure fournira services et plates-formes à l'ensemble de la communauté scientifique.

L'objectif de Biobanques est d'intégrer les collections d'échantillons biologiques, les technologies et les expertises associées afin de les pérenniser et de les enrichir, cela dans le respect des cadres éthiques et juridiques français et européens. Par ailleurs, l'infrastructure assure une coordination plus efficace des actions des biobanques et des mBRCs, ainsi que la valorisation des collections au sein de projets de recherche d'excellence, tant académiques que privés. Pour atteindre ces objectifs, l'infrastructure mutualisera les moyens et les compétences en développant des services communs tels que : affaires réglementaires et éthiques, qualité de l'échantillon, bioinformatique, unité de méthodologie et de biostatistique, ainsi que des plates-formes technologiques destinées à la communauté scientifique.

Parmi les services communs figure le service bioinformatique et bases de données, clés de voûte de la recherche biomédicale, notamment génomique et protéomique. Biobanques développera, dans un premier temps, un projet pilote afin d'équiper et de tester l'interopérabilité des bases de données cliniques et analytiques provenant de projets portant sur une sélection restreinte de maladies complexes et de maladies rares. Ce projet pilote permettra d'identifier la fiabilité et la validité des modules informatiques proposés ainsi que les goulots d'étranglement et les difficultés de leur mise en œuvre. Parmi celles-ci figure l'accès aux données cliniques pseudonymisées ou anonymisées des patients. Il faudrait développer une solution sécurisée permettant le transfert, à partir des services hospitaliers, d'une partie de ces données à des fins de recherche et cela dans le respect de la confidentialité et de la volonté du patient.

## Harmoniser les données des cohortes de pays européens

Par ailleurs, l'harmonisation par pathologie du contenu de ces bases est une nécessité pour collecter les données d'un nombre suffisant de patients, dépassant souvent les 10 000 sujets, afin d'atteindre la puissance statistique suffisante pour l'analyse des polymorphismes génétiques des maladies complexes. Une telle harmonisation peut être également utile dans la synthèse des informations provenant de différentes études ou encore dans le développement de nouveaux projets. Plusieurs initiatives dans cette perspective ont été entreprises au niveau international. Par exemple, l'harmonisation des données de l'étude Epic, qui associe des cohortes de plusieurs pays européens, est un élément essentiel de son succès ; il en est de même pour différentes autres études telles que GenomeEUTwin, Euralim ou encore Engage. De plus, le consortium P3G (Public Population Projects in Genomics) a entrepris depuis plusieurs années une harmonisation des bases épidémiologiques en association avec le projet européen Phoebe. Il a ainsi développé un outil pratique dénommé Datashaper[2], qui inclut deux séries de données, des variables primaires nécessaires à toute étude épidémiologique et des variables d'ajustement, et qui décline ces variables primaires en une série d'informations. Datashaper est un travail collaboratif impliquant plus de 25 cohortes (études longitudinales de population) internationales provenant de 14 pays. Biobanques, en concertation avec P3G, encouragera l'accès à Datashaper et l'utilisation des outils informatiques annexes développés par le consortium.

Biobanques a pour objectifs de porter les collections d'échantillons biologiques d'origine humaine et les collections microbiologiques à un nouveau niveau de coordination, de qualité et de valorisation en développant une infrastructure pérenne et reconnue à l'échelle nationale et internationale. L'infrastructure visera en particulier à atteindre les objectifs suivants :

- Accroître l'excellence scientifique et l'efficacité de la recherche française dans les sciences de la vie.
- Atteindre une masse critique suffisante en termes de recherche et d'investissements, et éviter la duplication des efforts en établissant des liens avec d'autres projets européens et internationaux, en améliorant la qualité et la standardisation des bioressources et de données associées, et en mettant en place l'interopérabilité des bases de données. Cela en étroite collaboration avec les infrastructures européennes BBMRI[3] et Embarc[4].
- Faciliter l'accès des chercheurs académiques et privés aux ressources biologiques et aux données associées afin de favoriser l'innovation et la compétitivité, accélérer la mise en place de partenariats public-privé.

**2**. http://www.datashaper.org/
**3**. http://www.bbmri.eu/
**4**. http://www.embarc.eu/