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.