AFCROs

LES ENTREPRISES DE LA RECHERCHE CLINIQUE

APPARIEMENT DE DONNÉES PRIMAIRES ET SECONDAIRES EN FRANCE : QUAND ? COMMENT ? ET POUR QUELS USAGES ?

Article rédigé par

Manon Belhassen, PELYON
Françoise Bugnard,
STEVE CONSULTANTS
Magali Lemaitre, HORIANA
Frédérique Maurel, IQVIA
Lise Radoszycki, CARENITY
Membres du groupe de travail
RWD de l'AFCROs

09/2023

L'appariement de données primaires et secondaires correspond au rapprochement de données de plusieurs bases sources distinctes. Les données primaires sont des informations spécifiquement collectées pour étudier un objectif particulier. Les données secondaires sont des informations qui ont déjà été collectées dans un but différent de celui de l'étude menée et qui sont à disposition pour une seconde utilisation. L'intérêt de ce rapprochement est de maximiser la qualité et la quantité de variables disponibles pour les analyses, et répondre ainsi de façon optimale à la question posée. Par exemple, une étude clinique dont l'objectif est d'évaluer le taux de réponse d'un traitement sur quelques mois sera peu informative pour le taux de survie.

Des informations non recueillies au cours d'une étude clinique ou observationnelle peuvent être obtenues dans le SNDS (Système National des Données de Santé), comme la consommation de soins avant la période d'étude ou pendant le suivi, et le taux de survie après la fin de l'étude.

A l'inverse, on peut chaîner les données du SNDS avec des registres nationaux, par exemple pour vérifier la représentativité des populations incluses, ou pour compléter les données du SNDS par des données cliniques et ainsi obtenir le profil et le parcours exhaustifs des patients.

Il est également possible d'enrichir les données du SNDS avec des données générées directement par les patients au travers de questionnaires (Patient-Reported Outcomes) afin de décrire le fardeau de la maladie, le vécu et les attentes des patients vis-à-vis de leur prise en charge. Une finalité de l'appariement peut également être le développement et la validation d'algorithmes de ciblage dans le SNDS, comme le propose l'initiative BOAS (Bibliothèque Ouverte d'Algorithme en Santé). Selon l'objectif poursuivi, il sera important de définir le moment le plus opportun pour réaliser l'appariement.

Quand prévoir l'appariement des données ?

Les données du SNDS sont mises à disposition annuellement avec un temps de latence variant entre 5 et 9 mois. On considère en effet de façon générale que les données d'une année sont disponibles au 2ème ou 3ème trimestre de l'année suivante. Ce temps nécessaire à la mise à disposition des données peut se traduire si on n'y prête pas garde par un décalage entre la disponibilité des données primaires et des données du SNDS et ainsi induire un délai de réalisation du traitement des données et de la disponibilité des résultats. Il convient donc de planifier au mieux les projets basés sur un appariement de différentes sources de données afin d'obtenir les résultats au moment voulu en fonction de l'objectif fixé, notamment dans le cadre de négociations avec les autorités.

Par exemple, dans le cas d'un appariement entre les données du SNDS et les données d'une étude observationnelle dont la période d'inclusion est de 6 mois, si l'on veut faire coïncider au mieux la disponibilité des données entre les deux sources de données, il est conseillé dans la mesure du possible de prévoir un début des inclusions suffisamment tôt une année donnée afin que le recrutement se termine sur cette même année.

Par ailleurs, la mise en œuvre d'un appariement entre des données primaires et secondaires nécessite de suivre un circuit réglementaire préétabli qui peut s'avérer plus ou moins long selon les cas de figure et le type de données appariées. De la même façon, une attention particulière doit être apportée à la planification des démarches réglementaires et à leur imbrication et interrelation. Il est par exemple possible de prévoir de scinder les circuits réglementaires afin de conduire la collecte des données primaires pendant la réalisation des soumissions réglementaires inhérentes au volet de l'étude concernant les données appariées, afin une nouvelle fois de faire concorder au mieux la disponibilité des données provenant des différentes sources.

Comment apparier les données ?

Un préalable indispensable pour le chainage de deux sources de données entre elles est l'existence de variables communes qui permettent de faire le rapprochement des individus. En pratique, il existe deux façons d'apparier des données primaires à des données secondaires :

1. Les appariements directs (également appelé « déterministes ») réalisés à partir d'une ou plusieurs variables permettant d'identifier clairement la personne dans les deux sources de données (numéro d'identification tel que le NIR[1] ou numéro de Sécurité Sociale) ;
2. Les appariements indirects (également appelé « probabilistes »), réalisés en l'absence de données communes directement identifiantes, au moyen de plusieurs variables communes suffisamment discriminantes (par exemple âge, sexe, dates et types de soins, prescriptions médicamenteuses, codes du producteur de soins, etc.).

Le circuit des appariements se fait en plusieurs étapes et peut nécessiter l'intervention d'intermédiaires de différentes natures selon la situation pour sa sécurisation (tiers de confiance pour la centralisation des données ou la reconstitution du NIR ou encore pour la mise en forme des fichiers d'identité).

Dans le cadre d'un chainage direct de données primaires avec les données du SNDS, la table des NIR associés à un numéro de sujet généré aléatoirement (différent de la base source, appelé identifiant temporaire « d'accrochage ») doit être transmise à l'équipe DEMEX de la CNAM (selon le process sécurisé mis en place par la CNAM) pour qu'elle puisse procéder à l'extraction des données du SNDS souhaitées correspondant aux NIR de la table transmise. La réconciliation peut ensuite être faite avec les données primaires grâce à l'identifiant d'accrochage. La procédure passe par un cryptage irréversible des informations identifiantes pour transformer le NIR en clair en identifiant SNDS crypté (NIR haché) au moyen de l'outil SAFE opéré par la CNAM (également appelée « circuit de pseudonymisation »).

[1] NIR : Numéro d'inscription au répertoire national d'identification des personnes physiques.

Dans le cas d'un appariement indirect, l'appariement est effectué à partir d'une liste de variables communes préalablement identifiées (exemple : sexe, dates de soins, année de naissance…). La recherche de correspondance est réalisée en général en plusieurs étapes par tâtonnement, afin d'obtenir d'une part un taux de patients appariés satisfaisant (pas ou peu d'attrition des sujets analysés) et d'autre part, une unicité (sujet unique correspondant). Dans le cas d'un appariement avec une étude clinique ou une étude observationnelle, l'identification des variables discriminantes potentielles doit être réalisée en amont de la mise en place de l'étude afin de s'assurer de leur recueil. L'envoi, la transmission et le traitement des données doivent toujours être réalisés en milieu sécurisé définis avec la CNAM et répondant aux exigences du référentiel de sécurité applicable au SNDS.

Dans tous les cas, réaliser un appariement nécessite que le projet soit d'intérêt public, qu'il ait obtenu un avis favorable du CESREES et une autorisation de la CNIL. Les sujets doivent au préalable être informés et ne pas s'opposer pas à la réutilisation de leurs données

Pour plus d'informations sur ces appariements et les process mis en place :

- Appariement de données avec le SNDS s'appuyant sur le numéro de sécurité sociale (NIR) : la CNIL publie un guide pratique | CNIL

- Schémas d'appariement et de circulation des données (health-data-hub.fr)

Comment informer les patients ?

Lors de la mise en place d'un projet nécessitant la réutilisation de données initialement collectées dans le cadre d'une autre finalité, il est nécessaire d'informer le patient de manière individuelle de cette nouvelle recherche. Cette lettre d'information patient doit spécifier l'objectif de l'étude, le responsable de traitement, le type de données utilisées, leur origine et la durée de réalisation du projet. Enfin, le patient peut s'opposer à la réutilisation de ses données dans le cadre de cette recherche, par conséquent, cette lettre précise les personnes à contacter afin d'exercer ses droits.

Dans le cadre d'un appariement aux données du SNDS, il est important d'expliquer de manière pédagogique ce que représente le SNDS, comment l'appariement est réalisé et de préciser que ces données seront traitées de façon pseudonymisée, par des personnes habilitées et dans un espace sécurisé.

En complément de cette note d'information individuelle, le responsable de traitement et le responsable de mise en œuvre du traitement doivent mettre en place une information collective sur leur site Internet reprenant l'ensemble des éléments présentés dans la note d'information individuelle.

6 CAS D'USAGE POUR LESQUELS L'APPARIEMENT DE DONNÉES PRIMAIRES ET SECONDAIRES EST PERTINENT

1. L'évaluation de la télésurveillance

Les systèmes de télésurveillance permettent de recueillir des données de surveillance clinique du patient. L'évaluation de ces systèmes demandée par les autorités nécessite d'apparier les données recueillies aux données du SNDS afin de disposer d'informations complémentaires sur l'état de santé du patient. Outre une évaluation épidémiologique de l'efficacité de ces dispositifs, l'appariement aux données du SNDS permet de conduire une évaluation économique. Cet appariement permet aussi d'allonger la durée de suivi des patients et d'évaluer l'efficacité des systèmes de télésurveillance à long terme.

De la même façon, les données recueillies dans le cadre de la télésurveillance sont souvent complétées par des données de type questionnaires de qualité de vie PREM, PROMs, de plus en plus considérées par les autorités dans le cadre de leur évaluation. Ainsi, l'appariement de l'ensemble de ces données permet de mettre en regard les données de qualité de vie avec des données épidémiologiques et cliniques et de fournir ainsi une vision complète de l'état de santé du patient.

2. L'évaluation des dispositifs médicaux

Selon le type de dispositif médical, l'évaluation de leur efficacité peut suivre le même process que celui d'un système de télésurveillance dans le cas par exemple de dispositif médical non implantable, connecté qui recueille des données en temps réel. Dans ce contexte, il est intéressant d'apparier les données recueillies à celles du SNDS pour en évaluer leur efficacité à long terme par exemple.

En revanche, dans le cas de dispositif médical implantable (prothèses, stents...), un appariement n'est pas forcément nécessaire pour en évaluer l'efficacité ou la sécurité. Ces éléments peuvent directement être appréhendés au travers des bases du SNDS.

3. L'évaluation de l'observance

Dans le domaine de l'observance thérapeutique, l'appariement entre sources de données différentes et complémentaires peut être très riche d'informations. Par exemple, il est possible de chainer des données patients provenant d'une étude de terrain recueillant des PROMs (et notamment des questionnaires autour de l'observance thérapeutique comme le questionnaire de Morisky), aux données du SNDS permettant de retrouver les délivrances de médicaments, voire à des données de dossiers médicaux informatisés permettant de retrouver les prescriptions initiales des médecins. Cela permet ainsi d'avoir une vue complète du parcours médicamenteux : prescription, dispensation, et questionnaire patients.

4. La description du fardeau de la maladie et de son impact sur la qualité de vie

Le fardeau de la maladie peut être décrit par divers indicateurs : espérance de vie, taux de mortalité, morbidité, auto-évaluation de la qualité de vie, etc. L'appariement de plusieurs sources de données est intéressant pour couvrir de façon plus complète les différentes composantes du fardeau d'une maladie. Ainsi, les données secondaires (registres, bases médico-économiques) peuvent être utilisées pour collecter des informations sur l'épidémiologie de la maladie (prévalence, incidence), le fardeau économique ou pour estimer l'APVP (Années Potentielles de Vie Perdues). La collecte prospective de PROMs permet de mesurer l'impact de la maladie sur la qualité de vie du point de vue du patient. Ces questionnaires d'auto-évaluation peuvent être génériques (EQ-5D, SF-12, SF-36) ou spécifiques à une maladie (DLQI, DQOL, SEP-59) ou à un symptôme (FSS, NRS, PSS).

5. Le suivi à long terme (OS, accès précoces)

Les données d'essais cliniques portent généralement sur des durées trop courtes, de par leur conception et/ou pour des raisons éthiques, pour apprécier l'effet des produits de santé sur la survie globale ou sans progression ou la survenue de tout type d'évènements sur le long terme. L'appariement de données d'essais cliniques aux données du SNDS permet de suivre sur le long terme les patients inclus dans l'essai pour compléter les données d'efficacité et de tolérance sur le long terme avec une exhaustivité des données, tout en allégeant le besoin de collecte des données par les centres investigateurs. Il en est de même, dans le cadre de l'accès précoce d'un produit de santé présumé innovant pour lequel il est demandé la mise en place d'une collecte de donnée mais dont la durée de recueil de données est limitée à la seule période d'accès précoce. L'appariement avec les données du SNDS permet de disposer de données exhaustives à long terme en termes d'efficacité et de tolérance des patients traités dans le cadre de ce dispositif et de couvrir d'autres dimensions d'intérêt (consommations de soins, séquences ultérieures de traitement…).

6. Validation d'algorithmes (BOAS)

L'identification de certaines pathologies peut s'avérer complexe à partir des seules données du SNDS, notamment lorsque celles-ci ne nécessitent pas systématiquement une hospitalisation ou la mise en place d'une ALD ou encore, si elles ne sont pas prises en charge par un traitement spécifique.
Dans ce contexte, le chaînage avec une base de données comportant le diagnostic clinique du patient permet d'affiner le développement d'algorithmes d'identification de la pathologie ou d'évaluer leur capacité à identifier les cas, soit la sensibilité. L'idéal est également d'avoir des patients non atteints de la pathologie dans la base de données cliniques afin d'évaluer aussi la spécificité de l'algorithme.

L'appariement de données cliniques et médico-administratives ouvre des perspectives intéressantes pour la recherche scientifique en permettant à la fois de développer des algorithmes robustes et d'apporter des résultats plus complets et précis quels que soient l'objectif de la recherche et l'aire thérapeutique ciblée.
Par ailleurs, dans le contexte actuel d'évolution permanente des données de santé, des aspects réglementaires qui permettent d'y accéder et de leur intérêt grandissant dans le domaine de la recherche scientifique, le développement d'entrepôts de données de santé (EDS) se généralise. En effet, dans le cas où l'intérêt scientifique n'englobe pas une unique étude mais pourrait permettre de réaliser plusieurs projets de recherche présentant des objectifs différents (algorithmie, économie...), il peut être pertinent pour le responsable de traitement de la base à apparier de réfléchir à la mise en place d'un EDS apparié au SNDS.