Big data en santé

Dossier réalisé en collaboration avec Rodolphe Thiebaut, directeur de l'équipe Statistiques pour la médecine translationnelle (unité 1219 Inserm/Inria), enseignant à l'Institut de santé publique d'épidémiologie et de développement (ISPED, Bordeaux), directeur de l'unité de soutien méthodologique à la recherche clinique et épidémiologique au CHU de Bordeaux et chercheur au Vaccine Research Institute (Créteil).

Dans le domaine de la santé, le *big data* (ou *données massives*) correspond à l'ensemble des données socio-démographiques et de santé, disponibles auprès de différentes sources qui les collectent pour diverses raisons. L'exploitation de ces données présente de nombreux intérêts : identification de facteurs de risque de maladie, aide au diagnostic, au choix et au suivi de l'efficacité des traitements, pharmacovigilance, épidémiologie… Elle n'en soulève pas moins de nombreux défis techniques et humains, et pose autant de questions éthiques.

En santé comme dans bien d'autres domaines, les progrès technologiques ont fait exploser la quantité d'informations recueillies à chaque instant. Ainsi, si dix ans ont été nécessaires pour obtenir la première séquence d'un génome humain, en 2003, il faut aujourd'hui moins d'une journée pour parvenir au même résultat. Cette accélération technologique fait croître le volume de données disponibles de manière exponentielle. Une aubaine pour la recherche en santé pour qui le big data est une source presque inépuisable de nouvelles connaissances, indispensables à l'innovation et aux progrès médicaux !

Un nombre important de sources et de types de données

La France possède environ 260 bases de données publiques dans le domaine de la santé, et le portail Epidémiologie-France recense jusqu'à 500 bases de données médico-économiques, cohortes, registres et études en cours.

© CC-BY 2.0 luckey_sun (via Flickr)

Les bases de données médico-administratives

Ces bases offrent des **données objectives et très exhaustives à l'échelle de larges populations**, avec peu de personnes perdues de vue en cours de suivi. Des atouts majeurs par rapport aux informations qui peuvent être recueillies lors d'études, poursuivies à court ou moyen terme, menées dans des populations spécifiques ou en nombre limité, et souvent fondées sur les déclarations des participants.

La plus riche des bases médico-administratives est le **SNIIRAM** (Système national d'information interrégimes de l'Assurance maladie). Dans cette base sont enregistrés tous les remboursements effectués par l'Assurance maladie pour chaque cotisant, tout au long de leur vie (biologie, médicaments, ambulances, consultations avec dates et noms des professionnels de santé vus, codes du type de maladie dans certains cas…). Ce système permet le suivi à long terme de données fiables.

Qui accède aux données du SNIIRAM ?
La base est actuellement accessible aux agences sanitaires et organismes publics de recherche à but non lucratif. En 2013, une cinquantaine de chercheurs l'a interrogée de manière régulière, réalisant plus de 17 000 requêtes, soit 30% de plus que l'année précédente.
Un arrêté du ministère de la Santé qui interdit l'accès à cette base aux organismes à but lucratif (compagnie d'assurances, laboratoire pharmaceutique…) a été jugé illégal par le Conseil d'État qui demande son annulation d'ici fin 2016. Par conséquent, toutes les structures voulant mener une étude d'intérêt général pourront bientôt accéder à ces données et les demandes devraient donc exploser dans les années à venir.

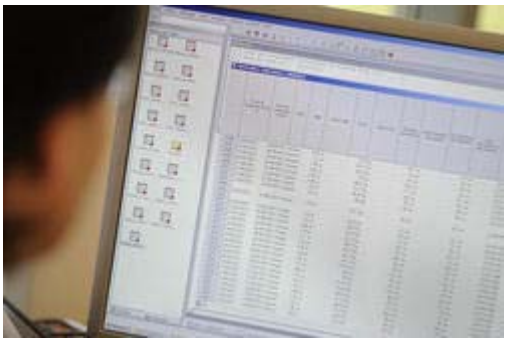
Il existe beaucoup d'autres bases médico-administratives, comme celle de l'**ATIH** (Agence technique de l'information sur l'hospitalisation) ou celles des caisses de retraite (dont la **CNAV**). Il existe également des bases gérées par des centres de recherche, notamment celle du **CépiDc** (Inserm) qui recense les causes médicales de décès en France depuis 1968.

Les cohortes

Une cohorte est un groupe de personnes partageant un certain nombre de caractéristiques communes, que des chercheurs suivent pendant un temps plus ou moins long afin d'identifier la survenue d'événements de santé (maladie ou dysfonctionnement de l'organisme) et des facteurs de risque ou de protection s'y rapportant.

Les organismes de recherche montent de grandes cohortes, incluant jusqu'à plusieurs dizaines de milliers de personnes, suivies pendant plusieurs années. C'est le cas par exemple des cohortes Constances, I-Share ou encore MAVIE et NutiNet-Santé, mises en œuvre en partenariat avec l'Inserm. La cohorte Constances, en cours de constitution, inclura à terme 200 000 adultes de 18 à 69 ans consultant dans des centres d'examens de santé de la Sécurité sociale. La cohorte I-Share incluera 30 000 étudiants des universités, suivis pendant 10 ans. L'observatoire MAVIE étudie les accidents de la vie quotidienne chez plus de 25

000 volontaires internautes. Quant à NutiNet-Santé, elle récolte une multitude de données sur le mode de vie, la santé et les habitudes alimentaires de 500 000 Français.



© Inserm/P. Delapierre

Vue d'écran et vérification de la collecte et de la cohérence des données prélevées sur les patients. Centre d'Investigation Clinique Plurithématique (CIC-P) Pierre Drouin, Institut Lorrain du Coeur et des Vaisseaux Louis Mathieu, CHU de Nancy

Toutes ces données récoltées permettent des études et une surveillance épidémiologiques, potentiellement à fort impact en santé publique.

Les études cliniques

Les laboratoires publics mènent par ailleurs de très nombreux travaux de recherche clinique, incluant des populations particulières de patients dont les profils de risque et les états de santé sont analysés. Or, le nombre de données collectées chez un même patient ne cesse de croitre, avec **des centaines d'informations recueillies chez un même individu, contre une dizaine il y a quelques années**.

En oncologie, des dizaines de paramètres cliniques, biologiques, d'imagerie et de génétique sont systématiquement recueillis. C'est aussi le cas pour le développement des vaccins. Ainsi, dans le cadre de l'essai clinique DALIA réalisé par *Vaccine Research Institute*, destiné à évaluer un vaccin thérapeutique contre le VIH, toutes les cellules immunitaires des patients ont été comptées grâce à la reconnaissance des marqueurs de surface, et leur fonctionnalité a été testée. Le protocole a généré environ 800 mesures par patient et par visite, sans compter l'étude de l'expression génétique de nombreux marqueurs (47 000 sondes/patient/visite) et du séquençage à haut débit du virus lui-même.

Les objets de santé connectés
Les objets de santé connectés génèrent également de très nombreuses données transmissibles et partageables : appareils mesurant le nombre de pas, la fréquence cardiaque, la glycémie, la pression artérielle… Ces données sont le plus souvent stockées et gérées par des géants d'internet ou *GAFAM* : Google, Apple, Facebook, Amazon et Microsoft.

Des défis techniques majeurs

Les énormes volumes de données désormais disponibles soulèvent des défis techniques concernant leur **stockage** et les **capacités d'exploitation**. Des programmes et des algorithmes informatiques et statistiques de plus en plus complexes s'avèrent nécessaires.

Les organismes de recherche disposent tous de serveurs de stockage et de supercalculateurs. Dans la plupart des cas, compte tenu de leur coût, ces plateformes sont mutualisées. C'est par exemple le cas du Mésocentre de calcul intensif aquitain (MCIA, Bordeaux), partagé par les universités de Bordeaux et les laboratoires CNRS, Inra, Inria et Inserm de la région. Autre exemple à Lyon, avec Platine, une plateforme européenne d'*immunomonitoring* gérée par plusieurs entreprises de biotechnologie ainsi que le Centre Léon Bérard de lutte contre le cancer et l'Inserm. Elle vise à aider les médecins à la décision thérapeutique en cancérologie et en infectiologie, en permettant l'analyse du statut immunologique initial des patients.

Autre problématique, les données massives sont assez fragmentées. **Les informations collectées sont en effet de plus en plus hétérogènes**, de par :

- leur nature (génomique, physiologique, biologique, clinique, sociale…),
- leur format (texte, valeurs numériques, signaux, images 2D et 3D, séquences génomiques…),
- leur dispersion au sein de plusieurs systèmes d'information (groupes hospitaliers, laboratoires de recherche, bases publiques…).

© CC-BY-SA 3.0 unported, Wikieditor243 (via Wikimedia Commons)

Data center

Pour rendre possible leur traitement et leur exploitation, ces informations complexes doivent être acquises de manière structurée, et codées avant de pouvoir être intégrées dans des bases ou des entrepôts de données. Des standards se développent, tel *I2b2* (pour *Informatics for Integrating Biology and the Bedside*), développé à Boston et désormais utilisé au CHU de Rennes, à Bordeaux ou encore à l'Hôpital européen Georges Pompidou (Paris). Ce système a par exemple été utilisé pour identifier et quantifier le risque accru d'infarctus du myocarde chez les patients sous Avandia, et a contribué au retrait du marché de ce médicament.

Grâce à ces standards, les hôpitaux et les centres de soins sont mieux armés pour compiler toutes les données collectées (pharmacie, biologie, imagerie, génomique, médico-économique, clinique...) dans des **entrepôts de données biomédicales**, interrogeables par les chercheurs via des interfaces web. De nombreuses équipes de recherche travaillent également sur des **plateformes intégrées**, pour apparier des bases et agréger leurs données avec celles de cohortes. Ainsi, le projet Hygie, conduit par l'Institut de recherche et de documentation en économie de la santé, apparie les bases SNIIRAM et SNGC (Système

national de gestion des carrières de l'Assurance retraite). L'objectif est de constituer un système d'information sur les indemnités journalières de sécurité sociale sur un échantillon de 800 000 personnes, qui servira à enrichir les fichiers de la cohorte CONSTANCES.

En pratique
Lorsqu'un chercheur souhaite démarrer une étude se fondant sur l'utilisation de données massives, il commence par identifier les bases qui lui sont utiles et demande un accès spécifique aux équipes ou organismes qui détiennent ces données. Il doit ensuite s'entourer de nombreuses compétences pour effectuer des méta-analyses intégrant toutes ces données. Pour l'essai DALIA par exemple, l'analyse des résultats a nécessité la contribution d'une cinquantaine de personnes issues de disciplines différentes : cliniciens, immunologistes, biologistes, virologistes, techniciens de laboratoire, assistants de recherche clinique, gestionnaires de bases de données, biostatisticiens ou encore bioinformaticiens.

Le big data, quelles utilités ?

Entreprises, organismes de recherche, à but lucratif ou non, scientifiques, médecins, industriels…. Le big data intéresse de très nombreux acteurs du monde de la santé car il permet de nombreux progrès médicaux.

Mieux prévenir et prendre en charge les maladies

Les données multidimensionnelles récoltées à long terme sur de larges populations, permettent d'**identifier des facteurs de risque** pour certaines maladies comme le cancer, le diabète, l'asthme ou encore les maladies neurodégénératives. Ces facteurs servent ensuite pour construire des messages de prévention, et mettre en place des programmes à destination des populations à risque.

Le big data permet en outre le développement de **systèmes d'aide au diagnostic** et d'outils permettant la **personnalisation des traitements**. Ces systèmes se fondent sur le traitement de grandes masses de données cliniques individuelles. Dans cette veine, le super-ordinateur Watson d'IBM permet par exemple d'analyser en quelques minutes le résultat du séquençage génomique de patients atteints de cancer, de comparer les données obtenues à celles déjà disponibles, et de proposer ainsi une stratégie thérapeutique personnalisée. En l'absence de cet outil, ce travail d'analyse prend plusieurs semaines. Les cliniques et hôpitaux intéressés passent un partenariat avec IBM qui détient ce super-ordinateur et fournit les résultats.

Le big data peut également permettre de **vérifier l'efficacité d'un traitement**. Par exemple, dans le domaine des vaccins, les cliniciens mesurent aujourd'hui des centaines de paramètres au cours des essais cliniques : comptages cellulaires, fonctionnalité cellulaire, expression de gènes d'intérêt... alors qu'il y a quelques années, on se limitait à la concentration des anticorps d'intérêt. À terme, cette évolution, les données massives qu'elle génère et la capacité à les analyser, pourrait permettre de vérifier qu'une vaccination a bien fonctionné au bout d'une heure seulement, à partir d'une micro goutte de sang.

Prédire des épidémies

Disposer de nombreuses informations sur l'état de santé des individus dans une région donnée permet de repérer l'élévation de l'incidence de maladies ou de comportements à risque, et d'alerter les autorités sanitaires.

Ainsi, le site HealthMap a pour objectif de prédire la survenue d'épidémies à partir de données provenant de nombreuses sources. Développé par des épidémiologistes et des informaticiens américains en 2006, ce site fonctionne en collectant les notes de départements sanitaires et d'organismes publics, les rapports officiels, des données internet… Le tout est mis à jour en continu pour **identifier des menaces sanitaires et alerter les populations**. Citons aussi le simulateur GLEAM, destiné à prédire la dissémination d'une épidémie en particulier, en exploitant les données de transport aérien.

En France, depuis 1984, le réseau Sentinelles suit plusieurs maladies infectieuses et alerte sur les épidémies grâce à la contribution de 1 300 médecins généralistes et d'une centaine de pédiatres répartis sur tout le territoire. Ces derniers rapportent au moins une fois par semaine le nombre de cas observés pour sept maladies transmissibles (diarrhée aiguë, maladie de Lyme, oreillons, syndromes grippaux, urétrite masculine, varicelle et zona) ainsi que les actes suicidaires. Les données sont transmises, via un réseau sécurisé, auprès de l'institut Pierre Louis d'Épidémiologie et de Santé Publique France, en collaboration avec l'Institut de veille sanitaire (InVS).

© Inserm/P. Latron

Améliorer la pharmacovigilance

L'analyse des données issues de cohortes ou des bases médico-économiques sur le long terme peut donc permettre d'observer beaucoup de phénomènes, et notamment de faire des rapprochements entre des traitements et la survenue d'événements en santé. Cette pratique permet de **repérer des événements indésirables graves** et d'alerter sur certains risques. En 2013, la base de données du SNIIRAM avait ainsi permis d'étudier le risque d'AVC et d'infarctus du myocarde chez les femmes utilisant une pilule contraceptive de 3ème génération.

Entre protection des données et avancée de la recherche :

Georgios Gropetis, ingénieur de recherche, responsable du centre de calcul de l'UMR-S 707 (réseau Sentinelles)

les défis éthiques du big data

Lors d'un essai clinique, un consentement est nécessaire avant le recueil de données de santé. De même, tout chercheur ou clinicien qui utilise des données du soin doit en informer le patient concerné et faire une déclaration auprès de la CNIL. Mais d'autres recueils se font à l'insu des contributeurs, notamment lors de recherches sur internet par mots clés ou lors de la transmission de données d'objets connectés. Cela pose évidemment des problèmes éthiques relatifs au souhait des citoyens de partager ou non ces données avec des tiers, ainsi que sur la préservation de l'anonymat.

Et de nombreuses autres questions se posent : faut-il conserver toutes les données ? Faut-il les mutualiser ? Qui doit les gérer et sous quelles conditions les partager ? Comment faire en sorte que Google, Apple, Facebook et Amazon ne s'approprient pas une partie d'entre elles ? Les enjeux sont de taille : risque de divulgation de la vie privée et conséquences pour la vie sociale, perte de confiance dans la puissance publique et la confidentialité de la recherche, harcèlement publicitaire... Ces problématiques font régulièrement l'objet d'avis de la part de comités d'éthiques, dont le Comité consultatif national d'éthique en France.

Les pouvoirs publics se sont également saisis de la question : la loi de modernisation de notre système de santé, promulguée le 26 janvier 2016, prévoit en effet l'**ouverture des données agrégées de santé à des fins de recherche, d'étude ou d'évaluation d'intérêt public**, à tout citoyen, professionnel de santé ou organisme (public ou privé) participant au fonctionnement du système de santé et aux soins. Cette ouverture est assortie de plusieurs conditions :

- les données ne doivent pas permettre l'identification des personnes concernées (la loi restreint drastiquement l'accès aux données à caractère personnel pouvant permettre l'identification d'une personne),
- les travaux ne doivent pas aboutir à la promotion de produits en direction des professionnels de santé ou d'établissements de santé, ni permettre l'exclusion de garanties des contrats d'assurance ou la modification de cotisations ou de primes d'assurance.

Pour y avoir accès, tout organisme de recherche ou d'étude désireux de mener un projet d'intérêt public doit soumettre ce dernier à l'Institut national des données de santé, composé entre autres de représentants de l'État, d'usagers de l'Assurance-maladie et de producteurs et d'utilisateurs publics et privés de données de santé. Le protocole de l'étude devra ensuite être validé par un comité scientifique, avant que la CNIL ne se prononce sur ses aspects relatifs au respect de la vie privée. Néanmoins, en juin 2016, les décrets d'application pour cette nouvelle organisation n'étaient toujours pas parus.

L'Inserm et le Système national des données de santé

La loi de modernisation du système de santé de janvier 2016 prévoit la création du Système national des données de santé (SNDS). Ce système sera notamment composé :

- des données de l'Assurance maladie (SNIIRAM),
- des données hospitalières (PMSI)
- des causes de décès (CépiDc-Inserm).

Il est prévu que la gouvernance de ce système inclue les producteurs de données, parmi lesquels l'Inserm. Plus concrètement, **l'Inserm devrait jouer le rôle d'opérateur d'extraction et de mise à disposition des données pour des traitements mis en œuvre à des fins de recherche**.

Marisol Touraine, ministre des Affaires sociales et de la Santé, a lancé en avril 2016 une consultation nationale en ligne sur le big data en santé. L'objectif est que chaque Français puisse donner son avis sur les objectifs souhaités pour les patients, les professionnels de santé, les industries, les assureurs ou la puissance publique, mais également sur les conditions dans lesquelles l'exploitation des données de santé est acceptable. Les conclusions sont attendues fin 2016. Les chercheurs plaident quant à eux pour une ouverture assez large des données, et un accès simplifié. Leur souhait est de parvenir à accélérer la recherche via des plateformes techniques adaptées, permettant de hauts niveaux de sécurité (ne collecter que des données ayant un intérêt potentiel pour le sujet de recherche, cloisonner les données identifiantes, chiffrer certaines informations, limiter les accès et la copie des informations...).

Pour aller plus loin

Data.gouv. fr - Plateforme ouverte des données publiques françaises
Institut national des données de santé