Les données de santé en France

Marie Zins, Marc Cuggia, Marcel Goldberg

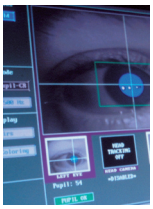
médecine/sciences 2021 ; 37 : 179-84

Les données de santé en France

Abondantes mais complexes

Marie Zins[1], Marc Cuggia[2], Marcel Goldberg[1]

> Alors que l'application de traçage des contacts (*contact tracing*) *StopCovid* (transformée à la mi-octobre 2020 en *TousAntiCovid*), débattue au Parlement[1] en raison des inquiétudes qu'elle suscitait concernant la confidentialité des données personnelles et les libertés individuelles du fait qu'elle permet d'alerter un utilisateur s'il s'est trouvé à proximité d'une personne atteinte de la COVID-19, a été adoptée par près de 12 millions de personnes[2], un dispositif concernant les données individuelles de santé, aux conséquences potentiellement beaucoup plus importantes pour les citoyens et leurs données personnelles, a commencé à se mettre en place suite à la Loi du 24 juillet 2019 (Loi n° 2019-774) relative à l'organisation et à la transformation du système de santé[3] : la *plateforme des données de santé*, communément appelée *Health Data Hub*, constituée sous la forme d'un groupement d'intérêt public (GIP). Il ne s'agit plus de simplement signaler qu'on a croisé une personne anonyme infectée par le SARS-Cov-2, mais de réunir, dans une infrastructure informatique unique, un immense ensemble de données personnelles particulièrement sensibles concernant la totalité de la population française. Ce projet suscite désormais un certain intérêt médiatique et un début d'inquiétude. Mais cette inquiétude ne concerne presque uniquement que le fait que ces données sont déposées et gérées dans un *cloud* appartenant à une société américaine, un nuage informatique qui tombe sous le coup de la loi américaine de 2018 dite « CLOUD act[4] », qui ouvre la possibilité d'un transfert des données personnelles vers les États-Unis, comme s'en est inquiété récemment le Conseil d'État.[5] Cet aspect est certes très important, mais il masque également de très nombreux enjeux liés au partage des données de santé, et qui sont largement méconnus de la population. Nous nous proposons de rappeler, tout d'abord, ce que sont les données de santé, ce qu'elles apportent et la nécessité d'en faciliter le partage, mais aussi les difficultés rencontrées pour leur accès et leur utilisation. Nous expliquerons ensuite, dans un deuxième article, en quoi cette *plateforme des données de santé*, telle qu'elle est conçue et pilotée par les pouvoirs publics pour répondre à ces difficultés et pour promouvoir l'intelligence artificielle en santé, est un projet qui soulève de fortes inquiétudes pour les citoyens et la société dans son ensemble. Même si les problèmes posés se présentent sous une forme différente selon les pays, notre propos concernera spécifiquement la situation en France. <



Vignette (Photo © Inserm-Michel Depardieu).

[1]UFR de médecine, Université de Paris, 16 avenue Paul-Vaillant Couturier, F-94800 Villejuif, France.
[2]UMR Inserm, Laboratoire traitement du signal et de l'image (LTSI) - Équipe Données massives en santé (*Health Big Data*), Modélisation des connaissances biomédicales, Université Rennes1, Faculté de médecine, Avenue du Professeur Léon-Bernard, 35043 Rennes Cedex, France.
marcel.goldberg@inserm.fr

Les données de santé : une explosion quantitative et qualitative

Pendant longtemps, les données de santé étaient constituées d'un nombre restreint de données collectées pour des besoins de gestion par diverses administrations : remboursement des soins pour l'assurance maladie et facturation hospitalière notamment. La recherche clinique et épidémiologique était également une source de production de données dans le cadre de projets de recherche ayant des objectifs spécifiques. Or, le paysage des données de santé

[1] Son utilisation a été votée le 27 mai 2020 par le Parlement.
[2] À la mi-décembre 2020.
[3] On se réfèrera particulièrement au Titre III, Chapitre Ier, Article 41, VIII (https://www.legifrance.gouv.fr/download/file/v1P1M3GXaBwvKuWy0CMq4zg8dfuYLob-Mvhwak3XtkyQ=/JOE_TEXTE).
[4] Le *Clarifying Lawful Overseas Use of Data Act* (CLOUD Act) est une loi des États-Unis sur l'accès aux données de communication (données personnelles) présentes dans le Cloud. Elle permet aux instances de justice fédérales, ou même locales, d'obliger les fournisseurs de services établis sur le territoire des États-Unis, par mandat ou assignation, à fournir les données relatives aux communications électroniques stockées sur des serveurs situés aux États-Unis ou dans des pays étrangers. Elle permet notamment de solliciter auprès des fournisseurs de services, opérant aux États-Unis, les communications personnelles d'un individu sans que celui-ci en soit informé, ni que son pays de résidence ne le soit, ni que le pays où sont stockées ces données ne le soit.
[5] Ordonnance du juge des référés du Conseil d'État du 13 octobre 2020, prise à la suite du mémoire rendu par la Commission nationale informatique et liberté (CNIL) qui avait été saisie pour avis par le Conseil suite à un recours de diverses associations et professionnels.

a profondément évolué. Depuis environ deux décennies, la santé connaît en effet une transformation numérique sans précédent [1], avec l'informatisation du dossier patient et, plus récemment, la production en routine de données de « omiques (OMICS) », d'imagerie ou issues de la santé connectée. L'énorme développement des capacités de calcul et de stockage des outils informatiques a permis un accroissement vertigineux de la quantité des données de différentes natures, disponibles et provenant de sources nouvelles, dans un contexte de soins comme de recherche.

Il faut souligner que les « données de santé » doivent être considérées dans une acception très large. Elles ne concernent pas uniquement des données médicales ou biologiques, mais incluent habituellement des données sur des éléments associés à la santé, comme les facteurs de risque (tabac, alcool, etc.) ou l'utilisation du système de soins par un individu donné.

À côté des données recueillies par les méthodes « classiques » (questionnaires, examens médicaux, dossiers hospitaliers, registres administratifs, etc.), les données mobilisées aujourd'hui proviennent en grande partie de sources qui n'étaient pas disponibles auparavant. Parmi celles-ci, les données OMICS constituent un enjeu majeur. Le séquençage de l'ADN devient un examen de routine, facilité par la techniques du NGS (*next generation sequencing*) qui en ont grandement réduit les coûts [2]. On estime que d'ici à 2025, le nombre de génomes humains séquencés sera de l'ordre de 100 millions, et pourrait même atteindre 2 milliards [3]. Les données OMICS, qui comprennent aussi les données de transcriptomique, de métabolomique et d'épigénétique, trouvent actuellement de nombreuses applications en médecine personnalisée. Le secteur de l'imagerie n'est pas en reste avec l'émergence de la radiomique[6] [4], en particulier en cancérologie. Dans cette approche quantitative, les nouvelles techniques d'acquisitions multimodales combinées à des algorithmes d'intelligence artificielle (IA) permettent de découvrir de nouveaux biomarqueurs diagnostiques, pronostiques ou de réponses thérapeutiques à partir du traitement de grands volumes de données.

(➜) Voir la série *Numérique et santé*, *m/s* n° 5 (mai 2018) à *m/s* n° 10 (octobre 2019)

L'accès *via internet* aux données produites par les objets en santé[7] constitue également une source de données en fort développement [5] (➜). Qu'il s'agisse de données issues d'applications de santé mobile (applications destinées aux patients, comme pour le contrôle de la tension artérielle ou de la glycémie), de dispositifs médicaux implantables communicants (défibrillateurs, prothèses connectées) ou d'appareils de bien-être (montres et balances connectées), ces données font l'objet d'investissements très importants par l'industrie de haute technologie et l'industrie pharmaceutique [6, 7] et ouvrent des perspectives de recherche en épidémiologie digitale [8].

Des données massives *a priori* hors champ santé, comme les images satellitaires, trouvent par exemple des applications épidémiologiques sur des questions de santé-environnement [9-11]. Des données individuelles de consommation d'énergie recueillies par des compteurs « intelligents » sont utilisées dans des projets de recherche sur les maladies chroniques ou le vieillissement, car ces données collectées au fil de l'eau contiennent des marqueurs très spécifiques des appareils électriques utilisés à domicile [12, 13]. Enfin, les données du *web* et des réseaux sociaux sont de plus en plus utilisées pour, par exemple, établir des modèles de surveillance épidémiologique [14, 15].

Les données médico-administratives issues de l'activité hospitalière ou liées au remboursement des soins existent depuis longtemps, mais leur usage pour la recherche et pour des études diverses, notamment en pharmaco-épidémiologie, s'est considérablement développé avec la mise en place d'entrepôts de données hospitalières et du système national des données de santé (SNDS). Créé par la loi du 26 janvier 2016 « Modernisation de notre système de santé »[8], le SNDS inclut notamment le système national d'information inter-régimes de l'assurance maladie (SNIIRAM).

Le cas du SNIIRAM-SNDS

Le SNIIRAM, socle majeur du SNDS, qui couvre la totalité des quelque 67 millions de personnes vivant en France, constitue certainement la plus importante base de données de santé au monde ; elle mérite une attention particulière du fait de son ampleur en termes de données et de couverture de la population. Le SNIIRAM réunit les bases de données de remboursement de soins, les données de séjours hospitaliers du programme de médicalisation des systèmes d'information (PMSI), les actes médicaux pratiqués en ville et à l'hôpital, ainsi que les causes de décès. Ces données, collectées de façon exhaustive pour toute la population avec une très grande précision, portent sur toutes les données de remboursements de soins, du plus banal, comme une consultation d'un généraliste ou une séance de détartrage dentaire, au plus intime, comme un diagnostic de cancer ou de Sida (syndrome d'immunodéficience acquise), une hospitalisation dans un établissement psychiatrique ou une interruption volontaire de grossesse (IVG).

Il existe cependant d'importantes limites aux données contenues dans le SNIIRAM. Ainsi, elles ne comportent de diagnostics qu'en cas d'hospitalisation ou de prise en charge au titre de l'ALD (affection de longue durée, qui permet une prise en charge intégrale des soins), mais ces diagnostics ne sont habituellement pas

[6] L'objectif de cette discipline est de mieux caractériser les tumeurs en exploitant de manière approfondie des données directement contenues dans les approches classiques d'imagerie.

[7] Des objets connectés ou non, utilisés en santé.

[8] On se réfèrera particulièrement à l'Article 193 du Chapitre V « Créer les conditions d'un accès ouvert aux données de santé » de la Loi n° 2016-41. (https://www.legifrance.gouv.fr/download/pdf?id=f1zqqKkO-FAUZH67_XjED1sDF ihSqtW46KWa2lSZzs=).

formellement validés ; les données précisent les actes effectués (on sait qu'à telle date, telle personne a eu une consultation de cardiologie, un dosage de glycémie, un examen d'anatomo-pathologie ou un scanner abdominal), mais on n'en a pas les résultats ; on ne trouve dans le SNIIRAM aucune donnée sur des facteurs de risque, comme par exemple le tabagisme, la consommation d'alcool ou le poids, ou sur la situation sociale des personnes (hormis le fait d'être bénéficiaire de la protection universelle maladie [PUMa], qui concerne des personnes ayant de faibles revenus).

En raison de son exhaustivité quant à la couverture de la population, le SNIIRAM est potentiellement à risque pour ce qui est de la confidentialité. En effet, malgré la pseudonymisation de ces données (c'est-à-dire la substitution d'un identifiant direct, comme le nom ou le numéro de sécurité sociale, par un identifiant non signifiant, comme un nombre généré aléatoirement), il peut être très simple d'identifier les personnes concernées par croisement de quelques variables faciles à connaître. Par exemple, il est quasiment certain qu'il n'y ait, au sein du SNIIRAM, qu'une seule personne qui soit une femme ayant un certain âge, résidant dans telle commune, hospitalisée tel jour dans telle clinique, ayant consulté tel cardiologue tel jour et ayant eu telle délivrance de médicaments tel autre jour dans telle pharmacie. Cette possibilité de « re-identification », inévitable dans toute base de données comportant des données individuelles, justifie le très contraignant encadrement législatif et réglementaire de l'accès aux données du SNIIRAM, que nous détaillons plus loin.

Le partage des données de santé, une nécessité

Jusqu'à une période récente, les données collectées pour des besoins de gestion ou pour réaliser des essais cliniques, des enquêtes épidémiologiques, etc., n'intéressaient que les organismes médico-administratifs (sécurité sociale, administration hospitalière, etc.) ou les équipes de cliniciens et de chercheurs qui les utilisaient pour leur propre usage, et ne suscitaient que très peu de demandes en dehors de ces sphères. Les données étaient de fait « fermées », et seul le responsable de la collecte pouvait les utiliser, même s'il arrivait qu'il autorise des personnes extérieures à l'équipe ou à l'administration concernée d'en bénéficier. Pour ce qui concerne la recherche (clinique ou épidémiologique), ce fonctionnement fermé correspondait, à de rares exceptions près, à des études portant sur des effectifs limités (quelques centaines, au plus quelques dizaines de milliers de sujets) et un petit nombre de données pour chaque sujet.

Ce mode de fonctionnement s'est révélé inadapté au nouveau contexte de production du nombre désormais gigantesque de données. La recherche en santé est en première ligne, à la fois comme utilisatrice mais aussi comme productrice de ces données. Ainsi, depuis le début des années 2000, on a vu, dans le domaine de l'épidémiologie, se mettre en place, dans différents pays, des études de taille beaucoup plus importante, avec notamment des cohortes composées de plusieurs centaines de milliers, voire de plus d'un million de participants, suivis pendant de très longues durées, et pour lesquels un nombre considérable de données est recueilli de façon prospective auprès de sources de plus en plus diversifiées, y compris des échantillons biologiques. Ainsi, la *UK Biobank* du Royaume-Uni est une cohorte de plus de 500 000 participants qui sont individuellement suivis et font l'objet de nombreux recueils de données [16]. En France, la cohorte *Constances* a recruté et suit plus de 200 000 volontaires [17]. Plus récemment, aux États-Unis, les *National Institutes of Health* (NIH) ont lancé l'initiative *All of Us* [18], une cohorte de plus d'un million de patients, dans laquelle sont collectées, de façon prospective, les données du dossier médical électronique, y compris l'imagerie médicale, ainsi que des données socio-comportementales et environnementales.

Cette énorme augmentation de la taille des cohortes correspond à des besoins scientifiques. En effet, la recherche sur les causes de nature environnementale, sociale, ou génétique des maladies est de plus en plus ciblée sur des risques relativement faibles, et des cohortes de très grande taille sont nécessaires pour assurer une puissance statistique suffisante pour comprendre le rôle de divers facteurs personnels et environnementaux et leur interaction avec des traits génétiques complexes. Cependant, même les plus grandes cohortes ne génèrent pas suffisamment de cas pour l'étude de maladies peu fréquentes ou pour l'analyse du rôle de l'exposition à des facteurs de risque à des niveaux peu élevés ; pour certaines situations complexes, il est donc souvent nécessaire de regrouper les données provenant de plusieurs cohortes au sein de consortiums internationaux. Le partage de données à grande échelle est tout aussi importante pour les recherches en santé publique et sur les services de santé, que pour l'analyse comparative internationale des déterminants sociaux de la santé. La mise en commun de données au sein de consortiums internationaux, comme ceux qui ont été à l'origine de la plupart des progrès récents de la génomique des populations, est une nécessité à la fois en raison de la complexité des phénomènes étudiés, mais aussi parce que ces recherches produisent des quantités énormes de données, qui nécessitent des moyens de stockage et d'analyse considérables et très coûteux.

Le besoin de croisement des données OMICS avec les données cliniques et environnementales était déjà décrit comme un enjeu majeur dans la stratégie nationale des NIH concernant la médecine 4P (personnalisée, préventive, prédictive, participative) en 2011 [19]. Dix ans plus tard, cette stratégie d'intégration et de partage de données multi-domaines et multi-échelles est à l'œuvre dans différents pays, en particulier en Amérique du Nord [20, 21], en République Populaire de Chine [22, 23], et dans plusieurs pays européens; dont la France [24].

À ces raisons scientifiques, qui s'inscrivent dans un mouvement plus général des sciences et des données

ouvertes (*open science* et *open data*), s'ajoutent des raisons économiques. Les gigantesques dispositifs épidémiologiques qui se sont mis en place ont un coût élevé, qu'il est devenu indispensable de mutualiser. La plupart des organismes de financement de la recherche souhaitent, voire exigent, que les données collectées grâce aux budgets qu'ils allouent puissent être utilisées le plus largement possible, et imposent en contrepartie qu'elles soient ouvertes à d'autres équipes que celles qui ont organisé leur recueil. Cela est d'autant plus justifié que ces très grandes études sont aujourd'hui de nature « généraliste », conçues pour couvrir un large éventail de problèmes de santé et de déterminants de nature diverse : médicale, sociale, environnementale, comportementale, professionnelle, etc. Aucune équipe, si importante et compétente soit-elle, n'a la capacité de pleinement exploiter scientifiquement tout le potentiel qu'offrent les vastes ensembles de données réunies qui, de fait, constituent des plateformes de recherche ouvertes permettant de développer de multiples projets à un coût marginal.

De nouveaux acteurs sont également apparus dans le monde des données de santé. Les industriels du médicament ont en effet, de leur côté, besoin de grands ensembles de données sur les populations pour évaluer les parts de marché de leurs produits, pour suivre leur utilisation et s'assurer de leur sécurité « en vie réelle » à long terme, au-delà des essais cliniques de durée limitée portant sur des populations sélectionnées qui ont permis d'obtenir l'autorisation de mise sur le marché (AMM). Il en est de même pour le développement des dispositifs médicaux. Les industriels du domaine ou les autorités régulatrices voient l'utilisation de données de vie réelle comme un levier pour apprécier le service médical rendu ou surveiller la survenue d'événements indésirables après la mise sur le marché des dispositifs. D'autres industriels ou acteurs économiques œuvrant dans divers domaines (tels que le secteur assurantiel) ont pris conscience de l'apport potentiel de ces données de santé pour le développement de leurs activités [25]. Les *data scientists*[9] y ont vu un matériau prometteur pour déployer leurs méthodes d'IA. Ils ont été suivis des GAFAM[10] et de *start-ups* souhaitant proposer leurs applications médicales.

L'encadrement législatif et réglementaire de l'accès et du partage des données de santé

En raison de leurs caractéristiques, un cadre législatif et réglementaire spécifique existe pour la collecte, la gestion et le partage des données de santé.

Rappelons tout d'abord que, selon le terme de la loi Informatique et libertés, la plupart des données de santé sont des données « à caractère personnel » (c'est-à-dire des données qui concernent des personnes directement ou indirectement identifiables). Il s'agit très souvent de données sensibles, dont la divulgation peut entraîner des effets néfastes pour les personnes concernées. Ce problème se pose en particulier pour le SNIIRAM, qui réunit de façon exhaustive, pour toute la population, des données individuelles de remboursement de soins, de séjours hospitaliers et les causes de décès. Or, comme nous l'avons indiqué, l'accès au SNIIRAM permet d'identifier aisément les personnes concernées par croisement de quelques variables faciles à connaître. La plupart des gens ignorent cette incroyable richesse des données du SNIIRAM, qui s'accroît régulièrement par l'inclusion des nouvelles sources de données (résultats d'examens biologiques, dossier médical partagé, etc.). Le recueil et l'usage des données de santé sont donc logiquement très fortement encadrés par divers textes législatifs et réglementaires nationaux (chapitre spécifique de la loi Informatique et libertés, loi relative aux recherches impliquant la personne humaine, dite loi Jardé, code pénal, multiples décrets, arrêtés et circulaires) et européens (règlement général sur la protection des données [RGPD]). En simplifiant, on peut dire que les grands principes qui organisent le recueil et l'usage des données de santé reposent essentiellement sur l'information des personnes, leur consentement éclairé au recueil et à l'utilisation de leurs données, leur droit de retrait et d'opposition. Plusieurs organismes sont chargés d'appliquer les textes et de veiller à leur respect : commission nationale de l'informatique et des libertés (CNIL), comité éthique et scientifique pour les recherches, les études et les évaluations dans le domaine de la santé (CESREES), comités de protection des personnes (CPP), délégués à la protection des données (DPO, *data protection officer*) des entreprises et des organismes.

Les difficultés pour l'accès et l'utilisation des données de santé

Les données de santé constituent une ressource qui peut être particulièrement utile pour la recherche et les études en épidémiologie sur le parcours de soins, en économie de la santé, pour la recherche sur les services de santé, pour des études évaluatives, etc. Il existe donc de forts enjeux pour faciliter l'usage de ces données.

Concernant le SNIIRAM

On rencontre deux situations qui correspondent à des modalités différentes d'accès aux données.

La première est celle où le chercheur travaillera uniquement sur les données contenues dans le SNIIRAM. Un exemple bien connu est l'étude des effets du *Mediator* [26] : les données de remboursement ont ainsi permis de sélectionner, au sein du SNIIRAM, tous les sujets diabétiques ayant eu des prescriptions de ce médicament, ainsi qu'un groupe de témoins. Les cas de valvulopathies cardiaques ont ensuite été recherchés

[9] Analystes de données en masse, ou *big data*.
[10] Les sociétés Google, Apple, Facebook, Amazon et Microsoft.

dans les données d'hospitalisation du PMSI. L'analyse a alors consisté à étudier l'association entre prescription du médicament et survenue d'une valvulopathie. Dans cet exemple, du fait de la faible fréquence des événements de santé étudiés, il a fallu travailler sur l'ensemble des 67 millions de personnes dont les données figurent dans la base de données. Dans d'autres situations, on peut se contenter d'utiliser un échantillon du SNIIRAM, constitué par tirage au sort d'environ une personne sur 100, appelé échantillon général des bénéficiaires (EGB), qui permet de travailler sur environ 660 000 personnes, et qui est d'accès plus facile que le SNIIRAM intégral.

La seconde situation est celle où on souhaite enrichir une base de données individuelles existante (une cohorte, un registre de maladie, un essai clinique, etc.) avec des données issues du SNIIRAM. Cette approche est d'un très grand intérêt, car elle permet de collecter des données précises et fiables sur des consommations de soins, des consultations de professionnels de santé ou des hospitalisations, avec l'avantage supplémentaire d'éviter certains biais et les individus perdus de vue dans les études longitudinales. Elle permet aussi de pallier certaines limites évoquées du SNIIRAM, en couplant données recueillies directement auprès des personnes et données médico-administratives. Cependant, ce rapprochement nécessite de disposer d'un identifiant commun entre le jeu de données à apparier et le SNIIRAM. Or, l'identifiant utilisé dans le SNIIRAM est un « pseudonyme », c'est-à-dire un identifiant crypté non signifiant calculé à partir du NIR (numéro d'inscription au répertoire national d'identification des personnes physiques, plus communément appelé numéro de sécurité sociale). Pour apparier et extraire les données du SNIIRAM, il faut donc disposer du NIR des sujets concernés afin de pouvoir calculer le pseudonyme correspondant. Il est aussi possible de réaliser un appariement probabiliste, en croisant des données communes ente la base de données concernée et le SNIIRAM (par exemple, l'âge, le sexe, la localisation, la date de consultation d'un spécialiste, etc.). Mais cette méthode est lourde, rarement possible car nécessitant de disposer de suffisamment de données communes précises et discriminantes, et elle n'est pas toujours efficace.

C'est donc essentiellement par l'utilisation du NIR qu'il est possible d'apparier un jeu de données au SNIIRAM. Or, jusqu'à la loi du 26 janvier 2016 de modernisation de notre système de santé et jusqu'au décret du 26 décembre 2016, qui en précise les modalités d'application, il n'était pratiquement pas possible d'utiliser le NIR : un décret du Conseil d'État, quasiment impossible à obtenir, était nécessaire... Et l'accès à ces données n'était pas autorisé aux entreprises privées à but lucratif. Ces dispositions limitaient donc très fortement l'usage des données du SNIIRAM. Les textes adoptés en 2016 ont donc eu pour buts de faciliter l'accès aux données de santé du SNIIRAM et de lever l'interdiction de les utiliser pour les entreprises privées.

D'autres difficultés proviennent des ressources nécessaires pour réaliser des projets utilisant les données de santé : lorsqu'il s'agit de données issues du SNDS, les contraintes de sécurité applicables impliquent des moyens informatiques et organisationnels que la plupart des équipes de recherche n'ont pas les moyens de réunir ; l'analyse de très vastes ensembles de données par des méthodes (notamment d'intelligence artificielle) très gourmandes en espace et en puissance de calcul requiert également des ressources informatiques adaptées et encore trop peu présentes dans les systèmes d'information de nos institutions.

Une autre difficulté, sans doute la plus critique, concerne les compétences nécessaires pour exploiter et interpréter correctement des données qui sont particulièrement complexes : important volume (des milliards de lignes de prestation disponibles), et architecture mal adaptée aux études portant sur des individus (les données sont réparties dans des dizaines de tables différentes en fonction du type de prestation, avec des clés de jointure variables). Cela s'explique par le fait qu'il s'agit de données de gestion, qui de plus nécessitent une très bonne connaissance du contexte juridique et technique du remboursement pour être exploitées correctement, car le codage des données est complexe et variable dans le temps. À titre d'exemples, les actes de détartrage dentaire avaient, jusqu'en 2014, le même code que pour une obturation[11] et trois codes différents étaient utilisés pour un frottis du col utérin, selon le lieu où il était réalisé. Ainsi, comment les diagnostics issus du PMSI peuvent-ils être utilisés correctement sans en connaître les règles de codage ? On imagine bien les erreurs qui peuvent être faites si on ignore toutes ces règles et leur évolution dans le temps.

Une bonne utilisation des données du SNIIRAM nécessite donc des compétences pluridisciplinaires qui incluent, en premier lieu, une compréhension du domaine sur lequel porte le traitement, mais également une connaissance fine des données utilisées, de leurs qualités et des processus qui les ont produites. Enfin, l'utilisation des données de santé requiert une connaissance du cadre éthique, déontologique et réglementaire en vigueur, ce dernier s'étant particulièrement durci depuis le RGPD [27]. Ainsi, face au risque de mésusage ou d'interprétation biaisée, l'adage « jamais seul face aux données » devrait prévaloir. Or ces compétences multidisciplinaires sont encore trop rares. Elles devraient être regroupées au plus près de l'expertise métier (c'est-à-dire des professionnels de santé ou de santé publique), au plus près des lieux de production des données (les établissements de santé et les unités de recherche en santé) et, en définitive, au plus près des patients, pour garder leur confiance.

[11] L'obturation d'une dent cariée consiste à enlever d'abord la partie cariée puis, après avoir nettoyé la cavité ainsi créée, remplir cette dernière d'un matériau dit d'obturation.

En dehors du SNIIRAM-SNDS

Il existe bien d'autres données de santé produites et gérées par de nombreux acteurs : services hospitaliers, organismes de recherche et universités, registres de maladies, enquêtes épidémiologiques et cohortes, mutuelles, etc., dont l'accès et l'utilisation sont difficiles, notamment en raison de la dispersion en de multiples systèmes d'information mis en place sans aucune coordination. Il n'est pas aisé pour un investigateur qui souhaite utiliser des données pertinentes pour son projet, de savoir où il peut les trouver, ce qu'elles couvrent réellement, quelles sont leurs caractéristiques. Chaque source pertinente a, de plus, ses propres règles d'accès. Il est donc nécessaire d'établir des accords avec chacune. Il faut aussi obtenir les autorisations réglementaires, qui relèvent de textes et d'organismes divers, selon le type d'investigation et la nature des données (notamment s'il s'agit ou pas de « recherches impliquant la personne humaine » telles que définies par la loi Jardé[12], ou de données génétiques).

Conclusion

L'accès aux données de santé et leur utilisation sont donc des opérations difficiles, qui nécessitent des moyens techniques et méthodologiques importants. La Plateforme des données de santé (PDS), plus communément appelée *Health Data Hub* (HDH), est une infrastructure officiellement créée par un arrêté ministériel du 30 novembre 2019. Elle est destinée à faciliter l'accès et l'utilisation des données de santé afin de favoriser la recherche. Le HDH a l'ambition de supprimer les différents obstacles que nous avons évoqués, en facilitant l'accès aux sources de données et en offrant un guichet unique pour les diverses démarches nécessaires, en apportant un soutien méthodologique pour l'utilisation des données et en mettant à disposition des ressources informatiques puissantes de haut niveau.

Cette ambition est tout à fait louable, mais certains choix techniques ou concernant la gouvernance du HDH posent des problèmes majeurs que nous détaillerons dans un prochain article. ◊

Health data in France: Abundant but complex

LIENS D'INTÉRÊT

Les auteurs déclarent n'avoir aucun lien d'intérêt concernant les données publiées dans cet article.

RÉFÉRENCES

1. DGOS. *Atlas des systèmes d'information hospitaliers*. Ministère des Solidarités et de la Santé, Dec. 18, 2020. https://solidarites-sante.gouv.fr/systeme-de-sante-et-medico-social/e-sante/sih/article/atlas-des-systemes-d-information-hospitaliers
2. Plan France Médecine Génomique 2025/Aviesan. https://www.aviesan.fr/aviesan/accueil/toute-l-actualite/plan-france-medecine-genomique-2025
3. Stephens ZD, Lee SY, Faghri F *et al.* Big Data: Astronomical or Genomical? *PLoS Biol* 2015 ; 13 : e1002195
4. Vande Perre S, Duron L, Milon A, *et al.* Radiomique : mode d'emploi. Méthodologie et exemples d'application en imagerie de la femme. *Imag Femme* 2019 ; 29 : 25-33.
5. Banerjee A, Chakraborty C, Kumar A, Biswas D. Handbook of data science approaches for biomedical engineering. *Emerging trends in IoT and big data analytics for biomedical and health care technologies*. New York : Elsevier 2019 : 35.
6. Les GAFAM continuent leur percée dans la santé. https://www.ticpharma.com/story/1079/les-gafam-continuent-leur-percee-dans-la-sante.html
7. Singh M, Sachan S, Singh A, Singh KK. Emergence of Pharmaceutical industry growth with industrial IoT approach. *Internet of things in pharma industry: possibilities and challenges*. New York : Elsevier 2020: 195-216.
8. Laboratoire d'épidémiologie digitale. https://www.campusbiotech.ch/fr/node/353
9. Nguyen QC, Huang Y, Kumar A, *et al.* Using 164 million google street view images to derive built environment predictors of COVID-19 cases. *Int J Environ Res Public Health* 2020 ; 17 : 6359.
10. Sharifi A. Yield prediction with machine learning algorithms and satellite images. *J Sci Food Agric* 2020 ; doi: 10.1002/jsfa.10696.
11. Bruzelius E, Le M, Kenny A, *et al.* Satellite images and machine learning can identify remote communities to facilitate access to health services. *J Am Med Inform Assoc* 2019 ; 26 : 806-12.
12. Patrono L, Primiceri P, Rametta I, *et al.* An innovative approach for monitoring elderly behavior by detecting home appliance's usage. 25th International Conference on Software, Telecommunications and Computer Networks (SoftCOM) 2017 : 1-7.
13. Fell JM, Kennard H, Huebner G, *et al.* Energising health: a review of the health and care applications of smart meter data. Technical report, May 2017. doi : 10.13140/RG.2.2.23987.63521.
14. Kogan NE, Clemente L, Liautaud P, *et al.* An early warning approach to monitor covid-19 activity with multiple digital traces in near real-time. *ArXiv* July 2020.
15. Poirier C, Lavenu A, Bertaud V, *et al.* Real time influenza monitoring using hospital big data in combination with machine learning methods: comparison study. *JMIR Public Health Surveill* 2018 ; 4 : e11361.
16. Sudlow C, Gallacher J, Allen N, *et al.* UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 2015 ; 12 : e1001779.
17. Zins M, Goldberg M, and the CONSTANCES team. The French CONSTANCES population-based cohort: design, inclusion and follow-up. *Eur J Epidemiol* 2015 ; 30 : 1317-28.
18. The All of Us Research Program Investigators. The All of Us research program. *N Engl J Med* 2019 ; 381 ; 7.
19. National Research Council (US) Committee on a framework for developing a new taxonomy of disease, toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease. Washington (DC) : National Academies Press (US), 2011.
20. PCORnet. The national patient-centered clinical research network. https://pcornet.org/
21. Bubela T, Genuis SK, Janjua NZ, *et al.* Medical information commons to support learning healthcare systems: examples from Canada. *J Law Med Ethics* 2019 ; 47 : 97-105.
22. Zhang L, Wang H, Li Q, *et al.* Big data and medical research in China. *BMJ* 2018 ; 360.
23. Wu F, Lu C, Zhu M, *et al.* Towards a new generation of artificial intelligence in China. *Nat Mach Intell* 2020 ; 2 : 312-6.
24. Cuggia M, Combes S. The French health data hub and the German medical informatics initiatives: two national projects to promote data sharing in healthcare. *Yearb Med Inform* 2019 ; 28 : 195-202.
25. Mamiko YA. The impact of big data and artificial intelligence (AI) in the insurance sector. 2020, p 36. http://www.oecd.org/finance/Impact-Big-Data-AI-in-the-Insurance-Sector.htm
26. Weill A, Païta M, Tuppin P, *et al.* Benfluorex and valvular heart disease: a cohort study of a million people with diabetes mellitus. *Pharmacoepidemiol Drug Saf* 2010 ; 19 : 1256-62.
27. CNIL. https://www.cnil.fr/fr/rgpd-de-quoi-parle-t-on.

[12] Loi du 5 mars 2012 relative aux recherches impliquant la personne humaine. (https://www.legifrance.gouv.fr/loda/id/JORFTEXT000025441587/2021-01-02/).

TIRÉS À PART

M. Goldberg