Les données de santé, un patrimoine commun qui doit servir à améliorer le bien-être de tous

Exhaustives et précises sur nos conditions de vie, nos soins..., de nombreuses informations sont collectées tous les jours mais restent peu ou pas disponibles pour la recherche, regrette **un collectif de professeurs de médecine**

Dans un contexte de pandémie mondiale, à l'heure où le prix Nobel de chimie d'Emmanuelle Charpentier *[qui mène ses recherches à l'étranger]* questionne l'avenir de la recherche en France, il est urgent de rappeler que nous disposons d'un immense potentiel sous-exploité : des millions de données administratives, sociales, médicales et environnementales.

En tant que médecins, chercheurs, épidémiologistes et économistes, nous travaillons tous les jours à résoudre des problématiques complexes pour améliorer le bien-être de tous. Nous cherchons à comprendre pourquoi la France a un écart d'espérance de vie en fonction du milieu social, comme par exemple entre ouvriers et cadres supérieurs, plus marqué que ses voisins européens ; comment le Levothyrox, les prothèses mammaires PIP ou le Mediator ont pu rester si longtemps en circulation malgré leur dangerosité ; pourquoi nos chances de mourir varient selon que l'on fait un AVC dans le Nord ou en Occitanie. Cependant, nous nous heurtons toujours à la même difficulté : l'accès limité aux données de santé. Pourtant, en France, des données exhaustives et précises sur nos conditions de vie, notre santé, les soins que nous recevons sont collectées tous les jours. Ces données sont nombreuses, souvent de qualité, mais peu ou pas disponibles pour la recherche. Nous ne comptons plus les travaux ralentis, voire abandonnés, faute d'un accès effectif aux données de santé. Les trois ou quatre ans pouvant être nécessaires à l'obtention d'un simple jeu de données sont totalement inacceptables.

> **IL EST INDISPENSABLE DE FOURNIR UN CADRE ÉTHIQUE, TECHNIQUE, HUMAIN ET FINANCIER, QUI ASSURE UNE MUTUALISATION AU SERVICE DE LA RECHERCHE**

Dépasser une logique propriétaire

Il est urgent pour notre pays que la recherche puisse confronter les données de santé, de recours aux soins, socio-économiques et environnementales. De façon très concrète, une actualisation et un croisement rapide des connaissances auraient permis d'identifier plus précocement les facteurs de vulnérabilité vis-à-vis de l'infection de Covid-19, de son évolution et du vécu du confinement. Les débats techniques et polémiques autour de la mobilisation des données de santé au service de la recherche ne doivent pas occulter l'enjeu principal : la santé de nos concitoyens et notre capacité à faire face aux grands enjeux présents et futurs qui s'y rapportent. Pour cela, il nous faut collectivement changer de regard sur ces données. Cesser de les considérer seulement comme un instrument de gestion administrative pour en faire un outil opérationnel de recherche. Dépasser une certaine logique propriétaire et accepter que leur utilité se révèle dans le partage. Nous devons les voir comme le patrimoine commun des citoyens et ne les utiliser que pour servir la société et la population.

Toutes les institutions de notre pays (hôpitaux, mairies, Assurance-maladie...) génèrent et gèrent des données sous des formes variées, avec des contrôles de qualité variables. Les organismes producteurs ne communiquent pas, et parfois ne désirent pas collaborer ou partager les données avec des tiers, les rendant difficilement accessibles. Tout se passe comme s'ils en étaient eux-mêmes propriétaires alors qu'elles appartiennent de fait aux citoyens.

Les raisons invoquées sont parfois légitimes, à titre d'exemple, ils souhaitent être crédités pour leur travail. A contrario, il n'est pas rare que d'autres raisons soient avancées, telles que la question de la protection des données ou de leur sécurité, dans un pays où la réglementation est l'une des plus strictes en la matière, comme un prétexte à réserver ces données à leurs seuls travaux. Veiller à valoriser leurs travaux est indispensable, accepter que les données leur soient réservées est impensable. Ces données sont un patrimoine commun devant servir à améliorer la santé de tous. Il est indispensable de fournir un cadre éthique, technique, humain et financier à même d'assurer cette mutualisation au service de la recherche. Ce cadre doit répondre à plusieurs impératifs : respecter la vie privée des citoyens ; respecter le travail de tous ceux qui participent à cette œuvre commune ; réserver cette mutualisation à la poursuite de l'intérêt général.

De façon très concrète, nous demandons une gouvernance claire, qui aura pour objectifs : rassembler les données de soins, sociales et environnementales ; faciliter leur accès et leur exploitation rapides ; créer un modèle de valorisation pour les chercheurs ou organismes qui collectent et partagent les données ; flécher les investissements vers la constitution et la pérennisation de notre patrimoine commun de données, en passant par le renforcement des équipes concernées. Il faut imposer le partage des données en conditionnant l'attribution des financements publics à leur mise à disposition facile et rapide pour la collectivité.

Permettre un système d'excellence

Le Ségur de la santé constitue l'occasion d'un geste fort pour amorcer ce mouvement en finançant le développement des entrepôts de données hospitaliers. De même, l'Agence nationale de la recherche pourrait prioriser les appels d'offres ayant pour objectifs la mutualisation et la valorisation scientifique des données. Enfin, les initiatives d'excellence et les trois instituts nationaux d'intelligence artificielle impliqués en santé sont particulièrement bien placés pour porter des initiatives dans ce domaine.

Ces demandes sont réalistes, nous avons toutes les qualités pour permettre un grand saut qualitatif de nos politiques de santé : un système de santé d'excellence, une protection sociale forte, un cadre éthique consensuel et une recherche de grande qualité. Collectivement nécessaire pour anticiper et suivre les crises sanitaires, assurer la sécurité des traitements et lutter contre les inégalités sociales ; individuellement utile pour prévenir les maladies, permettre des diagnostics précoces et une prise en charge adaptée tout au long de la vie, l'accessibilité des données de santé à la recherche n'est pas une question triviale. C'est un déterminant de l'accès aux soins pour la population et de la pertinence de la décision publique, de souveraineté politique. C'est une condition du libre exercice du débat démocratique. ■

Karine Chevreul, *directrice de l'unité Inserm 1123, professeure de santé publique et économie de la santé, université de Paris, Hôtel-Dieu, AP-HP ;* **Cyrille Delpierre,** *directeur de recherche Inserm, épidémiologiste, université de Toulouse-III ;* **Paul Dourgnon,** *directeur de recherche, économiste de la santé à l'Institut de recherche et documentation en économie de la santé (Irdes) ;* **Martine Gilard,** *professeure en cardiologie, université de Brest, CHU de Brest ;* **Michelle Kelly-Irving,** *chargée de recherche Inserm, épidémiologiste sociale, université de Toulouse-III ;* **Bertrand Lukacs,** *chirurgien urologue, hôpital Tenon, AP-HP ;* **Alexandre Mebazaa,** *directeur de l'unité Inserm 942, professeur en anesthésie-réanimation, université de Paris, hôpital Lariboisière, AP-HP ;* **Jean-Louis Pépin,** *directeur de l'unité Inserm UMR 1042, professeur en physiologie clinique, université et CHU Grenoble-Alpes.*

Marcel Goldberg et Marie Zins La plate-forme « Health Data Hub » pose des questions de sécurité majeures

Avec cette infrastructure informatique, qui permet d'héberger et de centraliser les informations de santé des Français à des fins de recherche médicale, il suffit de croiser quelques données simples pour identifier une personne, s'inquiètent les deux épidémiologistes

Le gouvernement a lancé un très ambitieux projet de « Health Data Hub » (HDH) visant à réunir l'ensemble des données disponibles sur la santé des Français, pour développer l'intelligence artificielle (IA) en santé. En effet, la situation actuelle est largement insatisfaisante en raison de la dispersion en de multiples systèmes d'information gérés sans coordination par de nombreux acteurs : hôpitaux, Sécurité sociale, organismes de recherche, universités, registres et enquêtes épidémiologiques, cohortes... On ne peut que souscrire aux objectifs de partage de données et de développement de l'IA en santé et se féliciter de la volonté politique de fournir des moyens conséquents.

Mais si l'intention est louable, réunir toutes les données dans une infrastructure informatique unique est extrêmement dangereux et largement inutile. Le fait de confier sa gestion à Microsoft a suscité de nombreux débats, mais on n'a pratiquement pas évoqué les très graves problèmes que pose le dispositif prévu, même s'il était géré sur une infrastructure nationale.

En effet, le HDH entend centraliser toute donnée collectée dans le cadre d'un acte remboursé par l'Assurance-maladie dans les hôpitaux, en médecine de ville, médecine du travail, pharmacies, services de protection maternelle et infantile, dépistage, enquêtes de santé... La centralisation des données concernant les aspects les plus intimes de la vie des 67 millions d'assurés sociaux chez un hébergeur unique pose des questions majeures de sécurité, car il suffit de croiser quelques données simples pour identifier une personne, avec des conséquences potentiellement très lourdes. La centralisation des données dans une seule infrastructure informatique peut permettre des mesures de sécurité accrues, mais les rend plus exposées aux attaques venant de l'extérieur comme de l'intérieur, avec des impacts plus grands en cas de rupture de confidentialité. Faire courir un tel danger aux personnes ne peut se justifier que si cela est indispensable. Or ce n'est pas le cas : non seulement un système centralisé est dangereux, mais il est largement inutile pour deux raisons essentielles.

La première tient à la qualité des différentes bases de données concernées. Construites dans des buts, des circonstances et avec des méthodes qui, pour la plupart, n'ont rien à voir entre elles, leur qualité et leur validité sont extrêmement variables : « big data » n'est pas synonyme de « good data ». Or les algorithmes d'intelligence artificielle ont besoin de données valides. Avant d'utiliser une base de données, un examen minutieux de ses caractéristiques et de sa qualité, impliquant ceux qui l'ont construite, est indispensable, sans quoi son intégration dans le HDH est inutile.

Aberration scientifique

La seconde raison est l'hétérogénéité de ces bases de données. Il ne suffit pas de regrouper des données de droite et de gauche pour les « faire parler ». Encore faut-il que les données soient interopérables, c'est-à-dire homogènes sur le plan sémantique. Par exemple, si on s'intéresse à l'insuffisance cardiaque, on peut trouver des données dans diverses sources : dossier de service de cardiologie, diagnostic de généraliste ou de cardiologue en ville, déclaration d'un sujet dans une enquête, réseaux sociaux... Mais, selon la source, ce terme n'a pas la même signification ni la même validité. Il faut connaître le contexte et les méthodes du recueil des données, la population dont elles sont issues, etc., le cas échéant le type d'appareil utilisé car on rencontre, par exemple, des électrocardiogrammes ou des images IRM provenant d'appareils différents. Et dans de nombreux cas, cette harmonisation s'avère impossible. C'est pourquoi les algorithmes d'IA sont le plus souvent développés sur une base de données unique.

Il arrive cependant que plusieurs bases de données puissent être rassemblées. Il faut alors les harmoniser. Mais ceci n'a de sens que pour des objectifs spécifiques de recherche et ne peut donc être réalisé qu'au cas par cas ; et implique un travail de comparaison et de définition des données, qui ne peut être réalisé que par les responsables des données concernées, qui disposent de l'expertise et de la connaissance approfondie des données, des conditions de leur recueil, des modalités de validation... Imaginer qu'il sera possible de développer des algorithmes d'IA à partir des données extrêmement hétérogènes uniquement parce qu'elles sont stockées dans un système informatique centralisé est donc une aberration scientifique et technique.

Analyse « distribuée »

Et même si toutes ces difficultés sont résolues et qu'on dispose de plusieurs bases de données véritablement interopérables, il n'est pas indispensable de les réunir dans la même infrastructure informatique. Il existe des méthodes d'analyse « distribuée » où des données gérées dans des systèmes informatiques différents sont exploitées en commun ; ces méthodes sont largement utilisées dans les cas où, pour des raisons de sécurité ou des raisons légales, les données ne doivent pas être transférées hors de leur propre environnement.

Développer le HDH en réunissant les données de 67 millions de Français dans une infrastructure informatique unique est donc une erreur fondamentale qui fait inutilement courir de graves dangers. Si les objectifs de partage de données et de développement de l'IA sont pleinement justifiés, plutôt que d'empiler aveuglément des bases de données hétéroclites, le HDH devrait se concentrer sur des activités réellement utiles, comme par exemple une cartographie analytique des bases de données disponibles, leur mise en réseau, la promotion de règles harmonisées de partage de données, etc. En gardant à l'esprit que les véritables difficultés se situent à la source même des données de santé, comme la crise sanitaire du Covid-19 l'a cruellement mis en évidence : absence de données provenant des Ehpad, insuffisance du nombre de spécialistes du codage des causes de décès, pour ne citer que les manques les plus voyants. ■

Marcel Goldberg *est professeur émérite d'épidémiologie et de santé publique à l'université Paris-Descartes ;* **Marie Zins** *est médecin épidémiologiste, enseignante-chercheuse à l'université Paris-Descartes*