Données massives, *big data* et santé publique

Dossier coordonné par
Margo Bernelin
Sonia Desmoulin
Chargées de recherche CNRS, UMR 6297 CNRS, université de Nantes
Thomas Lefèvre
Maître de conférences, praticien hospitalier, Iris-UMR 8156-997 CNRS Inserm EHESS université Sorbonne Paris Nord, service de médecine légale et sociale, hôpital Jean Verdier (AP-HP)

La quantité de données de santé concernant les personnes (poids, remboursements de soins, habitudes de vie, etc.) s'est accrue avec leur recueil systématique et à grande échelle. Parallèlement, il est devenu possible de traiter ces données massives et de livrer rapidement des informations autrefois difficiles d'accès, d'opérer des rapprochements inattendus et d'offrir des modèles prédictifs inédits. Cela sera-t-il révolutionnaire ou non pour la santé publique ?

Données massives, *big data* et santé publique : de quoi parle-t-on ?

Quelques définitions autour de la thématique de ce dossier avant d'aborder les enjeux que présente le *big data* pour les individus et pour la santé publique.

Santé publique et *big data* : concepts et définitions

Margo Bernelin
Sonia Desmoulin
Chargées de recherche CNRS, UMR 6297 CNRS, université de Nantes
Thomas Lefèvre
Maître de conférences, praticien hospitalier, Iris-UMR 8156-997 CNRS Inserm EHESS université Sorbonne Paris Nord

Santé publique et *big data* sont deux concepts aux définitions variables selon les contextes d'utilisation.

La santé publique

Nous parlons ici de santé publique au sens du domaine qui s'intéresse à la santé des populations, comme complément à la médecine, qui s'intéresse à la santé individuelle. La santé publique présente essentiellement une dimension organisationnelle et de prévention. Bien sûr, la santé des populations est liée à la santé des individus qui les composent, mais l'approche de santé publique est collective. On peut penser aux maladies infectieuses et à la vaccination, par exemple. Une maladie infectieuse de transmission interhumaine concerne l'individu et le groupe : il peut y avoir un intérêt à créer des institutions et des réseaux, à mobiliser des techniques (vaccins) et à adopter des politiques pour tenter de maîtriser, sinon d'éradiquer, une épidémie au niveau d'une population, sans pour autant que l'intérêt individuel direct soit évident pour chacun. Une personne qui pourrait être contaminée, puis véhiculer un virus sans en subir les effets, ou en ressentir des effets mineurs, pourra participer à sa diffusion, notamment vers des personnes qui pourront être symptomatiques, voire subir des conséquences plus graves. La santé publique va ainsi recouvrir l'organisation du système de santé, des professionnels de santé et l'élaboration de politiques publiques de santé ainsi que les moyens de leur mise en œuvre. Elle implique aussi les personnes dans ces différentes dimensions, notamment selon un principe de démocratie sanitaire affirmant que les citoyens doivent pouvoir contribuer, donner un avis, sur les décisions de santé publique.

Les données

Le terme de données recèle quant à lui plusieurs sens. On peut parler de données de la science : autour de cette notion gravitent celles de données probantes, de preuves ou de faits scientifiques. On peut également parler de données au sens d'une mesure, d'une donnée « brute » – même si aucune donnée n'est jamais « brute », étant le résultat d'une construction sociotechnique. Tous ces types de données peuvent être mobilisés dans une prise de décision. Les données au sens de mesures, encore rares il y a peu, semblent se multiplier ces dernières années. La question de leur accès, de leur utilisabilité et de leur utilité se pose alors.

Big data et données massives

Le terme de *big data*, dont l'utilisation large montre qu'il ne peut être réduit à celui de données massives, ne possède pas de définition consensuelle, *a*

fortiori en santé et en santé publique. Son apparition dans la littérature scientifique médicale pourrait être rapportée à un dossier de la revue *Nature* publié en septembre 2008. Le terme est souvent attribué à un document technique d'un cabinet de conseil américain (META group/Gartner), apparu en 2001. En réalité, il n'y est pas fait mention du terme. En revanche, on y voit apparaître la base de définitions reprises assez fréquemment dans différents domaines et par la presse généraliste : les définitions en « V ». Par exemple, les 3 V : volume, variété et vélocité (en français), pour caractériser des données d'utilisation de plus en plus courante, ou que l'on souhaiterait exploiter. *Volume*, pour une grande quantité soit de caractéristiques d'un individu, soit pour un grand nombre d'individus (une cohorte de plusieurs centaines de milliers d'individus), ou bien les deux à la fois. *Variété*, pour souligner que l'on va pouvoir exploiter des données de natures diverses, que l'on peut répartir entre données structurées et données non structurées. Schématiquement, des données structurées seraient représentées par un grand tableau bien défini, où toutes les colonnes correspondent à des mesures standardisées (une tension artérielle en mmHg, une taille en cm), et les lignes, autant de personnes ou patients. Des données non structurées seraient toute autre source de données. Cela peut être de la vidéo, de la parole, du texte libre. *Vélocité*, pour le fait que l'on va traiter ces données très rapidement, en « temps réel ». Ce critère est actuellement moins pertinent dans le champ de la santé, même s'il prend de l'importance dans les situations de gestion de crise, mais il a été très important dans un domaine précurseur pour l'usage du *big data*, à savoir la finance et le *trading* « haute fréquence » (le traitement extrêmement rapide de millions et milliards de transactions financières).

Un autre domaine où le *big data* a été très vite présent est celui du marketing et de la publicité : l'idée est qu'en utilisant et recoupant des données diverses et variées de nombreuses personnes, en particulier des données dites comportementales (activité physique : les pas, le rythme auquel on marche ; consommation : ce que dit notre ticket de caisse ; utilisation de réseaux sociaux) ou des données géolocalisées (adresse personnelle, coordonnées GPS), il devient possible de cerner d'une part nos préférences et d'autre part nos caractéristiques associées à ces préférences : la publicité « ciblée » est née. Plutôt que nous soyons tous exposés sur une page Internet ou sur le bord de la route à la même publicité, des publicités pour un produit plutôt qu'un autre s'afficheront, basées sur notre historique de navigation Internet, le contenu de nos emails ou encore les caractéristiques que nous aurons nous-mêmes rentrés dans notre profil d'utilisation d'un réseau social.

Le concept a fait le voyage depuis ces domaines vers la santé. On trouvera ainsi dans la littérature scientifique médicale tant du *big data* que l'utilisation « ciblée » de données, que l'on dénomme médecine 4 P (sur le modèle des 3 V et plus) : médecine préventive, prédictive, personnalisée et participative.

Les sources de données

Les sources de données en santé se multiplient : données recueillies, numérisées lors d'une hospitalisation ou d'une consultation médicale (diagnostic, traitement, examen médical ou biologique, d'imagerie…), données de consommation de soins (données enregistrées par l'assurance maladie, prescriptions), données d'objets connectés – qu'ils soient médicaux (tensiomètre, implant cardiaque) ou non médicaux à l'origine (pèse-personne, montre connectée, smartphone), et virtuellement toute autre source de données qui pourraient, dans un usage donné, être utiles : données sociodémographiques, géographiques… On comprend qu'en réalité, plus que le caractère massif des données, ce sont deux éléments essentiels qui vont potentiellement introduire une nouveauté : i) le fait de pouvoir accéder puis recouper des données qui d'habitude ne le sont pas, et ii) des moyens d'analyses et de traitement de ces données, comme par exemple « l'intelligence artificielle ». Cependant, ces données peuvent tout autant être analysées par des méthodes plus conventionnelles en médecine et en épidémiologie.

L'intelligence artificielle

L'intelligence artificielle trouve un nouveau souffle, un regain d'intérêt, en partie par le développement de nouveaux algorithmes, mais plus encore par l'accroissement des moyens de calculs (puissance et caractère répandu des ordinateurs) et surtout par l'accessibilité accrue de données couvrant de plus en plus de domaines de la vie quotidienne, y compris la santé. Un attrait et une efficacité renouvelée de l'intelligence artificielle – qui est aussi source de réticence de la part des professionnels de santé vis-à-vis de son utilisation – tiennent dans l'utilisation d'algorithmes qui vont « s'adapter » aux données, et établir des règles de décision non pas introduites *a priori* par l'homme, mais à partir de ce qui a été observé. Cela est bien sûr le fonctionnement théorique, et implique un certain nombre de limites, dont les biais d'apprentissage : un algorithme perpétuera, voire renforcera, les comportements observés et « appris » à partir des données. Ainsi, si les données ne concernaient que des patients de sexe masculin et d'origine caucasienne, rien ne dit que l'algorithme tiré de ces données saura correctement décider pour des patients de sexe féminin ou d'origine asiatique ou afro-caribéenne.

La santé publique de précision

Ces constats sont valables pour la médecine, centrée sur l'individu, mais peuvent l'être tout autant pour la santé publique, centrée sur les populations. On parle notamment depuis quelques années (2013) de santé publique de précision. On pourrait sans doute

tout autant parler de santé publique 4 P, celle-ci ayant dans ses attributions classiques la prévention, la participation (démocratie sanitaire, associations de patients) et la prédiction. Reste la personnalisation : en réalité, en santé publique, on peut chercher comme en marketing à « segmenter » les populations, c'est-à-dire à identifier certains ensembles de personnes qui présentent des caractéristiques similaires par rapport à un problème de santé. Le concept est voisin de celui de stratification du risque : tout le monde n'est pas exposé de la même façon aux mêmes risques de développer une pathologie donnée. À l'inverse, la personnalisation « vraie » – un diagnostic strictement unique, personnel, un traitement unique, personnel – n'existe pas et n'a probablement pas beaucoup de sens en général. De fait, la médecine 4 P est elle aussi basée sur une approche de groupe : le groupe des personnes partageant des caractéristiques similaires, impliquant un même traitement.

Pour l'heure, la santé publique de précision semble se démarquer selon deux grandes dimensions de la médecine 4 P :

- la précision est une précision en termes d'échelles des mesures, par exemple l'échelle spatiale – en effet, les bons résultats d'un indicateur, comme la mortalité infantile selon les régions, peuvent masquer des hétérogénéités géographiques majeures si l'on y regarde à une plus petite échelle que la région ou le pays, et la précision révèle alors une aggravation de l'indicateur en certains lieux, et une amélioration en d'autres, donc une aggravation des inégalités géographiques de santé ;
- ce qui a toujours distingué l'approche sociale de l'approche individuelle, à savoir qu'il existe des caractéristiques propres au collectif, au social, déterminants de l'état de santé individuel et populationnel, qui ne sauraient se réduire à la somme des caractéristiques individuelles. Il ne s'agit au fond que d'une réactualisation, voire une exacerbation de la tension classique entre préférences et approches individuelles, et préférences et approches collectives.

Données massives en santé publique : quels enjeux pour les personnes ?

Margo Bernelin
Sonia Desmoulin
Chargées de recherche CNRS, UMR 6297 CNRS, université de Nantes

Au cœur de la politique de santé publique, « *l'observation épidémiologique et la surveillance de l'état de santé des populations* » s'appuient désormais « *sur les nouveaux outils d'exploitation des données* », ainsi que le proclame Santé publique France sur son site Internet. La gestion de la crise sanitaire née de la propagation du virus SARS-Cov-2 (Covid 19) illustre bien l'intérêt de la collecte et du traitement des données massives pour la protection de la santé des populations, à des fins de diagnostics, de suivi de la progression pandémique et de recherche en santé publique. L'exemple montre aussi que sont traitées à la fois des données personnelles, comme le résultat des tests de dépistage PCR, et des données non personnelles, telles que les relevés de présence du virus dans les eaux usées de grandes villes. Pour être efficace, une telle approche suppose de collecter énormément de données, de natures diverses, et issues de plusieurs bases. Caractérisée notamment par les concepts de *volume* et de *variété*, cette démarche a logiquement des répercussions sur les droits des personnes. La protection de la vie privée et des données personnelles est difficile à tenir. La tension, bien connue en santé publique, entre protection des droits et libertés individuels et protection de la population est ainsi réactivée. En effet, s'appuyant en partie sur le traitement de données personnelles, l'usage des données massives en santé publique nécessite le concours des individus au risque de leur vie privée.

L'utilisation des données personnelles de santé

L'usage des données massives en santé publique s'appuie en partie sur la collecte et le traitement de données personnelles, c'est-à-dire de données qui permettent d'identifier directement ou indirectement des individus. Ainsi, noms, adresses, identifiants et autres numéros spécifiquement attachés à une personne sont autant de données personnelles, qui peuvent être utilisées en santé publique. Au sein de ces données, les données de santé tiennent une place particulière en ce qu'elles informent sur l'état de santé passé, présent ou futur d'un individu, qu'il s'agisse de sa santé mentale ou physique. Par exemple, les résultats sanguins ou de radiographie ou les données de remboursement des soins révèlent l'état de santé d'un individu et sont, par conséquent, des données personnelles de santé. Des données plus anodines peuvent également fournir des informations sur l'état de santé, comme une adresse qui mettrait en évidence l'exposition à un environnement pollué et au risque de développer des pathologies spécifiques. Le traitement de ces données variées en masse pourrait, par exemple, mettre en lumière des liens entre la prise d'un médicament, une zone géographique de domiciliation et l'expression d'effets secondaires.

Une double difficulté naît de cette démarche. La première tient à l'impératifde protection de la vie privée des personnes. Étant susceptibles de révéler des informations sur le comportement, l'état de santé et les conditions de

vie des individus, les données personnelles font l'objet d'un encadrement juridique spécifique (loi Informatique et libertés en France depuis 1978 ; Règlement européen sur la protection des données personnelles (RGPD) adopté en 2018). S'agissant des données de santé, les risques d'usages aboutissant à une discrimination néfaste pour la personne concernée sont importants. En conséquence, ces données sont considérées comme sensibles et ne doivent pas, en principe, être recueillies ni utilisées, sauf pour des motifs dérogatoires tels que les soins vitaux, les services de prestations ou de contrôle par les organismes d'assurance maladie, le traitement par les agences régionales de santé ou les traitements nécessaires à la recherche, aux études et évaluations dans le domaine de la santé. Très souvent recherché, le consentement exprès de la personne permet également le traitement de telles données. Cependant, la protection conférée par le consentement peut être discutée dès lors que la personne est en situation de vulnérabilité (un patient en attente de prise en charge ou de soins par exemple) et/ou que le consentement ne correspond en réalité qu'à un acte irréfléchi (comme un clic mille fois répété sur un ordinateur ou un téléphone portable). Ainsi, en matière sanitaire, un équilibre délicat doit être recherché entre protection des droits des personnes et objectif de protection de la vie et des populations.

La seconde difficulté vient de la possibilité d'explorer différentes bases de données et, en croisant les éléments ainsi collectés, de remonter de manière indirecte à des informations potentiellement sensibles pour les personnes. Dès lors, la distinction entre les données personnelles et les autres données est mise à l'épreuve. Les recherches en informatique montrent que les techniques d'anonymisation ne sont pas toujours efficaces (par exemple pour les données génétiques) ou qu'elles trouvent rapidement leurs limites lorsque plusieurs bases de données peuvent être croisées.

La tension entre protection des droits individuels et protection de la population

Si l'impératif de protection de la population semble ici prendre le pas sur le souci de préserver l'intimité des individus et les usages discriminatoires, cette préoccupation ne disparaît pas. Elle se traduit dans la mise en place de statuts spéciaux (hébergeurs de données de santé) et d'outils spécifiques (système centralisé sous la responsabilité d'autorités publiques). Cette tension entre protection individuelle et protection collective n'est pas nouvelle. En effet, dès le XX^{e} siècle des règles nationales imposaient aux médecins de déclarer les cas de rougeole ou de scarlatine. Les règlements sanitaires municipaux pouvaient également imposer aux hôteliers et aubergistes les mêmes obligations de déclaration. Ainsi, des données personnelles étaient déjà collectées afin d'endiguer une épidémie par l'organisation de quarantaines.

Aujourd'hui, la protection des données personnelles de santé devient d'autant plus difficile à assurer que les données des patients ne sont plus seulement traitées à un échelon local (celui du CHU par exemple), mais à un niveau national. Une grande partie des données est concentrée au sein de la plateforme nationale des données de santé Health Data Hub (données sur les soins pris en charge par l'assurance maladie, données de remboursement des soins…). À ces données pourraient s'en ajouter d'autres, plus hétérogènes : adresses, géolocalisation, contacts, photos, etc. Certains réseaux sociaux proposent, par exemple, de collecter des données de géolocalisation afin de suivre l'évolution de la pandémie de Covid-19. Ainsi, il ne s'agit plus simplement de déclarer aux institutions nationales une pathologie avérée, mais de collecter par différents biais un nombre important de données très diverses auprès de patients mais aussi, et surtout, de personnes en bonne santé.

Bien que les objectifs portés par la santé publique soient louables, le flot des données collectées, notamment sur les déplacements des malades ou sur les contacts avec des personnes à risque, interroge sur notre capacité à maintenir un cadre protecteur pour la vie privée et à éviter que les données collectées ne soient ensuite réutilisées à des fins discriminatoires ou commerciales. De graves failles de sécurité sont régulièrement signalées, comme dans l'exemple donné par l'association Greenbone Networks sur les données d'IRM et de scanner[1]. Or, les données peuvent intéresser de nombreux acteurs privés, comme les assureurs afin d'analyser le risque « santé » de chaque assuré, mais aussi les banques ou certains employeurs.

Par conséquent, le concours des individus à la santé publique à l'ère des données massives pose des questions particulières que le seul impératif de solidarité dans le partage des données ne saurait balayer. L'emploi des données massives a toutefois d'indéniables effets positifs sur les droits individuels, notamment en matière de droit à l'information, ainsi que le montre la création du « Tableau de bord de suivi de l'épidémie de Coronavirus en France », centralisant les données de Santé publique France, de la Direction générale de la santé et les données sur les transferts de patient. Un même outil peut servir des finalités collectives et de protection des personnes.

Conclusion

L'usage des données massives pour la santé publique replace plus que jamais l'individu au cœur de la promotion et de la protection de la santé de la population. Cependant, cette mobilisation n'est pas sans risques, notamment concernant le dévoilement d'informations sur le comportement, l'état de santé et la vie privée des personnes. Le thème de ce dossier invite donc à réinterroger le cadre et les objectifs de la santé publique, ainsi qu'à tester la robustesse des équilibres établis à la lumière du recours croissant aux données massives.

1. https://www.lemondeinformatique.fr/actualites/lire-plus-d-un-milliard-d-images-medicales-exposees-77688.html

Enjeux pour la santé publique du traitement des données massives diverses

Margo Bernelin
Sonia Desmoulin
Chargées de recherche CNRS, UMR 6297 CNRS, université de Nantes

« *La médecine de demain ne sera pas la même que celle d'aujourd'hui : prédictive, personnalisée, numérique, elle devra sans cesse s'adapter aux nouveaux enjeux et aux nouvelles technologies*[1]. » C'est par cette phrase que la Stratégie nationale de santé 2018-2022 indique vouloir faire de l'introduction du numérique en santé, et notamment des *données massives*[2], l'avenir des politiques publiques en santé et positionner la France dans un mouvement global. Ainsi, tant la médecine que la santé publique doivent bénéficier des avancées du numérique. Dans ce cadre, le développement d'outils algorithmiques exploitant des masses de données diverses au service de la santé des populations semble constituer, avec les nouvelles capacités d'archivage, une innovation phare de ce début de XXIe siècle. Cela se combine avec la possibilité d'exploiter l'ensemble des bases de données publiques et privées constituées notamment par la collecte de données issues d'objets connectés. Les données exploitées peuvent être très variées : signaux biologiques, dépenses de santé, observance des traitements, facteurs environnementaux, habitudes de vie, etc. Aussi, le volume des données traitées et leur diversité sans précédent ouvrent la voie à de possibles transformations majeures en santé publique.

De nombreuses perspectives

Les pouvoirs publics attendent des études menées sur les données collectées l'élaboration de nouveaux plans de prévention en santé, des analyses inédites en matière de risques épidémiques, des traitements mieux ciblés pour les patients et la découverte de voies de recherche insoupçonnées en santé publique, par exemple, de nouveaux facteurs de risques associés à des pathologies pour développer des mesures préventives. Le contexte pandémique né de la propagation du virus SARS-Cov-2 (Covid 19) a montré l'intérêt porté par les pouvoirs publics à l'analyse de données, afin d'anticiper l'évolution de la situation. En France, Epidemap, outil de modélisation issu de trois modèles traitant respectivement des données sur la répartition des habitants (bâtiment par bâtiment), des simulations de déplacements et de données d'épidémiologie a été conçu pour établir des scénarios de propagation du virus en fonction d'options de confinement à domicile ou de déconfinement. Il a participé à légitimer les mesures de gestion de crise[3]. Pour lutter contre la transmission virale, les autorités publiques ont également soutenu le développement d'applications mobiles ou de plateformes telles que StopCovid ou ContactCovid. Ces outils visent à collecter des données pour informer les personnes concernées d'un contact possible, durant les jours précédents, avec une personne dépistée ultérieurement, mais aussi pour informer les autorités publiques et les institutions sanitaires sur les parcours géographiques des personnes (dépistées ou non) et sur la détermination des cas contacts. Le relatif échec de l'application StopCovid (beaucoup moins utilisée en France qu'en Allemagne, par exemple) n'a pas remis en question la volonté du gouvernement français, puisqu'une nouvelle version de l'application mobile est prévue. Rebaptisée TousAntiCovid, elle inclut une dimension informative supplémentaire sur les risques sanitaires (symptômes de la maladie, statistiques…) et sur les lieux de dépistages, en plus de l'alerte sur les « cas contacts ».

Toutefois, ces évolutions s'accompagnent de nouvelles interrogations. En sus des enjeux de protection des droits et libertés individuels, en lien avec la protection de la vie privée et des données personnelles, des questions relatives aux frontières et aux acteurs de la santé publique sont ainsi posées. La délimitation des champs de la santé publique était déjà sujette à discussion, en ce qu'ils pouvaient être restreints à l'approche populationnelle et aux dimensions institutionnelle et organisationnelle, ou inclure plutôt, de manière large, tous les domaines couverts par le Code de la santé publique (y compris la médecine). L'approche par le croisement de données diverses en grand volume semble élargir encore les perspectives et donner une importance particulière au concept d'exposome récemment introduit dans le Code de la santé publique et voulant que « *l'intégration sur la vie entière de l'ensemble des expositions qui peuvent influencer la santé humaine*[4] » constitue un objectif primordial des politiques de santé publique. Si toutes sortes de données, constamment collectées, peuvent être compilées et croisées pour concevoir de nouvelles politiques et de nouvelles mesures de prévention, la détermination de ce qui relève spécifiquement de la santé publique devient ardue. Tout relève alors possiblement de la santé publique.

1. Stratégie nationale de santé 2018-2022. Ministère des Solidarités et de la Santé, Paris : 2017, p. 63.

2. Il n'existe pas de terme légal ou consensuel en France pour la traduction de l'expression big data. Deux expressions sont couramment employées : mégadonnées et données massives. Toutes deux ont l'inconvénient de ne pas suffisamment insister sur la caractéristique de diversité des données traitées qui, en sus de la rapidité de traitement grâce aux nouveaux outils informatiques et techniques algorithmiques, constitue la spécificité de l'approche par le big data. Sur le concept, voir le chapitre initial de ce dossier.

3. https://www.data.gouv.fr/fr/organizations/epidemap/

4. Art. L. 1411-1 CSP.

Une intervention grandissante du secteur privé

Quant à la dimension organisationnelle, elle est également bousculée par l'idée qu'il ne serait peut-être plus nécessaire de passer par des institutions et des réseaux créés par l'État ou par les autorités publiques pour traiter les alertes et transmettre les messages de prévention. La place du secteur privé dans les activités de collecte, de stockage et de traitement des données de santé (applications en ligne, plateformes nationales) interroge. Alors que par le passé la collecte et l'analyse des données à des fins de santé publique étaient essentiellement opérées par des services de l'État à des fins d'intérêt général, elles s'appuient aujourd'hui plus souvent sur le concours d'entreprises privées non spécialisées en santé. Les plateformes développées par des réseaux sociaux pour suivre la pandémie de Covid-19 en sont un exemple[5]. Désormais, les entreprises du numérique ou les assureurs s'impliquent au-delà de la simple collecte de données, par la transmission de messages à caractère sanitaire auprès de leurs abonnés-clients-assurés. Un tel constat n'est pas anodin. D'une part, une collecte si volumineuse de données personnelles par des sociétés privées enrichit encore un peu plus leur catalogue de données, faisant peser un risque accru sur la vie privée des citoyens dont les vies sont documentées. D'autre part, la qualité et la fiabilité des informations sanitaires offertes par ces entreprises peuvent être questionnées. En effet, comment s'assurer de la fiabilité du contenu d'une application qui partage des informations sur les moyens de lutte contre certaines pathologies ou virus ? Le passage de la donnée/mesure à l'information et à la donnée probante est ici crucial. Par exemple, les informations offertes par l'agence Santé publique France s'appuient sur le concours avéré de chercheurs dont les moyens d'investigation peuvent être retracés.

La place grandissante du secteur privé se fait également sentir dans les activités de conservation des données. Des entreprises privées telles que Google, Amazon ou Microsoft proposent leurs services. C'est dans ce contexte que le Health Data Hub, plateforme nationale des données de santé, appelée à jouer un rôle central pour la recherche en santé publique, a annoncé que la société étatsunienne Microsoft serait en charge de l'hébergement des données de santé, bien qu'elles soient souvent considérées comme un « trésor national ». Ce choix a été principalement justifié par les capacités importantes d'hébergement de cette société. Sous réserve que les conditions d'archivage respectent le droit européen en matière de protection des données personnelles, le droit français ne s'oppose pas à cette imbrication « public/privé », mais elle a pour conséquence de rendre l'État dépendant du secteur privé en matière de santé et notamment de santé publique. De nombreuses critiques ont été formulées à cet égard, pointant l'inopportunité de choisir une entreprise étrangère qui dispose déjà d'une puissance financière et commerciale considérable.

Au final, l'introduction des données massives diverses en santé publique paraît renouveler profondément cette discipline centenaire. En effet, les acteurs de la santé publique se diversifient notamment au profit du secteur privé et d'entreprises dont les activités étaient jusqu'alors fort éloignées du champ de la santé. La discipline semble également se renouveler dans ses pratiques, les contributions à ce dossier montrant que les recherches en santé publique évoluent sous l'attraction de la disponibilité nouvelle de données nombreuses et variées. Les objectifs de prévention des risques et de promotion de la santé de la population semblent néanmoins perdurer. Il reste toutefois à vérifier si, dans l'avenir, l'évolution vers toujours plus de stratification (en sous-groupes de populations en fonction de caractéristiques sociales, environnementales, physiologiques ou génétiques notamment) et vers une médecine « de précision » (adressant des messages préventifs et des propositions thérapeutiques individualisés) ne remettra pas en cause l'idée même d'une approche populationnelle globale et de choix collectifs.

5. Covid-19 Interactive Map & Dashboard, plateforme proposée aux abonnés de Facebook. https://covid-survey.dataforgood.fb.com/survey_and_map_data.html

Évolution ou nouvelle donne ?

L'usage des données issues du *big data* est-il une révolution pour la santé publique ? La diversité des sources, la réutilisation de données, l'intervention d'acteurs privés et la création d'un système national des données de santé transforment la santé publique.

Données dites « massives » et santé publique : une mise en perspective historique

Joël Coste
Université de Paris, École pratique des hautes études

Depuis sa conversion à la quantification au cours du XIX^e^ siècle, la santé publique n'a cessé de mobiliser des données dont le volume était en rapport avec les possibilités de calcul du moment. La mise en perspective historique opérée dans cet article permettra de faire la part de la rhétorique, répétitive, mettant en exergue les promesses, pour la santé publique, du recours aux données dites « massives », comme les éléments ou les enjeux plus originaux de la mobilisation de données génétiques, biologiques, comportementales ou encore de remboursements de soins que l'expression actuelle de « données massives » tend à amalgamer et dont le volume n'est probablement pas la caractéristique la plus originale ni la plus problématique.

L'expression *big data* est apparue pour la première fois au grand jour dans la revue *Nature* en septembre 2008. À l'occasion des dix ans de Google, *Nature* demanda en effet à des chercheurs et des industriels de champs différents quelles *technologies « pourraient autant changer le monde »* que Google à l'horizon 2018. Les « données massives » perçues comme susceptibles de « changer le monde » furent d'abord biologiques, génétiques et génomiques puis électro-physiologiques, avant d'inclure, dans les années qui suivirent, les dossiers médicaux, les données de remboursement de soins et enfin les données recueillies sur les réseaux sociaux ou fournies par les objets connectés.

La santé publique et, sa principale science pourvoyeuse de connaissances, l'épidémiologie, avaient vocation à s'intéresser aux *big data*. Elles exploitaient d'ailleurs depuis longtemps des *données de grande taille* sans avoir créé de nom pour celles-ci, et pour paraphraser Molière, elles les analysaient comme Monsieur Jourdain faisait de la prose.

Cet article retracera dans un premier temps l'usage par l'épidémiologie et la santé publique des données de grande taille, toujours aux limites des possibilités de calcul du moment, et rappellera les méthodes qu'elles ont contribué à développer pour leur analyse, avant de considérer dans un second temps les principaux problèmes posés par l'utilisation des *big data*, du moins des données regroupées sous ce terme depuis 2008. Certaines de ces utilisations, maîtrisées, ont déjà permis des développements pertinents, notamment en épidémiologie génétique et en pharmaco-épidémiologie. D'autres ont conduit à des échecs retentissants, et les dernières n'ont pour l'instant pas franchi l'étape de faisabilité ou de preuve du concept, malgré la com-

munication hyperbolique qui les accompagne souvent. Il sera montré qu'aux problèmes créés par la taille des données s'ajoutent des problèmes spécifiques d'autant plus ardus à résoudre que les données sont recueillies dans des champs plus éloignés de la biologie et de la médecine. Les aspects éthiques, particulièrement problématiques, de certaines utilisations proposées ne seront pas évoqués dans cet article, centré sur les questions épistémologiques.

Des statistiques sanitaires aux grandes cohortes et au Global Burden of Diseases, les données de grande taille au service de l'épidémiologie et de la santé publique

Il est d'usage de reconnaître dans les *Natural and Political Observations Made Upon the Bills of Mortality* de John Graunt (1662) l'acte de naissance des statistiques sanitaires. Exploitant des données recueillies pendant presque soixante ans, Graunt tabulait dans cet ouvrage les naissances, mariages et décès des Londoniens des deux sexes, par année et par paroisse, ainsi que les causes de décès, analysées en détail pour plus de 229 000 d'entre eux. Un ami de Graunt, le médecin William Petty, conduisit quelques années plus tard de grandes enquêtes, fortement quantitatives, sur les conditions démographiques, économiques et sanitaires de l'Irlande, qui furent à l'origine de l'« arithmétique politique » puis de la « statistique » – de l'allemand *Statistik*, un terme forgé en 1748 pour désigner les connaissances chiffrées nécessaires à l'État.

La statistique, et notamment la statistique sanitaire, se développa progressivement en Europe au XVIII[e] siècle – la Suède, par exemple, recueillit les causes de décès sur l'ensemble de son territoire à partir de 1749 –, mais ce fut certainement l'établissement du Registrar General Office pour l'Angleterre et le Pays de Galles, créé en 1837, et son département de statistique, dirigé jusqu'en 1879 par William Farr, qui permit à la santé publique de franchir une étape décisive dans l'usage de données quantitatives. Adossé au registre d'état civil et aux recensements réguliers d'une population de plus de 20 millions d'habitants, comportant 500 000 naissances et presque autant de décès chaque année – soit des données de taille alors inédite –, Farr réalisa de nombreuses études qui montrèrent le rôle des facteurs socioéconomiques et territoriaux de la santé. Farr entreprit aussi de *standardiser* le recueil des données, à commencer par la nomenclature utilisée par les médecins, et initia par là un mouvement qui conduisit à la mise au point de classifications internationales des maladies, qui furent utilisées dans la plupart des pays du monde dès les premières décennies du XX[e] siècle. Celles-ci fournissaient des instruments indispensables pour la réalisation d'études comparatives internationales de grande ampleur, dont le dernier avatar en date, le Global Burden of Diseases (GBD), présenta pour 2016 les indicateurs de 333 pathologies et états de santé dans 195 pays et territoires.

Une autre étape importante dans l'exploitation de volumineuses données pour la santé publique fut franchie après 1945, avec le développement de l'épidémiologie dite « moderne », celle des « facteurs de risque » et des maladies chroniques [40]. De grandes cohortes de sujets furent alors assemblées et suivies de nombreuses années, dont celle, emblématique, de Framingham, commencée en 1947 et concernant plus de 5 000 personnes de cette ville. Le nombre élevé de variables d'exposition à tester dans cette cohorte (28 mentionnées dès 1949) nécessita rapidement la mise au point de techniques statistiques d'analyse multivariée (d'abord de discrimination linéaire puis de régression logistique) qui ne purent être mises en œuvre qu'avec l'aide d'*ordinateurs*. Ces derniers accompagnèrent ensuite l'épidémiologie dans ses développements et permirent l'analyse de données de cohortes dépassant 100 000 (Pays-Bas, 1986), 500 000 (États-Unis, 1995), 1 320 000 (Royaume-Uni 1996-2001), voire 6 500 000 sujets pour la Cancer Epidemiology Descriptive Cohort Database regroupant 46 cohortes américaines (2015). Des registres de maladies concernant des milliers, voire des dizaines de milliers, de sujets furent également mis en place en Amérique du Nord et en Europe à partir des années 1970, certains d'entre eux recueillant des données de suivi nombreuses sur plusieurs années.

Les problèmes génériques créés par l'exploitation des données de grande taille en épidémiologie et santé publique

La tendance au gigantisme qui caractérise l'épidémiologie moderne, et qui s'est accélérée parallèlement à celle des capacités de calcul des ordinateurs dans les années 1990 et 2000, s'explique avant tout par la quête de la puissance statistique. Une fois les déterminants majeurs de la mortalité ou de la morbidité par cancer ou par maladies vasculaires identifiés et en partie contrôlés (tabagisme, hypercholestérolémie, hypertension artérielle, obésité), il devint nécessaire de recourir à des effectifs importants de sujets pour étudier des expositions et des manifestations de santé rares, ou encore mettre en évidence des effets faibles. Cette augmentation de puissance des analyses a eu pour contrepartie, pas toujours bien comprise, de permettre la mise en évidence de différences futiles ou bien non reproductibles en raison de l'erreur statistique de première espèce (ou « risque α ») et du phénomène d'*overfitting* (surajustement des modèles aux données). Deux autres conséquences de cette course à la puissance, cette fois liées aux conditions de recueil et d'agrégation de données volumineuses, ont été la création d'hétérogénéités artificielles, liées à des différences de recueil et surtout à l'inflation des données manquantes, concernant souvent des groupes particuliers (moins favorisés et moins alphabétisés, participant moins aux enquêtes) et responsable de biais de sélection que les modèles statistiques d'imputation des données manquantes, développées depuis les années 1980,

Les références entre crochets renvoient à la Bibliographie générale p. 57.

ne peuvent corriger entièrement. Ces biais peuvent être considérables, surtout dans les études portant sur des professionnels, sur des usagers des systèmes de soin ou encore chez des volontaires utilisant des technologies modernes comme Internet.

Les problèmes spécifiques des nouvelles « données massives »

Les problèmes liés au gigantisme des données évoqués précédemment se retrouvent naturellement dans l'analyse des *big data*. Jusqu'à présent toutefois, seul le champ de la recherche génétique, en charge de l'analyse des données « omics » (génomiques, épigénomiques, transcriptomiques, protéomiques, métabolomiques, etc.) a pris la mesure de ces problèmes et instauré – au début des années 2010, dans la perspective d'applications cliniques mais aussi après quelques erreurs retentissantes –, des règles de bonne pratique assez strictes, impliquant généralement le contrôle de l'erreur de première espèce et surtout la *réplication des résultats avant leur publication* – une procédure qui avait d'ailleurs été préconisée dès les années 1990 pour la construction des échelles de risque.

Les données administratives et cliniques recueillies en routine ont aussi fait l'objet de recommandations [43] mais celles-ci concernent essentiellement la description des méthodes dans les publications et sont moins normatives des pratiques de recherche. Les intérêts et les limites des données présentes dans les bases médico-administratives (en France les bases de l'Assurance maladie et du programme de médicalisation des systèmes d'information [PMSI]) pour la recherche épidémiologique ont été souvent soulignés ces dernières années [54]. Servant à la facturation, les données sont généralement fiables mais ne fournissent que des représentations pointillistes et surtout indirectes de l'état de santé des sujets, sauf en cas de maladie sévère (avec hospitalisations répétées) ou traitée un certain temps avec des médicaments remboursés. Les erreurs de mesure non différentielles sont donc importantes, et pas toujours bien prises en compte alors qu'elles réduisent la force des associations et la puissance statistique. L'absence dans ces bases des facteurs socioéconomiques a des conséquences beaucoup plus sérieuses, puisqu'ils sont des déterminants et des facteurs de confusion de nombreux phénomènes de santé. À ce jour, ce sont surtout les études de pharmaco-épidémiologie qui ont le mieux utilisé le potentiel des bases médico-administratives, en contournant un certain nombre de difficultés posées par celles-ci, au prix toutefois d'analyses longues et complexes pour en assurer la robustesse.

L'exploitation des données hospitalières n'en est quant à elle qu'à ses premiers balbutiements, nonobstant la communication hyperbolique de certains hôpitaux qui voudraient en tirer un profit financier. L'hétérogénéité des méthodes de recueil, l'absence de contrôle de la qualité, la faible standardisation du vocabulaire et des comptes rendus médicaux constituent de redoutables difficultés que les techniques de traitement automatisé ou d'« intelligence artificielle » (IA) – mieux vaudrait parler d'*algorithmique* – ne peuvent contourner qu'après de longs paramétrages. Quant aux techniques statistiques d'exploitation de ces données (classifications automatiques, forêts aléatoires, réseaux neuronaux, scores de propension de haute dimension, etc.), même parfaitement maîtrisées et mises en œuvre, elles ne peuvent en aucun cas corriger l'absence de données importantes et le caractère hautement sélectionné des populations fréquentant les structures de soin dans des parcours eux-mêmes déterminés par de nombreux facteurs qui échappent à l'enregistrement.

Les données recueillies sur les réseaux sociaux ou fournies par les objets connectés individuels sont encore à un stade plus préliminaire d'exploitation. L'échec de Google Flu Trends à prédire avec une raisonnable validité le nombre de consultations pour grippe aux États-Unis entre 2011 et 2013 puis l'abandon final du projet – bien plus discret que son lancement – en 2015 (comme celui de son produit dérivé Google Dengue Trends) ont conduit à reprendre la réflexion sur la nature et la pertinence des données à utiliser, et aussi sur le phénomène de surajustement des données (voir plus haut). La saisonnalité des matchs de basket peut ressembler à celle de la circulation de la grippe, mais vouloir prédire la survenue de la grippe par l'augmentation des recherches sur les matchs de basket, comme l'ont fait les informaticiens de Google [38], illustre bien les limites d'une approche exclusivement *data driven*, négligeant les déterminants causaux et les mécanismes biologiques qui produisent les événements pathologiques.

Au-delà des représentations et des perspectives hyperboliques et futuristes

Les représentations et perspectives hyperboliques et futuristes qui ont accompagné l'évocation des *big data* depuis 2008 pourraient sembler inédites. Elles furent toutefois déjà utilisées au temps des premiers usages de l'ordinateur au début des années 1960 : celui-ci allait, affirmait-on, remplacer le médecin et ses décisions hasardeuses, et l'IA allait surpasser l'intelligence humaine. On sait ce qu'il en est advenu de l'ordinateur médecin, et du programme de l'IA, réduit depuis à sa dimension « faible », mais efficace, la dimension algorithmique. Au-delà de cette rhétorique, répétitive, mettant en exergue les promesses des *big data*, restent donc de nouvelles sources de données permettant potentiellement de répondre à des questions pertinentes – les données ne posent pas de question – et dont l'analyse, si elle parvenait à être maîtrisée à l'exemple des études « omics » et de la pharmaco-épidémiologie, ne pourrait que contribuer à enrichir les connaissances épidémiologiques et à aider la décision en santé publique.

Mieux vaut bien sûr des données *grandes que petites*, mais le potentiel des données massives ne pourra être exploité que si les leçons de soixante-dix ans

POST-SCRIPTUM

Préparé avant la crise de la Covid-19, cet article n'a pas dû être corrigé lors de sa relecture à la fin de la première phase de celle-ci. Cette crise offrait des opportunités d'utilisation des *big data*. Toutefois, huit mois après son commencement, les données massives et l'IA n'ont apporté ni connaissance pertinente ni dispositif efficient pour la gestion de la crise, et les données n'ont éclairé la décision que dans la mesure où elles avaient été recueillies dans cet objectif et pouvaient être traitées rapidement. En France, c'est plutôt *l'absence de données disponibles* (sur l'origine géographique des sujets, l'activité des médecins généralistes, l'activité du secteur médicosocial…) et *l'impossibilité d'un traitement rapide de certaines d'entre elles* (les causes médicales de décès, un problème déjà souligné en 2003) qui ont été constatées, une nouvelle fois, lors de cette crise.

d'épidémiologie moderne sont retenues : assurer la représentativité des données ou minimiser les biais de sélection, limiter les erreurs de mesure, contrôler les phénomènes de confusion, maîtriser le « risque α » et s'assurer de la robustesse des résultats par leur réplication. De même, les conditions et la logique du recueil des données doivent être prises en compte et cela d'autant plus qu'elles ont été collectées à distance du champ biologique et médical. À l'exact opposé des annonces hyperboliques, il s'agirait donc plutôt d'une approche modeste, patiente, méticuleuse et respectueuse des données, préservée évidemment des liens d'intérêt et des utilisations mercantiles. Mal questionnées, mal analysées et imprudemment interprétées, les nouvelles données massives ne promettront que de grands échecs.

La « santé publique de précision » : un changement de paradigme pour la santé publique ou la perte de son âme ?

Élodie Giroux
Maître de conférences en philosophie des sciences et de la médecine, université Jean Moulin Lyon 3, Institut de recherches philosophiques de Lyon, EA4187

La « médecine personnalisée » désigne le fait de cibler le traitement et la prévention en fonction du profil, souvent génomique, de chaque individu. Depuis 2011, l'expression de « médecine de précision » se substitue progressivement à celle de « médecine personnalisée ». Cette dernière laisserait entendre à tort qu'il s'agit de développer des traitements spécifiques pour chaque individu alors qu'elle repose plutôt sur une stratification en sous-groupes. La médecine de précision est aussi davantage liée au recueil de données massives et aux technologies associées pour leur analyse. Dès 2013, la notion de « santé publique de précision » (SPP) a été proposée comme son complément. Cependant, n'y a-t-il pas une contradiction dans les termes même, ou tout au moins une tension forte ? Si parler de santé publique « personnalisée » paraît un oxymore, lui appliquer la notion de « précision » est-il davantage pertinent ? Le propre de son action et de son efficacité n'est-il pas d'être collective : par exemple la vaccination ou la réglementation sur le port de la ceinture en voiture ? Que pourrait apporter la « précision » qui ne dénaturerait pas cette spécificité de la santé publique ? En réalité le débat sur la pertinence de promouvoir la santé publique de précision dépend à la fois de ce qu'on entend par « précision » et par « santé publique ».

Une appellation problématique : vers une individualisation de la santé publique ?

En médecine, la précision se développe essentiellement en cancérologie, suite aux progrès de la connaissance au niveau moléculaire, eux-mêmes étroitement liés aux technologies du séquençage du génome. Décliner cette approche en santé publique conduit à prendre en compte l'hétérogénéité individuelle au niveau génomique afin de cibler les sous-populations les plus à risque. La « santé publique génomique », définie en 2012 par Cleeren et ses coauteurs [14] comme « *l'analyse de la manière dont la connaissance et les technologies basées sur le génome peuvent être intégrées dans les services de santé et la politique publique de manière responsable et efficace pour le bénéfice de la population* », est en effet le précurseur de ce qui est aujourd'hui désigné par santé publique de précision. À première vue, celle-ci semble donc renforcer le mouvement d'individualisation de la prévention. Elle s'inscrit dans la continuité de l'approche « facteurs de risque » de l'épidémiologie analytique qui se focalise sur des facteurs individuels biologiques et comportementaux, et désormais génomiques. Dans cette conception, on considère que la santé de la population est la somme des santés individuelles et qu'il est plus efficace d'agir au niveau individuel.

Or, la prise en compte des caractéristiques génomiques pour améliorer la prévention auprès des individus a-t-elle suffisamment fait ses preuves pour pouvoir être généralisée à la santé publique ? Cleeren et ses coauteurs mettent eux-mêmes en garde : « *La génétique est à double tranchant, elle peut conduire soit à renforcer, soit à réduire, les disparités de santé dans la population.* » Surtout, le propre de la santé publique n'est-il pas de repérer des facteurs de risque qui ne sont pas réductibles ou mesurables au niveau individuel ? Certains considèrent que la santé publique se caractérise avant tout par son mode collectif d'intervention et par l'analyse des causes de nature sociale, économique, environnementale et politique. C'est au niveau de la population, irréductible au niveau individuel, que se structurent les inégalités de santé dont l'analyse et la réduction sont l'un des enjeux majeurs de la santé publique. Se focaliser sur le niveau individuel risque de faire perdre ces éléments de vue.

Améliorer la santé publique : renforcer la stratégie ciblée ou du « haut risque »

Néanmoins, la santé publique de précision accorde de l'importance au niveau populationnel : Khoury, figure clé de ce courant, critique la médecine de précision et la médecine des 4 P (médecine préventive, prédictive,

Les références entre crochets renvoient à la Bibliographie générale p. 57.

personnalisée et participative) pour leur négligence de la perspective populationnelle et défend l'importance d'un cinquième P (population) [35, 36]. La santé publique de précision transposerait à ce niveau le principe de la médecine de précision : « *réaliser la bonne intervention, sur la bonne population, au bon moment* ». Renforcer la stratégie qui consiste à mieux cibler les sous-populations les plus à même de bénéficier d'une intervention, dite stratégie du « haut risque » selon la terminologie de l'épidémiologiste Rose, est loin d'être inutile. Les difficultés rencontrées en termes de rapport coût-bénéfice des politiques de dépistage massif de certains cancers conduisent à défendre une stratégie visant à écarter les personnes qui n'en tireraient pas forcément un bénéfice individuel.

Dans le cadre du recueil de données massives de nature pluridimensionnelle, la génomique n'est considérée que comme un moyen parmi d'autres pour mieux identifier les populations les plus à risque. La santé publique traditionnelle utilise déjà des critères d'âge, par exemple en recommandant le dépistage de l'hépatite C dans la sous-catégorie de personnes qui sont nées entre 1945 et 1965. Dans le cadre de la surveillance de maladies infectieuses, pouvoir tracer les individus contaminés grâce aux technologies de santé connectée apparaît déterminant pour réduire l'étendue du confinement : seules les personnes ayant été en contact avec ces cas sont mises en quarantaine. Mais dans cette perspective, on peut se demander ce qu'a de nouveau la santé publique de précision, en dehors de l'introduction des technologies associées à la génomique et au recueil massif de données individuelles. Car cette double stratégie « populationnelle » et du « haut risque » existe déjà en santé publique traditionnelle.

Les limites de la stratégie ciblée ou du « haut risque »

Ce qui pourrait néanmoins être considéré comme une évolution introduite par la santé publique de précision serait de donner la priorité à la stratégie du « haut risque ». Mais un certain nombre de présupposés se révèlent ici problématiques. Tout d'abord, on pense pouvoir réaliser des prédictions individuelles solides, c'est-à-dire extrapoler des prédictions de risque formulées au niveau de la population à des individus. Or ce n'est pas sans poser de redoutables difficultés ; et c'est en réalité au niveau de la population elle-même que ces prédictions sont le plus valides. Ensuite, on considère que ce genre de prédictions permettrait à chaque individu de modifier ses comportements. Or il a été montré que c'est loin d'être le cas. Enfin, on suppose aussi que le risque serait bien délimitable et catégorisable. Or nombre d'entre eux sont de « petits » risques continus et diffus (comme la pollution de l'air) ne permettant pas de discriminer quels individus sont le plus à risque. Ils sont pourtant ceux qui engendrent le plus grand nombre de pathologies.

C'est précisément cette difficulté à délimiter les risques au niveau individuel qui justifie, pour Rose, la centralité et la primauté de la stratégie populationnelle dans la santé publique. Elle permet d'assumer ce qu'il appelle le « paradoxe de la prévention » : un grand nombre de personnes dont le risque est faible donnent lieu à un plus grand nombre de cas de maladie qu'un petit nombre de personnes à haut risque. Elle est adaptée pour nombre de facteurs environnementaux ou sociaux dont l'effet est diffus et qui sont impliqués dans de nombreuses maladies chroniques. En outre, les facteurs de risque ciblés dans la stratégie du « haut risque » que promeut la santé publique de précision restent liés à l'individu, à sa biologie ou à son comportement. Or de nombreux travaux montrent que ces facteurs comptent pour une faible part de la variation dans le risque de maladie au niveau de la population. Les inégalités de santé sont essentiellement liées à des facteurs sociaux structurels et contextuels.

Reconceptualiser la précision à partir de la santé publique

En fait, pour Ostald et son coauteur [45], une telle approche de la santé publique de précision constitue en réalité une médecine de précision *pour la population* mais non pas une santé publique de précision proprement dite. En effet, pour eux, la santé publique a bien pour souci premier la causalité sociale, structurelle et contextuelle des inégalités de santé. Néanmoins, les auteurs partagent avec les promoteurs de la santé publique de précision le souci de renouveler la santé publique traditionnelle, dont les stratégies populationnelles et ciblées manquent d'efficacité, en particulier pour réduire les inégalités de santé. La source de cette inefficacité résiderait dans une insuffisante prise en compte de l'hétérogénéité de la position sociale. En effet, intrinsèquement multidimensionnelle, elle est appréhendée par divers indicateurs (éducation, revenu, profession, etc.) qui ne sont pas réductibles et peuvent interagir. Il importe donc à une santé publique de précision de prendre en compte la variabilité de ces influences pour améliorer la pertinence des interventions sur les inégalités de santé. La précision est alors envisagée comme une approche plus fine de la complexité du social.

De la médecine de précision, on retrouve le souci de l'hétérogénéité pour mieux cibler les sous-groupes qui ont besoin d'une intervention et atteindre ainsi une meilleure efficacité. Mais ici le but est de mieux comprendre les mécanismes par lesquels les inégalités se structurent. Et surtout, ces sous-groupes ne sont pas alors définis par la somme des positions sociales des individus, mais à partir du contexte social dans lequel ils sont incorporés. Par ailleurs, l'importance accordée aux données massives et à leur rôle prioritaire sur la théorisation, qui caractérise souvent l'approche de précision, est ici relativisée. Le rôle des théories sociales à partir desquelles la position sociale et les différenciations produites sont abordées est central. Dès lors, est-il encore pertinent de parler de « précision » et cela ne risque-t-il pas de prêter à confusion ?

Limites de la recherche de précision en santé publique

Bien que cette conception de la santé publique de précision soit séduisante, justifie-t-elle une refondation de la santé publique ? Mieux comprendre la complexité des mécanismes par lesquels les déterminants sociaux influent sur la santé s'inscrit dans la continuité de recherches en santé publique qui intègrent des approches systémiques. Mais surtout, cette conception se démarque de toute la littérature sur la santé publique de précision. Par suite, il semble qu'il y ait plus d'inconvénients que de bénéfices à conserver ce vocabulaire de la précision, associé aux notions d'individualisation, de stratégie ciblée et à la génomique.

Pour finir, il est important d'interroger la valeur et la pertinence d'une priorité donnée à la recherche de plus de précision pour améliorer la santé publique. L'approche de précision véhicule d'une part l'idée d'un privilège donné à la mesure quantitative, elle-même associée à celle de la supériorité des sciences naturelles sur les sciences humaines et, d'autre part, l'illusion que l'on pourrait se rapprocher d'une forme de certitude. Or l'intérêt de la santé publique ne tient-il pas à ce qu'elle est très pluridisciplinaire et qu'elle complète la biomédecine par des approches qualitatives de sciences humaines ? Surtout cette insistance sur la précision court le risque de laisser de côté les facteurs qui ne peuvent être ainsi mesurés et pour lesquels pourtant une intervention est efficace. Rose souligne que, dans le champ de la santé publique, rien ne peut jamais être certain et que la certitude ne saurait être un prérequis pour l'action. Si plus de précision c'est être avant tout attentif à l'individu, à l'hétérogénéité interindividuelle et à l'exactitude des résultats, et si c'est défendre la priorité de la connaissance sur l'action, les fondements de la santé publique sont remis en cause. Le propos n'est pas ici de faire l'apologie de l'imprécision en santé publique ni de défendre l'idée que la santé publique doit toujours privilégier le collectif sur l'individuel. Toutefois, se centrer sur l'objectif de précision est porteur d'implicites qui peuvent nuire à l'âme même de la santé publique, si on considère que la santé de la population dont elle s'occupe n'est pas réductible à la simple somme des santés des individus.

Données massives et santé publique : entre redéfinitions et ruptures normatives

Margo Bernelin
Chargée de recherche CNRS, UMR 6297 CNRS, université de Nantes

La stratégie nationale de santé 2018-2022, socle politique des projets de lois en matière de santé pour le quinquennat en cours, énonce que « *le développement des innovations numériques, technologiques et organisationnelles en santé est un enjeu clé pour l'évolution des pratiques professionnelles, l'accélération du virage ambulatoire, la qualité du suivi des patients chroniques ou le partage de l'information par les acteurs du système de santé et du médico-social*[1] ». Le numérique et le traitement des données ont ainsi pris une place centrale au sein des dispositifs juridiques mis en place depuis 2018. À cet égard, il est certain que la combinaison « numérique/données » permet des avancées importantes en matière de santé publique, qu'il s'agisse du suivi de la progression d'une épidémie, de la détection de facteurs de risques associés à des pathologies ou encore de la mise en lumière d'effets secondaires de médicaments. Le recours aux données massives en santé publique doit révolutionner la matière en favorisant des gains de temps dans la recherche, en faisant émerger de nouveaux champs de recherche et de nouvelles cibles de prévention. De tels bénéfices reposent sur la collecte et la conservation des données (des vivants mais également des défunts), dans des volumes sans précédents, par des opérateurs privés ou publics. Ils nécessitent le plus souvent une mise en commun des bases de données ainsi que leur exploitation par des algorithmes. Tandis que le champ de la santé publique se trouve, à tout le moins, transformé par ces nouvelles opportunités, qu'en est-il du droit en la matière ?

Une redéfinition des objectifs du droit de la santé publique

À l'image des transformations décrites, le droit se trouve plus que jamais orienté vers la collecte des données, les lois dernièrement votées en faisant une priorité nouvelle. En effet, jusqu'à présent, le traitement des données était pensé quasi uniquement par le prisme des données personnelles et encadré par la loi Informatique et libertés (LIL, 1978). Or, le Code de la santé publique (CSP) s'ouvre aujourd'hui à des dispositifs juridiques particuliers visant le traitement des données, qu'elles soient personnelles ou non, telles que les données d'activité des hôpitaux et les données scientifiques. Ainsi, alors qu'autrefois la question du traitement des données était rattachée aux nécessités de dénombrement, puis de vigilance (appliqué aux médicaments, aux matériaux, à la traçabilité), la collecte des données sort de ces

1. Ministère des Solidarités et de la Santé, Stratégie nationale de santé, déc. 2017, p. 63.

cadres spécifiques pour prendre une place plus générale et centrale au sein du Code de la santé publique. Cette transformation est mise en évidence par la création en 2016 du Système national des données de santé (SNDS), dont la composition et le fonctionnement sont régis par le Code de la santé publique. Le SNDS regroupait, en 2016, l'accès à cinq bases de données dont celle de l'Assurance maladie sur le remboursement des soins et la base de données nationale sur les causes de décès. Selon les règles gouvernant l'accès aux données, les données doivent être au service exclusif de la santé.

En 2019, le législateur poursuit ce mouvement enclenché en réformant le SNDS pour accorder une place toujours plus importante à la collecte et à la conservation des données. Pierre angulaire de la réforme : l'élargissement du nombre des données disponibles au sein du SNDS. Ainsi, aux bases de données déjà accessibles viennent s'ajouter, entre autres, toutes les données collectées à l'occasion de soins remboursés par l'Assurance maladie. Par conséquent, mesures de taille, de poids ou des résultats d'examens cliniques sont autant de données pouvant remonter au SNDS car consignées par les professionnels de santé lors de soins remboursés par l'Assurance maladie.

La réforme de 2019 est aussi l'occasion d'entériner un projet pilote lancé quelques mois plutôt : le Health Data Hub, plateforme nationale des données de santé chargée, notamment, de mettre à disposition les données du SNDS élargi, mais aussi de financer des projets de recherche sur ces mêmes données. Cette réforme n'est pas anodine : la collecte des données en grand nombre n'est plus un simple moyen au service de la santé publique au sein du Code de la santé publique, mais devient une fin en soi. À cet égard, la réforme simplifie la création d'entrepôts de données de santé. Ces derniers, créés par des hôpitaux, visent le regroupement de données diverses, y compris des données de santé. Ici, les nouvelles règles autorisent l'aspiration des données du SNDS à des fins d'enrichissement de tels entrepôts. Par conséquent, la réforme insiste sur la collecte des données en grand nombre, leur duplication au sein de bases existantes pour des recherches futures, et cela sans préciser davantage les types de recherches pouvant justifier un accès aux données.

Cette place centrale accordée à la collecte des données par le droit occulte même les difficultés de terrain, laissant en suspens la question du financement de ces collectes et de la mise en œuvre d'une interopérabilité dans l'accès aux données. De même la question des droits des individus est facilement évacuée au profit d'une anonymisation/pseudonymisation des données, pourtant largement faillibles.

Un droit perméable aux acteurs privés

Sous l'impulsion des données massives, le droit de la santé publique offre aujourd'hui une place plus importante aux acteurs privés alors même que la santé publique a longtemps été réservée aux services publics. Le légis-

Le SNDS en 2016	Le SNDS en 2019
• Les données d'analyse de l'activité des hôpitaux, du programme de médicalisation, des systèmes d'information.	• Les données d'analyse de l'activité des hôpitaux, du programme de médicalisation, des systèmes d'information.
• Les données du Système national d'information inter-régimes de l'Assurance maladie.	• Les données du Système national d'information inter-régimes de l'Assurance maladie.
• Les données sur les causes de décès.	• Les données sur les causes de décès.
• Les données médicosociales des maisons départementales des personnes handicapées.	• Les données médicosociales des maisons départementales des personnes handicapées.
• Un échantillon représentatif des données de remboursement par bénéficiaire transmises par des organismes d'assurance maladie complémentaire et défini en concertation avec leurs représentants.	• Un échantillon représentatif des données de remboursement par bénéficiaire transmises par des organismes d'assurance maladie complémentaire et défini en concertation avec leurs représentants.
	• Les données destinées aux professionnels et organismes de santé recueillies à l'occasion des activités donnant lieu à la prise en charge des frais de santé en matière de maladie ou de maternité et à la prise en charge des prestations en cas d'accident de travail et de maladie professionnelle.
	• Les données relatives à la perte d'autonomie lorsqu'elles sont appariées avec les bases de données précédemment citées.
	• Les données à caractère personnel des enquêtes dans le domaine de la santé lorsqu'elles sont appariées avec les bases de données précédemment citées.
	• Les données recueillies lors des visites médicales et de dépistage obligatoires effectuées par les médecins et infirmiers de l'Éducation nationale.
	• Les données recueillies par les services de protection maternelle et infantile dans le cadre de leurs missions.
	• Les données de santé recueillies lors des visites d'information et de prévention auprès des travailleurs.

lateur avait ainsi largement entendu confier les missions de prévention, de surveillance de la population à des agences sanitaires publiques, à des réseaux publics ou encore à des organismes, bien que de droit privé, à but non lucratif tels que les mutuelles. Cependant, l'écosystème autour du traitement des données massives a conduit à faire émerger la possibilité d'associer un plus grand nombre d'acteurs privés à la santé publique. Cette attraction vers le secteur privé a été mise en évidence par les discussions parlementaires de 2019 sur la création de la plateforme nationale des données de santé. La question posée était la suivante : cette plateforme, au rôle stratégique de pilotage de la collecte et de l'accès aux données en santé, devait-elle prendre la forme d'une structure publique ou d'une structure privée ? Certains députés avaient ainsi proposé la création d'une société par actions simplifiées (SAS), laquelle aurait pu être majoritairement détenue par l'État, garant de la finalité de la plateforme et gage de confiance pour le public. Porteuse de souplesse, une SAS devait offrir une attractivité plus importante pour la plateforme, capable de conclure des partenariats avec le public ou le privé plus rapidement qu'une institution publique. Cependant, les discussions parlementaires ont mis en lumière la crainte d'une défiance du public vis-à-vis d'une SAS dans la gestion de données personnelles aussi sensibles que celles relatives à la santé. Partant, c'est bien une structure publique qui fut privilégiée avec la création d'un *groupement d'intérêt public*.

La question de la présence d'acteurs privés au sein du champ de la santé publique fut relancée en décembre 2019 avec l'hébergement informatique des données du SNDS. Extrêmement volumineuses, les bases associées nécessitent, pour être regroupées et interrogées, d'être conservées par des hébergeurs disposant de capacités de conservation et de traitement des données considérables. Le choix s'est alors porté sur l'entreprise américaine Microsoft pour héberger ces données. Les critiques n'ont alors pas tardé, faisant valoir que les données de santé des Français devraient, pour plus de protection, être hébergées par une entreprise française ou à tout le moins européenne. Ces critiques ont également avancé que le SNDS se trouverait dépendant d'une entreprise possédant une capacité commerciale écrasante. Pourtant un tel choix n'est pas étonnant dès lors que le droit ne s'y oppose pas. En effet, le droit de la santé publique, qui encadre strictement l'hébergement des données de santé, imposant aux hébergeurs d'obtenir une certification, n'interdit pas qu'une entreprise privée, y compris étrangère, propose de tels services. Ainsi, du fait de l'introduction des données massives en santé publique, le code du même nom s'ouvre peu à peu à la présence d'acteurs privés, lesquels apparaissent comme des renforts, plus ou moins bien accueillis, de la puissance publique.

Pour conclure, la rencontre entre « données massives » et « santé publique » marque un véritable tournant pour le droit, dont les objectifs s'avèrent aménagés et réorientés vers la collecte des données et dont les acteurs se trouvent diversifiés.

Nicolas Savy
Maître de conférences à l'université Toulouse III, Institut de mathématiques de Toulouse, UMR 5219
Anne Mayère
Professeure à l'université Toulouse III, laboratoire Certop, UMR 5044, directrice adjointe de l'Iferiss
Anja Martin-Scholz
Maître de conférences à l'université Toulouse III, laboratoire Certop, UMR 5044
François Lambotte
Professeur à l'université catholique de Louvain, Institut langage et communication

La fabrique des données à l'épreuve des programmes de *big data*

Nous proposons d'interroger ici la fabrique de données dans le contexte du *big data*, en prenant exemple auprès de B. Latour [37], qui a investigué la fabrique du droit en observant la façon dont il se construit. C'est posé que, s'agissant des données, la question des finalités, du « pour quoi », n'est pas dissociable du « comment ». Les discussions autour des « données massives » laissent entendre l'avènement d'un nouveau régime de « fabrique des données » qui viendrait amplifier les précédents du fait d'évolutions techniques. Qu'est-ce qui différencie les données de santé, telles que produites selon les standards de recherche, des « données massives », qui sont au cœur de projets conséquents (Health Data Hub) ? Nous verrons que ces fabriques se différencient quant à la spécification des données, aux logiques de traitement, et à la question des « biais » inhérents à toute fabrique de données, nécessairement partielle et éventuellement partiale.

Avant d'aller plus en avant, évoquons la notion de *big data* et la différence avec le terme de données massives au cœur de ce dossier. Le premier V, volumétrie, de la règle des 5V couramment employée pour parler du *big data*, laisse entendre que les données massives y seraient incluses. Encore faut-il s'accorder sur la notion de « massives ». Selon G. Saporta [51], massive est entendue comme « trop gros pour entrer dans la machine », nécessitant une adaptation (si possible) des techniques et des modélisations statistiques usuelles. Par exemple, le volume d'information est tel qu'il est impossible de mettre en place, en une seule étape, une régression logistique. Cette technique, couramment employée en statistique, vise, pour une variable binaire (hypothèse 1 = malade/hypothèse 0 = non malade), à déterminer l'influence d'un ensemble de facteurs (l'âge, le sexe, le poids…) sur la probabilité d'observer l'hypothèse 1. Comme l'ont étudié les sociologues des sciences et

Les références entre crochets renvoient à la Bibliographie générale p. 57.

des techniques [9, 26], il n'est pas d'appellation « plus exacte » que d'autres, mais certaines qui s'imposent plus ou moins durablement sur ce terrain de luttes que constitue toute innovation éventuelle, traversé par des enjeux tant politiques, sociaux, qu'économiques et techniques.

« Stratégie pour comprendre » *versus* « stratégie pour prévoir »

Généralement, les données une fois rassemblées sont traitées au moyen de considérations statistiques. Sorti des aspects purement descriptifs, tout raisonnement statistique consiste en une confrontation des données observées sur un échantillon à une modélisation de la population dont l'échantillon est issu. On parle de modèle génératif. L. Breiman [7] distingue deux types de stratégies : la « stratégie pour comprendre » et la « stratégie pour prévoir ».

La « stratégie pour comprendre » repose sur un ensemble d'hypothèses faites sur le modèle génératif. Par exemple, on peut poser l'hypothèse d'une mortalité également distribuée entre les classes sociales ; le traitement statistique va permettre de distinguer s'il existe une différence significative de la mortalité entre les classes sociales. La « stratégie pour comprendre » consiste alors à inférer les paramètres dudit modèle à partir des données disponibles ; ainsi pourra-t-on inférer que les ouvriers présentent un risque plus élevé de mortalité précoce que les cadres supérieurs, avec des moyennes d'âge de décès distinctes. Ce type de modèle doit inclure un nombre raisonnable de variables précisées *ex ante*. La découverte d'un important « reste à expliquer » peut amener dans un second temps à revoir la spécification des variables retenues.

La « stratégie pour prévoir », quant à elle, ne cherche pas à spécifier *ex ante* le modèle génératif. Il s'agit d'une approche souvent algorithmique dont l'unique préoccupation est la précision de la prévision, c'est-à-dire la faculté de l'algorithme à retrouver la valeur « vraie » de la valeur à prévoir (à savoir, la valeur de l'objet étudié si sa mesure était possible de façon exhaustive et sans erreur). Reprenons l'exemple de la prédiction d'une maladie (oui/non) à partir d'un ensemble de variables explicatives. Une technique de *machine learning* fournira, pour un individu donné, une réponse de type maladie présente ou absente et ce sans passer par l'estimation de paramètres associés à chaque variable explicative. On est en présence d'un modèle dit « boîte noire » qui fournit – sans expliquer – une prédiction de la variable à expliquer (de façon souvent très performante d'ailleurs).

Il est à noter que ces deux approches ne sont pas forcément à mettre en opposition. Comprendre peut permettre de prévoir et prévoir peut fournir des outils de compréhension, ou du moins peut faire émerger des hypothèses pour comprendre. Cependant, la stratégie de recueil de données, les logiques de traitement ainsi que la nature des résultats obtenus sont différentes. Dès lors ces deux « fabriques » sont distinctes et non inclusives, au sens où une fabrique de données pour comprendre ne peut être une « simple extraction » d'une fabrique de données massives pour prévoir.

La fabrique des volumes de données : des logiques différenciées

L'approche « pour comprendre » est en particulier celle de la médecine fondée sur la preuve (*evidence-based medicine*). Cette démarche, devenue le standard en recherche en santé, est établie sur une production de données protocolarisée. Ce protocole recense notamment la volumétrie visée et le contour des données à recueillir. La logique consiste à travailler avec un ensemble délimité de données précisément caractérisées et tracées, voire certifiées. C'est donc la qualité des données, et ce qu'elle permet comme montée en généralité, qui est priorisée. Il s'agit de spécifier un ensemble de critères en amont de l'investigation de façon notamment à caractériser la population étudiée, les valeurs investiguées, et s'assurer que les mesures sont susceptibles d'en respecter les caractéristiques. Cela s'inscrit dans le paradigme sous-jacent des statistiques inférentielles (ensemble de méthodes consistant à généraliser à une population des conclusions tirées à partir des données d'un échantillon) afin de s'assurer que l'échantillon retenu peut apporter une connaissance fiable au regard de la question étudiée. Le protocole est établi en fonction du niveau de preuve visé. La finalité de la fabrique est une explication *causale* au phénomène étudié, c'est-à-dire s'il existe un (ou des) événements qui ont pour conséquence le phénomène étudié. La base de données qui vient l'alimenter est souvent, pour des questions scientifiques, organisationnelles et logistiques, de taille prédéfinie et en cela délimitée. Il s'agit notamment des cohortes, qui suivent un ensemble d'individus selon des critères et sur des variables précises dans une durée de plusieurs années ou décennies, ou des registres, qui recensent l'ensemble des patients atteints d'une pathologie sur un territoire donné.

La stratégie « pour prévoir » est associée à l'assemblage du plus grand nombre de données possible. Le contexte est donc celui des données massives. Ces données massives sont travaillées avec des techniques de *machine learning*. Dans cette quête de l'algorithme le plus performant en termes de prévision, il est courant d'utiliser un large spectre d'algorithmes, voire des combinaisons d'algorithmes. Les résultats s'expriment usuellement en termes de facteurs influençant la prévision. Ils mettent en lumière des *corrélations*, c'est-à-dire le degré de liaison entre deux variables, dont l'explication est très rarement causale. Les exemples de situation où une corrélation est prouvée sans qu'il y ait de relation causale sont légion. La performance attribuée à ces outils est largement liée à la quantité et à la fréquence des données ; le postulat étant que de l'accumulation de données peuvent se dégager des corrélations susceptibles de guider l'action (des pouvoirs publics, ou d'organisations marchandes).

Une formulation très différenciée des « biais de sélection » et de leur maîtrise

Quelle que soit la stratégie, on est donc potentiellement en présence d'un biais dans la conclusion lié à la sélection des patients. La « stratégie pour comprendre » est construite à partir de données protocolisées, les résultats obtenus sont conditionnels à ce protocole. Le « modèle pour prévoir » est construit à partir de données massives et les résultats sont donc eux aussi conditionnels aux données rassemblées. Il existe cependant des différences notables quant à la connaissance de ces biais et à leur prise en compte dans l'analyse et les résultats.

Dans l'approche « pour comprendre », ce biais, grâce aux définitions précises des critères d'inclusion/non-inclusion et aux éléments de design expérimental (par exemple pour une étude randomisée), est au moins réfléchi, au mieux pris en compte. La réflexion méthodologique est donc tout à fait majeure dans cette approche.

Dans l'approche « pour prévoir » avec des données massives, la performance des outils, basés sur des applications d'intelligence artificielle, repose massivement sur la volumétrie des données, qui laisse sous-entendre l'exhaustivité. Or, l'exhaustivité est très compliquée et coûteuse à obtenir, précisément parce qu'une partie de la population ou des critères recensés échappent aux formes de traçage les plus simples à systématiser. Par ailleurs, le risque est aussi de « chercher sous le lampadaire » en postulant que son « écologie » est significative de tout l'espace entre deux lampadaires. Dans cette approche, la question des biais de sélection reste omniprésente mais est plus sournoise. En effet, les modèles d'apprentissage s'enrichissent des données observées et fournissent pour celles non observées une prévision au mieux peu fiable. Ils sont donc dépendants de la qualité des données, or la réflexion méthodologique est repositionnée éventuellement en aval, voire écartée. De plus, ces masses de données sont souvent le fruit d'un processus d'assemblage de bases de données avec des échelles de mesure et des référentiels différents. On peut légitimement se poser la question de l'impact du contexte situé de production, et des hypothèses implicites relatives à la possibilité de combiner des données d'origines diverses dans ces agrégats de données.

Conclusion

Le succès grandissant des approches par « données massives » peut s'expliquer par plusieurs phénomènes. Tout d'abord, la capacité de stockage de la donnée est reléguée au second plan suite à des évolutions technologiques. Ensuite la performance des algorithmes de *machine learning* peut fonder l'espoir de résultats significatifs à venir. De plus, les discriminants usuels des « modèles pour comprendre » établis sur la statistique inférentielle deviennent vite inefficaces face à des échantillons massifs. Enfin, les promoteurs des *big data* n'ont de cesse d'annoncer de nouvelles formes de valorisation marchande de ces « gisements de données ».

Les discours portés par les promoteurs des *big data* convergent avec des logiques d'injonction à la performance et à l'efficacité dans les politiques publiques, en proposant de croiser de très nombreuses données afin d'identifier où il serait nécessaire de porter l'attention. Or les principes de traitement de données s'avèrent différenciés selon qu'il s'agit de logiques marchandes, qui peuvent se suffire d'une définition floue de la population de référence et raisonner sur des corrélations, alors que les démarches de recherche en santé publique et en épidémiologie vont requérir une relation de causalité démontrée sur une population qualifiée. Autrement dit, alors que les consommateurs d'Amazon peuvent se satisfaire de la piètre fiabilité de son algorithme de suggestion, il n'en est rien pour des décideurs en matière de politique publique s'agissant du dépistage d'une maladie.

L'exploration des différences entre les deux approches met en évidence leur éventuelle complémentarité. Cependant il s'agit de deux types de « fabriques de données » aux logiques et modalités de construction différenciées. Or les discours promotionnels des données massives laissent entendre que de grands assemblages de données pourraient servir les deux approches.

De tels discours nous semblent traduire une méconnaissance de ce que recouvre le « comment » de ces fabriques. La question est de savoir s'il importe de comprendre, de dégager des causalités vérifiées, pour guider l'action publique, ou si la priorité est au pilotage fondé sur des corrélations dont le caractère significatif et représentatif n'est pas nécessairement maîtrisé. Une question liée est le risque d'invisibilisation de toute une partie de la population, telle celle qui est concernée par les inégalités sociales de santé. Dans cette logique, certains acteurs pourraient privilégier l'accès aisé à des masses de données, comme les données collectées par les montres connectées, sans se poser la question relative à la représentativité et au profil sociodémographique des personnes en mesure d'acquérir ce type de technologie.

La pandémie de Covid-19 a remis en avant les questions liées aux données de santé et la nécessité de mieux connaître leur « fabrique » pour identifier les enjeux. Ce sont ainsi différentes configurations de données, différents principes de modélisation, des questions de biais et de maîtrise de ces biais qui sont à l'œuvre derrière la désignation homogénéisante des « données de santé ».

Big data et prédiction

La génétique peut être considérée comme le fer de lance de la « médecine des données » et pourrait apporter d'importants éléments en matière de santé publique. Mais quelle information transmettre au patient ? Avec quel encadrement juridique ?

Médecine génomique : vers une médecine prédictive ?

Sandra Mercier
Professeure des universités, praticienne hospitalière, service de génétique médicale, CHU de Nantes

Les références entre crochets renvoient à la Bibliographie générale p. 57.

Remerciements pour sa relecture à Guillaume Durand, maître de conférences en philosophie, département de philosophie, UFR Lettres et langages, Centre atlantique de philosophie (Caphi, EA 7463), université de Nantes, MSH Ange-Guépin (USR 3491, CNRS), Nantes.

« *Génétique, jusqu'où a-t-on le droit d'aller ?* » Cette question posée par le Comité consultatif national d'éthique (CCNE) en 2016 est de plus en plus d'actualité [16].

L'avènement du séquençage de nouvelle génération

Nous assistons en effet à une révolution technologique en génétique moléculaire avec l'avènement du séquençage haut débit ou séquençage de nouvelle génération (NGS, Next Generation Sequencing). Alors que l'analyse d'un seul gène pouvait prendre plusieurs mois (voire années) avec la technique de référence dite Sanger, le NGS permet aujourd'hui de séquencer, en une seule réaction, un panel de gènes (dizaines ou centaines de gènes) jusqu'à l'exome (région codante de nos 21 000 gènes, soit 1,5 % de notre ADN), voire l'ensemble de notre génome (100 % de notre ADN, 3×10^9 paires de bases).

L'analyse bio-informatique des données générées est fondamentale et complexe. Nous sommes en effet capables aujourd'hui de séquencer l'ensemble du génome d'un individu à des coûts qui n'ont jamais été aussi bas (moins de 1 000 dollars, analyse des résultats en sus). Les progrès liés au séquençage haut débit apportent un meilleur rendement diagnostique, fondamental pour les patients, en particulier porteurs d'une maladie rare, mais montrent également une grande complexité dans l'interprétation du génome humain. En effet, nous identifions beaucoup de variants de signification inconnue. L'interprétation des résultats reste délicate ; elle nécessite une confrontation clinico-biologique entre le médecin prescripteur et le généticien moléculariste et peut évoluer avec le temps en fonction des connaissances. La corrélation entre le génotype et le phénotype, ainsi que la prise en compte des antécédents familiaux connus chez le patient sont indispensables pour établir le bon diagnostic moléculaire.

Vers une médecine personnalisée

Le plan France médecine génomique 2025 prévoit l'analyse génomique de milliers de patients sur des plateformes nationales. Deux premières plateformes ont débuté les analyses, dans les régions Ile-de-France (SeqOIA) et Auvergne-Rhône-Alpes (Auragen). Nous aurons accès aux données des génomes de plus en plus de patients d'ici les prochaines années.

La médecine 4 P (prédictive, préventive, personnalisée, participative) vise à donner le bon traitement au bon patient au bon moment. Les progrès génétiques liés au NGS et la plus grande accessibilité de ces tests

permettent l'identification de « nouveaux gènes » ou « nouveaux variants » pour les patients porteurs de maladies rares et contribuent sensiblement à améliorer le rendement diagnostique pour ces patients. Il s'agit bien d'une médecine personnalisée où le test génétique va permettre de porter un diagnostic précis pour le patient avec une prise en charge et un conseil génétique adaptés. On entend aussi donner beaucoup de poids à la génétique pour une médecine « prédictive » et « préventive ». Nous sommes maintenant capables d'identifier des anomalies ou des prédispositions génétiques à certaines pathologies dans le cadre d'un diagnostic présymptomatique : par exemple, des prédispositions à certains cancers, certaines maladies neuromusculaires, cardiaques ou neurodégénératives. Dans les familles où la pathologie est déjà connue, il est possible de renseigner les apparentés qui le souhaitent sur leur statut génétique et de déterminer s'ils sont porteurs ou non de l'anomalie génétique. Selon les cas, un conseil génétique, une surveillance ou des mesures prophylactiques sont proposés. Le résultat d'un test génétique est rarement « prédictif » car, finalement, peu de pathologies génétiques recherchées en diagnostic présymptomatique ont une pénétrance complète ; il s'agit plus souvent d'une prédisposition : la personne ne va pas obligatoirement développer des signes de la maladie durant sa vie et, si elle en développe, il est souvent difficile de déterminer précisément l'âge de survenue de la maladie, son évolution et le pronostic. La participation du génotype dans cette médecine « prédictive » ne doit pas être considérée isolément sans prendre en compte l'individu dans sa globalité. En effet, le mode de vie, l'alimentation, l'environnement ont des répercussions sur l'épigénétique, qui régule l'expression de nos gènes, et sont donc des éléments déterminants dans l'expression de ces pathologies. Il serait réducteur de considérer que le phénotype d'un individu est déterminé essentiellement par son génotype. Cela pourrait conduire à des dérives, comme l'illustre notamment le film de science-fiction *Bienvenue à Gattaca*, réalisé par Andrew Niccol en 1997. Ces problématiques ne paraissent plus si futuristes aujourd'hui. Les avis n° 124, n° 129 et n° 130 du CCNE exposent parfaitement l'ensemble de ces problématiques liées à l'évolution des tests génétiques et mettent en garde contre des dérives potentielles, notamment des risques de discrimination [16].

Intérêt médical, risques de dérive et précautions à prendre

Par l'analyse de l'exome ou du génome d'un patient, nous avons potentiellement accès à l'identification de variants dans des gènes qui ne correspondent pas à la recherche diagnostique initiale. Il s'agit d'une donnée dite « secondaire ». L'ACMG (American College of Medical Genetics) a donné des recommandations concernant ces données secondaires et a établi une liste de 66 gènes à analyser si le patient le souhaite [34]. Ces gènes sont dits « actionnables », c'est-à-dire qu'une prévention et/ou un traitement peuvent être proposés si un variant pathogène est identifié. Il s'agit principalement de gènes impliqués dans les prédispositions aux cancers et à certaines pathologies cardiovasculaires (troubles du rythme cardiaque, cardiomyopathie, dissections artérielles). Cette question est très débattue en Europe et en France [16]. En effet, est-ce que le rendu de cette information sera bénéfique pour une famille déjà fragilisée par la maladie ? Sans doute mieux vaut-il prévenir que guérir, mais il ne faudrait pas non plus tomber dans l'excès inverse.

Nous connaissons le risque potentiel de telles prédispositions dans les familles où la maladie s'exprime, mais nous n'avons pas encore de données fiables sur un éventuel « surrisque » que conférerait la présence d'un même variant dans une famille sans antécédents familiaux. Une étude portant sur une cohorte de personnes âgées (plus de 80 ans) et sans maladies chroniques n'a pas retrouvé de différence significative de fréquence des variants dans la liste de ces gènes « actionnables » alors que l'on se serait attendu à une fréquence beaucoup plus faible [24]. Il existe des facteurs modificateurs, éventuellement protecteurs (génétiques, environnementaux…), qui vont interférer et peuvent modifier sensiblement le risque en population générale. À ce jour, la recherche des données secondaires n'est pas proposée dans le cadre diagnostique en France, mais elle l'est dans le cadre de la recherche (exemple du protocole Defidiag). Notre but en tant que médecin généticien est avant tout d'aider les patients et de ne pas être délétères dans l'annonce d'une prédisposition si celle-ci n'est pas avérée. Dans la balance bénéfices/risques, il faut prendre en compte le caractère anxiogène de ces données pour les patients. Le risque serait de générer à tort une anxiété en rendant toute la population « potentiellement » malade et de prévoir une prise en charge inadaptée.

Si l'on considère à l'avenir que cette recherche de données secondaires peut être proposée aux patients – donc que l'on estimerait un bénéfice médical à connaître ces données –, faudrait-il la proposer en population générale et pas seulement aux personnes malades par souci d'équité ? Dans ce cas, il faudrait pouvoir proposer une prise en charge adaptée, c'est-à-dire calquée sur le diagnostic présymptomatique comprenant un entretien psychologique et un délai de réflexion en amont du test génétique afin de préparer au mieux la personne à recevoir ses résultats. On touche là à une problématique plus large de santé publique concernant l'accessibilité, la faisabilité à grande échelle et le coût médico-économique de ces tests et de leurs conséquences en population générale.

Par ailleurs, on prend en compte également dans les données secondaires les tests préconceptionnels, à savoir la recherche pour un couple du risque de transmettre une maladie à sa descendance pour laquelle les membres de ce couple seraient porteurs sains (hétérozygotes). Le CCNE a proposé dans sa contribution à la révision de la loi de bioéthique (avis n° 129,

septembre 2018) qu'un test soit prescrit à toutes les personnes en âge de procréer qui le souhaitent, après une consultation de génétique pour les maladies héréditaires monogéniques graves. Le CCNE préconise aussi d'examiner les possibilités de l'extension du dépistage génétique à la population générale [16]. La révision de la loi de bioéthique en cours ne semble pas suivre ses propositions, mais ces tests préconceptionnels sont mis en place en Belgique et dans d'autres pays.

Enfin, les tests génétiques en libre accès se multiplient. Ces tests directement proposés aux consommateurs sont interdits par la législation française, mais sont autorisés dans d'autres pays, notamment aux États-Unis, et accessibles sur Internet. Le produit est payé en ligne et l'analyse génétique se fait à partir d'un prélèvement salivaire. Les résultats sont transmis sous forme d'un fichier numérique. La société 23andMe, par exemple, propose des analyses génétiques portant sur la prédisposition au cancer du sein et de l'ovaire, le risque de maladie d'Alzheimer, de Parkinson, etc. La FDA (Food and Drug Administration) avait interdit la commercialisation de ces tests en 2013 à la fois pour des raisons réglementaires, mais également en raison du manque de fiabilité et d'exhaustivité des tests, sans oublier les conséquences délétères potentielles de ces résultats controversés sur les individus. Ces tests sont de nouveau autorisés et les mises en garde sont un peu plus explicites, comme la part liée au mode de vie, à l'environnement et à d'autres facteurs génétiques. Mais nous nous interrogeons sur l'information qui sera comprise à partir de ces éléments bruts, sans contextualisation par rapport aux antécédents personnels et familiaux, et sur l'autonomie de la décision de la personne qui commande de telles analyses depuis son ordinateur [22]. Par exemple, cette personne va-t-elle comprendre que, dans le cas de la prédisposition au cancer du sein et de l'ovaire, seulement trois variations sont recherchées sur plus de trois mille variations pathogènes connues dans les gènes BRCA1 et BRCA2 et donc que l'analyse, ainsi que les résultats, sont très partiels ? La recherche de cette prédisposition en consultation d'oncogénétique passe par une information exhaustive, délivrée par un médecin spécialiste, un séquençage complet de ces gènes et un rendu des résultats en consultation ainsi qu'un accompagnement psychologique. Plus généralement, la prescription et le rendu de ces tests nécessitent un accompagnement global de la personne concernée par des professionnels spécialisés dans le domaine de la génétique. Une mauvaise interprétation des résultats a des conséquences désastreuses pour les patients et leur projet de vie. C'est le cas d'environ 40 % des tests en libre accès dont les résultats correspondent à des faux positifs liés à des artefacts techniques ou une mauvaise interprétation [53].

En conclusion, un encadrement éthique et juridique des pratiques en génétique doit être renouvelé régulièrement en tenant compte de l'évolution technologique rapide dans ce domaine, mais aussi de l'accessibilité des analyses du fait de la mondialisation. Ces questions nous concernent tous et nous amènent à repenser la société à laquelle nous aspirons pour demain. Il serait de toute façon illusoire de penser que la médecine génomique permettra de prédire précisément la santé et l'espérance de vie d'une personne alors que, dans la grande majorité des cas, celles-ci dépendent essentiellement de déterminants socioéconomiques et non génétiques.

Données génétiques massives : quelles règles juridiques et éthiques pour leur usage en santé publique ?

Emmanuelle Rial-Sebbag
Directrice de recherche Inserm, UMR 1027, Inserm université Toulouse III Paul Sabatier

Les nouvelles techniques génomiques (par exemple le séquençage complet du génome ou de l'exome) permettent dorénavant de générer des données massives pour un usage dans le soin (notamment grâce aux plateformes financées par le plan France médecine génomique 2025) mais également dans la recherche. Ce nouveau contexte technique soulève des enjeux très différents, que ce soit en termes d'acteurs (État, patients, population générale) ou d'obligations juridiques, mais également au regard de la production des données (leur qualité et leur fiabilité doivent être maximales afin d'atteindre un niveau de preuve suffisant pour concrétiser leur usage). Un niveau de complexité supplémentaire est atteint quand il s'agit d'utiliser les données génétiques massives dans un contexte de santé publique. En effet, notre droit met à la charge de l'État un certain nombre de possibilités pour déployer des actions de santé publique, sans toutefois faire explicitement référence à la génétique. Ainsi plusieurs concepts coexistent, souvent considérés comme proches mais fondamentalement différents (médecine prédictive, médecine préventive ou dépistage). Juridiquement et éthiquement ces approches sont basées sur des dimensions visant soit à développer des stratégies médicales individuelles

(tests génétiques pour le diagnostic), soit des stratégies, plus collectives que ces dernières, ciblant des groupes à risque (dépistage de pathologies dans des populations ciblées) ou la population générale (dépistage néonatal). Nous proposons de clarifier ces différentes approches de la production de données génétiques massives dans leurs dimensions juridiques en insistant sur les spécificités de leur usage en santé publique.

Les règles juridiques communes de la production et de l'utilisation des données génétiques

Quel que soit le contexte d'usage, la génération de données génétiques massives répond aux mêmes exigences que ce soit pour les droits des patients ou pour les droits attachés à la protection des données.

Tests génétiques et droits des patients

Les tests génétiques sont le support des données génétiques. En ce sens, le droit français vise essentiellement à protéger l'individu (et par extension sa famille, dont certains membres peuvent avoir un intérêt médical à être informés de l'existence de maladies génétiques à caractère héréditaire) face aux potentiels mésusages de ses informations génétiques. Ainsi, le Code civil et le Code de la santé publique dotent la personne dont sont issues les données génétiques de droits fondamentaux à l'occasion de la réalisation d'un test génétique (consentement écrit, information complète, conseil génétique, communication avec la famille, etc.) et cela à tous les moments de la chaîne de réalisation du test (prescription, rendu des résultats). Ce cadre juridique strict est construit au prisme des maladies monogéniques héréditaires[1] et tend, précisément, à considérer la protection accordée par le droit comme une prérogative individuelle (voire familiale) pour un diagnostic spécifique mais sans véritable référence à sa dimension collective. Ainsi, au regard des droits des patients, le Code de la santé publique (CSP) n'a pas encore totalement incorporé la dimension « de masse », qui, on le sait, vient pourtant bouleverser les notions d'information et de consentement – alors que toutes les données ne sont pas interprétables ou que les scientifiques ne connaissent pas totalement leur usage futur pour le soin ou pour la recherche (informer sur quoi ? comment ?) –, de retour de résultats (quels résultats ? communiqués à qui ?), ou encore de protection de la vie privée (données personnelles, données anonymes)[2]. En sus de la protection offerte par le droit des patients, le droit des données doit également être considéré.

La protection des données génétiques

Par essence, le concept de données génétiques est protéiforme sur le plan scientifique. En effet, celles-ci peuvent être classées soit en fonction de leur nature (un gène, plusieurs gènes, séquençage complet) soit en fonction de leur usage (médical ou préventif dans un contexte familial par exemple ; recherche des empreintes génétiques pour un usage judiciaire). En droit, la doctrine tend également à différencier la donnée génétique (donnée brute) de l'information génétique (donnée interprétée), impliquant dès lors l'application de cadres juridiques spécifiques. Toutefois, la science et le droit partagent le sentiment que la donnée génétique est particulière, car disant quelque chose de l'intime de la personne testée et potentiellement de sa famille, voire de son identité (ou sa possible identification). De ce fait, tant la réglementation européenne (Règlement général pour la protection des données personnelles, RGPD, (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016, entré en vigueur en mai 2018) que le droit français (loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés, LIL, modifiée) couvrent, par des dispositions particulières, la production et l'usage des données génétiques.

L'apport premier du RGPD n'est pas de doter les données génétiques d'un régime particulier (elles tombent sous le même régime juridique que les données de santé), mais plutôt de leur donner une définition. Ainsi, au sens du RGPD, les données, quand elles sont qualifiées de données personnelles (permettant une identification directe ou indirecte des personnes sources), sont des « *données à caractère personnel relatives aux caractéristiques génétiques héréditaires ou acquises d'une personne physique qui donnent des informations uniques sur la physiologie ou l'état de santé de cette personne physique et qui résultent, notamment, d'une analyse d'un échantillon biologique de la personne physique en question* » (RGPD, art. 4 al. 13). L'art. 9 du RGPD classe ces données parmi les données sensibles (au même titre donc que les données de santé), emportant par principe l'interdiction de leur traitement. Le RGPD pose, bien évidemment, un certain nombre d'exceptions à ce régime permettant de légitimer le recueil et l'usage des données génétiques pour la santé si le consentement de la personne est acquis, ou si le traitement est nécessaire pour des motifs d'intérêt public de recherche scientifique ou encore si les données sont nécessaires pour un traitement médical (y compris la prévention), une prise en charge sanitaire ou sociale, la gestion des systèmes et des services de soins de santé ou de protection sociale. Dans ce dernier cas (usage médical), une condition supplémentaire est requise puisque les personnes réalisant le traitement devront être soumises à une obligation de secret professionnel[3].

La loi française a incorporé l'ensemble de ces dispositions tout en rappelant spécifiquement que « *dans le cas où la recherche nécessite l'examen des caractéristiques génétiques, le consentement éclairé et exprès des per-*

1. Cela devrait évoluer dans la prochaine loi de bioéthique, qui vise à mieux distinguer génétique constitutionnelle et somatique.

2. Comité consultatif national d'éthique, avis 124 : « Réflexion éthique sur l'évolution des tests génétiques liée au séquençage de l'ADN humain à très haut débit », janv. 2016

3. À ces conditions générales devra s'ajouter l'ensemble des conditions nécessaires à la licéité du traitement des données personnelles. RGPD, art. 6.

sonnes concernées doit être obtenu préalablement à la mise en œuvre du traitement de données » (loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés, art. 75 al. 1). Ce cadre, bien qu'encore discuté en doctrine, semble clair pour la réalisation des tests dans le contexte médical individuel mais reste à clarifier pour une application en santé publique.

Une nécessaire clarification du cadre juridique des usages en santé publique

Si la réalisation des tests génétique est régie en France de manière uniforme, de nombreux défis restent à relever concernant leur usage en santé publique. L'action du législateur s'est concentrée sur l'application de principes de protection individuelle et reste timide pour son application en population. En effet, le spectre de la possible mise en œuvre de politiques eugénistes, de tri voire de sélection, n'a jamais engagé réellement les pouvoirs publics sur le déploiement de la génétique au service de la santé publique. L'apparition des données génétiques massives et de leur potentiel usage bénéfique au profit des populations vient rebattre les cartes de leur place au sein des politiques de santé publique. Parmi les actions de santé publique, rappelons que les actions de prévention relèvent de la compétence de l'État à titre principal, l'article L. 1411-1 du CSP en donne une définition large : « *la prévention collective et individuelle, tout au long de la vie, des maladies et de la douleur, des traumatismes et des pertes d'autonomie* ». Ces actions de prévention se déclinent, selon les définitions données par l'OMS, en prévention primaire (agit en amont de la maladie), secondaire (agit à un stade précoce de son évolution) et tertiaire (agit sur les complications et les risques de récidive). La génétique n'est pas spécifiquement convoquée à l'appui de ces stratégies de santé publique mais en fait partie [2] tout en posant des questions de nature singulière notamment pour le dépistage. Le dépistage « *consiste à identifier de manière présomptive à l'aide de tests, d'examens ou d'autres techniques susceptibles d'une application rapide, les sujets atteints d'une maladie ou d'une anomalie passée jusque-là inaperçue* » [56].

Les références entre crochets renvoient à la Bibliographie générale p. 57.

Les tests génétiques peuvent donc être utilisés à des fins de dépistage mais le Conseil de l'Europe en rappelle les principales conditions de mise en œuvre. Ainsi, un programme de dépistage ne pourrait se déployer que conformément au respect d'un certain nombre de droits fondamentaux, notamment son approbation par une autorité compétente et après une évaluation indépendante portant sur son acceptabilité sur le plan éthique[4]. Le Conseil de l'Europe a établi à cet égard des conditions de validité pour le déploiement d'une politique de dépistage (2008) pouvant se résumer à son utilité, sa faisabilité et son acceptation sociale [6].

La France met déjà en œuvre des dépistages ciblés (dans le cadre du diagnostic préimplantatoire et prénatal) mais également du dépistage en population pour des indications ciblées. La nouvelle stratégie présentée par le Plan France médecine génomique 2025 suggère l'émergence d'un nouveau modèle en population, ou pour des groupes à risque, par l'établissement de dépistage précoce. Le Comité national consultatif d'éthique (CCNE) s'est également prononcé sur cet élargissement dans son avis 129 (contribution du CCNE à la révision de la loi de bioéthique 2018-2019, sept. 2018) en ciblant deux questions clés : l'extension des modalités de dépistage préconceptionnel (dans les groupes à risque pour d'autres maladies que celles déjà existantes au sein de la famille et extension à la population générale) et le dépistage génétique en population. Dans les deux cas, le CCNE n'est pas défavorable à l'extension de l'usage de la génétique dans ces champs d'action de la santé publique mais sous le respect de conditions strictes (accompagnement par des conseillers en génétique, information claire et consentement, validation par des instances *ad hoc*, liste de gènes et techniques employée, etc.). Sur le plan du droit, s'il n'existe pas d'impossibilité de réaliser ces tests à large échelle (dans le respect des règles existantes), il n'en demeure pas moins que des modifications législatives devront être effectuées pour incorporer dans notre droit les modalités nouvelles de leur réalisation. Il semblerait que cette orientation n'ait pas été totalement suivie par le législateur puisque le projet de loi de révision de la loi de bioéthique n'a pas donné suite à ces élargissements concernant le dépistage général en population ou le dépistage généralisé préconceptionnel (la commission spéciale bioéthique du Sénat avait proposé cette possibilité à titre expérimental, modification qui n'a pas été à ce jour retenue). Cependant, le Sénat a fait droit à la demande concernant le dépistage néonatal puisqu'il a proposé, en première lecture, que ces tests puissent être offerts aux titulaires de l'autorité parentale.

Malgré les évolutions législatives en cours, les données génétiques massives ont du mal à trouver un régime juridique cohérent, entre protection de l'individu et bénéfice pour la collectivité. Les nouvelles techniques, telle que l'intelligence artificielle, viennent s'inviter au débat promettant de passionnantes controverses.

4. Conseil de l'Europe. Protocole additionnel à la Convention sur les droits de l'homme et la biomédecine relatif aux tests génétiques à des fins médicales, STCE n° 203, Strasbourg, 27 novembre 2008 (art. 19).

Big data et action publique

L'usage des données massives dans les domaines tels que l'hygiène, l'épidémiologie, la vigilance sanitaire permet d'entrevoir de nouvelles pistes de recherche et de nouveaux moyens d'action pour éclairer les décisions de la puissance publique.

Du Système national des données de santé au Health Data Hub : mise en œuvre et évolution

Ève Jullien
Conseillère juridique, Sous-direction de l'observation de la santé et de l'assurance maladie, Drees

Ouvrir l'accès aux données de santé collectées par des organismes publics afin de tirer profit des potentialités qu'elles offrent est un enjeu sanitaire majeur. Cette ouverture vise à accroître les connaissances relatives à l'offre de soins et à la prise en charge médico-sociale à destination aussi bien des professionnels de santé que des usagers et des citoyens, à contribuer à la recherche et à l'innovation en santé et enfin à favoriser la veille et la sécurité sanitaires, des objectifs dont l'épidémie de SARS-CoV-2 a révélé le caractère essentiel.

La création du Système national des données de santé (SNDS), par la loi de modernisation de notre système de santé[1], constitue la réponse du législateur, sous l'impulsion de la ministre des Affaires sociales et de la Santé. La France s'est ainsi dotée d'un dispositif de pointe par rapport à de nombreux États de l'Union européenne (UE), s'appuyant sur un cadre institutionnel, des acteurs et des procédures bien définis.

Le Système national des données de santé a été profondément réformé en 2019 pour tenir compte des besoins des acteurs concernés en termes de données, mais aussi d'un contexte politique favorable au développement du numérique en santé.

L'accès aux données médico-administratives du SNDS : une réglementation qui repose sur un équilibre entre ouverture des données et respect de la vie privée

Le Système national des données de santé historique, tel qu'il a été mis en place en 2016, a deux missions principales : il rassemble et met à disposition trois bases exhaustives au niveau national de données de santé publique préexistantes et indépendantes les unes des autres. Il est alimenté par :

- la base de données Sniiram (Système national d'information inter-régimes de l'Assurance maladie) contenant les données relatives à toutes les dépenses de l'Assurance maladie ;

1. Loi n° 2016-41 du 26 janvier 2016 de modernisation de notre système de santé. Elle institue le SNDS à l'article 193, dont les dispositions ont été ensuite intégrées aux articles L. 1461-1 à L. 1461-2 du Code de la santé publique.

- la base de données PMSI (Programme de médicalisation des systèmes d'information) contenant les données d'analyse de l'activité des établissements de santé ;
- la base de données du CépiDc (Centre d'épidémiologie sur les causes médicales de décès) contenant les données relatives aux causes de décès recueillies par les collectivités territoriales.

Des données médico-sociales liées au handicap, fournies par les maisons départementales des personnes handicapées viendront bientôt compléter cette base exhaustive[2].

Le SNDS, à l'image d'un entrepôt, regroupe donc des données de santé au sens strict, des données sur l'état de santé d'un patient (une affectation de longue durée par exemple ou un diagnostic d'hospitalisation) mais également des données portant sur le patient ou son professionnel de santé et enfin des informations relatives aux soins qu'il a consommés.

La Caisse nationale de l'assurance maladie (Cnam) est chargée du rassemblement et de la mise à disposition des données du Système national des données de santé historique. L'Institut national des données de santé (INDS) représentait le guichet unique pour le suivi des demandes de mise à disposition en vue de réaliser une étude, une recherche ou une évaluation dans le domaine de la santé. L'INDS opérait ainsi le suivi des formalités d'accès aux données du Système national des données de santé et ce conformément au cadre réglementaire.

La plupart des acteurs, publics et privés, devaient en effet déposer un dossier de demande d'accès présentant leur projet de recherche et accomplir des formalités obligatoires auprès d'un comité scientifique d'abord, puis de la Commission nationale de l'informatique et des libertés (Cnil). Un comité de protection des personnes (CPP), si le projet porte sur une recherche impliquant la personne humaine, ou le Comité d'expertise pour les recherches, les études et les évaluations dans le domaine de la santé (Cerees), pour celles n'impliquant pas la personne humaine, émet un avis sur les aspects méthodologiques et scientifiques du dossier avant que la Cnil ne se prononce sur le respect de la protection des données. Le Comité d'expertise sur l'intérêt public (CEIP) de l'INDS produit un avis sur ce point. Cette procédure permet de faciliter l'examen par la Cnil des projets de recherche et de réduire les délais d'instruction des demandes.

En revanche, certains organismes publics ou services chargés d'une mission de service public, dont la liste est prévue par voie réglementaire, bénéficient d'un accès dit permanent au Système national des données de santé et sont dispensés de formalités.

2. Un cinquième flux contenant des données de remboursement des complémentaires santé était prévu mais il n'a pas encore été mis en œuvre.

Le Système national des données de santé s'est construit au croisement de deux logiques, une logique de valorisation de cette masse considérable de données de santé ainsi collectée par des procédures homogénéisées et transparentes d'une part, et une logique de protection des patients par l'instauration de garde-fous juridiques et techniques d'autre part.

La réglementation encadre les motifs pour lesquels les données du Système national des données de santé peuvent être utilisées et interdit qu'elles servent à la promotion commerciale de produits de santé et à une éventuelle modification de cotisations ou primes d'assurance d'une personne présentant un risque. La loi formule également l'obligation de « pseudonymiser » les données du Système national des données de santé afin de rendre impossible l'identification directe un patient.

Des règles en matière de sécurité et confidentialité qui s'appliquent aux organismes ayant accès aux données du Système national des données de santé sont également prévues. Un référentiel de sécurité, propre au Système national des données de santé, a en outre été élaboré par le ministère chargé de la santé. Il précise les règles de sécurité que tout système utilisant des données du Système national des données de santé doit mettre en place en termes de procédure d'accès, d'habilitation des utilisateurs, de formations obligatoires et de traçage des traitements effectués, les données étant mises à disposition dans des bulles sécurisées.

Enfin, l'ensemble des principes applicables en matière de protection des données sont en vigueur, non seulement les droits que les personnes dont les données sont utilisées peuvent exercer mais également les obligations encadrant la collecte des données[3].

La loi d'adaptation du RGPD a maintenu les grands principes prévalant en France sur l'accès aux données de santé

Le règlement général sur la protection des données (RGPD), dont les dispositions sont en vigueur depuis le 25 mai 2018, approfondit et étend les droits et obligations à la charge des entités traitant des données à caractère personnel. Il accroît notamment la responsabilisation des acteurs en supprimant la plupart des formalités préalables et en faisant porter sur les

3. Les personnes ont le droit d'obtenir les informations relatives au traitement opéré (catégories de données collectées, identité du responsable de la collecte, objectifs du traitement, durée de conservation, destinataires, éventuels transferts ou réutilisations). Elles disposent également d'un droit d'accès (droit de connaître les données qui ont été traitées), de rectification (droit de corriger les données collectées) et d'opposition sous certaines conditions (droit de refuser que les données soient collectées).
S'agissant des obligations prévues par le RGPD, les données doivent être collectées et traitées de manière licite et loyale, pour des raisons déterminées et légitimes ; les données doivent être pertinentes, exactes et la durée de conservation doit être proportionnée à l'usage qui en est fait.

entités la charge de s'assurer, par elles-mêmes, que leurs traitements sont conformes aux dispositions en vigueur. Toutefois, s'agissant des données sensibles, dont les données de santé, le RGPD permet à chaque État de prévoir des procédures particulières.

La France, par la loi n° 2018-493 du 20 juin 2018 relative à la protection des données personnelles, a fait le choix de maintenir l'obligation d'obtenir une autorisation préalable de la Cnil pour traiter des données de santé dans tous les cas où aucun autre cadre ne permet d'y déroger (exceptions législatives, conformité à un référentiel…). Le rôle des instances consultatives (CPP, Cerees et CEIP) a été conservé. Cette volonté de conserver un haut niveau de protection des données de santé a conduit à l'instauration, par cette même loi, du comité d'audit du Système national des données de santé. Ce comité, présidé par le haut fonctionnaire de défense et de sécurité du ministère chargé de la santé, diligente des audits visant à s'assurer du bon respect des règles d'accès et d'usage du Système national des données de santé.

La réforme de 2019 élargit le périmètre du SNDS et modifie la gouvernance des données de santé afin que la France devienne pionnière dans la structuration des données de santé

Les évolutions du SNDS, introduites par la loi relative à l'organisation et à la transformation du système de santé[4], s'inscrivent dans la stratégie de développement du numérique et de l'intelligence artificielle, annoncée par le président de la République en mars 2018.

La loi du 24 juillet 2019 fonde l'extension du périmètre du Système national des données de santé à l'ensemble des données de santé dont le recueil est directement ou indirectement financé par des fonds publics, en particulier les données cliniques recueillies dans le cadre du parcours de soins, ce qui inclut les activités de prévention, de diagnostic, de soins et de suivi social et médico-social. De nouvelles sources de données sont ainsi ajoutées aux quatre bases historiques, l'appariement entre ces données devant faciliter leur structuration et démultiplier leurs usages. Ces nouvelles données permettent essentiellement d'apporter de l'information sur l'état de santé précis du patient (résultats d'examen, diagnostics des médecins) et sur ses déterminants de santé (situation sociale, indice de masse corporelle, consommation de tabac…), dont l'absence limitait les usages du Système national des données de santé.

Ces modifications s'accompagnent d'une refonte de la gouvernance des données de santé. L'Institut national des données de santé (INDS) est remplacé par la Plateforme des données de santé (PDS), également désignée par son appellation anglaise Health Data Hub, dont les missions sont élargies. Si elle conserve le rôle de guichet unique pour les demandes d'accès, la Plateforme peut en outre procéder pour le compte de tiers à des traitements de données et se voit confier un rôle de facilitateur, tant pour les usagers dans l'exercice de leurs droits que pour les porteurs de projets dans leur réalisation.

Une organisation bicéphale est instituée afin de mettre en œuvre le Système national des données de santé élargi, la Cnam et la PDS en deviennent les deux responsables. Le Cerees devient le Comité éthique et scientifique pour les recherches, les études et les évaluations dans le domaine de la santé (Cesrees), qui rend désormais des avis plus complets. Les demandes pour les recherches n'impliquant pas la personne humaine sont examinées par un seul comité qui se prononce sur les différents aspects (méthodologiques, éthiques et intérêt public) qui sont généralement imbriqués. Le système des accès permanent a, en revanche, été conservé.

Il est encore tôt pour évaluer l'impact des changements prévus par la loi du 24 juillet 2019 dès lors que tous les textes d'application ne sont pas encore entrés en vigueur et que la Plateforme des données de santé n'est pas complètement opérationnelle. Il importe toutefois de souligner que l'ambition de cette réforme est de faire de la France l'un des pays pionniers dans le domaine de la promotion et de la valorisation des données de santé. Au sein de l'Union européenne, seule la Finlande a mis en œuvre un système relativement comparable. Sans conteste, la refonte du Système national des données de santé confère à la France un atout alors que la Commission européenne a lancé en septembre 2019 une initiative destinée à créer un « espace européen de la donnée de santé ».

Exemple d'utilisation de données du Système national des données de santé

De nombreuses études mobilisent les données du SNDS. Cette base de données a déjà permis d'améliorer les connaissances sur le système de santé et ses acteurs mais également d'améliorer significativement la pharmacovigilance (depuis les affaires du Médiator ou de la Dépakine par exemple) ou encore de mieux suivre les parcours de soins et leurs conséquences. On peut par exemple citer l'étude de M. Coldéfy et C. Gandé [15] qui, à partir de la mobilisation des trois sources de données actuellement présentes dans le Système national des données de santé (Sniiram, PMSI et CépiDc), montre que la réduction de l'espérance de vie des individus suivis pour des troubles psychiques sévères atteint en moyenne 16 ans chez les hommes et 13 ans chez les femmes. Si ce phénomène était connu, seule la richesse des données du SNDS permet d'en objectiver l'ampleur, de le décliner par cause de mortalité ou troubles psychiatriques ou encore de suivre son évolution temporelle et géographique.

Les références entre crochets renvoient à la Bibliographie générale p. 57.

4. Loi n° 2019-774 du 24 juillet 2019 relative à l'organisation et à la transformation du système de santé (art. 41).

La prévention par la transparence dans le contrôle des infections liées aux soins

Lara Khoury
Ad. E.,
professeur agrégée,
faculté de droit,
université McGill
(Montréal, Canada)

La transparence et la communication ouverte participent-elles à la prévention en matière de sécurité des soins de santé offerts à la population ? Plusieurs le croient et aspirent à réduire les taux d'accidents évitables associés aux soins de santé par le biais de techniques de partage public des données liées à la sécurité sanitaire. C'est le cas notamment de la province canadienne du Québec, dont le gouvernement situe la transparence et la communication au cœur de la prévention des infections associées aux soins de santé (IAS) dans son plan d'action ministériel 2015-2020 sur la prévention et le contrôle des infections nosocomiales. Le Québec est à l'avant-garde de la divulgation des événements indésirables associés aux soins, à la fois aux *patients* et aux *comités de gestion des risques* œuvrant au sein de chaque établissement de soins. Toutefois, nous nous attarderons ici plutôt à la divulgation *au public* des taux d'infections associées aux soins de santé et à son effet possible sur l'amélioration de la qualité des soins offerts à la population.

La surveillance des infections associées aux soins de santé au Québec

La divulgation publique des taux d'IAS au Québec a lieu dans le cadre du Programme provincial de prévention et de contrôle des infections (PCI). Bien que le ministre de la Santé établisse les modalités de surveillance, la méthodologie et les outils, la surveillance est décentralisée et confiée aux établissements de santé. Toutefois, des structures locales et régionales font aussi partie du système québécois de surveillance des IAS (CS-PCI ; TC-PCI ; TRPIN).

La collecte de données locales est obligatoire pour sept IAS : les bactériémies nosocomiales panhospitalières (BACTOT), les bactériémies nosocomiales associées aux cathéters centraux aux soins intensifs (BACC-USI), les bactériémies associées aux accès veineux en hémodialyse (BAC-HD), les bactériémies à *Staphylococcus aureus* résistant à la méthicilline (BAC-SA), les diarrhées à *Clostridium difficile* (DACD), les infections à bacilles à Gram négatif producteurs de carbapénémases (BGNPC) et les infections à entérocoque résistant à la vancomycine (ERV). L'obligation de transmission des données est parfois restreinte à certains types d'établissements : par exemple, pour les ERV, les centres hospitaliers de soins généraux et spécialisés enregistrant plus de mille admissions par année ; pour les DACD, les centres hospitaliers de soins de courte durée.

Les établissements de santé communiquent ensuite leurs données locales à l'Institut national de santé publique du Québec (INSPQ) par le biais du Système d'information pour la surveillance des infections nosocomiales (SI-SPIN), qui permet de connaître à l'échelle provinciale l'incidence des infections associées aux soins de santé. Le SI-SPIN a comme objectif, notamment, de permettre aux établissements de comparer leur taux d'infection, de réduire le taux d'incidence des bactériémies au minimum et d'identifier des situations d'éclosion dans chaque établissement.

Le public n'a pas accès aux données du SI-SPIN. En revanche, les résultats nationaux de chaque programme de surveillance sont analysés et diffusés à la population dans les rapports du Comité sur les infections nosocomiales du Québec (CINQ), un comité faisant partie de l'Institut national de santé publique du Québec (INSPQ). Le CINQ produit annuellement un rapport général, ainsi que des rapports spécifiques pour chaque IAS surveillée. Alors que le rapport général (« Faits saillants ») adopte une forme et un langage accessibles au public, les rapports spécifiques (« Résultats de surveillance ») sont techniques et s'adressent de toute évidence plutôt aux acteurs du système de santé. Le rapport général du CINQ effectue des comparaisons dans le temps, nationales et internationales, basées sur des analyses statistiques. Il émet aussi des recommandations liées à la prévention et au contrôle des infections associées aux soins de santé à l'échelle provinciale, directement fondées sur les données recueillies. Il inclut en outre des données statistiques tel le taux d'incidence, la description des cas et la microbiologie. Des données spécifiques à chaque établissement de santé sont aussi fournies pour chaque IAS surveillée.

Outre le Québec, plusieurs provinces canadiennes partagent leurs taux d'infections associées aux soins de santé avec la population. Les taux et les tendances des IAS dans les établissements canadiens sont aussi accessibles publiquement par le biais du Programme canadien de surveillance des infections nosocomiales (PCSIN). La divulgation publique des indicateurs de sécurité des soins des établissements de santé au Canada — tels les taux d'IAS — s'insère dans un mouvement nord-américain favorable à la divulgation publique obligatoire des taux d'IAS. Ce mouvement a pris naissance aux États-Unis en 2003 à la suite d'une campagne majeure lancée par la Consumer Union, un lobby proconsommateur associé à la revue *Consumer Reports*. Cette campagne, probablement influencée par la publication du rapport de l'Institute of Medicine, *To Err is Human,* provoqua l'adoption dans de nombreux États américains de lois forçant les établissements de santé à faire rapport de leurs taux d'infection, publiquement pour la majorité.

Informer sur la qualité des données

Est-ce que la divulgation publique des infections associées aux soins de santé influe sur la sécurité sanitaire ? Difficile de le savoir. Un premier enjeu à cet égard concerne la qualité des données recueillies et partagées. L'effet du partage des données recueillies est évidemment tributaire de leur fiabilité. Pourtant, le CINQ n'émet presque aucun commentaire dans son rapport annuel général au sujet du bon fonctionnement du système de surveillance et de la qualité des données transmises à la population. Nous n'avons repéré par ailleurs aucune évaluation à cet égard provenant d'autres entités.

Une discussion de la qualité des données visant les taux d'infections associées aux soins de santé, accessible à la population et lui permettant de mieux comprendre les données transmises et leurs limites, serait opportune. Cela est d'autant plus vrai que la littérature américaine et canadienne documente les problèmes qui peuvent affecter la qualité des données sur la sécurité des soins. Cette littérature s'inquiète, par exemple, de la possibilité que les infections soient sous-rapportées ou mal rapportées ou qu'il y ait une absence d'uniformité dans les définitions de ces infections [52, 57]. Elle soulève le besoin de standardiser/ajuster les données pour tenir compte des différences dans les populations, de la portée et du type de procédures médicales concernées, des risques d'infection inhérents au patient, et de la complexité des soins offerts à chaque hôpital [23, 30]. Enfin, on note des enjeux de manque d'uniformité dans les techniques de surveillance [23, 30] ou leur utilisation inégale, ainsi que la nécessité de comprendre les limites de ces techniques. Ce ne sont ici que quelques exemples des mises en garde et critiques exprimées. Une évaluation des facteurs portant atteinte à la qualité des données transmises à la population nous apparaît requise pour leur assurer un véritable impact, et les résultats de cette analyse devraient être inclus dans le partage public des données. Le CINQ et l'INSPQ devraient aussi faire rapport à la population des ajustements périodiques du système de collecte de données, à la lumière de ces facteurs.

Divulgation des données et amélioration des soins

Un des moteurs du système de divulgation publique est la conviction que les établissements de santé amélioreront le contrôle des infections s'ils dévoilent publiquement leurs taux d'infections. Toutefois, certains auteurs remarquent qu'il n'existe que peu de preuves que la divulgation publique des IAS a une incidence sur les taux d'infections, ou même qu'elle améliore la qualité des soins apportés au patient, ou leur sécurité [30]. Certains effets sont constatés, cependant. Par exemple, Haustein *et al.* [30] notèrent en 2011 une baisse remarquable et inattendue des infections au SARM au Royaume-Uni après que la divulgation des taux de ces infections soit devenue obligatoire. Au Canada, Daneman *et al.* [20] observèrent en 2012 que la divulgation publique des infections à *C. difficile*, qui débuta en 2008 dans la province de l'Ontario, fut associée à une chute de 26 % de ces infections dans les hôpitaux, sauvant ainsi plus de cent vies par année. Avant la mise en place du système de divulgation obligatoire, soit entre 2002 et 2008, les infections causées par la bactérie *C. difficile* avaient été constamment en hausse. Enfin, aux États-Unis, Cardo *et al.* [8] constatèrent en 2005 que, depuis le début de la divulgation des IAS — à l'époque, celle-ci était volontaire et confidentielle —, les taux de bactériémies associées aux cathéters intravasculaires centraux, de pneumonies associées à la ventilation mécanique et les infections urinaires associées aux cathéters avaient aussi connu des baisses importantes.

Haustein *et al.* [30] posent l'hypothèse que la divulgation des taux d'infections associées aux soins de santé attire l'attention sur la situation et qu'elle agit comme un outil de renforcement externe, associé à un changement de culture organisationnelle et à une augmentation des activités préventives. Bien que ne pouvant pas déterminer exactement le mécanisme ayant entraîné un déclin des taux d'infections à la suite de la divulgation publique des infections à *C. difficile* en Ontario, les auteurs canadiens Daneman *et al.* [20] émettent l'hypothèse que cette divulgation aurait élevé les infections à *C. difficile* au rang de priorité et motivé les hôpitaux à mieux adhérer aux standards de prévention et de contrôle de ces infections. Aux États-Unis, on note que la demande de données publiques sur les taux d'infections associées aux soins de santé a mené à une amélioration par les hôpitaux de leurs pratiques en prévention des infections, afin de répondre aux attentes des législateurs et des patients, mais qu'il n'y aurait aucun lien clair entre l'imposition d'une obligation de divulgation publique et des procédures améliorées ou des réductions dans les taux d'infections subséquents [42].

Les références entre crochets renvoient à la Bibliographie générale p. 57.

Conclusion

La transparence et la communication ouverte et franche à l'égard des accidents médicaux sont des conditions nécessaires à une culture de prévention des accidents et de promotion de la sécurité des patients dans la délivrance des soins de santé. La divulgation des indicateurs de qualité a comme objectifs avoués non seulement de renforcer l'autonomie des patients en leur permettant de faire un choix éclairé des services de santé qui leur sont offerts, mais également d'améliorer la qualité des soins. Des recherches au Canada, aux États-Unis et au Royaume-Uni démontrent en effet — sans pouvoir établir un lien de causalité — que la divulgation publique des IAS semble contribuer à un changement de culture et d'attitude dans la prévention des risques au sein même des institutions de soins. Les entités participant à la surveillance et à la divulgation des données pertinentes doivent toutefois faire preuve de vigilance afin d'assurer la qualité des données et communiquer ouvertement au public les limites de celles-ci.

Données massives et prévention du risque épidémique

Paul Véron
Maître de conférences en droit privé à l'université de Nantes, Laboratoire droit et changement social (UMR 6297)

La découverte de l'origine microbienne ou virale des maladies infectieuses a constitué une avancée majeure pour la santé des populations au tournant du XX^e^ siècle. Le développement de la vaccination et des mesures d'hygiène ont ainsi permis un bond d'espérance de vie d'environ vingt ans entre 1900 et 1950. Grâce à ces progrès, certaines maladies infectieuses ont presque disparu en Europe. C'est le cas de la variole. En France, l'obligation vaccinale contre cette maladie, imposée à partir de 1902, a été supprimée par la loi du 2 juillet 1979. Toutefois, de nouvelles épidémies se sont répandues en Europe et dans le monde. Outre le sida apparu dans les années 1980, on peut citer le SARS, l'encéphalopathie spongiforme bovine (la vache folle), le virus Ebola, le virus Zika, différentes formes de grippes, ou très récemment le SARS-Cov2, nouveau Coronarivus dont la maladie est dénommée Covid-19.

Quels outils juridiques de gestion du risque épidémique ?

Le droit français compte plusieurs dispositifs visant à lutter contre les épidémies. Outre les recommandations ou obligations vaccinales, on citera le système des déclarations obligatoires des maladies contagieuses, les mesures de désinfection, les contrôles sanitaires aux frontières, les dispositifs d'isolement des malades. Plus récemment, l'état d'urgence sanitaire déclaré à la suite de la crise de la Covid-19 a été l'occasion d'un florilège de mesures, parfois extrêmement contraignantes pour les libertés, et visant à lutter contre la diffusion du virus : fermetures d'établissement et cessations d'activités, port du masque obligatoire et restrictions du droit de circulation, allant jusqu'au confinement à domicile.

L'usage de données en grand nombre : un phénomène nouveau ?

Champ traditionnel de l'intervention étatique, la santé publique représente selon l'OMS « *l'ensemble des efforts des institutions publiques pour améliorer, promouvoir, protéger et restaurer la santé de la population grâce à une action collective* ». Ces actions collectives passent traditionnellement par le recueil de données et l'établissement de statistiques, pour comprendre et mesurer l'état de santé des populations (par exemple identifier le taux de couverture vaccinale) ou pour surveiller la diffusion de maladies. Lors de la grande peste du XIV^e^ siècle existaient déjà des registres paroissiaux visant à recenser le nombre de malades sur les différents territoires du Royaume. La démarche n'est donc pas nouvelle.

Le recueil de données s'illustre notamment à travers le système des maladies à déclaration obligatoire. Le Code de la santé publique prévoit que les médecins et responsables des services et des laboratoires d'analyses de biologie médicale sont tenus de transmettre à l'autorité sanitaire les données individuelles dont ils disposent concernant deux grandes catégories de maladies : celles qui nécessitent une intervention urgente locale, nationale ou internationale, et celles dont la surveillance est nécessaire à la conduite et à l'évaluation de la politique de santé publique. En tout, une trentaine de maladies figurent sur la liste élaborée par les pouvoirs publics. La grande majorité d'entre elles sont des maladies infectieuses transmissibles. Sont concernées, entre autres, le chikungunya, le choléra, la dengue, la diphtérie, les fièvres hémorragiques africaines (dont Ebola), la fièvre jaune, le paludisme, la peste, la poliomyélite, la rage, la rougeole, la rubéole, la Creutzfeldt-Jakob, le typhus, le zika. Étonnamment, la Covid-19 n'a pas été ajoutée. Par ailleurs, depuis 1984, le réseau Sentinelles de recherche et de veille sanitaire suit plusieurs maladies infectieuses et alerte sur les épidémies grâce à la contribution de centaines de médecins généralistes et pédiatres répartis sur tout le territoire. Ces professionnels rapportent au moins une fois par semaine le nombre de cas observés pour plusieurs maladies transmissibles courantes, dont les syndromes grippaux, la varicelle et le zona. Les données sont transmises, *via* un réseau sécurisé, à l'Institut d'épidémiologie et de santé publique Pierre Louis (commun à l'Inserm et à la Sorbonne) et à l'agence nationale de santé publique Santé publique France, qui a absorbé en 2016 l'ancien Institut de veille sanitaire (InVS).

Des évolutions à la fois quantitatives et qualitatives dans l'utilisation des données peuvent néanmoins être relevées. D'une part, la lutte contre le risque épidémique associe aujourd'hui des acteurs privés. Des collaborations entre les autorités sanitaires et les entreprises privées ou multinationales se font jour, à l'image du programme Google Flu, fruit d'un partenariat entre l'institut fédéral de veille sanitaire américain formé par les Centers of Diseases Control and Prevention (CDC) et la société Google. D'autre part, on observe un double changement d'échelle et de nature des données utilisées, très diverses, y compris des données qui n'ont initialement pas été collectées pour une finalité d'ordre médical ou sanitaire. Avec Google Flu, il s'agissait de tirer profit de millions de traces issues des requêtes des utilisateurs du moteur de recherche sans que lesdites traces aient été à l'origine conservées pour cette utilité. Une autre illustration concerne l'usage des données détenues par les aéroports de différents pays sur les déplacements des usagers à une échelle

locale ou mondiale. Ainsi du simulateur GLEAM (Global Epidemic and Mobility Model), mis au point par une équipe de chercheurs et destiné à prédire la dissémination de certaines épidémies au niveau mondial, en exploitant entre autres les données de transport aérien, la densité de population et de flux de voyageurs entre différentes zones géographiques. Cet outil a notamment été utilisé, en coordination avec l'OMS, pendant la crise du virus Ebola afin d'évaluer en temps réel le risque d'importation de cas dans différents pays. D'autres outils pourraient être cités, dont le logiciel HealthMap, développé par des épidémiologistes et des informaticiens américains en 2006, avec pour objectif d'identifier la naissance de foyers d'épidémies dans le monde, en se fondant là encore sur des sources très diverses (notes de départements sanitaires et d'organismes publics, rapports officiels, données issues d'Internet et de réseaux sociaux, bulletins d'informations locaux, etc.). Au Yémen, un système informatique a été conçu pour permettre aux acteurs humanitaires d'anticiper les épidémies de choléra. Dans plusieurs endroits du pays, des pluies importantes submergent le réseau d'égout, endommagé par la guerre, ce qui favorise la propagation de la bactérie. Pour prédire le risque sur chaque zone du pays, le logiciel croise des données pluviométriques avec d'autres, relatives à la densité de population et à l'accès à l'eau potable.

L'évolution concerne également les modalités de collecte des données, avec notamment l'utilisation croissante d'objets connectés. On peut cette fois prendre des exemples plus locaux qui concernent un volume de données moins important, à l'image des expériences visant à comprendre le phénomène de diffusion d'une maladie infectieuse à l'échelle d'une ville, d'un hôpital, ou encore d'une cour d'école. Au CHU de Lyon, cinq cents patients et professionnels ont été équipés de capteurs électroniques (puces RFID) afin d'enregistrer l'ensemble de leurs contacts – plusieurs millions – sur une période de six mois, dans le but d'améliorer la compréhension de la dissémination des staphylocoques et éventuellement proposer de nouvelles stratégies d'hygiène. La récente crise sanitaire a également vu naître différentes applications de traçage numérique de « cas contacts » de personnes contaminées, dont StopCovid, ayant fait l'objet d'un encadrement réglementaire spécifique. Les personnes alertées – celles qui ont volontairement installé l'application – sont encouragées à se faire tester et le cas échéant à rester confinées, afin de réduire la propagation du virus. Le système repose sur une collecte automatisée de données de localisation ou relatives aux interactions sociales individuelles.

Quels bénéfices pour quels risques ?

L'exemple de Google Flu permet d'illustrer à la fois l'apport potentiel de l'utilisation des *big data* et de leur traitement algorithmique, mais surtout les biais, limites et risques qui peuvent y être associés. Ce programme a été créé par Google en 2008 dans le but de mieux anticiper et prévenir l'apparition de foyers de grippe sur le territoire américain. La grippe a en effet la particularité d'être un virus sujet à des mutations fréquentes, ce qui explique que les vaccins doivent constamment s'adapter. Sa diffusion rapide impose en outre une réactivité des autorités sanitaires sur deux plans : en amont, quant à la détection de la nouvelle souche du virus (c'est ce qui nous intéresse ici) ; en aval, sur le temps nécessaire à la fabrication du vaccin. Compte tenu de ces contraintes, il peut arriver qu'au moment où le nouveau vaccin est disponible, le virus en circulation soit différent ou encore que la vaccination de masse ne soit possible qu'après que le pic épidémique soit passé.

Aux États-Unis, les CDC (déjà évoqués) reposent sur un réseau de 150 laboratoires biologiques, qui transmettent les types, éventuellement les séquençages, des virus détectés, et un réseau d'environ 2 500 médecins couvrant le territoire des États-Unis et chargés de signaler à l'autorité sanitaire les maladies « semblables à la grippe ». Un échantillon de patients ainsi détectés est ensuite examiné de manière plus approfondie, afin de déterminer quelle proportion souffre réellement de la grippe. Si ce système est efficace, il suppose néanmoins que les patients prennent rendez-vous chez un médecin, le voient en consultation et que ce dernier rapporte avoir constaté un cas de grippe, ce qui peut prendre plusieurs jours ou semaines. C'est dans ce contexte que les CDC et Google se sont entendus pour développer un système d'alerte avancé avec pour but de donner, presque en temps réel, une indication sur l'avancée de la grippe. L'idée était la suivante : en se penchant sur la période 2003-2007, Google et les CDC ont observé une correspondance entre d'une part l'augmentation des requêtes en lien avec la grippe sur le moteur de recherche Google, et d'autre part, l'augmentation des déclarations des cas de grippe faites auprès des CDC. La confrontation des deux ensembles de données a fait ressortir une correspondance pour quarante-cinq termes de recherche ayant un rapport logique avec la grippe (complications, remèdes, symptômes, toux, etc.). En d'autres termes, grâce à cet algorithme, en observant la fréquence d'apparition de certains termes dans les requêtes, il serait possible de détecter le début de la phase de propagation de la grippe en temps quasi réel, et ainsi de gagner plusieurs semaines dans l'identification d'une nouvelle souche du virus.

Plusieurs limites de l'outil ont toutefois été relevées. Premièrement, il n'est pas prévu pour remplacer les déclarations faites par les médecins, pour au moins deux raisons. D'une part, le modèle a été élaboré à partir des données des CDC puisque l'algorithme repose sur la comparaison entre les requêtes générales des utilisateurs de Google et ces déclarations. D'autre part, le programme permet d'établir une corrélation entre le nombre de requêtes effectuées sur Google et le nombre de déclarations faites à l'organisme de veille sanitaire, mais pas un rapport de causalité. Il permet seulement de présumer une augmentation des cas de malades.

Deuxièmement, il peut exister un risque important de surestimation du risque. En effet, les recherches ne sont pas forcément effectuées par les personnes ressentant les symptômes de la grippe. Des paramètres internes et externes peuvent conduire à une augmentation du nombre des requêtes, tels que la forte médiatisation du sujet ou des déclarations faites par les pouvoirs publics. C'est précisément cette surestimation du risque qui a posé problème dans le cas de la pandémie grippale particulièrement virulente de l'hiver 2012-2013. En janvier 2013, à New York, le modèle prédit presque le double d'infections par rapport à ce que les médecins rapportent finalement. Ces prédictions faussées peuvent avoir des conséquences sur les comptes publics, par exemple des commandes massives de vaccins sur la base de ces estimations inexactes. Des changements dans le fonctionnement des moteurs de recherche intervenus postérieurement à la création du modèle – tels qu'un système de saisie semi-automatique – peuvent de même influer sur son utilisation.

Troisièmement, des tentatives de manipulation du modèle ne sont pas exclues : un fabricant de médicament antigrippal pourrait par exemple chercher à augmenter artificiellement le nombre de recherches d'un terme faisant partie du modèle.

À partir du début de l'année 2013, le dispositif Google d'alerte avancée sur la grippe a été adopté par vingt-neuf pays et étendu à une autre maladie, la dengue. Toutefois, en raison de ses résultats décevants, il n'est plus utilisé depuis 2015. En France, l'application StopCovid – reposant sur le volontariat – n'a guère rencontré davantage de succès. Moins de 3 % des Français l'avaient téléchargée au milieu de l'été 2020, nombre très insuffisant pour espérer impacter les chaînes de transmission du virus. Outre les questions relatives à la confidentialité et à la fiabilité des données recueillies, plusieurs limites de l'outil ont été soulignées, en particulier au regard du sous-équipement en téléphone mobile des personnes âgées, particulièrement exposées, et des enfants, le plus souvent porteurs asymptomatiques. De quoi relativiser les discours prometteurs des pouvoirs publics et de l'industrie du numérique pour ces dispositifs technologiques de prévention – et demain de surveillance ? – sanitaire.

Hygiène et éducation alimentaire à l'heure des *big data* : aide à la décision et gains pour la santé ?

Marine Friant-Perrot
Maître de conférences-HDR à la faculté de droit et de sciences politiques, université de Nantes

L'étude des effets de l'alimentation sur le corps remonte à l'Antiquité, mais l'éducation alimentaire et les principes de diététiques ont connu un regain d'intérêt aux XVIII^e^ et XIX^e^ siècles sous l'influence des hygiénistes. L'idée que l'environnement de vie et les habitudes individuelles peuvent largement influer sur le développement de maladies a ainsi favorisé la diffusion de principes d'hygiène publique inspirés du courant néohippocratique. Si la construction moderne du droit de la santé publique a marqué une rupture avec ce discours hygiéniste – ce n'est plus l'individu qui est comptable de sa santé envers la collectivité, mais bien cette dernière qui lui doit la protection de sa santé –, on a vu resurgir depuis les années 2000 divers « repères », « guides », « recommandations nutritionnelles » dans un contexte de forte augmentation du surpoids et de l'obésité, qui renouent avec ce modèle fondé sur la responsabilisation individuelle. Dans notre pays, près de la moitié des adultes et 17 % des enfants sont en surpoids (respectivement 17 % et 4 % sont obèses) et, au-delà des conséquences sanitaires (diabètes, maladies cardiovasculaires…), le coût financier en résultant est évalué par le Trésor public à 20 milliards d'euros. Parmi les mesures recommandées par l'OMS pour améliorer l'alimentation des populations figurent des leviers juridiques de nature à modifier les comportements alimentaires dans un sens plus vertueux (information et éducation nutritionnelles, taxation nutritionnelle, réglementation du marketing alimentaire, reformulation des produits…). Au sein de cette panoplie de mesures, la France comme l'Union européenne ont privilégié les mécanismes informationnels et éducatifs en se fondant sur le principe qu'une personne informée sur la composition nutritionnelle des aliments saura faire des choix conformes à sa santé. En témoigne le programme national relatif à la nutrition et à la santé (PNNS), dont la France s'est dotée dès le début des années 2000. Au plan européen, le règlement n° 1169/2011 dit Inco (information des consommateurs sur les denrées alimentaires) a suivi cette voie en prévoyant que toutes les denrées alimentaires préemballées doivent obligatoirement comporter une déclaration nutritionnelle qui doit aider le consommateur à opter pour une alimentation équilibrée. Cette déclaration peut être complétée par une forme d'étiquetage simplifiée. Il s'agit en France du Nutri-Score, logo

ou signal coloriel recommandé aux exploitants du secteur agroalimentaire par l'arrêté du 31 octobre 2017, et qui constitue une mesure phare du nouveau PNNS publié le 20 septembre 2019. Le mangeur informé de ses choix, voire incité à s'alimenter plus sainement, est ainsi perçu comme l'artisan de sa propre santé. Dans cette lutte contre les maladies chroniques d'origine nutritionnelle, la collecte massive de données sur la composition des aliments et sur les comportements alimentaires offre de nouveaux outils informationnels pour prévenir l'obésité et le surpoids [17]. Mais quelle sera sa véritable incidence sur les choix alimentaires et la santé nutritionnelle [31] ?

Des données massives nouvelles sur l'alimentation dans un contexte de renouvellement des champs classiques

La collecte et l'exploitation de données sur la composition des aliments et sur les comportements alimentaires connaissent une croissance exponentielle depuis les années 2000. À la suite de la crise sanitaire de la vache folle, il est apparu nécessaire de restaurer la confiance des consommateurs en améliorant la transparence sur les ingrédients composant les aliments et en instaurant une traçabilité obligatoire permettant d'identifier l'ensemble des maillons de la chaîne alimentaire. La place du numérique s'est alors accrue dans les États membres de l'Union européenne sous l'effet du règlement 178/2002 relatif à la législation alimentaire. Grâce au code-barres (EAN), toutes les informations relatives au produit, au fabricant et au pays d'origine sont recueillies et permettent de suivre l'itinéraire de la denrée alimentaire jusqu'au consommateur final. Au-delà de la traçabilité des opérateurs imposée par l'article 18 du règlement 178/2002 et de l'objectif initial lié à la préservation de la sécurité sanitaire, ce sont de véritables cartes d'identité numériques des aliments qui sont constituées sur les qualités nutritionnelles ou environnementales du produit.

Dans cette course à la transparence, on recense des bases de données collaboratives, comme Open Food Fact, qui a conclu un partenariat avec Santé publique France pour nourrir la base de calcul du Nutri-Score (plus de 700 000 aliments recensés), des start-up qui développent des applications smartphone comme Yuka ou des initiatives des opérateurs économiques eux-mêmes comme NumAlim. Les métabases de données répertorient les informations publiques qui figurent obligatoirement sur l'étiquetage du produit conformément au règlement Inco de 2011 (ingrédients dont allergènes et additifs, composition nutritionnelle…), mais aussi la présence de signes de qualité (bio…) ou des données privées collectées par les maillons de la filière agroalimentaire (présence de résidus de pesticides…).

Les consommateurs sont en attente de ces informations, notamment pour manger plus sainement car près de 35 % d'entre eux utilisent ces applications. Cette collecte de données sur la composition des aliments est complétée par de multiples informations sur les comportements alimentaires, qu'il s'agisse des bases de données de la recherche publique (cohorte Nutrinet Santé…), des données recueillies lors des soins (consultations de suivi de l'obésité…), des données répertoriées par les assureurs dans le cadre de contrat de complémentaires santé comportementales (programme Vitality…) ou des données collectées par les personnes elles-mêmes *via* les applications mobiles de coaching et d'éducation nutritionnelle (compteurs de calories, même à partir d'une photographie d'un plat, conseils nutritionnels et culinaires…) et *via* les objets connectés (pèse-personne, frigo intelligent…).

L'arrivée de ce flot de données coïncide avec une redéfinition des frontières entre le champ alimentaire et le champ médical. On assiste, d'une part, à une « médicalisation » du marché alimentaire avec l'augmentation de la consommation des compléments alimentaires et des allégations nutritionnelles et de santé qui vantent les effets positifs de certains aliments sur la santé. D'autre part, les frontières matérielles de la santé publique s'élargissent. Elles intègrent les déterminants de santé liés au mode de vie, notamment à la nutrition et à l'activité physique. Dans les actions de prévention des maladies nutritionnelles, les acteurs privés sont très présents et on assiste à une forme de « marchandisation » de la prévention en santé (marché des applications et des objets connectés du bien-être et de la santé, comme les pèse-personnes connectés, complémentaires santé comportementales). Ce double mouvement s'accompagne d'une porosité croissante entre le droit de la santé et le droit du marché. Selon une rhétorique commune, le « patient-consommateur » est responsabilisé dans la promotion et l'amélioration de sa santé par l'alimentation, il est « acteur du marché » et « acteur de sa santé », informé et éduqué, il est à même de choisir le régime alimentaire qui lui convient.

Incidence des *big data* sur les choix alimentaires

Dans ce contexte de « sur-responsabilisation » des personnes dans la gestion de leur alimentation et de leur mode de vie, la collecte de données massives a nécessairement une incidence sur les choix alimentaires. Les dispositifs juridiques mis en œuvre pour améliorer la qualité nutritionnelle de l'alimentation sont essentiellement centrés sur l'individu (informations par la déclaration nutritionnelle, Nutri-Score…) et ne modifient que timidement l'environnement alimentaire pour ne pas heurter les libertés économiques (par exemple la taxe soda en France…). Ils procèdent de l'idée, confortée par l'évolution des connaissances scientifiques (épigénétique, micronutrition…) et du marché des aliments santé, que les individus peuvent maîtriser leur alimentation et adopter un régime alimentaire individualisé vecteur d'une santé parfaite. Dans ce cadre, la transparence sur la composition nutritionnelle des aliments *via* le Nutri-Score

Les références entre crochets renvoient à la Bibliographie générale p. 57.

ou les applications comme Yuka, ainsi que la possibilité de mesurer, de comparer ses données nutritionnelles par les applications de *quantified self* constituent des outils normatifs qui incitent les individus à consommer les aliments les plus favorables à leur santé.

Ces possibilités liées aux données massives vont être déployées de manière encore plus importante avec la mise en place du « Store santé » au sein de l'espace numérique de santé (ENS). En vertu de l'article L. 1111-13-1 du Code de la santé publique, chaque usager pourra accéder à ses constantes de santé « *éventuellement produites par des applications ou objets connectés* ». Une place nouvelle est ainsi octroyée aux données dites de « bien-être » : formant un *continuum* avec les données recueillies lors des soins, elles sont recensées par le patient, qui alimente lui-même son espace numérique de santé et renforce ainsi sa capacité à améliorer son alimentation au bénéfice de sa santé.

Cet effet vertueux de la collecte de données sur ce que nous mangeons nécessite toutefois que chacun puisse se référer à une norme alimentaire définie de manière transparente, objective et fondée scientifiquement. Or, on assiste à une privatisation, voire une captation par le marché, du discours nutritionnel et de la prévention de l'obésité et du surpoids. Les forces créatrices de la norme alimentaire ne sont pas uniquement les autorités sanitaires, mais aussi les géants de l'agroalimentaire, qui tentent de développer leurs propres profils nutritionnels pour échapper au classement défavorable qu'établit par exemple le Nutri-Score pour leurs produits (par exemple tentative de mise en place de l'Evolved Nutrition Label, abandonné en 2018). Les messages sanitaires, les applications et les objets connectés conçus par les opérateurs privés ont aussi tendance à se focaliser sur la promotion de l'activité physique en minorant l'incidence de la consommation alimentaire comme cause de l'obésité et du surpoids.

Incidence des *big data* sur la santé nutritionnelle

Si les *big data* ont une incidence indéniable sur les choix alimentaires, il est plus complexe d'établir leur effet bénéfique sur la santé nutritionnelle.

Les risques sanitaires peuvent d'ores et déjà être mis en lumière. En premier lieu, la pertinence scientifique du processus connaît différentes failles. Il existe d'abord un décalage entre les déclarations et la réalité des consommations alimentaires qui ne peut pas toujours être décelé par les différents outils qui recueillent les données. Les applications sont parfois non conformes aux nouveaux repères nutritionnels, actualisés en 2017 (par exemple concernant la limitation de la consommation des jus de fruit ou des produits laitiers). Elles sont généralement opaques (par exemple Yuka qui ne communique pas le contenu détaillé de son algorithme) et simplificatrices (par exemple absence d'identification des légumes secs ou des produits céréaliers complets dans certaines applications alors même que leur consommation est recommandée selon les nouveaux repères). Certaines applications promeuvent des régimes amaigrissants alors même qu'ils sont de nature à accroître les risques nutritionnels et que leur interdiction a été préconisée par le HCSP en 2017. En second lieu, on peut craindre l'existence de pratiques commerciales trompeuses contraires à l'article L. 121-2 du Code de la consommation (art. 5213-3 et suivants du Code de la santé publique pour les dispositifs médicaux). À l'exception du régime juridique applicable aux allégations nutritionnelles et de santé, le contenu des messages relatifs à l'hygiène alimentaire ne fait l'objet que d'un contrôle *a posteriori*. Comme beaucoup d'acteurs ne sont pas issus du secteur de la santé (la moitié des entreprises commercialisant des applications en santé sont issues du monde de l'informatique [17]), le risque d'induire le consommateur en erreur, voire de l'amener à adopter des comportements alimentaires pathologiques comme l'anorexie est à craindre. En présence de liens capitalistiques avec des partenaires commerciaux, comme dans le cas des complémentaires santé comportementales qui récompensent en bons d'achat les comportements sains, ces dérives préjudiciables à la santé sont renforcées. Même si le gain escompté peut améliorer l'alimentation des consommateurs, des considérations commerciales interfèrent nécessairement avec la volonté d'améliorer la nutrition. En dernier lieu, une attention particulière doit être portée aux populations spécifiques (enfants et adolescents, femmes enceintes, personnes âgées, végétariens…). Les repères nutritionnels destinés à la population générale ne sont pas toujours adaptés à ces catégories particulières de personnes et, malgré l'existence de recommandations nutritionnelles validées scientifiquement, les applications et objets connectés ne tiennent généralement pas compte de ces besoins spécifiques. On peut aussi s'inquiéter du fait que les adolescents pourront dès 16 ans gérer leurs données de santé sur leur espace numérique de santé et accéder à leur magasin numérique d'applications santé sans considération de leur vulnérabilité, notamment face au diktat de minceur.

Pour s'assurer des gains pour la santé des données collectées sur l'alimentation, il serait nécessaire d'évaluer et de contrôler non seulement la capacité de ces données à orienter les choix alimentaires vers les aliments plus sains, mais aussi à réduire l'incidence des maladies nutritionnelles. Si le bénéfice sanitaire de la collecte et de l'exploitation des données est parfois contrôlé par les pouvoirs publics (Nutri-Score…), les exigences concernant les applications et dispositifs développés par les acteurs privés semblent faibles. La loi du 24 juillet 2019 relative à l'organisation et à la transformation du système de santé ne prévoit que le respect de référentiels de sécurité, d'interopérabilité et d'engagement éthique (art. L. 1111-13-1 du Code de la santé publique), sans exiger la preuve d'un bénéfice sanitaire. Cela obère très largement les espoirs suscités par le développement des *big data* dans le domaine de la nutrition.

Big data en santé : quel intérêt pour les vigilances sanitaires ?

Caroline Huchet-Kervella
Pharmacienne hospitalière

La dynamique des Trente Glorieuses a permis la mise sur le marché de produits de santé de plus en plus performants, mais aussi de plus en plus complexes. Les vigilances sanitaires ont été mises en place pour répondre aux besoins identifiés par des crises successives provoquées par les effets délétères de ces produits. C'est ainsi que les professionnels de santé ont inventé des systèmes pour quantifier les risques tandis que le législateur leur a défini un cadre juridique. Actuellement, le système repose sur l'extraction de résultat à partir de bases de données. Aujourd'hui, l'enjeu consiste à savoir si le traitement automatisé des données massives est un atout à la fois pour la santé publique et pour l'aide à la décision politique.

Les vigilances sanitaires

Quel est le contexte qui favorise la prise de conscience des risques associés aux soins ?

Après la Seconde Guerre mondiale, l'augmentation massive des citoyens qui accèdent aux thérapeutiques a fait émerger la notion de risque inhérent aux médicaments et a mis en lumière la nécessité de la gestion de crise. Par exemple, la crise autour des prescriptions de la Thalidomide a mis en lumière les problématiques d'effets indésirables liés aux médicaments (iatrogénie). L'usage massif de stupéfiants et de psychotropes a permis de documenter la plasticité des circuits des neuromédiateurs et d'objectiver les problématiques de dépendance à l'arrêt des traitements. Successivement, trois conférences de l'Organisation mondiale de la santé ont formulé ces problématiques à l'échelon mondial : en 1961 sur la dépendance aux stupéfiants, en 1963 autour de la gestion de la crise de la Thalidomide, en 1972 sur la dépendance aux psychotropes. La réponse scientifique et étatique à ces crises a été de mettre en place des systèmes de détection et de quantification des risques : les vigilances sanitaires.

Qu'est-ce qu'une vigilance sanitaire ?

Les vigilances sanitaires sont des dispositifs de surveillance épidémiologique qui se déroulent lors de la phase commerciale des produits de santé. Parmi les principaux produits de santé, nous retenons : les médicaments, les dispositifs médicaux, les dispositifs de diagnostic *in vitro*, les produits sanguins labiles, les tissus, les cellules, les organes, les produits cosmétiques, les produits de tatouage.

En dépit de l'absence d'une définition juridique, voici une définition globale des vigilances qui découle de la définition de l'OMS de la pharmacovigilance : les vigilances sont « *la résultante de la science et des activités relatives à la détection, l'évaluation, la compréhension et la prévention des effets indésirables et de tout autre problème lié à l'utilisation des produits de santé et des pratiques de soins* ».

Méthodologiquement, elles sont basées sur des outils épidémiologiques de veille sanitaire autour d'un dispositif d'alerte qui caractérise et valide le signal. Ainsi, la détection, la surveillance et la caractérisation des risques s'effectuent *a posteriori*, de façon prospective et spécifique. Cette évaluation des risques des produits de santé a pour but de permettre une prescription ou une utilisation la plus pertinente possible au regard de l'état actuel des connaissances scientifiques. Les missions des vigilances sanitaires sont complémentaires des autres dispositifs d'évaluation des produits de santé. En France, le législateur a choisi d'individualiser chaque vigilance et de lui attribuer une agence sanitaire référente. La codification législative des vigilances est un processus long de quatorze années qui a débuté en 1994.

L'objectif d'une vigilance est *in fine* de collecter le maximum de données vérifiables et inattendues sur les produits de santé une fois ceux-ci commercialisés. Cela vient en complément des études préalables à la commercialisation, qui visent à écarter des produits de santé présentant trop d'effets indésirables pour la santé humaine. La compilation de ces données est nécessaire pour établir les recommandations les plus adéquates au regard des connaissances scientifiques. Par retour d'information, les professionnels de santé peuvent améliorer la pertinence de leurs actes et des schémas de soins.

Que détecte une vigilance sanitaire ?

Les vigilances sanitaires détectent un effet indésirable ou un incident imputable à un produit de santé. Dans un premier temps, la déclaration ascendante du professionnel de santé ou du patient vers la structure experte a pour but de caractériser le signalement. À ce stade, le signal est validé et un degré d'imputabilité lui est attribué par des experts de la vigilance. Au niveau national, les agences sanitaires dédiées à chaque vigilance analysent l'ensemble des signalements. Lors de la mise en évidence de signaux convergents, l'agence émet alors une alerte descendante à destination des professionnels qui détaille la conduite à tenir adéquate en fonction du problème identifié.

Quelles sont les principales limites des vigilances sanitaires ?

La principale faiblesse du système réside dans son manque d'exhaustivité structurelle. En effet, il existe bien

une obligation légale (loi n° 2002-303 du 4 mars 2002) de déclaration des effets indésirables qui s'applique à l'ensemble des professionnels de santé. Contrairement au secteur libéral, l'effectivité de la mise en place des vigilances au sein des établissements de santé est un élément recherché depuis la première version de la certification. Par ailleurs, les vigilances figurent dans les contrats d'amélioration de la qualité et de l'efficience des soins (Caques), qui lient les établissements de santé et les agences régionales de santé depuis 2016. Ainsi, pour le secteur hospitalier, les vigilances sont abordées comme outil managérial au service de l'amélioration de la qualité des soins. Toutefois, il existe une vision janusienne des vigilances. D'un côté les experts, qui analysent les situations cliniques et rendent comptent aux agences. De l'autre les cellules qualités des établissements, qui analysent le *feed-back* négatif de la déclaration pour améliorer la qualité des soins. Par ailleurs, le fait de favoriser les échanges entre ces deux types d'expertise pourrait mettre en lumière des leviers managériaux motivationnels qui font sens.

Actuellement, l'acte de déclarer un effet indésirable ou un incident n'est pas assez valorisé dans le système de soins pour ce qu'il est. La capacité à détecter un effet indésirable est avant tout la capacité des professionnels de santé à questionner la situation dans laquelle se retrouve le patient par rapport à celle escomptée. C'est utiliser le doute comme moteur pour agir, comme le soulignent les travaux de la sociologue Anne-Chantal Hardy. Les résultats issus de l'enquête nationale sur les événements indésirables liés aux soins (Eneis) de 2004 et 2009 [41] mettent en évidence une sous-déclaration massive des événements indésirables comparable à ce qui est décrit dans les autres pays européens et nord-américains. Pour le manager de santé, ces travaux mettent en avant un manque de motivation pour agir, ce qui insinue une perte de sens probable pour l'action de déclarer.

Les références entre crochets renvoient à la Bibliographie générale p. 57.

Quels sont les principaux biais à la déclaration ?

Le mécanisme du doute, qui conduit à la remise en cause d'une étape du processus de soins, doit passer au travers de différents biais qui conduiront le professionnel à déclarer.

Le plus facilement appréhendable est le biais d'identification. En effet, pour mettre en lumière un effet indésirable, il faut l'identifier et établir un diagnostic différentiel qui fait soupçonner le lien causal entre le produit de santé et l'effet. En particulier, s'il est facile de mettre en évidence des effets indésirables fréquents décrits dans la littérature, il devient moins évident d'identifier la connexion de symptômes apparemment non reliables entre eux. En plus du biais d'identification, Anne-Chantal Hardy [29] identifie deux biais principaux à l'origine de la sous-déclaration.

Tout d'abord, le biais de connaissance qui est à l'origine d'un sentiment de trahison car il entraîne une remise en cause du savoir académique et de la transmission par les pairs. Ce sentiment de trahison est d'autant plus inconfortable que le professionnel a également le sentiment qu'il trahit la confiance que le patient lui a accordée.

Pour évacuer cet inconfort, un autre biais entre en jeu : celui de la minimisation, par lequel le professionnel dévalorise la plainte du patient. C'est comme si le fait de ne pas accorder de la valeur à la parole diminuait l'intensité de l'effet indésirable.

Enfin, il existe un biais structurel de sous-déclaration du fait de l'absence de valorisation et de reconnaissance de la mise en évidence d'effets indésirables au niveau des structures de soins. En effet, en dépit de la mise en place de la déclaration en ligne des effets indésirables, la déclaration exhaustive pour permettre une analyse pertinente prend du temps pour lequel il n'est pas prévu de contrepartie financière.

Bases de données et vigilances sanitaires

Quel est l'intérêt d'avoir recours aux bases de données massives dans le cadre des vigilances sanitaires ?

L'intérêt principal d'un recours aux données massives serait d'avoir accès à des données complémentaires pour renforcer la prise de décision. Aussi, le but serait d'identifier tous les signaux, y compris ceux que les biais précédemment évoqués ne permettent pas de déceler.

Au niveau national, une telle approche permettrait d'accroître la force des recommandations de la Haute Autorité de santé (HAS), mais aussi celles des sociétés savantes. Au niveau local, elle permettrait de mettre à disposition des professionnels de santé des informations plus complètes et transparentes au sujet des produits de santé.

Voici deux exemples qui illustrent les types de requêtes envisageables pour compléter les données manquantes. Dans le champ de la médecine libérale, un raccourcissement du délai entre deux consultations ou l'arrêt d'une thérapeutique généralement prescrite au long cours sont deux marqueurs de l'inattendu dans le processus de soins. Traiter de telles données pourrait faire remonter des signaux pertinents. Lors d'une prise en charge hospitalière, la présence d'un effet indésirable lié aux soins peut se détecter lorsque l'allongement de la durée de séjour est statistiquement significatif par rapport à celle attendue et définie par la tarification à l'activité dans les données agrégées des groupes homogènes de malades. Ainsi, au niveau de l'établissement de santé, il peut aussi être envisagé de définir à l'avance des mots-clés qui peuvent être le signe de la présence d'un effet indésirable. Une fois ce thésaurus défini, il serait à la base des requêtes dans les comptes rendus d'hospitalisation issus des résumés d'unités médicales et générés à chaque fin d'hospitalisation. Dans ces deux exemples, l'automatisation du traitement ne va toutefois pas de soi. La requête statistique générerait des séjours qu'il faudrait analyser manuellement pour confirmer le caractère iatrogène. Au vu des biais cognitifs évoqués

plus haut, il semble ardu de bâtir une intelligence artificielle capable de prendre en compte l'implicite dans les comptes rendus médicaux. L'augmentation significative des dossiers à analyser générerait d'importants besoins en ressources humaines pour pouvoir soit analyser, valider et imputer la déclaration, soit la réfuter.

En revanche, au niveau national, des projets de systèmes algorithmiques mobilisant les données massives au service de la pharmacovigilance sont déjà en cours de développement. En effet, cette vigilance est idéale pour mettre en place ce type de programme. Cela s'explique par une diffusion à large échelle des médicaments, par la connaissance des effets indésirables lors des tests qui précèdent la commercialisation. Enfin, les médicaments font partis des produits de santé qui présentent la période d'exploitation la plus longue. Depuis 2013, il est possible pour les chercheurs d'exploiter les données contenues dans le Système national inter-régimes de l'Assurance maladie (Sniiram) et désormais dans le Système national des données de santé (SNDS). L'objectif est de pouvoir répondre à des problématiques non solvables lors des études précliniques : études des signaux non signalés par les professionnels de santé, études chez des populations particulières (personnes sénescentes, grossesses) ou pour réaliser des études de pharmaco-épidémiologie. Par ailleurs l'avantage secondaire de ce type de recherche est de permettre de répondre aux problématiques de disparités territoriales de prescription et de documenter à l'échelle macro les différents leaders d'opinion de prescription. À partir de l'exploitation de ces données, des études pluridisciplinaires, associant des sociologues, pourraient mieux documenter les relations soignants-soignés autour de certaines thérapeutiques.

Quels sont les principaux obstacles méthodologiques ?

Dans le cadre de l'utilisation des bases de données, la définition de la faisabilité de la problématique est une étape fondamentale. En effet, l'architecture de la base de données permet de savoir quelles problématiques peuvent être résolues, en les distinguant de celles pour lesquelles il est impossible d'obtenir une réponse. Après validation de la problématique, il est nécessaire de retraiter les données pour permettre un usage statistique. Dans une démarche qualité (type ISO 9001), cette opération doit être décrite au préalable. Par ailleurs, la non-exhaustivité de certaines données est l'un des problèmes lors de la mise en place de base de données massives. Toujours dans une démarche qualité et en vue d'expliciter les possibles biais d'analyse, il est utile de préciser les modalités qui conduisent à ne pas retenir certaines données. L'idéal est donc de composer des équipes pluridisciplinaires (médecin, pharmacien, ingénieur, statisticien) autour d'une problématique commune.

Quels sont les principaux obstacles managériaux ?

L'arrivée des données massives dans les pratiques courantes va impliquer d'importants changements dans les pratiques des professionnels de santé. En effet, pour beaucoup de nos concitoyens, la méconnaissance de l'intelligence artificielle et des bases de données conduit à une réticence réactionnelle aux changements. Dans un contexte de perte de sens dans le travail, ces bouleversements technologiques induisent plus ou moins consciemment la peur de voir le travail de l'homme remplacé par celui de la machine. Bien que l'augmentation des données à traiter conduit à une probable augmentation des besoins en professionnels experts, les leviers motivationnels et la conduite du changement passent également par la prise en compte de la perte de sens au travail, qui s'étend au secteur de la santé et qui s'est accéléré avec la mise en place de la tarification à l'activité à partir de 2008. Les travaux de Nathalie Angelé-Halgand et Thierry Garrot [1] analysent les mécanismes de contrôle des ressources humaines pour satisfaire aux objectifs d'équilibre budgétaire. Ils font le parallèle entre l'introduction du New Public Management dans le monde de la santé et les mécanismes de régulation du panoptique de Bentham décrit par Michel Foucault [28]. Ils introduisent la notion de perte d'initiative des équipes soignantes, et par extension de sens, dans la prise en charges des patients. Les injonctions de devoir rendre compte budgétairement en temps quasi réel induit un biais d'attention des équipes soignantes. Cela se traduit jusque dans la façon de concevoir les relations entre soignants tout comme la relation soignant-soigné.

tribunes

Le SNDS, un outil au service des acteurs de terrain

L'exemple de l'étude du recours aux soins dentaires en Pays de la Loire

La connaissance des données de santé d'une population ou d'un territoire permet de faire émerger des besoins de santé et d'améliorer les parcours de soins proposés. Cet article présente l'utilisation des données de soins dentaires dans les Pays de la Loire.

Marie Dalichampt
Anne Tallec
Jean-François Buyck
Observatoire régional de la santé (ORS) des Pays de la Loire

La mise en place en 2017 du Système national des données de santé (SNDS) offre de nouvelles perspectives en matière d'action de santé publique. Ce système d'information permet en effet des analyses détaillées du recours aux soins et à la prévention, avec l'élaboration d'indicateurs déclinables par type de population, ou par niveau territorial fin. Ces analyses ouvrent de façon considérable le champ des possibles en matière d'identification des besoins, s'agissant de parcours de santé, de ciblage des populations ou de territoires prioritaires, puis de suivi des évolutions dans une logique évaluative.

Compte tenu de ces nouveaux enjeux, l'observatoire régional de la santé (ORS) des Pays de la Loire – qui comme l'ensemble des ORS dispose d'un accès large et permanent aux données du Système national des données de santé – s'est investi dans ce domaine en recrutant et formant son équipe à l'utilisation de ce système d'information particulièrement complexe.

En parallèle, afin de s'inscrire dans une logique d'action, mais aussi de mobiliser l'expertise métier indispensable à l'exploitation de ces données, notamment celles des bases de l'Assurance maladie, l'ORS a développé des collaborations avec les unions régionales des professionnels de santé libéraux (URPS), avec lesquelles il entretient des partenariats de longue date autour d'enquêtes sur les pratiques et conditions d'exercice [1].

Cette approche a rencontré une demande de l'URPS chirurgiens-dentistes des Pays de la Loire, qui souhaitait disposer d'une connaissance fine du recours au cabinet dentaire des enfants de la région. Une première étude a été produite en 2018 [2], et cette dynamique, à laquelle s'est alors associée l'Union française pour la santé bucco-dentaire (UFSBD), s'est poursuivie en 2019-2020 autour de trois nouvelles études, portant sur le recours des adultes âgés de 55 ans et plus [3], des personnes diabétiques [4], et des personnes traitées par biphosphonates (en cours). Pour chacune de ces études, un groupe de travail associant praticiens de terrain et spécialistes du Système national des données de santé a été mis en place pour élaborer des indicateurs pertinents et directement en lien avec les pratiques des professionnels de santé.

Un recours aux soins dentaires très en deçà des recommandations

Les analyses déjà réalisées sur trois des populations choisies par l'union régionale des professionnels de santé libéraux (enfants, seniors, personnes diabétiques) ont toutes montré un recours très insuffisant au cabinet dentaire, au regard des recommandations de la Haute Autorité de santé (HAS) d'un recours annuel minimum [5].

Le taux de recours annuel est très en deçà des 100 % souhaités : 61 % chez les 6-18 ans [2], 40 % chez les personnes diabétiques [4], et 47 % parmi les plus de 55 ans [3]. Chez ces derniers, le taux annuel de recours au cabinet dentaire décroît de façon continue à partir de 65 ans, et n'est plus que de 25 % au-delà de 90 ans.

De plus, le chaînage des données du Système national des données de santé, qui permet d'analyser l'ensemble des prestations de chaque bénéficiaire, met en évidence l'ampleur du non-recours sur plusieurs années

consécutives. En Pays de la Loire, près d'un enfant sur dix n'a bénéficié d'aucun recours bucco-dentaire préventif (ni examen bucco-dentaire [EBD], ni consultation, ni détartrage) entre 6 ans et 9 ans, et cette proportion atteint 25 % entre 14 ans et 17 ans. Et ce malgré le programme M'T dents de l'Assurance maladie, qui propose un examen bucco-dentaire sans avance de frais aux âges de 6, 9, 12, 15 et 18 ans, et à 3 ans depuis le 1er janvier 2019.

Chez les seniors, la situation est également très défavorable : 25 % des Ligériens âgés de 55 à 70 ans n'ont eu aucun recours au cabinet dentaire sur les années 2016 à 2018, et cette proportion dépasse 50 % au-delà de 90 ans.

Enfin, pour les personnes diabétiques, dont l'état de santé bucco-dentaire est étroitement lié au risque de déséquilibre et de complications du diabète, les résultats sont alarmants : plus d'un tiers des Ligériens pris en charge pour un diabète en 2015 n'ont eu aucun recours au cabinet dentaire au cours des trois années suivantes (2016-2018), et seulement 16 % ont eu un parcours conforme aux recommandations, c'est-à-dire au moins une consultation chacune des trois années.

Ces résultats illustrent bien l'importance des enjeux : malgré des recommandations anciennes, relayées par de nombreux acteurs de santé (Assurance maladie, sociétés savantes et associations d'usagers) et préconisant des soins bien remboursés par l'assurance maladie obligatoire, une part importante de la population a un recours aux soins dentaires très insuffisant, alors que l'impact de la santé bucco-dentaire sur la santé générale est désormais bien établi.

Des publics et territoires encore plus prioritaires que d'autres

Un grand nombre d'études ont montré que la santé bucco-dentaire constitue un excellent marqueur des inégalités sociales de santé. Les indicateurs étudiés ici, issus du Système national des données de santé, confirment ce constat. Chez les enfants, les différences les plus marquées concernent la fréquence du suivi préventif et l'âge du premier recours au cabinet dentaire : 30 % des enfants bénéficiant de la couverture maladie universelle complémentaire [CMU-C] n'ont jamais eu de recours avant 7 ans, contre 16 % de ceux qui n'en bénéficient pas [2]. Chez les personnes âgées de 55 ans et plus, la proportion de celles n'ayant eu aucun recours en trois ans atteint 43 % chez les bénéficiaires de la CMU-C ou de l'aide au paiement d'une complémentaire santé (ACS)[1], contre 29 % chez les personnes qui n'en bénéficient pas, à structure par âge équivalente [3]. Les écarts de recours selon le niveau social sont particulièrement importants pour les poses de prothèse fixe, soins à fort reste à charge pour la période concernée par l'étude, mais également pour les détartrages, pourtant bien remboursés.

Les enfants admis en affection de longue durée (ALD), qui sont le plus souvent atteints de maladies chroniques et sont pour certains en situation de handicap, présentent des indicateurs de recours encore plus dégradés que les autres enfants. Ainsi, 17 % n'ont eu aucune prestation de suivi bucco-dentaire entre 6 et 9 ans (10 % des enfants sans ALD). Lorsqu'un traitement orthodontique est commencé, il l'est plus tardivement chez les enfants en ALD, ce qui peut le rendre moins efficace : 42 % le débutent avant 10 ans et 32 % après 13 ans (contre respectivement 48 % et 25 % chez les enfants sans ALD).

Chez les personnes âgées de 75 ans et plus, pour lesquelles le recours au cabinet dentaire est globalement très insuffisant, le fait de résider ou non en établissement d'hébergement pour personnes âgées dépendantes (Ehpad) est un des principaux facteurs explicatifs d'une augmentation du risque de non-recours [3]. En effet, les analyses multivariées, ajustant sur l'âge, les caractéristiques sociales, l'état de santé et le niveau d'accessibilité potentielle localisée (APL) au chirurgien-dentiste libéral, montrent que l'association entre le non-recours au cabinet dentaire pendant au moins trois ans et le fait de résider en Ehpad est très significative, avec un *odds ratio* de près de 1,5 chez les personnes nouvellement arrivées en Ehpad, et qui s'élève à 2,5 chez celles hébergées depuis au moins deux années, comparées aux personnes vivant à leur domicile [3]. Ce résultat peut en partie être expliqué par un plus grand degré de dépendance des résidents en Ehpad, mais aussi par un éloignement à leur chirurgien-dentiste habituel du fait du déménagement vers l'Ehpad.

La déclinaison des différents indicateurs pour des zonages géographiques fins met par ailleurs en évidence des disparités territoriales de recours très marquées. Selon leur établissement public de coopération intercommunale (EPCI), la part des enfants qui bénéficient d'un parcours préventif bucco-dentaire régulier entre 6 et 9 ans varie de 30 à 60 % [2]. Les enfants domiciliés dans les établissements publics de coopération intercommunale de Sarthe, et plus particulièrement dans les intercommunalités les plus éloignées de la métropole du Mans, ont un recours globalement moins fréquent, moins précoce et moins régulier comparés aux enfants des autres établissements publics de coopération intercommunale de la région. À l'inverse, la Loire-Atlantique et la Vendée (où est né dans les années 1980 le bilan bucco-dentaire auquel a succédé le programme M'T dents) englobent la plupart des établissements publics de coopération intercommunale où les fréquences de parcours préventif chez les enfants sont les plus élevées. La Loire-Atlantique et la Vendée concentrent également la grande majorité des établissements publics de coopération intercommunale où le recours aux soins dentaires est plus satisfaisant au-delà de 55 ans, et chez les personnes prises en charge pour un diabète [3, 4]. Pour ces dernières, les écarts entre territoires sont considérables, avec une part de personnes diabétiques ayant eu un recours satisfaisant entre 2016 et 2018, c'est-à-dire chacune des trois années, qui varie de 5 à 24 %.

Agir, notamment localement dans le cadre des dynamiques interprofessionnelles

Les acteurs susceptibles de se saisir des données du Système national des données de santé pour mettre en place des démarches visant à améliorer ces parcours sont multiples (et non exclusifs les uns des autres).

Au plan national, l'Assurance maladie utilise depuis de nombreuses années le relais que constituent ses caisses locales pour mettre en œuvre et suivre, à partir de son système d'information (qui alimente le SNDS), des programmes de dépistages organisés et de prévention. L'invitation de chaque enfant aux examens bucco-dentaires gratuits du programme M'T dents en fait partie. Cette démarche pourrait également être étendue à des populations adultes cibles (par exemple à l'entrée en Ehpad, comme préconisé par l'Union française pour la santé bucco-dentaire).

Au plan régional, les URPS, qui ont pour mission de contribuer à l'organisation de l'offre de soins et à la politique régionale de santé, aux côtés des agences régionales de santé (ARS), peuvent également s'appuyer sur les données du SNDS pour faire émerger des projets d'actions. Les travaux sur le recours au cabinet dentaire menés

1. Depuis la réalisation de ces études, la CMU-C et l'ACS ont été remplacées par la complémentaire santé solidaire (CSS/C2S).

par l'ORS à la demande et avec le soutien de l'union régionale des professionnels de santé libéraux chirurgiens-dentistes en sont un premier exemple.

Au niveau local, le développement des coopérations interprofessionnelles et la mise en place récente des CPTS[2] (communautés professionnelles territoriales de santé) constituent de formidables opportunités. En effet, la mobilisation coordonnée des différents professionnels de santé, tant pour orienter les patients que pour partager de l'information à leur propos, est l'une des conditions de l'amélioration des parcours. L'étude concernant le recours au cabinet dentaire des personnes diabétiques montre, par exemple, que la proportion de celles ayant un parcours dentaire satisfaisant est sensiblement plus élevée parmi les personnes qui consultent régulièrement leur médecin généraliste, après ajustement sur les autres facteurs [4]. Cette association, bien qu'elle ne démontre pas de rapport de cause à effet, suggère le rôle central que peuvent jouer les généralistes dans l'adhésion de leurs patients aux recommandations de suivi bucco-dentaire. De même, il est important que le chirurgien-dentiste connaisse l'existence du diabète de son patient pour adapter sa prise en charge et lui rappeler l'importance d'un suivi régulier.

Les communautés professionnelles territoriales de santé, auxquelles a été confiée une responsabilité populationnelle, offrent désormais un cadre pertinent à ces dynamiques d'amélioration des parcours basée sur l'élaboration et le suivi d'indicateurs concernant les habitants de leur territoire.

Les indicateurs élaborés avec les professionnels de santé grâce aux données du Système national des données de santé permettent de définir des objectifs d'amélioration des parcours de soins concrets et adaptés au contexte local. Parce qu'ils correspondent à leur pratique au quotidien, ces indicateurs facilitent la mobilisation des professionnels de santé pour participer à des actions en vue d'atteindre ces objectifs. Les indicateurs peuvent cibler finement des populations en fonction de leur lieu d'habitation, de leur âge, de leur parcours de vie (entrée en Ehpad), de leur état de santé (ALD, diabète) ou encore de leur traitement (biphosphonates).

Enfin, la possibilité d'évaluer de façon rapide et fiable l'impact des actions menées en mesurant l'évolution des indicateurs du Système national des données de santé est d'un intérêt majeur. En effet, des résultats positifs permettent à la fois de valider la pertinence des actions, mais aussi de favoriser le maintien de la motivation des professionnels.

Quels enjeux pour les années à venir ?

La connaissance fine des données de santé d'une population ou d'un territoire est nécessaire pour faire émerger des besoins de santé et identifier des leviers d'amélioration. Mais cette connaissance n'est pas suffisante. Encore faut-il ensuite que les acteurs opérationnels, et notamment les professionnels de santé, s'en saisissent pour initier et mener des actions.

En Pays de la Loire et en matière de soins bucco-dentaires, une part du chemin a été parcourue à travers la mobilisation de l'union régionale des professionnels de santé libéraux chirurgiens-dentistes et de l'ORS autour de ces travaux, qui illustrent l'important besoin d'amélioration du recours aux soins. Mais un effort considérable de pédagogie et de communication autour de ces travaux doit encore être accompli, pour convaincre les différentes parties prenantes de la nécessité de mettre en place des actions, et parvenir à mobiliser les moyens humains et financiers que de telles dynamiques impliquent.

Ne pas tirer profit de l'apport des données du SNDS serait un immense gâchis. Ces données, mondialement enviées, constituent un levier majeur pour l'amélioration des parcours de soins, la réduction des inégalités sociales et territoriales de santé, et plus largement l'appropriation des enjeux de santé publique par les professionnels de santé.

2. Les CPTS émanent de l'initiative des acteurs de santé, en particulier des professionnels de santé de ville. Ce sont des équipes projets, s'inscrivant dans une approche populationnelle au sens où les différents acteurs acceptent de s'engager dans une réponse à un besoin de santé de leur territoire, qui peut impliquer pour eux de prendre part à des actions ou d'accueillir des patients sortant de leur exercice et de leur patientèle habituelle (instruction n° DGOS/R5/2016/392 du 2 décembre 2016 relative aux équipes de soins primaires [ESP] et aux communautés professionnelles territoriales de santé).

Références bibliographiques

1. ORS Pays de la Loire. Enquêtes et panels professionnels de santé [en ligne]. https://www.orspaysdelaloire.com/enquetes-et-panels-professionnels-de-sante
2. ORS Pays de la Loire, URPS chirurgiens-dentistes Pays de la Loire. *Recours au cabinet dentaire des enfants et des adolescents. Situation en Pays de la Loire et en France à partir d'une analyse des données du SNDS*. 2018, 76 p.
3. ORS Pays de la Loire, URPS chirurgiens-dentistes Pays de la Loire. *Recours au cabinet dentaire des adultes de 55 ans et plus. Situation en Pays de la Loire et en France à partir d'une analyse des données du SNDS*. 2019, 72 p.
4. ORS Pays de la Loire, URPS chirurgiens-dentistes Pays de la Loire. *Suivi bucco-dentaire des personnes diabétiques en Pays de la Loire à partir d'une analyse des données du SNDS*. À paraître en 2020, 28 p.
5. HAS. *Stratégie de prévention de la carie dentaire. Synthèse et recommandations*. 2010, 26 p.

tribunes

Big data, data reuse en santé : un chemin semé d'embûches nécessitant une approche pluridisciplinaire

Illustration de la complexité d'utilisation de données de santé et de la nécessaire collaboration de plusieurs professions.

Les références entre crochets renvoient à la Bibliographie générale p. 57.

Nous vous entraînerons avec nous dans un exemple de recherche ayant réussi. L'objet de cette recherche était de mettre en place un système d'intelligence artificielle (IA) permettant de prévenir les effets indésirables du médicament (EIM). Nous verrons que le chemin fut parsemé d'embûches. Ces embûches illustrent l'impossibilité de faire accomplir un tel travail par une machine seule, et l'absolue nécessité d'approches pluridisciplinaires.

Contexte et objectifs

En France, chaque hospitalisation de patient donne lieu à la collecte d'informations codées tels les diagnostics (par exemple « K37 » pour certaines appendicites) et les actes (par exemple « HHFA001 » pour certaines ablations chirurgicales de l'appendice). De plus, dans la plupart des hôpitaux, les médicaments sont prescrits *via* des logiciels et les résultats d'analyses de biologie médicale sont également transmis au service prescripteur par des logiciels spécifiques. Toutes ces données constituent des données massives ou *big data* (voir l'encadré de définitions) [3] : on dénombre en moyenne 100 résultats de biologie médicale par séjour (par exemple le taux d'hémoglobine) et 200 000 séjours par an dans certains CHU. Ces données servent respectivement à facturer le séjour à l'Assurance maladie et à soigner le patient. Elles

Quelques définitions simplifiées

Données massives/*big data*
Données volumineuses, comportant par exemple de nombreux individus et/ou de nombreuses variables, quelles que soient leur origine et leur exploitation.

Réutilisation de données/*data reuse*
Fait d'utiliser (pour de la recherche par exemple) des données qui ont initialement été collectées dans un autre but (le soin par exemple).

Intelligence artificielle (IA)
Fait de doter l'ordinateur de capacités visant à mimer le résultat d'une intelligence. Une IA peut s'appuyer sur un répertoire de règles écrites par un humain, un apprentissage automatisé (*machine learning*), ou un apprentissage par renforcement.

Effet indésirable du médicament (EIM)
Effet secondaire nocif et non voulu lié à la prise, la modification de dose ou l'arrêt d'un médicament (hormis erreurs d'administration et tentatives de suicide).

Emmanuel Chazard
Professeur des universités, faculté de médecine, Cerim ULR 2694, université de Lille, praticien hospitalier, CHU de Lille

peuvent cependant être réutilisées à des fins de recherche, on parle alors de réutilisation de données ou *data reuse*.

Notre objectif ici est de les réutiliser pour détecter automatiquement les circonstances causant des EIM. Une fois ces circonstances identifiées, nous bâtirons un logiciel d'IA [10] qui « surveillera » les prescriptions médicamenteuses et alertera le prescripteur en cas d'élévation du risque d'effet indésirable du médicament. Cet énoncé illustre le lien qu'entretiennent parfois *big data*, *data reuse* et intelligence artificielle. Cela dit, ce lien n'est pas obligatoire. La plupart du temps, le grand public parle de *big data* pour désigner le *data reuse* parce que les jeux de données dont la réutilisation est la plus prometteuse sont généralement de grande taille. Passées ces définitions, lançons-nous dans notre recherche !

Notre recherche, pas à pas

Construction d'un entrepôt de données

La première étape est « d'aspirer » les données disponibles et de les nettoyer (indifféremment de l'objectif). Copiées dans une nouvelle base de données plus simple, elles constituent un « entrepôt de données ». Première surprise, un certain nombre d'enregistrements sont « orphelins » et ne peuvent pas être rattachés à un séjour hospitalier… la complexité des systèmes d'information hospitaliers (SIH) est telle, qu'il est difficile de savoir si le lien existe mais est perdu lors de l'extraction, ou si l'information est d'emblée corrompue. Le chercheur, quant à lui, devra renoncer à tout analyser. Ensuite, les données comprennent de nombreuses erreurs. En voici un exemple : dans un hôpital, le libellé « hématies » est associé à des valeurs comprises entre 4 000 et 5 500/mm³. C'est mille fois trop faible, pourtant ces patients sont tous vivants. L'unité enregistrée est tout simplement fausse : pourtant, le médecin interprète le chiffre précisément mais sans regarder s'il s'agit de milliers ou millions, et parfois sans connaître précisément l'unité attendue. D'autres patients également ont un taux de potassium tournant aux alentours de 30 mmol/l, la norme devant rester inférieure à 4,5. Sont-ce des momies ? Non, simplement des patients dont on a dosé le taux de potassium urinaire, mais sans l'écrire clairement dans la base de données. Ces types d'erreurs n'ont généralement aucun impact sur le soin, car le soignant les corrige mentalement sans même s'en rendre compte. Elles sont critiques lorsqu'on analyse les données.

Leçon apprise

La construction d'un entrepôt de données requiert des informaticiens, mais aussi des médecins capables d'évaluer la qualité des données, et des analystes capables de dire quelles données sont indispensables ou peuvent être sacrifiées.

Analyse de données

Nos données sont enfin prêtes. Mais qu'en faire ? Certes, la construction d'un entrepôt de données nous a permis de passer de plus de mille tables de données d'un SIH à quelques dizaines de tables, mais les données restent complexes : hétérogènes (diagnostics, actes, médicaments, résultats d'analyses…), presque toujours manquantes et presque jamais par hasard (très peu de patients bénéficient d'une IRM de l'hypophyse… parce que leur hypophyse est normale), très complexes (près de 40 000 codes différents pour décrire les maladies, et encore cela ne suffit pas), et plus ou moins fiables !

Et pourtant, observons un médecin parcourir un dossier médical (figure 1) :

« Ce patient fait probablement une hémorragie : son **INR [International Normalized Ratio, l'un des indicateurs de la coagulation sanguine] augmente**, *témoignant d'une activité anticoagulante trop forte, et le lendemain son* **taux d'hémoglobine diminue**, *traduisant sans doute une hémorragie. Un* **antivitamine K** *(anticoagulant) était administré juste avant, et vraisemblablement potentialisé par du* **paracétamol à forte dose**. *Le médecin corrige ensuite cela en* **arrêtant l'anticoagulant** *et en* **introduisant de la vitamine K**, *son antidote. Par la suite, l'*INR **se normalise.** »

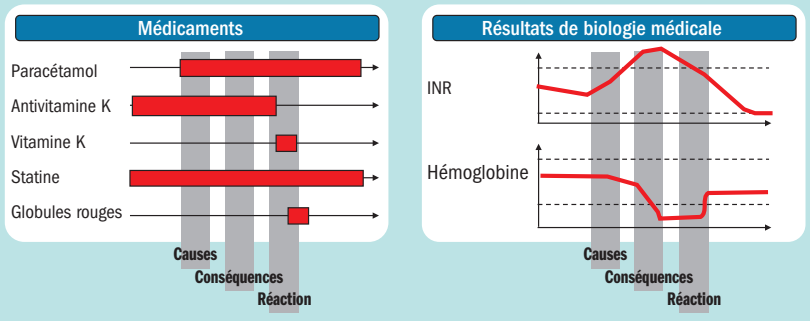
Quoi qu'on en dise, aucun ordinateur n'est capable de tenir un tel raisonnement. L'analyse de ce raisonnement nous permettra néanmoins de définir des « caractéristiques » (ou *features*) qu'il faudra calculer [13]. Il s'agit simplement d'appliquer tout un tas d'algorithmes (décidés par un expert) pour créer autant de variables simplifiées qu'il y avait de constatations **soulignées** ci-dessus, et plus encore par généralisation. Ces caractéristiques se présentent de manière plus simple que les données réelles disponibles, et elles sont « optimisées » pour porter du sens. Ainsi, grâce à l'extraction de caractéristiques, nos algorithmes de *machine learning* ne seront plus comme une poule face à un couteau.

Leçon apprise

L'extraction de caractéristiques nécessite en amont une excellente connaissance des données, une maîtrise de l'algorithmique de base, et en aval une connaissance des formes de données qui « fonctionnent » en *machine learning*.

figure 1

Interprétation de données brutes d'un patient qui présente une hémorragie sous anticoagulants



Lecture : le temps va de gauche à droite ; le commentaire est dans le texte

Prédiction par *machine learning*

L'extraction de caractéristiques nous a permis de construire un grand tableau, comprenant une seule ligne par patient, et des milliers de colonnes, représentant des variables simples. Nous avons supprimé la complexité des données. Il est à présent aisé d'utiliser des techniques d'apprentissage automatisé (*machine learning*) pour prédire automatiquement certaines variables (par exemple présenter une hémorragie) à l'aide de toutes les autres (avoir un anticoagulant, l'âge, le sexe, etc.). À ce jeu-là, les réseaux de neurones sont plutôt doués. Hélas, ils construisent pour ce faire des formules mathématiques de plusieurs pages, incompréhensibles par un humain. Cela leur vaut d'être qualifiés de « boîte noire ». En pratique, dans notre projet, ils s'avèrent inutilisables : comment alerter un médecin sur un risque, sans même être capable de livrer des arguments validés scientifiquement ? Nous préférons donc une technique moins vendeuse d'un point de vue marketing, mais plus lisible, comme les arbres de décision (figure 2). Les premiers arbres nous apprennent que… les patients plus âgés meurent davantage. Ce résultat était peut-être connu dès la Préhistoire ! Après des corrections de variables, les arbres suivants nous apprennent par exemple que le plus fort facteur de risque d'hyperglycémie, c'est d'avoir de l'insuline. Or l'insuline entraîne des hypoglycémies. C'est pourtant évident : association statistique ne signifie pas causalité. Il faut mettre en place des procédures automatisées et expertes de filtrage des associations découvertes par la machine. À la fin, nous tenons le bout (figure 2) : ainsi par exemple, sur sept patients présentant une insuffisance rénale et âgés de plus de 85 ans et ayant de la spironolactone, six (85 %) présentent ensuite une hyperkaliémie. L'association est techniquement, statistiquement et bibliographiquement valide. Pourtant, la relecture des dossiers nous apprend que seulement la moitié des pathologies sont réellement des effets imputables au médicament, car il arrive tout un tas d'autres choses à ces patients, invisibles dans les données.

Leçon apprise

La prédiction par *machine learning* n'est donc pas seulement un jeu de statisticien. Elle nécessite, outre l'extraction de caractéristiques, une forte expertise métier, des résultats analysables et critiquables par un humain, et une validation par retour aux cas réels.

Prévention en vie réelle

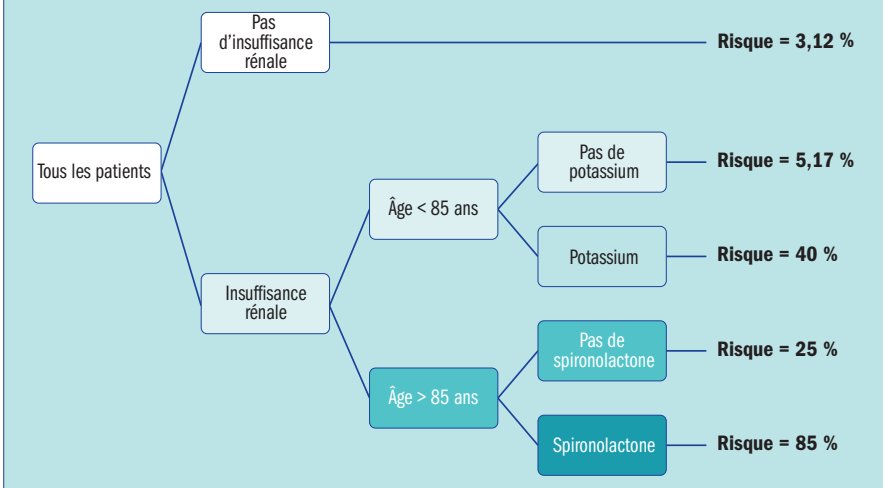
Au bout du processus de *machine learning*, nous tenons enfin un lot de 256 règles permettant de prédire la survenue d'EIM. Il « suffira » de les intégrer dans un système d'aide à la décision connecté au logiciel de prescription. Nous apprenons au passage que le niveau de risque (pourcentages dans la figure 2) varie fortement d'un service à l'autre. Plus généralement, les médecins maîtrisent très bien les médicaments de leur spécialité, et pour eux les alertes sont totalement inutiles : le fait est qu'elles leur pourrissent la vie sans diminuer les erreurs ! Inversement, c'est plutôt sur les situations plus rares qu'ils peuvent se tromper. Ainsi, par exemple, on ne verra jamais d'hyperkaliémie en néphrologie bien que tous les patients soient à risque, mais on peut en voir en pneumologie. Forts de ce constat, nous déployons le premier « SPC-CDSS » [12], c'est-à-dire le premier système d'aide à la décision filtré et contextualisé statistiquement. Bien que novateur et hautement valide, ce système ne sera pas utilisé par les cliniciens. Nous avions tout pris en compte, sauf le facteur humain. Bêtement, comme tant d'autres avant nous, nous avions engendré une monstruosité : un logiciel qui prend du temps à ceux qui en ont le moins, un logiciel qui rajoute des clics, des actions et des alertes à ceux auxquels tout le personnel hospitalier transfère insidieusement son travail : les médecins.

Leçon apprise

On ne peut pas se contenter de dire : « Le logiciel est bon donc ils l'utiliseront. » L'informatisation des hôpitaux a insidieusement transféré les tâches du personnel le moins qualifié vers le plus qualifié. Argumentant des économies, elle a augmenté le coût du travail pour une tâche donnée, et diminué la productivité pour certaines catégories de personnel hautement qualifié. Un bon logiciel d'IA est surtout un logiciel qui s'intégrera dans un *workflow* de manière à alléger la charge cognitive et l'agacement des professionnels qualifiés, et leur faire gagner du temps.

figure 2

Exemple d'arbre de décision (*machine learning*) prédisant automatiquement un risque d'hyperkaliémie



Conclusion

À travers ce cheminement, nous avons illustré les nombreux obstacles rencontrés, et la nécessité de faire collaborer au minimum informaticiens, statisticiens, spécialistes des facteurs humains, et surtout experts des données considérées (ici, des médecins). La vidéo citée en référence présente également ce cheminement [11].

tribunes

La position des institutions publiques françaises dans la promotion et l'utilisation des données en santé publique

La promotion et l'utilisation des données de santé sont organisées pour un service public, par des institutions qui peuvent être en concurrence.

Thomas Lefèvre
Maître de conférences, praticien hospitalier, Iris-UMR 8156-997 CNRS Inserm EHESS université Sorbonne Paris Nord
Sabine Guez
Post-doctorante, Iris-UMR 8156-997 CNRS Inserm EHESS université Sorbonne Paris Nord

Cette tribune n'a pas pour vocation de retracer une sociohistoire précise du rôle des institutions publiques françaises dans la promotion et l'utilisation des données en santé publique, mais d'en donner quelques éléments de repères et de questionnements.

Porter l'attention du privé vers le public en matière de données de santé

S'agissant du *big data*, de l'intelligence artificielle, plus largement de l'irruption du numérique dans notre quotidien, l'attention a surtout été focalisée sur les acteurs privés, comme Google, Amazon, Facebook, Apple ou Microsoft (Gafam), en particulier pour des raisons de souveraineté nationale et de protections juridiques variables et non garanties partout dans le monde de la même façon. Ces acteurs ont proposé, proposent encore régulièrement, de se positionner comme intermédiaires voire comme substituts à ce qui peut être considéré comme des fonctions régaliennes, au minimum des missions de service public. On peut penser à la fonction *safety check* de Facebook, en cas de catastrophe naturelle ou d'événement imprévu de portée collective, menaçant la vie des personnes, ou aux applications smartphone proposées par Google et Apple aux gouvernements dans le cas du *case tracking* dans le contexte de l'épidémie de Covid-19. En comparaison, la position des institutions publiques dans la promotion et l'utilisation des données n'a fait l'objet que de peu d'attention. Or en France l'organisation et l'utilisation des données en santé tiennent essentiellement à des initiatives des institutions publiques, aux niveaux gouvernemental et ministériel, depuis une quarantaine d'années.

Le pilotage médico-économique des établissements de santé basé sur la donnée

Une façon d'aborder le sujet est de partir de la création du PMSI (Programme de médicalisation du système d'information [des établissements de santé]) au début des années 1980, dont l'utilisation a été renforcée en 1996, puis en 2005. L'idée derrière le PMSI est la quantification et la standardisation d'un certain nombre d'informations ayant trait aux hospitalisations : les diagnostics, les durées de séjour. Conjointement, un travail de classification et de nomenclature se met en place, s'institutionnalise, notamment avec la création de l'ATIH (Agence technique de l'information sur l'hospitalisation) en 2000. Il s'agit d'une part d'identifier et de mettre à jour l'ensemble des GHM (les groupes homogènes de malades) à partir des données collectées dans les hôpitaux et, d'autre part, de faire coïncider une nomenclature médico-économique, sur laquelle la tarification à l'activité (la T2A) se base dès 2005. Des données concernant un grand nombre de malades et d'hospitalisations sont collectées ; par des approches algorithmiques, on identifie des groupes de malades qui se « ressemblent » tant en termes médicaux (diagnostics…) que de coûts du séjour en hôpital. Ces groupes permettent de « segmenter » la population hospitalière comme on le fait en marketing pour identifier des sous-populations de clients. La technique n'est en rien novatrice et est importée des pratiques du privé. Cela sert alors à donner

une assiette au budget d'un établissement de santé. Au sein de ces établissements, les départements d'information médicale (DIM) ont, dans beaucoup de cas, la tâche désormais principale de gérer le PMSI local, et de justifier le budget de leur établissement. Au niveau national, les données des PMSI locaux sont fédérées, consolidées.

Les données au-delà du PMSI et du Sniiram : la fédération des données, leur gestion et leur accessibilité

Le PMSI est donc une source de données historique en France. Une autre source de données elles aussi médico-administratives, gérée par une institution publique, est le Sniiram (Système national d'information inter-régimes de l'Assurance maladie). Elle est gérée par la Caisse nationale d'assurance maladie et fait partie de la base du Système national de données de santé (SNDS), regroupant de fait avant tout les données du PMSI et du Sniiram.

Simultanément, d'autres acteurs publics se sont manifestés de façon croissante quant à l'utilisation des données de santé, au-delà des données médico-administratives. On a ainsi assisté à l'émergence des entrepôts de données de santé des hôpitaux, tendant à rassembler l'ensemble des données dérivées de tous les examens et observations effectués lors d'un séjour hospitalier. Le panorama des organismes publics pouvant participer à l'écosystème de la production et de l'utilisation des données de santé en France se complète par les universités (facultés de médecine), les unités Inserm et les DIM déjà cités, ainsi que plusieurs organismes ministériels : Drees, Acoss, CNSA…

Nous avons donc ici le panorama classique d'une multitude de sources de données à réconcilier, néanmoins doublé d'une multitude d'acteurs, tous publics, aux intérêts concurrents. Cette concurrence est fréquemment réduite à l'argument d'une supposée propriété des données : l'hôpital propriétaire des données captées dans son établissement, l'Inserm propriétaire des données générées dans un cadre de recherche, la faculté propriétaire des données recueillis par son personnel hospitalo-universitaire – lui-même fréquemment affilié à une unité Inserm et exerçant en établissement hospitalier. Une solution organisationnelle et technique proposée pour une forme de convergence de ces sources de données, solution dont la teneur réelle reste néanmoins à préciser à l'épreuve de la pratique et du temps, s'incarne depuis le 30 novembre 2019 dans la plateforme de données de santé, le Health Data Hub, constitué en groupement d'intérêt public (GIP). Conçu comme un guichet unique facilitant l'accès et l'utilisation des données pour la santé pour différents acteurs, le Health Data Hub cristallise cependant plusieurs critiques d'autant plus aisément qu'il apparaît comme une personne identifiable plutôt que la multitude des acteurs. Ces critiques sont liées aux aspects de la nouvelle gestion publique (*new public management*). Un exemple est le recours à l'externalisation pour des tâches ou des missions importantes, comme l'hébergement des données attribué à Microsoft. Un autre exemple est celui de la conformité avec le référentiel de sécurité Système national des données de santé, qui par ricochet réglementaire, est applicable ou demandé désormais à des acteurs qui n'en ont pas les moyens, et les privant de données auxquelles ils avaient jusque-là accès, nécessaires à la réalisation de leurs missions.

Parallèlement, ce problème de convergence des sources de données, et de « propriété », aurait pu être au moins techniquement résolu par la création et l'utilisation du dossier médical partagé (DMP), annoncé par la loi du 13 août 2004 relative à l'assurance maladie : il s'agirait simplement d'un carnet de santé électronique, étendu à toute la vie et non plus aux seules premières années. Le projet ne prend pas, certainement pour des raisons de visions divergentes et de concurrences interinstitutionnelles. L'Agence des systèmes d'information partagée en santé (Asip, devenue Agence du numérique en santé en décembre 2019) est créée en 2009, dont une des missions est spécifiquement de relancer le projet du DMP. La Caisse nationale d'assurance maladie reprend le projet à son compte en 2015. Finalement, la dernière itération de ce concept devrait se réaliser dans l'Espace numérique de santé, défini par la loi du 24 juillet 2019, à partir du 1er janvier 2022.

La concurrence n'est pas limitée à une concurrence inter-institutions. Elle existe également au sein même des institutions. Un de ses aspects est une confrontation culturelle et liée à un changement dans les formations des élites françaises, où les écoles de management et de commerce sont depuis plusieurs années plus prisées que les écoles d'ingénieurs. Traditionnellement, au sein de l'institution publique, les personnes en charge des études et des statistiques – la statistique d'État, et ses extensions, *via* le captage et l'automatisation, la fédération des données produites et numérisées – sont majoritairement issues de l'École nationale de la statistique et de l'administration économique (Ensae). On observe, depuis quelques années, le recrutement de personnes plus souvent issues d'écoles privées de management et de commerce (HEC, Essec, Edhec…), voire la constitution interne de départements dédiés et distincts des départements des études et statistiques autour de ces recrutements. Ces nouveaux départements ou directions sont en général en charge de l'innovation ou de la transformation « digitale », et se retrouvent face aux départements historiques des systèmes d'information et des études statistiques. Il existe enfin probablement une concurrence, ou un frein à la convergence, sur le versant sanitaire : les experts en santé publique sont, en France, issus d'au moins trois horizons différents, parfois cumulatifs, mais dont les logiques institutionnelles sous-jacentes sont plutôt concurrentielles là aussi : les médecins de santé publique, les chercheurs en épidémiologie, et les diplômés de l'École des hautes études en santé publique (EHESP).

Globalement, on assiste donc à des efforts publics de concentration des données personnelles, captées à partir de différentes institutions publiques mais aussi depuis le secteur du privé *via* le secteur du travail, définie et imposée par la voie législative. Les efforts d'ouverture de ces données semblent réels, mais loin de reposer ou de favoriser une déconcentration, comme cela aurait pu, ou pourra être, en se basant sur la participation et l'accord de chaque citoyen, *via* un DMP ou tout autre dispositif semblable et acceptable par la société. À ce jour, tout sondage effectué pour mesurer l'acceptation des Français d'un tel dispositif a toujours pointé vers une adhésion massive au principe : quiconque passe de son généraliste aux urgences, puis à l'hospitalisation, puis d'un hôpital à un autre, d'un spécialiste à un autre, sait combien il serait utile de disposer d'un dossier unique.

La donnée pour décider, le fait scientifique et la santé publique

Le protectionnisme de l'État dès qu'il est question de données de santé, envers les autres nations, mais aussi envers une partie du secteur privé, parfois entre institutions publiques et enfin même envers les citoyens, est souvent justifié par le caractère sensible de ces données – on met en avant le caractère individuel, mais d'un point de vue national, elles sont aussi sensibles en cela qu'elles

renseignent sur la population –, le secret professionnel et la valeur des données pour l'industrie et la recherche. Il existe également la dimension que la récente épidémie de Covid-19 a mise en exergue : le besoin de disposer de connaissances scientifiques d'une part, mais également de données pour décider, pour gouverner. Plusieurs commentateurs auront relevé la porosité des champs lexicaux quant à la gestion de l'épidémie en cours, relevant tous des prérogatives régaliennes, à commencer par le langage martial : nous sommes en guerre contre le virus ; un état d'urgence sanitaire peut être déclaré, délivrant des pouvoirs spéciaux aux représentants de l'État. Dans le cadre de l'épidémie, les deux sources évoquées, supports de la décision éclairée, ont été prises en défaut, au sens de leur caractère utile, adapté et précis à un moment donné de l'épidémie : les connaissances scientifiques d'une part, la production de données utiles d'autre part. Ce qui est aux racines historiques de la statistique, dénombrer les morts et les naissances, a été défaillant, et les logiques concurrentielles entre institutions, un problème majeur : le CépiDc, unité Inserm unique dévolue au recueil et à l'étude des causes médicales de décès, en position d'asphyxie depuis plusieurs mois sinon années, a sans doute été un des seuls acteurs à avoir été privé de la manne financière ouverte à destination de la recherche et de l'industrie face à la Covid-19. L'écosystème en place n'a pas pu être mobilisé, renforcé, pour fonctionner et produire les données nécessaires. D'autres institutions publiques ont occupé le terrain de façon peu convaincante ou insuffisante pour documenter efficacement et rigoureusement les connaissances sur la mortalité, comme l'Insee, l'Ined ou Santé publique France. Et pour cause : aucune de ces institutions ne dispose des causes médicales de décès.

Symétriquement, le recueil de données de santé peut être, au-delà de son caractère informatif sur l'état de santé d'une population, le biais par lequel introduire une forme de gouvernement des individus, voire d'une police sanitaire, et la possibilité, parfois inédite, de rapprocher des institutions publiques, et même privées, dans cette optique de contrôle et de régulation des flux de population. Le cas de la santé publique de précision dans le cadre de la gestion de l'épidémie de Covid-19 est éclairant. Prenons l'exemple des applications téléphoniques de *tracking* (traçage des cas). La majorité des pays à hauts revenus s'est dotée d'une application de *tracking*. Aucune n'a été en position d'être sanitairement utile – la couverture nécessaire, c'est-à-dire la proportion de personnes installant et utilisant réellement l'application, est estimée à plus de 60 % pour être utile. Il semble que, dans le meilleur des cas, la proportion des personnes ayant téléchargé une telle application a été de 30 à 40 %. Le cas français (StopCovid) est une illustration classique de nos institutions : un outil propriétaire, développé par les acteurs habituels publics et industriels privilégiés, sans aucune évaluation extérieure. Dans d'autres pays, ce qu'il est important de souligner est alors l'articulation entre institutions derrière le déploiement de l'application, et son insertion dans une politique sanitaire plus large. Ainsi, en Chine, l'application délivrait ni plus ni moins des autorisations – un passeport – de circulation physique selon son niveau individuel de risque. Les algorithmes et les caractéristiques pris en compte pour élaborer le niveau de risque sont inconnus. En Corée du Sud, le système repose sur la circulation d'informations entre les acteurs du système de santé, la police ou encore les aéroports et le gouvernement.

Dans l'épidémie de Covid-19, jusqu'à présent, l'*evidence-based medicine* (EBM, médecine basée sur les preuves) a été mise à l'épreuve face à d'autres conceptions, dont l'école des pragmatiques, et semble ne pas avoir été en position de contribuer significativement à l'adaptation des politiques de santé. Le passage discret à une santé publique de précision, soutenant une *evidence-based policy* (EBP, politique basée sur les preuves), est, d'un point de vue sanitaire, encore moins convaincant.

Les problématiques autour des données de santé révèlent plus que jamais différentes tensions connues, que l'on peine à dénouer : les tensions au cœur de la santé publique, devant concilier intérêt individuel et populationnel ; les tensions inhérentes au *new public management*, et la porosité sélective, inconfortable, entre public et privé dans la répartition et la réalisation de missions de service public ; les tensions, enfin, entre acteurs participant à la production et à l'utilisation des données. En effet, la promotion et l'utilisation des données sont avant tout organisées par et pour un service public, dont les acteurs sont en concurrence. Parmi eux, les producteurs de données ne veulent pas être dépossédés d'une partie de leur travail, qu'ils n'ont pas les moyens de valoriser souvent seuls, la donnée étant présentée comme une valeur en soi, source d'avantage compétitif (entre hôpitaux par exemple).

Bibliographie générale

1. Angelé-Halgand N., Garrot T. « Discipliner par le chiffre : l'hôpital financiarisé au risque de la réification ? ». *Entrep. Hist.*, 8 octobre 2015, 79 (2), 41-58.

2. Aymé S. « Qu'est-ce que la médecine prédictive ? ». *ADSP*, mars 2001, 34, 18.

3. Baro E., Degoul S., Beuscart R., Chazard E. «Toward a literature-driven definition of big data in healthcare». *BioMed Research International*, 2015.

4. Bezes P., Musselin C. « Le New Public Management. Entre rationalisation et marchandisation ? » *In* : Boussaguet L., Jacquot S., Ravinet P. *et al*. *Une French Touch dans l'analyse des politiques publiques ?* Presses de Sciences Po, «Académique», 2015, 128-151.

5. Boelle P.-Y., Thiébaut R., Costagliola D. « Données massives, vous avez dit massives ? », *Questions de santé publique*, septembre 2015, 30.

6. Bouchard L., Blancquaert I. « Un cadre d'évaluation des technologies génétiques : le diagnostic et le dépistage des porteurs de la maladie de Steinert». *In* : Mélançon M. J., Gagné R. (dir). *Dépistage et diagnostic génétiques. Aspects cliniques, juridiques, éthiques et sociaux*. Les Presses de l'Université Laval, 1999.

7. Breiman L. « Statistical modeling : The two cultures». *Statistical Review*, 16 (3), 2001.

8. Cardo D. M. *et al*. « Mandatory reporting of hospital-acquired infections : Steps for success». *J. L. Med. & Ethics,* 2005, 33, 86.

9. Cardon D. *A quoi rêvent les algorithmes. Nos Vies à l'heure des big data*. Seuil, 2015.

10. Chazard E. « Intelligence artificielle et aide à la décision en santé». *In :* Rouet G. *Algorithmes et décisions publiques*. *Les Essentiels d'Hermès*. Paris : CNRS Éditions, 2019.

11. Chazard E. « Réutilisation de données hospitalières et intelligence artificielle : des données à l'intervention». Webinar Quantim Sestim. **https://www.youtube.com/watch?v=0oxwWaCyirI**, 2020, consulté le 10 septembre 2020.

12. Chazard E., Beuscart J.-B., Rochoy M., Dalleur O., Decaudin B., Odou B., Ficheur G. « Statistically prioritized and contextualized clinical decision support systems, the future of adverse drug events prevention ? ». *Studies in Health Technology and Informatics*, 2020, 270, 683-687.

13. Chazard E., Ficheur G., Caron A., Lamer A., Labreuche J., Cuggia M., Genin M., Bouzille G., Duhamel A. « Secondary use of healthcare structured data : The challenge of domain-knowledge based extraction of features». *Studies in Health Technology and Informatics*, 2018, 255, 15-19.

14. Cleeren E., Van der Heyden J., Brand A., Van Oyen H. « PH in the genomic era : Will Public Health Genomics contribute to major changes in the prevention of common diseases ? » *Archives of Public Health*, 5/12/2012, 69, 8, 1-12.

15. Coldefy M., Gandré C. « Personnes suivies pour des troubles psychiques sévères : une espérance de vie fortement réduite et une mortalité prématurée quadruplée». Irdes, *Questions d'économie de la santé*, sept. 2018, 237. **https://www.irdes.fr/recherche/questions-d-economie-de-la-sante/237-personnes-suivies-pour-des-troubles-psychiques-severes-une-esperance-de-vie-fortement-reduite.pdf**

16. Comité consultatif national d'éthique (CCNE). Avis n° 124 du 21/01/2016, « Réflexion éthique sur l'évolution des tests génétiques liée au séquençage de l'ADN humain à très haut débit». Avis n° 129 du 18/09/2018, « Contribution du CCNE à la révision de la loi de bioéthique 2018-2019». Avis n° 130 du 29/05/2019, « Données massives et santé : une nouvelle approche des enjeux éthiques».

17. Conseil national de la consommation. *Rapport sur les objets connectés en santé*. CNC : 7 juillet 2017, p. 5 **https://www.economie.gouv.fr/files/files/directions_services/cnc/avis/2017/Rapport-objets-connectes-sante070717.pdf**

18. Conseil national de l'ordre des médecins. *Médecins et patients dans le monde des data, des algorithmes et de l'intelligence artificielle*. Paris : janvier 2018.

19. Cuggia M., Polton D., Wainrib G., Combes S. *Health Data Hub : mission de préfiguration*. Paris : Ministère des Solidarités et de la Santé, 12 octobre 2018.

20. Daneman N. *et al*. « Reduction in Clostridium difficile infection rates after mandatory hospital public reporting : findings from a longitudinal cohort study in Canada». *PLoS Medicine*, 2012, 9, 7.

21. Delort P. *Le Big Data*. PUF, « Que sais-je ? », 2018.

22. Durand G., Guillet M., Mercier S. « Favoriser l'autonomie du patient face aux données additionnelles en médecine génomique». *Canadian Journal of Bioethics - Revue canadienne de bioéthique*, 2019, 2 (2), 135-142.

23. Edmond M. B., Bearman G. M. L. « Mandatory public reporting in the USA : An example to follow ? ». *Journal of Hospital Infection,* 2007, 65 (S2), 182.

24. Erikson G. A., Bodian D. L., Rueda M. *et al*. «Whole-genome sequencing of a healthy aging cohort». *Cell*, 5 mai 2016, 165 (4), 1002-11.

25. Fassin D. *Faire de la santé publique*. Presses de l'EHESP, 2008, 80 p.

26. Feenberg A. « (Re) penser la technique. Vers une technologie démocratique». *Revue du MAUSS*, Éditions La Découverte, 2004.

27. Ferretti L., Wymant C., Kendall M. *et al*. « Quantifying SARS-CovV-2 transmission suggests epidemic control with digital contact tracing». *Science,* 2020, n° 6491 : eabb6936 DOI : 10.1126/science. abb6936

28. Foucault M. *Surveiller et punir*. Gallimard, « Bibliothèque des Histoires», 1975.

29. Hardy A-C. « Oser l'incertain. L'imputation des effets indésirables médicamenteux». *Anthropologie & Santé*, 7 février 2019.

30. Haustein T. *et al.* « Use of benchmarking and public reporting for infection control in four high-income countries». *Lancet Infect Dis,* 2011, 11, 471.

Bibliographie générale

31. Haut Conseil de la santé publique. *Propositions pour une politique nationale nutrition santé à la hauteur des enjeux de santé publique en France. PNNS 2017-2021*. 2017.

32. Hecketsweiler C., Seckel H. « Coronavirus : en France, avoir un bilan final du nombre de morts prendra plusieurs mois ». *Lemonde.fr*, 2 mai 2020, mise à jour 3 mai 2020. https://www.lemonde.fr/planete/article/2020/05/02/coronavirus-en-france-avoir-un-bilan-final-prendra-plusieurs-mois_6038434_3244.html

33. Joyner M. J., Paneth N. « Seven questions for personalized medicine ». *JAMA*, 2015, 314 (10), 999-1000. doi : 10.1001/jama.2015.7725

34. Kalia S., Adelman K., Bale S. *et al*. « Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0) : A policy statement of the American College of Medical Genetics and Genomics ». *Genet Med,* 2017, 19 (2), 249-255. https://www.ncbi.nlm.nih.gov/clinvar/docs/acmg/

35. Khoury M. J., Gwinn M. L., Glasgow R. E., Kramer B. S. « A population approach to precision medicine ». *American Journal of Preventive Medicine*, 2/12/2012, 42 (6), 639-645.

36. Khoury M. J., Iademarco M. F., Riley W. T. « Precision public health for the era of precision medicine ». *American Journal of Preventive Medicine*, 2016, 50 (3), 398-401.

37. Latour B. *La Fabrique du droit, une ethnographie du Conseil d'Etat*. Éditions La Découverte, 2004.

38. Lazer D., Kennedy R., King G., Vespignani A. « The parable of Google Flu : Traps in big data analysis ». *Science,* 2014, 343 (6176), 1203-5.

39. Leca A. (dir.). *Le Risque épidémique*. Les Études hospitalières, 2003.

40. Leplège A., Bizouarn P., Coste J. (dir.). *De Galton à Rothman. Les Grands Textes de l'épidémiologie au xx^e siècle*. Paris : Hermann, 2011.

41. *L'Enquête nationale sur les événements indésirables liés aux soins (Eneis)*. Ministère des Solidarités et de la Santé. https://drees.solidarites-sante.gouv.fr/etudes-et-statistiques/open-data/etablissements-de-sante-sociaux-et-medico-sociaux/article/l-enquete-nationale-sur-les-evenements-indesirables-lies-aux-soins-eneis#Publications

42. Marsteller J. A. *et al.* « Evaluation the impact of mandatory public reporting on participation and performance in a program to reduce central line-associated bloodstream infections : Evidence from a national patient safety collaborative ». *American Journal of Infection Control,* 2014, 42, S209.

43. Nicholls S. G., Quach P., von Elm E., Guttmann A., Moher D., Petersen I., Sørensen H. T., Smeeth L., Langan S. M., Benchimol E. I. « The REporting of Studies Conducted Using Observational Routinely-Collected Health Data (RECORD) Statement : Methods for arriving at consensus and developing reporting guidelines ». *PLoS One,* 2015, 10. : e0125620.

44. Office parlementaire d'évaluation des choix scientifiques et technologiques. *L'Intelligence artificielle et les données de santé*. Paris : 21 mars 2019, n° 1795/n° 401.

45. Olstad D. L., McIntyre L. « Reconceptualising precision public health ». *BMJ Open,* 13/09/2019, 9. : e030279

46. Pailler L. « StopCovid : la santé publique au prix de nos libertés ? Brèves observations sur l'application de traçage numérique ». *Recueil Dalloz*, 2020, 935-936. hal-02569123

47. Pon D. et Coury A. *Stratégie de transformation du système de santé. Rapport final. Accélérer le virage numérique*. Ministère des Solidarités et de la Santé, 2018.

48. Py B. « De la surveillance des maladies à la surveillance des malades ». *Dalloz actualité*, 27 mai 2020.

49. Rist S., Mesnier T. (rapporteurs). *Rapport fait au nom de la commission des affaires sociales sur le projet de loi relatif à l'organisation et à la transformation du système de santé*. Commentaires d'articles et annexes. Paris : 14 mars 2019, vol. II, n° 1767.

50. Rose G. *The Strategy of Preventive Medicine*. Oxford : Oxford University Press, 1992.

51. Saporta G. « Quelle statistique pour les Big Data ? Entretien avec Gilbert Saporta ». *Statistique et Société*, Société française de statistique, 2017, 5 (1).

52. Stricof R. L. *et al.* « Lessons learned while implementing mandatory health care-associated infection reporting in New York State ». *J Public Health Management Practice,* 2013, 19, 4, 294.

53. Tandy-Connor S., Guiltinan J., Krempely K. *et al*. « False-positive results released by direct-to-consumer genetic tests highlight the importance of clinical confirmation testing for appropriate patient care ». *Genet Med,* 2018, 20, 1515-21.

54. Tuppin P., Rudant J., Constantinou P., Gastaldi-Ménager C., Rachas A., de Roquefeuil L., Maura G., Caillol H., Tajahmady A., Coste J., Gissot C., Weill A., Fagot-Campagna A. « Value of a national administrative database to guide public decisions : From the Système national d'information inter-régimes de l'Assurance maladie (SNIIRAM) to the Système national des données de santé (SNDS) in France ». *Revue d'épidemiologie et de santé publique,* 2017, 65, suppl. 4., S149-S167.

55. Weeramanthri T. S., Dawkins H. J. S., Baynam G. *et al*. « Editorial : Precision public health ». *Frontiers in Public Health,* 2018, 6, 1-3. doi : 10.3389/fpubh.2018.00121

56. Wilson J. M. G., Jungner G. *Principes et pratiques du dépistage des maladies*. OMS, 1970.

57. Wong E. S. *et al.* « Public disclosure of healthcare-associated infections : The role of the Society for Healthcare Epidemiology of America ». *Infection Control & Hospital Epidemiology,* 2005, 26 (2), 210.