Ministère des affaires sociales et de la santé

Colloque Big Data en santé

Quels usages ? Quelles solutions ?

Paris, 4 juillet 2016

Big Data en santé

Quels usages ? Quelles solutions ?

Marisol Touraine, Ministre des Affaires sociales et de la santé, a lancé le 10 septembre 2015 une réflexion sur le Big Data en santé. Cette réflexion a été menée selon quatre axes, approfondis par quatre groupes de travail[1] :

- le premier autour des usages existants ou potentiels du Big Data en santé ;
- Le deuxième sur les problématiques éthiques et juridiques ;
- Le troisième autour des infrastructures et systèmes d'information spécifiques ;
- Le quatrième sur les enjeux socio-économiques du big data en santé.

Les réflexions ont été étayées par de nombreuses rencontres avec divers acteurs du Big Data en santé, des structures de santé au sens large, des chercheurs, des acteurs du monde de la santé. Thomas Lefèvre a coordonné l'ensemble des travaux.

Lancement d'une consultation citoyenne

Les entretiens avec les acteurs associatifs ont révélé le manque de connaissances sur ce sujet vaste et émergent qu'est le Big Data en Santé, prérequis à la constitution d'un avis sur les enjeux qu'il recouvre. Le Ministère des Affaires sociales et de la Santé a donc sollicité, en parallèle des travaux du groupe de réflexion, le Secrétariat général pour la modernisation de l'action publique (SGMAP) et la Commission nationale du débat public (CNDP) afin de mettre en place un atelier citoyen autour de l'acceptabilité des usages du Big Data[2]. Le SGMAP a également mis en place une consultation en ligne sur le site faire-simple.gouv.fr « Partager ses données de santé : pour quels bénéfices et à quelles conditions ? », dont une synthèse sera présentée au colloque.

Contenu du document

Ce document reflète les travaux menés par les groupes. La première partie décrit le contexte et définit les concepts (donnée de santé et Big Data en santé) ; la deuxième partie présente une typologie des usages du Big Data en santé ; et la troisième partie examine les enjeux et les précautions nécessaires au déploiement du Big Data en santé.

[1] Sur le premier axe, le groupe de travail été animé par Catherine Duclos et Benjamin Sarda ; sur le deuxième par Laure Bédier, Sophie Nerbonne, Jean-Claude Ameisen et Dominique Thouvenin ; sur le troisième par Gilles Copin et Mathias Herberts ; sur le quatrième par Cyrille Delpierre et Béatrice Falise-Mirat.

[2] Avis de l'atelier citoyen, « Big data en santé – faut-il rendre accessibles les données de santé ? A qui ? Pour quoi faire ? A quelles conditions ? », 19 juin 2016, remis lors du colloque.

Première partie : contexte et définitions

Depuis le lancement de la réflexion, la législation en matière de données et d'accès aux données de santé a profondément évolué. L'article 193 de la loi de modernisation de notre système de santé du 26 janvier 2016 crée notamment un système national des données de santé (SNDS) qui centralisera les données de bases médico-administratives existantes en matière sanitaire et médico-sociale, et assurera leur mise à disposition selon des modalités simplifiées. Ce cadre juridique constitue ainsi un des premiers piliers du développement du Big data en santé par l'ouverture raisonnée des données médico–administratives issues dans un premier temps des remboursements de l'Assurance maladie, des séjours hospitaliers et des causes de décès, ultérieurement des organismes complémentaires et des maisons départementales des personnes handicapées.

Plus récemment, le projet de loi pour une République numérique a été adopté en première lecture par le Sénat le 3 mai 2016. Un de ses volets concerne la circulation des données et du savoir. À ce stade, le texte propose l'ouverture des données publiques, dès lors qu'elles ne sont pas couvertes par un secret légal. Ont également été votées la création d'un service public de la donnée et l'introduction de données d'intérêt général pour permettre leur réutilisation par tous. Cette ouverture des données publiques constituera aussi un facilitateur du déploiement du Big Data en santé, en permettant le rapprochement de bases de données de thématiques diverses.

Dans le contexte international de fort développement des données dans leur ensemble et des traitements qui en sont faits, le cadre juridique européen évolue également. Les principes énoncés dans la directive de 1995 sur la protection des données sont modernisés dans le nouveau règlement général sur la protection des données (RGPD) qui a été adopté le 14 avril 2016 et entrera en application en 2018[3]. Notamment, les données à caractère personnel bénéficieront désormais d'une protection renforcée, et un consentement clair sera requis pour pouvoir les traiter. L'information sur les traitements effectués devra être claire et complète. L'autodétermination informationnelle s'en trouve ainsi renforcée.

Bien que ces évolutions soient des préalables à l'essor du Big Data en santé, elles ne couvrent pas l'ensemble des enjeux juridiques qu'il soulève. Le travail du groupe de réflexion juridique, présenté infra, propose de bâtir un cadre de référence pour chacun des cas d'utilisation dégagés par le groupe des usages du Big Data en santé.

La réflexion menée par les quatre groupes s'insère dans un questionnement plus global autour de la santé et du numérique dans son acception la plus large. La Ministre des Affaires sociales et de la Santé a saisi, dès février 2014, le Conseil National du Numérique sur une possible mobilisation de la technologie numérique au service de la Stratégie Nationale de Santé. Parmi les recommandations du rapport *La santé, bien commun de la société numérique*, remis en octobre 2015, trois objectifs se dégagent :

- outiller les citoyens pour leur permettre d'être des acteurs à part entière du système de santé ;
- inciter les acteurs économiques détenteurs de données de santé à les partager de manière circonstanciée ;
- enfin, former les professionnels de santé aux enjeux et usages du numérique.

[3] https://www.cnil.fr/fr/reglement-europeen-protection-donnees

Plus récemment, le Premier Ministre a confié au professeur Yves Levy, Président Directeur Général de l'Inserm, le soin d'identifier les obstacles politiques, règlementaires, organisationnels et techniques à un séquençage à haut débit du génome humain entier, afin d'intégrer ces nouvelles données dans la pratique médicale. L'enjeu aujourd'hui n'est plus en effet de séquencer le génome humain mais bien de l'analyser en des temps raisonnables et de fournir une information utilisable par les praticiens[4]. Les problématiques posées dans ce cadre rejoignent les préoccupations éthiques, économiques et techniques étudiées par les quatre groupes mentionnés ci-dessus.

Comment définir le Big Data en santé ?

Définir ce qu'est le Big Data en santé est un pré-requis à la description de ses enjeux, mais le Big Data est un objet flou et évolutif.

Le terme « Big Data », en français « données massives », est apparu à la fin des années 1990 pour décrire les problèmes de stockage et de traitement rencontrés dans l'exploitation de données particulièrement volumineuses. Au début des années 2000, cette définition s'étoffe avec la notion des « 3 V » : le « Big Data » désigne alors communément des données de grand « volume » (connaissance de très nombreuses dimensions d'un grand nombre d'individus), ou d'une grande « variété » (grande diversité de sources et de formats), ou d'actualisation et de disponibilité très fréquente (« vélocité »). Wikipédia décrit en 2014 le Big Data de manière tout aussi large : « des données si volumineuses et complexes qu'il est difficile de les traiter par des applications traditionnelles ». Alors que le terme est maintenant largement utilisé dans les entreprises et dans les médias, le Big Data n'a toujours pas une définition précise et stabilisée. Il renvoie aujourd'hui comme hier aux limites techniques que ces données impliquent : des « Big Data » hier ne sont donc pas nécessairement des « Big Data » aujourd'hui. À ces contours mouvants s'ajoute le nombre grandissant de vocables (fr)anglais qui recoupent de près ou de loin le Big Data : Open Data, Smart Data, e-santé[5], etc. mais qui n'en sont pas synonymes.

La définition retenue dans ce travail repose sur deux aspects :

- le croisement de données à finalités différentes, cela afin d'enrichir une source par des données de contexte, ou d'exploiter le croisement de deux sources pour étudier l'interaction entre différents phénomènes. Ces données peuvent être de granularités diverses, par exemple l'une individuelle et l'autre agrégée au niveau de la commune. Ces croisements constitueront un vecteur de création d'information et de valeur.
- l'utilisation de techniques d'analyse et de traitements de données inhabituels dans le domaine de la santé[6]. Peuvent ainsi être utilisés des algorithmes spécifiques en vue d'obtenir une prédiction d'une maladie ou d'une complication donnée, de l'affluence dans un centre de soins, des effets indésirables d'un médicament, du type de pathologie que tel groupe de personnes pourrait développer... Des techniques adaptées qui répondent aux difficultés posées par la structuration spécifique de ces données peuvent également être mises en

[4] Yves Lévy, *France Médecine Génomique 2025 : permettre l'accès au diagnostic génétique sur tout le territoire",* remis au Premier ministre le 22 juin 2016.
[5] En anglais e-health.
[6] Elles l'étaient néanmoins dans d'autres domaines scientifiques avec des finalités et des questionnements différents. Elles relèvent par exemple de l'apprentissage statistique, de l'intelligence artificielle, de l'algorithmique, etc.

œuvre afin de fournir des réponses à des questions « classiques ». De ces outils d'analyse non traditionnels dans le monde de la santé naîtront également de la connaissance et des applications.

À ces deux aspects peuvent éventuellement s'ajouter les caractéristiques habituelles des Big Data, au sens des multiples V, volume, variété, vélocité.

Afin de clarifier les différences de concepts, il convient de préciser que l'Open Data ne relève pas nécessairement du Big Data. Comme son nom l'indique, l'Open Data désigne la « simple » mise à disposition de données. Aucun traitement, finalité ou caractéristique spécifique n'est induit par ce vocable si ce n'est le « libre accès ». Par nature, l'Open Data ne porte que sur des données strictement anonymes, pour lesquelles les individus dont les données sont extraites ne peuvent pas être reconnus ou retrouvés. Malgré la différence entre Big Data et Open Data, les données en accès libre se révèlent des facilitateurs de traitements « Big Data », voire même souvent un préalable, car elles permettent le croisement de données d'origines et de finalités différentes.

Le Big Data se nourrit donc de sources de données diverses, pouvant être des données personnelles nominatives, des données personnelles déidentifiées [7], des données individuelles rendues anonymes [8], ou encore des données agrégées, des données ou indicateurs de contexte (environnemental, socio-économique, démographique, météorologique, etc.)... À des fins d'illustration, on cite ici sans aucune prétention à l'exhaustivité (qui serait par définition illusoire) certaines sources potentielles de Big Data en santé :

- les données médico-administratives produites par l'Assurance maladie (Sniiram) et les hôpitaux (PMSI) ;
- les données figurant dans les dossiers médicaux, à l'hôpital notamment mais aussi en ville ;
- Les données détenues par des acteurs publics ou privés recueillies auprès de patients (essais cliniques notamment) ou de professionnels de santé ;
- les données générées par les objets connectés, les applications mobiles, les sites internet et moteurs de recherche ;
- les données de contexte, socio-économiques, géographiques, environnementales, etc.

A la lecture de cette simple énumération, émergent déjà quelques-unes des problématiques qui seront abordées dans la suite du document :

- de nombreuses données sont des données personnelles, directement ou indirectement identifiantes : le Big Data, pour se développer, doit donc pouvoir bénéficier d'un écosystème reposant sur la confiance de tous. Pour cela, il nécessite que soit recueilli et respecté le consentement éclairé des personnes, qu'on ait recours à des infrastructures et des technologies sûres et qu'enfin, soit mise en place une gouvernance adaptée ;

[7] C'est-à-dire des données dont on a retiré les identifiants directs (nom, adresse, numéro de sécurité sociale…) mais qui, à partir de croisements avec d'autres informations, permettent de retrouver les personnes.
[8] Dans ce cas, les informations sont suffisamment appauvries ou agrégées pour qu'on ne retrouve pas les personnes, même en procédant à des croisements avec des informations externes.

- une partie des données qui nourrissent le Big Data en santé sont des données dites non structurées[9] et/ou non qualifiées (recueil d'information partielle, information non validée, etc.). La capacité à développer des outils opérationnels et fiables de traitement de ces informations constitue un enjeu majeur pour la réussite du Big Data en langue française ;
- certains traitements du Big Data porteront sur des groupes ou des collectifs, à vocation de recherche. D'autres, qui seront probablement de plus en plus nombreux, auront une finalité d'aide à la décision, en appliquant des outils issus du Big Data à des patients identifiés. Le rôle des professionnels de santé, et au premier chef des médecins, et leur appropriation de ces nouveaux outils, sont donc primordiaux ;
- le Big Data, par la variété des données concernées et des usages possibles, nécessite des investissements importants : des investissements financiers dans les infrastructures et les technologies et des investissements dans les compétences humaines. Sa réussite reposera sur la création de collaborations et coopérations entre acteurs publics et acteurs privés.

Qu'appelle-t-on « données de santé » ?

La définition de l'état de santé retenue habituellement est celle de l'Organisation mondiale de la santé : « *la santé est un état complet de bien-être physique, mental et social et ne se limite pas à l'absence de maladie ou d'infirmité* ».

En cohérence avec cette approche, on définit traditionnellement une donnée de santé comme une donnée révélant des informations sur l'état de santé d'une personne, actuel ou futur.

Le développement du Big Data ouvre la possibilité que des données qui n'entraient pas jusqu'ici dans la conception d'une donnée de santé (c'est-à-dire qui ne révélaient pas, prises en elles-mêmes, des informations sur la santé) puissent, dans certains cas, révéler une information sur l'état de santé d'une personne. Les données issues d'applications pour coureurs en sont un exemple : comme l'a souligné l'autorité néerlandaise homologue de la CNIL, les variations de performance enregistrées dans ces applications peuvent informer sur la dégradation de l'état de santé.

On voit donc apparaître un schéma nouveau dans lequel ce n'est pas tant la donnée brute, prise en elle-même, qui peut être caractérisée comme une donnée de santé : ce sont les traitements menés sur les données qui révèlent, a posteriori, des informations sur la santé. C'est un renversement de perspective qui interroge notre conception juridique et analytique de la donnée de santé.

Le groupe de réflexion autour du Big Data en santé a finalement retenu la définition récemment adoptée par le règlement européen général sur la protection des données. L'article 4 définit comme « données concernant la santé », « *les données à caractère personnel relatives à la santé physique ou mentale d'une personne physique, y compris la prestation de services de soins de santé, qui révèlent des informations sur l'état de santé de cette personne* ». L'alinéa 35 du préambule détaille par ailleurs la conception de la santé retenue :

« Les données à caractère personnel concernant la santé devraient comprendre l'ensemble des données se rapportant à l'état de santé d'une personne concernée qui révèlent des informations sur l'état de santé physique ou mentale passé, présent ou futur de la personne concernée. Cela comprend

[9] Par exemple des données textuelles (compte-rendu opératoire, requête sur un moteur de recherche…) ou des images (radiologie, IRM…).

des informations sur la personne physique collectées lors de l'inscription de cette personne physique en vue de bénéficier de services de soins de santé ou lors de la prestation de ces services au sens de la directive 2011/24/UE du Parlement européen et du Conseil au bénéfice de cette personne physique ; un numéro, un symbole ou un élément spécifique attribué à une personne physique pour l'identifier de manière unique à des fins de santé ; des informations obtenues lors du test ou de l'examen d'une partie du corps ou d'une substance corporelle, y compris à partir de données génétiques et d'échantillons biologiques ; et toute information concernant, par exemple, une maladie, un handicap, un risque de maladie, les antécédents médicaux, un traitement clinique ou l'état physiologique ou biomédical de la personne concernée, indépendamment de sa source, qu'elle provienne par exemple d'un médecin ou d'un autre professionnel de la santé, d'un hôpital, d'un dispositif médical ou d'un test de diagnostic in vitro. »

La CNIL, saisie par la Ministre des affaires sociales et de la santé, considère que « *doivent être regardées comme des données de santé, d'une part, celles qui sont « par nature » relatives à l'état de santé d'une personne (celles issues de la relation de soin par exemple) et, d'autre part, celles qui le seraient compte tenu de leur destination (telles que celles issues de certains objets connectés)* ». En somme, une donnée de santé peut être considérée comme telle, qu'elle le soit par nature ou par intention. Cette définition met ainsi le traitement des données au cœur de l'analyse. Le questionnement des groupes de réflexion a donc conduit à caractériser les traitements envisageables, en établissant une typologie des usages (voir infra).

Big Data en santé : état des lieux et apports potentiels

Qu'elles soient issues d'acteurs privés, comme celles détenues, recueillies ou produites par exemple par les industriels des produits de santé, les assureurs complémentaires, les GAFAMS (Google, Apple, Facebook, Amazon, Microsoft, Samsung), ou d'acteurs publics, comme celles détenues ou produites par les établissements ou les professionnels de santé, les données de santé ont une valeur qui repose en grande partie sur les traitements qui en sont ou en seront faits.

Se sont ainsi posées différentes questions tout au long de la consultation : comment rendre les traitements de ces données de santé plus accessibles ? Comment rendre plus aisés leur mutualisation et leurs usages croisés? Sous quelles conditions notamment techniques, éthiques, économiques ? Avec quelles réserves ? Comment assurer leur confidentialité à court comme à long terme ? Comment garantir un traitement licite de ces données ? Comment préserver l'anonymat des personnes ? Quelle place le patient a-t-il dans l'utilisation de ses données ? Et quel est le vivier d'utilisations entrevues ?

Les projets de Big Data en santé sont rendus possibles grâce à la collecte massive de données de santé. A l'accumulation de données médico-économiques viennent s'ajouter des données issues des dossiers patients informatisés – même si ceux-ci ne sont pas encore généralisés.

La mise en place des entrepôts hospitaliers de données permet ainsi la concentration de données issues de leurs systèmes d'information. Celui de l'Hôpital Européen Georges Pompidou en constitue l'un des exemples emblématiques en France, mais d'autres entrepôts sont aussi en cours de construction comme dans les CHU brestois et rennais, ou au sein des Centres de Lutte Contre le Cancer (avec le projet ConSoRe d'Unicancer). La construction de ces entrepôts permet de fédérer des données initialement stockées dans de multiples bases de données ou transitant par des flux, et de

les interfacer pour rendre leur interrogation aisée. Ces entrepôts peuvent ainsi être utilisés pour identifier rapidement des patients présentant des caractéristiques définies afin d'évaluer la faisabilité de recherches cliniques, de constituer facilement des cohortes, de « fouiller » statistiquement les données pour faire émerger de la connaissance. L'Assistance publique – Hôpitaux de Paris déploie actuellement un entrepôt de données qui collectera les données de tous ses établissements.

Fédérer ces entrepôts de données permet d'étendre la capacité de constitution de cohortes. Des plateformes informatiques permettant ces usages ont été développées dans le cadre d'initiatives nationales ou internationales[10].

Les problématiques d'interopérabilité de ces fédérations d'entrepôts et de leur gouvernance font partie des questions cruciales à traiter, auxquelles s'ajoute la question de l'ouverture des données à l'extérieur des établissements. Les entrepôts de données sont constitués au sein d'établissements de santé et la question de leur partage est complexe.

Ces expériences montrent que, pour mener à bien des projets Big Data, il est nécessaire de développer l'interdisciplinarité, de mutualiser les connaissances, notamment médicales et biologiques, et les compétences technologiques, informatiques et en traitement de données des différents acteurs. Cela ne peut pas se faire sans financements conséquents ni sans pilotage et concertation.

Le Big Data en santé, par le croisement de bases de données autrefois ou encore aujourd'hui cloisonnées, par l'apport de méthodes mathématiques renouvelées, permettra d'aborder à nouveaux frais des questions anciennes sur les facteurs de risque et les déterminants de la santé. Dans le domaine de la santé mentale, de la santé au travail, du lien entre santé et environnement, de la polypathologie, dans l'étude des microbiotes, en épigénétique, etc, le champ des applications est vaste.

Le Big Data est aussi un moyen de rendre effective et productive la recherche sur le terrain, qui a certes un coût au déploiement mais pourrait permettre des gains de temps et de moyens sur le plus long terme, en s'orientant vers une médecine non plus fondée seulement sur la littérature scientifique et l'expérience, mais aussi sur la mesure et la donnée de vie réelle.

Le Big Data en santé pourrait faciliter par ailleurs la reproductibilité de la preuve, c'est-à-dire que les résultats peuvent être retrouvés en ayant recours aux mêmes algorithmes. Il sera ainsi complémentaire des études randomisées : celles-ci contribuent à mesurer l'effet causal de tel traitement sur telle pathologie, en raisonnant toutes choses égales par ailleurs, mais sur des effectifs peu nombreux, très sélectionnés et dans des conditions plus proches des conditions de laboratoire que de la vie réelle, ce qui ne permet pas de toujours bien évaluer les effets en vie réelle ni de mesurer les effets différenciés du traitement selon certaines caractéristiques individuelles.

Le Big Data a de plus la capacité d'identifier des signaux faibles ou rares, par la taille et la diversité des données disponibles. Dans certains cas, il permet enfin de découvrir des associations, génératrices de nouvelles questions de recherche, sans avoir formalisé *ex ante* un postulat.

[10] Voir par exemple le projet EHR4CR http://www.ehr4cr.eu ou le projet OHDSI http://www.ohdsi.org.

Deuxième partie : une typologie des usages du Big Data en santé

Les réflexions menées au sein du groupe de travail ont fait émerger une typologie des usages en sept axes qui délimite plus précisément les apports et les enjeux du Big Data en santé. Cette approche par usages a le mérite de rendre concrets les traitements et données concernés, et d'illustrer la diversité des enjeux qui se posent. Elle permet aussi de relier les usages du Big Data en santé à leur finalité, notion centrale dans les droits français et européen.

Au sein de ces catégories, certains usages pourront être considérés comme positifs et porteurs de progrès ; d'autres pourront donner lieu à débat, voire être interdits. C'était l'un des enjeux de la consultation en ligne et de l'atelier citoyen que de débattre des différents usages du Big Data en santé, de ses opportunités et de ses risques.

Pour une partie des données utilisées et des traitements envisagés, les dispositions de l'article 193 de la loi de santé créent un cadre qui permet de fixer les modalités de mise en place et de contrôle de la licéité des traitements. D'autres sources de données ou des usages centrés sur les personnes ne s'y insèrent en revanche guère, et méritent que soit posées les questions des cadres de régulation. Comme on le verra plus loin, les réflexions du groupe de travail l'ont conduit à identifier la possibilité de travailler sur des finalités de traitement élargies assises sur les usages décrits ici. Ces usages ont donc à la fois une vocation descriptive, pour aider à comprendre le Big Data et à en débattre, mais aussi – c'est l'un des principaux enjeux de ce travail - une vocation plus normative.

1. Les usages dans le cadre d'une relation entre un patient et un professionnel de santé

Le Big Data en santé pourra permettre le développement d'outils à disposition des professionnels de santé afin d'améliorer ou de faciliter la prise en charge de leurs patients. Ces outils peuvent intégrer une dimension d'aide à la décision guidée par les données ou contribuer à une meilleure connaissance du profil patient pour lui apporter la solution de prise en charge la plus adaptée. Il pourrait s'agir par exemple d'outils pour améliorer l'observance des traitements prescrits, usage alors dynamique avec transmission régulière de diverses mesures au médecin ; ou des outils statiques permettant l'adaptation du traitement au plus proche de la situation et des caractéristiques des patients.

Patients et professionnels de santé seraient donc concernés, avec des bénéfices attendus en termes de prévention et de qualité des soins dispensés.

2. Les usages centrés sur la personne, en dehors d'une relation entre un patient et un professionnel de santé

Cet axe recouvre l'utilisation d'objets connectés ou d'applications qui permettent aux individus d'être acteurs de leur santé. Ces objets sont à la fois producteurs de données mais aussi fournisseurs de conseils contextualisés exploitant des données disponibles. Ces conseils peuvent être le résultat de l'exploitation de forums partagés par une communauté, de l'analyse de données recueillies dans le cadre d'une surveillance d'une maladie (diabète, surpoids) ou contextualisant le patient (environnement, habitudes, déplacements...). L'utilisateur peut par exemple modifier son comportement selon les notifications envoyées par son application, éviter des comportements à

risques par une meilleure connaissance de son environnement (par exemple éviter une zone à forts allergènes si le patient est allergique, faire des choix plus éclairés).

3. Les usages de vigilance

Le Big Data pourrait être une aide à la pharmacovigilance, à la matériovigilance, ou encore à la surveillance épidémiologique. Avec une étude en temps réel des données, certains indicateurs pourraient être construits et suivis au plus près.

Par exemple, un relevé systématique des effets indésirables de tel médicament pourrait être effectué auprès des patients, complété par les relevés des médecins et d'autres bases de données, puis analysé par traitement relevant du Big Data, permettant ainsi une meilleure pharmacovigilance. Les patients asthmatiques pourraient ainsi être équipés d'un dispensateur de bronchodilatateur connecté qui remonte en temps réel de l'information spatiale sur les allergènes présents, et de l'information sur la bonne observance de son utilisation dans une perspective de santé publique.

Les acteurs de la veille sanitaire et de vigilance sont tout particulièrement concernés, ainsi que les patients qui pourraient collectivement bénéficier d'une meilleure vigilance.

4. Les usages de pilotage de l'organisation de l'offre de soins et les usages médico-administratifs, à diverses échelles

Le développement d'outils Big Data pourrait aider à une meilleure gestion et organisation du système de santé. Chaque niveau d'organisation et de régulation (micro, méso, macro) peut trouver intérêt à utiliser le Big Data pour améliorer l'efficience du système de santé et sa soutenabilité pour les finances publiques.

Pourrait par exemple être construite une interface, ou « fenêtre », lisible et pratique à partir des données de l'établissement, voire de l'ensemble des établissements. Seraient ainsi développés des algorithmes de gestion, de qualité et de performance de l'activité au niveau d'un professionnel de santé ou d'un établissement.

Pour un établissement de santé donné, il serait possible de développer une application qui fournisse une description de l'activité d'un de ses services par rapport à d'autres services similaires dans d'autres établissements. Cela pourrait par exemple permettre à un réanimateur de prédire la sortie précoce d'un de ses patients sur la base de son activité passée, mais aussi de comparer ses pratiques et résultats avec ceux d'autres praticiens.

À un niveau plus large, cela permettrait également un croisement des sources de données pour améliorer l'organisation de l'offre ambulatoire et pour approfondir la compréhension des parcours de soins.

L'ensemble des acteurs, établissements, professionnels de santé, patients et régulateurs seraient gagnants au développement de tels usages.

5. Les usages en matière de recherche

Par le rapprochement de bases de données, et par ses méthodes statistiques non traditionnelles, le Big Data en santé permet de traiter de l'information différemment, de répondre à d'autres

questions, ou d'éclairer de vieilles questions sous un nouveau jour. Il peut soulever par ailleurs des questions inconnues jusqu'alors par la mise au jour d'associations insoupçonnées. Il permet également d'examiner des questions auxquelles ne pouvaient pas être apportées de réponses faute de moyens techniques. Plusieurs pistes peuvent déjà être identifiées. L'une porte sur le traitement plus approfondi et généralisé des données présentes dans les dossiers médicaux hospitaliers, éventuellement encore enrichies d'autres bases de données. Une autre consiste à davantage utiliser les bases de données médico-administratives en les appariant avec d'autres données, plus médicales. Une troisième consiste à davantage ouvrir et mutualiser certaines données dont disposent les industries de santé.

Le Big Data permet donc d'élargir le champ des recherches en santé. Il peut s'agir de recherches menées tant par des acteurs publics que par des acteurs privés.

6. Les usages pédagogiques, en matière de formation

Les usages pédagogiques recouvrent des catégories juridiques différentes, selon qu'il s'agisse de formation initiale des professionnels de santé, de formation continue, ou d'éducation thérapeutique des patients. Cela peut aller de la construction de simulations où des données permettent d'affiner les conséquences simulées des réponses aux simulations, à l'utilisation de bases de données pour détecter des situations de prise en charge de moindre qualité pour faire évoluer les formations reçues par les professionnels de santé.

7. Les usages de marketing et de ciblage

Des traitements de données de type Big Data peuvent être mis en œuvre pour améliorer la perception des coûts des assureurs en mesurant et quantifiant mieux certains risques – il faut alors veiller à ce que cela ne s'exerce pas au détriment des assurés, et que cet usage ne donne pas lieu à une sélection ou une segmentation des assurés. Ces traitements peuvent également aider un établissement donné dans le développement stratégique d'un de ses secteurs d'activité.

8. Un huitième usage correspond aux usages non-répertoriés ou inconnus à ce jour

Même si les sept usages répertoriés couvrent un champ très large, on ne peut exclure que des usages non identifiés à ce jour apparaissent, portés par de nouvelles sources de données, de nouvelles technologies, de nouveaux acteurs de santé.

*

* *

Au-delà de cette typologie, le Big Data comporte d'importants enjeux, que la partie suivante s'attache à décrire.

Troisième partie : enjeux et précautions

Cette partie développe deux types d'enjeux : des enjeux transversaux au Big Data ; puis des enjeux plus spécifiques au Big Data en santé, techniques, juridiques et socio-économiques.

Enjeux transversaux au Big Data

Rôles des acteurs, sécurité et garanties

Dans les différents usages présentés ci-dessus, les acteurs en présence sont nombreux et variés : ce sont ceux qui ont en charge de collecter les données, mais aussi de les traiter, ce sont également ceux qui vont les utiliser ou en bénéficier.

Des enjeux de sécurité, de confidentialité et de durée de conservation des données se posent, au-delà des questions de capacité de stockage : comme le soulignent les dispositions de l'article 6 de la loi relative à l'informatique, aux fichiers et aux libertés, le stockage ne doit pas persister plus longtemps que la durée nécessaire à la finalité du traitement : l'élargissement éventuel des finalités avec le Big Data rend l'identification de cette durée plus complexe.

Qualité scientifique

Un des enjeux du Big data en santé est son positionnement et sa reconnaissance dans le monde de la recherche, actuellement scindée entre les essais cliniques randomisés, qui constituent à ce jour la preuve scientifique la plus puissante en termes d'analyse causale et d'administration de la preuve, et la recherche observationnelle ou l'épidémiologie. Les premiers répondent à une question de recherche précise en construisant deux échantillons : l'un témoin qui reçoit le placebo et l'autre, le traitement d'intérêt. Par comparaison des résultats est mesurée l'efficacité du traitement. Le second type de recherche utilise le plus souvent des cohortes ou des bases de données traditionnelles et tire sa légitimité des méthodes statistiques employées, classiques pour l'essentiel. Il répond là-aussi à une question bien déterminée. Le Big Data diffère dans son approche : il mesure le plus souvent des corrélations/associations et possède ainsi un avantage prédictif. Il ne pose pas nécessairement de questions précises mais « laisse parler les données ».

Le Big Data comporte également un enjeu sur la qualité, la robustesse et la validité des algorithmes à éprouver. Il requiert la présence de professionnels compétents qui nettoient, qualifient, manipulent les données et les rendent exploitables.

Les usages envisagés du Big Data en santé ont souvent pour but d'améliorer la prévention, la prise en charge, la qualité des soins, la connaissance de pathologies, de facteurs de risque, d'interactions médicamenteuses, d'effets indésirables. La validité des résultats du traitement est donc un enjeu central – qui fait sans doute du Big Data en santé l'un des plus sensibles au sein de tous les Big Data.

Pour autant, ce questionnement est comparable à celui qui apparaît face à des données « classiques », c'est-à-dire moins volumineuses ou issues de sources de données plus structurées. Il est probable qu'il s'agisse plus d'interrogations d'ampleur nouvelle que d'interrogations véritablement neuves.

Différents enjeux restent, dans l'un et l'autre cas, centraux :

- qualifier les données d'origine, leur finalité d'origine, les contextualiser, connaître leur pertinence et leurs limites. C'est la diversité des sources utilisables qui soulève cette question avec plus de force en matière de Big Data. C'est aussi, en matière de numérique, un enjeu de représentativité, dans la mesure où les données utilisées, par exemple par des applications téléphoniques, proviennent d'utilisateurs au profil spécifique ;
- mettre en œuvre le croisement ou le rapprochement des données : le volume, la diversité des sources peuvent augmenter le risque d'erreurs, d'approximations ou le recours à des imputations (en cas de données absentes ou de qualité jugée insuffisante) ;
- sélectionner et appliquer les traitements de données : il s'agit ici des enjeux liés au choix d'une méthode contre une autre, d'un algorithme donné, des propriétés, notamment statistiques, des résultats produits. L'image de l'algorithme comme « boîte noire » tend sans doute à faire apparaître cet enjeu comme nouveau ; pourtant il se pose également dans bien d'autres cas. L'enjeu est celui de s'y former ;
- interpréter et présenter les résultats, qu'il s'agisse d'analyses portant sur des déterminants, ou de prédictions.

Ces questions ont une importance à moduler selon le type d'usages : une erreur de prédiction d'événements rares pour un assureur n'est pas aussi lourd de conséquence humaine qu'une mauvaise prédiction d'effets indésirables dans le traitement d'un patient.

Ces questions appellent une réflexion éthique impliquant professionnels de santé, patients, industriels, *data scientists*, administrations, le rôle des uns et des autres et leur articulation devant pouvoir varier et s'ajuster selon les différentes catégories d'usages.

Enjeux éthiques

Le Comité consultatif national d'éthique a été saisi sur ces questions et remettra un avis ultérieurement.

Les enjeux éthiques liés au Big Data en santé sont nombreux : contrôle de leurs données par les citoyens, transparence des algorithmes, individualisation de la médecine, etc. Certains de ces enjeux sont approfondis dans la section sur les enjeux sociétaux.

Enjeux liés aux infrastructures et standards techniques

Les éléments qui suivent retracent les pistes de réflexion ouvertes par le groupe de travail, qui devront être approfondies dans des travaux ultérieurs.

1. Les sources de données à considérer

Les données produites ou collectées par les établissements de santé occupent une place particulière parmi toutes les données pouvant alimenter le Big Data en santé. Les hôpitaux présentent une forte hétérogénéité et diversité en termes de systèmes d'information, ce qui complexifie l'accessibilité réelle à ces données. Une des réponses possibles à cette hétérogénéité repose sur la constitution d'entrepôts de données au niveau des établissements ou groupes d'établissements (notamment au sein des groupements hospitaliers de territoire). L'efficacité de cette solution est cependant renforcée par l'existence d'une politique globale du Big Data en santé intégrant les autres sources de données, dans laquelle s'inscrirait le développement de tels entrepôts.

Les données issues de l'activité ambulatoire (qu'il s'agisse de cabinets de ville ou de maisons de santé pluridisciplinaires) présentent le même problème de morcellement et d'hétérogénéité que celles des établissements hospitaliers. Pour répondre à ce problème, la création d'un référentiel d'interopérabilité minimal combiné avec une approche « API »[11] constitue probablement l'approche la plus adaptée.

D'autres sources d'informations personnelles relevant du « *quantified self* », des objets connectés, des réseaux sociaux, des forums sont également à prendre en compte. La qualité de ces données demande également à être davantage qualifiée.

2. Options envisagées pour faciliter l'accès à ces données

Le groupe a considéré que la question de l'interopérabilité est centrale dès lors que l'on souhaite relier des informations issues de sources différentes. Considérant le morcellement des systèmes d'information et la multitude des sources de données, il a estimé nécessaire, pour instaurer un minimum d'interopérabilité, de se prononcer sur le choix de différents standards d'interopérabilité européens ou internationaux.

Il a également conseillé de s'appuyer sur les technologies, dont certaines existent déjà, qui permettent d'interconnecter des sources diverses sans pour autant devoir encoder les données d'origine avec des terminologies de référence.

Par ailleurs, pour chaîner les informations au niveau individuel, trois options ont été esquissées :

- centraliser les données en un ou plusieurs centres proches des sources de production, tels que des entrepôts de données gérées par un ou des hôpitaux : le stockage et la structuration des données se réalisent en un même lieu ;
- s'appuyer sur une démarche centrée sur la personne : chaque individu centralise à son niveau les données qui le concernent, quelle que soit leur source (par exemple les informations dérivées du système de soins, des objets connectés, du carnet de santé

[11] Une API (pour *Application Programming Interface*) est une interface de programmation qui permet la communication et l'échange de données entre applications.

numérique, …). La personne pourrait choisir elle-même l'hébergeur et gérer les accès à ses données ;

- développer une API fédératrice afin de faciliter la connexion entre différentes sources de données (bases nationales, entrepôts locaux, données individuelles…). Elle faciliterait l'accès à de multiples sources « à la demande » (sur la base d'un projet, et non de façon permanente).

Le groupe de travail estime que dans la pratique, les trois options se juxtaposent forcément. Cependant, l'action de la puissance publique doit faciliter l'interconnexion des données sans passage obligé par des « méga-bases », via la conception d'architectures décentralisées et normées. Cela implique la mise en place d'une API fédératrice pour connecter les sources de données entre elles, que ce soit au niveau des logiciels métier, des entrepôts ou d'autres sources de données.

Pour des raisons de capacité de calcul et de sécurité, le groupe a par ailleurs estimé que l'objectif cible est la conception de procédures et de calculs pouvant s'effectuer de manière distribuée, là où les données sont localisées, pour éviter d'avoir à rapatrier systématiquement les données en un seul lieu pour effectuer les traitements Big Data.

Enfin, ces options doivent en cible inclure un dispositif (de type portail, par exemple) permettant à la personne de gérer les autorisations d'accès et d'usage de ses données dans une logique d'autodétermination informationnelle.

3. L'infrastructure proposée, fondée sur une API fédératrice et une plateforme de chaînage et de traitement

L'option d'une API fédératrice appelle à l'émergence d'opérateurs que le groupe de travail propose d'appeler « intermédiaires de confiance », chargés notamment de faciliter les opérations de rapprochement de données, en garantissant à la fois la qualité et la loyauté des traitements vis-à-vis des responsables des différentes sources mises à contribution.

Les intermédiaires de confiance développeraient ou s'appuieraient sur des plateformes fédérant des données à la demande en utilisant l'API fédératrice. Les traitements resteraient évidemment possibles au niveau de chacune des sources existantes, l'objectif n'étant pas de brider les initiatives des acteurs gestionnaires des sources de données, ni leur autonomie, mais au contraire de favoriser les mises en commun en apportant une garantie à chacun sur la nature des traitements réalisés et en apportant le cas échéant les expertises nécessaires.

Cette API fédératrice devra intégrer une couche de sécurité conforme aux exigences de l'Agence des systèmes d'information partagés de santé (ASIP) afin de respecter la confidentialité des données et l'accès exclusif dans le cadre d'un acte médical (utilisation de la Carte de Professionnel de Santé [CPS] et de la carte Vitale, par exemple).

Enjeux juridiques : outils juridiques, recueil du consentement, solutions techniques

Les réflexions du groupe de travail ont notamment porté sur la manière de simplifier les démarches entre la CNIL et les acteurs impliqués, en s'inspirant des outils existants développés par la CNIL, notamment :

- des méthodologies de référence : deux sont adoptées, une en cours d'écriture. Elles portent sur des traitements opérés dans le cadre de recherches biomédicale, dans le cadre d'études non interventionnelles de performances menées sur les dispositifs médicaux in vitro, dans le cadre de recherches non-interventionnelles (en cours d'élaboration) ;
- des autorisations uniques, notamment pour les données recueillies dans le cadre de pharmacovigilance, d'échanges par message sécurisé de données de santé, de données recueillies dans le cadre des ATU/RTU (autorisations/recommandations temporaires d'utilisation des produits de santé), et dans le cadre du dépistage organisé du cancer du sein et du cancer colorectal.

Le groupe a considéré que la notion de famille de finalités, assises sur la typologie des usages décrite dans la partie précédente, semble pouvoir constituer une piste permise par les textes, et notamment le règlement européen : reste à construire cette réflexion et à l'inscrire dans un cadre juridique pertinent.

S'agissant du recueil du consentement, la typologie des usages présentés ne cloisonne pas a priori chaque exemple d'application du Big Data en santé dans une seule catégorie d'usages, et il existe de même une grande diversité d'applications possibles au sein de chaque usage. Par ailleurs, des données collectées à l'origine pour un usage donné peuvent devenir utiles à d'autres fins ou pour d'autres usages, par exemple lorsque d'autres données complémentaires deviennent disponibles.

Cette possibilité ouvre des questions plus ou moins complexes au regard des droits des usagers : s'il paraît relativement aisé, pour une application en lien direct avec le patient ou l'usager, de revenir vers lui pour recueillir son consentement afin d'utiliser ses données pour de nouvelles finalités (par exemple une application sur *smartphone* peut générer une notification de mise à jour des conditions générales d'utilisation), la mise à jour du consentement est plus difficile à réaliser dans d'autres cadres, par exemple pour une base de données gérée par un hôpital.

Au-delà de la possibilité technique de recueil du consentement se posent deux questions d'importance. La première porte sur la possibilité de donner un consentement éclairé : on sait que cette capacité est inégale selon les personnes et que se posent des questions de lisibilité et d'intelligibilité de ce à quoi on demande de consentir ; il y a là un enjeu éthique majeur pour assurer un développement du Big Data dans le respect des droits et de la dignité des personnes. La seconde questionne la pertinence d'interroger sur le consentement d'un usager pour une catégorie d'usages et pour un élargissement des finalités. On pourrait envisager de traiter cet aspect de façon différenciée selon les usages. Il en va en effet autrement d'une recherche clinique et d'une application de coaching en santé.

Pour avancer sur ces questions difficiles et capitales, la Ministre des affaires sociales et de la santé a sollicité la CNIL, d'une part sur la question de la diffusion de données pour de multiples usages,

d'autre part sur l'environnement réglementaire permettant d'introduire une plus grande fluidité dans la mise en place d'applications de Big Data.

Après avoir délibéré, la CNIL estime que la loi Informatique et libertés permet déjà le développement du Big Data à des fins scientifiques, sous réserve d'autorisation préalable de sa part. « *Le consentement des personnes doit néanmoins être recueilli lorsqu'il est requis par les textes. Le règlement européen permet le recueil d'un consentement pour une ou plusieurs finalités spécifiques, ce qui implique que les finalités aient été déterminées et que la personne en ait été préalablement informée. La question de la transparence vis-à-vis des personnes concernées est donc essentielle et pourrait être effective à travers des outils permettant à ces personnes d'avoir une vision globale de l'utilisation des données qui les concernent (de type tableau de bord) et d'exercer, par exemple, un droit d'opposition modulé tenant compte de leurs choix* ».

Par ailleurs, la CNIL a réfléchi aux méthodes et outils pouvant permettre le recours aux données dans les conditions de sécurité requises. Elle mentionne d'une part les méthodes de pseudonymisation, d'autre part des méthodes plus nouvelles comme les PIMS (*personal information management systems*), le recours à un intermédiaire de confiance chargé de centraliser les données, ou encore la mise en place d' « open algorithmes » et le recours aux API.

De manière générale, la CNIL estime que « *les outils juridiques existants, qu'ils soient en vigueur ou à venir, offrent un cadre permettant aux traitements de big data de trouver leur place dans l'éco-système* ».

Enjeux sociétaux, enjeux économiques

Le Big Data ne concerne évidemment pas seulement la santé. Il peut être porteur d'innovations majeures dans bien d'autres secteurs (bancaire, militaire, transports, environnement, industrie, agriculture…,). La santé est néanmoins l'un des secteurs le plus souvent cités dans les rapports, études et ouvrages consacrés au Big Data, et cela pour plusieurs raisons : d'abord, parce qu'il pourrait permettre de développer progressivement une médecine de plus en plus personnalisée et pourrait donc contribuer à des progrès diagnostics, thérapeutiques et préventifs ; ensuite, parce qu'il pourrait permettre de mieux organiser l'offre de soins et donc de réaliser des gains d'efficience et de dépenser mieux ; enfin, parce que le secteur de l'assurance, dans son ensemble, pourrait connaître des transformations du fait du Big Data, et donc aussi en particulier l'assurance en santé.

Ces différents axes de progrès ne vont pas sans questionnements, sociétaux et économiques. Ces questionnements visent moins à relativiser l'apport du Big Data ou à s'en défier, qu'à rappeler que toutes les innovations technologiques s'insèrent dans un contexte économique, social et politique qui appelle des débats, des régulations et des garde-fous.

Ainsi, il est important de rappeler que le Big Data a aussi ses limites : l'occurrence de telle pathologie ou de tel effet secondaire sera toujours prédite avec une marge d'erreur car toutes les caractéristiques individuelles pertinentes ne sont pas mesurées et sans doute pas mesurables. L'incertitude ira toujours de pair avec la prévision, ce qui rend hautement improbable l'existence à terme d'une médecine prédictive entièrement automatisée et personnalisée. Devant la persistance

de cette variabilité entre personnes, voire de la variabilité dans le temps pour une même personne, il faut former et informer les patients, afin de sensibiliser à l'absence de réponse absolue ; il faut également sensibiliser les professionnels de santé, pour lesquels la gestion de l'incertitude devant la maladie est déjà courante, mais qui devront intégrer à leur pratique des prédictions toujours plus fréquentes et plus individualisées. Il est ainsi indispensable d'expliquer, d'accompagner une annonce telle qu'une prédiction à 30 % du risque de développer une pathologie. Le rôle des médecins sera central, à la fois pour expliquer, mais aussi pour conseiller et orienter les patients.

Ce risque d'individualisation est aussi important en matière de conception assurantielle. Si des dispositions régissent strictement les modalités de segmentation des contrats selon des caractéristiques individuelles, si des dispositions existent et ont été renforcées dans la loi de modernisation de notre système de santé pour organiser, en matière de santé, un droit à l'oubli interdisant de stigmatiser les personnes ayant été concernées par des problèmes de santé spécifiques, si l'état de santé est lui-même reconnu comme un critère au vu duquel une situation de discrimination peut être caractérisée (cf. art. 225-1 du code pénal), cette individualisation met néanmoins ces principes en tension et appelle à adapter le cadre éthique et réglementaire. L'exemple le plus parlant est celui de l'assurance en santé. La tension entre l'individualisation rendue possible par des données de plus en plus nombreuses, et la mutualisation qui est au cœur même du principe de l'assurance, est aujourd'hui un objet de réflexion et de préoccupation pour l'ensemble des acteurs de l'assurance[12]. On peut par exemple craindre que l'individualisation du risque en santé constitue une menace sur l'accès à l'assurance emprunteur de ceux dont le profil génétique, par exemple, sera considéré comme un mauvais risque par les assureurs. Cela étant dit, tous les usages du Big Data par des assureurs santé ne concernent pas directement la tarification individuelle ; certains des usages prioritairement envisagés à ce jour portent sur des services d'accompagnement, de prévention et d'information. Un autre usage du Big Data en développement chez les assureurs (comme, d'ailleurs, au sein des organismes de protection sociale) concerne la détection de la fraude.

Un autre enjeu du Big Data, souvent mis en avant dans les débats, concerne la « gouvernance algorithmique[13] », avec l'accent mis sur la transparence des algorithmes et la capacité des professionnels, des patients, des citoyens, de la puissance publique à comprendre leurs ressorts, leur manière de fonctionner, les nouvelles formes de « gouvernementalité » qu'ils impliquent, le risque qui doit être maîtrisé d'y recourir trop systématiquement au détriment de l'interaction humaine ou de la délibération démocratique. Le Conseil d'État[14] a ainsi préconisé d'encadrer l'utilisation des algorithmes, pour assurer l'effectivité de l'intervention humaine dans la prise de décision au moyen d'algorithmes, pour mettre en place des garanties de transparence lorsque les algorithmes sont utilisés pour prendre des décisions à l'égard d'une personne, et pour détecter l'existence de discriminations illicites. Les enjeux sont donc réels et importants. Pour autant, ils ne sont pas véritablement nouveaux et renvoient plus largement à la compréhension par le patient ou l'assuré des logiques de prise de décisions des acteurs (professionnels de santé, assureurs publics et privés, pouvoirs publics, etc.) ; il n'est pas certain que l'aide à la décision par des algorithmes soit systématiquement moins transparente ou plus difficile à s'approprier que la prise de décision fondée sur les compétences, l'expérience, l'intuition, le hasard, les réseaux ou les préjugés.

[12] Voir par exemple le dossier *Big Data et assurance,* dans la Revue Risques n°95, septembre 2013.
[13] Voir notamment les travaux de Antoinette Rouvroy ; par exemple cette intervention : http://socio.revues.org/1251
[14] Conseil d'Etat, Etude annuelle 2014, *Le numérique et les droits fondamentaux*.

Si l'on peut craindre, par exemple, que des algorithmes opaques mettent en œuvre des processus cachés de discrimination (que ce soit dans l'accès à des contrats d'assurance, à des traitements ou à des services), on peut imaginer aussi que dans certains cas les algorithmes permettent de lever des discriminations existantes. À titre d'exemple, des études (qui doivent encore être approfondies) laissent penser qu'il existe des inégalités sociales dans l'accès (qui n'est pas régi par des algorithmes) à la liste d'inscription pour la greffe rénale, alors que, une fois que les personnes sont inscrites sur la liste, ces inégalités seraient bien moindres pour l'accès à la greffe elle-même – l'attribution des greffons disponibles reposant sur un score calculé par un algorithme. Un algorithme étant une production humaine, son aptitude à générer des inégalités est tout aussi réelle que pour des choix humains plus classiques. Mais il est possible d'objectiver et connaître les règles et choix qui régissent le fonctionnement des algorithmes, à condition qu'ils soient suffisamment transparents - ce qui peut poser par ailleurs des questions de propriété et d'équilibre entre l'intérêt privé et l'intérêt général en fonction des usages.

La question de la propriété ou du partage des données est, elle aussi, souvent au cœur du débat sur le Big Data. Plutôt qu'un droit de propriété, le Conseil d'Etat a préconisé un droit à l'autodétermination, tendant à « garantir en principe la capacité de l'individu à décider de la communication et de l'utilisation de ses données à caractère personnel ». D'autres acteurs se réfèrent à un droit de propriété, s'accompagnant donc de la vente des données, notamment par les patients et les citoyens eux-mêmes. Cette conception appelle trois types de commentaires.

D'une part, cela pourrait être en contradiction avec les principes de partage nécessaire au Big Data, puisque chaque acteur pourrait alors bénéficier seul de sa « propriété ». On ne peut néanmoins passer sous silence la crainte que les grands acteurs de l'internet et du numérique puissent dégager des bénéfices importants en exploitant les données des internautes qui, eux, n'y gagneraient rien ; on peut alors s'interroger sur la nécessité de rechercher un équilibre – la question étant de savoir si cet équilibre se trouve ou non dans la monétisation des données.

D'autre part, le Conseil d'Etat souligne que cela compliquerait l'exercice de la régulation par les pouvoirs publics.

Enfin, une vision propriétaire des données de santé est un déterminant majeur des modèles économiques qui pourraient émerger. Si une partie de la valeur réside certes dans la production et le stockage des données (raison pour laquelle des acteurs privés investissent dans la constitution et la vente de bases de données), l'essentiel de la valeur ajoutée se trouve en réalité dans les usages qu'on en fait : l'utilisation des données nécessite une compétence et un savoir stratégiques, sources d'une plus grande valeur – d'autant plus grande que les données se multiplient et se diversifient.

Néanmoins, on peut redouter que certains détenteurs de données, et donc de la matière première, s'organisent oligopolistiquement pour « contrôler » le marché et en tirer un bénéfice. C'est un effet pervers auquel le régulateur doit veiller, dans le domaine du Big Data en santé comme dans bien d'autres.

La valeur des usages, si elle est potentiellement immense, n'est le plus souvent pas connue a priori. Elle se découvre au fur et à mesure de la mise en œuvre des projets. Les modèles économiques du Big Data devront alors prendre en compte cette dimension, en développant et en inventant des règles de partage de la valeur a posteriori.

Ce qui est sûr, c'est que le Big Data, en santé comme dans les autres domaines du monde socio-économique, va augmenter la valeur, mais aussi la déplacer sur toute la chaîne de production de la santé. L'objectif prioritaire est que cette augmentation et ce déplacement aient lieu au bénéfice des citoyens et des patients ; mais c'est, aussi, l'occasion qu'ils bénéficient aux professionnels de santé et à la place internationale de la France dans l'industrie de la santé au sens large.

Les éléments qui précèdent incitent à considérer le Big Data en santé non comme un tout, mais d'une manière différenciée selon les acteurs et les finalités. Ce débat doit impliquer l'ensemble des citoyens. Les pouvoirs publics ont évidemment tout leur rôle à jouer, pour réguler les usages, garantir la qualité des référentiels de sécurité, favoriser l'interopérabilité des bases, orienter les financements publics, développer les formations dans les *datascience*, etc.

*

* *

La CNIL appelle à un déploiement du Big Data dans une stratégie cohérente, fruit d'une concertation avec les acteurs qui pourront mobiliser les outils adéquats en s'appuyant sur les attentes des citoyens.

L'avis de l'atelier citoyen recoupe les principaux enjeux identifiés dans ce document et met l'accent sur plusieurs catégories de risques :

- La protection des droits de la personne, au sens large : la protection de la vie privée, la protection contre les discriminations ;
- Le risque de creusement des inégalités ;
- Un enjeu autour de la capacité des patients à faire face à une responsabilisation accrue et à un flux de données ininterrompu sur leur santé ;
- Un risque d'une hyper-segmentation et d'une démutualisation en assurance.

Cet avis considère que les bénéfices attendus du Big Data en santé surpassent les risques, à condition de mettre en place les garde-fous nécessaires. Il appelle les pouvoirs publics à se saisir du sujet.

En cohérence avec ces positions, les travaux des groupes de travail et le colloque du 4 juillet 2016 organisé par le ministère des affaires sociales et de la santé, constituent des étapes importantes de la réflexion et de la concertation, qui auront vocation à se poursuivre et à se traduire en actions.