DOI : 10.4267/2042/54093

Le Big data en Santé préfigure-t-il la « médecine 3.0 » ?

Big Data in Health foreshadowing "3.0 medicine"?

Laurent Guigue, Christophe Richard

Membres du groupe de travail open data en santé/big data en santé du SYNTEC
christophe.richard@santeos.com

Résumé

En France, les données de santé constituent d'ores et déjà une manne partiellement exploitable à travers les entrepôts de données existants : Système national d'information inter-régimes de l'Assurance maladie (SNIIRAM), Programme de médicalisation des systèmes d'information (PMSI), registres…). D'autres sources de données voient le jour grâce à de nouveaux vecteurs d'information et à de nouveaux comportements individuels et collectifs, aboutissant à la production massive de données de santé sur internet ; il s'agit par exemple des réseaux sociaux, du crowd sourcing, des moteurs de recherche, des objets connectés (Quantified Self et télémédecine). L'exploitation de ces « Big Data » répond à de multiples objectifs tels que la veille sanitaire, la pharmacovigilance, l'épidémiologie, la médecine personnalisée, etc. Mais pour favoriser cette exploitation, les initiatives en rapport avec l'Open Data (ouverture des données de santé) doivent conduire à la définition d'un cadre éthique, sécuritaire et raisonné : à partir des données individuelles et personnelles, il sera ainsi possible de construire des outils au bénéfice de la collectivité et de la santé publique, sous condition du respect de la vie privée.

Mots-clés

Big data ; Open data; SNIIRAM (Système national d'information inter-régimes de l'Assurance maladie) ; Santé ; Epidémiologie ; Veille sanitaire ; Pharmacovigilance ; Crowd sourcing ; Quantified self ; Médecine personnalisée ; Anonymisation

Abstract

In France, health data are already partially exploitable through existing data warehouses (SNIIRAM, PMSI, registries…). Other data sources are emerging through new information channels and new individual and collective behaviors, leading to a mass production of health data on the Internet : for example, social networking, crowd sourcing, search engines, connected objects (Quantified Self and telemedicine). The exploitation of these « Big Data » addresses multiple objectives such as health monitoring, pharmacovigilance, epidemiology, personalized medicine, etc. But to support this, the initiatives related to Open (health) Data must lead to the definition of an ethical, safe and sensible framework: from the individual and personal information, it will be possible to build tools for the benefit of the community and the public health, while respecting the right to privacy.

Keywords

Big data ; Open data; SNIIRAM (Système national d'information inter-régimes de l'Assurance maladie) ; Health ; Epidemiology ; Health watch ; Drug monitoring ; Crowd sourcing ; Quantified self ; Personalized medicine ; Anonymisation

La donnée de santé : une matière première qui se bonifie avec le temps

La quantité de données issue de la prise en charge d'individus dans un cadre sanitaire, médical, médico-social ne cesse d'augmenter, de même que le nombre de sources de données disponibles. Si l'on associe

ce constat aux évolutions technologiques, chaque individu peut ainsi espérer bénéficier d'une médecine prédictive, préventive, personnalisée et participative.

Cette évolution (révolution ?) place le Big Data au cœur du système de santé tant l'origine de ces données est variée :

- Dossiers médicaux électroniques dans les cabinets médicaux, les hôpitaux, les centres d'imagerie, les laboratoires, les pharmacies et d'autres structures productrices de soins ;
- Données médico-économiques en rapport avec les dépenses de santé et la consommation de soins et de produits de santé (SNIIRAM, PMSI…) et plus généralement les bases de données publiques médico-administratives[1];
- Données de la recherche et de l'industrie pharmaceutique ;
- Données biométriques provenant de dispositifs et d'objets connectés qui surveillent automatiquement des paramètres tels que le poids, la tension artérielle, le taux de glucose dans le sang, etc. [2];
- Données saisies par les individus : préférences personnelles, niveaux de satisfaction, historique de consommation, données d'auto-mesure, etc. ;
- Données génomiques, apportant une réduction significative du temps et des coûts associés au séquençage génétique ;
- Données sur les autres déterminants de santé, tels que ceux liés à des facteurs socio-économiques et à l'environnement ;
- Données échangées à travers les médias sociaux.

Le Big Data en santé sera-t-il la clé de voute de la Santé Publique de demain ?

La gestion de ces données massives est un important levier pour une meilleure compréhension des maladies, du développement de médicaments et du traitement des patients. Une étude de l'Institut McKinsey a ainsi mis en avant en 2013 les principaux secteurs pour lesquels le Big Data apporterait de réels bénéfices :

- La prévention ciblée, permettant d'évoluer vers un mode de vie toujours favorable à la préservation de l'état de santé des individus ;
- L'aide au diagnostic et à la mise en place des soins adaptés à chaque patient tout en assurant sa sécurité, en tendant ainsi vers la médecine personnalisée ;
- L'optimisation du médicament pour obtenir l'impact clinique attendu, et l'optimisation des ressources médicales par la mise à disposition de professionnels adaptés au cas du patient ;
- La maîtrise des coûts pour une qualité de soin égale ou meilleure, en automatisant les remboursements et la détection de fraude ;
- L'innovation, en encourageant la R&D et la sécurité, une meilleure exploitation et diffusion des connaissances.

La France n'est pas absente de ces réflexions ; la Commission des affaires sociales de l'Assemblée nationale a saisi l'Office parlementaire d'évaluation des choix scientifiques et technologiques (OPECST) d'une demande d'étude sur les enjeux scientifiques, technologiques, éthiques et juridiques de la médecine personnalisée[3]. Il est notamment évoqué l'intérêt d'initier une réflexion *« sur les moyens de concilier d'une part, l'indispensable partage scientifique des données à partir de larges tailles d'échantillons permettant des méta-analyses et d'autre part, le respect de la confidentialité et de la sécurité de l'accès aux données […] des personnes »*, mais avec l'espoir que *« l'individu participe à la collectivité en mettant ses données au service de l'analyse globale et qu'avec beaucoup de chances il lui revienne quelque chose ».*

---

1. Cartographie des bases de données publiques en santé https://www.data.gouv.fr/dataset/cartographie-des-bases-de-donnees-publiques-en-sante.
2. Le plan européen "eHealth action plan 2012-2020"prévoit la parution en 2014 d'un livre VERT sur la « santé mobile » : http://epractice.eu/en/library/5362646
3. Rapport intermédiaire de l'OPECST sur les enjeux scientifiques, technologiques, sociaux et éthiques de la médecine personnalisée : http://www.assemblee-nationale.fr/14/cr-oecst/medecine_perso_rapport_provisoire.pdf.

Cette approche est d'autant plus chargée d'opportunité en termes de santé publique que la France est un des rares pays qui dispose de bases de données médico-sociales et économiques d'envergure nationale, centralisées, constituées et gérées, couvrant de façon exhaustive et permanente l'ensemble de la population dans divers domaines stratégiques (recours aux soins, hospitalisation, prestations et situation professionnelle et sociale[4]).

Hors de nos frontières, les pays scandinaves et le Canada ont mis au service de la santé publique et de la recherche leurs systèmes d'information médico-sociaux, créant ainsi de véritables « *Population Data Centers* », ouverts à la communauté scientifique.

Ces démarches ont facilité la réalisation de très nombreuses études de grande qualité (par exemple le centre mis en place à la *British Columbia University*[5]. Une approche «Big Data» sur ces bases de données laisse entrevoir des avancées scientifiques de premier ordre.

La surveillance épidémiologique : l'une des premières applications pratiques du Big Data

Parmi les pistes prometteuses figure la surveillance sanitaire, notamment épidémiologique, avec certains avantages par rapport aux systèmes traditionnels de veille, en particulier en termes de réactivité. De nouveaux services ont ainsi vu le jour, certains mis en œuvre par des industriels, d'autres par des institutions sanitaires pouvant disposer de jeux de données exploitables. Les producteurs de soins sont désormais aussi dans la course : la moitié des hôpitaux utilisera un logiciel d'analyse de pointe d'ici 2016, comparativement à 10 % aujourd'hui [1].

C'est ainsi que Google Flu Trends[6] est apparu en 2008-2009 pour le suivi de la grippe saisonnière à travers 18 pays, puisque les hashtags de Twitter ont été utilisés par la Food Standards Agency britannique pour surveiller les pics épidémiques de gastro-entérite durant l'hiver 2012-13[7]. En France, l'entreprise Celtipharm publie depuis peu sur openhealth.fr des cartes épidémiques réactualisées chaque jour, à partir des achats réalisés dans un réseau de plus de 4 000 pharmacies.

Un peu de prudence est néanmoins de mise dans l'exploitation et l'analyse qui sont faites de certaines données. Selon une étude publiée dans la revue *Science* et relayée par *The Register*, les prédictions de pics d'épidémie réalisées par Google seraient erronées pour 100 des 108 semaines écoulées depuis 2011. La raison invoquée est simple : « *presque tout le monde croit que le moindre rhume est une grippe* » et « *les internautes qui recherchent le terme «grippe» ne sont pas tous malades* ». Les chercheurs remettant en question la qualité du service rendu pensent même que « *l'outil peut générer un cercle vicieux en faisant croire aux internautes que la grippe est arrivée dans leur région et en les poussant à se croire malades de la grippe lorsqu'ils ont en réalité une maladie bien plus bénigne* » [2] !

La pharmacovigilance en temps réel est-elle possible ?

La pharmacovigilance représente un autre enjeu critique en termes de veille sanitaire, particulièrement mis en évidence lors de l'affaire du Mediator. Ici encore, l'accès à certaines sources de données et leur exploitation à des fins de santé publique pourraient apporter une surveillance plus réactive en générant des alertes et en permettant des prises de décisions des pouvoirs publics adaptées au contexte. Ceci pourrait s'inscrire par ailleurs dans une plus grande dynamique de transparence à l'égard du grand public[8].

---

4. Haut Conseil de la santé publique : Rapport Pour une meilleure utilisation des bases de données nationales pour la santé publique et la recherche : www.hcsp.fr/explore.cgi/hcspr20120309_bddadmination.pdf.
5. https://www.popdata.bc.ca/
6. http://www.google.org/flutrends/
7. http://tna.europarchive.org/20140107172105/http://blogs.food.gov.uk/science/entry/social_media_for_social_good
8. http://www.opendatasante.com/2014/02/22/723/

Le Big Data est-il soluble dans l'open data en santé?

Big Data n'est pas synonyme d'*Open Data* : les données de santé ne sont pas toutes accessibles. La rançon de cette sécurité - justifiée lorsqu'il s'agit de données à caractère personnel - se matérialise par des difficultés à pouvoir les exploiter. En France, après la circulation d'une pétition depuis janvier 2013 pour « libérer les données de santé » et la remise d'un rapport de l'IGAS (Inspection générale des affaires sociales) au ministère de la santé en octobre dernier[9], une démarche est désormais engagée afin d'intégrer l'accès aux données de santé dans la future loi de santé, de façon très encadrée : l'état a en effet souhaité l'ouverture d'un débat sur l'ouverture des données publiques de santé, porté par une commission[10] associant les différents acteurs concernés. Cette commission a été installée par Marisol Touraine, Ministre des affaires sociales et de la santé, le 21 novembre 2013 et a remis ses travaux le 9 juillet 2014 sous la forme d'un rapport et d'une doctrine visant à guider les décisions publiques[11].

L'accès aux données devra ainsi garantir le respect de la vie privée et de l'anonymat, et leur « ouverture » ne sera pas sans limite en raison des risques de ré-identification indirecte. Un dispositif juridique, technique et organisationnel en sécurisera donc l'accès et l'utilisation, à travers la mise en place d'une gouvernance adaptée. Les données du Système d'information inter-régimes de l'assurance-maladie (SNIIRAM) sont bien entendu concernées, puisqu'il s'agirait du plus vaste entrepôt de données de santé au monde, consolidant chaque année 500 millions d'actes médicaux et 11 millions de séjours hospitaliers.

Dans le cadre de la pharmacovigilance, un récent avis[12] de la *commission d'accès aux documents administratifs* (CADA) pourrait d'ailleurs avoir des conséquences sur les applications du Big Data en santé, et notamment sur la capacité à exploiter des données du SNIIRAM. La CADA indique en substance : *« que les données dont le [demandeur] sollicite la communication, si elles revêtent un caractère médical, ne constituent pas un extrait des données sources de la base mais correspondent, après traitement automatisé d'usage courant de ces données, à des informations anonymes et globales, par année et par département, ne permettant pas, compte tenu de leur niveau d'agrégation, l'identification, même indirecte, des patients ou des médecins concernés ».*

La CADA préconise donc, non pas un accès direct aux données *publiées* dans le SNIIRAM, mais à ce qu'il contient *après réalisation des traitements prévus et autorisés*.

La mise en œuvre de traitements automatisés, préservant la vie privée tout en donnant accès à d'immenses sources de données, ouvrirait ainsi de nouveaux horizons en termes de démocratie sanitaire, d'autonomisation du patient, d'efficacité de l'action publique et surtout d'innovation et de recherche.

Une médecine bijective et réflexive à la fois…

Les réseaux sociaux représentent manifestement une manne pour le Big Data, mais le « crowd sourcing médical » n'est pas seulement exploitable à travers les commentaires, les questions et les inquiétudes échangés par des internautes sans que ceux-ci soient nécessairement conscients que ces données vont servir à dépister une épidémie. Le crowd sourcing pourrait devenir réellement délibéré, motivé et ciblé : à la manière dont des patients sont enrôlés dans des essais cliniques, ils pourraient souhaiter devenir des contributeurs de masse de données anonymes permettant de développer ainsi un véritable Big Data de santé international au service de la recherche, dans le cadre d'une médecine plus *participative*.

En marge des médias sociaux, le grand public commence également à générer et partager des données personnelles de santé ou de « bien-être » à travers des dispositifs tels que des montres intelligentes et bracelets connectés qui surveillent le sommeil, l'exercice physique et la consommation de calories, la fréquence cardiaque, etc.

Parallèlement à l'arrivée de smartphones disposant de fonctions permettant à chacun de surveiller son état de santé par l'intermédiaire de capteurs, des constructeurs tels Apple et Samsung travaillent sur

9. Rapport de Pierre-Louis Bras, inspecteur général des affaires sociales (IGAS) et d'André LOTH, Directeur de projet Formation dans les domaines de la santé et des solidarités. Celui-ci a été rendu à la ministre de la santé et des affaires sociales en octobre 2013 : http://destinationsante.com/wp-content/uploads/2013/11/rapport-donnees-de-sante-2013.pdf.
10. http://www.drees.sante.gouv.fr/commission-open-data-sante,11250.html
11. http://www.drees.sante.gouv.fr/rapport-de-la-commission-open-data-en-sante,11323.html
12. Avis de la CADA: http://www.cada.fr/article20134348,20134348.html

des dispositifs et des technologies leur permettant de pénétrer le marché du « Quantified self » et de la santé connectée[13].

Bref, l'alimentation du Big Data à travers l'internet des objets semble inévitable, exploitant indifféremment les données provenant d'objets nomades, mais aussi du domicile. La domotique fait désormais bon ménage avec des solutions permettant d'assurer le maintien à domicile de personnes en situation de dépendance, âgées et/ou atteintes de pathologies chroniques. Différents services peuvent ainsi se décliner à travers un « habitat intelligent et connecté », de la téléconsultation à la télésurveillance, en passant par des systèmes assurant la coordination de soins et d'aides à domicile, la livraison de repas, des services de coaching nutritionnel ou destinés à l'éducation thérapeutique, etc. Les supermarchés peuvent aussi connaître les habitudes alimentaires des consommateurs, et même le réfrigérateur de la cuisine peut être connecté à l'internet afin de suivre ce qui est consommé et ce qui doit être réapprovisionné.

De la médecine individualisée à la médecine personnalisée

Dans un autre domaine, l'étude du génome et des relations entre les maladies et leurs facteurs de risque génétiques et environnementaux permet d'explorer le champ de la médecine personnalisée en optimisant le traitement d'un individu donné. Mais le fait est que les technologies de l'information ne peuvent pas (encore) suivre le rythme de l'explosion des données, en particulier pour assurer un « suivi génétique global » des cancers qui nécessiterait le recoupement des caractéristiques de millions de tumeurs. Les jeux de données recueillis par les cliniciens sont tout simplement bien trop volumineux pour être partagés ou échangés sur les réseaux en place. A titre illustratif, aux États-Unis, en estimant un volume de 100 Go pour une seule tumeur, appliqué à près de 2 millions de nouveaux cas de cancer par an et 14 millions de personnes en vie atteintes d'un cancer, on totaliserait des dizaines voire des centaines de péta-octets de données par an (soit des centaines de millions de giga-octets) [3].

Perspectives et limites

Outre la protection de la vie privée, il reste encore de nombreux défis à résoudre, et celui de la liaison entre les données figure en bonne place. Alors que le Web est actuellement un réseau de documents (pages web), certains envisagent l'élaboration d'un « Web de données ». Traditionnellement, les données de santé sont reliées entre elles grâce aux identifiants des patients, mais ce nouveau web sémantique offre désormais la possibilité de relier des données hétérogènes, qui pourraient inclure des personnes, des lieux, des produits ou des concepts. Les méthodes et les technologies pouvant contribuer à cette évolution du Web n'en sont qu'à leurs débuts et sont particulièrement ambitieuses [4]. Or, le nombre d'analystes spécialisés est encore très insuffisant, en particulier sur les aspects liés à la santé des individus et à la qualité de leur prise en charge. Dans un premier temps, la priorité a en effet été plutôt accordée à la valeur médico-économique des données et non à leur valeur médicale et sanitaire [5]. Ce qui peut se justifier quand on constate que les applications du Big Data pourraient contribuer à réduire de 300 milliards de dollars les coûts liés au système de santé américain grâce à la prévention et à la médecine personnalisée [6].

Le Big Data serait-il le Graal en permettant d'améliorer la santé tout en baissant les dépenses de santé ?

De la causalité à la corrélation

La sécurité représente un enjeu majeur en raison de la capacité à identifier un individu à travers des caractéristiques pourtant déclarées anonymisées (par exemple génétiques), ou le recoupement de ces caractéristiques avec d'autres bases de données publiques. Dans l'état de l'Illinois, des données anonymisées de sorties d'hospitalisation, de recensement et de listes électorales ont ainsi pu être

13. http://www.macplus.net/depeche-76214-avec-sami-samsung-veut-s-installer-dans-les-bracelets-sante http://www.cnetfrance.fr/produits/ios-8-iwatch-et-iphone-6-comment-apple-veut-revolutionner-les-objets-connectes-39798845.htm

recoupées avec des données génétiques également anonymisées, permettant de ré-identifier (entre autres) 50 % des patients atteints de maladie de Huntington. Certains scientifiques vont même jusqu'à affirmer que « l'anonymat est devenu algorithmiquement impossible »[14].

Le cœur du problème ne réside pas tant dans les mesures à prendre pour protéger les données que dans l'identification des nouvelles menaces induites par le Big Data, les outils permettant d'exploiter les données étant en train d'apparaître avant les mécanismes de sécurité pour nous en protéger.

Il existe donc un véritable cadre éthique et sécuritaire à définir au niveau international afin de garantir le bon usage de ces données, sans risque pour la vie privée [7], et il convient de se prémunir contre toute dictature des données qui nous conduirait à des principes de précaution « prédictifs » puisqu'en raison de l'immensité des données, des décisions pourraient être prises non plus par les humains, mais par des machines à l'image des dérives du trading haute fréquence dans le secteur boursier[15].

Délivrer éthiquement et durablement plus de valeur et de qualité au meilleur coût

Nous sommes à la veille d'une révolution en Santé. Il est évident que les données de santé ont une valeur, qu'elles sont une opportunité et un enjeu majeur en santé et qu'elles constituent une ressource pour les outils de Big Data.

Le Big Data progresse dans un écosystème naissant et pas encore stabilisé mais les « paramètres sont connus » et maîtrisables.

Le débat ne doit pas opposer bénéfices « individuels » aux bénéfices « collectifs ». Chacun devra s'attacher à construire une économie du système qui ne sera pas basée sur la « vente des données », mais sur l'utilisation des enseignements qui découleront des traitements effectués.

Si l'objectif est de transformer les données en connaissances et de coproduire des protocoles de soins[16] personnalisés en temps réel pour transformer la médecine curative en une médecine préventive, les efforts n'auront pas été vains.

Références

1. Khan I.. U.S. Hospital Health Data Analytics Market - Growing EHR Adoption Fuels A New Era in Analytics. Frost & Sullivan, 7 Aug 2012. http://www.frost.com/c/10046/sublib/display-report.do?id=NA03-01-00-00-00.
2. Chirgwin R. Google flu-finding service diagnosed with 'big data hubris' - Bad data contagion overwhelms prediction service. The Register, 23 Mar 2014. http://www.theregister.co.uk/2014/03/23/goggle_flu_foo_fubar/.
3. Manos D. Health IT not keeping pace with big data. Healthcare IT News, February 12, 2014 http://www.healthcareitnews.com/news/health-it-not-keeping-pace-big-data.
4. Pope C, Halford S, Tinati R, Weal M J. What's the big fuss about 'big data'? Health Serv Res Policy, April 2014;19:67-68. http://hsr.sagepub.com/content/19/2/67.long.
5. Walberg R. Value of big data in health care is measured not just in dollars, but in lives. Financial Post, February 5, 2014 http://business.financialpost.com/2014/02/05/value-of-big-data-in-health-care-is-measured-not-just-in-dollars-but-in-lives/?__lsa=c350-1536.
6. Groves P, Kayyali B, Knott D. The 'big-data' revolution in healthcare: Accelerating value and innovation. McKinsey & Company, January 2013 http://www.mckinsey.com/insights/health_systems/~/media/7764A72F70184C8EA88D805092D72D58.ashx.
7. Malin B, Sweeney LJ. How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. Biomed Inform 2004;37:179-92. http://www.j-biomed-inform.com/article/S1532-0464(04)00053-X/fulltext.

Liens d'intérêt : aucun

14. http://www.technologyreview.com/news/514351/has-big-data-made-anonymity-impossible/
15. http://patrimoine.lesechos.fr/patrimoine/placement/actu/0203432670468-nouvelle-charge-contre-le-trading-haute-frequence-663587.php
16. http://www.decisionsante.com/derniere-minute/article/big-data-quand-les-donnees-vont-bousculer-lhopital/?tx_ttnews[backPid]=1&cHash=13b8d87888