HEALTHCARE
DATA INSTITUTE

INTERNATIONAL THINK TANK
DEDICATED TO BIG DATA IN HEALTHCARE

BIG DATA ET PRÉVENTION
DE LA PRÉDICTION À LA DÉMONSTRATION

NOVEMBRE 2016

BIG DATA
ET PRÉVENTION
DE LA PRÉDICTION À LA DÉMONSTRATION

Remerciements aux rédacteurs de ce Livre blanc :

- **Isabelle Barbier-Feraud**, Pagamon
- **Jeanne Bossi Malafosse**, avocate
- **Patrice Bouexel**, TeraData
- **Catherine Commaille-Chapus**, Open Health Company
- **Anne Gimalac**, Aston Life Sciences
- **Guillaume Jeannerod**, Epiconcept
- **Magali Léo**, CISS
- **Bruno Leroy**, Sanofi
- **Bernard Nordlinger**, AP-HP, Académie de médecine
- **Michel Paoli**, InterMutuelles Assistance
- **Pablo Prados**, Sanofi
- **Jean-Yves Robin**, Open Health Company

INTRODUCTION

Vouloir prévenir les épidémies, identifier le plus en amont possible les facteurs de risque pour éviter les maladies ou ralentir leur développement, prédire dès la naissance des prédispositions à telle ou telle infection… Autant d'ambitions mille fois exprimées par les scientifiques et les représentants du corps médical.

Jusqu'à présent il était répondu à ces questionnements par le recours aux méthodes traditionnelles de la recherche qui ont et sont toujours en grande partie fondées sur la notion de reproductibilité : à partir d'une hypothèse formulée sur le fondement de laquelle sont collectées des données, une observation permet d'opérer une relation entre des données et d'en déduire de nouvelles hypothèses ou de nouveaux traitements qui seront reproduits à des cas similaires.

Depuis plusieurs années, cette recherche s'est enrichie par d'autres sources très diverses prenant ainsi en compte le milieu dans lequel les individus évoluent : données environnementales, données socio-professionnelles, etc.

Cette situation est aujourd'hui réinterrogée par l'irruption des nouvelles techniques de gestion des données qui doivent s'adapter à la production massive et exponentielle de données caractérisées par le phénomène du Big Data – nous reviendrons sur la définition de ce terme.

Cette production massive de données peut-elle avoir un effet sur la prévention des maladies, le Big Data a-t-il un rôle à jouer, parmi d'autres, pour améliorer le niveau sanitaire de la population et

permettre ainsi aux politiques de disposer des bonnes informations au bon moment pour définir une politique de santé publique plus intelligente ?

Le développement du numérique au cours des dernières années a également conduit à une véritable « datification » de nos sociétés. Les données sont partout et constituent une matière première précieuse pour la création de nouvelles connaissances et un enjeu mondial de croissance majeur pour de nombreux pays.

Il se caractérise également dans le secteur de la santé par l'informatisation croissante du secteur professionnel, notamment des activités de soins et de prévention, de recherche en sciences de la vie, de gestion des systèmes de santé et par l'implication croissante des patients.

Les phénomènes de mobilité, le développement des objets et dispositifs médicaux connectés contribuent aussi à la croissance exponentielle du volume de données produit, le phénomène du Big Data ne faisant que traduire les spécificités du traitement de grands volumes de données avec leurs exigences de rapidité de traitement, d'hétérogénéité des données et de création de valeur particulière.

Les données de santé représentent donc des enjeux singuliers et un potentiel de création de valeur dont la concrétisation dépendra de la capacité des pays à organiser le développement d'un écosystème facilitant leur exploitation tout en garantissant le respect de la vie privée et la confidentialité des données personnelles.

Les États sont en effet aujourd'hui confrontés à des enjeux majeurs pour lesquels le traitement des données de santé peut jouer et joue déjà un rôle essentiel : des enjeux de santé publique, des enjeux de qualité des soins, des enjeux de transparence et de démocratie sanitaire, des enjeux d'efficience des systèmes de santé dans un contexte généralisé de croissance des dépenses de santé et des enjeux d'innovation et de croissance dans des domaines aussi variés et importants que la médecine personnelle ou médecine de précision et les technologies de l'information.

Le Healthcare Data Institute qui a entamé depuis plus de deux ans maintenant une réflexion sur le sujet du Big Data et de la san-

té a constitué un groupe de travail spécialement dédié au sujet du Big Data et la prévention. Son objectif a été de retenir quelques exemples qui illustrent le rôle que peut jouer le Big Data dans la prévention sanitaire, d'identifier les facteurs qui favorisent son développement et ceux qui les freinent et de proposer quelques pistes de réflexions qui pourraient s'avérer utiles à la décision qu'elle soit de nature privée comme publique.

Le travail que traduit ce livre blanc est le fruit de réflexions et de travaux de personnes aux métiers différents dans des environnements professionnels privés comme publics. Il n'a évidemment pas vocation à être exhaustif sur le sujet, mais davantage à attirer l'attention sur des domaines qui, aujourd'hui intègrent les nouvelles potentialités du Big Data et, dans le même temps, mettent en exergue certaines problématiques de nature différente. ■

DÉFINITIONS

Pour présenter ce travail, il a été nécessaire de s'accorder sur deux définitions importantes, celle du Big Data et celle de la prévention.

Le Big Data désigne des ensembles de données qui deviennent si volumineux qu'ils sont difficiles à traiter avec les seuls outils de gestion de base de données ou les outils classiques de gestion de l'information. Le Big Data désigne aussi l'ensemble des technologies, infrastructures et services permettant la collecte, le stockage et l'analyse de données recueillies et produites en nombre croissant, grâce à des traitements automatisés et le recours aux technologies de l'intelligence artificielle[1].

Il est d'usage d'illustrer cette définition avec l'image des « 3V » qui caractérisent le Big Data : explosion du **volume** des données, **variété** des données structurées et non structurées produites par de multiples sources et **vélocité** de l'information et vitesse de traitement simultanée.

À ces trois V, s'ajoute un quatrième, la **valeur** que représentent ces données pour l'entreprise ou l'individu. La donnée est devenue la matière première de l'univers numérique.

On perçoit aisément aujourd'hui la confusion souvent faite entre le terme utilisé de Big Data et l'usage de méthodes d'analyse de données comme le Data Mining, ce qui a pour conséquence d'utiliser le terme de Big Data de façon très extensive pour désigner l'ensemble des nouvelles méthodes d'analyse de données.

Le groupe de travail a toutefois délibérément pris le parti d'adopter cette définition large du terme Big Data pour adapter son champ d'analyse à la réalité du sujet : la question étant davantage aujourd'hui de savoir comment constituer des bases de données intelligentes pour arriver à faire du Big Data.

1. Rapport de l'Institut Montaigne sur Big Data et objets connectés 2015.

Dès lors, dans ce Livre blanc, le terme de Big Data sera retenu de façon souple comme regroupant l'ensemble des techniques et méthodes utilisées aujourd'hui pour analyser des masses de données de plus en plus volumineuses et produites par des sources variées.

S'agissant de **la prévention**, qui est également un terme vaste, nous avons retenu la définition donnée par l'Organisation mondiale de la santé (OMS). Selon cette organisation, la prévention est « *l'ensemble des mesures visant à éviter ou réduire le nombre ou la gravité des maladies ou accidents* ».

Trois types de prévention sont ainsi distingués :

• La prévention primaire qui recouvre « *l'ensemble des actes destinés à diminuer l'incidence d'une maladie, donc à réduire l'apparition de nouveaux cas* ». Elle fait appel à des mesures de prévention individuelle (hygiène corporelle, alimentation…) et/ou collective (distribution d'eau potable, vaccination…).

• La prévention secondaire qui recouvre « *tous les actes destinés à réduire la prévalence d'une maladie donc à réduire sa durée d'évolution* ». Elle comprend le dépistage et le traitement des premières atteintes.

• La prévention tertiaire qui concerne « *tous les actes destinés à diminuer la prévalence des incapacités chroniques ou des récidives dans la population donc à réduire les invalidités fonctionnelles dues à la maladie* ». Elle a pour objectif de favoriser la réinsertion sociale et professionnelle après la maladie. Cette définition étend la prévention aux soins de réadaptation. ■

SOMMAIRE

Les sujets abordés dans ce Livre blanc à travers une série d'articles s'organisent autour de thèmes horizontaux et transversaux. Chaque article peut être lu de façon indépendante.

I. Un premier thème aborde le rôle du Big Data et de la prévention personnalisée à partir du sujet de la génomique.

II. Le deuxième thème s'attache à la prévention populationnelle universelle et illustre les possibilités de gestion des données de vie réelle au service de la santé publique et de la prévention des risques sanitaires.

III. Le troisième thème retenu traite du sujet essentiel de l'évolution du paysage technologique comme un facteur clé de succès pour « faire parler les données ».

IV. Le quatrième thème aborde la question de la définition d'un nouveau paradigme pour les acteurs.

Le choix de ces quatre sujets a permis aux membres du groupe de travail d'identifier pour chacun d'eux des problématiques communes et transversales qu'il a paru intéressant d'étudier et qui peuvent donner lieu à de recommandations.

Il s'agit de l'article **V.**, sujet important de la nature anonyme ou pseudonymisée des données utilisées et du bon niveau d'agrégation à retenir, et de l'article **VI.** sur les sujets éthiques et juridiques abordés ici essentiellement à travers le thème des patients et de l'ouverture des données.

ARTICLE 1
BIG DATA & GÉNOMIQUE : QUEL AVENIR DANS LE TRAITEMENT DU CANCER ?

1. CONTEXTE & TECHNOLOGIE

Il se dit que nous sommes entrés dans l'ère de la Médecine 6P (Personnalisée, de Précision, Participative, Préventive, Prédictive et orientée Patient) fondée sur une approche génomique et moléculaire de la maladie qui offre l'opportunité de :

- mieux comprendre les mécanismes pathologiques ;
- identifier de nouvelles cibles thérapeutiques ;
- identifier les facteurs de risque ;
- accompagner le diagnostic et la prise de décision médicale ;
- personnaliser le traitement.

Qu'en est-il ?

Pour une pathologie donnée, le séquençage de tout ou partie du génome d'un patient permet d'identifier les différences d'orthographe (appelés variants) en comparaison du génome d'individus qui n'ont pas la pathologie en question. Ces dernières années, des avancées technologiques considérables dans le domaine du séquençage à haut débit (méthode de déchiffrage du génome) ont permis de réduire temps et coûts ainsi que de faciliter le stockage et l'analyse de ces données génomiques de masse. Cela pourrait laisser présager que bientôt le séquençage du génome entier (et pas seulement de certaines parties ciblées) des individus et des patients se ferait en routine, ouvrant ainsi leurs applications à visée autant de recherche et que de thérapie…

Où en sommes-nous réellement aujourd'hui ? Peut-on vraiment parler de Big Data dans le domaine de la génomique ? Y parviendrons-nous dans un futur proche ou lointain ? Pour quels résultats et avec quels obstacles à surmonter d'ici là ?

Dans la suite de notre exposé, nous prenons le parti de décrire l'exemple du domaine du cancer.

2. LA RECHERCHE SCIENTIFIQUE ET LE TRAITEMENT MÉDICAL DU CANCER AUJOURD'HUI

Des progrès considérables se poursuivent actuellement dans la compréhension et le traitement des cancers, fondés à la fois sur une meilleure compréhension de l'impact de facteurs environnementaux, sur l'observation clinique, sur des essais cliniques de nouveaux traitements pharmaceutiques, sur les progrès en immunothérapie (notamment pour les mélanomes) et en biopsies liquides (et pas seulement tissulaires).

Nous assistons également au développement de l'analyse génomique des tumeurs elles-mêmes et du génome des patients. À titre d'exemple à l'Institut national du cancer (INCA) français, 70 000 nouveaux patients chaque année voient leur génome être séquencé de façon ciblée sur certaines régions, puis être analysé en routine afin d'identifier des anomalies moléculaires et de personnaliser leur traitement pour gagner en efficacité et diminuer les effets secondaires. Il est prévu que d'ici 2020 le séquençage du génome entier se fasse en routine, ce qui devrait ainsi faciliter une compréhension systémique de la maladie.

De plus, nous assistons à un changement dans la façon de développer de nouvelles molécules issues de la recherche académique et industrielle, et de démontrer leur efficacité au cours d'essais cliniques réglementés. Il s'avère que dorénavant les patients recrutés dans les essais cliniques de Phases I, II et III peuvent être sélectionnés par rapport à leur profil génomique. D'anciennes études de Phase III sont notamment ré-analysées à la lumière de la découverte de nouveaux **bio-marqueurs** (quelques marqueurs génétiques spécifiques exprimés par la tumeur) pour évaluer l'intérêt de thérapies ciblées pour les tumeurs de patients ayant un profil génomique associé à une meilleure réponse au traitement, en d'autres termes qui permettent de prédire comment une tumeur répondra à un traitement donné.

L'idée est de disposer de données pour administrer aux patients le traitement le mieux adapté possible à leur maladie, pour augmenter leur chance de guérison. En effet, il est désormais bien établi que tous les patients atteints de cancer ne réagissent pas de la même façon à un traitement. Cela dépend à la fois de caractéristiques propres aux patients et à leur tumeur. Nous recherchons ainsi des bio-marqueurs corrélés à une meilleure réponse à un traitement, et des bio-marqueurs permettant d'estimer le risque de rechute d'un cancer.

Nous parlons de **bio-marqueurs pronostiques** quand ils nous permettent de mieux prédire la durée de vie ou durée de vie sans récidive du patient, comme un autre moyen d'améliorer sa prise en charge.

Dans le cas du cancer colorectal, nous avons pu établir que seuls les patients dont la tumeur porte une version normale (non mutée) d'un gène nommé RAS peuvent bénéficier des effets du cétuximab et panitumumab, deux nouveaux médicaments (anticorps monoclonaux) de thérapie ciblée. Nous parlons dans ce cas de **bio-marqueurs prédictifs** qui nous permettent de mieux prédire l'efficacité d'un traitement.

Or ici, nous ne pouvons toujours pas considérer être dans une approche Big Data de la génomique, en terme de données de masse ou de diversité de sources notamment. Même si l'on sait aujourd'hui croiser les informations génétiques qui caractérisent la tumeur et la réponse à une thérapie pour quelques centaines ou milliers d'individus, le choix du traitement basé sur le séquençage du génome n'est pas encore entré en routine systématique dans la pratique médicale.

À l'heure actuelle, l'application des Big Data en génomique n'occupe qu'une petite place parmi les méthodes de recherche et traitement des maladies, notamment le cancer.

Des initiatives dispersées vont tout de même dans ce sens. En voici quelques exemples (liste non exhaustive) : en Angleterre (*The 100 000 genomes project*), en Allemagne (génotypage systématique des tumeurs au *DKFZ* - German Cancer Research Center à Heidelberg, un modèle à suivre), aux États-Unis (*CancerLinQ*, *Cancer Moonshot*), en France (*France Médecine Génomique 2025*, INSERM/

Quest Diagnostics « *BRCA Share* » pour améliorer le diagnostic de prédisposition aux cancers du sein et de l'ovaire), en Suisse (*Personalized Health*), en Europe (*SPECTAcolor*, *Million European Genomes Alliance*, *Sophia Genetics**), etc.

*Il est intéressant d'observer un nouveau courant d'application de Big Data, à la frontière entre la biotechnologie, la génomique, la santé digitale et l'intelligence artificielle. Par le biais d'algorithmes sophistiqués, 170 hôpitaux dans 28 pays européens utilisent à l'heure actuelle des services d'analyse et d'interprétation de profils génomiques de leurs patients du cancer (à partir de données mutualisées) afin d'accompagner la prise de décision de leurs médecins concernant le degré de pathogénicité de leurs mutations génétiques. Le but ici n'est pas que l'algorithme remplace le diagnostic par le médecin, mais est plutôt d'accélérer le diagnostic de cas « simples » afin de libérer du temps au médecin pour le diagnostic de cas plus complexes, mettant en jeu l'accumulation de plusieurs mutations génétiques.

3. QUEL POURRAIT ÊTRE L'APPORT DE L'ANALYSE DES DONNÉES GÉNOMIQUES DE MASSE DANS CE DISPOSITIF ? QUE POURRAIT-ON ESPÉRER DES BIG DATA ?

En épidémiologie, l'analyse des données de masse, en croisant les données génétiques / génomiques des patients et de leurs tumeurs, avec des données externes environnementales (géographie, alimentation, pollution atmosphérique, rayonnements UV, etc.), des données de santé connectée, etc., améliorerait considérablement l'efficacité des méthodes actuelles, permettrait de mieux identifier les facteurs de risque qui influencent la survenue des cancers et donc de contribuer à leur prévention.

Tout progrès dans ce domaine dépendra de la possibilité de croiser les données cliniques et génétiques de tumeurs et de patients,

donc de la constitution de bio-banques et bases de données prospectives et rétrospectives, et de leur accès non restreint bien que strictement réglementé en termes de protection de la donnée de santé. **On sera réellement rentré dans l'ère du Big Data en génomique quand on saura accumuler les données génomiques et diagnostiques de grand nombre de patients, les données cliniques avant/pendant/après traitement, etc., permettant de définir des corrélations potentielles entre génotypes et phénotypes et ainsi de prédire l'influence de facteurs environnementaux, la susceptibilité à des maladies ou incidents thérapeutiques. Ces bases de données prospectives et rétrospectives seront dédiées à la fois à la recherche, au diagnostic et au traitement, et pourront permettre d'adresser des questions soulevées en s'affranchissant d'un but initial de collecte.**

Pr Pierre-Laurent PUIG, auditionné par le groupe de travail, médecin au service de biochimie de l'hôpital européen Georges Pompidou (Paris) et directeur de l'unité de recherche INSERM UMR-S775 à l'université de médecine Paris Descartes. Sujet de recherche : Les cancers colorectaux.

« Aujourd'hui, la génération de données génomiques est facile et pas chère. En termes d'application, nous ne pouvons plus dissocier la recherche scientifique du traitement médical, car nous analysons les données de séquençage brutes liées pour caractériser une pathologie spécifique et pour adapter le traitement des patients atteints par cette pathologie. Pour autant, en France, nous faisons face à un frein réglementaire (ni technologique, ni financier) de la protection des données santé qui n'autorise pas à croiser et à corréler les données génomiques et les données cliniques du dossier médical, et qui n'autorise pas d'utiliser les données à des fins autres que celles pour lesquelles elles ont été générées et collectées. Les bases de données sont constituées en rétrospectif. Le prospectif, accumulant des données collectées sans but premier, sera possible quand nous pourrons mettre en place des dossiers médicaux électroniques qui seront dès le départ centralisés et constitués à des fins de gestion et de recherche sans but préétabli. »

4. QUELS SONT LES OBSTACLES À SURMONTER POUR PARVENIR À UNE MÉDECINE 6P BASÉE NOTAMMENT SUR LE BIG DATA ET LE DIAGNOSTIC MOLÉCULAIRE ?

Nous listons ici ce qui nous semble être les principaux défis technologiques, organisationnels, politiques, économiques.

A. DÉFIS TECHNOLOGIQUES

Le marché du logiciel et de l'algorithme requiert :

1) l'ouverture aux données digitalisées / numérisées (fondamental pour la traçabilité, même si cela induit de nouveaux risques) ;
2) la connaissance métier ;
3) la standardisation, la répétabilité.

Il s'avère donc nécessaire de :

• Développer des dossiers électroniques patients standardisés intégrant données génomiques et données cliniques classiques.

• Créer l'écosystème permettant de faire conjuguer les compétences d'ingénieurs, d'informaticiens, de généticiens moléculaires, de médecins généticiens, etc.

• Développer des algorithmes, des applications, des processus de confidentialité et de sécurisation des données afin de permettre aux médecins praticiens de rendre un verdict, un diagnostic et un schéma thérapeutique clairs.

• Faire valider les algorithmes par tous ces experts pour en garantir la robustesse, la facilité d'utilisation et l'applicabilité (IBM Watson n'est applicable que pour la pratique médicale américaine !).

Mais aussi de :

• Mettre en place un réseau multidisciplinaire pour faire face aux besoins exponentiels de séquençage à haut débit, de collecte, de stockage, d'analyse à grande échelle, de décryptage des bases de données les rendant compréhensibles par le praticien.

B. DÉFIS ORGANISATIONNELS

• Garantir la confidentialité des données génétiques et veiller au respect d'un cadre réglementaire pour éviter les discriminations.

• Préparer les médecins à la révolution génomique qui s'annonce, les former au changement de pratique, les préparer au décloisonnement des spécialités médicales (la maladie n'étant plus considérée sous l'angle de sa localisation organique, mais sous l'angle moléculaire et cellulaire).

• Développer les nouveaux outils avec les praticiens eux-mêmes afin d'en faire des ambassadeurs qui contribueront à faire évoluer la pratique et à concevoir la médecine du futur.

• Mettre en œuvre un travail en synergie avec d'autres métiers (bio-informatique, statistiques…) pour déterminer le protocole le plus adapté.

• Faire évoluer la relation médecin / patient pour accompagner les patients exclus des sous-groupes pour lesquels un médicament a été ciblé, pour accompagner les patients à forte prédisposition génétique, pour éduquer les patients à de nouveaux comportements préventifs, pour une meilleure compréhension du patient permettant son consentement éclairé.

• Réorganiser la recherche scientifique et technologique en fonction de l'évolution des pratiques médicales et cliniques, penser à de nouveaux montages publics / privés, etc.

C. DÉFIS POLITIQUES ET ÉCONOMIQUES

• Faire évoluer le système de santé, déterminer les conditions et niveaux de prise en charge par l'assurance maladie.

• Constituer une filière industrielle autour de la santé génomique et du digital, de mobiliser l'industrie du séquençage, du stockage de données et de l'instrumentalisation scientifique, de concevoir un nouveau modèle pour l'industrie pharmaceutique.

• Rapprocher, le plus en amont possible de la chaîne de valeur, les acteurs comme l'industrie pharmaceutique, avec les structures hospitalo-universitaires, pour avoir accès aux génomes de cohortes de patients et identifier des bio-marqueurs associés aux pathologies.

ARTICLE II
DATA ET PRÉVENTION DE LA SANTÉ DES POPULATIONS

En santé comme dans tous les domaines, les progrès technologiques ont fait exploser la quantité d'informations recueillies à chaque instant. L'exploitation et l'analyse de ces volumes croissants de données disponibles auprès d'une variété de sources qui les collectent pour diverses raisons sont porteuses d'espoirs et de progrès considérables en matière d'avancées scientifiques et ouvrent des perspectives nouvelles en matière de prévention de la santé des populations et de maîtrise des risques sanitaires.

La prévention de la santé des populations est une des fonctions de la santé publique, définie comme la « science et l'art de favoriser la santé, de prévenir les maladies et de prolonger la vie grâce aux efforts organisés de la société[2] ». La prévention recouvre dans cette acception l'ensemble des mesures ayant pour but d'éviter la survenue de maladies et d'accidents ou de réduire leur nombre, leur gravité et leurs conséquences.

Le Big Data se fonde sur la capacité à collecter, agréger et traiter des données issues d'une multiplicité de sources hétérogènes. La variété des données collectées (structurées, non structurées, etc.) et des sources (bases publiques / privées, données médico-administratives, données de santé / données d'environnement, etc.), leur volume, la vélocité avec laquelle elles sont recueillies et traitées sont au cœur des démarches de Big Data . La diffusion des objets connectés ainsi que le développement des pratiques de géolocalisation contribuent à l'accroissement considérable du volume de données disponibles. L'analyse de ces données massives, rendue possible tant par la numérisation croissante des activités de la société que par l'accroissement des capacités de stockage et de traitement, révolutionne les approches traditionnelles en matière de prévention.

2. D Nutbeam, OMS, 1998.

L'analyse des données a toujours été au cœur de la compréhension des phénomènes sanitaires ; la connaissance des autorités publiques, nécessaire à l'action, se fonde historiquement sur des actions d'observation épidémiologique, qui visent à observer et mesurer les événements sanitaires touchant une population donnée, à les expliquer, à mesurer l'impact des mesures prises pour les traiter.

Ce qui est nouveau aujourd'hui avec les techniques du Big Data , c'est cette capacité nouvelle à agréger et traiter des volumes massifs de données, issues de sources différentes, couplée au fait qu'est désormais disponible un patrimoine de données numériques considérable, tant dans le champ sanitaire (données de soins, données médico-administratives, données produites par les patients eux-mêmes via les objets connectés, etc.) que plus largement sur les conditions socio-économiques, culturelles, environnementales qui sont des déterminants clés de la santé des populations, de sorte qu'il n'est plus nécessairement besoin de produire ces données de façon spécifique. Le règlement européen sur la protection des données personnelles adopté le 27 avril 2016 consacre ainsi, pour la première fois, la notion de « finalité compatible », autorisant la réutilisation des données pour une finalité autre que celle pour laquelle elles ont été collectées initialement dès lors que cette finalité est compatible avec la finalité première.

Là où les études en santé nécessitaient jusqu'alors la collecte de données de façon *ad hoc*, sur la base de critères à partir desquels étaient renseignées les bases de données, souvent sur des plages de temps longues ; des données produites à l'occasion d'actes de soins ou à des fins de remboursement, ou encore par les patients-citoyens eux-mêmes via les objets connectés ou les réseaux sociaux sont désormais disponibles et constituent une source presque inépuisable pour l'identification des facteurs de risque de maladie, la sécurité sanitaire ou l'épidémiologie.

Le traitement et l'analyse de ces données massives encore largement sous-utilisées offrent des perspectives de production de connaissances démultipliées. Ces analyses peuvent contribuer à la prévention des maladies, des épidémies, à la surveillance et à la veille sanitaire en permettant de mieux comprendre les déterminants socio-économiques et environnementaux de la santé des populations,

de détecter des événements de santé inhabituels susceptibles de constituer une alerte de santé publique, éventuellement d'établir des liens avec des facteurs d'exposition. Elles peuvent permettre de cibler les efforts de prévention vers les groupes de populations pour lesquels ces mesures sont le plus efficaces. Elles peuvent contribuer au suivi et à l'évaluation d'actions de santé publique.

Les paragraphes qui suivent illustrent la variété des possibilités ouvertes par le Big Data en matière de prévention de la santé des populations, de veille et de sécurité sanitaire.

1. LES PERSPECTIVES D'EXPLOITATION BIG DATA EN PHARMACOVIGILANCE

Dans un pays qui se classe parmi les plus forts consommateurs européens de médicaments, la question des effets indésirables liés à leur consommation et de la surveillance de leur bon usage s'impose de plus en plus comme un sujet de santé publique majeur. Les récentes crises sanitaires (Mediator, pilules de 3e et de 4e génération, Depakine, etc.), dont l'impact tant en termes de santé publique qu'en termes économiques est à l'évidence considérable, illustrent les difficultés auxquelles doivent faire face les autorités sanitaires et la solidarité nationale face à un mésusage des médicaments plus massif qu'ailleurs doublé d'un déficit de données précises sur les conditions d'utilisation des médicaments[3]. Laisser perdurer des mésusages ou découvrir avec dix ans de décalage des effets graves est à la fois dramatique en termes de santé des populations et extrêmement coûteux pour la collectivité. Au niveau européen, le poids des effets indésirables liés aux médicaments est estimé à 197 000 morts par an dans l'UE. L'impact économique est évalué à 79 milliards d'euros par an[4].

3. B. Bégaud, D. Costagliola, *Rapport sur la surveillance et la promotion du bon usage du médicament en France,* mission sur la pharmacosurveillance confiée par la ministre des Affaires sociales et de la Santé, madame Marisol Touraine, le 26 février 2013.

4. H. Pontes, M. Clément, V. Rollason, *Safety Signal Detection : The Relevance of Literature Review*, Spinger International Publishing Switzerland 2014.

La surveillance de la sécurité du médicament repose essentiellement sur les actions d'évaluation pouvant être menées postérieurement à sa commercialisation et sa diffusion dans la population générale, auprès de groupes de populations plus larges et plus divers dans leurs caractéristiques que ceux des essais cliniques, et pour lesquels l'utilisation du médicament (durée de traitement, posologie, observance, etc.) est elle-même beaucoup plus variable[5]. L'analyse des données issues de cohortes ou des bases médico-économiques sur le long terme peut permettre de détecter des signaux et de faire des rapprochements entre la survenue d'un événement de santé et l'exposition à tel ou tel traitement et déclencher une alerte sur de possibles effets indésirables des médicaments ou sur une utilisation hors de l'indication pour laquelle tel ou tel produit de santé a reçu une autorisation de mise sur le marché.

La rapidité d'accès à la connaissance en la matière est tout à fait clé. Dans le cas de la détection d'événements indésirables graves, les délais en années qui peuvent s'écouler entre la suspicion d'un phénomène et l'établissement d'un niveau de preuve suffisant peuvent s'avérer très préjudiciables. Les techniques de fouille de type Big Data peuvent être mobilisées au service de ces analyses et permettre une détection précoce des signaux d'alerte tout à fait déterminante en santé publique.

2. LE BIG DATA DANS LA PRÉVENTION ET LE SUIVI DES ÉPIDÉMIES

La disponibilité de données massives issues de sources différentes, en temps réel ou quasi réel, permet de renseigner sur l'état de santé d'une population dans une zone géographique donnée. Le croisement et l'analyse de ces données, combinés aux techniques mathématiques de modélisation, peuvent permettre de repérer des ruptures de tendance annonciatrices de l'élévation de l'incidence de maladies ou de comportements et conduire à anticiper une évolution probable de l'état sanitaire de cette population.

5. J.L. Faillie, F. Montastruc, J.L. Montastruc, A. Pariente, *L'Apport de la pharmaco-épidémiologie à la pharmacovigilance*, 2016.

Ainsi, à côté des dispositifs traditionnels de recueil d'informations et d'alerte sur les maladies infectieuses et les épidémies (comme le réseau Sentinelles en France, composé de médecins répartis sur le territoire qui remontent chaque semaine les cas observés pour un certain nombre de maladies transmissibles), se développent des modèles épidémiologiques de suivi de la propagation spatio-temporelle des phénomènes sanitaires comme les épidémies à partir de l'utilisation des données massives.

Depuis un peu plus d'une dizaine d'années, l'usage du Big Data a démarré au sein de la nouvelle agence de santé publique qui regroupe l'InVS, l'INPES et l'EPRUS et qui a pour mission de protéger la santé des populations. Cet usage réside avant tout sur l'utilisation de données en provenance de sources de données définies et connues.

Dans les 3 V qu'on associe fréquemment au Big Data, c'est celui de Variété qui est prépondérant.

En effet, à partir de sources de données validées et stables comme le PMSI, le SNIIRAM, le CepiDC ou encore les données en provenance des registres des cancers des indicateurs robustes dans le temps et validés scientifiquement ont été élaborés.

Sursaud (Surveillance sanitaire des urgences et des décès), le système de surveillance syndromique mis en place après la canicule de 2003 et fondé sur des données en provenance des services d'urgence, de SOS Médecins, des données de mortalité et des certificats de décès en est un exemple.

Ces indicateurs permettent de produire le BQA (Bulletin quotidien des alertes) destiné aux autorités sanitaires, dont le ministre en charge de la santé.

Un tel système est performant pour détecter un événement sanitaire inattendu, estimer l'impact d'un événement environnemental ou sociétal, surveiller des pathologies en dehors de tout événement, ou détecter précocement un événement sanitaire prédéfini, tel qu'une épidémie saisonnière, en mesurer l'impact et les conséquences.

Si les différents systèmes mis en place sont efficaces pour détecter un événement ou confirmer une tendance, ils ne permettent pas ou peu d'en expliquer la causalité dans les délais imposés par la réponse à ces derniers et la pression politique.

Pour rappel, en Île-de-France en mars 2014, une augmentation de l'asthme a été constatée et attribuée dans un premier temps au pic de pollution qui sévissait en même temps. *A posteriori*, il s'avère qu'une quantité plus importante qu'à l'accoutumée de pollen circulait dans l'air à ce moment et que la pollution n'était qu'un tiers facteur.

L'exemple de *Google Flu Trend*, destiné à suivre l'évolution de la grippe dans le monde à partir de l'analyse de certains mots-clés du moteur de recherche, a été maintes fois mis en avant. Une analyse comparative des méthodes de détection avec d'autres modalités de suivi de la propagation du syndrome grippal a montré les limites de ce modèle quant à la détermination de l'amplitude du phénomène. Google a été contraint de réintégrer des données provenant des centres américains de contrôle et de prévention des maladies (CDC) et le service a depuis été fermé.

L'utilisation des données des téléphones mobiles est un exemple tout à fait intéressant, ces données peuvent en effet permettre de décrire avec un grand niveau de précision les effets que peuvent avoir les rassemblements et les mouvements de population sur la propagation d'une épidémie comme le choléra ou la malaria. La compréhension de ces phénomènes qui permet l'identification de « points de transmission » d'une épidémie est riche de promesses pour participer à l'éradication de maladies infectieuses[6].

D'autres initiatives comme l'étude de la fréquence de consultation de certaines pages de Wikipédia (la page sur la bronchiolite par exemple) ou l'augmentation de mots-clés sur Twitter permettent de détecter des signaux faibles.

S'il convient, comme le montre ces exemples, de rester prudent sur la validation de ces analyses, qui exige des observations recueillies à partir de longues séries dans le temps et dans l'espace, et si ces modèles n'ont pas encore révélé leur plein potentiel, il n'en reste pas moins que s'ouvrent là des perspectives tout à fait considérables en termes de prévention et de contrôle des pathologies notamment infectieuses.

6. Flavio Finger, Tina Genolet, Lorenzo Mari, Guillaume Constantin de Magny, Noël Magloire Manga, Andrea Rinaldo and Enrico Bertuzzoa, *Mobile phone Data highlights the role of mass gatherings in the spreading of cholera outbreaks.*

Puissance des algorithmes et augmentation des sources de données apporteront avec le temps la garantie que tout événement sanitaire sera détectable.

En revanche, il est et il sera primordial de garder du temps et de l'expertise médicale portée par des femmes et des hommes de terrain pour faire le tri dans la masse d'événements détectés et comprendre et associer de manière fiable un événement sanitaire et sa cause, ce qui n'est pas forcément compatible avec la société de l'information qui se développe au même rythme que ces technologies.

3. CROISEMENT DE DONNÉES CLINIQUES ET SOCIO-DÉMOGRAPHIQUES : VERS UNE AMÉLIORATION DU FACTEUR PRÉDICTIF EN MATIÈRE D'ADHÉRENCE ?

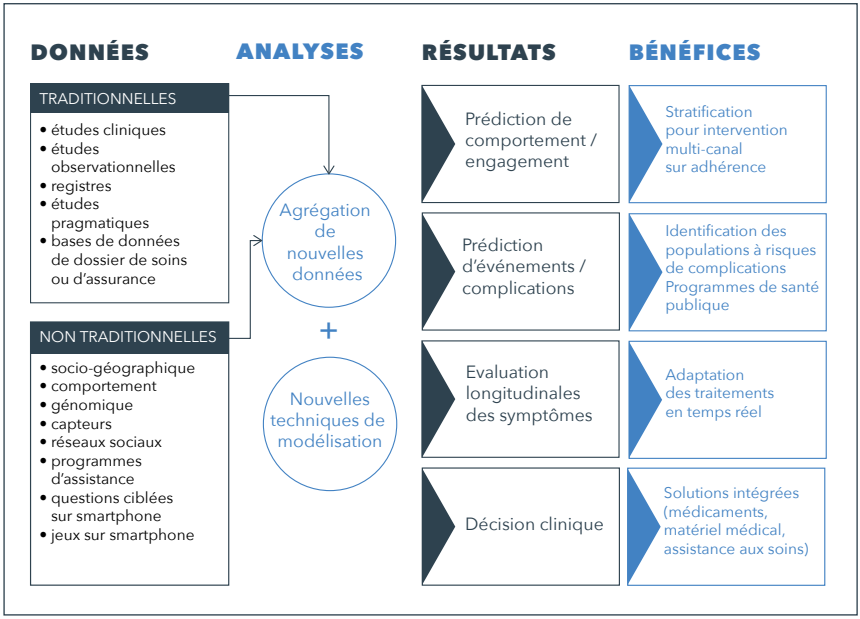
Les données sociogéographiques, traditionnellement non utilisées en santé, pourraient être intégrées dans la conception de programmes d'intervention de santé publique plus personnalisés, par la modélisation utilisant des techniques avancées d'apprentissage machine.

Ces techniques permettent de développer des modèles de prédiction d'adhérence au traitement, et de survenue de complications comme les événements cardiovasculaires aigus ou les épisodes d'hypoglycémie sévère chez le patient souffrant de maladies chroniques comme le diabète. Les programmes d'interventions de santé publique pourraient alors être optimisés en privilégiant l'utilisation de certaines ressources pour les catégories de populations les plus exposées à ces complications, fondés sur leur risque de non-adhérence au traitement évalué en fonction de certaines caractéristiques géographiques ou de comportement socio-économique. Des modèles similaires à ceux utilisés dans la grande consommation pour prédire un comportement d'achat sont actuellement testés dans cette optique. La plupart des systèmes de recommandations font des suggestions de choix personnalisées, fondés sur l'expérience des comportements d'achats précédents. Ces approches n'ont pas

encore été largement utilisées en santé publique et représentent une source possible d'amélioration de la qualité des soins.

Cette approche nécessite de lier tout d'abord de façon anonyme des données de remboursement avec des données médicales démographiques et d'évolution de la maladie pour identifier des facteurs prédictifs d'adhérence au traitement et de complications des maladies concernées. Les données socio-économiques et géographiques sont ensuite rajoutées dans l'analyse pour comparer ou améliorer cette prédiction et ainsi guider l'intensification ou la simplification des programmes d'intervention.

De nombreuses données socio-économiques et géographiques sont désormais disponibles en accès libre dans certains pays comme les États-Unis (données de recensement, dépenses des consommateurs, étude nationale des salaires, environnement alimentaire, enquête de logement). Un point critique néanmoins est la taille de la région concernée pour maintenir l'anonymat des données médicales. Des données de comportement du patient peuvent aussi être obtenues à partir d'applications digitales, permettant par exemple le suivi de l'activité sociale et physique à partir d'un smartphone, collec-



tées directement, ou à l'aide de questionnaires simplifiés administrés de manière régulière.

Les types de données analysables, les opportunités et l'impact possible sur la prise en charge des patients sont résumés dans le graphe ci-dessus.

4. BIG DATA ET INDIVIDUALISATION DE LA PRÉVENTION

Le champ de la prévention recouvre aussi bien les actions collectives visant à protéger la santé des personnes par le développement d'un environnement physique et social favorable à la santé (veille sanitaire, vaccination, dépistages de certaines maladies, préservation de la qualité de l'air et de l'eau, etc.) que par la promotion des comportements individuels favorables à la santé (prévention de l'alcoolisme et du tabagisme, promotion de l'activité physique et conseils nutritionnels « manger bouger », etc.).

La prévention s'adresse donc à l'ensemble de la population et s'exerce dans tous les lieux de vie des individus : au domicile, au travail, à l'école, dans la cité, etc.

Historiquement, la mise en œuvre des programmes de prévention se fait par des actions de masse qui portent bien souvent la dénomination de « campagne » :

- campagne de vaccination ;
- campagne de dépistage ;
- campagne de sensibilisation vis-à-vis de comportements à risque, etc.

Ces campagnes, menées par les organisations en charge de la santé des populations sont peu spécifiques dans le sens où elles s'adressent au plus grand nombre, souvent sur la base d'études de type coût/efficacité. Le bénéfice individuel est alors une conséquence du bénéfice populationnel :

- Informer toute la population via un message radio ou télévisé pour lancer la campagne de vaccination contre la grippe tout en insistant sur la population des personnes âgées.

• Proposer systématiquement tous les deux ans une mammographie à l'ensemble des femmes de 50 à 74 ans.

Dans les deux exemples cités ci-dessus, on voit que l'on va solliciter très largement pour augmenter l'impact populationnel. Il y a peu de ciblage, les personnes concernées sont censées se reconnaître dans le dispositif mis en place.

Pour les campagnes de dépistage, en cas d'un test positif, le suivi médical et les éventuels traitements seront individualisés dans un parcours de soins personnalisé en aval de l'action de dépistage.

Aujourd'hui, la disponibilité des données et l'intelligence que l'on peut obtenir en croisant ces dernières permettent de s'adresser au plus grand nombre tout en ciblant les messages, en réorientant les personnes à risque, ou en excluant les personnes non concernées.

En utilisant les mêmes techniques que celles apparues avec Internet pour la publicité, il est possible de faire passer le bon message à la bonne personne en tenant compte de différentes informations très personnelles : antécédents, consommation de médicaments, informations génétiques ou comportement.

On peut citer le changement qui est en train de se produire dans le programme de dépistage du cancer du sein. Initialement, c'est un programme vertical dédié à une population identifiée : les femmes de 50 à 74 ans à qui on propose un test de dépistage tous les deux ans. Ce programme va être optimisé grâce à la disponibilité des données génétiques des femmes ayant des prédispositions héréditaires de cancer du sein.

Le dispositif va s'adapter pour que le dépistage systématique et de masse devienne un dépistage de masse « individualisé » tenant compte des facteurs de risque de chacune avec des plans personnalisés de suivi adaptés. L'INCa (Institut national du cancer) travaille actuellement sur la mise en place de tels programmes.

Plus les données seront disponibles, plus le ciblage sera fin et plus la « proposition de prévention » sera précise et adaptée.

Ces possibilités offertes par le Big Data sont la promesse d'une amélioration importante des programmes de santé publique.

Cette promesse d'une meilleure efficacité et d'économies substantielles a été très bien comprise et intégrée par des organismes comme les fournisseurs d'objets connectés dans le domaine du bien-être, les membres du GAFAM, les mutuelles et plus récemment, par l'assurance maladie (programme « santé active »).

Un coaching individuel peut être fait par des systèmes d'intelligence artificielle en fonction de son comportement. Dans le cadre des mutuelles, un « bon comportement » peut faire l'objet d'une rétribution sous la forme d'un cadeau voire de la modification du montant de la prime dans certains pays comme les États-Unis.

De la mutualisation des risques, principe fondateur des programmes de santé publique et des systèmes d'assurance maladie au sens large, la tentation d'utiliser des modèles qui pointeront les « mauvais comportements » d'individus, ou proposeront une exclusion du système des personnes à risque va devenir de plus en plus grande.

On imagine sans peine que l'on peut se diriger vers des modèles élitistes et *behaviouristes* promus par ceux qui gèrent les données à des fins de rentabilité pour la société et pour ces derniers.

Plus que jamais se pose la question du rôle des différents acteurs : gouvernements, institutions en charge de la santé des populations, experts, sociétés et organismes qui possèdent les données.

Pour ces programmes de prévention, le mandat et les responsabilités de chacun doivent être précisés clairement et maintenus pour éviter des dérives, qui au final iraient à l'encontre de leur raison d'être.

5. CONCLUSION : ENJEUX MÉTHODOLOGIQUES ET ÉTHIQUES

Le croisement d'un nombre croissant de données sanitaires et extra-sanitaires conduit à générer des hypothèses en rapport avec la santé et ses déterminants environnementaux et socio-économiques et permet potentiellement l'accès à de nouvelles connaissances. Ces

données sont issues d'une multiplicité de bases structurées ou non structurées, avec des méthodes de mesure variables selon les bases et selon les années, des données manquantes, des sources, des origines disciplinaires et des modes de recueil extrêmement variés. La question de la méthode dans ce contexte et celle du contrôle de l'information deviennent tout à fait centrales pour notre système de santé publique[7].

De même, la disponibilité des données et les possibilités que cela emporte en termes d'évaluation de l'efficacité des actions de santé publique devraient conduire à une nouvelle responsabilité des acteurs. Le suivi des données en temps réel peut permettre d'objectiver et de mesurer l'efficacité des actions prises et le cas échéant de les ajuster. *A contrario*, il devient impossible d'ignorer l'absence de résultats. Ainsi en va-t-il par exemple de l'échec des campagnes de vaccination contre la grippe saisonnière reconduites à l'identique d'année en année alors même que les résultats en termes de couverture des populations restent très en deçà des recommandations de l'OMS, avec les conséquences induites en termes de surmortalité liée à l'épidémie chaque année.

Les données, dès lors qu'elles sont disponibles et susceptibles d'éclairer sur l'efficience de l'action, de détecter des signaux, de suivre les épidémies et les crises sanitaires, ne peuvent plus rester inertes. Leur accès doit être ouvert à une pluralité d'acteurs à des fins de santé publique, en favorisant la porosité entre les acteurs en charge de la politique de santé et le monde des Data-sciences, en développant des modes de régulation et de coopération modernes entre la sphère publique et les acteurs du monde économique, porteurs d'innovation, en soutenant l'ouverture et l'usage des données de santé, dans des conditions de contrôle garantes du respect de la vie privée, au service de la prévention des maladies et de la sécurité sanitaire.

7. C. Dimeglio, C. Delpierre, N. Savy, Y. Lang, *Big Data et santé publique : plus que jamais, les enjeux de la connaissance*, ADSP N°93, décembre 2015.

ARTICLE III
L'ÉVOLUTION DU PAYSAGE TECHNOLOGIQUE COMME UN FACTEUR CLÉ DE SUCCÈS POUR « FAIRE PARLER LES DONNÉES »

1. INTRODUCTION

Pourquoi l'architecture des *SI* apporte de nouvelles perspectives aux « Mégadonnées[8] » ?

Les technologies spécifiques au Big Data n'apportent pas en soi de nouveauté dans les concepts de traitement des données en eux-mêmes, mais plutôt dans la façon dont ils sont intégrés pour d'une part répondre aux nécessités spécifiques de « Volume, Vélocité, Variété » et d'autre part offrir de nouvelles opportunités en matière de « Variabilité, Véracité, Vitesse, Visualisation[9] ».

En effet, face aux Big Data, les technologies traditionnelles sont limitées par leur rigidité en termes de format des données (« Variété »), comme par la non-scalabilité de leurs outils (« Volume »), à la multiplicité et l'évolution des sources de données disponibles et au nombre croissant de techniques d'analyses nécessaires (« Vélocité »).

Le Big Data est finalement plus une évolution qu'une révolution. Le Big Data d'aujourd'hui sera la « Small Data » de demain. Il n'en reste pas moins que le cœur demeure « la donnée » .

Si une définition devait en être faite, le paysage technologique Big Data désigne l'ensemble des **technologies**, **architectures, infrastructures et procédures** permettant de très rapidement **capter**, **traiter** et **analyser**

8. Terme français dont l'usage est recommandé par le Journal Officiel du 22 août 2014.

9. Ces exigences sont souvent désignées comme les « sept V » ou les « 7 V ».

de larges quantités et contenus hétérogènes et changeants pour en extraire les informations pertinentes, et ce de façon industrielle.

Centré sur les « mégadonnées », ce paysage technologique doit être évolutif pour s'ajuster de façon agile et rapide :

• à la mise à disposition de nouvelles sources de données : compatibilité des systèmes internes et externes (cloud), migration, Internet of Things (IoT)… ;

• aux formats divers des données, des modèles, des ontologies et des sémantiques : données structurées, non structurées, modèles réglementaires ;

• aux besoins d'adaptation des flux de données : ajustement avec les processus opérationnels ou décisionnels… ;

• aux traitements hétérogènes des données : capacité de stockage, aire de stockage, agrégation, requête, traçabilité… ;

• aux besoins variés d'analyse : profondeur d'analyses sur plusieurs années, mise en résonance de différents domaines de données pour répondre à de nouvelles questions… ;

• aux changements de contextes opérationnels et exigences réglementaires : adaptation des modèles aux variations du marché, modèles prédictifs… ;

• aux besoins de visualisation et de communication personnalisées : données adaptées au contexte métier et au profil d'utilisateur (patient, autorité réglementaire, personnel de santé, etc.) ;

• etc.

Ce paysage est constitué de briques technologiques complémentaires qui répondent chacune à un besoin technique ou fonctionnel. Par leur intégration, elles constituent un paysage technologique qui atteint des niveaux de capacité et de performance de traitement nécessaires aux « mégadonnées ». Ainsi, les 7 V[10] définissent l'horizon de ce paysage technologique.

10. Les 7 V sont « Volume, Vélocité, Variété, Variabilité, Véracité, Vitesse, Visualisation ».

Comment la conception des *SI* doit-elle prendre en compte les fonctions de production de la connaissance ?

Pour définir un tel paysage, les principaux sujets à considérer sont alors :

• la disponibilité et granularité des données indispensables au champ d'application recherché ;

• les formats de données à intégrer / visualiser ;

• les volumes adéquats à centraliser, les niveaux de confiance acceptables des données à traiter, les exigences de fraîcheur de ces données, la profondeur des données (historique) ;

• les types de traitement à implémenter, les besoins de mettre au point les algorithmes par apprentissage ;

• les vitesses d'exécution appropriées des traitements ;

• les débits des flux d'information à prendre en compte ;

• la capacité de consommer cette donnée en self-service ;

• le besoin d'espace de prototypage, d'évaluation de nouveaux scénarios en mode agile de type « bac à sable » ou « Data lab » ;

• la variété des différents outils de Data Mining, de restitution et de visualisation ;

• les niveaux de services attendus par les différents consommateurs de ces données ;

• la stratégie de gestion adoptée pour le traitement de ces données (dupliquées, isolées ou centralisées) ;

• les bonnes pratiques mises à disposition pour répondre à ces exigences (modèle de données métiers…).

Et sur chacun de ces sujets, il faut se poser les questions suivantes :

• Pour qui ?

• Pour quel objectif « métier » ?

A. LA VARIÉTÉ

Ce que la technologie apporte…

La richesse des données non structurées est enfin exploitable.

L'un des enjeux du Big Data est d'analyser les données non structurées (ex : vidéo, enregistrements vocaux, textes non numérisés…), car leur production est plus abondante et leur exploitation à plus haute valeur ajoutée.

Ici, ce n'est pas le « **Big** » qui est important mais plutôt le « **Smart** » . Par exemple, pour comprendre un document rédigé en langage naturel, le choix s'orientera sur les technologies et standards du « **Web sémantique** » pour effectuer des analyses sémantiques.

… et ce que cela peut signifier pour la prévention

Les scientifiques peuvent désormais travailler sur des données nouvelles, quel que soit leur format d'origine.

Ainsi, Mayo Clinic a mis en œuvre les outils d'analyses textuelles pour disposer d'une meilleure perception des notes en texte libre issues des millions de dossiers médicaux disponibles dans leurs systèmes. Six types de messages HL7 sont maintenant collectés en quasi-temps réel, indexés, transformés, analysés et mis à disposition de leurs personnels.

B. LE VOLUME

Ce que la technologie apporte…

Le volume permet l'exhaustivité

En 2019, le potentiel de stockage de tous les ordinateurs du monde devrait être de 20 exaoctets soit 10^{18} octets.

Sébastien Verger, directeur technique d'EMC prédisait en 2013 que « *les volumes de données vont augmenter d'un facteur trente d'ici à 2020 pour atteindre 35 zetaoctets (soit 10^{21} octets) au niveau mondial.* »

Les scientifiques, habitués à la rareté des informations et aux échantillonnages réduits, changent d'échelle : de « rares », les informations passent à « ordinaires » ; de « x % », les échantillons passent à « tout », cependant parfois au détriment de la qualité de l'information.

… et ce que cela peut signifier pour la prévention

L'exhaustivité permet d'être moins exigeant sur l'exactitude des informations. Les données sont souvent en désordre, de qualités variables et issues d'innombrables sources. Les informations issues de données non structurées ont souvent une faible densité, mais la valeur réside dans leur quantité.

Travailler sur des données complètes, même imparfaites, permet de prendre du recul vis-à-vis du principe de causalité, souvent source d'erreur et de mauvaise interprétation.

La biologie est un domaine où l'hypothèse est à l'origine de tout raisonnement. Le protocole est en effet toujours le même : on émet une hypothèse de départ puis on vérifie son exactitude ou son inexactitude.

À titre d'exemple, plutôt que de dégager une hypothèse à partir d'une observation d'une protéine, l'exploitation de « Data » permet d'identifier une tendance probable à partir d'une masse de données aux sujets de nombreuses molécules. C'est ce sur quoi travaille l'américain Ernest Fraenkel, chercheur en biologie au MIT, qui essaie de construire un modèle unique qui incorpore tous les types de données : « *Ma plus grande innovation a été de proposer une interprétation holistique de la donnée.* ». Ernest Fraenkel vient ainsi bouleverser les codes de la biologie avec cette nouvelle approche.

C. LA VÉLOCITÉ

Ce que la technologie apporte…

Associé au Big Data, le Data Mining devient le « Big Data Mining »…

Les outils d'analyse des données (statistiques et Data Mining) et d'analyse de texte (text-mining) sont essentiels pour exploiter les gisements d'information.

Les outils de Data Mining actuels permettent d'analyser de nombreuses données, mais sur un échantillon considéré comme représentatif. Ces outils sont limités par leur temps de traitement et/ou par leur capacité de traitement.

Par l'association des outils du Data Mining à ceux du Big Data (capacité de stockage immense, rapidité d'exécution des traitements), le Big Data Mining permet de traiter l'intégralité des données avec des temps de traitement devenus acceptables et superposables, et d'évoluer du descriptif vers le prédictif et le prescriptif.

La capacité à combiner toutes ces données pour plus de transversalité et d'agilité…

Avec les technologies du Big Data, la donnée peut ainsi être factorisée, mise en résonance dans un éco-système analytique offrant une transversalité permettant de répondre aux enjeux du métier avec plus d'agilité. La même donnée est ensuite consommée, sous des prismes métiers distincts au travers des outils de Data Mining et de business intelligence par des profils métiers différents.

… et ce que cela peut signifier pour la prévention

Les scientifiques ont besoin de générer des traitements statistiques très rapidement et d'en conserver l'historique.

Les effets indésirables liés aux médicaments entraîneraient en France 10 000 décès par an. Afin de constituer des règles d'alertes basées sur les erreurs passées, le projet européen PSIP (Patient Safety Through Intelligent Procedures in Medication) propose entre autres de générer ces règles d'après la fouille automatisée des données (Data Mining).

Considérons maintenant cela sous l'angle du paysage technologique. Les fournisseurs d'IRM ont la capacité d'assurer une qualité de services et un taux de disponibilité de leurs équipements sensibles en forte hausse au bénéfice direct des praticiens et des patients. En effet, GE Healthcare s'appuie sur des fonctions d'analyses avancées et divers modèles prédictifs appliqués aux données transmises par leurs équipements en temps réels pour faire de la maintenance prédictive. L'intervention a donc lieu en amont, avant qu'une interruption de service n'apparaisse. (source : TeraData)

D. LE RATIO « VOLUME / VÉLOCITÉ / VITESSE »

Ce que la technologie apporte…

Une augmentation considérable des volumes traités sans diminution des performances

Les données échangées sur Internet, provenant parfois des outils connectés, sont stockées sur des serveurs et des disques durs. Elles peuvent être collectées directement. La scalabilité est la capacité d›un système à maintenir ses fonctionnalités et ses performances en cas de forte montée en charge. Mais les modèles de scalabilité ne sont pas linéaires et le seul ajout de capacité de stockage ne permet pas toujours d'améliorer les performances.

Les solutions traditionnelles du marché (IBM, EMC, etc.) répondent aux 3 V du Big Data mais chacune suit son propre modèle d'implémentation du stockage distribué : *Cluster File System, Parallel File System*. Cependant ces solutions n'ont pas les mêmes performances ni les mêmes capacités d'évolution face à une montée en charge

(quand la capacité de stockage des disques a augmenté de 100 000, le débit n'a augmenté que de 100).

Ainsi, l'architecture décisionnelle « traditionnelle » avec sa base de données n'est plus l'unique architecture de référence. C'est en réfléchissant au-delà du carcan des architectures traditionnelles que des acteurs (les grands du Web notamment) ont trouvé des solutions. Il existe à présent 3 architectures de référence complémentaires à maîtriser : base de données, In Memory et massivement parallèle.

… et ce que cela peut signifier pour la prévention

Avoir un site Web toujours disponible, collecter très rapidement des masses colossales d'information, faire face à des pics d'affluences sur ce site, assurer la sécurité et la sauvegarde de sa plate-forme, c'est une exigence de tout site Web d'un acteur de la santé qui souhaite diffuser ou collecter des informations aux personnels médicaux et / ou aux patients, tout en sécurisant le flux d'information.

Johnson & Johnson, l'un des principaux fabricants d'implants de hanches et genoux va s'appuyer sur Watson pour créer un service de conciergerie pour les patients afin de les préparer aux mieux à l'intervention et les aider dans les suites post-opératoires.

La dernière génération de pacemakers est connectée : ils envoient en temps réel à l'hôpital ou au médecin des données qui permettent de réagir rapidement en cas d'anomalie ou tout au moins de contrôler à distance l'état de santé du malade.

E. LA VARIABILITÉ ET LA VÉRACITÉ

Ce que la technologie apporte…

Les algorithmes pour le « quoi »

Un principe du Big Data est de découvrir des corrélations, des tendances, mais sans en expliquer l'origine. Une corrélation ne fait que

quantifier la relation statistique entre deux valeurs. Elle sera forte, si une valeur a de grandes chances de changer quand l'autre est modifiée. Les corrélations n'apportent aucune certitude, que des probabilités. Elles ne disent pas pourquoi quelque chose se produit, mais simplement qu'elle se produit.

Ici, le Big Data s'attaque au « quoi » et non au « pourquoi ». Mais avec une grande précision. En conséquence, il n'est plus besoin d'émettre des hypothèses à vérifier : il suffit de laisser parler les données et d'observer leurs connexions dont on n'aurait parfois même pas soupçonné l'existence.

Mais, selon Frank Pasquale, de l'université de Yale (*The Black Box Society*), la neutralité serait un mythe et la manipulation des algorithmes une réalité.

Partant de ce constat, les technologies du Big Data ont pour objectif de faire évoluer cette neutralité puisqu'elles s'appliquent à un champ d'information beaucoup plus large et donc de prendre du recul.

… et ce que cela peut signifier pour la prévention

En décembre dernier, Google a annoncé la création de Verily, nouvelle filiale regroupant ses projets dans le domaine médical et basée sur l'ancienne entité Google Life Science. Dotée d'équipes hardware, software, clinique et scientifique, Verily travaille sur des plates-formes, des produits et des algorithmes destinés à dégager les causes profondes des maladies et à analyser les traitements les plus adaptés afin de mieux les diagnostiquer et les soigner.

Attention toutefois, il faut prendre garde à toujours qualifier, contextualiser et relativiser correctement les données, car le niveau de responsabilité est élevé en matière de santé.

En 2009, en pleine pandémie de grippe H1N1, le ministère de la Santé américain a demandé l'aide de Google. En localisant sur une carte la position et la provenance des mots-clés tapés dans le célèbre moteur de recherche, les ingénieurs ont pratiquement réussi à dessiner l'évolution de l'épidémie… à quelques écarts prêts : le mot

« grippe » dans le moteur de recherche traduisait une simple inquiétude à l'approche de l'hiver des internautes.

F. LA VISUALISATION

Ce que la technologie apporte…

La visualisation pour les interactions

Les technologies du Big Data permettent de consommer l'information en toute autonomie, de disposer d'un maximum d'interactivité, d'avoir une instantanéité des réponses.

La visualisation peut permettre de détecter des tendances invisibles de prime abord, et de mettre en valeur les données les plus probantes.

… et ce que cela peut signifier pour la prévention

La visualisation n'est pas qu'un mode de représentation, mais également un moyen d'encourager les parties prenantes à s'approprier les données, à les étudier, à les partager, et à mieux se comprendre. Outil d'aide à la compréhension et à la décision, il peut devenir un moyen de communication.

Ce problème complexe de visualisation de données est au cœur d'un programme d'exploration sur la culture de la santé porté par une fondation américaine ; par exemple, quelle visualisation est la plus adaptée pour communiquer sur l'importance de prévenir les maladies cardiovasculaires ? Des infographistes n'ayant aucune connaissance en prévention ont ainsi planché sur une liste de seize scénarios de communication sur des risques médicaux. En bout de chaîne, citoyens et patients ont jugé les différentes propositions.

G. LES DIFFICULTÉS ACTUELLES

Dans un contexte technologique encore peu mature :

• difficulté d'expression des besoins du demandeur vers le fournisseur ;

• difficulté de compréhension du fournisseur du plan de route à moyen terme ou long terme du demandeur ;

• difficulté du demandeur à identifier le 1er projet structurant et les étapes successives conduisant au plan de route à moyen ou long terme ;

• sous-estimation des efforts d'alignement et de nettoyage des données ;

• multitude de technologies et de combinaisons technologiques ;

• expertise naissante des intégrateurs ;

• manque de recul sur les capacités industrielles des plates-formes open-source à servir les exigences opérationnelles et besoins métiers ;

• sous-estimation des besoins de gouvernance des données ;

• difficulté à obtenir un sponsor à la direction générale en mesure de porter ce type de projet transverse.

H. CONCLUSION

• Les possibilités fonctionnelles qu'apporte le Big Data sont immenses.

• La réglementation reste à ce jour ouverte aux propositions d'innovation technologique.

• Les principaux challenges restent aujourd'hui organisationnels ou éthiques.

I. UNE INFRASTRUCTURE FÉDÉRATRICE… ESSENTIELLE POUR LE BIG DATA

Ce que la technologie apporte…

Gérer un environnement aussi riche et complexe nécessite une nouvelle approche du paysage technologique.

Parlons alors d'un « écosystème » capable de traiter efficacement les données dans toutes ses expressions et dimensions, de supporter l'usage d'outils de visualisation et d'analyses innovants, d'optimiser les processus d'intégration de données et de s'adapter immédiatement et sans compromis aux différentes contraintes (techniques, opérationnelles et réglementaires).

Le traitement parallèle et au plus près de la donnée (In-Database) revêt un caractère indispensable et une caractéristique incontournable de l'infrastructure Big Data. Les traitements s'exécutent donc en parallèle, garantie unique de temps de réponses et d'engagements de niveaux de services indispensables pour servir des utilisateurs dont l'exigence augmente avec l'expansion des données à disposition.

La flexibilité et les capacités industrielles intrinsèques d'une telle plateforme autorisent donc l'usage combiné de plusieurs technologies de traitement des données. Cet ensemble modulaire, agile et cohérent est parfaitement décrit par les analystes tels que Gartner sous la terminologie de « Logical Data Warehouse Architecture Model ».

Ce modèle d'architecture permet aux organisations d'accéder à la donnée, de l'affiner et de la restituer de manière fiable tout en garantissant simultanément la disponibilité du système, la performance, la concurrence et la variété des traitements.

Sélectionner l'outil approprié pour répondre au problème posé est un principe clé des bonnes pratiques d'ingénierie. Sur ce principe fondamental, le concept d'un écosystème Big Data s'attache à disposer de la bonne technologie pour exécuter le bon traitement. L'adage « *aux bons artisans les bons outils* » prend donc tout son sens. Il n'existe pas actuellement une technologie capable de résoudre tous les problèmes.

Cette architecture agile distingue 5 composants :

1) le Lac de Données (Data Lake) ;

2) la solution Exploratoire (Data Discovery) ;

3) un Entrepôt de Données Integrées (Integrated Data Warehouse) ;

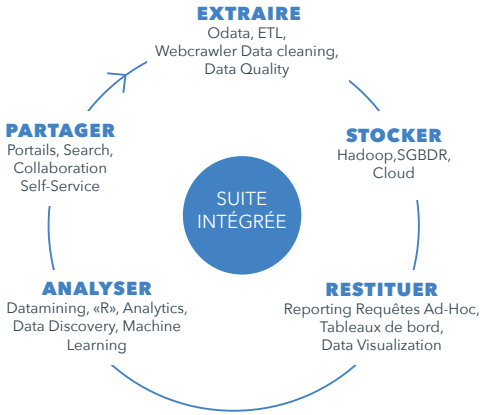
4) la couche d'Intégration des Données (Data Integration Layer) ;

5) un moteur éprouvé d'allocation des ressources (Integrated Workload Management).

Le tout étant totalement intégré et interopérable.

La clé est de comprendre quelles parties d'une charge de travail analytique devraient être allouées à chacune des technologies au sein de l'architecture et d'orchestrer l'interaction de ces différentes briques. En effet, la charge de travail analytique peut impliquer plusieurs sources et types de données nécessitant une dispersion des traitements sur des technologies différentes de l'écosystème.

Par exemple, l'ensemble des briques de cette architecture sera sollicité pour la corrélation des données de comportements issues de l'application mobile de suivi des patients diabétiques de Type 2 (activités, prise de constantes, niveau de glucose, évaluation du bien-être, repas et calories…) avec les données du site Web communautaire, les données de prescription d'une population de patients et le dossier médical individuel. La brique Data Lake sera plutôt sollicitée pour intégrer les données issues du smartphone ou du site Web communautaire, la brique Data Discovery s'appuyant sur la couche d›intégration pour accéder de façon transparente aux données du Data Lake et de l'Integrated Data Warehouse et appliquer les différents algorithmes d'analyses de comportement et enfin la brique Integrated Data Warehouse pour les données plus structurées de prescription, du dossier patient ou encore récupérer les résultats de la phase de Discovery en vue d'une opérationnalisation de ces résultats vis-à-vis des patients par exemple.



Les différentes étapes du traitement des Big Data. Source : le CXP 2015

Data Lake

Le composant Data Lake abrite la matière première, données brutes non raffinées, avec une faible densité d'information, ingérées sans schéma strict d'organisation pour être ensuite transformées et corrélées afin d'apporter de nouvelles perspectives et de la valeur métier.

Capturer toutes les données continuellement est un but implicite à atteindre car toute donnée ignorée et par conséquent non exploitée ne révélera jamais son potentiel. La mise en œuvre d'un Data Lake réussi doit répondre à ces caractéristiques clés :

- la montée en charge (scalabilité) ;
- le faible coût de la donnée stockée par terabyte ;
- le support des types de données structurées, semi-structurées et non-structurées ;
- l'ouverture aux solutions et outils externes.

Data Discovery

Le cas d'usage type pour la fonction de Data Discovery relève des activités de R&D effectuées par les data-scientistes et analystes métiers dans le cadre d'une démarche analytique d'exploration et de modélisation statistique. Elle se distingue par sa grande agilité à corréler des données issues des briques Data Lake, Integrated Data Warehouse ou de sources externes, d'itérer rapidement afin de trouver et de révéler rapidement des faits cachés et des corrélations avérées tout en offrant aux utilisateurs métiers la possibilité de consommer ces résultats sous la forme d'applications réutilisables.

La mise en œuvre d'une solution de Data Discovery réussie doit répondre à ces caractéristiques clés :

- Le large éventail de techniques analytiques disponibles (analyses statistiques, textuelles, analyses de graphes, analyses de parcours, analyses de modèles, etc.) et des techniques d'exécutions de scripts (R, SAS, SQL, MapReduce, etc.).
- L'efficacité du processus d'acquisition, préparation, analyse et visualisation des données.

• Le support des types de données structurées, semi-structurées et non-structurées.

• L'ouverture aux solutions et outils externes.

Integrated Data Warehouse

La brique Integrated Data Warehouse intègre et déploie la donnée en tant que produit fini auprès des utilisateurs métiers dans une optique opérationnelle de prise de décisions. Son rôle est de fédérer, mutualiser une seule version de la donnée, servant de multiples usages (gestion multi-domaines) et des exigences utilisateurs variées. La transversalité recherchée sera d'autant plus efficace que le système d'information sera réellement intégré dans sa globalité.

Les exigences liées à la gestion multi domaines imposent à la solution de pouvoir assurer une performance constante matérialisée par des engagements contractuels de niveau de services vis-à-vis des utilisateurs finaux.

La solution doit également permettre de mixer des données de production avec des données non certifiées au sein de Data Labs ou Sandbox analytiques. En d'autres termes, des laboratoires de données en libre-service permettant aux utilisateurs de se créer un espace de prototypage, d'évaluation dans lequel ils pourront tester, vérifier et valider des hypothèses en combinant des données de production certifiées (en lecture seule) avec des nouvelles sources d›information (fichier plat, données du monde réel, données issues du Data Lake, ou du Data Discovery).

La mise en œuvre d'une brique Integrated Data Warehouse réussie doit répondre à ces caractéristiques clés :

• une solution hautement véloce bâtie sur une technologie de traitement massivement parallèle ;

• la robustesse et la haute disponibilité de la solution ;

• la montée en charge linéaire au fur et à mesure de l'évolution des besoins et contraintes ;

• des fonctions analytiques exécutées au plus près des données (in-Database) ;

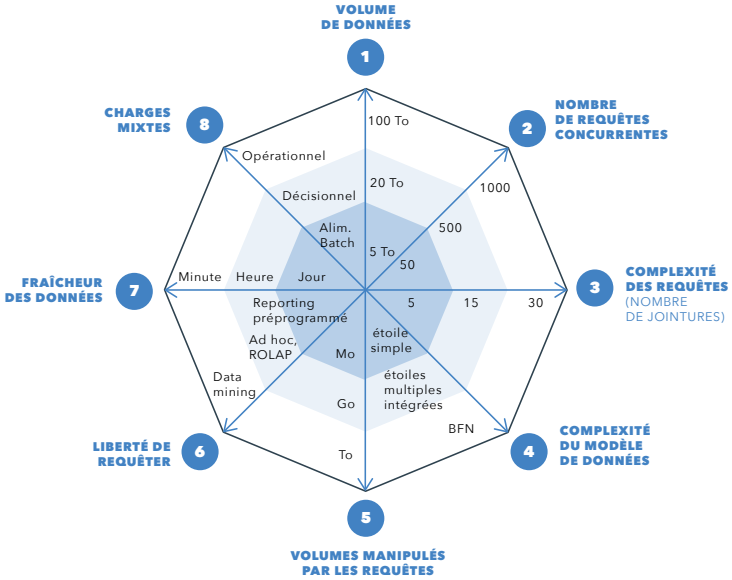
• le support des types de données structurées, semi-structurées et non-structurées ;

• des fonctions d'historisation, de traçabilité, d'auditabilité et de sécurité avancées ;

• la flexibilité des espaces en libre service (Data Lab ou Sandbox) ;

• l'ouverture aux solutions et outils externes ;

• la capacité de la solution à absorber simultanément des contraintes opérationnelles (volume, nombre de requêtes simultanées, performance).



Évolution simultanées sur plusieurs dimensions

Intégration Data Layer

Pour être productifs, les utilisateurs ont toutefois besoin de technologies qui les affranchissent de la nécessité de transférer les données de système à système, ou même de connaître les mécanismes de déplacement des données.

Cette brique revêt une importance particulière à l'heure où les technologies open-source deviennent une composante majeure des environnements Big Data.

Force est de constater l'association de différentes technologies provenant de différents fournisseurs au sein des systèmes analytiques et opérationnels. Ces technologies, non intégrées, sont souvent considérées par les utilisateurs en tant que systèmes distincts. Cet ensemble arrive donc difficilement à s'échanger des données et à opérer les traitements inter-composants.

Le composant Integration Data Layer assure la couche d'interopérabilité et donc une interaction bi-directionnelle transparente aux données et aux traitements analytiques entre les différents composants de l'écosystème Big Data. Il se doit de gérer l'exécution des requêtes qui utilisent plusieurs moteurs analytiques et entrepôts de données (Data Lake, Data Warehouse, etc.), de manière optimisée et sécurisée et d'intégrer ainsi de multiples systèmes pour n'en constituer qu'un seul.

La mise en œuvre d'une brique Integrated Data Layer réussie doit répondre à ces caractéristiques clés :

- traitements bi-directionnels ;
- fonction d'optimisation lors de l'exécution de requêtes sur de multiples systèmes et environnements technologiques hétérogènes ;
- intégration avec des environnements open-source ;
- possibilité d'interroger l'écosystème à partir de n'importe quel composant (Data Lake vers Integrated Data Warehouse, Data Discovery vers Data Lake, Integrated Data Warehouse vers Data Lake, etc.) ;
- mouvements de données ou duplication de données limités.

Integrated Workload Management

Une gestion efficace de la charge de travail est cruciale pour fournir aux utilisateurs de l'entreprise des réponses à leurs questions en temps réel et tenir les engagements de performance et de niveau de services des infrastructures Big Data. La brique Integrated Workload Management est relative à l'allocation dynamique des ressources.

Son rôle est de simplifier et automatiser la gestion des données, d'optimiser l'allocation des ressources machines et de faire cohabiter des usages très différents sur la même plateforme. Cela permet à chaque tâche de pouvoir s'exécuter dans des conditions optimales, et de garantir les différents niveaux de services au regard des variations opérationnelles.

Grâce à une « hiérarchisation déterministe », le moteur assure que les tâches les plus importantes sont réalisées en priorité. Ses mécanismes permettent en outre de stabiliser les temps de réponse et d'optimiser l'utilisation des ressources lors des pics, qu'ils soient ou non planifiés.

La mise en œuvre d'une brique Integrated Workload Management réussie doit répondre à ces caractéristiques clés :

• vision holistique de l'écosystème Big Data (Data Lake, Data Discovery et Integrated Data Warehouse) ;

• changement dynamique des priorités sans arrêt des requêtes ou du système ;

• maturité du moteur d'allocation dynamique des ressources ;

• assignation hiérarchisée des tâches (priorités) ;

• fonctions de filtres, de gestion des exceptions, de planification ;

• définition des critères d'assignation (règles) ;

• possibilité d'affectation de niveaux de services ;

• intégration dans un outil de surveillance intégré.

… et ce que cela peut signifier pour la prévention

Le projet GLIDE (Global Integrated Drug Development Environment) mis en œuvre en 2015 au sein du groupe pharmaceutique Roche combine une multitude de données internes et externes au sein d'un même écosystème (études cliniques / données de nouveaux traitements – généralement des Datasets SAS, des données de laboratoires – sang, radiographies, électro-encéphalogrammes (EEG) ; données médicales, données génétiques, bio-marqueurs ou encore données du monde réel).

Cet écosystème analytique a pour effets de réduire significativement les temps de traitements (de plusieurs jours à quelques heures ou minutes), mais aussi de permettre aux équipes R&D de développer des thérapies mieux ciblées, de diminuer les effets indésirables des nouvelles molécules ou encore, à partir de tout cet historique d'observations de repositionner un médicament à de nouvelles fins thérapeutiques.

J. GLOSSAIRE

Données non structurées : textes issus de traitements de textes ou de contenus de messages électroniques, fichiers audio de sons, de voix, images fixes ou vidéos.

Données structurées : feuilles de tableurs organisées en tableau, bases de données où les données sont organisées en tables reliées entre elles.

Big Data : littéralement « grosses données ». Expression anglophone désignant les ensembles de données tellement volumineux qu'ils deviennent difficiles à travailler avec des outils classiques de gestion de base de données. Au sein du HDI, le terme a un sens plus élargi et désigne toutes les données quels que soient leurs sources, leurs formats et leurs volumes.

Smart Data : se focalise sur les données pertinentes par rapport à ses propres objectifs.

Data Mining : détection d'information dans une base de données.

Outils capables de détecter l'information cachée « au plus profond » de la « mine de données ». Cela ne concerne pas les systèmes d'interrogation de base de données, ni les tableurs, ni les systèmes statistiques, ni même les systèmes d'analyse de données traditionnels.

Le Data Mining suit plusieurs approches :

• **Approche « vérification » :** l'utilisateur a l'intuition ou l'idée générale du type d'information qu'il peut obtenir de ses données. Il tire alors profit de sa base de données en « quantifiant » son intuition. Il est clair que les données extraites, et les décisions qui en découlent,

dépendent exclusivement de l'intuition de l'utilisateur concernant les paramètres importants du problème (âge, géographie, etc.), intuition qui est souvent correcte mais non exhaustive.

• **Approche « découverte » (Advanced Data Mining) ou recherche de l'information cachée :** l'utilisateur comprend que la quantité de données dont il dispose étant considérable, la détection optimale et exhaustive des structures ou relations importantes est totalement hors de portée de l'utilisateur humain. Il doit alors s'appuyer sur des méthodes avancées d'analyse de données pour détecter l'information cachée (dont il se peut qu'elle soit la plus intéressante).

Text-mining : ensemble de méthodes, de techniques et d'outils pour exploiter les documents non structurés. Le text-mining s'appuie sur des techniques d'analyse linguistique.

Architecture : désigne la structure générale inhérente à un système informatique, l'organisation des différents éléments du système (logiciels et/ou matériels et/ou humains et/ou informations) et des relations entre les éléments.

Voici différentes architectures :

• **Architecture décisionnelle « traditionnelle » (Oracle, SQL Server, MySQL, Informatica, Datastage…)** : cette architecture est performante lorsque le volume de données à transférer entre chaque étape reste limité.

• **Architecture In Memory (Qlikview, ActivePivot, HANA…) :** cette architecture permet d'offrir des services d'analyses performants, voire en temps réel (mise à jour des données et re-calcul au fil de l'eau des agrégats) ainsi que des services de simulation.

• **Architecture massivement parallèle (Hadoop, TeraData) :** cette architecture permet de stocker une quantité immense de données (sans limites) et de manière élastique.

Scalabilité : Capacité d'un système à fonctionner correctement avec des charges de travail plus importantes.

2. LE « KNOWLEDGE BY DESIGN » OU LA SÉMANTIQUE COMME CLÉ DE LA VALEUR.

Les systèmes d'information du secteur de la santé devraient toujours pouvoir contribuer à la santé publique, à soutenir notre système de santé et à produire de nouvelles connaissances. Cela n'est possible qu'en définissant un certain nombre de principes et de règles constitutifs d'un cadre cohérent et stable, multidimensionnel, technologique, juridique, économique, organisationnel, institutionnel, politique. Dans cette démarche d' « urbanisation », la production des données, leur nature et les conditions de leur réutilisabilité constituent un investissement majeur pour l›avenir. La création de valeur attendue par la numérisation du secteur et la production de données massive passent donc par une ambition forte en termes d'interopérabilité.

La sémantique, source d'une création de valeur infinie : le « knowledge by design »

Le déploiement des infrastructures de l'Internet et l'organisation des différents protocoles de communication bénéficient à tous les secteurs d'activité. Les enjeux spécifiques au secteur de la santé ne sont pas là, mais bien dans les couches plus « métier » au premier rang desquelles l'interopérabilité sémantique, la principale source de valeur avec l'émergence de l'Internet des Objets. Et celle-ci constitue, pour la santé, un levier complexe à appréhender tant il est porteur de promesses considérables. Produire des informations dématérialisées permet certes des gains de temps et des économies directes. Mais faire en sorte que ces données soient porteuses de sens, interprétables, analysables, réutilisables ouvre des perspectives tout simplement infinies. Les référentiels sémantiques sont l'imprimerie du XXI^{e} siècle et vont permettre de produire et véhiculer l'essentiel des connaissances des prochains siècles. La production des données connaît une croissance considérable et cette production doit donc bénéficier au plus vite de référentiels tant il sera impossible de revenir en arrière sur ces données. L'enjeu de pouvoir automatiser tout ou partie du traitement de ces données provenant de sources multiples et produites dans des contextes et pour des finalités à

chaque fois différents devient un impératif d'une urgence extrême. Le déploiement d'une informatique qui continuerait comme c'est encore souvent le cas aujourd'hui à dématérialiser des données non structurées et non réutilisables doit être remplacé au plus vite par la diffusion de référentiels sémantiques en lieu et place pas exemple de documents JPEG inexploitables. Sur un plan plus macro-économique, la capacité à réutiliser les données revient à mutualiser les coûts de leur production et à multiplier la création de valeur. Le droit de réutiliser les données pour des finalités dites « compatibles » introduit par le nouveau règlement européen est une mesure aux effets considérables. Le modèle qui jusqu'ici imposait de produire une donnée de façon spécifique pour un usage dédié le plus souvent unique devra désormais être réservé à des études *ad hoc* et faisant appel à des notions insuffisamment répandues pour pouvoir utiliser un vocabulaire courant.

L'intégration sémantique des différents systèmes constitue un enjeu majeur qui doit être pris en compte en amont pour permettre une approche dite « translationnelle » établissant un pont permanent entre activités de soins d'une part et recherche de l'autre. Ce concept de *knowledge by design* par analogie au *privacy by design* promu dans le champ de la sécurité doit faire en sorte que l'essentiel des données produites soit porteur d'un sens commun au service notamment de la santé publique.

Les autres pays s'en préoccupent comme en atteste la création du consortium CDISC (Clinical Data Interchange Standards Consortium) par la Food and Drug Administration américaine qui vise à mutualiser les données des essais cliniques autour de référentiels sémantiques communs et libres d'accès. Ajoutons l'enjeu linguistique stratégique. Les 70 États constitutifs de la francophonie et les 220 millions de francophones dans le monde peuvent prétendre à ne pas voir se généraliser les seules codifications anglo-saxonnes.

Chaque donnée de santé peut potentiellement contribuer à la prévention dès lors que son sens est clair et que la réutilisation de la donnée est prise en compte dès la conception du système.

ARTICLE IV
LA DÉFINITION D'UN NOUVEAU PARADIGME POUR LES ACTEURS

1. BIG DATA : UN NOUVEAU PARADIGME POUR L'INDUSTRIE PHARMACEUTIQUE

L'industrie pharmaceutique comprend aujourd'hui que la quantité de données de santé disponibles, combinée avec des algorithmes et les capacités d'analyse actuelles, constitue un facteur majeur de l'émergence de nouvelles pratiques en épidémiologie, médecine personnalisée, prévention, de même que pour la recherche et le développement de nouveaux services de santé qui vont transformer la prise en charge des patients.

Un grand nombre de laboratoires pharmaceutiques développent des produits médicaux destinés aux patients souffrant de maladies chroniques, qui impliquent un suivi sur le long terme. Le médicament seul n'est plus vu comme une variable isolée du traitement. Il fait partie d'une prise en charge élargie qui a un impact décisif sur l'observance à long terme du traitement par le patient.

Le médicament est désormais juste un élément parmi d'autres des solutions de santé intégrées, qui font partie de la stratégie *beyond-the-pill* (« au-delà du cachet »).

Pour s'assurer de l'efficacité d'un médicament, l'industrie pharmaceutique a besoin de s'équiper d'outils pour mesurer les résultats – l'amélioration des résultats cliniques, la prévention des effets indésirables et des événements graves possibles chez les patients. Dans un modèle centré sur le patient, les statistiques jouent un rôle vital. Elles contribuent à la compréhension du comportement du patient et de ce fait permettent de connaître les facteurs qui déterminent l'observance du traitement par le patient[11].

11. Entretien avec Pascale Witz, *Healthcare Data Institute Newsletter*, juin 2015.

Les laboratoires pharmaceutiques vont dans le sens de développer et de fournir des services de santé innovants aux patients : produits médicinaux, mais aussi les solutions de santé qui les accompagnent. Beaucoup de ces solutions impliquent la collecte et l'analyse des données en temps réel avec des algorithmes adéquats, dans le but de prendre des décisions rapides de traitement.

La technologie va changer l'expérience du patient (en améliorant la prise en charge et la collaboration de celui-ci) et, en associant l'analyse de grandes quantités de données, va permettre d'initier un nouveau processus de prise en charge et de soin. L'amélioration des résultats cliniques des patients est l'objectif qui doit être atteint pour consolider un *business model* adéquat. En conséquence, la technologie couplée au suivi thérapeutique aura plus d'impact, par rapport à la technologie seule (cf. schéma 1).

Schéma 1 : La technologie des objets connectés est en train de changer l'expérience du patient, et la collecte et l'analyse des Big Data facilitent le processus. Dans le schéma, un exemple d'une approche de soins intégrée pour suivre et aider des patients atteints de problèmes respiratoires.

Plusieurs logiciels innovants de données de santé ont déjà dépassé le stade du rapport rétroactif (collecte de données) pour incorporer des fonctionnalités prédictives (analyse de données) qui permettent d'alerter les patients d'un effet indésirable potentiel.

En parallèle, de récentes avancées technologiques ont facilité la collecte et l'analyse d'informations venant de sources multiples – un avantage majeur dans le domaine de la santé, sachant que les données personnelles d'un patient peuvent venir de diverses technologies de mesure, dans beaucoup de cas, avec peu ou aucune participation active du patient. Par exemple, des capteurs portables, ingérables ou implantables, et des applications digitales portables peuvent collecter et transmettre des données de façon autonome ; l'analyse de données en presque temps réel permettra aux professionnels de la santé de devancer et d'éviter les crises imminentes des patients.

A. LE SUIVI DU PATIENT ET LA PRÉVENTION SECONDAIRE INDIVIDUELLE – MALADIES CHRONIQUES ET SANTÉ DIGITALE

L'intérêt de la prévention inclut aussi bien la prévention prospective des populations, que la prévention individuelle secondaire, cette dernière visant spécifiquement à éviter ou à anticiper les événements graves, les récidives et les crises chez les patients souffrant de maladies chroniques (telles que le diabète, l'asthme, l'arthrite rhumatoïde et les maladies cardiovasculaires).

La mise en œuvre de la prévention individuelle secondaire, nécessite généralement des solutions de santé intégrées qui comprennent des dispositifs médicaux, des capteurs portables, des applications pour smartphones, la collecte et l'analyse des données de santé en temps réel avec des algorithmes appropriés, et le retour d'information en temps réel, la notification des patients ainsi que des professionnels de santé, pour permettre des interventions en cas d'urgence (cf. schéma 2).

Les dispositifs intelligents (tels que le monitoring de la pression sanguine sans fil, les dispositifs pour la collecte passive et active de données, les capteurs bracelet pour le suivi de la qualité du sommeil) sont des composants courants de plates-formes avancées de surveillance et de participation des patients, qui ont pour but d'améliorer la prise en charge de cesderniers, et ainsi d'obtenir de meilleurs résultats cliniques et d'abaisser le coût total du traitement.

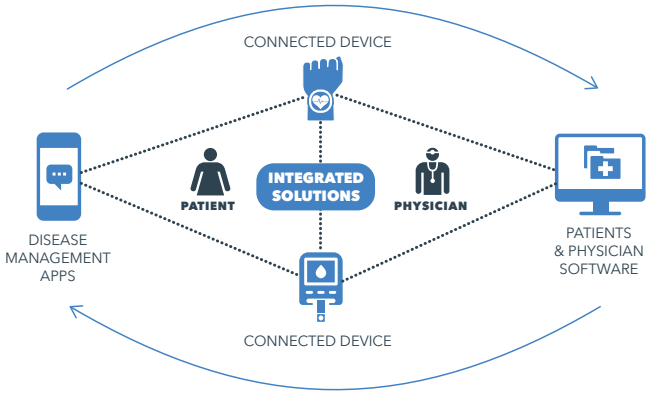


Schéma 2 : Les objets connectés, capteurs, les logiciels et l'analyse de données sont des composants nécessaires de solutions de santé intégrées.

Les médecins et les patients commencent à utiliser des outils digitaux qui permettent de gérer la santé de façon plus efficace. Dans le futur, les laboratoires pharmaceutiques et autres parties prenantes vont adopter les technologies existantes et en développer de nouvelles. Dans la foulée, ils vont aider à créer un *écosystème digital de la santé plus connecté* qui bénéficiera à tout le monde :

• **Les laboratoires pharmaceutiques** créent des applications et d'autres plates-formes pour mieux gérer les maladies et les traitements – fournissant des informations concernant les maux, médicaments, dosages adéquats, résultats cliniques attendus, objectifs de santé, et choix de vie sains.

• **Quand un médecin** prescrit un médicament, il ou elle prescrira aussi l'application digitale.

• **Les patients** peuvent ainsi collecter et suivre leurs propres données en temps réel. Ils peuvent les envoyer à leur médecin traitant pour une évaluation approfondie, et faire des ajustements par eux-mêmes sur le traitement médicamenteux ou l'hygiène de vie.

• **Les patients** peuvent aussi envoyer leurs données aux laboratoires pharmaceutiques, qui pourront alors utiliser ces informations pour développer des traitements, des services de santé améliorés et plus ciblés sur les patients, et ainsi conduire des recherches en adéquation.

B. EXEMPLES D'INITIATIVES DE SOLUTION DE SANTÉ INTÉGRÉE UTILISANT L'ANALYSE DE DONNÉES EN TEMPS RÉEL

Ginger.io offre une application mobile dans laquelle les patients atteints de maladies spécifiques souhaitent, en collaboration avec les médecins, être suivis via leurs téléphones portables et assistés avec des thérapies comportementales de santé. L'application enregistre les données des appels, SMS, localisation géographique, et même l'activité physique. Les patients répondent également à des questionnaires fournis sur leurs téléphones portables. L'application *Ginger.io* intègre les données des patients avec la recherche publique sur la santé comportementale du National Institutes of Health et d'autres sources. Les informations obtenues peuvent révéler, par exemple, un manque de mouvement ou d'autre activité qui signalerait que le patient ne se sent pas bien physiquement, et qu'il a des cycles de sommeil irréguliers (révélés par des SMS ou des appels tardifs) qui pourront signaler qu'une crise d'anxiété est imminente[12].

Diabète

Il n'y a aucune règle absolue concernant la dose d'insuline à injecter. Quand un patient commence un traitement pour le diabète, il ou elle doit procéder par étapes (titration). Le défi pour le patient diabétique est de maintenir le taux du sucre dans le sang. L'hypoglycémie (sous-dosage) ou son opposé, l'hyperglycémie (sur-dosage) peuvent être la source de problèmes graves.

Un plan de traitement solide va inclure une solution de monitoring continu du niveau de glucose, en se fiant à des dispositifs médicaux connectés et à des capacités analytiques de données afin de prévenir une crise hypoglycémique bien avant la présence de symptômes cliniques. La solution *Diabeo* de Sanofi est un exemple d'une technologie avancée de solution de télémédecine pour l'ajustement individualisé du dosage d'insuline, basé sur les mesures glycémiques, les régimes des patients et l'activité physique.

12. *Unlocking the full potential of data analytics for the benefit of all,* Healthcare Data Institue 2015.

Le monitoring mobile des signes vitaux des patients diabétiques avec des objets portables connectés – type bracelet – permet de partager les données en temps presque réel avec les professionnels de la santé du patient, et ainsi anticiper le risque de complications futures, y compris le pied diabétique et les maladies oculaires.

Un pancréas bionique automatique et portable, qui utilise un capteur relié à un smartphone et avec un logiciel algorithmique, a démontré plus de performance que les pompes à insuline traditionnelles, selon un article publié dans le *New England Journal of Medicine*, juillet 2014.

Livres blancs intéressants

- *Predictive Medicine Depends on Analytics*, Harvard Business School
- *The Big Data revolution in healthcare*, McKinsey & Company

2. LE SECTEUR DE L'ASSURANCE ET LE BIG DATA : UNE SANTÉ PRÉDICTIVE DONC PERSONNALISÉE ?

A. LE CONTEXTE

Nous vivons une ère de révolution et de mutations technologiques définie comme étant la troisième révolution industrielle. À l'origine de cette révolution se trouve la donnée par analogie parfois comparée à ce qu'a pu représenter hier le charbon, puis le pétrole dans les mutations conduisant à la production de masse toujours au cœur de notre société actuelle.

Le pétrole brut, comme la donnée brute, en soi n'est rien. Raffiné et transformé, il a donné naissance à la révolution pétrochimique et à la révolution des transports actuels. Correctement contextualisée et analysée au travers d'algorithmes bénéficiant de l'accroissement continu de la puissance de calcul croissant, la donnée brute va donner lieu à une formidable création de valeur.

Cette révolution s'appuie sur les capacités du Big Data. Celles-ci reposent tout d'abord sur un accroissement exponentiel du volume

(V) de données disponibles dues à la multiplication (**V pour Variété**) des objets (smartphone, PC, montres, consoles) et des capteurs (altimétrie, domotique, véhicules, électroménager…) connectés. L'ensemble de ces interconnexions croissantes définit le concept d'Internet des Objets (IoT) ou d'Internet du Tout Connecté dont résulte un doublement du volume (**V pour Volume**) des données disponibles tous les trois ans. Certains comparent cet Internet des Objets à la nouvelle machine à vapeur du XXIe siècle.

La conséquence de cette « déferlante » est la mise en données du monde ou la digitalisation du réel. À noter que l'analogie avec le charbon ou le pétrole, ressources finies, s'arrête là. La donnée est une ressource inépuisable.

Le Big Data repose également sur la puissance actuelle (**V pour Vélocité**) des capacités de stockage, des capacités de calcul et sur la sophistication des algorithmes.

Pour la première fois dans l'histoire de l'humanité, l'homme dispose d'un volume de données et d'une puissance de calcul qui lui permettent de s'affranchir des contraintes de l'échantillonnage, de l'exactitude des mesures ou de la nécessité de s'astreindre à travailler sur des données limitées analysées au moyen du calcul statistique habituel.

Si cette révolution de la donnée va bouleverser tous les secteurs et organisations économiques, **trois secteurs sont particulièrement impactés : Le logement** (domotique, économie d'énergie, électroménager), **la mobilité** (véhicules connectées et autonomes) **et la santé** (efficience des soins, vieillissement, suivi des maladies chroniques).

Ces secteurs sont pour les assureurs des marchés primordiaux et ils jouent sur ceux-ci un rôle majeur.

Au carrefour de ces marchés, **les assureurs se voient donc dans l'obligation** au travers de l'Internet des Objets et du Big Data **de repenser leurs offres de services, leurs organisations internes et leur *business model***. Et ceci afin d'une part, de ne pas être submergés et, d'autre part, d'être en situation d 'exploiter l'extraordinaire opportunité de création de Valeur (le quatrième V du Big Data après celui de Volume, Variété et Vélocité) qui se présente à eux.

B. IMPACTS ET PERSPECTIVES

Les assureurs sont historiquement des collecteurs de données sur lesquelles ils procèdent à des analyses. Ces analyses s'appuient sur les sinistres constatés, sur un grand ensemble de données sociétales et sur une mutualisation du risque permettant le calcul des primes.

L'analyse et l'évaluation du risque prennent jusqu'à présent peu en compte l'usage du bien assuré, que celui-ci soit matériel, (assurance IARD) ou immatériel comme la santé (assurance santé). **Les nouvelles techniques** de collecte de données, la banalisation des capteurs connectés, la digitalisation de la santé avec son flux croissant de données s'y rapportant et l'introduction d'algorithmes *ad hoc* capables de s'adapter afin de suivre les changements de comportement d'un assuré **vont bouleverser le modèle historique du calcul actuariel.**

Pour les assureurs, **il s'agit de faire converger la connaissance agrégée historique de leurs assurés et sinistres** (données off-line) **avec les données du nouvel environnement digital** (données on-line) désormais accessibles (fréquentation de sites Web, activité sur les réseaux sociaux, domotique, boîtiers embarqués, devices connectés…). Cette fusion de l'off-line et de l'on-line permet un ciblage par l'étude des signaux faibles et un suivi prédictif et personnalisé des assurés.

Cette nouvelle utilisation des données impacte toute la chaîne de valeur de l'assurance, à savoir :

- la conception des produits et la tarification ;
- la connaissance du client et du marché ;
- la lutte contre la fraude ;
- la gestion des sinistres ;
- les nouveaux services.

Une précision préalable toutefois : en France le stockage et l'utilisation de ces données personnelles sont encadrés par une réglementation très stricte, particulièrement pour ce qui concerne les données de santé, ce qui limite fortement le champ des possibles pour les assureurs.

La conception des produits, la tarification, la connaissance du client et des marchés

Pour la conception des produits, pour affiner la connaissance du marché et des clients, le marketing dispose ou va disposer d'outils sans cesse plus performants visant à améliorer la segmentation, la personnalisation et l'identification des assurés.

Comme expliqué plus haut, la multiplication des capteurs, le rapprochement et le croisement des données, la puissance des algorithmes permettent désormais non plus de s'interroger directement et seulement sur la seule valeur de la sinistralité, mais sur l'usage ou sur le comportement constaté sur le périmètre assuré.

Les exemples et possibilités sont innombrables. Des offres apparaissent déjà permettant de différencier en termes de prime certains comportements de conducteurs. Aux USA, certaines négociations d'assurance santé dépendent des résultats relevés par les objets connectés.

De même, l'étude des signaux faibles issus de l'IoT permet par exemple de réduire le risque d'attrition (perte d'un assuré) qui peut se produire lors d'un changement de véhicule (détection de consultations inhabituelles de sites Web d'offres automobiles) ou de logement (navigation sur des sites immobiliers) et d'anticiper en lui proposant de nouvelles offres adaptées.

La tarification va bénéficier de toute la puissance des outils nouveaux. Les variables habituelles (sexe, âge, type de voiture ou de logement, etc.) s'enrichissent des données de comportement et d'usage qui permettent une segmentation de l'offre et une adaptation en temps réel de la prime.

De fait, les assureurs passent d'une vision descriptive des événements sur laquelle était construit le calcul actuariel à une vision prédictive qui permet d'ajuster la gestion du risque au plus près de la personne.

La lutte anti-fraude

Ce sujet est primordial pour les assureurs. Près de 20 % des cas de fraude ne sont pas détectés. Leur coût pèse sur le poids de la sinistralité et donc sur le calcul des primes. Le Big Data, par son analyse des données et des signaux faibles, par sa capacité à détecter des comportements inhabituels et des usages inadaptés permet d'envisager une baisse significative de ce coût. La marge dégagée, en dehors d'un retour sur investissement, permet d'imaginer une baisse des cotisations et / ou la conception de nouvelles offres.

La gestion des sinistres

Outre la réduction de sinistralité attendue par le suivi et la connaissance personnalisée précédemment évoquée, la digitalisation croissante des *process*, et des flux, la réorganisation interne qui en découle, autorisent une optimisation de la gestion des sinistres qui entre dans le cadre des gains de productivité.

À noter que pour ce qui concerne la France, la mise en œuvre de la Carte Vitale associée aux télétransmissions des CPAM vers les assureurs complémentaires santé ont déjà largement permis cette optimisation.

Les nouveaux services

Ces nouveaux services découlent des nouveaux outils marketing qui identifient les nouvelles offres et besoins découlant du suivi prédictif et personnalisé. Ces nouveaux services améliorent le CRM, accroissent la fidélisation, réduisent l'attrition, favorisent une réduction du coût des campagnes de recrutement. Ils participent également fortement à une stratégie de différenciation et à la recherche d'un avantage comparatif.

En résumé, les assureurs peuvent attendre de l'Internet des Objets, des outils connectés et du Big Data des gains de productivité et une amélioration de leur qualité opérationnelle. Ils peuvent surtout affiner leurs offres et leurs services au bénéfice de l'assuré (gain de bien-être, réduction ou ajustement des primes, etc.).

Sur ce plan, la législation française très protectrice pour le consommateur (CNIL notamment) pourrait devenir un atout en permettant d'organiser la collecte et la mise à disposition de toutes les données nécessaires, **d'établir une relation de confiance entre les assureurs et leurs assurés.**

Ces derniers doivent rester maîtres de leurs données et doivent être informés (informations dans les espaces client des sites Web de l'assureur, campagne en *push* de communication et d'information ciblées) et convaincus des usages et des finalités de la collecte faite.

La transparence permet d'établir la confiance. La confiance d'autoriser la collecte. La collecte permet d'atteindre le volume critique permettant les analyses.

Les analyses conduisent à des diagnostics personnalisés ou prédictifs par le biais d'algorithmes qui sont porteurs de l'expertise et des secrets industriels de l'assureur.

Ces algorithmes deviendront d'ailleurs le cœur de la valeur ajoutée d'un assureur et la garantie de sa performance sur le marché, dans le strict respect du cadre législatif et d'une éthique systématiquement en faveur des individus.

Assurances et santé

Sur le périmètre de la santé, on doit noter la pression qui repose sur les assureurs santé. Dans la plupart des pays développés, compte tenu du poids croissant des maladies chroniques, les dépenses de santé croissent plus vite que le PIB. En France, cette part se situe au-delà de 11 % et on a assisté à un quasi-triplement sur les trois dernières décennies.

Cette augmentation s'explique par le vieillissement croissant de la population (donc plus de maladies chroniques à traiter sur de plus longues périodes), par le coût élevé des nouveaux traitements (nouvelles molécules en oncologie) et des nouvelles techniques d'exploration ou interventionnelles (imagerie, cathétérismes, greffes).

L'irruption proche en pratique quotidienne des nanotechnologies et des usages de la génomique ne va certainement pas inverser cette tendance.

Ces dépenses en inflation continue sont essentiellement consommées par le curatif et ne laissent qu'une faible part à la prévention (2 % à peu près).

En conséquence, la pression augmente sur la part complémentaire dévolue aux assureurs santé. Cette tension se constate également dans les domaines de la dépendance, du handicap, du vieillissement.

Cette tendance peut se décliner également dans des pays où le rôle de l'assurance santé de nature privée est plus significatif qu'en France. La problématique sera alors de garder le montant de la prime à un niveau acceptable pour le payeur.

En conséquence, les assureurs santé sont à la recherche de nouveaux *business models* susceptibles de répondre à ces difficultés. Dans ce contexte, l'Internet des Objets et le Big Data, par leurs capacités de prédictivité et d'individualisation du risque sont vus par le secteur comme des outils nouveaux, capables de générer de la valeur tout en respectant le modèle historique de la mutualisation. En effet, une tarification plus précise du risque d'un individu au sein d'un groupe n'empêche pas la mutualisation (au sein d'une entreprise, d'une branche, d'un secteur d'activité ou encore sur des principes liés à la non-discrimination), au contraire de nouvelles formes de mutualisation bien plus efficientes devraient apparaître.

Par ailleurs, **les assureurs complémentaires,** *via* leur rôle de financeur, doivent pouvoir **devenir des acteurs plus importants dans l'interaction patient / système de soins,** une meilleure connaissance des assurés se traduisant par la mise à disposition de services ou d'options encore plus adaptés.

Il est donc nécessaire de prendre acte que, dans le cadre de l'économie numérique, l'introduction progressive et croissante des TIC et des NBIC induit un bouleversement très profond de cette interaction : évolution des paradigmes du soin vers la prévention et le bien-être, mutations dans l'organisation du système sanitaire, nouveau comportement du citoyen-patient (parfois identifié comme un consomm'acteur) impliqué et connecté (réseaux sociaux, sites Web…), modification de la relation patient-professionnel de santé, etc.

En résumé, la digitalisation et la mise en données de la santé autorisent de nouvelles formes :

• de coordination dans le suivi des patients (plate-forme de services médico-sociale, sites Internet…) ;

• de collaboration entre les professionnels de santé en interaction approfondie et connexion permanente ;

• de gestion par le citoyen patient, y compris en mobilité, de ses informations de santé (stockage, partage, synthèse, surveillance, alerte, transmission…) ;

• de partage et de distribution de la connaissance entre les acteurs y compris les assureurs et d'éventuels nouveaux entrants (GAFA).

Ces évolutions se traduisent ou vont se traduire par :

• **une optimisation du parcours de soins ;**

• **la réduction des hospitalisations ;**

• **l'amélioration de l'observance thérapeutique ;**

• **une meilleure allocation des ressources destinées à la prévention.**

Financeurs du risque et financeurs du système, les assureurs santé ne peuvent que s'impliquer et soutenir ces tendances portées par l'Internet des Objets et les outils du Big Data.

Dans le même temps, le concept de santé et le mode de gestion de celle-ci évoluent et s'étendent. De fait, à partir de la bien connue définition de l'OMS (un état complet de bien-être physique, mental et social…), une mutation s'accomplit conduisant du concept historique purement biomédical à une approche holistique qui vise à appréhender tous les besoins d'une personne (affectifs, sanitaires, nutritionnels, culturels…). Les déterminants de la santé deviennent multiples et larges (comportement, environnement, etc.).

On doit donc apprendre à gérer la santé comme une ressource rare selon cette approche holistique qui privilégiera en particulier plus le préventif que le curatif.

La promotion de la santé ainsi mise en place favorise le développement de la prévention et particulièrement de la notion de bien-être. C'est dans ces deux secteurs (prévention, bien-être) que les capteurs connectés et les analyses Big Data sont particulièrement pertinents et performants.

Cette voie (gestion du capital santé réduisant ainsi le risque santé, usage extensif de capteurs connectés et analyses Big Data) est particulièrement identifiée et favorisée par les assureurs qui attendent des impacts positifs sur toute leur chaîne de valeur, à savoir :

• Un gisement d'économie possible (surveillance des maladies chroniques dans le cadre de la prévention secondaire ou tertiaire, baisse de la sinistralité à travers des jours d'hospitalisation gagnés, amélioration de l'observance médicamenteuse, baisse du coût indu des prestations évitables…).

• Un renforcement de la relation client (mieux connu, donc mieux suivi).

• La possibilité de nouveaux services (baisse du taux d'attrition, meilleure fidélisation, politique de différenciation, recherche d'avantages comparatifs). On citera par exemple : suivi de l'entrée / sortie d'hospitalisation, prévention des troubles musculo-squelettiques, lutte contre le stress au travail, coaching nutritionnel ou de l'activité physique, etc.

• L'ajustement des primes, gain de productivité. se traduisant par une amélioration de la qualité opérationnelle.

L'aboutissement de cette tendance initiée par les outils connectés et la mise en données de la santé va s'accélérer d'une part avec l'arrivée des nanotechnologies, mais surtout et particulièrement avec la place croissante prise par la génomique.

L'analyse et le décryptage du génome humain connaissent un processus d'automatisation croissant et de réduction de coût significatif autorisant son usage croissant dans les soins. La puissance des technologies mises en œuvre par le Big Data permet la détection des hauts potentiels de risque et l'optimisation de la réponse thérapeutique et assurantielle.

Cette révolution génomique parachève un changement de paradigme fondamental à savoir **le passage d'une médecine standardisée** (le même traitement pour telle pathologie quelle que soit la population atteinte) **vers une médecine personnalisée** (une maladie, un patient, un environnement, un traitement) également dénommée médecine de précision. C'est dans ce paysage nouveau que s'inscrit l'activité assurantielle de demain.

La question de la confiance, déjà évoquée précédemment, devient ici particulièrement sensible. L'assureur devra certes respecter les lois et cadres législatifs relatifs aux données de santé et à la protection de la vie privée.

Il devra aussi faire preuve de transparence et savoir convaincre l'assuré de la pertinence de sa démarche et du retour positif qu'il peut attendre de la collecte et de l'analyse de toutes les données collectées.

Il appartient aussi aux autorités publiques de faire évoluer notre cadre juridique actuel pour préserver tant la liberté et les droits des patients que la compétitivité des assureurs Français par rapport à leurs concurrents internationaux.

ARTICLE V
DONNÉES ANONYMISÉES, DONNÉES PSEUDONYMISÉES : QUEL NIVEAU D'AGRÉGATION ?

Les termes d'« anonymisation » et de « pseudonymisation » désignent des procédés désormais courants appliqués aux données de santé permettant soit de couper le lien entre l'identité de la personne et les données qui la concernent, soit d'être en mesure de « chaîner » les informations de cet individu sans en connaître l'identité.

L'anonymisation (ou désidentification) des données à caractère personnel désigne la méthode et le résultat du traitement de données à caractère personnel dans le but d'empêcher irréversiblement l'identification de la personne concernée. D'une manière générale, il ne suffit donc pas de supprimer directement des éléments qui sont, en eux-mêmes, identifiants pour garantir que toute identification de la personne n'est plus possible. Une solution d'anonymisation efficace doit empêcher la réidentification, ce qui ne se limite pas simplement à empêcher l'individualisation (isoler un individu dans un ensemble de données, retrouver le nom et/ou l'adresse d'une personne) mais également la corrélation (relier entre eux des ensembles de données distinctes concernant un même individu) et l'inférence (déduire de cet ensemble de données des informations sur un individu).

La pseudonymisation est une technique consistant à remplacer un attribut (généralement un attribut unique) par un autre dans un enregistrement. Le résultat de la pseudonymisation peut être indépendant de la valeur initiale (comme dans le cas d'un numéro aléatoire généré par le responsable du traitement ou d'un nom choisi par la personne concernée) ou il peut être dérivé des valeurs originales d'un attribut ou d'un ensemble d'attributs, par exemple au moyen d'une fonction de hachage ou d'un système de chiffrement. La personne physique est donc toujours susceptible d'être identifiée indirectement. Par conséquent, la pseudonymisation ne permet pas, à elle seule, de produire un ensemble de données anonymes, elle réduit le risque de mise en corrélation d'un ensemble de données avec l'identité originale d'une

personne concernée ; à ce titre, c'est une mesure de sécurité utile, puisqu'elle réduit les risques pour les personnes concernées, mais non une méthode d'anonymisation.

Ces définitions sont conformes à celles de l'avis du G 29 du 10 avril 2014[13] sur le sujet et bien sûr à celles du nouveau règlement européen sur la protection des données personnelles du 27 avril 2016[14].

Au regard de la protection des données à caractère personnel, les données pseudonymisées restent des données à caractère personnel et ne visent pas à exclure toute autre mesure de protection des données.

Il est intéressant de noter que dans le texte du règlement européen qui sera demain la loi de tous les États membres de l'Union européenne en matière de protection des données personnelles, la pseudonymisation est également visée au point 4 de l'article 6 sur la licéité du traitement comme une des garanties appropriées que le responsable de traitement peut mettre en œuvre pour justifier la possibilité de traiter des données collectées initialement pour une finalité déterminée pour une autre finalité considérée alors comme compatible.

Cette notion nouvelle de compatibilité est importante sur le plan juridique pour justifier l'utilisation ultérieure de données pour des finalités différentes de celles prévues initialement.

Or, la multiplication des données de sources différentes relatives à un même individu et les nouvelles capacités de traitement de ces données et notamment le « Data Maning » modifient considérablement la notion de réversibilité de l'anonymisation des données personnelles (ré-identification). Des données considérées comme anonymes à un moment donné peuvent présenter plus tard un risque élevé de ré-identification du fait de l'apparition de nouvelles techniques ou de

13. Opinion 05/2014 on Anonymisation Technique.

14. Aux termes du 5) de l'article 4 du règlement européen, la pseudonymisation est « *le traitement de données à caractère personnel de telle façon que celles-ci ne puissent être attribuées à une personne concernée précise sans avoir recours à des informations supplémentaires, pour autant que ces informations supplémentaires soient conservées séparément et soumises à des mesures techniques et organisationnelles afin de garantir que les données à caractère personnel ne sont pas attribuées à une personne physique identifiée ou identifiable.* »

nouvelles sources de données, particulièrement dans un contexte marqué par le Big Data. Les techniques les plus sûres d'anonymisation restent l'agrégation de données qui transforment des données individuelles en données collectives. Mais ces techniques interdisent nombre de traitements ultérieurs. Il est donc souvent légitime de préserver le caractère individuel des données tout en maîtrisant le risque de ré-identification des personnes.

L'anonymisation (action irréversible) reste souhaitable chaque fois qu'elle est possible et aboutit à des données impersonnelles. Dans tous les autres cas, les données individuelles doivent être considérées comme peudonymisées (ou indirectement nominatives) et présentent un risque plus ou moins élevé de ré-identification d'une part et de divulgation d'autre part. C'est l'évaluation de ces deux risques (ré-identification et divulgation) au regard de la sensibilité des données traitées qui doit conduire à des mesures de sécurité appropriées.

La distinction entre pseudonymisation et anonymisation est-elle suffisante à l'heure du Big Data ?

Aujourd'hui en effet le Data Mining qui consiste à traiter des données volumineuses à l'aide de techniques de calcul puissantes utilise soit des données anonymisées, soit des données pseudonymisées, et très rarement, pour ne pas dire jamais des données directement nominatives.

Or pour certaines études, les données anonymisées peuvent s'avérer d'un intérêt limité au regard des objectifs poursuivis, l'agrégation étant insuffisante pour traduire des phénomènes de santé que seules des données plus fines permettraient de révéler.

S'agissant de la pseudonymisation, celle-ci peut se heurter aux conditions actuelles du respect des règles de protection des données personnelles telles que définies en France par la loi informatique et libertés et demain en Europe par le règlement européen précité.

Comment en effet déterminer à l'avance la finalité précise pour laquelle on souhaite collecter des données et les traiter alors que ce qui caractérise justement les nouvelles modalités de recherche comme le Big Data, c'est justement de traiter des données sans connaître à l'avance la finalité ?

S'ajoute au respect de ce principe, l'obligation de formalités préalables auprès de la CNIL dont on sait la complexité et la longueur. La CNIL a de plus la faculté de certifier les processus d'anonymisation des données à caractère personnel, en vue notamment de la réutilisation d'informations publiques mises en ligne[15].

Si les textes donnent aujourd'hui quelques pistes, elles restent limitées et ne permettent pas de prendre en compte la réalité des besoins.

15. Article 11 de la loi n° 78-17 du 6 janvier 1978 modifiée par la loi 2016-41 du 26 janvier 2016 : « *Elle peut certifier ou homologuer et publier des référentiels ou des méthodologies générales aux fins de certification de la conformité à la présente loi de processus d'anonymisation des données à caractère personnel, notamment en vue de la réutilisation d'informations publiques mises en ligne dans les conditions prévues au titre II du livre III du code des relations entre le public et l'administration.* »

ARTICLE VI
LIBÉRER LES DONNÉES : PATIENTS, LES USAGERS AU CŒUR DE LA « DISRUPTION »

Grâce à la collecte et l'analyse de données massives, les techniques de prévention, de traitement, de diagnostic et de suivi des patients évoluent de façon accélérée depuis, que la santé mobile, notamment, s'offre sur les stores de téléchargement de nos smartphones.

Le recueil de ces données, communément appelé Big Data, est sans doute sur le point de transformer la méthodologie des chercheurs ainsi que les approches des différents acteurs institutionnels de la maîtrise des risques et des dépenses sanitaires dans notre pays.

Si la notion de Big Data est aujourd'hui la traduction « fun » de l'existence d'une collecte intensive et industrialisée des données de santé, elle ne dit toutefois rien des modalités de l'accès à ces mêmes informations. Pour cela, un autre anglicisme est souvent associé : celui de l'Open Data. Ce concept, né d'un mouvement sociétal favorable à un maximum de transparence et de démocratie citoyenne, répond au « droit de savoir » des usagers, mais pas seulement. L'Open Data en santé est une philosophie fondée sur le libre accès aux données de santé numériques, quel qu'en soit l'auteur. Elle suppose une diffusée structurée de ces données, selon une méthode et une licence ouverte garantissant son libre accès et sa réutilisation par tous, sans restriction technique, juridique ou financière.

Notre droit a récemment évolué pour moderniser notre système de santé et redéfinir, en ce sens, un processus d'accès aux données de santé Système national des données de santé (SNDS), composé des bases de données suivantes :

• les données produites dans le cadre du Programme de médicalisation des systèmes d'information (PMSI) qui traduisent l'activité des établissements de santé, publics ou privés ;

• les données du système national d'information interrégimes de l'as-

surance maladie (SNIIRAM) produites par les organismes gérant un régime de base d'assurance maladie ;

• le registre national des causes de décès ;

• les données produites par les maisons départementales des personnes handicapées sous l'autorité de la Caisse nationale de solidarité pour l'autonomie ;

• un échantillon représentatif des données de remboursement par bénéficiaire transmises par des organismes d'assurance maladie complémentaire et défini en concertation avec leurs représentants.

Le SNDS a pour finalité la mise à disposition des données pour contribuer à :

• l'information sur la santé ainsi que sur l'offre de soins, la prise en charge médico-sociale et leur qualités ;

• la définition, à la mise en œuvre et à l'évaluation des politiques de santé et de protection sociale ;

• la connaissance des dépenses de santé, des dépenses d'assurance maladie et des dépenses médico-sociales ;

• l'information des professionnels, des structures et des établissements de santé ou médico-sociaux sur leur activité ;

• la surveillance, à la veille et à la sécurité sanitaires ;

• la recherche, aux études, à l'évaluation et à l'innovation dans les domaines de la santé et de la prise en charge médico-sociale.

Les données du SNDS qui font l'objet d'une mise à la disposition du public sont traitées pour prendre la forme de statistiques agrégées ou de données individuelles constituées de telle sorte que l'identification, directe ou indirecte, des personnes concernées y est impossible. Ces données sont mises à disposition gratuitement. La réutilisation de ces données ne peut avoir ni pour objet ni pour effet d'identifier les personnes concernées.

L'accès aux données personnelles du SNDS par les organismes d'études et de recherches poursuivant un but lucratif (notamment les personnes produisant ou commercialisant des produits à finalité sanitaire, les établissements de crédit, les entreprises exerçant une

activité d'assurance directe ou de réassurance et les intermédiaires d'assurance) ne peut être autorisé que pour permettre des traitements à des fins de recherche, d'étude ou d'évaluation contribuant à l'une des six finalités susmentionnées et répondant à un motif d'intérêt public. Dans ce cadre, ces accès ne sont autorisés que « dans la mesure où ces actions sont rendues strictement nécessaires par les finalités de la recherche, de l'étude ou de l'évaluation ou par les missions de l'organisme concerné » et seules les données nécessaires à ce traitement peuvent être utilisées. Au-delà, une demande d'autorisation spécifique doit être réalisée auprès de la CNIL.

Le nouvel Institut national des données de santé (INDS) reçoit les demandes d'autorisation qui devront ensuite être soumises à l'avis du comité d'expertise pour les recherches, les études et les évaluations dans le domaine de la santé. Ce comité se prononcera sur la méthodologie retenue, la nécessité du recours à des données à caractère personnel, la pertinence de celles-ci par rapport à la finalité poursuivie et, le cas échéant, la qualité scientifique du projet. Sur saisine de la CNIL ou de sa propre initiative, l'avis de l'INDS pourra être requis sur le caractère d'intérêt public que présente la recherche, l'étude ou l'évaluation justifiant la demande de traitement. La CNIL appréciera ensuite le projet au regard des principes de protection des données personnelles et de l'intérêt que présente cette demande.

L'INDS publiera l'autorisation de la CNIL, la déclaration des intérêts, puis les résultats et la méthode.

L'UNAF, la FNATH et le CISS rappellent la nécessité d'intégrer les représentants des malades et des usagers au sein des comités stratégiques et comités de suivi qui devraient être créés.

C'est donc ainsi que le droit français envisage la dimension opérationnelle du Big Data en santé, estimant que les projets d'études des acteurs à but lucratif doivent se conformer à une rigueur augmentée pour accéder aux données du SNDS du fait de leur approche commerciale.

Communiqué de presse du HDI de septembre 2016

ACCÉDER AUX DONNÉES DE SANTÉ EN FRANCE EST UN PARCOURS LONG, COMPLEXE ET INÉGALITAIRE

Paris, le 26/09/16 - La loi du 26 janvier 2016 portant réforme de notre système de santé, affiche une volonté d'ouverture des données de santé, mais peine à convaincre sur l'ambition que doit se donner la France en la matière. Le Healthcare Data Institute alerte sur la complexité du dispositif mis en place et formule des propositions pour la lever.

Le traitement des données de santé devient une source potentielle de création de valeurs considérables en particulier pour la recherche scientifique et l'acquisition de nouvelles connaissances, de la gestion des soins à la définition des politiques de santé. En créant un nouveau Titre VI consacré à la mise à disposition des données de santé, la loi du 26 janvier 2016 portant réforme de notre système de soins semble montrer de fortes ambitions.

« *Force est de constater qu'en l'état du texte et avant la publication de ses décrets d'application, le dispositif mis en place n'apparaît en l'état ni opérationnel, ni de nature à permettre l'ouverture des données qu'il annonce en organisant une procédure lourde et complexe, particulièrement pour les entreprises du secteur privé* », explique Isabelle Hilali, présidente du Healthcare Data Institute.

Lourdeur accrue et absence de visibilité

Les membres du Healthcare Data Institute constatent une lourdeur accrue depuis le vote de la loi du 26 janvier 2016.

Au moins six étapes devront être franchies pour pouvoir mettre en œuvre un traitement de recherche, pas moins de six autorités sont susceptibles d'intervenir dans le processus (INDS, CNAMTS, CNIL, Comité d'expertise, Laboratoire de recherches ou bureau d'études, tiers de confiance...).

« *Les acteurs, qu'ils soient institutionnels ou économiques, ont besoin pour le développement de leur mission et de leur activité, de stabilité. Difficile de mener un projet industriel ou un grand projet national sereinement dans un contexte juridique qui évolue à l'envers d'autres textes et surtout contre le mouvement que porte aujourd'hui le numérique* », rappelle

Jeanne Bossi Malafosse, avocat à la Cour, Counsel chez DLA PIPER et membre du Healthcare Data Institute.

Des inégalités d'accès aux données

Pour les membres du Healthcare Data Institute, des inégalités d'accès aux données sont encore présentes dans les nouvelles conditions posées par l'article 193 de la loi du 26 janvier 2016. Elles témoignent d'un manque de confiance, voire d'une défiance à l'égard des acteurs privés qui sont pourtant au cœur de l'innovation dans le secteur de la santé aujourd'hui comme hier.

Le risque se traduit par une baisse de compétitivité de la France en matière d'analyse des données de santé. Certaines études sont aujourd'hui d'abord conduites avec des données étrangères, plus faciles d'accès et sans qu'aucune atteinte aux droits et libertés n'ait été constatée. Le risque est également de voir se développer des bases de données parallèles dévalorisant ainsi les bases de données nationales qui sont pourtant d'une grande valeur scientifique.

Gageons que les textes d'application de la loi mettront en place une gouvernance du Système national des données de santé qui soit ouverte, à l'heure où la richesse vient d'une association de tous les acteurs privés comme publics.

Ces constats conduisent donc aujourd'hui les membres du Healthcare Data Institute à demander une révision de ces dispositions.

Les quatre propositions du Healthcare Data Institute

• **Aligner** les procédures entre acteurs publics et privés dès lors qu'un intérêt de santé publique est poursuivi et que des garanties appropriées sont prises.

• **S'inscrire** dans un processus de contrôle fondé sur les principes qui commandent désormais la protection des données personnelles dans l'ensemble des pays de l'Union européenne (*Accountability et Privacy by design*) : contrôle *a priori* allégé et défini en fonction d'une étude de risque, contrôle *a posteriori* renforcé, sanctions beaucoup plus lourdes en cas de manquements constatés. À cet égard, les possibilités de sanctions de la CNIL doivent être renforcées.

• **Distinguer** selon la nature des données pour ne pas soumettre aux mêmes contraintes des données qui ne portent en elles qu'un faible risque de ré-identification. À cet égard, il ne faut pas confondre l'intensité du risque (les conséquences de sa survenue) avec la probabilité de la survenue de ce même risque.

• **Rendre véritablement effectives** les possibilités reconnues par les textes à la CNIL de simplifier les procédures (autorisations uniques, méthodologies de référence...).

> Lire la position détaillée du Healthcare Data Institute.

Mais le SNDS sera-t-il réellement le temple des données de santé ?

Aujourd'hui, grâce aux nombreux capteurs des smartphones et la diffusion galopante des applications santé et bien-être, les données de santé se disséminent un peu partout dans un paysage bien plus vaste et impressionniste que le seul SNDS dont on vient de voir la conception administrative.

Les tenants traditionnels du Big Data que sont notamment la CNAMTS et les établissements de santé du fait de leurs codages, n'ont, de fait, plus le monopole des données de santé. Et cette dispersion tient moins à l'Open Data qu'à la multiplicité des offreurs de services dématérialisés. Les données de santé se déclinent sous plusieurs aspects : médico-administratives dans une approche originelle, brutes et signifiantes lorsqu'elles sont pré-requises pour des applis qui permettent tantôt de suivre des constantes physiologiques, tantôt de géolocaliser les utilisateurs pour les situer par rapport à un établissement de santé, tantôt d'améliorer l'adhésion au traitement, quand il ne s'agit pas de proposer un coach minceur, un compagnon numérique pour mieux vieillir, un veilleur de nuit connecté, etc.

L'une des questions les plus prégnantes par rapport aux données collectées dans le cadre du *quantified self* porte sur leur avenir. Que deviennent, en effet, ces masses d'informations collectées avec le consentement plus ou moins éclairé des utilisateurs ? À qui profitent-elles réellement ? Et en quoi le recours au concept de l'Open

Data peut-il être utile pour ces nouveaux usages que nous désignons comme tels depuis plusieurs années maintenant ?

La proportion de données fournies par capteurs personnels dans l'ensemble des informations stockées devrait passer de 10 % à près de 90 % au cours de la prochaine décennie.

Les enjeux, pour les détenteurs et les hébergeurs de l'ensemble de ces données se mesurent au regard de la profitabilité que celles-ci représentent pour les promoteurs du bien-être et de la santé augmentés, mais pas seulement. Car si la valeur des données de santé excède, sur le « marché noir », celle des données bancaires, c'est bien parce qu'elles présentent un attrait particulier pour de grands comptes plus ou moins scrupuleux et bienveillants.

Après les cyberattaques subies par des laboratoires d'analyses médicales ou des établissements de santé, les hébergeurs de données requises par des applis santé ne seront-ils pas les prochains à se faire odieusement rançonner ?

Le droit positif est-il assez robuste pour garantir la sécurité des données de santé « éclatées » ? Plus globalement, sommes-nous outillés pour tirer parti, collectivement et individuellement du Big Data tout en protégeant les droits individuels des personnes ?

Le nouveau règlement européen sur la protection des données personnelles, paru au Journal officiel de l'Union européenne le 4 mai 2016 et qui entrera en application en 2018 devrait permettre à l'Europe de s'adapter aux nouvelles réalités du numérique en reconnaissant notamment :

• L'obligation de mettre à disposition d'une information claire, intelligible et aisément accessible aux personnes concernées par les traitements de données.

• L'expression du consentement : les utilisateurs doivent être informés de l'usage de leurs données et doivent en principe donner leur accord pour le traitement de celles-ci, ou pouvoir s'y opposer. La charge de la preuve du consentement incombe au responsable de traitement. La matérialisation de ce consentement doit être non ambiguë.

• Le droit à la portabilité des données permettant à une personne de récupérer les données qu'elle a fournies sous une forme aisément ré-

utilisable, et, le cas échéant, de les transférer ensuite à un tiers. Il s'agit ici de redonner aux personnes la maîtrise de leurs données, et de compenser en partie l'asymétrie entre le responsable de traitement et la personne concernée.

• Le droit, pour les associations actives dans le domaine de la protection des droits et libertés des personnes en matière de protection des données, d'introduire des recours collectifs en matière de protection des données personnelles.

• Le droit, pour toute personne ayant subi un dommage matériel ou moral du fait d'une violation du présent règlement, d'obtenir du responsable du traitement ou du sous-traitant réparation du préjudice subi.

Ce règlement européen fait ainsi évoluer le droit d'accès des personnes vers un droit à la portabilité en vertu duquel elles disposent désormais d'un « droit de retour » et d'une réappropriation des données les concernant. Le Big Data devient ainsi utile aux personnes qui deviennent en partie « décideurs » de l'utilisation de leurs données. À partir de ce précieux matériau, nombre de services pourront, demain, proposer des solutions à forte valeur ajoutée aux personnes qui, à bon droit, sont resituées au centre des flux.

Ces nouvelles approches, fortement individualisées, réorienteront la recherche vers des solutions adaptées aux personnes, à leur mode de vie, à leurs contraintes et permettront de définir avec finesse leurs micro-objectifs et les méthodes motivationnelles.

Plus collectivement, les données du Big Data, de l'Open Data et du Self Data pourraient s'avérer essentielles pour la recherche épidémiologique, dès lors qu'elles permettront aux professionnels de mieux connaître l'état de santé de la population, d'ajuster le traitement administré au patient en observant des modèles à plus grande échelle, ou de tirer de nouvelles conclusions, par exemple sur le rapport entre l'évolution d'une pathologie et les facteurs environnementaux.

Une meilleure exploitation des données de santé pourrait, de surcroît, entraîner des gains de productivité et des réductions de coûts dans le secteur de la santé (aux États-Unis, les prévisions sont de 300 milliards de dollars, en valeur, par an).

On mesure encore mal combien le Big Data pourrait modifier la conception que nous avons, tous, de notre rapport à la santé et de ce que devrait être le pilotage rationnel et efficient de notre politique de santé. Mais il y a tout lieu de croire que nous sommes à l'aube d'une révolution fondamentale que certains nomment « disruption », soutenue par la digitalisation de notre économie de service, par la prospérité de la miniaturisation et par la forte capitalisation des sociétés informatiques qui entraînent, dans leurs sillages, nombre de start-up innovantes.

Les droits des personnes doivent être perçus comme des ressources utiles dans cette nouvelle économie qui pourrait demain être dominée par la science des algorithmes. C'est, en tout état de cause, à cette condition que le Big Data profitera à tous et en premier lieux aux individus eux-mêmes.

À PROPOS DU
HEALTHCARE DATA INSTITUTE

Créé en 2014, le Healthcare Data Institute est le premier Think Tank international consacré au Big Data dans le domaine de la santé, il agit comme un catalyseur d'idées et de projets autour du Big Data dans l'écosystème santé.

Le conseil d'administration du Healthcare Data Institute rassemble des représentants d'Aviesan, Axa, Caisse des Dépôts et Consignation, CEA, CRI, Groupe Elsan, McKinsey&Company, Orange Healthcare Quintiles-IMS et Sanofi.

HEALTHCARE
DATA INSTITUTE

CONTACT
Pierre-Yves ARNOUX
office@healthcaredatainstitute.com
+33 (0)6 07 13 77 13

21, rue Jasmin
75016 PARIS - FRANCE

@HCDATAINSTITUTE
healthcaredatainstitute.com