Big data et santé publique : plus que jamais, les enjeux de la connaissance

Chloé Dimeglio
Université Toulouse III UMR 1027, CHU Toulouse Service d'épidémiologie
Cyrille Delpierre
Inserm UMR1027
Nicolas Savy
Institut mathématiques de Toulouse, Université Toulouse III
Thierry Lang
Université Toulouse III UMR 1027, CHU Toulouse Service d'épidémiologie, Institut fédératif d'études et de recherche santé société (Iferiss)

Remerciements : le travail de ce groupe est issu d'un projet de recherche soutenu par la Région Midi-Pyrénées sous le numéro 15/06/12.01

De nombreuses bases de données existent dans le domaine de la santé. Quelles données intégrer ? Comment s'assurer de leur fiabilité ? Pour quelle utilité ? avec quels risques ?

« *Le Big Data c'est comme le sexe à l'adolescence : Tout le monde en parle, personne ne sait vraiment comment ça marche. Tout le monde pense que les autres le font, donc tout le monde prétend le faire. Les seuls qui n'en parlent pas sont ceux qui l'ont déjà fait car leur première fois ne s'est pas très bien passée.* »

C'est ainsi que Dan Ariely, professeur de psychologie et d'économie comportementale à l'Université Duke (USA), parlait du Big Data en 2013 [1]. Pour établir finalement un constat simple : personne ne comprend réellement ce qu'est le Big Data.

Qu'est-ce que le Big Data ?

Si on cherche à se détacher du fantasme adolescent, force nous est de reconnaître une chose : du *Data Mining* au *Cloud-Computing*, le terme de Big Data recouvre aujourd'hui une grande variété de techniques et d'innovations, qu'elles touchent au stockage de données ou à leur traitement. Dans cette diversité de termes et d'utilisation, la grille de lecture de Doug Laney fait encore foi [3] : la notion de Big Data est associée aux 3V, à savoir le Volume (qui rend compte de l'utilisation massive des données), la Variété (qui touche à l'hétérogénéité des contenus), et la Vélocité (qui est associée à la vitesse de traitement des données). Le Big Data c'est donc le traitement rapide de données hétérogènes massives. Cette notion, très largement exploitée et vulgarisée par les GAFA (Google Apple Facebook Amazon) omet cependant une dimension importante. À l'heure où le volume de données – qu'elles soient structurées ou non structurées – devient difficilement exploitable avec des solutions classiques, il est crucial de parvenir à utiliser des données de sources diverses et de natures diverses pour produire une « autre » information certes, mais une information qui soit également plus complète et plus fiable. Aux 3V du Big Data semble ainsi manquer celui de la Véracité de l'information.

Dans son rapport de novembre 2014 sur les sciences du vivant, le cabinet Ernst and Young établit que la santé est le secteur où les Big Data ont le plus de potentiel [2]. Effectivement, on vérifie aisément que la plupart des problématiques liées à la définition et à la mise en place d'une si populaire « e-santé » sont indissociables des Big Data. Google, *via* Google Flu Trends, prétendait pouvoir prévoir la diffusion des épidémies par l'exploitation de gros volumes de données de santé. Calico a été créé pour permettre le dépistage des maladies par l'utilisation de gros volumes de données génétiques.

Cependant, cet horizon si étendu, ouvert à la santé *via* les Big Data, reste aujourd'hui largement sous-exploité. En santé tout particulièrement, la notion de Big Data demeure centrée sur l'utilisation de gros volumes de données, souvent biologiques (génome, omic), à visée individuelle. Ainsi vise-t-elle la médecine prédictive, la détermination de risques individuels ou la décision diagnostique par la démultiplication des données biologiques individuelles, pour

une santé « à la carte », popularisée par le terme de « médecine personnalisée ».

Or, la notion de santé n'est pas entièrement recouverte par celle de pathologie, et donc encore moins par cette réduction à la détermination biologique d'un risque pathologique. Pour approcher une compréhension globale des phénomènes de santé, il s'agit donc de prendre en compte ses déterminants larges, environnementaux, socio-économiques, psychologiques, biologiques, qui font de la santé un domaine complexe, interdisciplinaire et trans-dimensionnel.

Cet aspect pose la question de la valeur de la donnée, qui doit alors être ajoutée aux V du Big Data précédemment exposés (Volume, Variété, Vélocité, Véracité, Valeur). Dans quel but et pour quoi faire ces données sont-elles produites ? Pour répondre à quelle question ? Le modèle de la santé choisi devient alors un enjeu primordial lorsqu'on invoque le Big Data en santé. Loin d'écarter les hypothèses *a priori*, n'être guidé que par les données, le Big Data en santé impose de se positionner sur le modèle choisi pour définir la santé. Ce Big Data est alors conditionné par des hypothèses et constitue un moyen de mettre en lumière une vision spécifique de la santé.

La génération d'hypothèses en rapport avec la santé et ses déterminants larges, environnementaux, socio-économiques suppose donc de mettre en connexion de nombreuses données sanitaires et extra-sanitaires. Celles-ci sont contenues dans un grand nombre de bases de données sous des formes plus ou moins structurées, avec des méthodes de mesure variables, des données manquantes, des origines disciplinaires, sectorielles et des modes de recueil extrêmement variés.

Les modèles de la santé, les choix de méthode, et le contrôle de l'information deviennent donc des enjeux essentiels pour notre système de santé publique. À l'heure où émergent des groupes de travail sur ces différentes problématiques, il nous semble essentiel d'intégrer toutes les approches dans un dialogue serré, incluant les questions méthodologiques, sociétales et éthiques. Cette question ne peut échapper à une approche transversale et interdisciplinaire.

Les enjeux de méthode se sont développés avec l'ouverture des bases de données de santé. Des applications concernent déjà l'exploitation de bases de données existantes, appariées sur un identifiant commun (le Répertoire national d'identification des personnes physiques de l'Insee [RNIPP] ou l'identifiant du Système national d'information inter-régimes de l'Assurance maladie [SNIIRAM], par exemple). Toutefois, le croisement de ces bases de données se heurte à des données manquantes et surtout à des modes de mesure variables selon les bases et probablement les années. De façon générale, pour approcher une définition globale et opérationnelle du Big Data en santé, il s'agit alors de résoudre les problèmes de mesure, de gestion de données déclaratives, c'est-à-dire de source, de type et de nature variés. Dans ce contexte, la fusion de bases de données apparaît comme un élément central dans le domaine des Big Data en santé. Cette mise en place d'un processus de recoupement de bases de données différentes, hétérogènes s'inscrit dans un processus opérationnel et prospectif du Big Data en santé. De la même manière les données issues d'objets connectés, de développement rapide dans le domaine de la santé, peuvent également être perçues comme des variables longitudinales et doivent, à ce titre, être intégrées dans les problématiques à traiter. Les applications et les besoins dans le domaine de la santé sont donc multiples.

Il importe également de s'interroger sur les conséquences sociales et sanitaires que peuvent avoir ces développements techniques. Les expériences en matière d'innovations technologiques suggèrent qu'il est nécessaire d'être attentif à leur impact sur la santé, aux inégalités sociales de santé et aux populations exclues. L'accès au numérique reste une préoccupation, mais l'exemple des téléphones portables, en Afrique ou parmi les réfugiés syriens, montre que les technologies peuvent être aidantes, y compris dans des situations d'extrême pauvreté. Une tache aveugle pourrait en revanche concerner des populations exclues de la collecte des données, comme les étrangers sans-papiers, les populations précaires, sans abri… Se pose aussi la question de la couverture de la population, de la représentativité de cette dernière au regard des outils numériques utilisés, avec toutes les limites potentielles que de tels travaux conduits sur des populations finalement sélectionnées pourraient avoir en termes de santé publique.

S'assurer de la véracité et du contrôle des informations produites

Pouvoir produire l'information soulève le plus souvent la question du Volume et de la Vélocité, donc principalement celle des moyens techniques informatiques, mais la possibilité de l'interpréter et de la diffuser pose la question de considérer l'information en santé comme un bien commun. L'équilibre entre l'action publique, le secteur privé et le rôle de l'État est essentiel à définir pour s'interroger sur le mode de production de l'information. Quelles sont les méthodes mises en œuvre, pour quels algorithmes ? Sur quels modèles et quelles hypothèses mathématiques seront basées les informations produites ? La difficulté croissante, y compris pour les mathématiciens et les statisticiens, à comprendre les modèles utilisés et leurs hypothèses porte en germe un risque de perte de contrôle de l'information. Ce danger est déjà largement manifeste dans le secteur des assurances où il est reconnu que les algorithmes mis en œuvre permettent effectivement d'adapter les produits aux situations individuelles sans qu'il y ait une quelconque maîtrise des hypothèses et des méthodes statistiques les sous-tendant.

Malgré les techniques d'imputation des données manquantes, y compris dans les techniques de fusion de données, la qualité et le modèle qui orientent le recueil des données façonnent l'information produite. Des travaux sociologiques qui ont suivi la mise en place du dossier informatisé dans des opérations zéro papier ont bien montré que des pans entiers de l'information nécessaire au soin continuaient de circuler sous forme écrite, de post-it ou de feuilles glissées dans la poche [4]… Cette information ne sera pas disponible dans une analyse étiquetée Big Data, pas plus que des données sur le lien santé travail ou emploi, si elles ne sont dans aucun système d'information. Rappelons que seulement 10 % des maladies à caractère professionnel sont aujourd'hui reconnues comme telles [5]. Pour les soins, l'enjeu est de ne pas abandonner la prise en charge des malades (le « care ») et ne pas résumer les soins à la prescription de traitements optimisés, fussent-ils personnalisés.

Une question centrale est donc celle de la véracité de l'information produite et peut-être surtout des moyens de la contrôler. Les choix techniques précédents ne sont

Références

1. Ariely D. https://twitter.com/danariely/status/287952257926971392
2. Ernst & Young. *Étude (Big) data - Où en sont les entreprises françaises ?* 2014. http://www.ey.com/Publication/vwLUAssets/EY-etude-big-data-2014/$FILE/EY-etude-big-data-2014.pdf
3. Laney D. « 3D Data Management : Controlling Data Volume, Velocity, and Variety ». *Application Delivery Strategies* Meta Group, 2001, http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf
4. Marrast P., Zaraté P. « Annotation management: a Group decision support system for nurses tasks ». *J Decision Systems*, 2015, vol. 24, no 2, p. 105-116.
5. Rivière S., Cadéac-Birman H., Chevalier A., Valenty M. « Surveillance de la santé au travail : exemple de la "Quinzaine Maladie à caractère professionnel (MCP)" en Midi-Pyrénées, France, 2006 ». *BEH,* 2008 :32.

pas dissociables d'une réflexion sur les modèles de la santé. Le « modèle français » de la santé repose depuis plus d'un siècle sur les soins curatifs, dans une relation singulière entre soignant et soigné. Cette vision individuelle se traduit dans l'information annoncée par les Big Data comme porteuses d'une médecine « personnalisée », productrice de normes et de normativité, soulignant la responsabilité individuelle des comportements et *in fine* substituant l'assurance à la solidarité. Le pouvoir de contrôle social de l'information produite, de ciblage de population est loin d'être anodin.

Le potentiel des Big Data est pourtant fécond pour également souligner et documenter le rôle des déterminants sociaux et sociétaux de la santé, une orientation dans laquelle le système de santé français s'est engagé, avec la prise en compte des déterminants sociaux et des inégalités sociales de santé. Replacer la santé dans un contexte social, économique et sociétal grâce à des données hors du système de santé est une démarche en faveur de laquelle le HCSP a plaidé, dans ses travaux, pour suivre l'évolution des inégalités sociales de santé.

Ce traitement simultané de différentes bases soulève bien sûr la question de la confidentialité des données. Cette garantie pourrait ne plus être assurée à l'ère de l'interopérabilité des bases de données qui pourrait mécaniquement rendre identifiables des données originairement anonymes. Les conséquences sur les libertés, la vie privée, des pratiques commerciales ou discriminatoires, par exemple dans le domaine de l'assurance santé ne sont pas de simples effets secondaires marginaux.

Ces développements de méthodes sont potentiellement extraordinairement féconds pour arriver à générer des connaissances dans une optique intersectorielle et interdisciplinaire, à partir de variables de natures différentes. L'approche s'inscrirait donc à la fois dans une vision globale des mécanismes de santé mais également dans une démarche d'une production qualitative de l'information, orientant le Big Data en santé vers deux nouveaux V, Véracité et Valeur. « *Traiter un grand nombre d'informations variées plus rapidement certes, mais à quoi bon si ça n'est pas pour améliorer la valeur de l'information utile ?* ». Pour revenir à la métaphore adolescente de Dan Ariely : « *Le Big Data en Santé, c'est bien comme le sexe à l'adolescence : ça demeure une expérience fondatrice* ».