UNIVERSITÉ LILLE 2 DROIT ET SANTÉ

FACULTÉ DE MÉDECINE HENRI WAREMBOURG

Année : 2015

THÈSE POUR LE DIPLÔME D'ÉTAT
DE DOCTEUR EN MÉDECINE

Vers une définition des Big data en santé basée sur la littérature

Présentée et soutenue publiquement le 11 mai 2015 à 16h
au Pôle Formation

Par Émilie Baro

_____

JURY
Président :
Monsieur le Professeur Régis Beuscart
Assesseurs :
Monsieur le Professeur Jean-Louis Salomez
Monsieur le Professeur Philippe Amouyel
Monsieur le Professeur Alain Duhamel
Directeur de Thèse :
Monsieur le Docteur Emmanuel Chazard

_____

Avertissement

La Faculté n'entend donner aucune approbation aux opinions émises dans les thèses : celles-ci sont propres à leurs auteurs.

Liste des abréviations

ADN	Acide désoxyribonucléique
CHRU	Centre Hospitalier Régional Universitaire
Cnamts	Caisse nationale de l'Assurance maladie des travailleurs salariés
IC95 %	Intervalle de confiance à 95 %
Log	Logarithme décimal
MeSH	*Medical Subject Headings*
PMC	PubMed Central
PMSI	Programme de médicalisation des systèmes d'information
SADM	Systèmes d'aide à la décision médicale
SNIIRAM	Système national d'information inter-régimes de l'Assurance maladie
XML	*Extensible Markup Language* (langage balisé extensible)

Table des matières

RÉSUMÉ

Contexte : Le terme « Big data » émerge récemment dans la littérature scientifique. Ce terme n'est pas encore référencé dans le MeSH (*Medical Subject Headings*). Or son usage semble ambigu et les propriétés attribuées à ce terme par les auteurs varient selon les articles. L'objectif de ce travail est de proposer une définition du terme « Big data » à partir d'une revue de la littérature incluant les articles mentionnant ce terme et de décrire systématiquement les propriétés rattachées à ce terme par les auteurs.

Méthode : Nous avons conduit une recherche systématique de la base de données PubMed de tous les articles publiés jusqu'au 9 mai 2014 en utilisant le terme de recherche « Big data ». Ces articles ont été classés en domaines d'études. Le nombre d'individus statistiques (*n*) et le nombre de variables (*p*) ont été relevés pour les articles décrivant un jeu de données. Nous avons également considéré les caractéristiques attribuées aux Big data par les auteurs. En s'appuyant sur cette analyse, une définition des Big data a été proposée.

Résultats : Cent quatre-vingt-seize articles ont été inclus. Trois principales catégories d'études ont été identifiées : les spécialités « omiques », les spécialités médicales et la santé publique. Les Big data peuvent être définies comme des données avec un *Log(n * p)* supérieur ou égal à 7. Les propriétés des Big data sont ses grandes variétés de données et leur importante vélocité. Les Big data soulèvent des défis concernant la véracité, la gestion des données, l'extraction d'informations utiles, le partage des informations et l'existence d'experts humains ayant à la fois des compétences cliniques et analytiques. L'émergence des Big data nécessitent la

création de nouvelles méthodes de calcul qui optimisent la gestion de données. Les concepts reliés sont la réutilisation des données (*data reuse*), la possibilité de connaissances erronées et la question de la confidentialité des données.

Conclusion : Les Big data sont définies par le volume. La taille des données qui les qualifie de « Big data » va probablement augmenter avec le temps. Les Big data ne doivent pas être confondues avec le *data reuse :* les données peuvent être massives sans être forcément réutilisées dans un autre objectif, par exemple dans le cas des spécialités « omiques ». Inversement, des données peuvent être réutilisées sans être nécessairement de grande dimension. C'est le cas par exemple de l'utilisation secondaire du dossier patient informatisé.

INTRODUCTION

I. Le contexte des Big data

A. Les Big data : volume ou technologie ?

Le terme anglo-saxon « Big data » n'a pas d'équivalent en français. On parle parfois de « données massives », de « données de grande dimension » ou de « volumes massifs de données ». Mais on parle également de concept, de phénomène, ou encore de discipline Big data. Ce terme désigne le traitement automatisé de grandes quantités de données pour en extraire des informations. À cette fin, on recourt à des nouvelles procédures d'analyse. Le terme « Big data » est aussi devenu synonyme de « data analysis » (analyse des données).

Les Big data correspondent donc à la fois à une masse considérable de données, mais aussi à la technologie mise en œuvre pour les gérer dans le but d'en extraire des connaissances.

B. Aux origines des Big data

C'est en 1944 que le problème de stockage et d'accès aux données a été mentionné pour la première fois par Fremont Rider, un bibliothécaire de la Wesleyan University (1). Il avait calculé que la taille des bibliothèques universitaires américaines doublait en moyenne tous les seize ans. Compte tenu de ce taux de

croissance, il avait estimé que la bibliothèque de Yale comporterait en 2040 environ 200 millions d'ouvrages, ce qui représenterait 10 000 kilomètres de rayons.

Le terme « Big data » aurait été employé pour la première fois en 1997 dans un article publié par des chercheurs américains à la NASA (*National Aeronautics and Space Administration*). Ils affirmaient alors que l'augmentation du volume des données devenait problématique pour les systèmes informatiques de l'époque. C'est ce qu'ils ont appelé le « problème des Big data » (2).

À la fin des années 1990, la puissance du matériel informatique s'est considérablement développé. Les sources de données se sont multipliées : l'Internet, les réseaux sociaux, la téléphonie mobile. Au début des années 2000, l'avènement d'outils comme le *cloud computing*[1] a permis de stocker des données à moindre coût. Depuis, l'espace sur le *cloud* a fortement augmenté et a contribué à l'essor du phénomène « Big data » (3).

C. Usages et opportunités des Big data

L'augmentation des données et le développement d'outils informatiques permettant de les analyser offrent d'innombrables possibilités. Les Big data se rencontrent dans des domaines aussi variés que les sciences, le marketing, l'industrie, la finance, les transports, l'écologie, l'éducation, mais aussi la santé.

Des modèles mathématiques et des algorithmes permettent de faire de l'analyse prédictive grâce aux données : l'analyse de masses de données passées et présentes permet ainsi de prédire, avec un certain degré de certitude, des comportements ou des événements (4).

Par exemple, en marketing, l'analyse des Big data permet de mieux comprendre l'utilisation des services par les usagers et d'affiner l'offre.

1. Cf. définition en Annexe 2.

Dans le domaine des transports, les données provenant des « pass » de transports en commun, des vélos et des voitures « communes », mais aussi de la géolocalisation de personnes ou de voitures, sont utilisées pour modéliser les déplacements des populations afin d'adapter les infrastructures et les services (5).

Dans le domaine des sciences, les analyses des Big data sont des outils qui ne se substituent pas à la compréhension des scientifiques : elles attirent l'attention sur des corrélations détectées afin que ces derniers recherchent ensuite des explications causales. En science, les Big data peuvent donc être vues comme un outil, à l'image d'un microscope, pour faire progresser la connaissance (6).

II. Les Big data dans le domaine de la santé

A. Big data et data reuse

Une distinction doit être faite entre les termes « Big data » et « Data reuse ». D'une part, dans le domaine de la santé, les activités transactionnelles génèrent d'importants volumes de données, que l'on appelle les « Big data ». Ces données sont recueillies informatiquement en routine pour répondre à une question définie initialement lors du recueil des données. Il s'agit par exemple des résultats d'analyses biologiques de patients, ou encore les médicaments prescrits et administrés.

D'autre part, on peut s'intéresser à la réutilisation de ces données à des fins scientifiques. La réutilisation de ces données déjà recueillies est qualifiée de « data reuse » (ou « data re-use »). Ce terme désigne le fait d'utiliser des données dans un but différent de celui qui était prévu initialement lors de la collecte des données (7), il désigne donc le fait d'utiliser dans un but décisionnel des données qui étaient recueillies dans un but transactionnel. L'intérêt est de valoriser au maximum les

données et d'obtenir de nouveaux résultats répondant à une nouvelle question, et cela à moindre coût.

Habituellement, les données collectées en routine puis réutilisées sont caractérisées par un volume important. Les Big data sont donc généralement une propriété des données pouvant être réutilisées.

B. Exemples de possibilités de réutilisation des données

1. En génomique

Le terme « génomique » désigne une discipline de la biologie qui étudie la structure et le fonctionnement de l'ensemble du patrimoine génétique (génome) d'un individu, contrairement à la génétique classique qui s'intéresse à un gène ou un groupe de gènes particulier (8). Cette discipline est née avec le projet de séquençage du génome humain entre 1993 et 2003. Le génome est souvent comparé à une encyclopédie dont les différents volumes seraient les chromosomes (23 pour les humains). Les gènes seraient les phrases contenues dans ces volumes et ces phrases seraient écrites dans un langage génétique représenté par quatre bases (adénine, thymine, cytosine et guanine, désignées par leurs initiales ATCG). Le génome humain comporte approximativement 3 milliards de paires de bases (9), la variation entre deux génomes humains étant d'environ une base pour 1000 (10). La recherche génomique gère donc des quantités massives de données qui nécessitent une grande capacité de calcul et de stockage. Grâce aux progrès technologiques, il est désormais possible de séquencer un génome entier en quelques heures et pour quelques milliers d'euros (11).

Le croisement de données génétiques, d'informations médicales, d'informations sur le comportement des individus et sur leur environnement (type de profession, lieu d'habitation, etc.) et le développement de nouveaux outils d'analyses

pourraient permettre de mieux comprendre l'évolution de pathologies, d'améliorer les mesures de prévention ou encore les protocoles de soins. C'est la combinaison de ces informations qui permettrait d'ouvrir la porte à la « médecine préventive et personnalisée » (12). Cette pratique consiste à traiter chaque patient de façon individualisée en fonction des spécificités génétiques et biologiques de sa maladie, de l'environnement du patient et de son mode de vie.

2. Détection d'épidémie

La réutilisation de données massives peut permettre une anticipation précise de la survenue d'épidémie. Par exemple, *HealthMap* est un outil de cartographie en ligne développé par des chercheurs de l'hôpital pour enfants de Boston et de l'Université d'Harvard en 2006. Cet outil s'appuie sur des données issues de sites d'informations en ligne, de réseaux sociaux, de blogs et de sites gouvernementaux, afin de détecter des épidémies naissantes (13).

Le *Google Flu Trends* est un autre exemple de réutilisation de données massives pour détecter précocement la survenue d'épidémie par l'analyse des recherches des internautes sur Google (14).

3. Aide au diagnostic médical et à la décision thérapeutique

Les systèmes d'aide à la décision médicale (SADM) sont des applications informatiques dont le but est de fournir aux cliniciens les informations décrivant la situation clinique d'un patient ainsi que les connaissances appropriées à cette situation, afin d'améliorer la qualité des soins et la santé des patients (15). Les SADM sont apparus au début des années 1990 aux États-Unis. Mais ces outils ne pouvaient alors pas rassembler et synthétiser l'ensemble des connaissances scientifiques, ils ne prenaient pas en compte l'évolution des connaissances

médicales et ne permettaient pas de personnaliser le traitement en fonction des spécificités du patient. C'est à ce niveau que la réutilisation des données massives pourraient jouer un rôle important en permettant d'exploiter très rapidement différents types de données structurées et non structurées, générées par des sources multiples. Une des méthodes disponibles est l'analyse sémantique, c'est-à-dire la capacité à comprendre un document rédigé en langage naturel (par exemple un compte-rendu opératoire) pour en tirer des données exploitables.

Par exemple, le projet européen PSIP (*Patient safety through intelligent procedures in medication* – La sécurité des patients par des procédures intelligentes de prescription médicamenteuse) initié par le CHRU de Lille a pour objectif de développer des applications informatiques innovantes capables de détecter automatiquement des situations à risque d'effets indésirables liés aux médicaments et de fournir aux professionnels de santé des informations pertinentes pour les aider à prévenir ces effets indésirables (16).

Un autre exemple est le programme informatique d'intelligence artificielle Watson conçu par IBM qui pourrait trouver une application dans le domaine médical. En comparant les données médicales (analyses biologiques, résultats d'imagerie, etc.) et personnelles (âge, antécédents, etc.) du patient et les connaissances emmagasinées (littérature scientifique), Watson propose un diagnostic, suggère un traitement et donne le niveau de fiabilité de sa réponse (5).

4. Autres opportunités de réutilisation des données à l'hôpital

En France, les hôpitaux publics et privés participent déjà, sur le plan légal, à la création et au partage de données anonymisées en les transmettant dans le cadre du Programme de médicalisation des systèmes d'information (PMSI). Le nombre de séjours hospitaliers en 2011 s'élevait à 17 millions (17). La codification des actes

avec le PMSI et le dossier patient informatisé ouvrent donc la voie à de nombreuses analyses.

Aujourd'hui, des cabinets de conseil en management et stratégie proposent, en s'appuyant sur l'analyse des données disponibles au sein des hôpitaux, de développer des processus d'amélioration continue en matière de gestion et de parvenir à une meilleure compréhension des trajectoires des patients, permettant ainsi une gestion optimisée des ressources disponibles (18).

C. Investissement dans la réutilisation des données en santé

1. À l'étranger

De nombreux pays investissent dans la réutilisation des données dans le domaine de la santé.

Au Royaume-Uni, le *National Cancer Registration Service* dispose d'une base de données qui permet de retracer le suivi médical complet de patients atteints de cancer. Les informations relatives à l'évolution de la maladie et la réponse au traitement peuvent être reliées aux analyses moléculaires et génomiques des patients (19).

Aux États-Unis, le projet « CATCH » est basé sur un partenariat entre le MIT (*Massachusetts Institute of Technology*) et le *Massachusetts General Hospital*. Son but est de développer des outils d'analyse permettant de combiner des informations médicales, comportementales et génétiques de patients atteints de certaines maladies chroniques, afin de mieux comprendre ces maladies (20).

À Singapour, un hôpital utilise l'analyse prédictive pour diminuer les taux de réadmissions à l'hôpital. Des données concernant des patients réadmis à l'hôpital à plus de deux reprises dans un intervalle de six mois sont analysées et servent à l'élaboration d'un modèle prédictif. L'analyse permet de produire une liste des

patients qui sont à risque d'être réadmis dans les mois suivants. Le modèle prédictif permet d'anticiper la demande de soins et ainsi améliorer la coordination de la prise en charge des patients après un séjour à l'hôpital, notamment par des soins préventifs à domicile (21).

2. En France

En France, du fait principalement des contraintes associées à la protection des données, de nombreuses données relatives à la santé ne sont pas valorisées. Le peu de recours aux analyses de données dans la gestion et la prise de décision s'explique également par le cloisonnement des données. Les données récoltées par deux administrations différentes sont parfois difficilement compatibles. Cette incompatibilité s'explique par le fait que les données administratives ne sont généralement pas recueillies à des fins d'analyse, mais pour la gestion interne (5).

Le Gouvernement français a validé en juillet 2014 une « feuille de route Big data » (22). Celle-ci s'inscrit dans un effort de soutien de l'État de 60 millions d'euros sur trois ans au développement de la filière Big data en France. La Caisse nationale de l'assurance maladie des travailleurs salariés (Cnamts) et l'École polytechnique ont signé fin 2014 une convention de partenariat de recherche et développement pour une durée de trois ans qui vise à favoriser le développement des technologies du Big data appliqué au domaine de la santé. Cette collaboration a pour ambition de déployer de nouvelles pistes d'exploitations des données du Système national d'information inter-régimes de l'Assurance Maladie (SNIIRAM). Créé en 1999 par la loi de financement de la sécurité sociale, le SNIIRAM regroupe des données complètes et détaillées sur le parcours des patients et l'organisation du système de soins. Il rassemble les données de remboursements et d'hospitalisation des bénéficiaires de l'ensemble des régimes d'Assurance maladie obligatoire en France. Ce partenariat entre la Cnamts et l'École polytechnique abordera un programme de

développement d'algorithmes définis au regard des missions de la Cnamts. La détection d'anomalies en pharmaco-épidémiologie, l'identification de facteurs utiles à l'analyse des parcours de soins et la lutte contre la fraude font partie des thèmes de recherche identifiés.

La Cnamts a diffusé en décembre 2014 un premier jeu de données de santé anonymisées réutilisables sur le portail *open data* du Gouvernement (23). L'*open data* est un processus d'ouverture des données publiques ou privées pour les rendre disponibles à l'ensemble de la population sans restriction juridique, technique ou financière. L'*open data* contribue à l'augmentation des données disponibles à l'analyse (5). Cette démarche menée par la Cnamts constitue une première étape pour la mise en *open data* de bases de données de santé.

III. Objectif du travail

Le MeSH (*Medical Subject Headings*), thésaurus biomédical de référence, est un outil d'indexation, de catalogage et d'interrogation des bases de données de la NLM (*National Library of Medicine*, Bethesda, États-Unis), notamment MEDLINE/PubMed. Malgré l'utilisation croissante du terme « Big data » dans le domaine de la santé, il n'existe pas aujourd'hui de définition dans le MeSH de ce terme. Or une définition précise et sans ambiguïté est nécessaire pour une compréhension partagée du terme Big data. L'objectif de ce travail est de proposer une définition des Big data en santé au travers d'une revue de la littérature.

MATÉRIELS ET MÉTHODES

I. Stratégie de recherche

Pour cette revue de la littérature, nous avons conduit une recherche systématique sur la base de données PubMed de tous les articles publiés jusqu'au 9 mai 2014 en utilisant le terme de recherche « Big data ». Pour être exhaustif, nous n'avons pas défini de date de début. Nous avons utilisé l'équation de recherche suivante :

(big data[Title/Abstract]) AND ("1900/01/01"[Date - Publication]:"2014/05/09"[Date – Publication])

L'éligibilité de chaque article a été déterminée après lecture de son titre et de son résumé. Les articles ont été exclus s'ils ne concernaient pas directement le domaine de la santé ou si les Big data n'étaient pas le sujet de l'article.

L'accès au texte intégral des articles a nécessité l'utilisation de services de recherche en ligne : les accès gratuits au texte intégral de PMC (PubMed Central), Google, Google Scholar et le Service commun de documentation (SCD) de l'Université de Lille. Si l'article n'était pas disponible via ces services, le premier auteur de l'article a été contacté directement. Les articles non retrouvés en version intégrale ont été exclus. Les articles dans leur version intégrale ont ensuite été lus.

Les articles inclus dans l'analyse ont été classés en trois catégories selon qu'ils décrivent un jeu de données, qu'il s'agisse de dissertations, ou de revues de la littérature.

II. Collecte des données

Les informations suivantes ont été collectées pour chaque article : le titre, l'année de publication, le titre du journal, sa spécialité, le type d'article (articles décrivant un jeu de données, dissertation ou revue de la littérature), le domaine d'étude et les caractéristiques attribuées par les auteurs aux Big data et à la réutilisation des données (*data reuse*).

Pour les articles décrivait un jeu de données, le nombre d'individus statistiques (*n*) et le nombre de variables (*p*) ont également été relevés. Il convient de préciser que le nombre d'individus statistiques *n* ne décrit pas forcément des personnes physiques, mais peut correspondre également à d'autres éléments d'études comme des séquences de gènes par exemple. Le nombre de variables *p* peut être, par exemple, le nombre de propriétés physico-chimiques utilisées pour classer les acides aminés (24), les indicateurs de rendement utilisés pour évaluer un modèle de performance (25), ou bien encore le nombre de caractéristiques des demandes de remboursement de soins médicaux. Dans ce dernier cas, le nombre d'individus *n* est représenté par le nombre de demandes de remboursement de soins (26).

III. Analyse statistique et classification

Les analyses statistiques descriptives ont été réalisées avec le logiciel R (27). Dans ce travail, la notation « Log » correspond au logarithme décimal (c'est-à-dire le logarithme de base 10) et la notation « IC95 % » correspond à l'intervalle de confiance à 95 %. Les intervalles de confiance à 95 % des variables binaires ont été calculés en utilisant la loi binomiale.

A. Évolution du nombre d'articles publiés mentionnant le terme « Big data » en santé

L'analyse de l'évolution des publications sur les Big data en santé s'est basée sur deux représentations graphiques : l'une décrivant la publication annuelle des articles inclus dans notre revue de la littérature, l'autre représentant la publication annuelle des articles décrivant un jeu de données. Le nombre annuel de journaux scientifiques distincts concernés a aussi été étudié.

B. Évolution de la taille des Big data en santé

Pour les articles décrivant un jeu de données, l'effectif des données que les auteurs qualifient de « Big data » a été estimé par le logarithme décimal du produit du nombre d'individus statistiques (*n*) et du nombre de variables (*p*), c'est-à-dire *Log(n * p),* en fonction du temps.

C. Nombre d'individus (*n*) et de variables (*p*) dans chaque champ d'études

L'analyse des nombres *n* et *p* a reposé sur le calcul et l'estimation graphique de la densité de probabilité de *Log(n)*, *Log(p)* et *Log(n * p)* pour chaque champ d'étude. Enfin, la relation entre *Log(p)* et *Log(n)* a été représentée pour les différents champs d'étude.

IV. Caractéristiques des Big data

Les caractéristiques attribuées aux Big data par les auteurs ont été relevées au fur et à mesure de la lecture de tous les articles inclus dans notre analyse. Ces caractéristiques ont ensuite été classées en catégories. Les citations des auteurs des articles inclus ont été traduites de l'anglais pour ce travail.

V. Proposition de définition des Big data en santé

Une différence a été faite entre définition, propriétés et concepts reliés.

Il est difficile de trouver un consensus clair sur le mot « définition », tant en sciences qu'en philosophie. Étymologiquement, définir signifie : délimiter ce qui est de ce qui n'est pas (28). Selon le dictionnaire Le Petit Robert, une définition est l'action de préciser une idée. Il s'agit de la détermination précise et concrète des caractères distinctifs d'un être. Une définition correspond à l'ensemble des propriétés essentielles de quelque chose. Ainsi, tout élément qui valide la définition appartient à l'ensemble, et réciproquement tout élément de l'ensemble valide la définition.

Une propriété est un caractère distinctif qui appartient à quelque chose mais qui ne lui appartient pas toujours exclusivement. C'est une qualité particulière d'une chose ou d'une personne. Par conséquent, tout élément de l'ensemble valide la propriété, mais la réciproque est fausse : un élément qui valide la propriété ne fait pas nécessairement partie de l'ensemble.

Un concept relié est une représentation mentale générale et abstraite d'un objet. C'est une idée générale de quelque chose. Ainsi, aucune des deux relations décrites plus haut n'est toujours vraie, mais elles le sont parfois. Un élément qui valide le concept relié appartient parfois à l'ensemble, et réciproquement les éléments de l'ensemble valident souvent le concept relié. La différence entre définition, propriété et concept relié est illustrée dans la figure 9.

Ainsi, parmi les articles inclus, un jeu de données correspondant à la définition était qualifié de « Big data », et se trouvait alors avoir les propriétés proposées. Inversement, un jeu de données ayant quelques-unes ou toutes les propriétés listées n'était pas nécessairement qualifié de « Big data ». Enfin, les concepts reliés correspondaient à des propriétés qui n'étaient pas systématiquement associées aux Big data.

Sur la base des résultats de la revue de la littérature, une limite inférieure du volume des données qualifiées de « Big data » a été proposée. Ce seuil a résulté d'une discussion entre les auteurs de cette revue de la littérature et a tenu compte de la taille des jeux de données des articles inclus, mais aussi des caractéristiques attribuées aux Big data par les auteurs de l'ensemble des articles inclus.

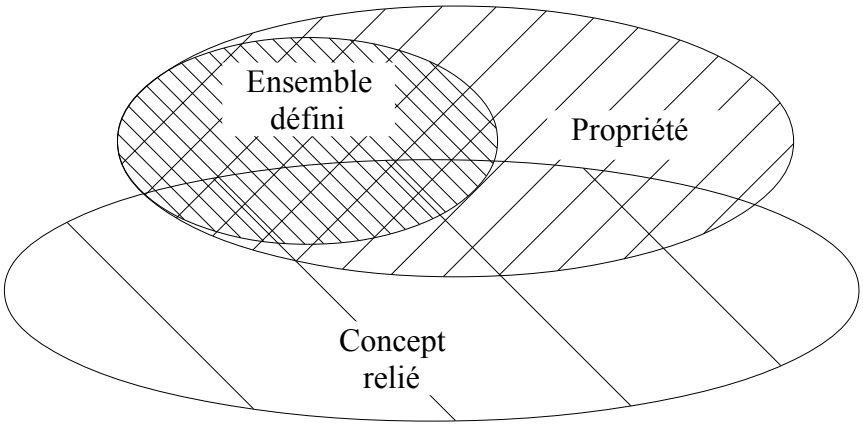

Figure 9 : Illustration de la différence entre définition, propriété et concept relié

RÉSULTATS

I. Stratégie de recherche

La recherche sur PubMed a permis d'identifier 330 articles. Après lecture des titres et des résumés, 94 articles ont été exclus. Un total de 236 articles ont été inclus pour lecture du texte intégral des articles. Dix-huit articles étaient indisponibles. Les versions intégrales des 218 articles restant ont été lues. Vingt-deux articles ont été exclus pour les raisons suivantes :

- articles non directement en lien avec la santé (18 articles)
- articles dont les Big data n'étaient pas le sujet de l'article (4 articles).

Parmi les 196 articles inclus, 48 décrivaient un jeu de données, 121 étaient des dissertations, et 27 des revues de la littérature. La figure 1 décrit les résultats obtenus à chaque étape de la recherche bibliographique.

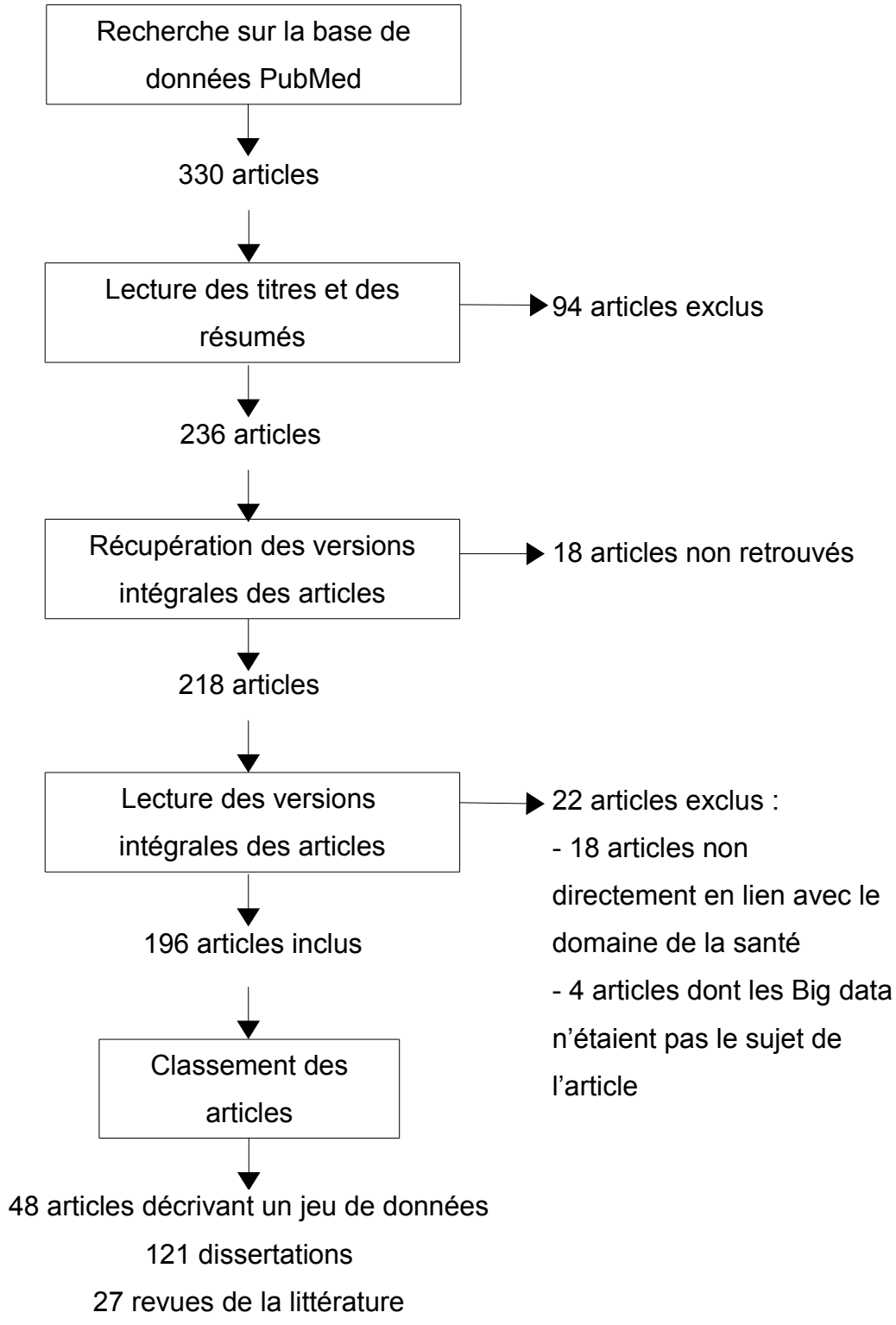

Figure 1 : Diagramme de flux de la revue de la littérature

II. Collecte des données

Le nombre d'article par domaine d'étude pour tous les articles inclus est présenté dans le tableau 1. Trois principales catégories ont été identifiées : les sciences « omiques »[2], les spécialités médicales et la santé publique. Le terme « omique » désigne les domaines de la biologie dont le nom se termine par « -omique », comme la génomique, la métabolomique et la protéomique.

La catégorie la plus représentée est la santé publique avec 83 articles (42 %, IC95 % =[35 ; 50]), suivis par les spécialités médicales avec 66 articles (33 %, IC95 % =[27 ; 41]), puis les sciences « Omiques » avec 44 articles (22 %, IC95 % =[17 ; 29]) et enfin les autres domaines non classés dans ces 3 catégories (éthique et philosophie) avec 3 articles (2 %, IC95 % =[0 ; 4]).

Le nombre d'articles par champs d'étude parmi les 48 articles décrivant un jeu de données est présenté dans le tableau 2. Parmi ces 48 articles, les spécialités « omiques » sont le principal domaine représenté avec 23 articles (48 %, IC95 % = [33 ; 63]), suivis par les spécialités médicales (endocrinologie, infectiologie, neurologie et imagerie médicale) avec 15 articles (31 %, IC95 % = [19 ; 46], puis la santé publique (bioinformatique, dossier patient informatisé, épidémiologie, pharmacovigilance et autre champ de la santé publique) avec 10 articles (21 %, IC95 % = [10 ; 35]).

2. Cf. définition en Annexe 2.

Tableau 1 : Nombre d'articles par domaine d'étude et spécialité pour l'ensemble des articles inclus

Domaine d'étude	Nombre d'articles	Proportion	Intervalle de confiance à 95 %
Santé publique			
Bioinformatique	36		
Dossier patient informatisé	6		
Épidémiologie	8		
Pharmacovigilance	1		
Autre santé publique	32		
Total santé publique	**83**	**42 %**	**[35 ; 50]**
Spécialités médicales			
Anesthésie – réanimation	6		
Biologie	14		
Dermatologie	1		
Endocrinologie	2		
Gastro-entérologie	1		
Imagerie médicale	8		
Immunologie	2		
Infectiologie	1		
Médecine du sport	1		
Néphrologie	1		
Neurologie	10		
Odontologie	1		
Oncologie	7		
Orthopédie	1		
Pédiatrie	2		
Pharmacologie	2		
Pneumologie	1		
Psychologie et psychiatrie	5		
Total spécialités médicales	**66**	**33 %**	**[27 ; 41]**
« Omiques »			
Génomique	36		
Métabolomique	1		

Protéomique	7		
Total « Omiques »	**44**	**22 %**	**[17 ; 29]**
Autres			
Éthique	2		
Philosophie	1		
Total « autres »	**3**	**2 %**	**[0 ; 4]**
Total	**196**		

Tableau 2 : Nombre d'articles par domaine d'étude et spécialité parmi les 48 articles décrivant un jeu de données

Domaine d'étude	Nombre d'articles	Proportion	Intervalle de confiance à 95 %
« Omiques »			
Génomique	18		
Métabolomique	1		
Protéomique	4		
Total « Omiques »	**23**	**48 %**	**[33 ; 63]**
Spécialités médicales			
Endocrinologie	2		
Imagerie médicale	3		
Immunologie	1		
Infectiologie	1		
Neurologie	8		
Total spécialités médicales	**15**	**31 %**	**[19 ; 46]**
Santé publique			
Bioinformatique	3		
Dossier patient informatisé	1		
Épidémiologie	2		
Pharmacovigilance	1		
Autre santé publique	3		
Total santé publique	**10**	**21 %**	**[10 ; 35]**
Total	**48**		

III. Analyse statistique et classification

A. Évolution du nombre d'articles publiés mentionnant le terme « Big data » en santé

La figure 2 montre l'évolution de la publication des articles mentionnant le terme « Big data » dans le domaine de la santé de 2003 à 2013. La publication annuelle d'articles est passée d'une publication en 2003 à 79 en 2013. De la même façon, on observe une augmentation de la publication annuelle d'articles portant sur un jeu de données (figure 3). Les 196 articles inclus dans notre revue de la littérature ont été publiés dans 34 journaux scientifiques différents. Parmi ces journaux, un seul journal a publié des articles en 2008, alors qu'ils étaient 68 en 2013.

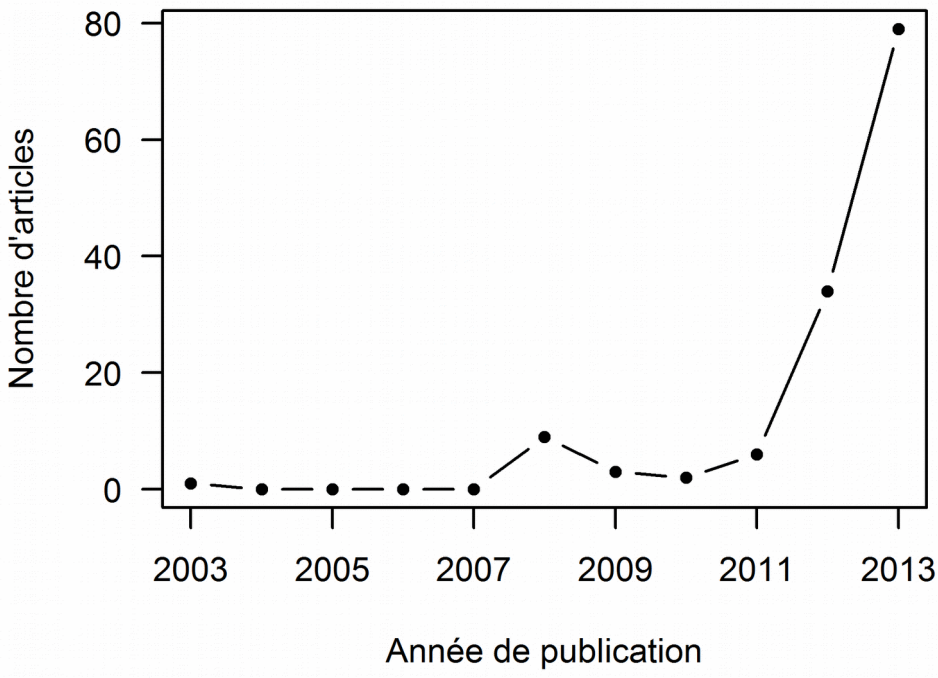

Figure 2 : Nombre d'articles mentionnant le terme « Big data » dans le domaine de la santé par an (années entières uniquement)

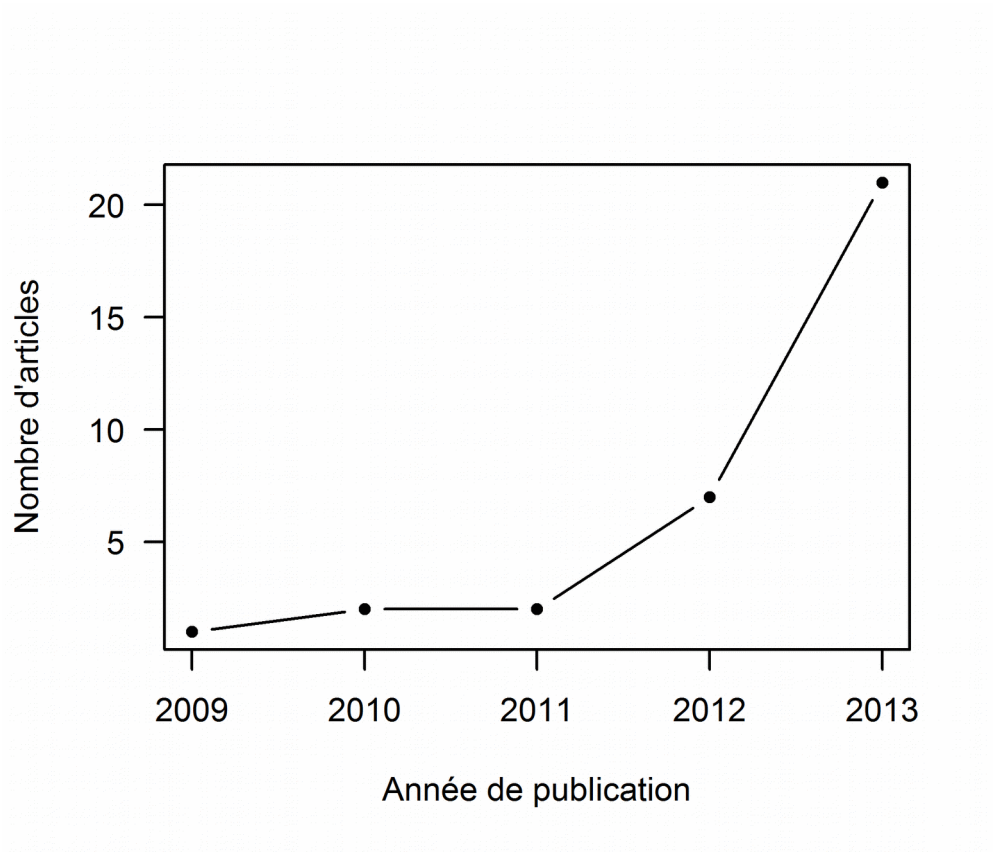

Figure 3 : Nombre d'articles portant sur un jeu de données mentionnant le terme « Big data » dans le domaine de la santé par an (années entières uniquement)

B. Évolution de la taille des Big data en santé

La figure 4 illustre le logarithme décimal du produit du nombre d'individus statistiques et du nombre de variables (*Log(n * p))* pour chaque année de publication des articles portant sur un jeu de données. On observe une augmentation non significative de 0,43 article par an (valeur-p = 0,34).

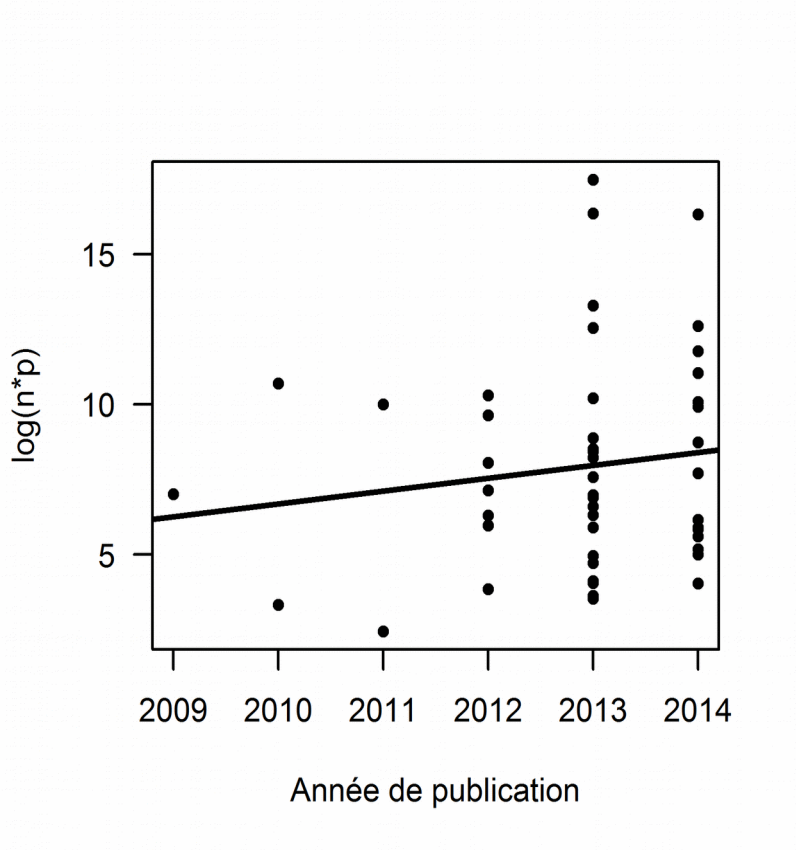

Figure 4 : *Log(n * p)* par année de publication. La ligne continue représente la régression linéaire (valeur-p = 0,34)

C. Nombre d'individus (*n*) et de variables (*p*) dans chaque champ d'études

Les figures 5, 6 et 7 représentent la densité de probabilité de *Log(n)*, *Log(p)* et *Log(n * p)*, respectivement pour les spécialités « omiques », les spécialités médicales, la santé publique et pour l'ensemble des articles. On peut remarquer que *Log(n * p)* est inférieur à 7 dans 23 études sur 48 (48 %, IC95 % =[33 ; 63]).

La figure 8 représente *Log(p)* en fonction de *Log(n)* pour les spécialités « omiques », les spécialités médicales et la santé publique. Cette figure suggère les différences suivantes entre les différentes catégories :

– les études dans le domaine « omique » concernent des données massives collectées sur un nombre relativement limité d'individus : *n* relativement petit par rapport à *p*.

– les études du domaine « santé publique » concernent un nombre important d'individus et un faible nombre de variables : grand *n*, petit *p*.

– les études du domaine « spécialités médicales » sont caractérisées par un nombre élevé d'individus et de variables : grand *n*, grand *p*.

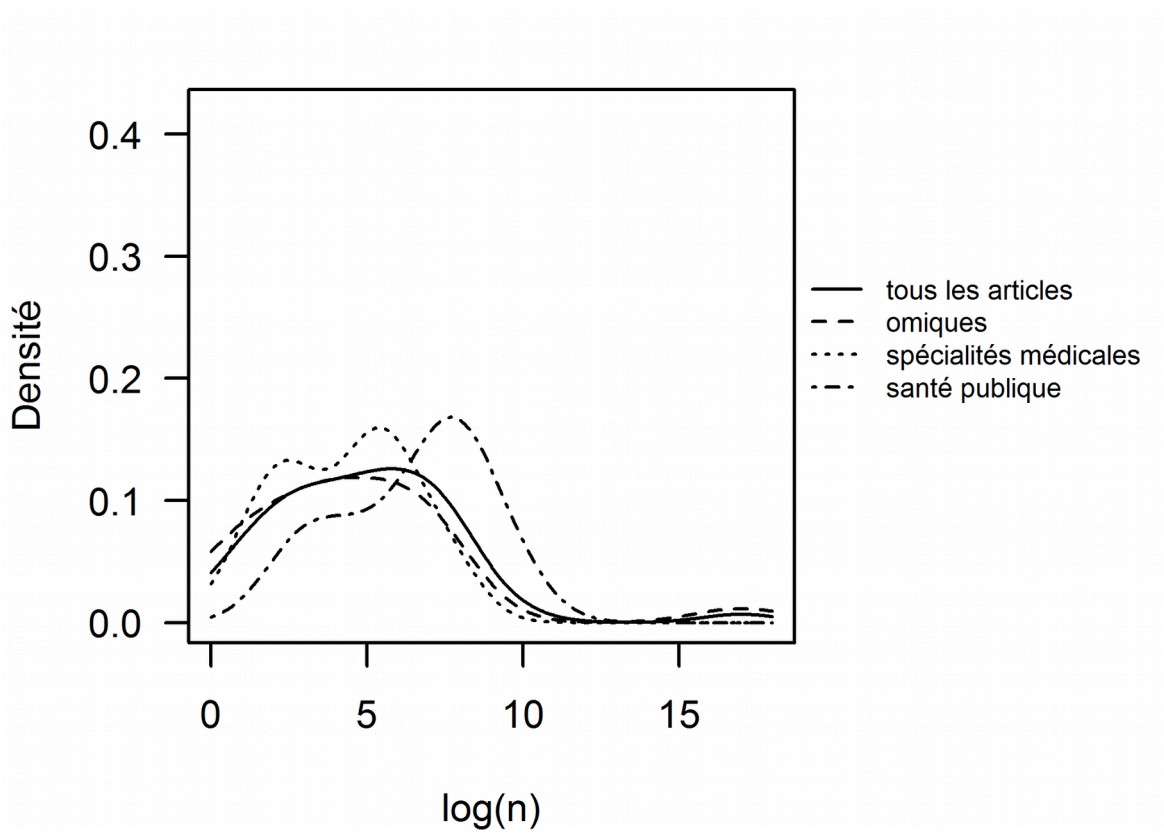

Figure 5 : Estimation graphique de la densité de probabilité de *Log(n)* pour les spécialités « omiques », les spécialités médicales, la santé publique et pour l'ensemble des articles

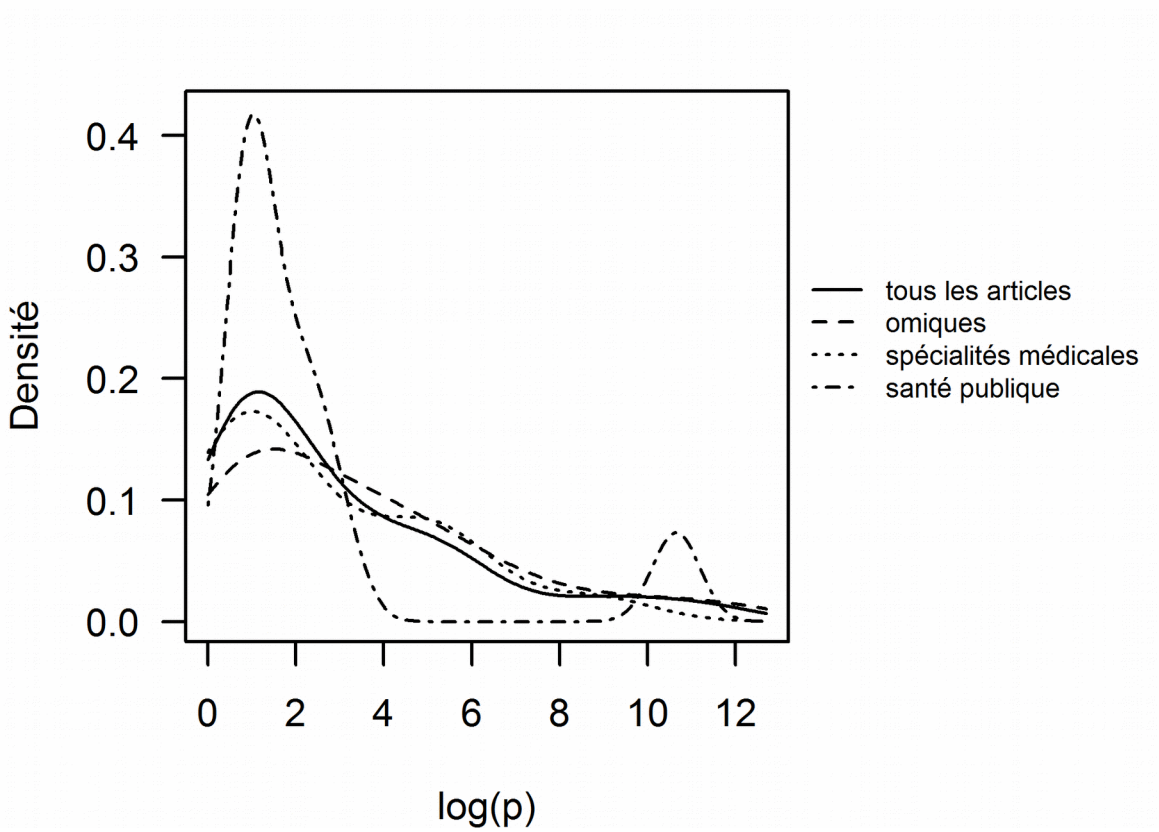

Figure 6 : Estimation graphique de la densité de probabilité de *Log(p)* pour les spécialités « omiques », les spécialités médicales, la santé publique et pour l'ensemble des articles

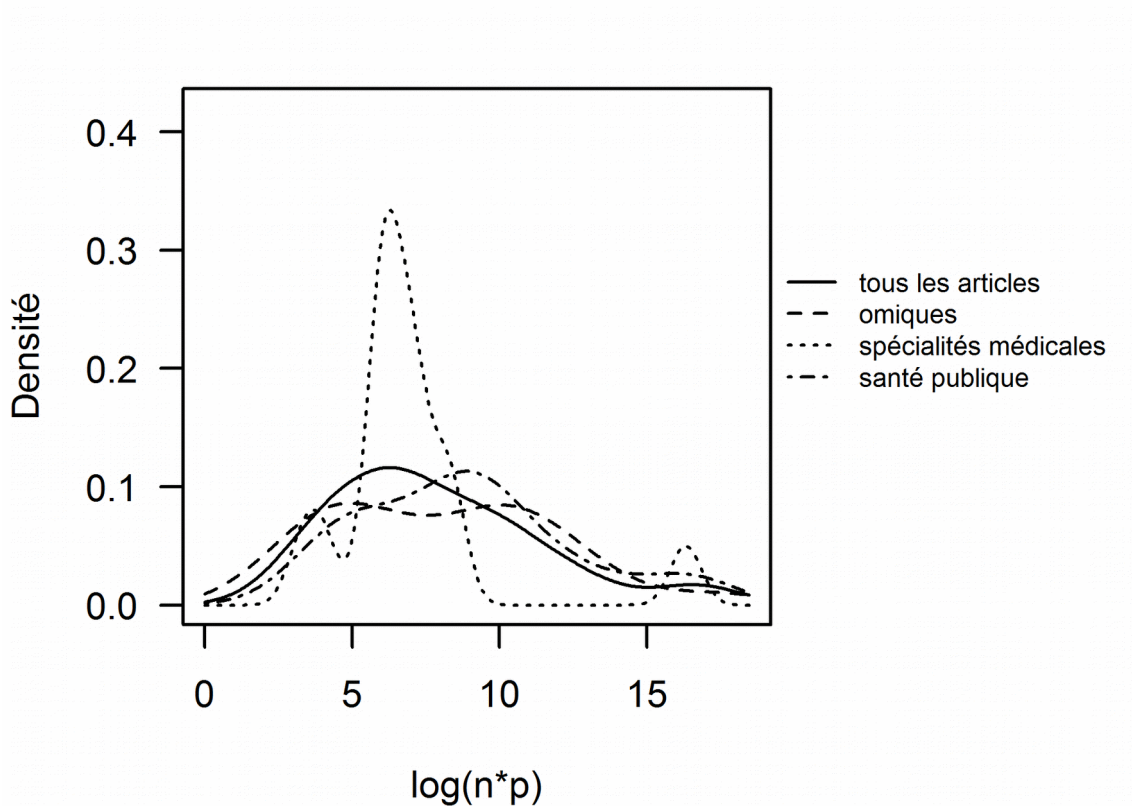

Figure 7 : Estimation graphique de la densité de probabilité en fonction de *Log(n * p)* pour les spécialités « omiques », les spécialités médicales, la santé publique et pour l'ensemble des articles

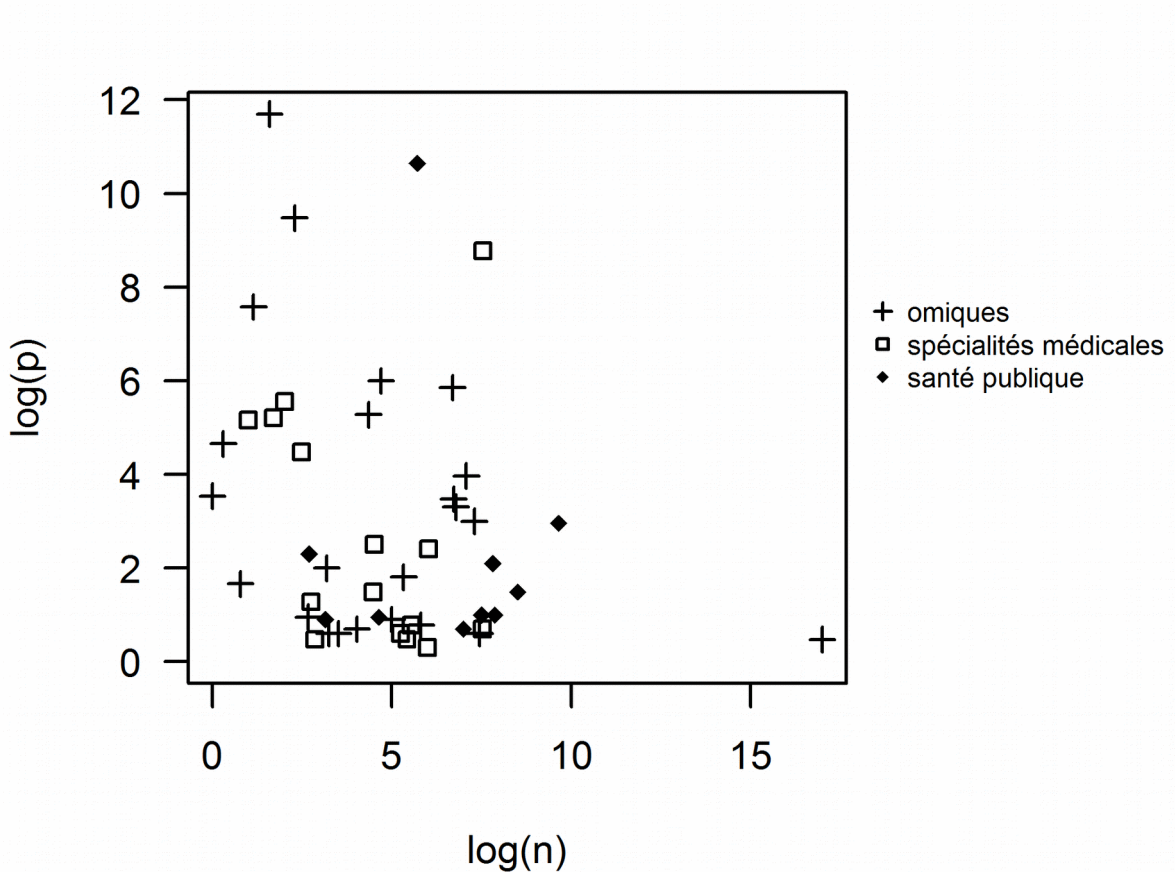

Figure 8 : *Log(p)* en fonction de *Log(n)* pour les spécialités « omiques », les spécialités médicales et la santé publique. Chaque pictogramme représente un article

D. Caractéristiques des Big data

Les principales caractéristiques concernant des Big data retrouvées dans les articles sont leur taille importante et leur complexité (26,29–37). Les Big data « ne concernent pas seulement l'ampleur et l'étendue des nouveaux jeux de données mais aussi leur complexité croissante » (35). Pour décrire la complexité des Big data, une approche largement utilisée est celle des trois « V » : le volume, la variété et la vélocité (26,38–45). « Les Big data sont un terme utilisé pour décrire les données dont le traitement est problématique du fait de leur taille (volume), de la fréquence de

leur mise à jour (vélocité), ou de leur diversité (variété) » (38). La véracité est un quatrième « V » parfois ajouté pour décrire un challenge posé par les Big data (37,43,46–48). Certains auteurs mentionnent un cinquième « V » : la valorisation (46,49).

1. Volume

Le volume est la principale caractéristique mentionnée par les auteurs (26,32,36,41,43,46,50,51). « Les notions de volume (ampleur et /ou étendue) (…) reconnues comme étant les caractéristiques des Big data » (41). « En ce qui concerne le volume, cela se traduit aujourd'hui en téraoctets (10^{12} octets), pétaoctets (10^{15} octets), ou exaoctets (10^{18} octets) » (26). « Le volume, de grandes quantités de données se multipliant rapidement et qui n'ont jamais été disponibles auparavant » (45). Certains auteurs mentionnent un seuil pour les Big data sans clairement le définir (26,52) « À quoi peut correspondre le « big » de Big data ? (…) la taille est un terme relatif quand il s'agit de données » (52). « Ces données sont incontestablement massives (de l'ordre de 10^{17}) » (41). Les données utilisées en épidémiologie (…) dépassent en fait à peine le seuil des big data » (26).

2. Variété

La variété est une autre caractéristique importante des Big data (26,45,46,50,51,53–55). En effet, les Big data proviennent de sources variées (43,56). La variété se traduit en « l'agrégation de données provenant de sources très diverses ou le regroupement de données provenant de sources indépendantes » (26). Les données non structurées, par exemple les données de texte libre (26,32,57) et les images (52,58–60), représentent un défi important. Dans le domaine de la santé, « les données prennent de nombreuses formes, y compris des nombres, des textes, des données codifiées, des graphiques, des images, des mesures physiologiques (signaux) et des sons. Les professionnels de santé se fient

à tous leurs sens, y compris l'odeur, pour recueillir des informations sur les individus » (32). Dans le domaine de la santé, « le nombre de données non structurées dépassent largement le nombre de données structurées » (54). « Le dossier patient informatisé génère des données massives posant le défi de la conversion de larges données non structurées en actions utiles pour les patients » (61). Les Big data « peuvent s'écarter des traditionnelles données structurées (organisées en lignes et colonnes) et peuvent être représentées en données semi-structurées tel que des fichiers de données XML, ou des données non structurées y compris les fichiers non hiérarchiques (« flat files ») qui ne sont pas compatibles avec les méthodes traditionnelles de données » (53). Ces données « ne sont pas assez structurées pour être analysées avec des techniques d'analyse traditionnelles de base de données relationnelles » (51).

De plus, les Big data peuvent être « volatiles, c'est-à-dire changeantes et disponibles uniquement pendant une durée limitée » (43).

3. Vélocité

L'augmentation rapide des données est une autre caractéristique des Big data (26,41,43,45,46,51,62). Il s'agit de « données en temps réel ou en temps quasi-réel » (45). « La vélocité correspond à la vitesse à laquelle les données d'aujourd'hui sont générées et traitées » (51).

4. Défi de véracité

La véracité des données vient ensuite : les Big data peuvent être difficile à valider (37,46–48). « Les Big data doivent être interprétées avec précaution, et dans leur contexte pour être utiles en médecine » (47). Les Big data ont une faible véracité et ne peuvent jamais « être exactes à 100 % » (48).

5. Défis à toutes les étapes de la gestion des données

Les Big data soulèvent des défis à toutes les étapes de la gestion des données : le recueil des données (52), la saisie (26,57,63–65), la collecte (40,66), le stockage (26,40,52,63,64,67–73), la gestion des données (40,63,65,74,75), le traitement des données (29,32,39,46,67,68,71,72,76,77), l'analyse des données (26,40,51–53,59,63–65,69–75,78–80) le partage des données et la publication des résultats (65). Les Big data « posent des difficultés dans la saisie des données, leur stockage, l'épuration des données, leur analyse, leur visualisation et leur partage » (63). Les Big data sont également difficiles à valoriser (46,49) : les Big data « ne sont pas simplement de grands volumes de données, elles se déplacent rapidement, sont difficiles à valider et à valoriser » (46).

6. Défi dans la conception d'outils de gestion des données

Trouver de nouvelles méthodes statistiques et informatiques de gestion est un autre défi soulevé par les Big data (53,63,70,71,79,81,82). Les Big data requièrent « un changement de perspective, d'infrastructures et de méthodes pour la collecte et l'analyse des données » (82). Des méthodes de visualisation permettant de comprendre les données doivent être créées (52,63,64,77). Pour que les Big data produisent du sens, « la création de nouveaux outils et méthodes de collecte, d'analyse, de visualisation et d'exploitation des données » (52) seront nécessaires.

7. Défi dans l'extraction d'informations utiles

Plusieurs auteurs soulignent l'importance d'extraire des informations utiles des Big data (50,64,83,84) et soulève la question de la façon dont on pourrait extraire du sens de ces données : les Big data créent « des défis sur la façon d'interpréter les données (…) pour en extraire des informations et des recommandations tout en supprimant le bruit et les données erronées » (39).

8. Défi pour faciliter l'accès aux données

De nombreux auteurs soulignent la nécessité d'identifier des moyens pour faciliter l'accès à l'information et leur partage (26,35,50,54,63–66,69,70,73,82,83,85–87). Il est nécessaire de promouvoir « la collaboration entre les scientifiques » (66). Ces auteurs considèrent que les données devraient être disponibles en *open source* pour mieux comparer les données.

9. Des experts humains en nombre insuffisant

Certains auteurs mentionnent le fait qu'il n'y a pas encore assez d'experts humains ayant à la fois des connaissances cliniques et en informatique et statistiques (50,88) : « Une sorte de personne hybride ayant des connaissances cliniques et mathématique. Nous n'avons pas encore assez de ces personnes en place » (50)

10. Réutilisation des données : « data reuse »

Certains auteurs mentionnent le fait que les Big data peuvent être des données qui sont recueillies de façon habituelle sans être utilisées immédiatement : « De gigantesques masses de données sont recueillies informatiquement en routine sans être analysées immédiatement, mais simplement parce que c'est à prix abordable. Ces données, on l'espère, répondront un jour à des questions, dont la plupart restent à être posées » (40). Les auteurs mettent en évidence le fait que les Big data sont souvent un usage secondaire des données, que nous pouvons appeler *data reuse* (la réutilisation de données) (34,40,41,61,85,89–92).

11. Possibilité de connaissances erronées

Certains auteurs soulignent le fait qu'extraire des connaissances des Big data peut mener à des faux résultats et à de fausses conclusions (93–95) : « Les résultats exploratoires émergeant des Big data sont tout aussi susceptibles d'être faux » (95). Nous ne pouvons extraire de la connaissance des Big data sans connaître le

contexte dans lequel les données ont été collectées : « La taille des données n'est pas suffisante en elle-même pour faire de l'épidémiologie » (94).

12. La question de la confidentialité des données

Une préoccupation mentionnée par plusieurs auteurs est la question de la confidentialité des données : « La facilité croissante avec laquelle les données peuvent être utilisées et réutilisées soulève la question de la protection des données personnelles et du consentement éclairé » (96). La capacité à « protéger les données personnelles à l'ère du Big data est devenue limitée » (59). Même si les données sont anonymisées, il reste un risque résiduel de ré-identification. En effet, l'identité des individus peut être retrouvée en manipulant les données grâce à des techniques de couplage de données (48,59,86,97). Le torrent de données soulève des questions éthiques (35). « L'utilisation répandue du dossier patient informatisé et la nécessité de partager les données pour mesurer la qualité des soins (…) met en lumière le problème de la protection des données concernant les patients » (86). « La capacité de découvrir des informations sur l'ADN à partir de sources non en lien avec l'ADN posent le problème de l'anonymisation des données au-delà du domaine de la confidentialité des données génotypiques et a ainsi des conséquences potentiellement importantes pour les règles de protection des données en recherche scientifique » (59).

E. Proposition de définition des Big data en santé

Une définition des Big data a été établie en se basant sur les résultats de notre revue de la littérature. Nous considérons que les Big data devraient exclusivement être définies par le volume. Il s'agit en effet de la principale caractéristique des Big data mentionnée par les auteurs. Nous proposons que des données peuvent être

considérées comme des Big data uniquement si le *Log(n * p)* est supérieur ou égal à 7. Les propriétés des Big data sont les suivantes :

– grande variété,

– grande vélocité,

– défi de véracité,

– défi à toutes les étapes de la gestion de données,

– défi dans la conception d'outils de gestion des données,

– défi dans l'extraction d'informations utiles,

– défi pour le partage des données,

– défi pour trouver des experts humains.

L'ensemble de ces termes qualifiant les propriétés des Big data ont été définis plus haut. Ces éléments ont été classés comme propriétés des Big data, car ils ne qualifient pas uniquement les Big data : ce sont des caractéristiques qui appartiennent aux Big data, sans toujours leur appartenir exclusivement.

Les concepts reliés aux Big data sont les suivants :

– la réutilisation des données (*data reuse*),

– la possibilité de connaissances erronées,

– la question de la confidentialité des données.

Ces éléments ont été classés comme concepts reliés, car il s'agit d'idées générales sur les Big data qui peuvent parfois qualifier les Big data. La définition, les propriétés et les concepts reliés aux Big data sont présentés dans le tableau 3.

Tableau 3 : Définition, propriétés et concepts reliés aux Big data en santé

Définition	Volume : *Log(n * p)* ≥ 7
Propriétés	Grande variété Grande vélocité Défi de véracité Défi à toutes les étapes de la gestion de données Défi dans la conception d'outils de gestion des données Défi dans l'extraction d'informations utiles Défi pour le partage des données Défi pour trouver des experts humains
Concepts reliés	La réutilisation des données (*data reuse*) La possibilité de connaissances erronées La question de la confidentialité des données

DISCUSSION

I. Résultats principaux

L'objectif de ce travail était de proposer une définition des Big data en santé au travers d'une revue de la littérature. La recherche documentaire a été réalisée sur PubMed et a permis de sélectionner 196 articles publiés jusqu'à mai 2014.

Les Big data sont caractérisées par le volume des données avec un seuil de *Log(n * p)* supérieur ou égal à 10^7, *n* représentant le nombre d'individus statistiques et *p* le nombre de variables.

Les propriétés des Big data sont les suivantes :

– la variété, notamment la faible part de données structurées dans le domaine de la santé,

– la vélocité, c'est-à-dire la vitesse d'accumulation et de traitement des données,

– un défi de véracité, qui est un des obstacles à surmonter,

– un défi à toutes les étapes de la gestion de données,

– un défi dans la conception d'outils de gestion des données : traiter de telles quantités de données nécessite de mettre au point de nouveaux outils d'exploitation de données,

– un défi dans l'extraction d'informations utiles,

– un défi pour le partage des données,

– un défi pour trouver des experts humains ayant à la fois des connaissances cliniques et en informatique et statistiques.

Les concepts reliés des Big data sont les suivants :

– la réutilisation des données (*data reuse*),

– la possibilité de connaissances erronées,

– la question de la confidentialité des données.

Nous avons défini la réutilisation des données (*data reuse*) comme concept relié des Big data, car nous pensons qu'il y a une confusion entre ces deux termes : le *data reuse* correspond à la réutilisation de données déjà recueillies dans le but de les valoriser alors que les Big data sont liées au volume des données collectées. En effet, les données peuvent être de taille importante sans être réutilisées : c'est le cas des spécialités « omiques » par exemple. Inversement, des données peuvent être réutilisées sans être forcément de volume important, comme l'utilisation secondaire des données des dossiers médicaux informatisés.

II. Points forts et points faibles

A. La revue de la littérature

Cette recherche systématique sur PubMed a permis de constituer un recensement relativement complet des articles mentionnant le terme « Big data » dans le domaine de la santé. Cependant, nous avons probablement omis les articles qui concernaient les Big data mais qui n'ont pas été inclus dans notre recherche, car le terme « Big data » n'apparaissait pas dans le résumé ou dans les mots-clés de l'article.

Par ailleurs, il faut également tenir compte de l'effet de mode des Big data. En effet, le terme « Big data » est désormais utilisé pour qualifier des données qui, autrefois, n'auraient pas été considérées comme des Big data.

Bien qu'il existe d'autres bases de données de littérature scientifique, seule la base de données PubMed a été interrogée. PubMed est en effet le principal moteur de recherche de données bibliographiques des domaines de la biologie et de la médecine et demeure la base de données de référence pour les sciences biomédicales (98).

B. La proposition de définition

Comme il n'y a actuellement pas de définition des Big data en santé, les résultats de la revue de la littérature peuvent être eux-mêmes erronés. C'est une limite de cette approche inductive caractérisée par le fait que nous avons utilisé des observations pour construire une définition. Le problème de la définition du seuil du volume des Big data illustre bien cette difficulté : le seuil de 10^7 peut paraître en désaccord avec les résultats de la figure 7. En effet, cette définition des Big data est seulement le résultat d'une discussion entre les auteurs de cette revue de la littérature. Elle a été décidée sur la base des résultats du nombre d'individus et de variables observés dans les études décrivant un jeu de données, mais aussi en tenant compte des caractéristiques des Big data mentionnées par les auteurs de tous les articles inclus dans la revue de la littérature. Ainsi, par exemple, les auteurs sont unanimes concernant les difficultés de traitement des Big data. Il aurait donc été difficile d'admettre des $Log(n * p)$ supérieurs ou égaux à 6, ce qui concerne pourtant 35 % des articles dans notre recherche documentaire. En effet, cet ordre de grandeur est facile à traiter de nos jours, même avec un simple tableur sur un ordinateur de bureau. Cependant, ce seuil choisi de 10^7 suppose que la moitié des

articles décrivant une base de données dans cette revue de la littérature appellent à tort leurs données « Big data ».

Il est donc difficile d'indiquer à partir de quel volume de données on se trouve dans une situation de type Big data. De plus, la taille des données qui les qualifie de « Big data » va probablement continuer d'augmenter. En effet, les infrastructures informatiques et les méthodes de traitement nécessaires pour analyser ces données massives vont probablement s'améliorer avec le temps.

III. Perspectives

A. Recherche translationnelle

Les Big data présentent de nombreuses opportunités pour la « recherche translationnelle[3] » (*translational research*), c'est-à-dire l'application des conclusions de la recherche fondamentale à la pratique clinique. L'informatique sera la clé du succès de la recherche translationnelle (99). Comme Shah l'a affirmé, « l'informatique translationnelle est sur le point de révolutionner la santé humaine et les soins en utilisant des mesures à grande échelle sur les individus. Des approches centrées sur des données massives vont permettre de découvrir des tendances et de faire des prévisions cliniques pertinentes » (100). Selon Chen, le *cloud computing* pourrait être un outil pour faciliter la recherche en informatique translationnelle (87). L'informatique est nécessaire pour exploiter pleinement le potentiel des données en santé. L'émergence de nouveaux outils devrait pouvoir transformer les données de santé en connaissances, et ainsi apporter un bénéfice pour les patients.

3. Cf. définition en Annexe 2.

B. Data scientist : un métier d'avenir ?

Le *Data scientist* présente une triple compétence opérationnelle : la maîtrise des statistiques et notamment des techniques de « data mining[4] » (exploration de données), la maîtrise des outils informatiques de gestion de bases de données, et un savoir-faire dans le secteur d'application des données analysées (en médecine clinique par exemple) (55).

C. Autres défis des Big data en santé

Les avantages attendus des Big data pour améliorer le pilotage des systèmes de santé et la mise en œuvre d'une médecine de plus en plus ciblée soulèvent néanmoins un certain nombre de questions. Comment déterminer qui peut ou doit avoir accès à ces données de santé et dans quelles conditions ? Comment conjuguer intérêt collectif et protection des personnes ? Comment se prémunir contre toute utilisation commerciale, voire frauduleuse de ces données ?

1. Confidentialité des données

a) Sécurité des systèmes de stockage et de traitement

Le développement de l'analyse des Big data doit s'accompagner d'un questionnement relatif à la protection des données. Une fois acquises, les Big data sont stockées dans des centres de traitement de données (*data centers*[5]). Ceux-ci se sont imposés avec le développement du *cloud computing* et sont actuellement localisés en grande majorité aux États-Unis ou au Royaume-Uni. Cette réalité matérielle pose un problème particulier pour les données liées à la santé, domaine où les exigences de sécurité sont essentielles (5). La gestion de ces données demeure donc un challenge significatif, non seulement du fait des volumétries, mais

4. Cf. définition en Annexe 2.
5. Cf. définition en Annexe 2.

également de la nature des informations médicales, par définition sensibles et propriété du patient et devant donc être particulièrement protégées (22).

b) Protection de la vie privée

Grâce aux algorithmes, on peut désormais combiner des bases de données pour dresser le portrait d'un individu et prédire son comportement et ses besoins futurs. Un combat est mené par des associations de consommateurs pour que les individus puissent avoir le contrôle sur leurs propres informations. Il est de la responsabilité des scientifiques d'aider les citoyens et les législateurs qui s'interrogent sur les limites à mettre. Ceux qui élaborent des modèles d'utilisation des données doivent montrer scientifiquement l'intérêt de le faire, les difficultés qui se présentent pour l'anonymisation de ces données et quantifier les dangers auxquels on s'expose en partageant des données (6).

En France, l'usage des données à caractère personnel est réglementé par la loi "Informatique et libertés" (101). Cette loi stipule que les données personnelles doivent être collectées et traitées pour des finalités déterminées, explicites et légitimes. Seules les données pertinentes pour un usage défini peuvent donc être collectées. Leur durée de conservation ne doit pas excéder le temps nécessaire à l'atteinte des objectifs pour lesquels elles sont collectées. Passé ce délai, prévaut le « droit à l'oubli » ou l'obligation de destruction des données. La loi s'applique également si les données ne sont pas enregistrées mais traitées en temps réel.

Avec l'analyse des Big data, il est cependant difficile d'anticiper quel usage il en sera fait. La collecte ciblée et le principe de suppression entrent par ailleurs en contradiction avec la nécessité d'un volume de données le plus important possible (102).

2. Risques de mauvaise utilisation des données personnelles

Il existe également un risque de mauvaise utilisation des données personnelles (12). En permettant de mieux anticiper les comportements, mais aussi l'apparition de maladies associées à des profils génétiques, les Big data pourraient aussi être utilisées par les services de santé ou les compagnies d'assurance pour refuser des traitements ou des clients, ou encore surveiller les comportements des assurés (103).

3. Validité scientifique et dimension humaine

La valeur scientifique de données traitées informatiquement peut être mise en question, car elles peuvent entraîner des erreurs d'interprétation. L'application des Big data dans le domaine de la santé pose également la question de la dimension humaine d'une telle démarche. Les médecins pourront légitimement se montrer méfiants à l'égard des Big data et du rôle qu'elles sont amenées à jouer dans la décision médicale. Ils seront en droit d'exiger que la question suivante soit abordée : la rigueur algorithmique de l'ordinateur va-t-elle se substituer à l'expérience et à la sensibilité humaine ? (104)

D. Fossé entre potentiel et réalisation

On observe un enthousiasme grandissant pour les Big data. La promesse est réelle, mais il y a actuellement un fossé entre ses potentiels et ses réalisations.

Le terme « Big data » circule aujourd'hui comme un mantra. Sami Coll, sociologue à l'Université de Lausanne, auteur d'un livre consacré aux cartes de fidélité dans la grande distribution, souligne le fait suivant : « On affirme qu'en maniant des masses de données, on va redresser l'économie, prévenir des catastrophes, traiter des pathologies. On entre désormais dans une logique du Big data pour tout, qui relève clairement de la foi » (105). Hervé Fischer, professeur à

l'Université du Québec à Montréal, philosophe, observateur de la civilisation numérique, fait le même constat : « Il y a dans nos esprits beaucoup plus de magie qu'il n'y en a jamais eu. On se dit qu'avec le numérique, on va développer la démocratie dans le monde entier, qu'on va supprimer tous les fléaux. Nous vivons un moment d'exaltation de notre tendance primitive à la pensée magique » (106). L'auteur tente de mettre en évidence les mythes et l'effervescence magique de l'âge du numérique (107). Dans l'article intitulé *Critical Questions for Big Data*, les auteurs mentionnent la part de mythologie dans le phénomène Big data, en évoquant une « large croyance selon laquelle de grands ensembles de données promettent une forme supérieure d'intelligence et de connaissance, susceptible d'offrir un tout nouveau regard empreint de vérité, d'objectivité et de précision » (108).

CONCLUSION

Cette étude a permis de se pencher sur la question de la définition des Big data dans le domaine de la santé. Les Big data correspondent à de grands volumes de données structurées ou non structurées, difficilement gérables avec des solutions classiques de stockage et de traitement. Ces données proviennent de sources variées et peuvent être produites en temps réel.

Les données sont des gisements potentiels de valeur. Elles sont porteuses de sens et de connaissance. On s'interroge sur la façon de traiter ces données afin d'en tirer des informations utiles, pour en extraire des connaissances. Ces connaissances sont transformées en actions qui visent à apporter des bénéfices pour les patients.

Données → Informations → Connaissances → Actions → Bénéfices pour le patient

Les Big data sont à l'ordre du jour de nombreuses publications mais les technologies permettant d'analyser ces données sont encore naissantes, fortement évolutives. Le nombre de données continue à croître et les outils d'analyse vont se perfectionner.

Néanmoins, les utilisations potentielles des Big data sont restées jusqu'à présent des possibilités théoriques en raison de multiples barrières, notamment la confidentialité des données et la question de la propriété des données.

RÉFÉRENCES BIBLIOGRAPHIQUES

1. Bell D. The Coming Of Post-industrial Society. New York: Basic Books; 1973.

2. Cox M, Ellsworth D. Managing big data for scientific visualization. ACM Siggraph. 1997;21.

3. Brasseur C. Enjeux et usages du Big Data : technologies, méthodes et mise en œuvre. Paris: Lavoisier; 2013.

4. Siegel E. Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die. Hoboken (NJ): John Wiley & Sons; 2013.

5. Hamel M-P, Marguerit D. Analyse des big data. Quels usages, quels défis ? Commis Général À Strat À Prospect [Internet]. 2013 Nov [cited 2015 Mar 16]. Available from: http://www.strategie.gouv.fr/publications/analyse-big-data-usages-defis

6. Blondel V. Big Data. Paris: La Recherche. 2013 Dec;482:28-30.

7. Zimmerman AS. New Knowledge from Old Data: The Role of Standards in the Sharing and Reuse of Ecological Data. Sci Technol Hum Values [Internet]. 2008 Feb 5 [cited 2015 Mar 16]. Available from: http://sth.sagepub.com/content/early/2008/02/05/0162243907306704

8. Neuvial P. Vers une médecine personnalisée grâce à la recherche en génomique. Variances. 2013 Oct;48:31–33.

9. Watson JD, Watson, Gilman M. Recombinant DNA. 2nd ed. New York: W.H. Freeman & Company; 1992.

10. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature. 2001 Feb 15;409(6822):928–33.

11. Jordan B. Séquençage de nouvelle génération : déjà en clinique ? Médecine/sciences. 2011;27:1127–30.

12. Pentland A, Reid TG, Heibeck T. Revolutionizing Medicine and Public Health. World Innovation Summit for Health [Internet]. 2013 Feb 5 [cited 2015 Mar 16]. Available from: http://kit.mit.edu/sites/default/files/documents/WISH_BigData_Report.pdf

13. Freifeld CC, Mandl KD, Reis BY, Brownstein JS. HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. J Am Med Inform Assoc JAMIA. 2008 Apr;15(2):150–7.

14. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009 Feb 19;457(7232):1012–4.

15. HAS. Etude SADM, Systèmes informatiques d'Aide à la Décision Médicale [Internet]. 2010 Jul [cited 2015 Feb 4]. Available from: http://www.has-sante.fr/portail/jcms/c_1021245/fr/systemes-informatiques-d-aide-a-la-decision-medicale

16. Projet Européen PSIP. CHRU de Lille - Dossier de presse [Internet]. 2011 [cited 2015 Feb 4]. Available from: http://www.chru-lille.fr/~gapicms/fichiers/678/communiques/projet%20psip%20dossier.pdf

17. Dupays S, Natali J-P. L'évolution des volumes d'activité des établissements de santé : description, déterminants et prévision. Inspection générale des affaires sociales; 2013 Sep p. 92.

18. Guidoni D. Le BIG DATA et la RFID au service des soins hospitaliers. Kurt Salmon; [Internet]. 2014 [cited 2015 Feb 4]. Available from: http://www.kurtsalmon.com/fr-fr/dt/vertical-insight/1163/Le-BIG-DATA-et-la-RFID-au-service-des-soins-hospitaliers

19. The National Cancer Registration Service – Eastern Office [Internet]. [cited 2015 Feb 4]. Available from: http://ecric.nhs.uk/

20. CATCH Health [Internet]. [cited 2015 Feb 4]. Available from: http://catch-health.org/Home_Page.html

21. Hospitals prevent relapse with predictive analytics | SAS Media Releases [Internet]. [cited 2015 Feb 4]. Available from: https://www.sas.com/offices/asiapacific/singapore/news/KTPH/Analytics.html

22. Hermelin P, Bourdoncle F. La feuille de route Big Data. Ministère L'économie Ind Numér. 2014 Jul;40.

23. Data.gouv.fr [Internet]. [cited 2015 Feb 5]. Available from: https://www.data.gouv.fr/fr/

24. Zou Q, Wang Z, Guan X, Liu B, Wu Y, Lin Z. An approach for identifying cytokines based on a novel ensemble classifier. BioMed Res Int. 2013;2013:686090.

25. Zhao L, Wong L, Lu L, Hoi SCH, Li J. B-cell epitope prediction through a graph model. BMC Bioinformatics. 2012;13 Suppl 17:S20.

26. Berger ML, Doban V. Big data, advanced analytics and the future of comparative effectiveness research. J Comp Eff Res. 2014 Mar;3(2):167–76.

27. R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing [Internet]. 2014 [cited 2015 Feb 4]. Available from: http://www.R-project.org/

28. Ouvrier-Buffet C. Construction de définitions / construction de concept : vers une situation fondamentale pour la construction de définitions en mathématiques [These]. Grenoble I: Université Joseph Fourier; 2003

29. Mallon WJ. Big data. J Shoulder Elb Surg Am Shoulder Elb Surg Al. 2013

Sep;22(9):1153.

30. Salcido RS. Big data and disruptive innovation in wound care. Adv Skin Wound Care. 2013 Aug;26(8):344.

31. Ketchersid T. Big data in nephrology: friend or foe? Blood Purif. 2013;36(3-4):160–4.

32. Hovenga EJS, Grain H. Health data and data governance. Stud Health Technol Inform. 2013;193:67–92.

33. Müller H, Hanbury A, Al Shorbaji N. Health information search to deal with the exploding amount of health information produced. Methods Inf Med. 2012 Dec 4;51(6):516–8.

34. Porche DJ. Men's Health Big Data. Am J Mens Health. 2014 May;8(3):189.

35. Callebaut W. Scientific perspectivism: A philosopher of science's response to the challenge of big data biology. Stud Hist Philos Biol Biomed Sci. 2012 Mar;43(1):69–80.

36. Fan J, Liu H. Statistical analysis of big data on pharmacogenomics. Adv Drug Deliv Rev. 2013 Jun 30;65(7):987–1000.

37. Lupşe O-S, Crişan-Vida M, Stoicu-Tivadar L, Bernard E. Supporting diagnosis and treatment in medical care based on big data processing. Stud Health Technol Inform. 2014;197:65–9.

38. Hay SI, George DB, Moyes CL, Brownstein JS. Big data opportunities for global infectious disease surveillance. PLoS Med. 2013;10(4):e1001413.

39. Hamilton B. Impacts of big data. Potential is huge, so are challenges. Health Manag Technol. 2013 Aug;34(8):12–3.

40. Markowetz A, Błaszkiewicz K, Montag C, Switala C, Schlaepfer TE. Psycho-Informatics: Big Data shaping modern psychometrics. Med Hypotheses. 2014 Apr;82(4):405–11.

41. Chute CG, Ullman-Cullere M, Wood GM, Lin SM, He M, Pathak J. Some experiences and opportunities for big data in translational research. Genet Med Off J Am Coll Med Genet. 2013 Oct;15(10):802–9.

42. Kao RR, Haydon DT, Lycett SJ, Murcia PR. Supersize me: how whole-genome sequencing and big data are transforming epidemiology. Trends Microbiol. 2014 Mar 21.

43. Ola O, Sedig K. The challenge of big data in public health: an opportunity for visual analytics. Online J Public Health Inform. 2014;5(3):223.

44. Gardner E. The HIT approach to big data. Health Data Manag. 2013 Mar;21(3):34, 36, 38 passim.

45. Moore KD, Eyestone K, Coddington DC. The big deal about big data. Healthc Financ Manag J Healthc Financ Manag Assoc. 2013 Aug;67(8):60–6, 68.

46. Dereli T, Coşkun Y, Kolker E, Güner O, Ağırbaşlı M, Ozdemir V. Big data and ethics review for health systems research in LMICs: understanding risk, uncertainty and

ignorance-and catching the black swans? Am J Bioeth AJOB. 2014 Feb;14(2):48–50.

47. Litman RS. Complications of laryngeal masks in children: big data comes to pediatric anesthesia. Anesthesiology. 2013 Dec;119(6):1239–40.

48. Ward JC. Oncology reimbursement in the era of personalized medicine and big data. J Oncol Pract Am Soc Clin Oncol. 2014 Mar 1;10(2):83–6.

49. Özdemir V, Badr KF, Dove ES, Endrenyi L, Geraci CJ, Hotez PJ, et al. Crowd-funded micro-grants for genomics and "big data": an actionable idea connecting small (artisan) science, infrastructure science, and citizen philanthropy. Omics J Integr Biol. 2013 Apr;17(4):161–72.

50. Harnessing big data. How to achieve value. Hosp Health Netw AHA. 2014 Feb;88(2):61–71.

51. Jee K, Kim G-H. Potentiality of big data in the medical sector: focus on how to reshape the healthcare system. Healthc Inform Res. 2013 Jun;19(2):79–85.

52. Van Horn JD, Toga AW. Human neuroimaging as a "Big Data" science. Brain Imaging Behav. 2013 Oct 10.

53. O'Driscoll A, Daugelaite J, Sleator RD. "Big data", Hadoop and cloud computing in genomics. J Biomed Inform. 2013 Oct;46(5):774–81.

54. Buyer's brief: cognitive computing in the age of big data. Healthc Financ Manag J Healthc Financ Manag Assoc. 2014 Apr;68(4):35–6.

55. Davenport TH, Patil DJ. Data scientist: the sexiest job of the 21st century. Harv Bus Rev. 2012 Oct;90(10):70–6, 128.

56. Khoury MJ, Lam TK, Ioannidis JPA, Hartge P, Spitz MR, Buring JE, et al. Transforming epidemiology for 21st century medicine and public health. Cancer Epidemiol Biomark Prev Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol. 2013 Apr;22(4):508–16.

57. Bonney S. HIM's role in managing big data: Turning data collected by an EHR into information. J AHIMA Am Health Inf Manag Assoc. 2013 Sep;84(9):62–4.

58. Jayapandian CP, Chen C-H, Bozorgi A, Lhatoo SD, Zhang G-Q, Sahoo SS. Cloudwave: distributed processing of "big data" from electrophysiological recordings for epilepsy clinical research using hadoop. AMIA Annu Symp Proc AMIA Symp AMIA Symp. 2013;2013:691–700.

59. Schadt EE. The changing privacy landscape in the era of big data. Mol Syst Biol. 2012;8:612.

60. Aji A, Wang F, Saltz JH. Towards Building a High Performance Spatial Query System for Large Scale Medical Imaging Data. Proc ACM SIGSPATIAL Int Conf Adv Inf. 2012 Nov 6;2012:309–18.

61. Matheson GO, Klügl M, Engebretsen L, Bendiksen F, Blair SN, Börjesson M, et al. Prevention and management of noncommunicable disease: the IOC Consensus Statement,

Lausanne 2013. Clin J Sport Med Off J Can Acad Sport Med. 2013 Nov;23(6):419–29.

62. Afendi FM, Ono N, Nakamura Y, Nakamura K, Darusman LK, Kibinge N, et al. Data Mining Methods for Omics and Knowledge of Crude Medicinal Plants toward Big Data Biology. Comput Struct Biotechnol J. 2013;4:e201301010.

63. Mohr DC, Burns MN, Schueller SM, Clarke G, Klinkman M. Behavioral intervention technologies: evidence review and recommendations for future research in mental health. Gen Hosp Psychiatry. 2013 Aug;35(4):332–8.

64. Ansermino JM. From the Journal archives: Improving patient outcomes in the era of Big Data. Can J Anaesth J Can Anesth. 2014 Mar 29.

65. Klingström T, Soldatova L, Stevens R, Roos TE, Swertz MA, Müller KM, et al. Workshop on laboratory protocol standards for the Molecular Methods Database. New Biotechnol. 2013 Jan 25;30(2):109–13.

66. Mervis J. U.S. science policy. Agencies rally to tackle big data. Science. 2012 Apr 6;336(6077):22.

67. Mohammed Y, Mostovenko E, Henneman AA, Marissen RJ, Deelder AM, Palmblad M. Cloud parallel processing of tandem mass spectrometry based proteomics data. J Proteome Res. 2012 Oct 5;11(10):5101–8.

68. Karlsson J, Trelles O. MAPI: a software framework for distributed biomedical applications. J Biomed Semant. 2013;4(1):4.

69. Bower MR, Stead M, Brinkmann BH, Dufendach K, Worrell GA. Metadata and annotations for multi-scale electrophysiological data. Conf Proc Annu Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Conf. 2009;2009:2811–4.

70. Ranganathan S, Schönbach C, Kelso J, Rost B, Nathan S, Tan TW. Towards big data science in the decade ahead from ten years of InCoB and the 1st ISCB-Asia Joint Conference. BMC Bioinformatics. 2011 Nov 30;12 Suppl 13:S1.

71. DiLeo MV, Strahan GD, den Bakker M, Hoekenga OA. Weighted correlation network analysis (WGCNA) applied to the tomato fruit metabolome. PloS One. 2011;6(10):e26683.

72. Greene CS, Tan J, Ung M, Moore JH, Cheng C. Big Data Bioinformatics. J Cell Physiol. 2014 May 6.

73. Dai L, Gao X, Guo Y, Xiao J, Zhang Z. Bioinformatics clouds for big data manipulation. Biol Direct. 2012;7:43; discussion 43.

74. Maclean D, Kamoun S. Big data in small places. Nat Biotechnol. 2012 Jan;30(1):33–4.

75. Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA J Am Med Assoc. 2013 Apr 3;309(13):1351–2.

76. Marx V. Biology: The big challenges of big data. Nature. 2013 Jun 13;498(7453):255–60.

77. Schadt EE, Linderman MD, Sorenson J, Lee L, Nolan GP. Computational solutions to

large-scale data management and analysis. Nat Rev Genet. 2010 Sep;11(9):647–57.

78. Cole JB, Newman S, Foertter F, Aguilar I, Coffey M. Breeding and Genetics Symposium: really big data: processing and analysis of very large data sets. J Anim Sci. 2012 Mar;90(3):723–33.

79. Finding correlations in big data. Nat Biotechnol. 2012 Apr;30(4):334–5.

80. Kolker E, Stewart E, Ozdemir V. Opportunities and challenges for the life sciences community. Omics J Integr Biol. 2012 Mar;16(3):138–47.

81. Troiano RP, McClain JJ, Brychta RJ, Chen KY. Evolution of accelerometer methods for physical activity research. Br J Sports Med. 2014 Apr 29.

82. Feldmann E, Liebeskind DS. Developing Precision Stroke Imaging. Front Neurol. 2014;5:29.

83. Green DE, Rapp EJ. Can big data lead us to big savings? Radiogr Rev Publ Radiol Soc N Am Inc. 2013 May;33(3):859–60.

84. Huberman BA. Sociology of science: Big data deserve a bigger audience. Nature. 2012 Feb 16;482(7385):308.

85. Lynch C. Big data: How do your data grow? Nature. 2008 Sep 4;455(7209):28–9.

86. White SE. De-identification and the sharing of big data. J AHIMA Am Health Inf Manag Assoc. 2013 Apr;84(4):44–7.

87. Chen J, Qian F, Yan W, Shen B. Translational biomedical informatics in the cloud: present and future. BioMed Res Int. 2013;2013:658925.

88. Mavandadi S, Dimitrov S, Feng S, Yu F, Yu R, Sikora U, et al. Crowd-sourced BioGames: managing the big data problem for next-generation lab-on-a-chip platforms. Lab Chip. 2012 Oct 21;12(20):4102–6.

89. Maps, “Big Data,”and Case Reports. Glob Adv Health Med Improv Healthc Outcomes Worldw. 2012 Jul;1(3):5–7.

90. Hoffman S, Podgurski A. Big bad data: law, public health, and biomedical databases. J Law Med Ethics J Am Soc Law Med Ethics. 2013 Mar;41 Suppl 1:56–60.

91. Cockfield J, Su K, Robbins KA. MOBBED: a computational data infrastructure for handling large collections of event-rich time series datasets in MATLAB. Front Neuroinformatics. 2013;7:20.

92. Martin SF, Falkenberg H, Dyrlund TF, Khoudoli GA, Mageean CJ, Linding R. PROTEINCHALLENGE: crowd sourcing in proteomics analysis and software development. J Proteomics. 2013 Aug 2;88:41–6.

93. Lindenmayer DB, Likens GE. Analysis: don’t do big-data science backwards. Nature. 2013 Jul 18;499(7458):284.

94. Toh S, Platt R. Big data in epidemiology: too big to fail? Epidemiol Camb Mass. 2013

Nov;24(6):939.

95. Castellanos FX, Di Martino A, Craddock RC, Mehta AD, Milham MP. Clinical applications of the functional connectome. NeuroImage. 2013 Oct 15;80:527–40.

96. Currie J. “Big data” versus “big brother”: on the appropriate use of large-scale data collections in pediatrics. Pediatrics. 2013 Apr;131 Suppl 2:S127–32.

97. Docherty A. Big Data - ethical perspectives. Anaesthesia. 2014 Apr;69(4):390–1.

98. Lu Z. PubMed and beyond: a survey of web tools for searching biomedical literature. Database. 2011 Jan 1;2011:baq036.

99. Shen B, Teschendorff AE, Zhi D, Xia J. Biomedical Data Integration, Modeling, and Simulation in the Era of Big Data and Translational Medicine. BioMed Res Int. 2014 Jul 24;2014:e731546.

100. Shah NH. Translational bioinformatics embraces big data. Yearb Med Inform. 2012;7(1):130–4.

101. Loi n° 78-17 du 6 janvier 1978 relative à l’informatique, aux fichiers et aux libertés.

102. Levallois-Barth C. Big data et protection des données personnelles : un défi (quasi) impossible ? Télécom. 2013 Jul;169.

103. Reynaudi M, Sauneron S. Médecine prédictive : les balbutiements d’un concept aux enjeux considérables. Centre d’analyse stratégique [Internet]. 2012 Oct [cited 2015 Feb 4]. Available from: http://archives.strategie.gouv.fr/content/medecine-predictive-les-balbutiements-dun-concept-aux-enjeux-considerables-note-danalyse-289

104. Gauthier T, Miquelard-Garnier G, Camilleri J-P. Big data et santé (3/3) Validité scientifique, éthique et dimension humaine. Les Echos. 2012 May 11.

105. Coll S. Surveiller et récompenser. Les cartes de fidélité qui nous gouvernent. Zürich: Seismo; 2015.

106. Ulmi N. Civilisation numérique, le retour du primitif. Le Temps. 2015 May 1.

107. Fischer H. La pensée magique du Net. Paris: François Bourin; 2014.

108. Boyd D, Crawford K. Critical Questions for Big Data. Inf Commun Soc. 2012 Jun 1;15(5):662–79.

ANNEXES

Annexe 1 : Article « Toward a Literature-Driven Definition of Big Data in Healthcare »

Cet article a été publié en libre accès dans la revue BioMed Research International, dans un numéro spécial intitulé « Big Data Approaches for Biomedical Informatics ». L'article est téléchargeable à ce lien : http://www.hindawi.com/journals/bmri/aa/639021/

Référence :

Baro E, Degoul S, Beuscart R, Chazard E. Toward a Literature-Driven Definition of Big Data in Healthcare. Biomed Res Int. Forthcoming 2015.

Abstract:

Objective. The aim of this study was to provide a definition of big data in healthcare. **Methods**. A systematic search of PubMed literature published until May 9, 2014 was conducted. We noted the number of statistical individuals (n) and the number of variables (p) for all papers describing a dataset. These papers were classified into fields of study. Characteristics attributed to big data by authors were also considered. Based on this analysis, a definition of big data was proposed. **Results**. A total of 196 papers were included. Big data can be defined as datasets with Log(n*p)≥7. Properties of big data are its great variety and high velocity. Big data raises challenges on veracity, on all aspects of the workflow, on extracting meaningful information and on sharing information. Big data requires new computational methods that optimize data management. Related concepts are data reuse, false knowledge discovery, and privacy issues. **Conclusion**. Big data is defined by volume. Big data should not be confused with data reuse: data can be big without being reused for another purpose, e.g., in omics. Inversely, data can be reused without being necessarily big, e.g., secondary use of Electronic Medical Records (EMR) data.

1. Introduction

The 21st century is an era of big data involving all aspects of human life, including biology and medicine [1]. With the advance in genomics, proteomics, metabolomics, and other types of omics technologies during the past decades, a tremendous amount of data related to molecular biology has been produced [2]. In addition, the transition from paper medical records to EHR systems has led to an exponential growth of data [3]. As a result, big data provides a wonderful opportunity for physicians, epidemiologists and health policy experts to make data-driven decisions that will ultimately improve patient care [3]. As Margolis stated, “Big data are not only a new reality for the biomedical scientist, but an imperative that must be understood and used effectively in the quest for new knowledge” [4].

To date, however, the term "big data" does not have a proper definition in the MeSH (Medical Subject Headings) database yet. A precise, well-formed, and unambiguous definition is a requirement for a shared understanding of the term big data. The objective of this work is to provide a definition of big data in healthcare through a review of literature.

2. Material and Methods

2.1 Search Strategy

For this literature review, we conducted a systematic search of the PubMed database for all papers published until May 9, 2014 using the keywords "big data". To be fully inclusive, we did not define a start date. We used the following PubMed query:

(a) (big data[Title/Abstract]) AND ("1900/01/01"[Date - Publication]: "2014/05/09"[Date - Publication])

Titles and abstracts were reviewed by a human for eligibility. Papers were excluded if they were not directly related to healthcare or if big data was not found to be the topic of the paper.

We then attempted to retrieve the full-text papers. We used online search facilities (the Free PMC database, Google and Google Scholar), resources and services of the Lille University library, and tried to directly contact the first or corresponding author. Full-text papers were then read.

Each of the remaining papers was included in the analysis and classified either as a paper describing a dataset, a dissertation, or a review of literature.

2.2 Data collection process

For each paper, we collected the following information: title, year of publication, journal title, specialty area, type of paper (paper using a dataset, dissertation, and literature review), the field of study, and characteristics given by authors to big data and to data reuse. In case the paper dealt with a dataset, we also collected the number of statistical individuals (n) and the number of variables (p). It should be noted that the number of statistical individuals n is not necessarily physical persons, but can also be,

e.g., gene sequences. The number of variables p could be, e.g., the number of physicochemical properties used to classify amino acids [5], the performance metrics adopted to evaluate model performance [6], or the number of features of medical claims. In this last case, the number of individuals n is represented by the number of records of medical claims [7].

2.3 Analysis and classification

Statistical analyses were performed with R statistical computing software [8]. In this paper, the notation "Log" denotes the decimal (or common, or decadic) logarithm, and the notation "CI95" denotes 95% confidence intervals. CI95 of binary variables were computed using the Binomial Law.

2.3.1 Time evolution of publication about big data in healthcare

To analyze the evolution of publication in healthcare, we draw a graph showing the annual publication of papers included in our review and a graph showing the annual publication of papers which were describing a dataset. We also noted the number of journals which published papers about big data in healthcare per year.

2.3.2 Time evolution of the size of big data in healthcare

In order to see the evolution of what authors refer to as "big data", from papers describing a dataset, we plotted the decimal logarithm of the product of the number of statistical individuals (n) and the number of variables (p), Log(n * p), as a function of the year.

2.3.3 Number of individuals and variables in each field of study

The numbers n and p were analyzed with respect to the field of study. To this end, the probability density functions of Log(n), Log(p), and Log(n * p) were plotted with respect to fields of study. Finally, Log(p) as a function of Log(n) was plotted with respect to fields of study.

2.4 Characteristics of big data

Characteristics attributed to big data by the authors in free text were noted as reading all the papers included in the analysis and were then sorted out by categories.

2.5 Proposal of a definition of big data

We then gathered to propose a definition of big data in healthcare. A difference was made between definition, properties, and related concepts. A dataset that matches the definition qualifies as “big data”, and thus has the properties that are proposed. Conversely, a dataset that has some or all of the listed properties does not necessarily qualify as “big data”. Finally, related concepts refer to properties that are not systematically related to big data.

We attempted to bring out a threshold of the volume of big data on the basis of findings from this literature review. The threshold resulted from a discussion between the authors of this paper, taking into account sizes of actual datasets, but also properties that are attributed to big data by the authors of the papers included in this literature review.

3. Results

3.1 Search strategy

The search query yielded 330 papers. After reading titles and abstracts, 94 papers were excluded. A total of 236 papers were included for full-text review. Eighteen papers were unavailable. The full-texts of the remaining 218 papers were read. After applying the exclusion criteria, 22 papers were excluded, leaving 196 papers. Papers were excluded due to the following reasons: papers not directly related to healthcare (18 papers), and papers in which big data was not the topic of the paper (4 papers). Of the 196 papers left for inclusion, there were 48 papers describing a dataset, 121 dissertations, and 27 reviews of literature. Figure 1 shows a detailed description of the search strategy and results.

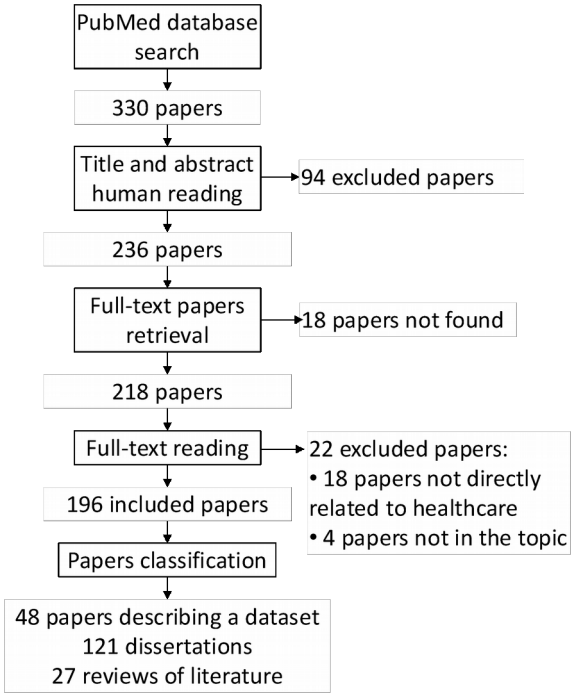

Figure 1: Flowchart of the literature review.

3.2 Data collection process

The number of papers by field of study among the 48 papers describing a dataset is listed in Table 1.

Among the 48 papers describing a dataset, three main categories of studies were identified: omics, medical specialties and public health. The term "omics" refers to biology fields of study ending in -omics, such as genomics, metabolomics, or proteomics. The main area represented is omics: 23 papers (48%, CI95 = [33; 63]). It is followed by medical specialties (endocrinology, infectiology, immunology, neurology, and imaging): 15 papers (31%, CI95 = [19; 46]), and public health (bioinformatics, Electronic Health Records (EHR), epidemiology, pharmacovigilance, and public health): 10 papers (21%, CI95 = [10; 35]).

3.3 Analysis and classification

3.3.1 Time evolution of publication about big data in healthcare

Figure 2 shows the evolution of the publication of papers about big data in healthcare from 2003 to 2013. Annual publication of papers about big data in healthcare increased from 1 in 2003 to 79 in 2013. In the same way, an increase in the annual publication of papers describing a dataset can be observed (Figure 3). The 196 papers included in our review were published in 134 different journals. Among these journals, one journal published papers about big data in healthcare in 2008. There were 68 in 2013.

Table 1: Number of papers by field of study among the 48 papers describing a dataset.

Field of study	Number of papers
Omics	
Genomics	18
Metabolomics	1
Proteomics	4
Medical specialties	
Endocrinology	2
Imaging	3
Immunology	1
Infectiology	1
Neurology	8
Public health	
Bioinformatics	3
EHR*	1
Epidemiology	2
Pharmacovigilance	1
Public health	3

*EHR: Electronic Health Records

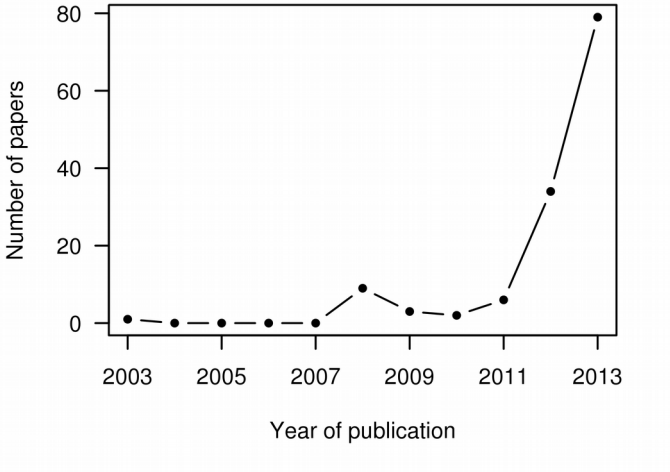

Figure 2: Number of papers about big data in healthcare published per year (full years only).

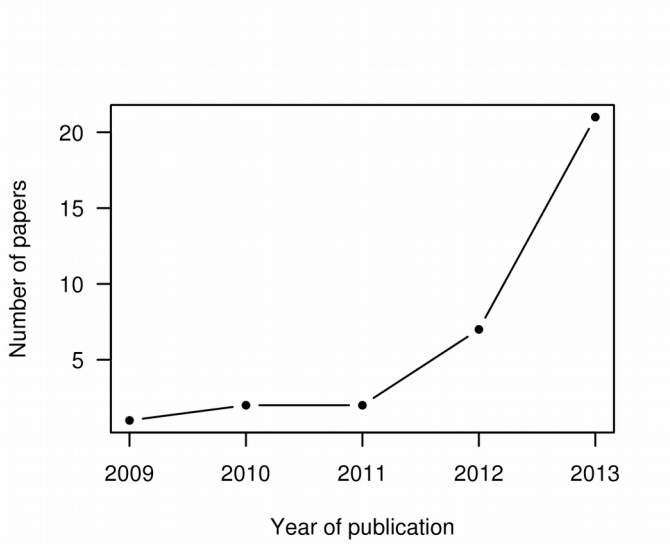

Figure 3: Number of papers about big data in healthcare describing a dataset per year (full years only).

3.3.2 Time evolution of the size of big data in healthcare

Figure 4 illustrates the decimal logarithm of the number of statistical individuals multiplied by the number of variables (Log(n * p)) for each year of publication of the papers that describe a dataset. We observe a non-significant increase of 0.43 per year (p value=0.34).

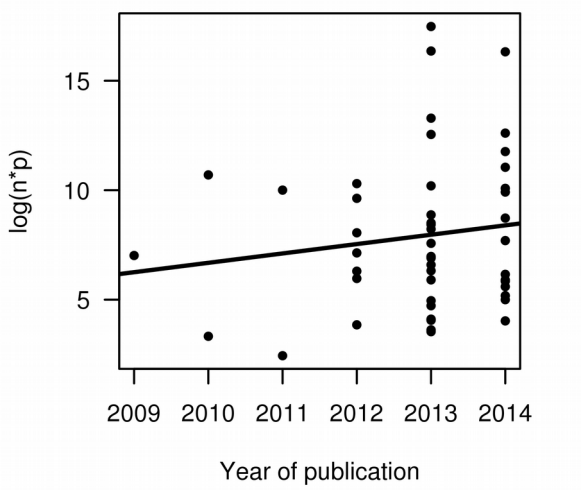

Figure 4: Log(n*p) per year of publication. The continuous line represents the linear regression (p=0.34).

3.3.3 Number of individuals and variables in each field of study

Figure 5, 6 and 7 represent the probability density function of Log(n), Log(p) and Log(n * p), respectively, for omics, medical specialties, public health and for all papers. It can be pointed out that Log(n * p) is inferior to 7 in 23 studies out of 48 (48%, CI95= [33; 63]).

Figure 8 shows Log(p) as a function of Log(n) for omics, medical specialties, and public health. This figure suggests the following differences between omics, medical specialties, and public health categories:

(i) big data in omics concern massive data collected on a limited number of individuals: small n, high p,

(ii) public health studies concern an important number of individuals and a low number of variables: high n, small p,

(iii) medical specialties are characterized by an important number of individuals and variables: high n, high p.

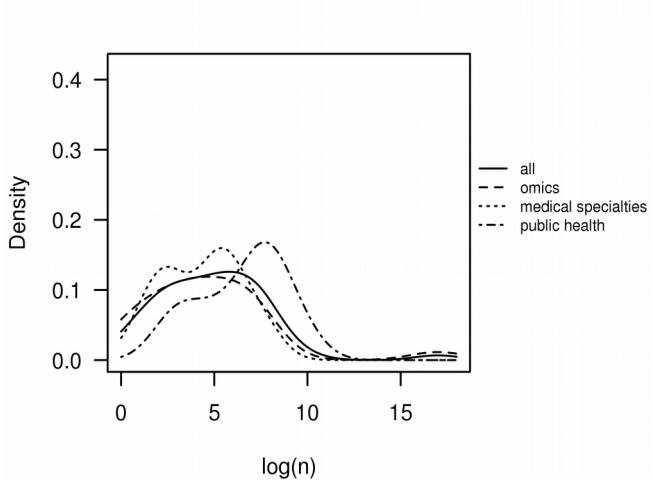

Figure 5: Representation of the probability density function of Log(n) for omics, medical specialties, public health and all fields together.

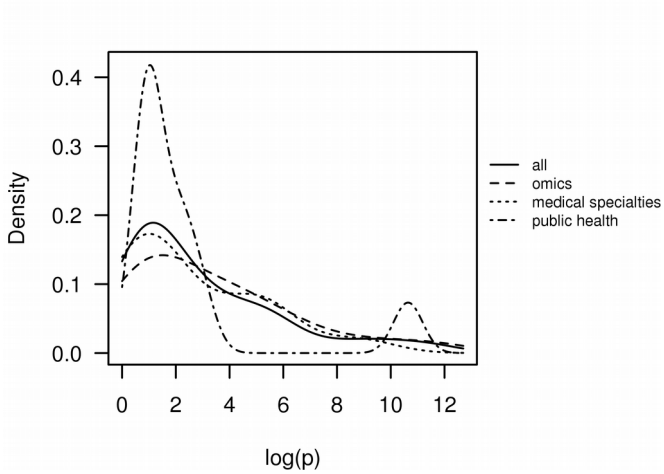

Figure 6: Representation of the probability density function of Log(p) for omics, medical specialties, public health and all fields together.

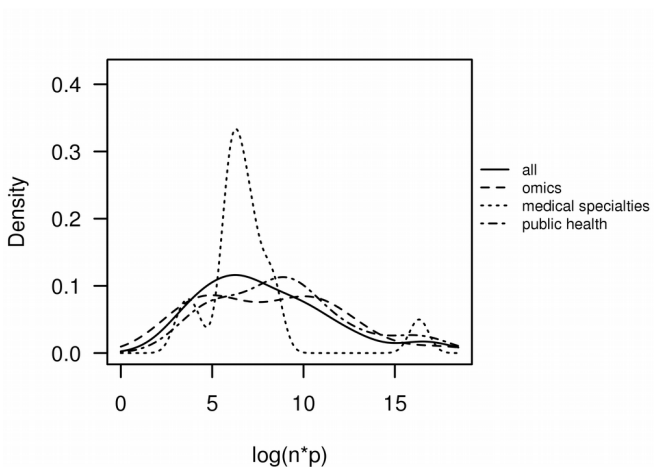

Figure 7: Representation of the probability density function of Log(n * p) for omics, medical specialties, public health, and all fields together.

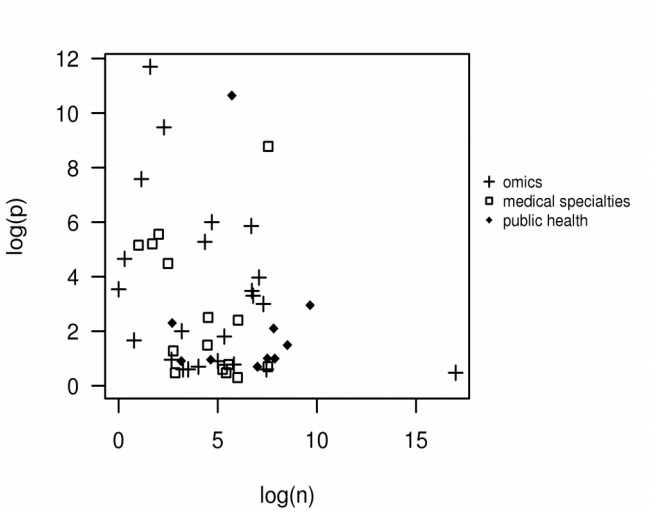

Figure 8: Log(p) as a function of Log(n) for omics, medical specialties, and public health. Each pictogram stands for one paper.

3.4 Characteristics of big data

The main characteristic about big data found in the papers is its massive size and complexity [7, 9–17]. Big data concern "not only the sheer scale and breadth of the new data sets but also their increasing complexity" [15]. Widely used notions to describe the complexity of big data are the three "Vs": volume, variety, velocity [7, 18–25]."Big Data is a term used to describe information assemblages that make conventional data, or database, processing problematic due to any

combination of their size (volume), frequency of update (velocity), or diversity (variety)" [18]. Veracity is a fourth "V" sometimes added to describe big data challenge [17, 23, 26–28]. Some authors mention a fifth "V": valorization [26, 29].

3.4.1 Volume

Volume is the main characteristic mentioned by authors [7, 12, 16, 21, 23, 26, 30, 31]."These correspond to the well-accepted notions of volume (breadth and/ or depth) (…) recognized as the hallmarks of big data." [21] "For volume, this translates today into terabytes (1012 bytes), petabytes (1015 bytes) or exabytes (1018 bytes)" [7]. "Volume - much greater amounts of rapidly multiplying data than were ever previously available" [25]. Some authors mention a big data threshold without clearly defining it [7, 32]: "How big is "Big"? (…) size is a relative term when it comes to data" [32]. "Those data are unquestionably 'big' (order 1017)" [21]. Data sets used "in epidemiology (…) in fact barely pass the "big data" threshold" [7].

3.4.2 Variety

Variety is another important characteristic of big data [7, 25, 26, 30, 31, 33–35]. Indeed, big data comes from various sources [23, 36]. Variety translates into "aggregation of widely disparate sources of data or mash-ups of data derived from independent sources" [7]. Unstructured data, e.g., free text data [7, 12, 37] and images [32, 38–40] are particularly a big challenge. In healthcare, "data take many forms including numbers, text, coded data, graphics, images, physiological measures (signals), and sound. Healthcare professionals rely on all their senses, including smell, to collect assessment data from individuals" [12]. In this area, "unstructured data is expected to exponentially outpace structured data" [34]. "Electronic Medical Records (EMR) generate massive data sets, offering the challenge of how to convert largely unstructured by-products of health care delivery into useful assets for patients' insight" [41]. Big data "can deviate from traditional structured data (organized in rows and columns) and can be represented as semi-structured data such as XML, or unstructured data including flat files which are not compliant with traditional database methods" [33]. These data are "unstructured for analysis using conventional relational database techniques" [31].

Moreover, big data can be "volatile, i.e. changing, and available only for a limited amount of time" [23].

3.4.3 Velocity

Accelerated increase of data is another attribute of big data [7, 21, 23, 25, 26, 31, 42]. It is "data at or near real-time" [25]. "Velocity refers to the enormous frequency with which today's data is generated, delivered, and processed" [31].

3.4.4 Challenge on veracity

Veracity comes next: big data can be difficult to validate [17, 26–28]. "Big data must be interpreted with caution, and in context, if it is to be clinically useful" [27]. It has a low veracity. Big data can never "be 100% accurate" [28].

3.4.5 Challenges on all aspects of the workflow

Big data raises challenges on all aspects of the workflow: from amassing [32], capturing [7, 37, 43–45], collecting [20, 46], storing [7, 20, 32, 43, 44, 47–53], data management [20, 43, 45, 54, 55], processing [9, 12, 19, 26, 47, 48, 51, 52, 56, 57], analyzing [7, 20, 31–33, 39, 43–45, 49–55, 58–60], to peer-reviewed publications of results [45]. Big data "creates difficulties in data capture, storage, cleaning, analytics, visualization and sharing" [43]. Big data is also difficult to valorize [26, 29]: big data "is not merely large in volume; it also moves rapidly, is difficult to validate and valorize" [26].

3.4.6 Challenges on statistical and computational methods

Finding new statistical and computational methods is another challenge raised by big data [33, 43, 50, 51, 59, 61, 62]. Big data requires "a change of perspective, infrastructure, and methods for data collection and analyses" [62]. Visualization methods

that allow us to understand the data need to be created [32, 43, 44, 57]. To make sense of big data, "the further creation of new tools and services for data discovery, integration, analysis, and visualization" [32] will be required.

3.4.7 Challenges on extracting meaningful information

Several authors emphasize the fact that it is necessary to derive useful information of these data [30, 44, 63, 64] and raise the question of how the data could be meaningfully interpreted: big data creates "challenges around how to meaningfully interpret the data - much of it not described using consistent standards or metadata - into information and recommendations while eliminating noise and erroneous data" [19].

3.4.8 Challenges on facilitating information access and sharing

Many authors highlight the necessity of identifying ways to facilitate information access and sharing [7, 15, 30, 34, 43–46, 49, 50 ,53, 62, 63, 65–67]. It is necessary to promote "collaboration among scientists" [46]. Data must be made more readily available from more open sources to better compare data.

3.4.9 Not enough human experts

Some authors mention the fact that the number of available human experts who have both clinical and analytic knowledge is not sufficient yet [30, 68]: "the role needs some sort of hybrid person that has clinical knowledge and analytic knowledge. We are experiencing a drought in terms of analytic experience. We don’t have enough of those people in place yet" [30].

3.4.10 Data reuse

Some authors mention the fact that big data can be data that are commonly collected without an immediate use: "Massive amounts of data are commonly collected without an immediate business case, but simply because it is affordable. This data, so it is hoped, will later answer questions, most of which yet have to arise" [20]. They put into light the fact that big data are often a secondary use of data, which we can call data reuse [14, 20, 21, 41, 65, 69–72].

3.4.11 False knowledge discovery

Some authors highlight the fact that deriving knowledge from big data can lead to false results and to conclusions that are wrong [73–75]: "Exploratory results emerging from Big Data are no less likely to be false" [75]. We can’t extract knowledge from big data without knowing the context in which data sets were collected: "big size is not enough for credible epidemiology" [74].

3.4.12 Privacy issues

One concern mentioned by several authors is privacy issues: "the increasing ease with which data may be used and reused has increased concerns about privacy and informed consent" [76]. The ability "to protect individual privacy in the era of big data has become limited" [39]. Even if large databases use pseudonymised personal confidential data that have been anonymised, they retain a residual risk of re-identification. Indeed, the identity of individuals can be determined by manipulating databases through data linkage techniques [28, 39, 66, 77]. The data torrent poses ethical challenges [15]. "The widespread implementation of EHRs and the need to share data to measure quality and manage accountable care organizations (ACOs) brings to light all of the privacy issues surrounding sharing patient data" [66]. "The ability to derive DNA-based information from non-DNA-based sources generalizes the issue of data de-identification beyond the area of genotypic data privacy and has thus potentially important consequences for privacy rules in scientific research" [39].

3.5 Proposal of a definition of Big Data

A definition of big data was established on the basis of findings from the literature review. We consider that big data should exclusively be defined by volume, and we propose that a dataset could be qualified as "big dataset" only if Log(n * p) is superior or equal to 7.

Properties of big data can be listed as follows:

- great variety,
- high velocity,
- challenge on veracity,
- challenge on all aspects of the workflow,
- challenge on computational methods,
- challenge on extracting meaningful information,
- challenge on sharing data,
- challenge on finding human experts.

Related concepts of big data are:

- data reuse,
- false knowledge discovery,
- privacy issues.

The definition of big data is summed up in Table 2.

Table 2: Definition of big data in healthcare.

Définition	Volume : $Log(n * p) \geq 7$
Properties	Great variety
	High velocity
	Challenge on veracity
	Challenge on all aspects of the workflow
	Challenge on computational methods
	Challenge on extracting meaningful information
	Challenge on sharing data
	Challenge on finding human experts
Related concepts	Data reuse
	False knowledge discovery
	Privacy issues

4. Discussion

In this work, through a detailed literature review, we tried to provide a current and quantitative definition of big data. We performed a literature review of 196 papers published until May 2014. Finally, we proposed a definition of big data in healthcare.

This systematic search should ensure that we accumulate a relatively complete census of relevant literature of big data in healthcare. However, we may have missed papers that do use big data in the research but were not included in our query because the term was not mentioned in the abstract or keywords of the paper. Those papers could be less and less frequent in the future.

Nevertheless, as there is no definition of big data, the literature can itself be wrong. It is a limitation of this inductive approach: we use observations to build a definition. The problem of defining a threshold illustrates this difficulty: the threshold of 107 may appear in disagreement with the results of Figure 7. This definition of big data is simply the result of a discussion between the authors of this literature review. It has been decided based on the results of the number of individuals and of variables found in the studies describing a dataset, but it has also taken into account the characteristics of big data mentioned by the authors of all the papers included in this literature review. Thus, for example, we can consider that the problems related to computational methods do not exist for Log(n * p) inferior to 7, even when the analysis is performed with a simple spreadsheet instead of a statistical software calling for high computational capacities. However, this proposal suggests that half of the studies describing a dataset in this literature review wrongly call their dataset big data. As everyone talks about the challenges of computing and data processing, considering what we know today in practice about software and computers, it would have been difficult to admit a threshold of Log(n * p) superior or equal to 6 (although such a threshold already excludes 35% of the studies of our review), because we know that, nowadays, such size of data is easy to deal with.

It should also be pointed out that there is an undeniable current trend of big data, which leads to the fact that the term "big data" is now used to qualify datasets that, in the past, would not have been called this way. Moreover, we can consider that the size of datasets that qualify as big data may keep on increasing due to the main property of big data, which is the challenge on data processing and the fact that computational infrastructure that is required to process these large-scale datasets may progress with time.

Data reuse has been defined as a related concept of big data because we think that there might be some confusion between these two terms: data reuse is the fact of using for decisional purposes data that were collected routinely for transactional purposes, whereas big data is related to the size of the data collection. Indeed, data can be big without being reused for another purpose: this is the case of omics for example. Inversely, data can be reused without being necessarily big, such as secondary use of data from Electronic Medical Records (EMR).

Big data presents many opportunities for translational studies, and informatics will be the key for successful translational research [78]. As Shah stated, “translational informatics is ready to revolutionize human health and healthcare using large-scale measurements on individuals. Data-centric approaches that compute on massive amounts of data to discover patterns and to make clinically relevant predictions will gain adoption” [79]. Cloud computing could be an enabling tool to facilitate translational bioinformatics research [67].

Informatics is needed to fully harness the potential of health data and new tools are emerging to translate health data into knowledge for improved healthcare.

Annexe 2 : Définitions

Cloud computing : Le *cloud computing* est l'accès via un réseau de télécommunications (généralement Internet), à la demande et en libre-service, à des ressources informatiques partagées configurables, en exploitant la puissance de calcul ou de stockage de serveurs informatiques distants. Il s'agit donc d'une délocalisation de l'infrastructure informatique.

Data center : Un centre de traitement de données est un site physique sur lequel se trouvent regroupés des équipements constituants du système d'information de l'entreprise (ordinateurs centraux, serveurs, baies de stockage, équipements réseaux et de télécommunications, etc.). Il peut être interne et/ou externe à l'entreprise.

Data mining : L'exploration de données, connue aussi sous l'expression de fouille de données, forage de données, ou prospection de données, a pour objet l'extraction d'un savoir ou d'une connaissance à partir de grandes quantités de données, par des méthodes automatiques ou semi-automatiques. Elle utilise un ensemble d'algorithmes issus des statistiques, de l'intelligence artificielle ou de l'informatique, pour construire des modèles à partir des données, c'est-à-dire trouver des structures intéressantes ou des motifs selon des critères fixés au préalable, et en extraire des connaissances utiles.

Omiques : Le terme « omique » désigne les domaines de la biologie dont le nom se termine par « -omique », comme la génomique, la métabolomique et la protéomique. Le suffixe « -ome » indique les objets d'étude de ces domaines de la biologie. Dans notre exemple, il s'agit respectivement du génome, du métabolome et du protéome. Ce terme est traduit du mot anglo-saxon « omics ». Le terme « omique » est utilisé dans ce travail par volonté de traduction, bien que le mot anglo-saxon « omics » soit fréquemment utilisé par les scientifiques.

Recherche translationnelle : La recherche translationnelle (*translational research* en anglais) est une discipline scientifique émergente qui vise à traduire en applications concrètes (les sciences appliquées) les théories scientifiques et les découvertes de laboratoire. Il s'agit de « traduire » (*translate* en anglais) les connaissances en application. Ce terme est le plus souvent utilisé en sciences médicales et vise à traduire les découvertes de laboratoire en pratique clinique ou en intervention de santé publique.

AUTEUR : BARO Émilie

Date de Soutenance : 11 mai 2015

Titre de la Thèse : Vers une définition des Big data en santé basée sur la littérature

Thèse – Médecine – Lille 2015

Cadre de classement : Santé publique

Mots-clés : big data, exploration de données, réutilisation des données, omiques

Résumé :

Vers une définition des Big data en santé basée sur la littérature

Contexte : Le terme « Big data » émerge récemment dans la littérature scientifique. Ce terme n'est pas encore référencé dans le MeSH (*Medical Subject Headings*). Or son usage semble ambigu et les propriétés attribuées à ce terme par les auteurs varient selon les articles. L'objectif de ce travail est de proposer une définition du terme « Big data » à partir d'une revue de la littérature incluant les articles mentionnant ce terme et de décrire systématiquement les propriétés rattachées à ce terme par les auteurs.

Méthode : Nous avons conduit une recherche systématique de la base de données PubMed de tous les articles publiés jusqu'au 9 mai 2014 en utilisant le terme de recherche « Big data ». Ces articles ont été classés en domaines d'études. Le nombre d'individus statistiques (*n*) et le nombre de variables (*p*) ont été relevés pour les articles décrivant un jeu de données. Nous avons également considéré les caractéristiques attribuées aux Big data par les auteurs. En s'appuyant sur cette analyse, une définition des Big data a été proposée.

Résultats : Cent quatre-vingt-seize articles ont été inclus. Trois principales catégories d'études ont été identifiées : les spécialités « omiques », les spécialités médicales et la santé publique. Les Big data peuvent être définies comme des données avec un *Log(n * p)* supérieur ou égal à 7. Les propriétés des Big data sont ses grandes variétés de données et leur importante vélocité. Les Big data soulèvent des défis concernant la véracité, la gestion des données, l'extraction d'informations utiles, le partage des informations et l'existence d'experts humains ayant à la fois des compétences cliniques et analytiques. L'émergence des Big data nécessitent la création de nouvelles méthodes de calcul qui optimisent la gestion de données. Les concepts reliés sont la réutilisation des données (*data reuse*), la possibilité de connaissances erronées et la question de la confidentialité des données.

Conclusion : Les Big data sont définies par le volume. La taille des données qui les qualifie de « Big data » va probablement augmenter avec le temps. Les Big data ne doivent pas être confondues avec le *data reuse :* les données peuvent être massives sans être forcément réutilisées dans un autre objectif, par exemple dans le cas des spécialités « omiques ». Inversement, des données peuvent être réutilisées sans être nécessairement de grande dimension. C'est le cas par exemple de l'utilisation secondaire du dossier patient informatisé.

Composition du Jury :

Président : Monsieur le Professeur Beuscart

Assesseurs : Monsieur le Professeur Salomez

Monsieur le Professeur Amouyel

Monsieur le Professeur Duhamel

Monsieur le Docteur Chazard