Institut de Recherche en Santé Publique
GIS-IReSP

Questions de santé publique

N°30 – septembre 2015 https://doi.org/10.1051/qsp/2015030

Le terme de données massives décrit le recueil, la gestion et l'analyse de données de sources hétérogènes, d'un volume important, avec une grande vitesse de génération. À côté des données omiques, les données de capteurs, les smartphones et les réseaux sociaux créent une énorme quantité d'informations utilisables en épidémiologie. Les méthodes statistiques classiques ont des limites dans ce contexte qu'on soit dans la situation où l'on a un grand nombre de sujets et un grand nombre de variables ou dans celle où le nombre de variables est bien plus grand que le nombre de sujets. Les méthodes *data-driven* ou *hypothesis-driven* peuvent être utilisées pour réduire l'information et aider à l'interprétation des associations mises en évidence. Nous illustrons les aléas et les succès de quelques approches développées sur les données massives dans le champ des maladies transmissibles. Il faut garder à l'esprit que les données massives ne sont pas une solution magique pour l'interprétation causale des associations, au cœur de la démarche épidémiologique.

Données massives, vous avez dit données massives ?

Pierre-Yves Boelle[1], Rodolphe Thiébaut[2], Dominique Costagliola[1]

[1] Sorbonne Universités, UPMC Univ Paris 06, Inserm, Institut Pierre Louis d'épidémiologie et de Santé Publique (IPLESP UMRS 1136), Paris, France
[2] Inserm, U1219, Bordeaux ; INRIA, SISTM, Talence ; Université de Bordeaux, ISPED, Bordeaux ; CHU de Bordeaux, USMR, Bordeaux ; Vaccine Research Institute, Créteil, France
pierre-yves.boelle@upmc.fr – rodolphe.thiebaut@isped.u-bordeaux2.fr – dcostagliola@ccde.chups.jussieu.fr

Le terme de *Big Data* ou encore données massives est très utilisé de nos jours que ce soit dans la presse grand public, la presse scientifique ou la presse économique. Mais que recouvre ce concept ?

BIG DATA MAIS ENCORE ?

En sciences de l'information, ce terme décrit le recueil et la gestion de bases de données se distinguant par un volume important, une variété des types de données de sources hétérogènes et une grande vitesse de génération [1]. En biologie et en médecine, l'exemple type est celui des données génomiques, produites de plus en plus rapidement, avec un détail toujours plus fin et pour des coûts de moins en moins élevés [2], et qui ont notamment été utilisées dans les *genome wide association studies* (GWAS : association phénotype/génotype à l'échelle du génome entier). Cette tendance se poursuit aujourd'hui avec une analyse systématique des « -omiques »[1] : épigénome (les éléments non codés dans l'ADN contrôlant l'expression des gènes), méthylome (points de méthylation sur l'ADN), transcriptome (l'ensemble des ARN issus de la transcription), métabolome (les métabolites présents dans un échantillon biologique), protéome (l'ensemble des protéines exprimées dans une cellule ou un tissu à un instant donné). Mais, à une toute autre échelle, les technologies modernes, smartphones et réseaux sociaux notamment, sont aussi la source d'une énorme quantité d'informations utilisables en épidémiologie, informations qu'il aurait été impossible de recueillir par des approches traditionnelles. L'exposome, concept introduit en 2005 [3], définit par exemple l'ensemble des expositions environnementales (par exemple, régime alimentaire, pollution sous toutes ses formes, mode de vie, infections, les traitements médicamenteux, stress, etc.) auxquelles un individu est confronté de sa naissance jusqu'à son décès, et ayant un impact sur l'environnement chimique interne et sur la santé [4]. La prise en compte de l'ensemble de ces données dans les études épidémiologiques suscite des attentes et des espoirs en termes de compréhension des causes et des mécanismes des maladies comme pour la personnalisation du suivi médical. Dans cet article, nous présentons quelques problèmes d'analyses posés par les données massives, illustrons l'impact des données des réseaux sociaux et objets connectés dans des études actuelles, et concluons finalement sur la notion de causalité épidémiologique dans le contexte des données massives.

BEAUCOUP DE VARIABLES SUR BEAUCOUP DE SUJETS, QUELS PROBLÈMES D'ANALYSES ?

À l'heure actuelle, les données massives en épidémiologie confrontent les chercheurs à deux situations. Dans la première, le nombre d'individus et le nombre de variables mesurées sont importants (> milliers d'individus, > milliers de variables). Il s'agit typiquement des études GWAS dans lesquelles on recherche une association entre des variants génétiques et des phénotypes [5] ou bien des bases de données comme celles de l'assurance maladie (SNIIRAM ou EGB [Echantillon Généraliste Bénéficiaire]) [6]. Dans la seconde, le nombre de variables est beaucoup plus important que le nombre d'individus (peu d'individus, > milliers de variables). Cette situation est courante avec les études translationnelles[2] qui peuvent être ancillaires[3] à un essai clinique et destinées à explorer des hypothèses nécessitant des mesures spécifiques comme la transcriptomique ou la protéomique chez un petit nombre d'individus [2]. Ces deux situations exposent l'épidémiologiste à des difficultés alors inédites dans cette discipline. La première est de pouvoir gérer le volume des données recueillies, tant pour le stockage que pour la transmission et le calcul, nécessitant un équipement informatique performant. Une autre difficulté, moins triviale, est de gérer la complexité des données. En effet, le nombre important de variables engendre des difficultés de visualisation nécessitant à la fois un équipement adapté (mémoire vive) et des approches spécifiques (des équipes de recherche travaillant spécifiquement sur ce sujet). Les méthodes statistiques usuelles pour explorer les associations ont de nombreuses limites dans un tel contexte. D'une part, un grand nombre de variables engendre une multiplicité de tests statistiques, conduisant à beaucoup plus de corrélations « significatives » à tort que d'associations existantes en réalité (à risque de première espèce fixé) [7]. D'autre part, de la taille importante de l'échantillon liée au grand nombre d'individus découle le fait que la puissance statistique de toute comparaison est forte. Ce n'est pas un problème en soi mais cela nécessite de prendre en compte l'interprétation clinique de la taille de l'effet car beaucoup d'associations seront potentiellement statistiquement significatives. Les méthodes exploratoires dites *data-driven*, supposant l'absence d'hypothèses a priori permettent de faire de la réduction des données pour guider l'interprétation des associations entre les différentes variables. Il est aussi possible d'utiliser des méthodes dites *hypothesis-driven*, nécessitant l'élaboration d'hypothèses préalables, en interaction étroites avec les différents scientifiques impliqués dans les études. Les deux approches ne sont pas mutuellement exclusives mais plutôt complémentaires. Il reste critique de prendre en compte la séquence temporelle des données, ce qui nécessite des développements des méthodes *data-driven* [8]. Enfin, lorsque le nombre de variables est bien plus grand que le nombre d'individus, des méthodes statistiques spécifiques doivent être utilisées pour éviter un problème de sur-ajustement où la prédiction est parfaite dans l'étude mais très mauvaise avec un nouvel échantillon indépendant de celui qui a servi à établir le modèle.

1. Les sciences « omiques » désignent les champs de la biologie qui s'intéressent aux interactions entre des ensembles vivants complexes (populations, individus, cellules, protéines, ARN, ADN) en prenant en compte leur environnement.

2. La recherche translationnelle assure un continuum entre recherche fondamentale et recherche clinique, afin de favoriser l'application concrète des résultats de la recherche fondamentale.

3. L'étude ancillaire porte sur un petit nombre de cas (qui font partie d'un essai thérapeutique plus large) afin d'évaluer l'effet du médicament sur un point particulier.

DONNÉES MASSIVES, OBJETS CONNECTÉS, RÉSEAUX SOCIAUX

Les données massives utilisées aujourd'hui proviennent de deux sources : des innovations technologiques spécifiques qui permettent d'acquérir des données détaillées (puces génomiques, senseurs GPS pour mesurer les déplacements ou l'activité des personnes, etc.), ou des activités dont la finalité n'est pas épidémiologique mais le devient par une utilisation secondaire (données de remboursement, séjours hospitaliers, etc.). Un domaine emblématique de la santé publique, celui des maladies transmissibles, a vu de nombreuses tentatives d'utilisation de telles données ces dernières années, et permet d'illustrer l'utilisation de données issues d'internet et de bases de données commerciales comme celle d'objets techniques innovants.

Un premier exemple, qui a été largement commenté, est celui de l'utilisation des moteurs de recherche et de sites sociaux comme alternative à la surveillance épidémiologique traditionnelle. Partant de l'idée que la recherche en ligne de symptômes ou de traitements pourrait augmenter lors d'épidémies, Google a développé un algorithme sélectionnant les requêtes les plus utilisées lors d'épidémies de grippe passées [9]. En suivant prospectivement le volume de ces requêtes au cours du temps, on peut alors proposer un nouvel indicateur épidémique généré purement *in silico*[4], basé sur les millions de requêtes des internautes. L'enthousiasme initial autour de cette approche, suscité par la bonne corrélation entre requêtes internet et indicateurs épidémiologiques plus traditionnels, a cependant dû être tempéré : l'algorithme, développé en 2008, n'a pas détecté la survenue de la grippe H1N1 en 2009 et, plus globalement, ce système a mené à une surestimation presque systématique du niveau de circulation du virus grippal au cours du temps [10]. Dans ce cadre, l'utilisation de données massives n'a donc pas été une alternative, mais présente un complément important à une approche plus traditionnelle.

Mais l'utilisation de données massives a aussi pu mener à des résultats plus décisifs, par exemple sur la dissémination globale de pathogènes. Dès les années 1970, des travaux avaient montré l'importance de la mobilité des populations dans la propagation des épidémies en reliant les flux de passagers entre 50 aéroports dans le monde à la pandémie de grippe de 1957 [11]. Aujourd'hui, les développeurs de GLEAM (www.gleamviz.org), un des simulateurs les plus aboutis pour étudier la dissémination globale d'une épidémie, utilisent les connexions de 3 800 aéroports couvrant plus de 99 % du trafic mondial et 50 000 000 de vols chaque année. Pour de nombreuses maladies émergentes (grippe H1N1, SARS, MERS-CoV, Chikungunya, Ebola) ces modèles ont montré une très bonne capacité à prédire la séquence temporelle des introductions à partir du pays source [12].

À une échelle plus locale, la mobilité observée à partir de la localisation des téléphones portables a permis de mieux comprendre les conséquences sanitaires d'un désastre, comme le tremblement de terre à Haïti [13] et l'épidémie de choléra qui a suivi ; elle a aussi été proposée comme clé pour comprendre l'extension de l'épidémie d'Ebola qui a touché des zones beaucoup plus intensément connectées en 2014 que dans toutes les épidémies précédentes [14].

Finalement, l'utilisation d'objets technologiques nouveaux dans le cadre d'études observationnelles a également permis d'obtenir des données jamais auparavant disponibles pour mieux comprendre des phénomènes épidémiques. Par exemple, en équipant 600 personnes avec un récepteur GPS dans une ville du Pérou, la cartographie des mouvements de population a été dressée avec une grande précision. Ceci a permis d'étudier la propagation de la dengue à l'échelle intra-citadine [15]. Enfin, l'équipement de 500 personnes dans un hôpital, patients et professionnels, avec des senseurs électroniques permettant d'enregistrer leurs contacts a permis de documenter plusieurs millions de contacts au cours de 6 mois de suivi [16]. De telles mesures peuvent aider à comprendre la dissémination des staphylocoques, et possiblement à proposer de nouvelles stratégies d'hygiène.

BIG DATA IS A POWERFUL TOOL FOR INFERRING CORRELATIONS, NOT A MAGIC WAND FOR INFERRING CAUSALITY[5]

Les associations entre variables mises en évidence dans l'analyse des données massives sont une mine d'or pour générer et étayer des hypothèses épidémiologiques, c'est pourquoi il faut favoriser l'utilisation d'innovations technologiques et l'analyse secondaire de données collectées dans un contexte différent. Mais le pouvoir des nombres n'entraîne pas la simplification de la démarche épidémiologique, centrée sur l'interprétation causale des associations observées. Au contraire, la question posée devra être encore mieux explicitée sur la base d'une théorie conceptuelle pour permettre l'interprétation. Et comme dans d'autres disciplines, il faudra valider le rôle de facteurs identifiés dans de telles analyses par exemple en montrant que l'altération de ce facteur permet de diminuer le risque de la maladie associée. C'est un message que notre communauté doit porter en direction des autres disciplines et du grand public par rapport aux attentes vis-à-vis des données massives en santé. ■

4. C'est-à-dire à partir de modèles informatiques.

5. « Les données massives sont des outils puissants de déductions de corrélations, mais il ne s'agit pas de baguettes magiques permettant de déduire des causalités », March 29, 2013, Steamrolled by Big Data by Gary Marcus, The New Yorker.

Directeur de la publication : Geneviève Chêne. **Rédacteur en chef :** Claire-Isabelle Coquin. **Questions de santé publique** est une publication de l'Institut de Recherche en Santé Publique. **Éditeur :** Éditions EDK (Groupe EDP/Sciences), 109, avenue Aristide Briand, 92541 Montrouge Cedex, France Tél. : 01 41 17 74 05 fax : 01 49 85 03 45. E-mail : edk@edk.fr. **Création/maquette/infographie :** Isabelle Benoit. Dépôt légal : à parution – ISSN : 1967-5615

N°30 – septembre 2015

RÉFÉRENCES

[1] Trelles O, Prins P, Snir M, Jansen RC. Big data, but are we ready? *Nat Rev Genet* 2011 ; 12 : 224.

[2] Thiébaut R, Hejblum B, Richert L. L'analyse des Big Data en santé publique. *Rev Epidemiol Sante Publ* 2014 ; 62 : 1-4.

[3] Wild CP. Complementing the genome with an « exposome » : the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiol Biomark Prev* 2005 ; 14 : 1847-50.

[4] Rappaport SM, Smith MT. Environment and disease risks. *Science* 2010 ; 330 : 460-1.

[5] Limou S, Le Clerc S, Coulonges C, Carpentier W, Dina C, Delaneau O, Labib T, Taing L, Sladek R, Deveau C, Ratsimandresy R, Montes M, Spadoni JL, Lelièvre JD, Lévy Y, Therwath A, Schächter F, Matsuda F, Gut I, Froguel P, Delfraissy JF, Hercberg S, Zagury JF, ANRS Genomic Group. Genomewide association study of an AIDS-nonprogression cohort emphasizes the role played by HLA genes (ANRS Genomewide Association Study 02). *J Infect Dis* 2009 ; 199 : 419-26.

[6] Orriols L, Delorme B, Gadegbeku B, Tricotel A, Contrand B, Laumon B, Salmi LR, Lagarde E ; CESIR research group. Prescription medicines and the risk of road traffic crashes : a French registry-based study. *PLoS Med* 2010 ; 7 : e1000366.

[7] Hochberg Y, Benjamini Y. Controlling the false discovery rat : a practical and powerful approach to multiple testing. *JR Stat Soc Ser B* 1995 ; 57 : 289-300.

[8] Binder H, Blettner M. Big data in medical science : a biostatistical view. *Dtsch Arztebl Int* 2015 ; 112 : 137-42.

[9] Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. *Nature* 2009 ; 457 : 1012-4.

[10] Lazer D, Kennedy R, King G, Vespignani A. Big data. The parable of Google flu : traps in big data analysis. *Science* 2014 ; 343 : 1203-5.

[11] Rvachev LA, Longini IM. A mathematical model for the global spread of influenza. *Math Biosci* 1985 ; 75 : 3-22.

[12] Balcan D, Hu H, Goncalves B, Bajardi P, Poletto V, Ramasco JJ, Paolotti D, Perra N, Tizzoni M, Van den Broeck W, Colizza V, Vespignani A. Seasonal transmission potential and activity peaks of the new influenza A (H1N1) : a Monte Carlo likelihood analysis based on human mobility. *BMC Med* 2009 ; 7 : 45.

[13] Lu X, Bengtsson L, Holme P. Predictability of population displacement after the 2010 Haiti earthquake. *Proc Natl Acad Sci USA* 2012 ; 109 : 11576-81.

[14] Wesolowski A, Buckee CO, Bengtsson L, Wetter E, Lu X, Tatem AJ. Containing the Ebola outbreak : the potential and challenge of mobile network data. *PLoS Curr* 2014 ; 6.

[15] Paz-Soldan V, Morrison AC, Elder JP, Kochel TJ, Scott TW, Kitron U. Usefulness of commercially available GPS data-loggers for tracking human movement and exposure to dengue virus. *Int J Health Geographr* 2009 ; 8 : 68.

[16] Obadia T, Silhol R, Opatowski L, Temime L, Legrand J, Thiébaut AC, Herrmann JL, Fleury É, Guillemot D, Boëlle PY, I-Bird Study Group. Detailed contact data and the dissemination of Staphylococcus aureus in hospitals. *PLoS Comput Biol* 2015 ; 11 : e1004170.

Actualités 2015-2016 de l'Institut Thématique Multi-Organisme Santé publique et de l'Institut de Recherche en Santé Publique

L'Institut Thématique Multi-Organisme (ITMO) Santé Publique au sein de l'Alliance pour les sciences de la vie et de la santé (AVIESAN) et l'IRESP ont une démarche très active de soutien des chercheurs pour l'exploitation des bases de données à travers des actions telles que le Workshop santé numérique, le Club Cohortes ou le Cohort Innovation Day, et mettent à disposition des outils comme le Portail Épidémiologie-France, qui répertorie l'ensemble des bases de données françaises pour la recherche en santé (plus de 900 bases recensées).

Tout en mobilisant le financement de recherches exploitant des bases de données existantes, l'ITMO Santé Publique et l'IRESP favorisent aujourd'hui l'organisation d'une nouvelle étape pour l'accès et l'exploitation des données de santé à des fins de recherche.

Dans la continuité de l'appui depuis longtemps proposé par l'Inserm aux chercheurs, à travers la mise à disposition et l'aide apportée à l'exploitation des données sur les causes médicales de décès ou à travers l'accès par un guichet unique aux données du SNIIRAM, l'Inserm s'est engagé dans la préparation d'une infrastructure capable d'apporter un appui technique, juridique, éthique et scientifique aux chercheurs pour l'accès facilité au Système National des Données de Santé (SNDS) créé par l'article 193 de la loi n° 2016-41 du 26 janvier 2016 de modernisation de notre système de santé. L'objectif est de promouvoir l'utilisation par les chercheurs des données de santé au bénéfice de la santé des populations et dans le respect de la confidentialité des données personnelles.

Présentation de l'Institut de Recherche en Santé Publique

L'Institut de Recherche en Santé Publique (IReSP) est un groupement d'intérêt scientifique créé en 2007 par une convention entre 23 partenaires, acteurs de la recherche en Santé Publique (voir ci-dessous). Son objectif général est de constituer une communauté scientifique de taille internationale capable de répondre au développement souhaité de la recherche en Santé Publique et de contribuer aux nouveaux dispositifs mis en place par la loi du 9 août 2004 relative à la politique de Santé Publique. Pour atteindre cet objectif, le GIS-IReSP s'appuie sur une mutualisation des compétences et des moyens de ses partenaires. Le GIS-IReSP est dirigé par Geneviève Chêne, professeur de santé publique.

Les domaines de recherche soutenus sont les suivants :

- Fonctionnement du système de santé
- Politiques publiques et santé
- Interaction entre les déterminants de la santé

Les modalités d'actions du GIS sont :

- Lancement d'appels à projets ciblés
- Aide à l'émergence d'équipes de recherche
- Mutualisation d'outils pour la recherche en Santé Publique
- Constitution de groupes de travail sur des sujets émergents
- Aide à la mise en place et à l'exploitation de grandes enquêtes et de grandes bases de données
- Valorisation et communication

Afin de pallier le manque de visibilité des résultats de la recherche en Santé Publique en France, l'IReSP a décidé de créer ce bulletin trimestriel à large diffusion intitulé *Questions de Santé Publique*. Chaque trimestre, un sujet de recherche en Santé Publique intéressant le grand public est traité par un chercheur.

Les partenaires de l'IReSP

Ministères (Ministère de la Santé [DGS et DREES], Ministère délégué à la Recherche), Opérateurs de la recherche en Santé Publique (CNRS, Inserm, IRD, INED, EHESP, UDESCA, CPU, Institut Pasteur, CNAM, Sciences Po), Agences et opérateurs de la Santé Publique (InVS, HAS, ANSM, ANSES, EFS, ABM, INPES, INCa), Organismes de protection sociale (CNAMTS, RSI, CNSA).

Site internet : www.iresp.net

Cet article ainsi que les précédents numéros de *Questions de Santé Publique* sont téléchargeables sur le site internet de l'IReSP : www.iresp.net