Gynécologie Obstétrique & Fertilité 44 (2016) 375–376

Disponible en ligne sur
ScienceDirect
www.sciencedirect.com

Elsevier Masson France
EM|consulte
www.em-consulte.com

Éditorial

Cancer du sein et Big Data : évolution ou révolution ?

Breast cancer and Big Data: Evolution or revolution?

INFO ARTICLE

Mots clés :
Cancer du sein
Big Data
Machine learning
Données de santé

Keywords:
Breast cancer
Big Data
Machine learning
Health data

Les origines du Big Data remontent à 1941, date à laquelle les premières références ont été faites à la notion d'« explosion de l'information » dans l'Oxford Dictionary of English. James Maar a mis en exergue dès 1996 dans un rapport de la National Academy of Sciences la notion de « massive data set » (jeux de données massives) [1]. Mais c'est seulement en 1997 que le terme précis de Big Data a fait son apparition dans un article de la bibliothèque numérique de l'Association for Computing Machinery [2], faisant référence au challenge technique que représente l'analyse de grands ensembles de données. Le terme Big Data a récemment été introduit dans les dictionnaires français avec son équivalent officiel « mégadonnées » proposé par la Commission générale de terminologie et de néologie [3]. Il est depuis utilisé pour désigner « des données structurées ou non, dont le très grand volume requiert des outils d'analyse adaptés ». Les géants du Web (Google, Amazon, Facebook, Apple, Twitter) ont développé depuis dix ans de tels outils, permettant ainsi d'assurer un coût marginal constant d'exploitation des données, indépendamment de leur volume.

Aujourd'hui, le Big Data se caractérise par 5 « V » : volume, vitesse, variété, véracité et valeur des données exploitées. La chute des prix de stockage et l'augmentation des capacités de calcul sont à l'origine des gros Volumes et de la grande Vitesse de traitement des données. La Variété des données (images, textes, bases de données, objets connectés, etc.) est principalement due à la digitalisation croissante des supports d'information. Enfin, la Véracité des données, dont découle la valeur des travaux, constitue un enjeu central pour tout projet d'analyse automatisée des données. En effet, un algorithme est d'autant plus performant que les données sont nombreuses, exactes, et bien adaptées à la question à résoudre. Multiplier les sources et les croisements sans se soucier de la qualité des données ne peut que mener à des résultats erronés. Le développement du Big Data s'est accompagné de l'apparition des « Open Data » qui correspondent à des données générées et conservées par différents organismes et mises à la disposition des citoyens et des entreprises.

Les 5 « V » sont cependant insuffisants pour caractériser l'essence de l'innovation apportée par le Big Data. En effet, celle-ci provient avant tout de la combinaison des outils permettant de gérer ces 5 « V » avec un sous-domaine de l'intelligence artificielle dénommé « machine learning » (apprentissage automatique). Ce dernier permet de construire des algorithmes capables d'accumuler de la connaissance et de l'intelligence à partir d'expériences, sans être humainement guidés au cours de leur apprentissage, ni explicitement programmés pour gérer telle ou telle tâche particulière, d'où leur rôle central dans la chaîne de valeur de la donnée. La maîtrise de ces algorithmes est au cœur du métier de *data scientist*.

Les outils Big Data ont permis le lancement de nombreux projets médicaux fondés sur l'exploitation de données massives, à l'image de l'algorithme de « Support Vector Machine » permettant, à partir de l'analyse de 368 gènes, de discriminer les tumeurs mammaires basales de pronostic péjoratif de celles dont le pronostic est plus favorable [4]. Les *computer-aided diagnosis* (CAD), qui peuvent aider les radiologues pour l'interprétation des mammographies, en sont un autre exemple [5]. Plus récemment, le projet Senometry [6], porté conjointement par des médecins et des *data scientists*, vise à analyser pour 10 000 patientes atteintes d'un cancer du sein, et suivies pendant plusieurs décennies, des données non structurées provenant de leur histoire personnelle, de l'imagerie (scanner, IRM, mammographies, échographie, scintigraphie, imagerie par émission de positrons, etc.), de la biologie, de l'analyse anatomo-pathologique (caractéristiques tumorales, facteurs prédictifs et pronostiques), des thérapeutiques et de leur évolution. L'analyse croisée de ces multiples données devrait apporter un certain nombre de réponses à des questions non résolues en sénologie dans ses différents domaines (épidémiologie descriptive, analyse des facteurs de risque et protecteurs, évaluation de nouvelles pratiques, détermination plus précise du pronostic, etc.).

À la lumière des possibilités offertes par ce type de projets Big Data, le temps médical prend une autre dimension. À titre

http://dx.doi.org/10.1016/j.gyobfe.2016.05.011
1297-9589/© 2016 Elsevier Masson SAS. Tous droits réservés.

© 2018 Elsevier Masson SAS. Tous droits réservés. - Document téléchargé le 30/01/2018 par bongiorni marc (423486). Il est interdit et illégal de diffuser ce document.

d'exemple, il aura fallu presque 30 années de suivi de cohortes pour prouver que le travail de nuit constitue un facteur de risque de cancer du sein et quantifier ce risque [7]. Les technologies Big Data devraient permettre de répondre à ce type de questions en très peu de temps en analysant les données existantes, avec un impact économique important (réduction du coût des études) et une applicabilité immédiate en santé publique. Ces nouvelles technologies vont probablement accélérer les résultats de la recherche médicale et réduire l'écart qui existe aujourd'hui entre la temporalité du malade, qui a besoin de réponses immédiates et la temporalité de la recherche « classique » [8].

De manière plus globale, l'accès rapide à des jeux massifs de données pourrait changer nos paradigmes médicaux. Alors que le raisonnement médical traditionnel consistait à émettre une hypothèse puis à la vérifier sur des séries de patientes, l'arrivée du Big Data permet dans certains cas une démarche originale où la découverte de corrélations inattendues est postérieure à la récolte des données (c'est le concept de sérendipité).

Le Big Data aura probablement à l'avenir un impact fort sur la compréhension et le traitement du cancer du sein, à condition que la collecte des données et leur finalité soient dès aujourd'hui guidées par des médecins et des data scientists. À ce titre, le Big Data ne se conçoit pas sans interdisciplinarité, avec apprentissage d'une sémantique commune entre ces deux métiers. Les unités de sénologie, par leur organisation historiquement transverse avec mise en place des réunions pluridisciplinaires, constituent un cadre adapté pour de telles collaborations.

En plus des médecins et des data scientists, les projets Big Data doivent impliquer les patientes, et de manière plus générale la société civile. En effet, la question centrale qui se pose dans tout projet d'analyse des données de santé est celle du respect de la vie privée. Des techniques éprouvées d'anonymisation et de pseudonymisation au sens de la Commission nationale de l'informatique et des libertés (CNIL) existent et permettent de répondre à cet enjeu [9]. Une problématique additionnelle pour l'application des technologies Big Data aux données de santé est liée aux contraintes sur les finalités de récolte et d'utilisation de ces données. L'article 6-2 de la loi n° 78-17 du 6 janvier 1978, relative à l'informatique, aux fichiers et aux libertés, modifiée en 2004, impose que l'utilisation des données personnelles en France soit encadrée par « des finalités déterminées, explicites et légitimes » définies lors de leur récolte. Cette logique peut se heurter à celle des projets Big Data où les données sont d'abord collectées massivement sans finalité précise autre que la « recherche médicale », pour ensuite être traitées dans le cadre de cas d'usage précisés a posteriori. Cependant, ce même article élargit le champ des usages possibles des données collectées en indiquant que les données doivent simplement « ne pas être traitées ultérieurement de manière incompatible avec ces finalités », cette vérification de compatibilité avec la finalité de « recherche médicale » étant dévolue à la CNIL.

Pour conclure, il y a un réel essor des projets impliquant le Big Data en médecine. L'Ordre des médecins a d'ailleurs largement débattu de la place des médecins à l'ère des nouvelles technologies lors de son 2nd Congrès du 29 octobre 2015. Les institutions politico-économiques se sont aussi emparées de cet outil pour élargir son application, à l'image de la Commission Innovation 2030 [10] où le Big Data et la médecine individualisée figurent parmi les sept domaines avec un potentiel d'innovation majeur pour la France. À l'avenir, il est important que nos jeunes médecins s'impliquent dans cette thématique de recherche car ils seront à l'origine de la collecte des données médicales, devront contribuer à définir les cadres éthique et juridique ainsi que les objectifs des projets, et par la suite intégrer les résultats des études Big Data à leurs pratiques quotidiennes, en particulier avec les patientes.

Déclaration de liens d'intérêts

C.M. déclare faire partie de l'essai clinique Senometry (Big Data et cancer du sein) des Hôpitaux Universitaires de Strasbourg.

K.N. et I.I. sont data scientists à Quantmetry et font partie de l'essai clinique Senometry (Big Data et cancer du sein) des Hôpitaux Universitaires de Strasbourg.

Références

[1] Committee on Applied and Theoretical Statistics, Board on Mathematical Sciences, Commission on Physical Sciences, Mathematics, and Applications, National Research Council. Massive data sets: proceedings of a workshop. Washington, DC: National Academy Press; 1996.
[2] Cox M, Ellsworth D. Application-controlled demand paging for out-of-core visualization. In: Proceedings of the 8th Conference on Visualization'97; 1997.
[3] Journal officiel 2014;89:13972.
[4] Sabatier R, Finetti P, Cervera N, Lambaudie E, Esterni B, Mamessier E, et al. A gene expression signature identifies two prognostic subgroups of basal breast cancer. Breast Cancer Res Treat 2011;126(2):407–20.
[5] Dheeba J, Singh NA, Selvi ST. Computer-aided detection of breast cancer on mammograms: a swarm intelligence optimized wavelet neural network approach. J Biomed Inform 2014;49:45–52.
[6] Hacking Health Camp : créer le futur de la santé. [Available: http://www.bpifrance.fr/Vivez-Bpifrance/Actualites/Hacking-Health-Camp-creer-le-futur-de-la-sante-14616].
[7] Benabu JC, Stoll F, Gonzalez M, Mathelin C. Travail de nuit, travail posté : facteur de risque du cancer du sein ? Gynecol Obstet Fertil 2015;43(12):791–9.
[8] Shrager JM, Tenenbaum JM. Rapid learning for precision oncology. Nat Rev Clin Oncol 2014;11(2):109–18.
[9] Commission Nationale de l'Informatique et des Libertés. La sécurité des données personnelles. Les guides de la CNIL; 2010, https://www.cnil.fr/sites/default/files/typo/document/Guide_securite-VD.pdf.
[10] Commission Innovation 2030. Un principe et sept ambitions pour l'innovation; 2013, http://www.ladocumentationfrancaise.fr/var/storage/rapports-publics/134000682.pdf.

C. Mathelin (MD, PhD) *
Unité de sénologie, hôpital de Hautepierre, hôpitaux universitaires de Strasbourg, 1, avenue Molière, 67098 Strasbourg cedex 09, France

K. Neuberger, I. Ibnouhsein
Quantmetry, Data Science Consulting, 55, rue La Boétie, 75008 Paris, France

*Auteur correspondant
Adresse e-mail : Carole.Mathelin@chru-strasbourg.fr (C. Mathelin)

Reçu le 6 mars 2016
Disponible sur Internet le 28 juin 2016

© 2018 Elsevier Masson SAS. Tous droits réservés. - Document téléchargé le 30/01/2018 par bongiorni marc (423486). Il est interdit et illégal de diffuser ce document.