PRINCIPES DE BASE DES MÉGADONNÉES

Centre de collaboration nationale des maladies infectieuses
National Collaborating Centre for Infectious Diseases

En santé publique, l'utilisation des mégadonnées pour la planification est encore émergente et les questions sont courantes. Nous commençons par des questions fondamentales : c'est quoi les « mégadonnées », qu'est-ce qui les distingue des autres données, et comment sont-elles utilisées?

LES MÉGADONNÉES, C'EST QUOI?

Les mégadonnées sont loin d'être nouvelles, mais elles continuent d'évoluer avec les progrès de la technologie. Cela a donné lieu à de nombreuses définitions différentes et à un manque de clarté conceptuelle. Une définition commune des mégadonnées est la suivante :

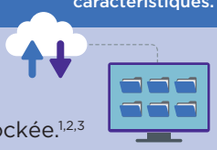
Des ensembles de données très vastes et diversifiés qui sont analysés par ordinateur à grande vitesse pour révéler des modèles, des tendances et des associations.[1,2]

QU'EST-CE QUI DISTINGUE LES MÉGADONNÉES? – LES « 3 VS »

La plupart des définitions des mégadonnées font référence à trois caractéristiques.

1 VOLUME

Une quantité de données en constante augmentation peut être collectée et stockée.[1,2,3]



Les nouvelles technologies ont permis aux chercheurs de collecter et d'analyser des ensembles de données massifs.

En 2017, plus de 2,5 quintillions d'octets de données ont été créés chaque jour, soit 300 fois plus qu'en 2005.[4]

2 VARIÉTÉ

Il existe une abondance de sources de mégadonnées, qui se répartissent en 3 catégories : [1,2,3]

Les données de santé peuvent désormais être extraites de diverses sources : L'Internet des objets, les appareils mobiles, les médias sociaux, les requêtes des moteurs de recherche ou les données du commerce électronique.[5]

STRUCTURÉES

SEMI-STRUCTURÉES

NON STRUCTURÉES

P. ex., des fichiers Excel avec des champs prédéfinis contenant des informations démographiques ou des données de localisation provenant de téléphones cellulaires.[6]

P. ex., photos, vidéos, fichiers audio, contenu des médias sociaux, sites web, images satellites, réponses à des sondages ouverts.[6]

L'intégration de plusieurs ensembles de données dans une analyse permet aux chercheurs d'enrichir leurs résultats et d'identifier de nouvelles tendances et associations.[2]

Moins de 20 % de toutes les données sont structurées et les données non structurées représentent une proportion croissante.[6]

3 VITESSE

La vitesse à laquelle les données sont générées est élevée.[1,2,3]

Les téléphones cellulaires et les dispositifs portables fournissent des données à grande vitesse, produites en continu et collectées en temps réel.[7]

Les données sont générées à des vitesses très différentes, selon la source.

COMMENT LES MÉGADONNÉES SONT-ELLES UTILISÉES?

DONNÉES À HAUT VOLUME

Une étude récente qui a analysé le contenu de 2,8 millions de tweets relatifs à la pandémie de COVID-19 aide la santé publique à comprendre les besoins d'information du public et à réagir plus rapidement et de manière plus appropriée.[8]

ENSEMBLES DE DONNÉES MULTIPLES

Les études qui rassemblent des données sur les voyages aériens commerciaux, la localisation et la mobilité des personnes, et les restrictions de voyage permettent de prévoir la propagation du COVID-19 et d'évaluer les effets des restrictions de voyage par rapport à d'autres mesures de santé publique.[9,10]

DONNÉES À HAUTE VITESSE

Une étude qui a analysé les données des téléphones cellulaires aide la santé publique à démêler la relation entre les mouvements des personnes et les facteurs socio-économiques, et à comprendre comment les restrictions de santé publique peuvent affecter différentes populations différemment et à des moments différents.[11]

RÉFÉRENCES

1. Mooney SJ, Pejaver V. « Big data in public health: terminology, machine learning, and privacy. » *Annu. Rev. Public Health.* 2018 Apr 1;39:95-112.
2. Fuller D, Buote R, Stanley K. « A glossary for big data in population and public health: discussion and commentary on terminology and research methods. » *J Epidemiol Community Health.* 2017 Nov 1;71(11):1113-7.
3. Ylijoki O, Porras J. « Perspectives to definition of big data: a mapping study and discussion. » *J. Innov. Manag.* 2016 May 4;4(1):69-91.
4. Herschel R, Miori VM. « Ethics & big data. » *Technology in Society.* 2017;49:31-36.
5. Jia Q, Guo Y, Wang G & Barnes SJ. « Big data analytics in the fight against major public health incidents (including COVID-19): a conceptual framework. » *Int. J. Environ. Res. Public Health.* 2020;17(17): 6161.
6. Marr B. « What's the difference between structured, semi-structured and unstructured data? » *Forbes*. 2019 Oct 18.
7. Badr HS, Du H, Marshall M, Dong E, Squire MM, & Gardner LM. « Association between mobility patterns and COVID-19 transmission in the USA: a mathematical modelling study. » *The Lancet Infect. Dis.*. 2020; 20(11):1247-1254.
8. Abd-Alrazaq A, Alhuwail D, Househ M, Hamdi M, Shah Z. « Top concerns of tweeters during the COVID-19 pandemic: infoveillance study. » *J.Med. Internet Res.* 2020;22(4):e19016.
9. Watts A, Au NH, Thomas-Bachli A, Forsyth J, Mayah O, Popescu S. & Bogoch II. « Potential for inter-state spread of Covid-19 from Arizona, USA: analysis of mobile device location and commercial flight data. » *J. Travel Med.* 2020, 1–3.
10. Chinazzi M, Davis JT, Ajelli M, et al. « The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. » *Science.* 2020;368(6489):395–400.
11. Long JA, Ren C. « Associations between mobility and socio-economic indicators vary across the timeline of the Covid-19 pandemic. » *Computers, environment and urban systems.* 2022 (91)1.