

Santé publique numérique, sciences des données, intelligence artificielle, données massives

Où va l'épidémiologie ?

Rodolphe Thiébaud

Université de Bordeaux, CHU de Bordeaux, Inserm U1219, Bordeaux Population Health, Inria équipe-projet SISTM, Bordeaux, France

rodolphe.thiebaud@u-bordeaux.fr

Cet article s'inscrit dans la suite du point de vue historique écrit par Marcel Goldberg pour ce numéro hors-série de *médecine/sciences*, un de ces épidémiologistes qui ont marqué ma carrière. Il s'agit ici d'une vision tout à fait subjective de l'avenir de l'épidémiologie dont certains éléments, et peut-être bientôt tous (suite à l'apprentissage issu de cet article et d'autres), peuvent se retrouver en interrogeant un grand modèle de langage (c'est-à-dire : ChatGPT d'OpenAI, Le Chat de Mistral...).

La santé publique numérique

La santé numérique est un terme très utilisé se référant aux aspects numériques des interventions pour le soin ou la prévention avec une vision assez large allant de l'utilisation des données à la télémédecine¹. La terminologie de santé publique numérique est moins utilisée. Il s'agit en fait de la transformation numérique de la santé publique. Ce terme fait référence au fait que la digitalisation du monde a conduit à une production massive de données, source d'une transformation de la recherche en santé publique et donc de l'épidémiologie. Il y a donc d'une part une digitalisation des données, par exemple les données des dossiers cliniques de patients ou les données de remboursement des médicaments. Aujourd'hui ces informations sont le plus souvent acquises d'emblée au format numérique et organisées dans des entrepôts de données. On retrouve par exemple les données de l'assurance maladie avec une partie des données hospitalières (du programme de médicalisation des systèmes d'information ou PMSI) et des causes de décès (CepiDC) dans le système national des données de santé (SNDS), couvrant la quasi-totalité de la population française. De nouveaux types de données sont également apparus, notamment avec les réseaux sociaux. Les smartphones, les montres connectées et autres capteurs (internet des objets - *internet of things*) offrent de nouvelles formes de recueil de données, parfois en continu. Les progrès biotechnologiques ont aussi révolutionné les informations disponibles en imagerie médicale du fait de la qualité et de la quantité d'images produites (radio-

logie conventionnelle, tomodensitométrie et imagerie par résonance magnétique). En biologie, il y a à la fois une production plus dense et rapide des informations (par exemple le multiplexage) et également de nouvelles mesures, notamment grâce au séquençage de l'ADN. On parle des « omiques » avec les données de génomique (ADN), transcriptomique (ARN-messager), protéomique (peptides) permettant de mesurer des dizaines de milliers de marqueurs dans un même prélèvement chez un individu, reflétant des mécanismes physiopathologiques ou les micro-organismes présents, le microbiome [1]. Dans la même idée, ces techniques sont utilisées pour les mesures de marqueurs environnementaux, l'exposome². Chacune de ces approches évolue encore, à l'instar du transcriptome mesuré maintenant au sein de cellules uniques (*single-cell*) et localisé au niveau intracellulaire (*spatial transcriptomics*). Nous sommes donc face à une véritable situation de données massives (*big data*) à la fois par le nombre de personnes concernées (disponibles dans le système national des données de santé, ou dans les entrepôts de données de santé hospitaliers, par exemple), et par le nombre de caractéristiques mesurées (données omiques). Cette situation est transformante à la fois pour la méthodologie à mettre en œuvre et pour les questions de santé publique adressables.

La science des données en santé publique, dont l'épidémiologie

L'analyse de ces nouvelles données engendre de nouveaux défis. Par exemple, l'organisation de l'information issue de centaines de logiciels

¹ Voir par exemple <https://pepr-santenum.fr/>

² Voir la série thématique « Exposome » de *m/s* : <https://www.medecinesciences.org/fr/component/toc?task=topic&id=1529>

Rodolphe Thiébaut

Professeur de santé publique à l'Université de Bordeaux et directeur du centre de recherche *Bordeaux Population Health* (Inserm U1219). Médecin de santé publique de formation, il est titulaire d'un doctorat d'université en biostatistique. Il a commencé sa carrière comme chercheur à l'Inserm en 2002, avant de devenir directeur de recherche, puis professeur des universités. Il a effectué un séjour de recherche à Londres en immunobiologie, ce qui a nourri ses travaux sur l'analyse statistique de données complexes en immunologie et infectiologie, notamment dans le domaine du VIH et des vaccins. Il a créé l'équipe SISTM (statistiques pour la médecine translationnelle), reconnue équipe-projet par l'Inria, et collabore avec des institutions de pointe telles que le *Vaccine Research Institute*. Il est également à la tête du service d'information médicale du CHU de Bordeaux, où il pilote l'entrepôt de données de santé EDS@NOVA.

Très engagé dans la formation, il a coordonné le parcours *Epidémiologie* du master de santé publique de l'Institut de santé publique, d'épidémiologie et de développement (ISPED) pendant 11 ans et a créé le parcours *Public Health Data Science* en 2018 dans le cadre de l'École universitaire de recherche en santé publique numérique de Bordeaux.



hospitaliers afin qu'elle soit exploitable nécessite des méthodes issues de l'informatique en santé (c'est-à-dire : ontologies). L'analyse de milliers de marqueurs mesurés dans une étude de quelques dizaines ou centaines d'individus confronte les chercheurs au fléau de la grande dimension. C'est pourquoi il est nécessaire de convoquer plusieurs disciplines pour réellement exploiter ces données massives. La science des données en santé publique repose sur les statistiques (plus souvent, les biostatistiques), l'informatique (bio-informatique et informatique en santé) et l'épidémiologie. L'épidémiologie couvre certes plusieurs domaines applicatifs (épidémiologie clinique, environnementale, nutritionnelle, infectieuse...) mais elle représente surtout la discipline méthodologique qui fait le lien entre la question de santé publique et les différentes méthodes à mettre en œuvre afin d'assurer une réponse valide à la question. Il s'agit de se préoccuper de l'échantillonnage, du schéma d'étude, de la validité des données aussi différentes que le séquençage ou le recueil de la qualité de vie (l'adage « *garbage in, garbage out* »³ est toujours pertinent), des critères d'évaluation... Les approches sont descriptives, prédictives ou causales [2]. L'inférence causale⁴ a progressé en tant que champ de recherche et dans sa mise en œuvre dans les études épidémiologiques. On peut citer les approches contrefactuelles imitant une randomisation grâce à la pondération des données dont l'implémentation est généralisée [3]. D'autres approches, notamment mécanistiques basées sur des systèmes d'équations différentielles, restent anecdotiques notamment du fait de la difficulté d'implémentation [4, 5]. L'inférence causale en grande dimension reste un défi de recherche majeur. Les méthodes de prédiction connaissent aussi une évolution marquée par la disponibilité des données et de nouvelles méthodes d'apprentissage machine, en particulier celles basées sur les réseaux de neurones. L'intelligence artificielle (IA) est

sans doute de loin le terme le plus utilisé pour faire référence à des méthodes voire des concepts très différents. Dans le contexte de l'analyse de données, l'IA fait souvent référence aux algorithmes basés sur les réseaux de neurones dont l'apprentissage profond s'avère très performant dans l'analyse des données d'image ou de voix notamment [6, 7]. L'IA générative et notamment les grands modèles de langage (« *large language models* ») ont plus récemment été mis en avant notamment avec ChatGPT [8]. En réalité, son apport en épidémiologie a souvent été présenté avec des résultats assez décevants lors de comparaisons avec des approches plus classiques d'apprentissage statistique ou machine [9]. Cependant, ces nouveaux algorithmes présentent l'avantage de pouvoir parcourir des espaces de données de dimension impossible à gérer avec des algorithmes conventionnels, et de retrouver des relations très complexes entre les éléments de cet espace. De plus, l'efficacité de ces algorithmes, par exemple pour les revues systématiques ou l'aide au codage de programmes pourrait transformer la pratique quotidienne des chercheurs.

La complexité de ces approches engendre de multiples défis. Tout d'abord les moyens nécessaires pour leur mise en œuvre sont beaucoup plus importants quantitativement (temps, personnes) et qualitativement (pluridisciplinarité). S'assurer de la reproductibilité et de la véracité des résultats (et donc de la disponibilité des codes et des données), pour éviter les fraudes, répond à la même problématique. Au-delà de l'augmentation des moyens nécessaires tant humains que d'infrastructure (problématique bien comprise par les géants du secteur numérique), l'enjeu est sans doute la formation tant initiale que continue des professionnels concernés. Cette formation nécessite d'être transformée et amplifiée. Elle

³ Littéralement : « poubelle à l'entrée, poubelle à la sortie ». Née dans le domaine informatique, cette expression signifie que si les données d'entrée dans un système sont de mauvaise qualité ou incorrectes, les résultats produits par ce système le seront également (ndlr).

⁴ L'inférence causale se réfère à une approche visant une relation de cause à effet entre des données observationnelles (ndlr).



doit être transformée car elle couvre à la fois plusieurs domaines scientifiques et des méthodes variées. L'exhaustivité est impossible. Elle doit être amplifiée, ce qui nécessite plus d'interactions entre les formations disciplinaires, l'épidémiologie étant par essence la discipline centrale propice à établir le lien, comme cela a été fait depuis longtemps avec les biostatistiques. À noter que la formation des formateurs est actuellement indispensable et le sera sans doute vu la progression des connaissances sur le sujet.

Les opportunités

En mesurant des informations à de multiples niveaux, tels que l'exposome environnemental, les habitudes de vie ou les biomarqueurs sanguins (c'est-à-dire : protéomique, métabolomique), l'ambition est souvent de répondre à plusieurs questions et, surtout, de comprendre l'impact de l'exposome sur la santé des populations [10]. Cette compréhension apporte des arguments de relation causale mais ouvre aussi des pistes d'intervention. Les exemples d'impact *in fine* sur la santé des populations, à l'instar des signatures génomiques pour guider le traitement du cancer du sein, ne sont pas nombreux, sans doute du fait de la complexité des données générées et de leur analyse.

L'utilisation secondaire des données est extrêmement prometteuse. Il s'agit de réutiliser des données initialement collectées à d'autres fins que la recherche. Par exemple, les données de remboursement de médicaments par l'assurance maladie ou des mutuelles, ou les données de soins issues des entrepôts de données hospitaliers ou libéraux⁵. Dans le contexte de la crise de la COVID-19, de nombreux résultats ont été générés à partir de ce type de données, que ce soit pour la mesure de l'efficacité vaccinale [11] ou de la tolérance aux vaccins [12]. La place de ces approches au regard des études plus classiques, notamment de cohortes observationnelles, reste à définir. Des cohortes créées à partir d'internet, à l'instar de Nutrinet, sont un exemple de réussite, mais l'enjeu de la qualité initiale de l'information reste un défi en vue de l'usage secondaire des données. Par exemple, la reconstitution de la temporalité des événements à partir de comptes-rendus d'hospitalisation est particulièrement difficile. Tout comme les approches multi-cohortes, les études multi-entrepôts commencent via des approches fédérées (par exemple avec les consortiums 4CE⁶ et DARWIN⁷) et l'interopérabilité entre les systèmes (par exemple avec l'OMOP⁸). L'exploitation secondaire de ces données est aussi proposée pour la surveillance épidémiologique [13] (tel que le projet ORCHI-

DÉE⁹), bien que cela soit encore un réel défi, comme l'est la surveillance à partir des données fournies par les réseaux sociaux (l'info-épidémiologie), et le fameux exemple de Google *Flu Trends*¹⁰ [14]. Les deux avantages principaux de l'usage secondaire des données sont la disponibilité potentielle en temps réel et la couverture populationnelle, encore faut-il que la gouvernance d'accès à ces données le permette. Cette transformation numérique se prête parfaitement à l'évolution de sujets de recherche tels que la santé mondiale (*Global Health*) ou le concept d'une seule santé (*One Health*), pourvu que les systèmes d'information soient également appropriés dans les pays concernés [15]. De même, le lien entre les données de santé et les données environnementales, dont les données satellitaires et météorologiques, permettra de mieux évaluer l'impact du changement climatique sur la santé.

Au-delà de la disponibilité des données, la disponibilité de nouveaux algorithmes offre de nouvelles opportunités. Par exemple, l'épidémiologie clinique pourrait être augmentée par l'IA, à l'image de l'Homme augmenté, à la fois pour le repérage des participants à inclure dans un essai, le recueil de données, jusqu'au schéma d'étude adaptatif (par exemple à base d'apprentissage par renforcement). L'épidémiologie est aussi mise à l'épreuve pour l'évaluation des interventions numériques (aide au diagnostic, interventions personnalisées) qui dépassent l'apport immédiat diagnostique et/ou clinique en s'intéressant aux modifications de l'organisation des soins, à l'amélioration de la santé des populations avec une intervention qui se modifie au cours du temps du fait de l'utilisation d'algorithmes apprenant [16]. Tout comme les approches en santé publique, ici aussi la pluridisciplinarité va au-delà de la science des données puisqu'elle convoque les sciences cognitives (notamment l'étude de l'interaction Homme-machine), les sciences humaines et sociales et les sciences politiques. À quoi bon développer un logiciel de chatbot pour orienter un patient s'il n'est pas utilisé ?

Conclusion

L'épidémiologie constitue la clé de voûte des approches pluridisciplinaires de la science des données nécessaire à la transformation de la santé publique par le

⁵ <https://www.p4dp.fr/>

⁶ Le Consortium 4CE (*consortium for characterizing COVID-19 by electronic health records*), est une collaboration internationale de chercheurs et de scientifiques dont l'objectif est d'analyser et de caractériser les données provenant des dossiers de santé électroniques (EHR) pour mieux comprendre et gérer la COVID-19. Cela inclut l'étude des facteurs de risque, des traitements efficaces, et des impacts sur la santé à long terme associés à cette maladie (ndlr).

⁷ DARWIN (*Data analysis and real world interrogation network*), est une initiative de l'agence européenne des médicaments qui centralise l'analyse de données issues de dossiers hospitaliers, dossiers de santé primaires, registres de maladies, biobanques, etc, partout en Europe, afin d'en extraire des preuves utiles pour la régulation des médicaments (ndlr).

⁸ L'OMOP (*observational medical outcomes partnership*) est une initiative de recherche visant à standardiser la manière dont les données médicales sont collectées et analysées à des fins de recherche clinique et de sécurité des médicaments (ndlr).

⁹ Le projet Orchidée (Organisation d'un réseau de centres hospitaliers impliqués dans la surveillance épidémiologique et la réponse aux urgences), coordonné par Santé publique France, a pour objectif de mettre en place une surveillance épidémiologique multi-thématiques, à partir de données provenant de l'ensemble des services des hôpitaux participant au réseau (ndlr).

¹⁰ Google Flu Trends était un projet de Google qui utilisait les données de recherche Google pour estimer et prédire les tendances de la grippe dans différentes régions. Le projet a été interrompu en 2015 en raison de critiques concernant la précision des prédictions par rapport aux données réelles de surveillance de la grippe (ndlr).

numérique. Former aux concepts fondamentaux dans une ère d'information aussi dynamique est le premier défi à relever. Les données et les algorithmes existent mais il reste à les exploiter pour véritablement impacter la santé des populations. L'exploitation des données, notamment à usage secondaire, nécessite d'y avoir accès, et donc un cadre réglementaire mais flexible. Elle nécessite aussi, et surtout, la confiance du citoyen, renforçant l'intérêt de l'épidémiologie participative, où le citoyen contribue à la construction et à la communication autour des entrepôts de données [17]. ♦

Digital public health, data science, artificial intelligence, big data: where is epidemiology heading?

REMERCIEMENTS

Cet article est signé seul mais ce point de vue est le fruit des multiples collaborations avec tous les collègues que je remercie. Plusieurs financements du programme France 2030 ont permis le soutien des activités de recherche et d'enseignement évoqués dans cet article : le programme SMATCH du PEPR Santé numérique (ANR-22-PESN-0003), l'École Universitaire de Recherche Digital Public Health (17-EURE-00193), et le programme Compétences et Métiers d'Avenir CAP-Santé numérique (ANR-23-CMA-0018).

LIENS D'INTÉRÊT

L'auteur déclare n'avoir aucun lien d'intérêt concernant les données publiées dans cet article.

RÉFÉRENCES

- Ottmann M. Ces virus qui nous habitent et qui nous visitent : le virome humain. *Med Sci (Paris)* 2022; 38 : 1028-38.
- Hernán MA, John H, and Healy B. A second chance to get causal inference right: a classification of data science tasks. *CHANCE* 2019; 32 : 42-9.
- Hernan MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *Am J Epidemiol* 2016; 183 : 758-64.
- Aalen OO, Roysland K, Gran JM, Ledergerber B. Causality, mediation and time: a dynamic viewpoint. *J R Stat Soc Ser A Stat Soc* 2012; 175 : 831-61.
- Prague M, Commenges D, Gran JM, et al. Dynamic models for estimating the effect of HAART on CD4 in observational studies: Application to the Aquitaine Cohort and the Swiss HIV Cohort Study. *Biometrics* 2017; 73 : 294-304.
- Benani A. Histoire et perspectives de l'intelligence artificielle. *Med Sci (Paris)* 2024; 40 : 283-6.
- Jean A. Une breve introduction a l'intelligence artificielle. *Med Sci (Paris)* 2020; 36 : 1059-67.
- Thiebaut R, Hejblum B, Mougin F, et al. ChatGPT and beyond with artificial intelligence (AI) in health: Lessons to be learned. *Joint Bone Spine* 2023; 90 : 105607.
- Ferté T, Dutartre D, Hejblum B, et al. Reservoir Computing for Short High-Dimensional Time Series: an Application to SARS-CoV-2 Hospitalization Forecast. *Proc Mach Learn Res* 2024 : 13570-91.
- Lefevre-Arbogast S, Chaker J, Mercier F, et al. Assessing the contribution of the chemical exposome to neurodegenerative disease. *Nat Neurosci* 2024; 27 : 812-21.
- Dagan N, Barda N, Kepten E, et al. BNT162b2 mRNA Covid-19 vaccine in a nationwide mass vaccination setting. *N Engl J Med* 2021; 384 : 1412-23.
- Semenzato L, Le Vu S, Botton J, et al. Long-term prognosis of patients with myocarditis attributed to COVID-19 mRNA vaccination, SARS-CoV-2 infection, or conventional etiologies. *Jama* 2024; 332 : 1367-77.
- Bouzellé G, Poirier C, Campillo-Gimenez B, et al. Leveraging hospital big data to monitor flu epidemics. *Comput Methods Programs Biomed* 2018; 154 : 153-60.
- 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.
- Ganser I, Thiébaut R, Buckeridge DL. Global variations in event-based surveillance for disease outbreak detection: time series analysis. *JMIR Public Health Surveill* 2022; 8 : e36211.
- Boers M, Rochereau A, Stuwe L, et al. Classification grid and evidence matrix for evaluating digital medical devices under the European union landscape. *NPJ Digit Med* 2025; 8 : 304.
- Mathieu F, Lanta M, Bungener M. L'apport des patients dans la relecture des protocoles de recherche. *Med Sci (Paris)* 2023; 39 : 558-63.



**Avec m/s, vivez en direct
les progrès et débats
de la biologie et de la médecine**

CHAQUE MOIS / AVEC LES ARTICLES DE RÉFÉRENCE DE M/S
CHAQUE JOUR / SUR WWW.MEDECINESCIENCES.ORG

Abonnez-vous sur

www.medecinesciences.org