APRÈS LE DÉLUGE DES DONNÉES DE SANTÉ

Données massives et médico-administratives

Nouveaux enjeux pour l'épidémiologie

Grégoire Rey
Directeur du CépiDc (Centre d'épidémiologie sur les causes médicales de décès)
Inserm, Le Kremlin-Bicêtre

Résumé

La production de données massives, le développement de méthodes adaptées à leur traitement et le mouvement d'ouverture et de partage des données publiques viennent modifier les métiers et les pratiques de l'épidémiologie. Le Système National des Données de Santé (SNDS) créé en 2016 est emblématique de ces enjeux. Il peut être réutilisé avec un coût limité, sans perdu de vue, sans biais de mémoire, et pour des pathologies et des expositions rares. Cependant, sa complexité peut fréquemment amener à de nombreuses erreurs d'exploitation et son chaînage avec d'autres bases de données accroît sensiblement son intérêt. Seul le développement d'outils mutualisés de documentation, d'exploitation statistique et de chaînage permettrait d'en tirer le meilleur parti. Ce sont les ambitions du Health Data Hub, créé par la loi en 2019, et dont la mission s'étend à la quasi-totalité des données personnelles de santé produites par le service public.
Mots-clés : données massives, méthodes épidémiologiques, SNDS, données administratives des demandes de remboursement des soins de santé.

Abstract

Massive and administrative health data : new challenges for epidemiology

The production of massive data, the development of methods to process them and the open data and data sharing movement are changing the professions and practices of epidemiology. The National System of Health Data (SNDS) created in 2016 is emblematic of these issues. It can be reused with a limited cost, without lost to follow up, memory bias, and for rare pathologies and exposures. However, its complexity can frequently lead to many operating errors and its linkage with other databases significantly increases its interest. Only the development of shared tools for documentation, statistical exploitation and chaining would make the best of it. These are the ambitions of the Health Data Hub. The hub was created by law in 2019, and its mission extends to almost all personal health data produced by the public service.
Keywords : *Massive data, Epidemiologic methods, SNDS, Healthcare data*

Production, exploitation et partage des données de santé

Au cours des vingt dernières années, le volume et la diversité des données disponibles pour des traitements de recherche dans le domaine des sciences de la vie et de la santé, et donc pour l'épidémiologie, a connu un accroissement sans précédent. Celui-ci a été et est toujours alimenté par :

- le développement des systèmes d'information en santé et la dématérialisation des documents administratifs en général, et dans le domaine de la santé en particulier,
- la génération de données « omiques », dont font partie les données génomiques (issues du décodage du génome), transcriptomiques (issues de la transcription du génome en ARN) et protéomiques (issues des protéines coder par le génome),
- la production de données biologiques, physiologiques et environnementales (exposome), en particulier avec l'utilisation de plus en plus généralisée de GPS et d'objets connectés.

Données massives et développement des outils informatiques et statistiques

Les progrès dans la production de données se sont accompagnés de développement des moyens de calculs et des méthodes statistiques tout aussi spectaculaires. Par exemple, l'évolution des méthodes d'analyse de l'effet des vagues de chaleur sur la mortalité en France, et plus particulièrement la comparaison de l'analyse des vagues de 1976 et de 2006, est à cet égard édifiante. Des données équivalentes étaient produites sur les deux périodes, la totalité des enregistrements provenant des certificats de décès et traitée depuis 1968 par l'Inserm, et les données de températures enregistrées par Météo France dans des stations couvrant l'ensemble du territoire. Pour autant, malgré l'observation d'une hausse de 22 % de la mortalité dans des communiqués de presse de l'Inserm en octobre 1976, dont on sait aujourd'hui qu'elle correspondait à plus de 5 000 décès en excès[1], le lien avec la chaleur n'était alors que suggéré. Trente ans plus tard, pour analyser la vague de chaleur sensiblement plus intense survenue entre le 11 et le 28 juillet 2006, il a été possible de construire et d'utiliser un modèle prédictif basé sur les températures observées pour calculer non seulement le nombre de décès en excès (environ 2200), mais également le nombre de décès en excès que l'on aurait attendus si la relation entre chaleur et mortalité était restée identique à celle d'avant le traumatisme de la canicule de 2003 (environs 6500 décès). Cette différence de 4300 décès a pu ainsi être interprétée comme une baisse de la vulnérabilité de la population française à la chaleur due notamment à la mise en place d'un système de prévention et d'alerte en 2004[2]. Ainsi, les outils

[1] Rey G, Jougla E, Fouillet A, *et al*. The impact of major heat waves on all-cause and cause-specific mortality in France from 1971 to 2003. *Int Arch Occup Env Health,* 2007, 80, 7 : 615-26.

[2] Fouillet A, Rey G, Wagner V, *et al.* Has the impact of heat waves on mortality changed in France since the European heat wave of summer 2003 ? A study of the 2006 heat wave. *Int J Epidemiol,* 2008 ; 37, 2 : 309-17.

statistiques permettant d'accroître les connaissances pour améliorer l'aide à la décision et évaluer les interventions ont considérablement progressé.

*L'*Open Data *dans le champ de la santé*

Enfin, sur les dix dernières années, le mouvement dit *Open Data* a, peu à peu, gagné le domaine de la santé. Ce courant idéologique s'inscrit dans une tendance qui considère l'information publique comme un bien commun dont la diffusion est d'intérêt public et général. Suivant cette logique dans le domaine de la santé, la transparence des informations produites doit nécessairement, in fine, donner lieu à une amélioration des décisions et de l'état de santé général. Il est largement admis que l'*Open Data* est un moyen nécessaire pour parvenir à un haut niveau de reproductibilité des résultats d'analyse, et donc d'intégrité scientifique[3]. En revanche, l'*Open Data* n'est pas suffisante pour garantir le bon usage des données, si on considère que certaines d'entre elles peuvent être biaisées ou induire des comportements contre productifs. Les Anglo-Saxons le signalent régulièrement à propos des indicateurs de qualité des soins hospitaliers[4]. Le principe vertueux de l'*Open Data* repose donc sur le présupposé que tout indicateur couramment utilisé fait nécessairement l'objet d'un débat public raisonnable et intelligible, guidant spontanément à une amélioration des outils et méthodes. Sur un plan plus économique, l'*Open Data* est un excellent moyen de limiter le gâchis des données, bien souvent très coûteuses à produire et insuffisamment utilisées par les producteurs eux-mêmes. L'accélération de l'innovation dans les usages des données est de plus vue comme un facteur de croissance économique important.

En pratique, en France, dans le champ de la santé, ce mouvement a donné lieu à la création d'une commission *Open Data* en santé, dont les conclusions dans son rapport de 2014[5] ont été à l'origine de l'article 193 de la loi de modernisation de notre système de santé. Cette loi crée le système national des données de santé (SNDS), regroupement et chaînages de bases de données de santé fondées sur des procédures administratives, dans le but d'ouvrir l'accès à ces données pour le grand public, les administrations et la recherche, les études et les évaluations. Allant, encore plus loin, l'article 41 de la loi relative à l'organisation et à la transformation du système de santé de 2019 étend le périmètre du SNDS notamment à toutes les données recueillies dans le cadre de soins remboursés par l'assurance maladie et aux données d'enquête dans le domaine de la santé chaînées avec les données administratives de santé.

Le partage de données et la recherche académique

Dans le champ de la recherche académique, les financements accordés aux projets et l'évaluation des équipes de recherche ont fait intervenir de façon croissante le critère du partage des données. C'est notamment le cas des investissements d'avenir accordés pour les grandes cohortes, ou de l'évaluation des registres, deux cadres qui demandent aux porteurs la définition de critères ouverts et transparents pour permettre à des équipes tierces d'accéder aux données produites. Parallèlement les institutions de recherche internationales comme l'Inserm ont pris des engagements publics en ce sens[6]. Les revues scientifiques exigent de plus en plus souvent un engagement à pouvoir transmettre

[3] Research Integrity, Open Science, and Health Policy, en ligne [https://www.bmj.com/content/363/bmj.k4309/rr-0] (consulté le 11 mai 2019).

[4] Shahian DM, Iezzoni LI, Meyer GS, *et al.* Hospital-wide mortality as a quality metric: conceptual and methodological challenges. *Am J Med Qual Off J Am Coll Med Qual*, avr. 2012, 27, 2 : 112-23.

[5] Commission open data en santé : rapport, en ligne, [http://www.ladocumentationfrancaise.fr/rapports-publics/144000397/index.shtml] (consulté en juillet 2014).

[6] Walport M, Brest P. Sharing research data to improve public health. *Lancet Lond Engl*, 12 fév. 2011 ; 377 (9765) : 537-9.

les données pour reproduire les analyses, et demandent parfois directement une transmission des données. Les principes FAIR (pour *Findable*, *Accessible*, *Interoperable* et *Reusable*) guident, autant que possible, la démarche d'ouverture et de partage de ces données[7].

Ces évolutions changent les métiers de l'épidémiologie, qui historiquement consistait souvent pour une étude à ce qu'une même équipe formalise un questionnement scientifique plus ou moins vaste, élabore et mette en œuvre un protocole de recueil de données dédié et assure son analyse. L'équipe avait alors en son sein la connaissance fine de la donnée, de ces limites potentielles, et des procédures permettant sa mise en qualité. La réutilisation scientifique de données complexes et imparfaites produites par des tiers suppose la mise en œuvre de démarches méthodologiques nouvelles.

Nous donnons dans la suite de cet article une illustration de ces enjeux appliqués à l'utilisation du SNDS, dans son acception médico-administrative définie par la loi de modernisation de notre système de santé du 26 janvier 2016. Des exemples sont donnés de problématiques associées à la production des données des causes médicales de décès.

Le SNDS pour la recherche

Le Système National des Données de Santé (SNDS) regroupe et chaine les données suivantes :

- le Système National d'Information Inter-Régimes de l'Assurance Maladie (SNIIRAM) et en son sein le DCIR (Datamart Consommation Inter-Régime). Celui-ci est constitué essentiellement des données de remboursement de soins issues des consultations médicales, incluant les honoraires pratiqués, et des feuilles de soins numérisées ou dématérialisées, permettant notamment de connaître la liste des médicaments achetés et remboursés. Le système, géré par la Caisse Nationale d'Assurance Maladie (Cnam), permet également de connaître le régime d'appartenance et la présence d'une Affection de Longue Durée (ALD),
- le Programme de Médicalisation des Systèmes d'Information (PMSI) est une base de données collectée par l'Agence Technique de l'Information Hospitalière (ATIH), elle rassemble les informations concernant les diagnostics des patients, les actes prodigués lors des séjours, et leur origine et leur destination, à des fins de calcul des enveloppes financières annuelles attribuées aux établissements de santé publics et privés,
- la Base des Causes Médicales de Décès (BCMD), qui contient les diagnostics et entités nosologiques mentionnés sur les certificats de décès, collectée et traitée par l'Institut National de la Santé et de la Recherche Médicale (Inserm),
- les données médico-sociales relatives aux personnes handicapées, issues des Maisons Départementales des Personnes Handicapées (MDPH) et rassemblées par la Caisse Nationale de Solidarité pour l'Autonomie (CNSA), ces données n'étant pas encore intégralement standardisées et centralisées au moment de la rédaction de cet article.

Ces données, dès lors qu'elles sont produites, sont pseudonymisées (elles ne contiennent pas d'identifiant direct en clair) et transmises à la Cnam, organisme responsable du traitement SNDS.

L'accès à ces données est réglementé de façon précise[8]. Dans tous les cas, les accès, les extractions, les traitements et les appariements aux données du SNDS doivent se faire dans un cadre conforme à un référentiel de sécurité défini par un arrêté. Ce haut niveau de sécurité résulte de la

[7] Wilkinson MD, Dumontier M, Aalbersberg IjJ, *et al*. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*, 15 mars 2016 ; 3 : 160018.

[8] Lesaulnier F. Recherche en santé et protection des données personnelles à l'heure du Règlement général relatif à la protection des données. *Médecine Droit*, 2019, 158 : 103-11. [http://www.sciencedirect.com/science/article/pii/S124673911830112X] (consulté le 15 nov 2018).

sensibilité de données de santé recueillies en routine, couvrant la quasi-totalité de la population Française.

Outre le mouvement *Open Data*, la mise en qualité progressive de ces données[9] et le nombre croissant d'études et de recherches menées ont peu à peu mis en évidence leur fort potentiel pour la recherche en santé publique, notamment en épidémiologie[10]. En effet, contrairement aux données d'enquêtes épidémiologiques nécessitant un recueil auprès des personnes, elles ne sont soumises ni à des biais de mémoire, ni à des risques de perdus de vue. C'est pourquoi le nombre de publications portant sur le SNIIRAM a crû à un rythme très soutenu au cours des années 2010. Entre autres publications, des travaux retentissants ont été menés sur le nombre de décès associés à la prise de Médiator hors autorisation de mise sur le marché[11], sur le lien entre prise de contraceptif oraux et le risque d'embolie pulmonaire, d'AVC et d'infarctus[12], ou sur le risque d'hospitalisation et de décès associé à la prise de Baclofène pour traiter l'alcoolo-dépendance[13]. Comme les exemples donnés ci-dessus le suggèrent, de nombreuses études sur le SNIIRAM ont été menées par ou en collaboration avec la Cnam. D'autres équipes ont, peu à peu, acquis une bonne connaissance de ces données, mais malgré une documentation et des outils de communication renforcés (forums, formations et réunions régulières), le risque d'obtenir des résultats biaisés par des artefacts de production demeure élevé. Cela est d'autant plus vrai lorsque l'on considère les études portant sur les multiples composantes du SNDS.

Enjeux méthodologiques liés aux données du SNDS

Le SNDS est un système de données médico-administratives. À l'exception des données sur les causes médicales de décès, la finalité de leur production n'est pas l'exploitation statistique à des fins d'épidémiologie.

Qualité et comparabilité

Pourtant, même les données des causes médicales de décès doivent être traitées avec précaution. En effet, les données finalement en base sont d'abord tributaires de la qualité et de la comparabilité des déclarations faites par les médecins certificateurs. Or l'information déclarée dépend de la connaissance que le médecin a du cas, du niveau de précision qu'il peut déclarer, et éventuellement de la transmission de cette information à l'Inserm. Pour exemple des conséquences d'une comparabilité limitée à l'étape de la transmission de l'information, la distribution spatiale du nombre de décès par suicide, tels qu'ils sont enregistrés à l'Inserm, est biaisée par rapport à la distribution réelle. Les zones géographiques dans lesquelles la proportion de suicides est la plus élevée

[9] Tuppin P, Rudant J, Constantinou P, *et al.* Value of a national administrative database to guide public decisions : From the système national d'information interrégimes de l'Assurance Maladie (SNIIRAM) to the système national des données de santé (SNDS) in France. *Rev Epidemiol Sante Publique,* oct. 2017 ; 65 Suppl 4 : S149-67.

[10] Weill A, Rudant J, Coste J. Utilisation des données de l'assurance maladie française pour étudier l'usage et les effets des médicaments en vie réelle : revue de 216 articles publiés entre 2007 et 2016. *Rev Epidemiol Sante Publique*, 1er juin 2017 ; 65 : S127.

[11] Weill A, Païta M, Tuppin P, *et al.* Benfluorex and valvular heart disease : a cohort study of a million people with diabetes mellitus. *Pharmacoepidemiol Drug Saf*, déc. 2010 ; 19, 12 : 1256-62.

[12] Weill A, Dalichampt M, Raguideau F, *et al.* Low dose oestrogen combined oral contraception and risk of pulmonary embolism, stroke, and myocardial infarction in five million French women: cohort study. *BMJ*, 10 mai 2016 ; 353: i2002.

[13] Chaignot C, Zureik M, Rey G, Dray-Spira R, *et al*. Risk of hospitalisation and death related to baclofen for alcohol use disorders: Comparison with nalmefene, acamprosate, and naltrexone in a cohort study of 165 334 patients between 2009 and 2015 in France. *Pharmacoepidemiol Drug Saf*, nov. 2018 ; 27, 11 : 1239-48.

(Bretagne, Franche-Comté) sont les régions dans lesquelles les causes de décès sont les mieux renseignées, tandis que les zones dans lesquelles cette proportion est la plus faible (Île de France, Rhône-Alpes) sont celles dans lesquelles les Instituts Médico-Légaux renvoient le moins d'information, générant une forte proportion de données manquantes[14]. Dans ce cas, une analyse des données brutes induirait des erreurs d'appréciation, voire des prises de décision sur le ciblage d'une politique de prévention, erronée. Plus généralement, l'identification et la correction de ce type de biais nécessitent des études spécifiques guidées par la connaissance fine du processus de production de la donnée au plus près du terrain[15]. Ce questionnement typique d'une approche épidémiologique rigoureuse doit être posé régulièrement, et généralisé à l'ensemble des données du SNDS.

Documentation

La donnée sur les causes de décès est codée par l'Inserm en suivant les règles de la Classification Internationale des Maladies (CIM). Cette classification évolue avec la connaissance médicale, et il est fréquent que ces évolutions impactent la statistique finale en France et à l'étranger[16]. On constate par exemple que le passage de la neuvième à la dixième révision de la CIM est à l'origine d'une baisse de 30 % du nombre de décès ayant la pneumonie comme cause initiale. En tant qu'acteur de la mise en œuvre des règles de la CIM, le producteur des données identifie aisément l'origine de cette évolution. En revanche, la diffusion de ces données à une équipe tierce pose la question de la formation des équipes et de la documentation de ces évolutions. Cela est d'autant plus difficile pour l'Inserm s'il s'agit de faire accéder l'utilisateur à toutes les évolutions de codage, et pour la Cnam et l'ATIH s'il s'agit de documenter toutes les raisons possibles d'évolutions de remboursement ou d'applications des multitudes de règles. Les enjeux de la constitution d'une documentation structurée et d'outils pour faciliter sa consultation sont donc essentiels.

Identification de phénotype

Dans le cadre d'études épidémiologiques, le chercheur souhaite souvent pouvoir identifier la date d'incidence d'une pathologie, son stade ou sa gravité. Dans le SNDS, ces informations sont recueillies suivant différents objectifs et sont rarement identifiables de façon claire et univoque. Il est alors nécessaire de combiner les informations de différentes sources (médecine de ville et hospitalisation) pour constituer l'approximation d'une variable d'incidence avec une fiabilité acceptable. Ces combinaisons d'information sont également appelées algorithmes d'identification de phénotype. Ces derniers sont aujourd'hui définis le plus souvent à dires d'experts. Différentes initiatives visant à généraliser ou centraliser la constitution d'algorithmes ont été menées par la Cnam[17] ou le réseau Redsiam[18]. La validation et/ou l'élaboration de ces algorithmes en population générale sont idéalement obtenues en présence d'un *gold standard* obtenu par le chaînage des

[14] Richaud-Eyraud E, Rondet C, Rey G. Transmission of death certificates to CepiDc-Inserm related to suspicious deaths, in France, since 2000. *Rev Epidemiol Sante Publique*, mars 2018 ; 66 (2) : 125-133

[15] Richaud-Eyraud E, Gigonzac V, Rondet C, *et al*. État des lieux des pratiques et de la rédaction des certificats de décès par les instituts médicolégaux en France, en 2016, dans la perspective de la mise en place d'un volet complémentaire du certificat de décès. *Rev Médecine Légale*, 1er févr. 2018 ; 9, 1 : 1-9.

[16] Rey G, Aouba A, Pavillon G, *et al*. Cause-specific mortality time series analysis : a general method to detect and correct for abrupt data production changes. *Popul Health Metr*, 2011 ; 9 : 52.

[17] Cartographie des pathologies et des dépenses, [https://www.ameli.fr/l-assurance-maladie/statistiques-et-publications/etudes-en-sante-publique/cartographie-des-pathologies-et-des-depenses/index.php] (consulté le 21 nov. 2017).

[18] Goldberg M, Carton M, Doussin A, *et al*. Le réseau REDSIAM (Réseau données Sniiram) -Spécial REDSIAM. *Rev Epidemiol Sante Publique,* 1er oct. 2017 ; 65 : S144-8.

données des patients identifiés par un registre de morbidité[19] avec les données de ces patients dans le SNDS, et possiblement par l'application de méthodes d'apprentissage statistique[20].

Incomplétude des champs couverts

Malgré leur volume, certaines données essentielles à l'épidémiologie sont absentes ou très insuffisamment renseignées dans le SNDS. C'est notamment le cas de données relatives au mode de vie, au niveau socio-économique, et aux diverses expositions environnementales. Une approximation du niveau socio-économique calculé à l'échelle de la commune de domicile est régulièrement utilisée[21]. Bien qu'intéressant pour de nombreux traitements portant par exemple sur les inégalités sociospatiales de santé, cette approximation est très insuffisante pour des analyses fines. Des chaînages avec des données sociales individuelles, comme les données du recensement ou les données administratives de l'assurance vieillesse sont absolument nécessaires[22].

Dépendance des biais au protocole de l'étude

Pour toutes les raisons évoquées plus haut, il existe des écarts aussi bien sur les variables à expliquer (survenue de maladie ou état de santé) que sur les variables explicatives (soin, expositions diverses) entre les informations que l'on souhaite mesurer et les informations réellement mesurées. Ces écarts sont susceptibles de créer des biais lorsque la mesure d'une association se fait dans une finalité d'interprétation causale, comme c'est souvent le cas en épidémiologie[23]. Sans viser une qualité parfaite de l'information, inexistante dans l'absolue, il est nécessaire d'interroger ce type de biais pour les différents usages des données et en fonction des objectifs poursuivis par les études. À défaut de pouvoir éliminer ces biais, l'incertitude que ces écarts génèrent sur les estimations doit être prise en compte.

Méthodes statistiques associées

À l'international, les données médico-administratives suscitent tout autant d'intérêt. En raison de la grande quantité d'informations qu'elles peuvent chainer pour chaque individu et de la taille des populations couvertes, elles permettent de construire des populations exposées et non exposées, par exemple à la prise d'un médicament, présentant de nombreuses caractéristiques communes et considérées comme comparables, pour émuler des pseudo-interventions contrôlées[24]. Les méthodes statistiques développées à cette fin ne permettent pas de s'affranchir de toutes les précautions évoquées dans les paragraphes précédents, mais elles sont innovantes, souples et avec un minimum d'hypothèses non vérifiées, visant à tirer le meilleur parti de la grande dimension des données[25].

---

19 Défini comme un recueil continu et exhaustif de données nominatives, intéressant un ou plusieurs événements de santé, dans une population géographiquement définie.

20 S'appuyant sur des corpus de données, les plus souvent volumineux, pour optimiser avec un minimum d'hypothèses sur la structure des données la capacité prédictive d'un algorithme.

21 Rey G, Jougla E, Fouillet A, Hemon D. Ecological association between a deprivation index and mortality in France over the period 1997-2001: variations with spatial scale, degree of urbanicity, age, gender and cause of death. *BMC Public Health*, 2009 ; 9 : 33.

22 Rey G. Mesures des inégalités socio-spatiales de santé. Séminaire inégalités sociales de santé, 3 déc. 2015, DREES, Paris.

23 Hernán MA, Robins J. Causal Inference. Boca Raton: Chapman & Hall/CRC, à paraître.

24 Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *Am J Epidemiol*, 15 avr. 2016 ; 183, 8 : 758-64.

25 Van der Laan MJ, Rose S. *Targeted learning: causal inference for observational and experimental data*. New York, Springer, 2011.

Leur mise en œuvre fait appel à des techniques d'apprentissage statistique nécessitant des infrastructures informatiques de calcul hautes performances. Il est important de préciser ici que le résultat optimal obtenu par ces nouvelles méthodes peut être très insatisfaisant si les biais évoqués ci-dessus, ou d'autres que nous n'avons pas la place de discuter ici, sont trop importants. C'est toujours le rôle de l'épidémiologiste d'en évaluer la pertinence.

Enjeux organisationnels

Pour résumer les éléments permettant de capitaliser l'ouverture des données de santé, en particulier celles du SNDS, il est indispensable de disposer :

- d'une expertise très fine sur les données produites, et une capacité à documenter cette connaissance et à la faire partager en structurant l'information pour la rendre interrogeable, autonomisant ainsi les chercheurs dans leur exploitation sur une finalité donnée,
- d'outils informatiques et de gestion des accès aux données permettant de respecter la confidentialité des traitements en suivant la réglementation et les référentiels de sécurité applicables, tout en accédant à des ressources de calculs hautes performances permettant de traiter des données volumineuses avec des méthodes innovantes,
- d'un cadre de gouvernance et de réglementation simplifiée des données pour permettre notamment de réaliser des chaînages entre bases de données, et ainsi obtenir tout autant des données couvrant des champs complémentaires (ex. données sociales) que des données d'incidence faisant office de *gold standards* (ex. données de registres).

Ces trois éléments sont en étroite interaction, typiquement parce que les règles de gouvernance et la réglementation des données sont souvent associées à des principes de sécurité des traitements et d'expertise des producteurs de données. Aucune équipe ne dispose aujourd'hui à la fois de l'expertise épidémiologique et de compétences expertes sur les données produites, sur la réglementation, et en système d'information, et l'obtention pour chaque équipe d'épidémiologie de l'ensemble de ces compétences demanderait des moyens considérables. Ces constats incitent naturellement à identifier le besoin d'un cadre structurant, multidisciplinaire et mutualisé.

C'est dans ce contexte que les annonces présidentielles de mars 2018 sur la stratégie nationale en intelligence artificielle ont permis la mise en place d'un *Health Data Hub*, appelé Plateforme des Données de Santé dans la loi de 2019 relative à l'organisation et à la transformation du système de santé. Cette plateforme prendra la forme d'un Groupement d'Intérêt Public (GIP), aura pour ambition de permettre le partage et l'appariement beaucoup plus généralisé de données de santé, et proposera des outils facilitant leur exploitation dans des conditions sécurisées.

Si elles sont mises au service de la recherche en épidémiologie, s'appuyant notamment sur le savoir accumulé et les initiatives existantes, ces évolutions pourraient largement faciliter la mise en place de projets ambitieux dont les résultats seront très utiles à la santé publique.

L'auteur déclare n'avoir aucun lien d'intérêt.