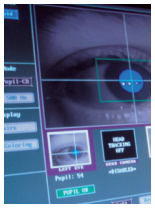
médecine/sciences 2021 ; 37 : 277-81

Le *Health Data Hub* (fin)

De multiples problèmes et des solutions alternatives ?

Marcel Goldberg, Marie Zins



UFR de médecine,
Université de Paris,
16 avenue Paul-Vaillant-Couturier,
F-94800 Villejuif, France.
marcel.goldberg@inserm.fr

Des choix qui posent de multiples problèmes

Le projet de *Health Data Hub* ne tient pas compte des ressources existantes

Les contraintes de sécurité applicables aux données de santé, notamment celles du Système national des données de santé (SNDS) et du Règlement général sur la protection des données (RGPD), impliquent d'importants moyens informatiques et organisationnels. La proposition du *Health Data Hub* (HDH) d'offrir une plateforme hautement sécurisée pour stocker et gérer les données correspond donc à un véritable besoin, car la plupart des équipes utilisatrices de données de santé n'ont pas les ressources nécessaires pour assurer cette protection. On peut cependant s'interroger sur la nécessité de construire *de novo* un tel dispositif, forcément très coûteux. Il existe en effet une offre publique qui accueille déjà de nombreuses bases de données, dont des bases de données de santé, et qui répond tout à fait à ce type de besoin : le CASD (Centre d'accès sécurisé aux données), un Groupement d'intérêt public (GIP) rassemblant l'État, représenté par l'Institut national de la statistique et des études économiques (Insee), le Groupe des écoles nationales d'économie et statistique (GENES), le CNRS, l'École polytechnique et l'École des hautes études commerciales de Paris (HEC Paris) [1]. Le CASD offre des services de « bulles sécurisées », dédiées pour l'hébergement et le traitement de données sensibles par des *data scientists* et des chercheurs dûment autorisés et authentifiés. Ce GIP gère depuis plusieurs années des accès sécurisés aux données confidentielles de l'Insee, aux données fiscales, ainsi que, de plus en plus, à des données de santé (PMSI [Programme de médicalisation des systèmes d'information], ainsi que plusieurs cohortes). Il héberge les données qui lui ont été confiées dans un environnement informatique contrôlé et sécurisé, situé en France, et est certifié Hébergeur de données de santé et ISO 27001 (norme internationale de sécurité de l'information). Ces certifications concernent la fourniture de services sécurisés d'infrastructure d'hébergement et de traitement de données *via* le contrôle d'accès biométrique et de connexion chiffrée depuis un boitier dédié (*SD-Box*) installé dans des établissements ayant signé un contrat avec le CASD. Ce dispositif est également homologué au référentiel de sécurité des données de santé pour l'accès aux données du SNDS. L'accès par un boitier sécurisé dédié offre l'avantage de garantir une sécurité élevée de bout en bout, évitant ainsi à l'utilisateur d'avoir à investir lourdement pour homologuer son poste de travail et son environnement (réseau et systèmes). Plus récemment, le CASD a été certifié selon la nouvelle norme ISO 27701, qui concerne la protection des données personnelles intégrant notamment les exigences du RGPD. Ces certifications ont pour objet d'apporter des garanties de sécurité aux producteurs de données et à la CNIL (Commission nationale de l'informatique et des libertés), particulièrement vigilants à ce que l'usage des données soit réalisé dans un cadre offrant un niveau de sécurité approprié. On peut signaler que d'autres acteurs, publics (comme l'Inserm) et privés, ont également des projets comparables.

Un projet qui met en danger la confidentialité des données personnelles

Le choix du HDH d'utiliser le nuage (*cloud*) AZURE de *Microsoft* pour le stockage et l'exploitation de toutes les données de santé de l'ensemble de la population

Vignette (Photo © Inserm-Michel Depardieu).

française a créé un important émoi et a mobilisé de nombreux acteurs, jusque sur les bancs du Parlement. Ce choix est inexplicable, pour des raisons évidentes de souveraineté nationale et de sécurité. Le danger potentiel majeur du choix du *cloud* de *Microsoft* provient, comme cela a été abondamment relevé, du fait de la loi américaine dite *CLOUD Act* : il s'agit d'une loi fédérale des États-Unis qui permet aux forces de l'ordre américaines de contraindre les fournisseurs de services américains à fournir les données stockées sur des serveurs, qu'ils soient situés aux États-Unis ou dans des pays étrangers, et donc d'obtenir les données personnelles d'un individu sans que celui-ci en soit informé, ni que son pays de résidence, ni que le pays où sont stockées ces données ne le soient. Pour cette raison, la Cour de Justice européenne a rendu le 16 juillet 2020, un arrêt dans lequel elle invalide le *Privacy Shield* (un dispositif qui permettait auparavant le transfert de données personnelles aux États-Unis) au motif qu'il n'offre pas en réalité, de garanties suffisantes au regard du Règlement général sur la protection des données à caractère personnel. Dans son ordonnance du 13 octobre 2020, le Conseil d'État reconnaît l'existence d'un risque de transfert de données issues du HDH vers les États-Unis. La décision du HDH de gérer le SNDS sur le *cloud* AZURE a été justifiée par l'urgence, en la prétendue absence de solution nationale, et par le fait que les données seraient cryptées de façon inviolable. Ces divers arguments sont tout à fait faux : on ne voit pas quelle est l'urgence absolue de réunir dans une plateforme de *Microsoft* des données qui sont aujourd'hui déjà accessibles et qui peuvent être mobilisées très rapidement en cas d'urgence, comme cela se produit sans difficulté majeure pendant l'actuelle épidémie de COVID-19 ; il existe, on l'a vu, une offre française publique fonctionnelle ; et tous les spécialistes de la sécurité informatique, ainsi que la CNIL, sont unanimes pour reconnaître que *Microsoft* pourra déchiffrer les données cryptées stockées dans son nuage. Ce choix est donc absolument incompréhensible, sinon par une méconnaissance des particularités des données de santé et la volonté d'un effet d'annonce. Par ailleurs, en dehors même du problème du *cloud* de *Microsoft*, la CNIL a rendu, le 29 octobre 2020, un avis extrêmement sévère sur de nombreux points du projet de décret fixant les modalités de fonctionnement du HDH concernant le SNDS [2].

Mais même s'il était géré par une infrastructure nationale, un système centralisé est non seulement inutile, comme on l'a vu, mais il est potentiellement dangereux. Il s'agit d'un choix des responsables du HDH, alors que le rapport de préfiguration du HDH privilégiait une structure en réseau avec des nœuds périphériques. Les systèmes centralisés sont les plus à risque d'attaques malveillantes, car les conséquences néfastes sont proportionnelles au nombre de personnes concernées, au caractère sensible et à l'abondance des données concernant les personnes dont les données sont enregistrées. On peut par ailleurs remarquer que l'approche centralisée choisie par le HDH est aujourd'hui techniquement largement dépassée, et que des solutions décentralisées sont de plus en plus souvent utilisées dans les situations où existent de fortes contraintes de confidentialité et de sécurité, ce qui est le cas des données de santé. Dans une configuration décentralisée, les données restent au niveau des sources, c'est-à-dire, par exemple, au niveau des hôpitaux qui les produisent. Dans cette configuration, ce ne sont pas les données qui doivent être déplacées pour être centralisées, mais les algorithmes qui « migrent » vers les sources pour s'exécuter. Dans les applications d'intelligence artificielle (IA), les algorithmes réalisent la phase d'apprentissage de façon fédérée [3]. Cette approche a de multiples avantages, dont celui d'éviter une trop forte centralisation des données, et il existe actuellement des méthodes, comme la plateforme d'analyse *DataSHIELD* [4] largement utilisée dans le cadre de consortiums internationaux de cohortes, qui permettent de respecter les contraintes de confidentialité et de sécurité des données, puisque celles-ci ne sont jamais transférées hors de leur support local.

Des attaques malveillantes contre le HDH auraient d'autant plus de gravité que le problème du risque de ré-identification des personnes n'est absolument pas résolu par leur pseudonymisation. En effet, comme nous l'avons décrit [6] (➜) à partir du moment où on peut avoir accès aux données individuelles, le croisement d'un petit nombre de données concernant certaines caractéristiques des personnes permet de les ré-identifier. Ce problème n'est pas spécifique du HDH, mais ses conséquences sont très amplifiées par le fait que la totalité de la population y serait enregistrée (on est donc sûr qu'une personne dont on veut connaître les données y est bien présente, ce qui n'est pas le cas d'une base qui ne concerne qu'un nombre limité de personnes, comme une cohorte par exemple) et que la multiplication des données de diverses sources concernant ces personnes facilitera grandement leur identification et la possibilité de conséquences néfastes pour elles.

(➜) Voir le Repères de M. Zins *et al.*, *m/s* n° 2, février 2021, page 179

Ce projet pose des problèmes de responsabilité des données et de gouvernance

Le HDH indique que, dès leur arrivée sur la plateforme technique, les données sont placées sous sa responsabilité juridique, et qu'il *« est responsable du traitement relatif à leur stockage, leur organisation et leur mise à disposition »* [5]. Cela implique que les responsables des sources de données n'auront plus la maîtrise des conditions de mise à disposition, car cette dernière sera gérée par le HDH. Ces règles établies par le HDH sont la source de divers problèmes.

• **Le consentement éclairé :** les textes nationaux et européens imposent que les personnes dont on recueille des données donnent un consentement « éclairé », c'est-à-dire qu'elles ont préalablement été informées des buts du recueil, des conditions de leur gestion et de leur utilisation. En pratique, le consentement n'est pas exigé lorsque le recueil des données a pour but la

gestion financière, notamment pour le remboursement ou la prise en charge de soins, puisque ces données sont nécessaires pour que les personnes en bénéficient. Il n'en est pas de même dans un contexte où les personnes consentent au recueil de leurs données sans qu'elles en retirent un bénéfice personnel, notamment dans le cadre de projets de recherche et de production de connaissances. L'exemple des cohortes épidémiologiques illustre bien les problèmes techniques et scientifiques, juridiques et déontologiques majeurs soulevés par les règles établies par le HDH. En effet, ces cohortes (comme d'autres dispositifs d'enquête auprès des personnes) sont constituées de personnes volontaires, qui ont accepté de fournir leurs données (parfois les plus intimes) dans le cadre d'un contrat moral et juridique avec les responsables institutionnels et scientifiques des cohortes, sur la base d'un engagement de ceux-ci concernant les objectifs scientifiques, la confidentialité, l'information sur les projets de recherche, la possibilité d'opposition et de retrait, etc. L'adhésion des volontaires implique une confiance dans le respect de ces règles de gestion et d'utilisation de leurs données. Un des éléments sur lesquels est fondée cette confiance est l'information des volontaires sur l'utilisation de leurs données, afin notamment de leur permettre de s'y opposer s'ils le souhaitent, comme, par exemple, si l'utilisateur est un organisme privé à but lucratif, comme le permettent les textes actuels. L'information vers les volontaires relève de la responsabilité des responsables institutionnels et scientifiques des cohortes, mais le fonctionnement proposé par le HDH ne permet plus à ceux-ci de maîtriser l'ensemble des décisions en lien avec l'utilisation des données confiées par les volontaires : comment les responsables des cohortes pourraient-ils leur fournir une information sérieuse permettant d'exercer leur droit d'opposition, alors qu'eux-mêmes n'ont pas été consultés sur les projets et n'en ont pas une vision claire ? Au total, la gouvernance du HDH concernant l'accès aux données ne permettra plus aux responsables des cohortes de respecter leurs engagements vis-à-vis des volontaires qui y participent et constitue une rupture de contrat qui fragilisera la nécessaire relation de confiance indispensable pour leur participation.

• **L'intérêt public :** les textes prévoient que toute personne ou structure, publique ou privée, à but lucratif ou non lucratif, peut accéder aux données du SNDS en vue de réaliser une étude, une recherche ou une évaluation présentant un « intérêt public ». Cet accès est conditionné à autorisation de la CNIL, après avis du Comité éthique et scientifique pour les recherches, les études et les évaluations dans le domaine de la santé (CESREES) qui doit juger de la qualité scientifique des demandes et de leur intérêt public. Cette notion d'intérêt public avait été introduite, en relation avec l'ouverture du SNDS à des entreprises privées à but lucratif afin d'éviter des usages abusifs des données. Un *Comité d'expertise sur l'intérêt public*, sous l'égide de l'INDS (Institut national des données de santé), avait été mis en place pour élaborer une doctrine et préciser comment ce critère devait être pris en compte dans l'instruction des demandes, sans avoir pu dégager de conclusions opérationnelles. On devine en effet les problèmes potentiels pour décider si l'intérêt financier d'une entreprise ou la recherche d'économies d'un établissement de santé privé relèvent bien de l'intérêt public…

À l'origine, cette disposition ne concernait que l'accès au SNIIRAM (Système national d'information inter-régimes de l'Assurance maladie) qui réunit des données collectées sans accord explicite des personnes ; avec l'extension du périmètre du SNDS, elle s'applique également désormais à toutes les données réunies dans le HDH, donc à d'autres sources de données constituées après un consentement éclairé des personnes pour un usage spécifique. Pour respecter le contrat passé avec les volontaires, il faut donc pouvoir les informer de façon individuelle sur les nouvelles utilisations envisagées de leurs données, sur la qualité des demandeurs, et leur laisser la maîtrise de leur propres données en s'opposant à leur utilisation le cas échéant. Or, comme on l'a vu, les procédures imposées par le HDH ne permettent pas d'apporter aux volontaires une information individuelle précise et complète.

• **L'indépendance du CESREES :** celui-ci est en charge de rendre des avis sur les projets de recherche et d'études nécessitant le recours à des données personnelles de santé, préalablement à l'autorisation de la CNIL. Le secrétariat du Comité est assuré par le HDH, qui est, par ailleurs, guichet unique pour les demandes d'accès, et responsable juridique des données. Alors que l'organisme qui auparavant jouait le rôle d'évaluation des projets (le Comité d'expertise pour les recherches, les études et les évaluations dans le domaine de la santé, ou CEERES) était rattaché au ministère chargé de la Recherche, qui assurait son secrétariat sans avoir de lien avec les organismes de gestion des données de santé, le CESREES, qui l'a remplacé en 2020, dépend maintenant du HDH, qui a la responsabilité des données de santé, et dont l'objectif explicite est d'ouvrir celles-ci le plus largement possible pour des projets d'IA, y compris à des organismes privés. Cette organisation est clairement une source potentielle de conflits d'intérêt.

• **La qualité scientifique de l'instruction des demandes et de la réalisation des projets :** l'accès aux données des cohortes et autres bases de données de santé, par des équipes publiques ou privées, est actuellement conditionné (outre les autorisations réglementaires) par une évaluation de la qualité et de l'intérêt scientifiques des projets, réalisée par une instance scientifique *ad hoc* mise en place auprès de chaque source de données (Comité d'accès, Conseil scientifique, etc.), qui a une connaissance approfondie de celle-là. À l'avenir, ce sera le HDH qui donnera le feu vert et l'accès aux données, sans qu'il soit prévu une instruction de la demande par les responsables des sources de données et leurs instances scientifiques, ni même leur information. Or, le HDH ne peut évidemment pas avoir la même connaissance intime des cohortes, des conditions du recueil de

données, de leur signification et de leurs éventuelles limites, que leurs responsables scientifiques et les personnels qui gèrent les données au quotidien. Il y aura donc inévitablement une dégradation de la qualité de l'instruction scientifique des demandes d'accès. Il en sera de même pour la mise en œuvre des projets par les chercheurs, car les données concernées sont des données construites et structurées, dont l'accès nécessite un accompagnement pour en réaliser l'extraction et permettre une mise à disposition de données pertinentes sur le plan scientifique. L'expérience montre que les chercheurs qui souhaitent utiliser les données ont impérativement besoin d'une aide pour finaliser leurs demandes de données, car les bases concernées sont complexes, ainsi que pour leur utilisation, qui requiert une expertise technique et scientifique et un savoir-faire spécifique à chaque source. Les personnels du HDH, qui ne participent pas au fonctionnement de celles-là et ne peuvent en avoir une connaissance approfondie, ne pourront pas apporter le soutien méthodologique et scientifique indispensable pour le choix des variables à extraire et des sous-populations pertinentes, pour documenter la qualité des données, les données manquantes et les données imputées, etc. Par rapport à l'existant, il y aura donc aussi une dégradation de la qualité des projets qui pourront être réalisés.

• **La gouvernance de l'accès aux données est remise en cause :** chaque base de données possède sa propre gouvernance et ses propres règles d'accès. Ainsi, toutes les cohortes ont établi des chartes fixant les règles d'ouverture des données, dictées par des normes juridiques qui n'ont pas été abrogées, ainsi que des accords de consortium entre partenaires de ces cohortes, qui fixent les règles de fonctionnement et formalisent juridiquement les modalités de la collaboration. Les règles imposées par le HDH ne tiennent pas compte des règles de gouvernance existantes, qu'elles contredisent souvent.

Pour un *Health Data Hub* réellement vertueux

Face à une monstrueuse infrastructure informatique inutile et potentiellement dangereuse…

Conçu à l'origine pour faciliter l'accès et l'utilisation des données du SNIIRAM, qui constituaient la quasi-totalité du SNDS « historique » lors de sa création, le HDH est devenu progressivement, sous la forme que lui donnent actuellement ses responsables, un monstrueux dispositif, inutile et potentiellement dangereux.

Inutile, car si le HDH a pour justification principale de permettre l'exploitation massive et simultanée de l'ensemble de bases de données, qui serait rendue possible grâce à leur réunion, le stockage permanent de données très diverses et hétéroclites au sein d'une même infrastructure informatique ne résout en rien les problèmes liés à la qualité des données et à leur interopérabilité, qui nécessite une harmonisation préalable qui ne peut se faire qu'au cas par cas, en fonction d'objectifs spécifiques d'analyse. La centralisation des bases de données diverses ne correspond à aucun besoin scientifique, et il n'y a donc aucune valeur scientifique ajoutée à stocker et gérer de façon permanente toutes les données que l'insatiable Moloch veut avaler.

Inutile aussi, puisque le CASD propose déjà une solution publique associant des institutions scientifiques majeures, offrant les mêmes prestations pour le stockage et l'exploitation de données sensibles, parfaitement sécurisée, fonctionnelle, accueillant déjà de nombreuses bases de données publiques et privées. De plus, ce GIP, qui gère ces données sur sa propre infrastructure informatique, ne présente pas les risques d'intrusion liés au stockage de données dans un nuage privé américain.

Potentiellement dangereux, car il réunit, dans une infrastructure unique, une énorme quantité d'informations d'origines diverses concernant la santé, la situation sociale et économique, les comportements, les habitudes de vie, etc., c'est-à-dire la vie la plus intime de la totalité de la population vivant en France. Quelles que soient les précautions techniques prises, il sera impossible de garantir une étanchéité totale entre les différentes sources de données concernant les mêmes personnes, puisque le but du HDH est justement de permettre de tels rapprochements de données. Même les pays les plus totalitaires ne disposent pas en permanence d'un tel ensemble de données aussi exhaustives sur leur population entière, dont les usages potentiellement néfastes peuvent avoir des conséquences incalculables, tant pour les personnes que pour la société. De plus, on l'a vu, le choix aberrant de stocker cet ensemble de données dans un *cloud* américain ne fait qu'augmenter les craintes qu'on peut avoir sur la confidentialité des données personnelles.

… pour un *Health Data Hub* vraiment utile

La principale raison qui a présidé à la décision de développer le HDH est de constituer un gigantesque entrepôt de données permettant de déployer aisément des méthodes et des applications d'intelligence artificielle (IA) gourmandes en données massives. Pour les raisons que nous avons explicitées, il s'agit d'une vision de *data scientists* ne prenant pas en compte les caractéristiques spécifiques des données de santé et leur écosystème, ni les besoins en termes de recherche biomédicale et de santé publique. Qu'il soit clair que ce que nous mettons en cause n'est pas l'utilisation de l'IA en santé, mais bien la fausse bonne solution imposée par le HDH.

Reste l'objectif de faciliter l'accès aux données de santé et leur utilisation. Il est en effet clair que la situation actuelle, caractérisée par une extrême dispersion des données de santé en de multiples systèmes d'information mis en place et gérés par de nombreux acteurs sans aucune coordination, n'est pas satisfaisante. Sans qu'il ne soit aucunement besoin de stocker toutes ces données dans une infrastructure informatique unique, le HDH pourrait jouer un rôle très utile à travers des activités et des prestations diverses. Il pourrait réaliser

une cartographie des bases de données, incluant une analyse de leur qualité et de leurs limites ; constituer et mettre à jour régulièrement un catalogue bien documenté des données disponibles ; servir de point d'entrée pour les demandes de données ; conseiller et orienter vers leurs responsables ; apporter une aide pour les démarches réglementaires ; développer et mettre à disposition des outils pour faciliter l'analyse des données du SNIIRAM, qui est une base de données très complexe et difficile à utiliser. Il pourrait aussi proposer des espaces de travail pour des analyses nécessitant de très fortes ressources informatiques en termes de volume et de performances, dans lesquels des bases de données pourraient être temporairement hébergées, le temps de réaliser les analyses. Dans cet esprit, le HDH pourrait implanter une plateforme *DataSHIELD* qui permet l'analyse centralisée de plusieurs bases de données décentralisées qui restent chacune sur son propre support informatique.

Il pourrait aussi réaliser des appariements de plusieurs sources de données. En effet, l'ouverture des bases de données administratives (SNDS, mais aussi les bases de l'Assurance vieillesse, la base fiscale, etc.) permet d'enrichir les données d'enquêtes en population, d'essais cliniques, de cohortes, etc. Ces appariements présentent diverses difficultés (notamment la sécurité des données, l'identification des personnes à apparier, la structure complexe des bases de données administratives et autres) qui expliquent qu'ils soient encore peu mis en œuvre ; là aussi, le HDH pourrait jouer un rôle important en prenant en charge ce type d'appariements, comme le fait actuellement la CNAM (Caisse nationale de l'Assurance maladie) pour l'appariement de cohortes avec le SNIIRAM. Il faut souligner que techniquement, ce type d'opération repose sur des tables de correspondance d'identifiants pseudonymisés des bases de données à apparier, ou sur des appariements probabilistes, et qu'il ne sert absolument à rien que les bases de données à apparier soient gérées de façon permanente dans le même système informatique.

Conclusion

Il faut rappeler, enfin, que le véritable problème des données de santé en France est la pauvreté des moyens disponibles pour les collecter, les valider, les gérer et les mettre à disposition, comme la crise sanitaire du Covid-19 l'a cruellement révélé : absence de toute donnée provenant des EHPAD (Établissement d'hébergement pour personnes âgées dépendantes), insuffisance du nombre de spécialistes du codage des causes de décès, pour ne citer que les manques les plus voyants. Sauf lorsqu'elles sont un sous-produit d'activités de gestion économique, comme c'est, par exemple, le cas pour les données de remboursement de soins ou de séjours hospitaliers, le financement de dispositifs, comme les registres de maladie, les cohortes épidémiologiques, les enquêtes de santé diverses, etc., qui sont censés abonder le HDH, ont des budgets presque toujours insuffisants et non pérennes. Il en est de même au niveau des organismes de régulation des données de santé, qui manquent de personnel qualifié en nombre suffisant pour l'instruction des dossiers et leur accompagnement juridique et technique.

La décision de développer le HDH selon ses orientations actuelles est une erreur stratégique fondamentale de pouvoirs publics aveuglés par la prétendue toute puissance de la technologie et le mirage du patrimoine des données de santé, qui ne demanderait qu'à être valorisé par les méthodes sophistiquées mises en œuvre par de brillants *data scientists*. Les données de santé sont particulières, ceux qui les travaillent de longue date le savent bien. Le véritable enjeu n'est pas la facilitation de leur traitement, mais il réside avant tout en amont, dans la qualité des données produites, comme le rappelle le vieil adage bien connu en informatique, GIGO : *garbage in, garbage out*[1].

Les très importants budgets alloués pour la mise en place et le fonctionnement du HDH seraient utilisés de façon beaucoup plus efficace en renforçant les moyens de recueil et de gestion de données de qualité, et en instaurant une véritable coordination des données de santé. Encore faudrait-il que l'État et les organismes concernés se dotent de structures travaillant en étroite coopération avec ceux qui collectent et gèrent les données, et ne se contentent pas de produire des textes réglementaires, et d'effets d'annonce politique totalement déconnectés de la réalité et des véritables besoins. ◊

Health Data Hub: Multiple problems and alternative solutions?

LIENS D'INTÉRÊT

Les auteurs déclarent n'avoir aucun lien d'intérêt concernant les données publiées dans cet article.

RÉFÉRENCES

1. Le Centre d'accès sécurisé aux données. https://www.casd.eu/
2. CNIL. Projet de décret relatif au SNDS. https://cdn2.nextinpact.com/medias/avis-cnil-snds.pdf
3. Zerka F, Barakat S, Walsh S, *et al.* Systematic review of privacy-preserving distributed machine learning from federated databases in health care. *JCO Clin Cancer Inform* 2020 ; 4 : 184-200.
4. Gaye A, Marcon Y, Isaevaet J, *al.* DataSHIELD: taking the analysis to the data, not the data to the analysis. *Int J Epidemiol* 2014 ; 43 : 1929-44.
5. Health Data Hub. Les engagements de la plateforme des données de santé vis-à-vis de ses partenaires responsables des données. https://www.acteurspublics.fr/upload/media/default/0001/27/bb953de13e9e7bba37a04c5df91982dc14a3d4eb.pdf
6. Zins M, Cuggia M, Goldberg M. Les données de santé en France : abondantes mais complexes. *Med Sci (Paris)* 2021 ; 37 : 179-84.

TIRÉS À PART

M. Goldberg

[1] Si les données en entrée sont de mauvaise qualité, alors le résultat, en sortie, sera de mauvaise qualité.