Données massives, *big data* et santé publique : de quoi parle-t-on ?

Quelques définitions autour de la thématique de ce dossier avant d'aborder les enjeux que présente le *big data* pour les individus et pour la santé publique.

Santé publique et *big data* : concepts et définitions

Margo Bernelin
Sonia Desmoulin
Chargées de recherche CNRS, UMR 6297 CNRS, université de Nantes
Thomas Lefèvre
Maître de conférences, praticien hospitalier, Iris-UMR 8156-997 CNRS Inserm EHESS université Sorbonne Paris Nord

Santé publique et *big data* sont deux concepts aux définitions variables selon les contextes d'utilisation.

La santé publique

Nous parlons ici de santé publique au sens du domaine qui s'intéresse à la santé des populations, comme complément à la médecine, qui s'intéresse à la santé individuelle. La santé publique présente essentiellement une dimension organisationnelle et de prévention. Bien sûr, la santé des populations est liée à la santé des individus qui les composent, mais l'approche de santé publique est collective. On peut penser aux maladies infectieuses et à la vaccination, par exemple. Une maladie infectieuse de transmission interhumaine concerne l'individu et le groupe : il peut y avoir un intérêt à créer des institutions et des réseaux, à mobiliser des techniques (vaccins) et à adopter des politiques pour tenter de maîtriser, sinon d'éradiquer, une épidémie au niveau d'une population, sans pour autant que l'intérêt individuel direct soit évident pour chacun. Une personne qui pourrait être contaminée, puis véhiculer un virus sans en subir les effets, ou en ressentir des effets mineurs, pourra participer à sa diffusion, notamment vers des personnes qui pourront être symptomatiques, voire subir des conséquences plus graves. La santé publique va ainsi recouvrir l'organisation du système de santé, des professionnels de santé et l'élaboration de politiques publiques de santé ainsi que les moyens de leur mise en œuvre. Elle implique aussi les personnes dans ces différentes dimensions, notamment selon un principe de démocratie sanitaire affirmant que les citoyens doivent pouvoir contribuer, donner un avis, sur les décisions de santé publique.

Les données

Le terme de données recèle quant à lui plusieurs sens. On peut parler de données de la science : autour de cette notion gravitent celles de données probantes, de preuves ou de faits scientifiques. On peut également parler de données au sens d'une mesure, d'une donnée « brute » – même si aucune donnée n'est jamais « brute », étant le résultat d'une construction sociotechnique. Tous ces types de données peuvent être mobilisés dans une prise de décision. Les données au sens de mesures, encore rares il y a peu, semblent se multiplier ces dernières années. La question de leur accès, de leur utilisabilité et de leur utilité se pose alors.

Big data et données massives

Le terme de *big data*, dont l'utilisation large montre qu'il ne peut être réduit à celui de données massives, ne possède pas de définition consensuelle, *a*

fortiori en santé et en santé publique. Son apparition dans la littérature scientifique médicale pourrait être rapportée à un dossier de la revue *Nature* publié en septembre 2008. Le terme est souvent attribué à un document technique d'un cabinet de conseil américain (META group/Gartner), apparu en 2001. En réalité, il n'y est pas fait mention du terme. En revanche, on y voit apparaître la base de définitions reprises assez fréquemment dans différents domaines et par la presse généraliste : les définitions en « V ». Par exemple, les 3 V : volume, variété et vélocité (en français), pour caractériser des données d'utilisation de plus en plus courante, ou que l'on souhaiterait exploiter. *Volume*, pour une grande quantité soit de caractéristiques d'un individu, soit pour un grand nombre d'individus (une cohorte de plusieurs centaines de milliers d'individus), ou bien les deux à la fois. *Variété*, pour souligner que l'on va pouvoir exploiter des données de natures diverses, que l'on peut répartir entre données structurées et données non structurées. Schématiquement, des données structurées seraient représentées par un grand tableau bien défini, où toutes les colonnes correspondent à des mesures standardisées (une tension artérielle en mmHg, une taille en cm), et les lignes, autant de personnes ou patients. Des données non structurées seraient toute autre source de données. Cela peut être de la vidéo, de la parole, du texte libre. *Vélocité*, pour le fait que l'on va traiter ces données très rapidement, en « temps réel ». Ce critère est actuellement moins pertinent dans le champ de la santé, même s'il prend de l'importance dans les situations de gestion de crise, mais il a été très important dans un domaine précurseur pour l'usage du *big data*, à savoir la finance et le *trading* « haute fréquence » (le traitement extrêmement rapide de millions et milliards de transactions financières).

Un autre domaine où le *big data* a été très vite présent est celui du marketing et de la publicité : l'idée est qu'en utilisant et recoupant des données diverses et variées de nombreuses personnes, en particulier des données dites comportementales (activité physique : les pas, le rythme auquel on marche ; consommation : ce que dit notre ticket de caisse ; utilisation de réseaux sociaux) ou des données géolocalisées (adresse personnelle, coordonnées GPS), il devient possible de cerner d'une part nos préférences et d'autre part nos caractéristiques associées à ces préférences : la publicité « ciblée » est née. Plutôt que nous soyons tous exposés sur une page Internet ou sur le bord de la route à la même publicité, des publicités pour un produit plutôt qu'un autre s'afficheront, basées sur notre historique de navigation Internet, le contenu de nos emails ou encore les caractéristiques que nous aurons nous-mêmes rentrés dans notre profil d'utilisation d'un réseau social.

Le concept a fait le voyage depuis ces domaines vers la santé. On trouvera ainsi dans la littérature scientifique médicale tant du *big data* que l'utilisation « ciblée » de données, que l'on dénomme médecine 4 P (sur le modèle des 3 V et plus) : médecine préventive, prédictive, personnalisée et participative.

Les sources de données

Les sources de données en santé se multiplient : données recueillies, numérisées lors d'une hospitalisation ou d'une consultation médicale (diagnostic, traitement, examen médical ou biologique, d'imagerie…), données de consommation de soins (données enregistrées par l'assurance maladie, prescriptions), données d'objets connectés – qu'ils soient médicaux (tensiomètre, implant cardiaque) ou non médicaux à l'origine (pèse-personne, montre connectée, smartphone), et virtuellement toute autre source de données qui pourraient, dans un usage donné, être utiles : données sociodémographiques, géographiques… On comprend qu'en réalité, plus que le caractère massif des données, ce sont deux éléments essentiels qui vont potentiellement introduire une nouveauté : i) le fait de pouvoir accéder puis recouper des données qui d'habitude ne le sont pas, et ii) des moyens d'analyses et de traitement de ces données, comme par exemple « l'intelligence artificielle ». Cependant, ces données peuvent tout autant être analysées par des méthodes plus conventionnelles en médecine et en épidémiologie.

L'intelligence artificielle

L'intelligence artificielle trouve un nouveau souffle, un regain d'intérêt, en partie par le développement de nouveaux algorithmes, mais plus encore par l'accroissement des moyens de calculs (puissance et caractère répandu des ordinateurs) et surtout par l'accessibilité accrue de données couvrant de plus en plus de domaines de la vie quotidienne, y compris la santé. Un attrait et une efficacité renouvelée de l'intelligence artificielle – qui est aussi source de réticence de la part des professionnels de santé vis-à-vis de son utilisation – tiennent dans l'utilisation d'algorithmes qui vont « s'adapter » aux données, et établir des règles de décision non pas introduites *a priori* par l'homme, mais à partir de ce qui a été observé. Cela est bien sûr le fonctionnement théorique, et implique un certain nombre de limites, dont les biais d'apprentissage : un algorithme perpétuera, voire renforcera, les comportements observés et « appris » à partir des données. Ainsi, si les données ne concernaient que des patients de sexe masculin et d'origine caucasienne, rien ne dit que l'algorithme tiré de ces données saura correctement décider pour des patients de sexe féminin ou d'origine asiatique ou afro-caribéenne.

La santé publique de précision

Ces constats sont valables pour la médecine, centrée sur l'individu, mais peuvent l'être tout autant pour la santé publique, centrée sur les populations. On parle notamment depuis quelques années (2013) de santé publique de précision. On pourrait sans doute

tout autant parler de santé publique 4 P, celle-ci ayant dans ses attributions classiques la prévention, la participation (démocratie sanitaire, associations de patients) et la prédiction. Reste la personnalisation : en réalité, en santé publique, on peut chercher comme en marketing à « segmenter » les populations, c'est-à-dire à identifier certains ensembles de personnes qui présentent des caractéristiques similaires par rapport à un problème de santé. Le concept est voisin de celui de stratification du risque : tout le monde n'est pas exposé de la même façon aux mêmes risques de développer une pathologie donnée. À l'inverse, la personnalisation « vraie » – un diagnostic strictement unique, personnel, un traitement unique, personnel – n'existe pas et n'a probablement pas beaucoup de sens en général. De fait, la médecine 4 P est elle aussi basée sur une approche de groupe : le groupe des personnes partageant des caractéristiques similaires, impliquant un même traitement.

Pour l'heure, la santé publique de précision semble se démarquer selon deux grandes dimensions de la médecine 4 P :

- la précision est une précision en termes d'échelles des mesures, par exemple l'échelle spatiale – en effet, les bons résultats d'un indicateur, comme la mortalité infantile selon les régions, peuvent masquer des hétérogénéités géographiques majeures si l'on y regarde à une plus petite échelle que la région ou le pays, et la précision révèle alors une aggravation de l'indicateur en certains lieux, et une amélioration en d'autres, donc une aggravation des inégalités géographiques de santé ;
- ce qui a toujours distingué l'approche sociale de l'approche individuelle, à savoir qu'il existe des caractéristiques propres au collectif, au social, déterminants de l'état de santé individuel et populationnel, qui ne sauraient se réduire à la somme des caractéristiques individuelles. Il ne s'agit au fond que d'une réactualisation, voire une exacerbation de la tension classique entre préférences et approches individuelles, et préférences et approches collectives.

Données massives en santé publique : quels enjeux pour les personnes ?

Margo Bernelin
Sonia Desmoulin
Chargées de recherche CNRS, UMR 6297 CNRS, université de Nantes

Au cœur de la politique de santé publique, « *l'observation épidémiologique et la surveillance de l'état de santé des populations* » s'appuient désormais « *sur les nouveaux outils d'exploitation des données* », ainsi que le proclame Santé publique France sur son site Internet. La gestion de la crise sanitaire née de la propagation du virus SARS-Cov-2 (Covid 19) illustre bien l'intérêt de la collecte et du traitement des données massives pour la protection de la santé des populations, à des fins de diagnostics, de suivi de la progression pandémique et de recherche en santé publique. L'exemple montre aussi que sont traitées à la fois des données personnelles, comme le résultat des tests de dépistage PCR, et des données non personnelles, telles que les relevés de présence du virus dans les eaux usées de grandes villes. Pour être efficace, une telle approche suppose de collecter énormément de données, de natures diverses, et issues de plusieurs bases. Caractérisée notamment par les concepts de *volume* et de *variété*, cette démarche a logiquement des répercussions sur les droits des personnes. La protection de la vie privée et des données personnelles est difficile à tenir. La tension, bien connue en santé publique, entre protection des droits et libertés individuels et protection de la population est ainsi réactivée. En effet, s'appuyant en partie sur le traitement de données personnelles, l'usage des données massives en santé publique nécessite le concours des individus au risque de leur vie privée.

L'utilisation des données personnelles de santé

L'usage des données massives en santé publique s'appuie en partie sur la collecte et le traitement de données personnelles, c'est-à-dire de données qui permettent d'identifier directement ou indirectement des individus. Ainsi, noms, adresses, identifiants et autres numéros spécifiquement attachés à une personne sont autant de données personnelles, qui peuvent être utilisées en santé publique. Au sein de ces données, les données de santé tiennent une place particulière en ce qu'elles informent sur l'état de santé passé, présent ou futur d'un individu, qu'il s'agisse de sa santé mentale ou physique. Par exemple, les résultats sanguins ou de radiographie ou les données de remboursement des soins révèlent l'état de santé d'un individu et sont, par conséquent, des données personnelles de santé. Des données plus anodines peuvent également fournir des informations sur l'état de santé, comme une adresse qui mettrait en évidence l'exposition à un environnement pollué et au risque de développer des pathologies spécifiques. Le traitement de ces données variées en masse pourrait, par exemple, mettre en lumière des liens entre la prise d'un médicament, une zone géographique de domiciliation et l'expression d'effets secondaires.

Une double difficulté naît de cette démarche. La première tient à l'impératifde protection de la vie privée des personnes. Étant susceptibles de révéler des informations sur le comportement, l'état de santé et les conditions de