considérer des critères liés à la littératie : il s'agit tout d'abord de s'assurer que le langage soit accessible et compréhensible pour tous et que la navigation soit facile et conforme à la culture de recherche de l'utilisateur ; ensuite, que l'application développe l'accessibilité du contenu par exemple en utilisant de l'audio et de la vidéo mais aussi en privilégiant les activités ludo-éducatives (« la gamification »).

Conclusion

En conclusion, les enjeux de l'usage des Ocas en prévention sont multiples. Certes, les Ocas sont intéressants car ils permettent de rendre accessible physiquement une offre d'accompagnement qui peut être inéquitable si elle reste essentiellement humaine. Certes, les différentes modalités de communication (écrit, audio, vidéo) et le caractère « gamifié » des applications peuvent être un bon levier en éducation pour la santé. Certes, enfin, le développement des Ocas en prévention s'inscrit dans l'air du temps, donnant un souffle intéressant à la prévention. En revanche, il s'agit d'être vigilant à considérer ces outils non pas comme des outils en tant que tels mais comme une stratégie d'intervention à part entière [2] et exiger que leur évaluation tout comme leur conception s'appuient sur les meilleures preuves disponibles et bénéficient d'une analyse compréhensive. Concrètement, il s'agit de comprendre comment et dans quelles conditions les Ocas sont susceptibles d'accompagner le changement de comportement dans une perspective de réduction des inégalités sociales de santé. C'est en cela qu'ils deviendront de vraies modalités innovantes en prévention.

Big data et intelligence artificielle en santé publique

Rodolphe Thiébaut
Université de Bordeaux, Institut de santé publique, d'épidémiologie et de développement (Isped), Inserm Bordeaux Population Health, Institut national de la recherche en informatique et automatique (Inria), Statistics in System Biology and Translational Medicine (SISTM), pôle de santé publique du CHU de Bordeaux

De quoi parle-t-on ?

Les données massives ou big data sont devenues omniprésentes en santé publique. Les essais cliniques qui étudient l'effet d'un traitement ou une intervention intègrent maintenant, dans l'essai principal ou dans des études ancillaires, la mesure de centaine de milliers de marqueurs grâce aux nouvelles technologies telles que le séquençage génomique ou la cytométrie de masse quantifiant les populations cellulaires. Ainsi, des millions de mesures sont effectuées chez un même individu réalisant ce que certains appellent un phénotypage profond. Les données deviennent également massives de par la taille des populations étudiées. L'accès à des entrepôts de données hospitaliers, aux bases de données de remboursement du médicament de l'assurance maladie, grâce au Système national des données de santé (SNDS), permet d'analyser des millions de patients pour répondre à une question posée. D'autres données, potentiellement très volumineuses, sont disponibles *via* Internet à l'instar des traces numériques sur les réseaux sociaux, des requêtes faites sur les moteurs de recherche… Un exemple fameux est la tentative de prédire la survenue d'épisodes d'épidémies de grippe grâce aux requêtes effectuées sur le moteur de recherche Google. Les recherches continuent aujourd'hui pour exploiter les informations disponibles sur les moteurs de recherche et les réseaux sociaux : c'est l'épidémiologie digitale.

Les big data sont classiquement définies à travers le prisme des 5 V : volume, variété, vélocité, véracité, valeur. Le volume est bien évidemment lié au caractère massif des données, en particulier quand il s'agit de données de séquençage ou d'imagerie. Le principal défi est d'ordre matériel pour le stockage et le déplacement des données. La variété fait référence aux différents types de données disponibles. Cette variété engendre une complexité et une spécificité associées à chaque type de données. Par exemple, chaque donnée biologique demande un dialogue avec un spécialiste (transcriptomique, protéomique, métabolomique…) pour bien comprendre comment elle a été générée. La vélocité fait référence à plusieurs aspects, dont la rapidité de déplacement des données et surtout la rapidité d'analyse. Dans certains cas, des volumes massifs de données arrivant en continu nécessitent d'être analysés en direct. Quelle que soit la situation, les analyses de données massives nécessitent un temps beaucoup plus important du fait de leur complexité, des temps de calcul, et parce que souvent plusieurs questions sont posées à cause de la pluralité des informations disponibles.

Le quatrième et le cinquième V ne sont pas toujours repris et pourtant représentent des aspects importants. La véracité des données (ou validité) est une problématique majeure. En effet, rien n'est possible sans une qualité des données minimale. Pourtant, en récupérant des informations obtenues à d'autres fins qu'une recherche en santé publique, la qualité d'information n'est pas toujours optimale et peut nécessiter un travail spécifique d'extraction d'information. Les entrepôts de données hospitaliers sont un exemple caricatural pour lesquels il est nécessaire de mettre en œuvre un traitement automatique de la langue (TAL) afin d'extraire les informations de textes libres issus des comptes rendus et d'organiser les connaissances à l'aide d'ontologies. Enfin, la valeur des données concerne au départ

l'impact économique de l'exploitation d'une information de qualité. Cette notion se généralise aisément dans la notion de l'impact attendu des big data.

Plus la taille des données à analyser augmente plus les outils d'analyse utilisés sont autonomes. En effet, en dimension réduite, les méthodes d'apprentissage statistique classiques, dont les modèles de régression par exemple, sont très efficientes et donnent des résultats très satisfaisants du fait de leur interprétabilité et des indicateurs estimés (c'est-à-dire le risque relatif et son intervalle de confiance). L'inconvénient de ces méthodes est qu'elles nécessitent que l'opérateur définisse le modèle (quelles variables sont intégrées, quelles formes de liens). Bien entendu, si le modèle est mal défini, cela compromet l'efficience de l'approche, qui renvoie alors des estimations biaisées, donc fausses. Dans le cadre de la grande dimension, la définition manuelle des modèles est donc difficile. Une solution se dessine avec l'apprentissage machine, qui se veut plus autonome que l'apprentissage statistique. Par exemple, une technique appelée « forêt aléatoire » permettra de prendre en compte des formes d'associations différentes entre les variables sans devoir les formuler explicitement (par exemple linéaires ou non).

Les méthodes de type réseaux de neurones constituent les méthodes les plus autonomes utilisées actuellement. Ainsi, par définition, cette autonomisation de l'analyse des données s'inscrit dans les approches d'intelligence artificielle. Par ailleurs, les méthodes d'intelligence artificielle comme l'apprentissage profond nécessitent souvent une grande quantité d'information pour la phase d'apprentissage (d'un algorithme prédictif). D'autres évolutions technologiques ont suivi l'arrivée des données massives, dont le stockage avec le *Cloud computing*, l'organisation des bases de données adaptées aux données non structurées (*Hadoop*) et la mise au point de méthodes de calcul adaptées (*MapReduce*).

Au total, la génération et l'accès à des données de plus en plus volumineuses et complexes se sont accompagnés de progrès importants dans plusieurs disciplines (informatique, statistique, épidémiologie), qui sont mobilisées ensemble pour exploiter ces nouvelles sources d'information, donnant lieu à ce qu'on appelle la science des données. L'ensemble de ces méthodes s'inscrit dans l'intelligence artificielle au sens où les algorithmes prédominent à l'analyste pour explorer ces espaces de très grande dimension.

Faut-il avoir peur ?

L'intelligence artificielle, autrement dit l'homme laissant la place à la machine, engendre des craintes immédiates de remplacement de l'intelligence humaine. Une machine remplacera-t-elle le médecin ? le statisticien ? l'épidémiologiste ? En fait, les algorithmes, bien que plus ou moins autonomes, sont efficaces sur une tâche spécifique mais sont incapables de généralisation. Un algorithme donné pourra très efficacement reconnaître des lésions cancéreuses sur une image mais il ne remplacera pas la prise en charge globale du clinicien, qui intègre bien plus d'informations. Il n'est pas exclu d'aller vers une intelligence artificielle plus générale, qui en serait capable. Cependant, elle ne remplacera pas la relation humaine médecin-patient malgré les progrès de la recherche sur l'interaction homme-machine. Le médecin est donc « augmenté », plus performant mais non remplacé. Le professionnel de santé publique est également « augmenté ». Il pourra détecter plus tôt le début des épidémies, recueillir en direct la réaction d'une population à un événement donné…

Les opportunités

Les opportunités sont en fait nombreuses. L'arrivée des données génomiques a accéléré l'objectif de personnaliser les interventions, à l'instar du traitement du cancer du sein. Pourquoi ne pas détecter très précocement le risque suicidaire à partir de la trace numérique des individus ? L'exploitation de l'ensemble des données historiques hospitalières et de ville, des données mesurées à l'entrée d'une hospitalisation d'urgence permettra-t-elle de mieux prendre en charge un patient ? Il faut reconnaître qu'aujourd'hui les espoirs sont plus importants que les résultats. Alors est-ce un leurre ?

Les enjeux

Il existe en fait de nombreux obstacles pour que les données massives et l'intelligence artificielle impactent réellement la santé publique. Sur le plan méthodologique, il y a plusieurs enjeux : la gestion et l'accès aux données, l'exploitation de ces données, l'évaluation des outils développés. Idéalement, pour répondre à une question donnée, il faudrait que toutes les données pertinentes soient disponibles, valides et accessibles. Bien entendu, le premier frein est éthique et réglementaire. N'importe qui ne peut pas accéder à n'importe quelles données et il ne le faut sans doute pas. Le récent règlement européen pour la protection des données (RGPD) donne un cadre réglementaire en Europe. Les considérations éthiques sont majeures et nécessitent d'être correctement traitées puisque, sous cette condition, elles permettront un accès potentiellement facilité. En effet, la confiance du citoyen est un élément clé pour accéder à ses données. Si on veut pouvoir réellement exploiter la valeur de ces big data, il est nécessaire de permettre au plus grand nombre de chercheurs d'y accéder afin de relever le défi de leur exploitation. C'est dans cet esprit que la ministre des Solidarités et de la Santé, Agnès Buzyn, a annoncé la création d'un *Health Data Hub* (**https://www.health-data-hub.fr**) de façon à organiser l'accès aux données de santé à l'échelle nationale.

Cependant, l'autre défi est le temps-personne nécessaire pour développer et mettre en œuvre les méthodes pertinentes à l'exploitation des données massives. Il faut plus de *data scientists* qui passent plus de temps sur les projets. Cela signifie qu'il faut avant tout former plus de professionnels, aguerris aux nouvelles technologies, à l'instar de ce que nous proposons dans

le cadre de l'Institut de santé publique, d'épidémiologie et de développement (Isped) à Bordeaux, et de l'école universitaire de recherche Digital Public Health (https://digital-public-health.u-bordeaux.fr) avec le nouveau programme de Master « Public Health Data Science ». En outre, il faut bien tenir compte de la durée nécessaire pour l'analyse des données massives afin de pouvoir les exploiter réellement. Là où trois mois équivalent temps-plein d'un statisticien était suffisant, douze mois d'un *data scientist* peuvent être nécessaires ! Il s'agit au départ d'une considération pragmatique, mais dont l'implication est majeure. Ne pas se donner les moyens d'analyser les données massives comme il se doit peut compromettre l'intégrité scientifique du projet.

Enfin, les considérations matérielles sont également de mise. Où stocker les données ? dans le Cloud d'Amazon ? Les prix sont attractifs mais nous avons affaire à des données de santé et de santé publique qui nécessitent une sécurisation et une gouvernance d'accès relevant de la plus haute autorité. Si les données générées par le service public restent dans le service public, alors il faut que l'État décide d'investir à large échelle.

En conclusion, les données massives en santé publique et les développements en cours en intelligence artificielle constituent une réelle opportunité d'effet levier pour la surveillance, la prévention et l'intervention en santé publique. Cependant, l'organisation des données et de leur accès, leur exploitation et l'évaluation des approches nécessitent des moyens supplémentaires et conséquents qui permettront à la fois d'impacter la santé publique tout en assurant l'intégrité scientifique et le respect du citoyen.