lateur avait ainsi largement entendu confier les missions de prévention, de surveillance de la population à des agences sanitaires publiques, à des réseaux publics ou encore à des organismes, bien que de droit privé, à but non lucratif tels que les mutuelles. Cependant, l'écosystème autour du traitement des données massives a conduit à faire émerger la possibilité d'associer un plus grand nombre d'acteurs privés à la santé publique. Cette attraction vers le secteur privé a été mise en évidence par les discussions parlementaires de 2019 sur la création de la plateforme nationale des données de santé. La question posée était la suivante : cette plateforme, au rôle stratégique de pilotage de la collecte et de l'accès aux données en santé, devait-elle prendre la forme d'une structure publique ou d'une structure privée ? Certains députés avaient ainsi proposé la création d'une société par actions simplifiées (SAS), laquelle aurait pu être majoritairement détenue par l'État, garant de la finalité de la plateforme et gage de confiance pour le public. Porteuse de souplesse, une SAS devait offrir une attractivité plus importante pour la plateforme, capable de conclure des partenariats avec le public ou le privé plus rapidement qu'une institution publique. Cependant, les discussions parlementaires ont mis en lumière la crainte d'une défiance du public vis-à-vis d'une SAS dans la gestion de données personnelles aussi sensibles que celles relatives à la santé. Partant, c'est bien une structure publique qui fut privilégiée avec la création d'un *groupement d'intérêt public*.

La question de la présence d'acteurs privés au sein du champ de la santé publique fut relancée en décembre 2019 avec l'hébergement informatique des données du SNDS. Extrêmement volumineuses, les bases associées nécessitent, pour être regroupées et interrogées, d'être conservées par des hébergeurs disposant de capacités de conservation et de traitement des données considérables. Le choix s'est alors porté sur l'entreprise américaine Microsoft pour héberger ces données. Les critiques n'ont alors pas tardé, faisant valoir que les données de santé des Français devraient, pour plus de protection, être hébergées par une entreprise française ou à tout le moins européenne. Ces critiques ont également avancé que le SNDS se trouverait dépendant d'une entreprise possédant une capacité commerciale écrasante. Pourtant un tel choix n'est pas étonnant dès lors que le droit ne s'y oppose pas. En effet, le droit de la santé publique, qui encadre strictement l'hébergement des données de santé, imposant aux hébergeurs d'obtenir une certification, n'interdit pas qu'une entreprise privée, y compris étrangère, propose de tels services. Ainsi, du fait de l'introduction des données massives en santé publique, le code du même nom s'ouvre peu à peu à la présence d'acteurs privés, lesquels apparaissent comme des renforts, plus ou moins bien accueillis, de la puissance publique.

Pour conclure, la rencontre entre « données massives » et « santé publique » marque un véritable tournant pour le droit, dont les objectifs s'avèrent aménagés et réorientés vers la collecte des données et dont les acteurs se trouvent diversifiés.

Nicolas Savy
Maître de conférences à l'université Toulouse III, Institut de mathématiques de Toulouse, UMR 5219
Anne Mayère
Professeure à l'université Toulouse III, laboratoire Certop, UMR 5044, directrice adjointe de l'Iferiss
Anja Martin-Scholz
Maître de conférences à l'université Toulouse III, laboratoire Certop, UMR 5044
François Lambotte
Professeur à l'université catholique de Louvain, Institut langage et communication

La fabrique des données à l'épreuve des programmes de *big data*

Nous proposons d'interroger ici la fabrique de données dans le contexte du *big data*, en prenant exemple auprès de B. Latour [37], qui a investigué la fabrique du droit en observant la façon dont il se construit. C'est posé que, s'agissant des données, la question des finalités, du « pour quoi », n'est pas dissociable du « comment ». Les discussions autour des « données massives » laissent entendre l'avènement d'un nouveau régime de « fabrique des données » qui viendrait amplifier les précédents du fait d'évolutions techniques. Qu'est-ce qui différencie les données de santé, telles que produites selon les standards de recherche, des « données massives », qui sont au cœur de projets conséquents (Health Data Hub) ? Nous verrons que ces fabriques se différencient quant à la spécification des données, aux logiques de traitement, et à la question des « biais » inhérents à toute fabrique de données, nécessairement partielle et éventuellement partiale.

Avant d'aller plus en avant, évoquons la notion de *big data* et la différence avec le terme de données massives au cœur de ce dossier. Le premier V, volumétrie, de la règle des 5V couramment employée pour parler du *big data*, laisse entendre que les données massives y seraient incluses. Encore faut-il s'accorder sur la notion de « massives ». Selon G. Saporta [51], massive est entendue comme « trop gros pour entrer dans la machine », nécessitant une adaptation (si possible) des techniques et des modélisations statistiques usuelles. Par exemple, le volume d'information est tel qu'il est impossible de mettre en place, en une seule étape, une régression logistique. Cette technique, couramment employée en statistique, vise, pour une variable binaire (hypothèse 1 = malade/hypothèse 0 = non malade), à déterminer l'influence d'un ensemble de facteurs (l'âge, le sexe, le poids…) sur la probabilité d'observer l'hypothèse 1. Comme l'ont étudié les sociologues des sciences et

Les références entre crochets renvoient à la Bibliographie générale p. 57.

des techniques [9, 26], il n'est pas d'appellation « plus exacte » que d'autres, mais certaines qui s'imposent plus ou moins durablement sur ce terrain de luttes que constitue toute innovation éventuelle, traversé par des enjeux tant politiques, sociaux, qu'économiques et techniques.

« Stratégie pour comprendre » *versus* « stratégie pour prévoir »

Généralement, les données une fois rassemblées sont traitées au moyen de considérations statistiques. Sorti des aspects purement descriptifs, tout raisonnement statistique consiste en une confrontation des données observées sur un échantillon à une modélisation de la population dont l'échantillon est issu. On parle de modèle génératif. L. Breiman [7] distingue deux types de stratégies : la « stratégie pour comprendre » et la « stratégie pour prévoir ».

La « stratégie pour comprendre » repose sur un ensemble d'hypothèses faites sur le modèle génératif. Par exemple, on peut poser l'hypothèse d'une mortalité également distribuée entre les classes sociales ; le traitement statistique va permettre de distinguer s'il existe une différence significative de la mortalité entre les classes sociales. La « stratégie pour comprendre » consiste alors à inférer les paramètres dudit modèle à partir des données disponibles ; ainsi pourra-t-on inférer que les ouvriers présentent un risque plus élevé de mortalité précoce que les cadres supérieurs, avec des moyennes d'âge de décès distinctes. Ce type de modèle doit inclure un nombre raisonnable de variables précisées *ex ante*. La découverte d'un important « reste à expliquer » peut amener dans un second temps à revoir la spécification des variables retenues.

La « stratégie pour prévoir », quant à elle, ne cherche pas à spécifier *ex ante* le modèle génératif. Il s'agit d'une approche souvent algorithmique dont l'unique préoccupation est la précision de la prévision, c'est-à-dire la faculté de l'algorithme à retrouver la valeur « vraie » de la valeur à prévoir (à savoir, la valeur de l'objet étudié si sa mesure était possible de façon exhaustive et sans erreur). Reprenons l'exemple de la prédiction d'une maladie (oui/non) à partir d'un ensemble de variables explicatives. Une technique de *machine learning* fournira, pour un individu donné, une réponse de type maladie présente ou absente et ce sans passer par l'estimation de paramètres associés à chaque variable explicative. On est en présence d'un modèle dit « boîte noire » qui fournit – sans expliquer – une prédiction de la variable à expliquer (de façon souvent très performante d'ailleurs).

Il est à noter que ces deux approches ne sont pas forcément à mettre en opposition. Comprendre peut permettre de prévoir et prévoir peut fournir des outils de compréhension, ou du moins peut faire émerger des hypothèses pour comprendre. Cependant, la stratégie de recueil de données, les logiques de traitement ainsi que la nature des résultats obtenus sont différentes. Dès lors ces deux « fabriques » sont distinctes et non inclusives, au sens où une fabrique de données pour comprendre ne peut être une « simple extraction » d'une fabrique de données massives pour prévoir.

La fabrique des volumes de données : des logiques différenciées

L'approche « pour comprendre » est en particulier celle de la médecine fondée sur la preuve (*evidence-based medicine*). Cette démarche, devenue le standard en recherche en santé, est établie sur une production de données protocolarisée. Ce protocole recense notamment la volumétrie visée et le contour des données à recueillir. La logique consiste à travailler avec un ensemble délimité de données précisément caractérisées et tracées, voire certifiées. C'est donc la qualité des données, et ce qu'elle permet comme montée en généralité, qui est priorisée. Il s'agit de spécifier un ensemble de critères en amont de l'investigation de façon notamment à caractériser la population étudiée, les valeurs investiguées, et s'assurer que les mesures sont susceptibles d'en respecter les caractéristiques. Cela s'inscrit dans le paradigme sous-jacent des statistiques inférentielles (ensemble de méthodes consistant à généraliser à une population des conclusions tirées à partir des données d'un échantillon) afin de s'assurer que l'échantillon retenu peut apporter une connaissance fiable au regard de la question étudiée. Le protocole est établi en fonction du niveau de preuve visé. La finalité de la fabrique est une explication *causale* au phénomène étudié, c'est-à-dire s'il existe un (ou des) événements qui ont pour conséquence le phénomène étudié. La base de données qui vient l'alimenter est souvent, pour des questions scientifiques, organisationnelles et logistiques, de taille prédéfinie et en cela délimitée. Il s'agit notamment des cohortes, qui suivent un ensemble d'individus selon des critères et sur des variables précises dans une durée de plusieurs années ou décennies, ou des registres, qui recensent l'ensemble des patients atteints d'une pathologie sur un territoire donné.

La stratégie « pour prévoir » est associée à l'assemblage du plus grand nombre de données possible. Le contexte est donc celui des données massives. Ces données massives sont travaillées avec des techniques de *machine learning*. Dans cette quête de l'algorithme le plus performant en termes de prévision, il est courant d'utiliser un large spectre d'algorithmes, voire des combinaisons d'algorithmes. Les résultats s'expriment usuellement en termes de facteurs influençant la prévision. Ils mettent en lumière des *corrélations*, c'est-à-dire le degré de liaison entre deux variables, dont l'explication est très rarement causale. Les exemples de situation où une corrélation est prouvée sans qu'il y ait de relation causale sont légion. La performance attribuée à ces outils est largement liée à la quantité et à la fréquence des données ; le postulat étant que de l'accumulation de données peuvent se dégager des corrélations susceptibles de guider l'action (des pouvoirs publics, ou d'organisations marchandes).

Une formulation très différenciée des « biais de sélection » et de leur maîtrise

Quelle que soit la stratégie, on est donc potentiellement en présence d'un biais dans la conclusion lié à la sélection des patients. La « stratégie pour comprendre » est construite à partir de données protocolisées, les résultats obtenus sont conditionnels à ce protocole. Le « modèle pour prévoir » est construit à partir de données massives et les résultats sont donc eux aussi conditionnels aux données rassemblées. Il existe cependant des différences notables quant à la connaissance de ces biais et à leur prise en compte dans l'analyse et les résultats.

Dans l'approche « pour comprendre », ce biais, grâce aux définitions précises des critères d'inclusion/non-inclusion et aux éléments de design expérimental (par exemple pour une étude randomisée), est au moins réfléchi, au mieux pris en compte. La réflexion méthodologique est donc tout à fait majeure dans cette approche.

Dans l'approche « pour prévoir » avec des données massives, la performance des outils, basés sur des applications d'intelligence artificielle, repose massivement sur la volumétrie des données, qui laisse sous-entendre l'exhaustivité. Or, l'exhaustivité est très compliquée et coûteuse à obtenir, précisément parce qu'une partie de la population ou des critères recensés échappent aux formes de traçage les plus simples à systématiser. Par ailleurs, le risque est aussi de « chercher sous le lampadaire » en postulant que son « écologie » est significative de tout l'espace entre deux lampadaires. Dans cette approche, la question des biais de sélection reste omniprésente mais est plus sournoise. En effet, les modèles d'apprentissage s'enrichissent des données observées et fournissent pour celles non observées une prévision au mieux peu fiable. Ils sont donc dépendants de la qualité des données, or la réflexion méthodologique est repositionnée éventuellement en aval, voire écartée. De plus, ces masses de données sont souvent le fruit d'un processus d'assemblage de bases de données avec des échelles de mesure et des référentiels différents. On peut légitimement se poser la question de l'impact du contexte situé de production, et des hypothèses implicites relatives à la possibilité de combiner des données d'origines diverses dans ces agrégats de données.

Conclusion

Le succès grandissant des approches par « données massives » peut s'expliquer par plusieurs phénomènes. Tout d'abord, la capacité de stockage de la donnée est reléguée au second plan suite à des évolutions technologiques. Ensuite la performance des algorithmes de *machine learning* peut fonder l'espoir de résultats significatifs à venir. De plus, les discriminants usuels des « modèles pour comprendre » établis sur la statistique inférentielle deviennent vite inefficaces face à des échantillons massifs. Enfin, les promoteurs des *big data* n'ont de cesse d'annoncer de nouvelles formes de valorisation marchande de ces « gisements de données ».

Les discours portés par les promoteurs des *big data* convergent avec des logiques d'injonction à la performance et à l'efficacité dans les politiques publiques, en proposant de croiser de très nombreuses données afin d'identifier où il serait nécessaire de porter l'attention. Or les principes de traitement de données s'avèrent différenciés selon qu'il s'agit de logiques marchandes, qui peuvent se suffire d'une définition floue de la population de référence et raisonner sur des corrélations, alors que les démarches de recherche en santé publique et en épidémiologie vont requérir une relation de causalité démontrée sur une population qualifiée. Autrement dit, alors que les consommateurs d'Amazon peuvent se satisfaire de la piètre fiabilité de son algorithme de suggestion, il n'en est rien pour des décideurs en matière de politique publique s'agissant du dépistage d'une maladie.

L'exploration des différences entre les deux approches met en évidence leur éventuelle complémentarité. Cependant il s'agit de deux types de « fabriques de données » aux logiques et modalités de construction différenciées. Or les discours promotionnels des données massives laissent entendre que de grands assemblages de données pourraient servir les deux approches.

De tels discours nous semblent traduire une méconnaissance de ce que recouvre le « comment » de ces fabriques. La question est de savoir s'il importe de comprendre, de dégager des causalités vérifiées, pour guider l'action publique, ou si la priorité est au pilotage fondé sur des corrélations dont le caractère significatif et représentatif n'est pas nécessairement maîtrisé. Une question liée est le risque d'invisibilisation de toute une partie de la population, telle celle qui est concernée par les inégalités sociales de santé. Dans cette logique, certains acteurs pourraient privilégier l'accès aisé à des masses de données, comme les données collectées par les montres connectées, sans se poser la question relative à la représentativité et au profil sociodémographique des personnes en mesure d'acquérir ce type de technologie.

La pandémie de Covid-19 a remis en avant les questions liées aux données de santé et la nécessité de mieux connaître leur « fabrique » pour identifier les enjeux. Ce sont ainsi différentes configurations de données, différents principes de modélisation, des questions de biais et de maîtrise de ces biais qui sont à l'œuvre derrière la désignation homogénéisante des « données de santé ».