

# Évolution ou nouvelle donne ?

**L'usage des données issues du *big data* est-il une révolution pour la santé publique ? La diversité des sources, la réutilisation de données, l'intervention d'acteurs privés et la création d'un système national des données de santé transforment la santé publique.**

## Données dites « massives » et santé publique : une mise en perspective historique

**Joël Coste**
Université de Paris, École pratique des hautes études

Depuis sa conversion à la quantification au cours du XIX^e^ siècle, la santé publique n'a cessé de mobiliser des données dont le volume était en rapport avec les possibilités de calcul du moment. La mise en perspective historique opérée dans cet article permettra de faire la part de la rhétorique, répétitive, mettant en exergue les promesses, pour la santé publique, du recours aux données dites « massives », comme les éléments ou les enjeux plus originaux de la mobilisation de données génétiques, biologiques, comportementales ou encore de remboursements de soins que l'expression actuelle de « données massives » tend à amalgamer et dont le volume n'est probablement pas la caractéristique la plus originale ni la plus problématique.

L'expression *big data* est apparue pour la première fois au grand jour dans la revue *Nature* en septembre 2008. À l'occasion des dix ans de Google, *Nature* demanda en effet à des chercheurs et des industriels de champs différents quelles *technologies « pourraient autant changer le monde »* que Google à l'horizon 2018. Les « données massives » perçues comme susceptibles de « changer le monde » furent d'abord biologiques, génétiques et génomiques puis électro-physiologiques, avant d'inclure, dans les années qui suivirent, les dossiers médicaux, les données de remboursement de soins et enfin les données recueillies sur les réseaux sociaux ou fournies par les objets connectés.

La santé publique et, sa principale science pourvoyeuse de connaissances, l'épidémiologie, avaient vocation à s'intéresser aux *big data*. Elles exploitaient d'ailleurs depuis longtemps des *données de grande taille* sans avoir créé de nom pour celles-ci, et pour paraphraser Molière, elles les analysaient comme Monsieur Jourdain faisait de la prose.

Cet article retracera dans un premier temps l'usage par l'épidémiologie et la santé publique des données de grande taille, toujours aux limites des possibilités de calcul du moment, et rappellera les méthodes qu'elles ont contribué à développer pour leur analyse, avant de considérer dans un second temps les principaux problèmes posés par l'utilisation des *big data*, du moins des données regroupées sous ce terme depuis 2008. Certaines de ces utilisations, maîtrisées, ont déjà permis des développements pertinents, notamment en épidémiologie génétique et en pharmaco-épidémiologie. D'autres ont conduit à des échecs retentissants, et les dernières n'ont pour l'instant pas franchi l'étape de faisabilité ou de preuve du concept, malgré la com-

munication hyperbolique qui les accompagne souvent. Il sera montré qu'aux problèmes créés par la taille des données s'ajoutent des problèmes spécifiques d'autant plus ardus à résoudre que les données sont recueillies dans des champs plus éloignés de la biologie et de la médecine. Les aspects éthiques, particulièrement problématiques, de certaines utilisations proposées ne seront pas évoqués dans cet article, centré sur les questions épistémologiques.

## Des statistiques sanitaires aux grandes cohortes et au Global Burden of Diseases, les données de grande taille au service de l'épidémiologie et de la santé publique

Il est d'usage de reconnaître dans les *Natural and Political Observations Made Upon the Bills of Mortality* de John Graunt (1662) l'acte de naissance des statistiques sanitaires. Exploitant des données recueillies pendant presque soixante ans, Graunt tabulait dans cet ouvrage les naissances, mariages et décès des Londoniens des deux sexes, par année et par paroisse, ainsi que les causes de décès, analysées en détail pour plus de 229 000 d'entre eux. Un ami de Graunt, le médecin William Petty, conduisit quelques années plus tard de grandes enquêtes, fortement quantitatives, sur les conditions démographiques, économiques et sanitaires de l'Irlande, qui furent à l'origine de l'« arithmétique politique » puis de la « statistique » – de l'allemand *Statistik*, un terme forgé en 1748 pour désigner les connaissances chiffrées nécessaires à l'État.

La statistique, et notamment la statistique sanitaire, se développa progressivement en Europe au XVIII[e] siècle – la Suède, par exemple, recueillit les causes de décès sur l'ensemble de son territoire à partir de 1749 –, mais ce fut certainement l'établissement du Registrar General Office pour l'Angleterre et le Pays de Galles, créé en 1837, et son département de statistique, dirigé jusqu'en 1879 par William Farr, qui permit à la santé publique de franchir une étape décisive dans l'usage de données quantitatives. Adossé au registre d'état civil et aux recensements réguliers d'une population de plus de 20 millions d'habitants, comportant 500 000 naissances et presque autant de décès chaque année – soit des données de taille alors inédite –, Farr réalisa de nombreuses études qui montrèrent le rôle des facteurs socioéconomiques et territoriaux de la santé. Farr entreprit aussi de *standardiser* le recueil des données, à commencer par la nomenclature utilisée par les médecins, et initia par là un mouvement qui conduisit à la mise au point de classifications internationales des maladies, qui furent utilisées dans la plupart des pays du monde dès les premières décennies du XX[e] siècle. Celles-ci fournissaient des instruments indispensables pour la réalisation d'études comparatives internationales de grande ampleur, dont le dernier avatar en date, le Global Burden of Diseases (GBD), présenta pour 2016 les indicateurs de 333 pathologies et états de santé dans 195 pays et territoires.

Une autre étape importante dans l'exploitation de volumineuses données pour la santé publique fut franchie après 1945, avec le développement de l'épidémiologie dite « moderne », celle des « facteurs de risque » et des maladies chroniques [40]. De grandes cohortes de sujets furent alors assemblées et suivies de nombreuses années, dont celle, emblématique, de Framingham, commencée en 1947 et concernant plus de 5 000 personnes de cette ville. Le nombre élevé de variables d'exposition à tester dans cette cohorte (28 mentionnées dès 1949) nécessita rapidement la mise au point de techniques statistiques d'analyse multivariée (d'abord de discrimination linéaire puis de régression logistique) qui ne purent être mises en œuvre qu'avec l'aide d'*ordinateurs*. Ces derniers accompagnèrent ensuite l'épidémiologie dans ses développements et permirent l'analyse de données de cohortes dépassant 100 000 (Pays-Bas, 1986), 500 000 (États-Unis, 1995), 1 320 000 (Royaume-Uni 1996-2001), voire 6 500 000 sujets pour la Cancer Epidemiology Descriptive Cohort Database regroupant 46 cohortes américaines (2015). Des registres de maladies concernant des milliers, voire des dizaines de milliers, de sujets furent également mis en place en Amérique du Nord et en Europe à partir des années 1970, certains d'entre eux recueillant des données de suivi nombreuses sur plusieurs années.

## Les problèmes génériques créés par l'exploitation des données de grande taille en épidémiologie et santé publique

La tendance au gigantisme qui caractérise l'épidémiologie moderne, et qui s'est accélérée parallèlement à celle des capacités de calcul des ordinateurs dans les années 1990 et 2000, s'explique avant tout par la quête de la puissance statistique. Une fois les déterminants majeurs de la mortalité ou de la morbidité par cancer ou par maladies vasculaires identifiés et en partie contrôlés (tabagisme, hypercholestérolémie, hypertension artérielle, obésité), il devint nécessaire de recourir à des effectifs importants de sujets pour étudier des expositions et des manifestations de santé rares, ou encore mettre en évidence des effets faibles. Cette augmentation de puissance des analyses a eu pour contrepartie, pas toujours bien comprise, de permettre la mise en évidence de différences futiles ou bien non reproductibles en raison de l'erreur statistique de première espèce (ou « risque α ») et du phénomène d'*overfitting* (surajustement des modèles aux données). Deux autres conséquences de cette course à la puissance, cette fois liées aux conditions de recueil et d'agrégation de données volumineuses, ont été la création d'hétérogénéités artificielles, liées à des différences de recueil et surtout à l'inflation des données manquantes, concernant souvent des groupes particuliers (moins favorisés et moins alphabétisés, participant moins aux enquêtes) et responsable de biais de sélection que les modèles statistiques d'imputation des données manquantes, développées depuis les années 1980,

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



ne peuvent corriger entièrement. Ces biais peuvent être considérables, surtout dans les études portant sur des professionnels, sur des usagers des systèmes de soin ou encore chez des volontaires utilisant des technologies modernes comme Internet.

### Les problèmes spécifiques des nouvelles « données massives »

Les problèmes liés au gigantisme des données évoqués précédemment se retrouvent naturellement dans l'analyse des *big data*. Jusqu'à présent toutefois, seul le champ de la recherche génétique, en charge de l'analyse des données « omics » (génomiques, épigénomiques, transcriptomiques, protéomiques, métabolomiques, etc.) a pris la mesure de ces problèmes et instauré – au début des années 2010, dans la perspective d'applications cliniques mais aussi après quelques erreurs retentissantes –, des règles de bonne pratique assez strictes, impliquant généralement le contrôle de l'erreur de première espèce et surtout la *réplication des résultats avant leur publication* – une procédure qui avait d'ailleurs été préconisée dès les années 1990 pour la construction des échelles de risque.

Les données administratives et cliniques recueillies en routine ont aussi fait l'objet de recommandations [43] mais celles-ci concernent essentiellement la description des méthodes dans les publications et sont moins normatives des pratiques de recherche. Les intérêts et les limites des données présentes dans les bases médico-administratives (en France les bases de l'Assurance maladie et du programme de médicalisation des systèmes d'information [PMSI]) pour la recherche épidémiologique ont été souvent soulignés ces dernières années [54]. Servant à la facturation, les données sont généralement fiables mais ne fournissent que des représentations pointillistes et surtout indirectes de l'état de santé des sujets, sauf en cas de maladie sévère (avec hospitalisations répétées) ou traitée un certain temps avec des médicaments remboursés. Les erreurs de mesure non différentielles sont donc importantes, et pas toujours bien prises en compte alors qu'elles réduisent la force des associations et la puissance statistique. L'absence dans ces bases des facteurs socioéconomiques a des conséquences beaucoup plus sérieuses, puisqu'ils sont des déterminants et des facteurs de confusion de nombreux phénomènes de santé. À ce jour, ce sont surtout les études de pharmaco-épidémiologie qui ont le mieux utilisé le potentiel des bases médico-administratives, en contournant un certain nombre de difficultés posées par celles-ci, au prix toutefois d'analyses longues et complexes pour en assurer la robustesse.

L'exploitation des données hospitalières n'en est quant à elle qu'à ses premiers balbutiements, nonobstant la communication hyperbolique de certains hôpitaux qui voudraient en tirer un profit financier. L'hétérogénéité des méthodes de recueil, l'absence de contrôle de la qualité, la faible standardisation du vocabulaire et des comptes rendus médicaux constituent de redoutables difficultés que les techniques de traitement automatisé ou d'« intelligence artificielle » (IA) – mieux vaudrait parler d'*algorithmique* – ne peuvent contourner qu'après de longs paramétrages. Quant aux techniques statistiques d'exploitation de ces données (classifications automatiques, forêts aléatoires, réseaux neuronaux, scores de propension de haute dimension, etc.), même parfaitement maîtrisées et mises en œuvre, elles ne peuvent en aucun cas corriger l'absence de données importantes et le caractère hautement sélectionné des populations fréquentant les structures de soin dans des parcours eux-mêmes déterminés par de nombreux facteurs qui échappent à l'enregistrement.

Les données recueillies sur les réseaux sociaux ou fournies par les objets connectés individuels sont encore à un stade plus préliminaire d'exploitation. L'échec de Google Flu Trends à prédire avec une raisonnable validité le nombre de consultations pour grippe aux États-Unis entre 2011 et 2013 puis l'abandon final du projet – bien plus discret que son lancement – en 2015 (comme celui de son produit dérivé Google Dengue Trends) ont conduit à reprendre la réflexion sur la nature et la pertinence des données à utiliser, et aussi sur le phénomène de surajustement des données (voir plus haut). La saisonnalité des matchs de basket peut ressembler à celle de la circulation de la grippe, mais vouloir prédire la survenue de la grippe par l'augmentation des recherches sur les matchs de basket, comme l'ont fait les informaticiens de Google [38], illustre bien les limites d'une approche exclusivement *data driven*, négligeant les déterminants causaux et les mécanismes biologiques qui produisent les événements pathologiques.

### Au-delà des représentations et des perspectives hyperboliques et futuristes

Les représentations et perspectives hyperboliques et futuristes qui ont accompagné l'évocation des *big data* depuis 2008 pourraient sembler inédites. Elles furent toutefois déjà utilisées au temps des premiers usages de l'ordinateur au début des années 1960 : celui-ci allait, affirmait-on, remplacer le médecin et ses décisions hasardeuses, et l'IA allait surpasser l'intelligence humaine. On sait ce qu'il en est advenu de l'ordinateur médecin, et du programme de l'IA, réduit depuis à sa dimension « faible », mais efficace, la dimension algorithmique. Au-delà de cette rhétorique, répétitive, mettant en exergue les promesses des *big data*, restent donc de nouvelles sources de données permettant potentiellement de répondre à des questions pertinentes – les données ne posent pas de question – et dont l'analyse, si elle parvenait à être maîtrisée à l'exemple des études « omics » et de la pharmaco-épidémiologie, ne pourrait que contribuer à enrichir les connaissances épidémiologiques et à aider la décision en santé publique.

Mieux vaut bien sûr des données *grandes que petites*, mais le potentiel des données massives ne pourra être exploité que si les leçons de soixante-dix ans

*POST-SCRIPTUM*

Préparé avant la crise de la Covid-19, cet article n'a pas dû être corrigé lors de sa relecture à la fin de la première phase de celle-ci. Cette crise offrait des opportunités d'utilisation des *big data*. Toutefois, huit mois après son commencement, les données massives et l'IA n'ont apporté ni connaissance pertinente ni dispositif efficient pour la gestion de la crise, et les données n'ont éclairé la décision que dans la mesure où elles avaient été recueillies dans cet objectif et pouvaient être traitées rapidement. En France, c'est plutôt *l'absence de données disponibles* (sur l'origine géographique des sujets, l'activité des médecins généralistes, l'activité du secteur médicosocial…) et *l'impossibilité d'un traitement rapide de certaines d'entre elles* (les causes médicales de décès, un problème déjà souligné en 2003) qui ont été constatées, une nouvelle fois, lors de cette crise.

d'épidémiologie moderne sont retenues : assurer la représentativité des données ou minimiser les biais de sélection, limiter les erreurs de mesure, contrôler les phénomènes de confusion, maîtriser le « risque α » et s'assurer de la robustesse des résultats par leur réplication. De même, les conditions et la logique du recueil des données doivent être prises en compte et cela d'autant plus qu'elles ont été collectées à distance du champ biologique et médical. À l'exact opposé des annonces hyperboliques, il s'agirait donc plutôt d'une approche modeste, patiente, méticuleuse et respectueuse des données, préservée évidemment des liens d'intérêt et des utilisations mercantiles. Mal questionnées, mal analysées et imprudemment interprétées, les nouvelles données massives ne promettront que de grands échecs.

# La « santé publique de précision » : un changement de paradigme pour la santé publique ou la perte de son âme ?

**Élodie Giroux**
Maître de conférences en philosophie des sciences et de la médecine, université Jean Moulin Lyon 3, Institut de recherches philosophiques de Lyon, EA4187

La « médecine personnalisée » désigne le fait de cibler le traitement et la prévention en fonction du profil, souvent génomique, de chaque individu. Depuis 2011, l'expression de « médecine de précision » se substitue progressivement à celle de « médecine personnalisée ». Cette dernière laisserait entendre à tort qu'il s'agit de développer des traitements spécifiques pour chaque individu alors qu'elle repose plutôt sur une stratification en sous-groupes. La médecine de précision est aussi davantage liée au recueil de données massives et aux technologies associées pour leur analyse. Dès 2013, la notion de « santé publique de précision » (SPP) a été proposée comme son complément. Cependant, n'y a-t-il pas une contradiction dans les termes même, ou tout au moins une tension forte ? Si parler de santé publique « personnalisée » paraît un oxymore, lui appliquer la notion de « précision » est-il davantage pertinent ? Le propre de son action et de son efficacité n'est-il pas d'être collective : par exemple la vaccination ou la réglementation sur le port de la ceinture en voiture ? Que pourrait apporter la « précision » qui ne dénaturerait pas cette spécificité de la santé publique ? En réalité le débat sur la pertinence de promouvoir la santé publique de précision dépend à la fois de ce qu'on entend par « précision » et par « santé publique ».

## Une appellation problématique : vers une individualisation de la santé publique ?

En médecine, la précision se développe essentiellement en cancérologie, suite aux progrès de la connaissance au niveau moléculaire, eux-mêmes étroitement liés aux technologies du séquençage du génome. Décliner cette approche en santé publique conduit à prendre en compte l'hétérogénéité individuelle au niveau génomique afin de cibler les sous-populations les plus à risque. La « santé publique génomique », définie en 2012 par Cleeren et ses coauteurs [14] comme « *l'analyse de la manière dont la connaissance et les technologies basées sur le génome peuvent être intégrées dans les services de santé et la politique publique de manière responsable et efficace pour le bénéfice de la population* », est en effet le précurseur de ce qui est aujourd'hui désigné par santé publique de précision. À première vue, celle-ci semble donc renforcer le mouvement d'individualisation de la prévention. Elle s'inscrit dans la continuité de l'approche « facteurs de risque » de l'épidémiologie analytique qui se focalise sur des facteurs individuels biologiques et comportementaux, et désormais génomiques. Dans cette conception, on considère que la santé de la population est la somme des santés individuelles et qu'il est plus efficace d'agir au niveau individuel.

Or, la prise en compte des caractéristiques génomiques pour améliorer la prévention auprès des individus a-t-elle suffisamment fait ses preuves pour pouvoir être généralisée à la santé publique ? Cleeren et ses coauteurs mettent eux-mêmes en garde : « *La génétique est à double tranchant, elle peut conduire soit à renforcer, soit à réduire, les disparités de santé dans la population.* » Surtout, le propre de la santé publique n'est-il pas de repérer des facteurs de risque qui ne sont pas réductibles ou mesurables au niveau individuel ? Certains considèrent que la santé publique se caractérise avant tout par son mode collectif d'intervention et par l'analyse des causes de nature sociale, économique, environnementale et politique. C'est au niveau de la population, irréductible au niveau individuel, que se structurent les inégalités de santé dont l'analyse et la réduction sont l'un des enjeux majeurs de la santé publique. Se focaliser sur le niveau individuel risque de faire perdre ces éléments de vue.

## Améliorer la santé publique : renforcer la stratégie ciblée ou du « haut risque »

Néanmoins, la santé publique de précision accorde de l'importance au niveau populationnel : Khoury, figure clé de ce courant, critique la médecine de précision et la médecine des 4 P (médecine préventive, prédictive,

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



personnalisée et participative) pour leur négligence de la perspective populationnelle et défend l'importance d'un cinquième P (population) [35, 36]. La santé publique de précision transposerait à ce niveau le principe de la médecine de précision : « *réaliser la bonne intervention, sur la bonne population, au bon moment* ». Renforcer la stratégie qui consiste à mieux cibler les sous-populations les plus à même de bénéficier d'une intervention, dite stratégie du « haut risque » selon la terminologie de l'épidémiologiste Rose, est loin d'être inutile. Les difficultés rencontrées en termes de rapport coût-bénéfice des politiques de dépistage massif de certains cancers conduisent à défendre une stratégie visant à écarter les personnes qui n'en tireraient pas forcément un bénéfice individuel.

Dans le cadre du recueil de données massives de nature pluridimensionnelle, la génomique n'est considérée que comme un moyen parmi d'autres pour mieux identifier les populations les plus à risque. La santé publique traditionnelle utilise déjà des critères d'âge, par exemple en recommandant le dépistage de l'hépatite C dans la sous-catégorie de personnes qui sont nées entre 1945 et 1965. Dans le cadre de la surveillance de maladies infectieuses, pouvoir tracer les individus contaminés grâce aux technologies de santé connectée apparaît déterminant pour réduire l'étendue du confinement : seules les personnes ayant été en contact avec ces cas sont mises en quarantaine. Mais dans cette perspective, on peut se demander ce qu'a de nouveau la santé publique de précision, en dehors de l'introduction des technologies associées à la génomique et au recueil massif de données individuelles. Car cette double stratégie « populationnelle » et du « haut risque » existe déjà en santé publique traditionnelle.

## Les limites de la stratégie ciblée ou du « haut risque »

Ce qui pourrait néanmoins être considéré comme une évolution introduite par la santé publique de précision serait de donner la priorité à la stratégie du « haut risque ». Mais un certain nombre de présupposés se révèlent ici problématiques. Tout d'abord, on pense pouvoir réaliser des prédictions individuelles solides, c'est-à-dire extrapoler des prédictions de risque formulées au niveau de la population à des individus. Or ce n'est pas sans poser de redoutables difficultés ; et c'est en réalité au niveau de la population elle-même que ces prédictions sont le plus valides. Ensuite, on considère que ce genre de prédictions permettrait à chaque individu de modifier ses comportements. Or il a été montré que c'est loin d'être le cas. Enfin, on suppose aussi que le risque serait bien délimitable et catégorisable. Or nombre d'entre eux sont de « petits » risques continus et diffus (comme la pollution de l'air) ne permettant pas de discriminer quels individus sont le plus à risque. Ils sont pourtant ceux qui engendrent le plus grand nombre de pathologies.

C'est précisément cette difficulté à délimiter les risques au niveau individuel qui justifie, pour Rose, la centralité et la primauté de la stratégie populationnelle dans la santé publique. Elle permet d'assumer ce qu'il appelle le « paradoxe de la prévention » : un grand nombre de personnes dont le risque est faible donnent lieu à un plus grand nombre de cas de maladie qu'un petit nombre de personnes à haut risque. Elle est adaptée pour nombre de facteurs environnementaux ou sociaux dont l'effet est diffus et qui sont impliqués dans de nombreuses maladies chroniques. En outre, les facteurs de risque ciblés dans la stratégie du « haut risque » que promeut la santé publique de précision restent liés à l'individu, à sa biologie ou à son comportement. Or de nombreux travaux montrent que ces facteurs comptent pour une faible part de la variation dans le risque de maladie au niveau de la population. Les inégalités de santé sont essentiellement liées à des facteurs sociaux structurels et contextuels.

## Reconceptualiser la précision à partir de la santé publique

En fait, pour Ostald et son coauteur [45], une telle approche de la santé publique de précision constitue en réalité une médecine de précision *pour la population* mais non pas une santé publique de précision proprement dite. En effet, pour eux, la santé publique a bien pour souci premier la causalité sociale, structurelle et contextuelle des inégalités de santé. Néanmoins, les auteurs partagent avec les promoteurs de la santé publique de précision le souci de renouveler la santé publique traditionnelle, dont les stratégies populationnelles et ciblées manquent d'efficacité, en particulier pour réduire les inégalités de santé. La source de cette inefficacité résiderait dans une insuffisante prise en compte de l'hétérogénéité de la position sociale. En effet, intrinsèquement multidimensionnelle, elle est appréhendée par divers indicateurs (éducation, revenu, profession, etc.) qui ne sont pas réductibles et peuvent interagir. Il importe donc à une santé publique de précision de prendre en compte la variabilité de ces influences pour améliorer la pertinence des interventions sur les inégalités de santé. La précision est alors envisagée comme une approche plus fine de la complexité du social.

De la médecine de précision, on retrouve le souci de l'hétérogénéité pour mieux cibler les sous-groupes qui ont besoin d'une intervention et atteindre ainsi une meilleure efficacité. Mais ici le but est de mieux comprendre les mécanismes par lesquels les inégalités se structurent. Et surtout, ces sous-groupes ne sont pas alors définis par la somme des positions sociales des individus, mais à partir du contexte social dans lequel ils sont incorporés. Par ailleurs, l'importance accordée aux données massives et à leur rôle prioritaire sur la théorisation, qui caractérise souvent l'approche de précision, est ici relativisée. Le rôle des théories sociales à partir desquelles la position sociale et les différenciations produites sont abordées est central. Dès lors, est-il encore pertinent de parler de « précision » et cela ne risque-t-il pas de prêter à confusion ?

### Limites de la recherche de précision en santé publique

Bien que cette conception de la santé publique de précision soit séduisante, justifie-t-elle une refondation de la santé publique ? Mieux comprendre la complexité des mécanismes par lesquels les déterminants sociaux influent sur la santé s'inscrit dans la continuité de recherches en santé publique qui intègrent des approches systémiques. Mais surtout, cette conception se démarque de toute la littérature sur la santé publique de précision. Par suite, il semble qu'il y ait plus d'inconvénients que de bénéfices à conserver ce vocabulaire de la précision, associé aux notions d'individualisation, de stratégie ciblée et à la génomique.

Pour finir, il est important d'interroger la valeur et la pertinence d'une priorité donnée à la recherche de plus de précision pour améliorer la santé publique. L'approche de précision véhicule d'une part l'idée d'un privilège donné à la mesure quantitative, elle-même associée à celle de la supériorité des sciences naturelles sur les sciences humaines et, d'autre part, l'illusion que l'on pourrait se rapprocher d'une forme de certitude. Or l'intérêt de la santé publique ne tient-il pas à ce qu'elle est très pluridisciplinaire et qu'elle complète la biomédecine par des approches qualitatives de sciences humaines ? Surtout cette insistance sur la précision court le risque de laisser de côté les facteurs qui ne peuvent être ainsi mesurés et pour lesquels pourtant une intervention est efficace. Rose souligne que, dans le champ de la santé publique, rien ne peut jamais être certain et que la certitude ne saurait être un prérequis pour l'action. Si plus de précision c'est être avant tout attentif à l'individu, à l'hétérogénéité interindividuelle et à l'exactitude des résultats, et si c'est défendre la priorité de la connaissance sur l'action, les fondements de la santé publique sont remis en cause. Le propos n'est pas ici de faire l'apologie de l'imprécision en santé publique ni de défendre l'idée que la santé publique doit toujours privilégier le collectif sur l'individuel. Toutefois, se centrer sur l'objectif de précision est porteur d'implicites qui peuvent nuire à l'âme même de la santé publique, si on considère que la santé de la population dont elle s'occupe n'est pas réductible à la simple somme des santés des individus.

# Données massives et santé publique : entre redéfinitions et ruptures normatives

**Margo Bernelin**
Chargée de recherche CNRS, UMR 6297 CNRS, université de Nantes

La stratégie nationale de santé 2018-2022, socle politique des projets de lois en matière de santé pour le quinquennat en cours, énonce que « *le développement des innovations numériques, technologiques et organisationnelles en santé est un enjeu clé pour l'évolution des pratiques professionnelles, l'accélération du virage ambulatoire, la qualité du suivi des patients chroniques ou le partage de l'information par les acteurs du système de santé et du médico-social*[1] ». Le numérique et le traitement des données ont ainsi pris une place centrale au sein des dispositifs juridiques mis en place depuis 2018. À cet égard, il est certain que la combinaison « numérique/données » permet des avancées importantes en matière de santé publique, qu'il s'agisse du suivi de la progression d'une épidémie, de la détection de facteurs de risques associés à des pathologies ou encore de la mise en lumière d'effets secondaires de médicaments. Le recours aux données massives en santé publique doit révolutionner la matière en favorisant des gains de temps dans la recherche, en faisant émerger de nouveaux champs de recherche et de nouvelles cibles de prévention. De tels bénéfices reposent sur la collecte et la conservation des données (des vivants mais également des défunts), dans des volumes sans précédents, par des opérateurs privés ou publics. Ils nécessitent le plus souvent une mise en commun des bases de données ainsi que leur exploitation par des algorithmes. Tandis que le champ de la santé publique se trouve, à tout le moins, transformé par ces nouvelles opportunités, qu'en est-il du droit en la matière ?

### Une redéfinition des objectifs du droit de la santé publique

À l'image des transformations décrites, le droit se trouve plus que jamais orienté vers la collecte des données, les lois dernièrement votées en faisant une priorité nouvelle. En effet, jusqu'à présent, le traitement des données était pensé quasi uniquement par le prisme des données personnelles et encadré par la loi Informatique et libertés (LIL, 1978). Or, le Code de la santé publique (CSP) s'ouvre aujourd'hui à des dispositifs juridiques particuliers visant le traitement des données, qu'elles soient personnelles ou non, telles que les données d'activité des hôpitaux et les données scientifiques. Ainsi, alors qu'autrefois la question du traitement des données était rattachée aux nécessités de dénombrement, puis de vigilance (appliqué aux médicaments, aux matériaux, à la traçabilité), la collecte des données sort de ces

**1**. Ministère des Solidarités et de la Santé, Stratégie nationale de santé, déc. 2017, p. 63.



cadres spécifiques pour prendre une place plus générale et centrale au sein du Code de la santé publique. Cette transformation est mise en évidence par la création en 2016 du Système national des données de santé (SNDS), dont la composition et le fonctionnement sont régis par le Code de la santé publique. Le SNDS regroupait, en 2016, l'accès à cinq bases de données dont celle de l'Assurance maladie sur le remboursement des soins et la base de données nationale sur les causes de décès. Selon les règles gouvernant l'accès aux données, les données doivent être au service exclusif de la santé.

En 2019, le législateur poursuit ce mouvement enclenché en réformant le SNDS pour accorder une place toujours plus importante à la collecte et à la conservation des données. Pierre angulaire de la réforme : l'élargissement du nombre des données disponibles au sein du SNDS. Ainsi, aux bases de données déjà accessibles viennent s'ajouter, entre autres, toutes les données collectées à l'occasion de soins remboursés par l'Assurance maladie. Par conséquent, mesures de taille, de poids ou des résultats d'examens cliniques sont autant de données pouvant remonter au SNDS car consignées par les professionnels de santé lors de soins remboursés par l'Assurance maladie.

La réforme de 2019 est aussi l'occasion d'entériner un projet pilote lancé quelques mois plutôt : le Health Data Hub, plateforme nationale des données de santé chargée, notamment, de mettre à disposition les données du SNDS élargi, mais aussi de financer des projets de recherche sur ces mêmes données. Cette réforme n'est pas anodine : la collecte des données en grand nombre n'est plus un simple moyen au service de la santé publique au sein du Code de la santé publique, mais devient une fin en soi. À cet égard, la réforme simplifie la création d'entrepôts de données de santé. Ces derniers, créés par des hôpitaux, visent le regroupement de données diverses, y compris des données de santé. Ici, les nouvelles règles autorisent l'aspiration des données du SNDS à des fins d'enrichissement de tels entrepôts. Par conséquent, la réforme insiste sur la collecte des données en grand nombre, leur duplication au sein de bases existantes pour des recherches futures, et cela sans préciser davantage les types de recherches pouvant justifier un accès aux données.

Cette place centrale accordée à la collecte des données par le droit occulte même les difficultés de terrain, laissant en suspens la question du financement de ces collectes et de la mise en œuvre d'une interopérabilité dans l'accès aux données. De même la question des droits des individus est facilement évacuée au profit d'une anonymisation/pseudonymisation des données, pourtant largement faillibles.

## Un droit perméable aux acteurs privés

Sous l'impulsion des données massives, le droit de la santé publique offre aujourd'hui une place plus importante aux acteurs privés alors même que la santé publique a longtemps été réservée aux services publics. Le légis-

| Le SNDS en 2016 | Le SNDS en 2019 |
|---|---|
| • Les données d'analyse de l'activité des hôpitaux, du programme de médicalisation, des systèmes d'information. | • Les données d'analyse de l'activité des hôpitaux, du programme de médicalisation, des systèmes d'information. |
| • Les données du Système national d'information inter-régimes de l'Assurance maladie. | • Les données du Système national d'information inter-régimes de l'Assurance maladie. |
| • Les données sur les causes de décès. | • Les données sur les causes de décès. |
| • Les données médicosociales des maisons départementales des personnes handicapées. | • Les données médicosociales des maisons départementales des personnes handicapées. |
| • Un échantillon représentatif des données de remboursement par bénéficiaire transmises par des organismes d'assurance maladie complémentaire et défini en concertation avec leurs représentants. | • Un échantillon représentatif des données de remboursement par bénéficiaire transmises par des organismes d'assurance maladie complémentaire et défini en concertation avec leurs représentants. |
| | • Les données destinées aux professionnels et organismes de santé recueillies à l'occasion des activités donnant lieu à la prise en charge des frais de santé en matière de maladie ou de maternité et à la prise en charge des prestations en cas d'accident de travail et de maladie professionnelle. |
| | • Les données relatives à la perte d'autonomie lorsqu'elles sont appariées avec les bases de données précédemment citées. |
| | • Les données à caractère personnel des enquêtes dans le domaine de la santé lorsqu'elles sont appariées avec les bases de données précédemment citées. |
| | • Les données recueillies lors des visites médicales et de dépistage obligatoires effectuées par les médecins et infirmiers de l'Éducation nationale. |
| | • Les données recueillies par les services de protection maternelle et infantile dans le cadre de leurs missions. |
| | • Les données de santé recueillies lors des visites d'information et de prévention auprès des travailleurs. |

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 ».