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.