tribunes

Le SNDS, un outil au service des acteurs de terrain

L'exemple de l'étude du recours aux soins dentaires en Pays de la Loire

La connaissance des données de santé d'une population ou d'un territoire permet de faire émerger des besoins de santé et d'améliorer les parcours de soins proposés. Cet article présente l'utilisation des données de soins dentaires dans les Pays de la Loire.

Marie Dalichampt
Anne Tallec
Jean-François Buyck
Observatoire régional de la santé (ORS) des Pays de la Loire

La mise en place en 2017 du Système national des données de santé (SNDS) offre de nouvelles perspectives en matière d'action de santé publique. Ce système d'information permet en effet des analyses détaillées du recours aux soins et à la prévention, avec l'élaboration d'indicateurs déclinables par type de population, ou par niveau territorial fin. Ces analyses ouvrent de façon considérable le champ des possibles en matière d'identification des besoins, s'agissant de parcours de santé, de ciblage des populations ou de territoires prioritaires, puis de suivi des évolutions dans une logique évaluative.

Compte tenu de ces nouveaux enjeux, l'observatoire régional de la santé (ORS) des Pays de la Loire – qui comme l'ensemble des ORS dispose d'un accès large et permanent aux données du Système national des données de santé – s'est investi dans ce domaine en recrutant et formant son équipe à l'utilisation de ce système d'information particulièrement complexe.

En parallèle, afin de s'inscrire dans une logique d'action, mais aussi de mobiliser l'expertise métier indispensable à l'exploitation de ces données, notamment celles des bases de l'Assurance maladie, l'ORS a développé des collaborations avec les unions régionales des professionnels de santé libéraux (URPS), avec lesquelles il entretient des partenariats de longue date autour d'enquêtes sur les pratiques et conditions d'exercice [1].

Cette approche a rencontré une demande de l'URPS chirurgiens-dentistes des Pays de la Loire, qui souhaitait disposer d'une connaissance fine du recours au cabinet dentaire des enfants de la région. Une première étude a été produite en 2018 [2], et cette dynamique, à laquelle s'est alors associée l'Union française pour la santé bucco-dentaire (UFSBD), s'est poursuivie en 2019-2020 autour de trois nouvelles études, portant sur le recours des adultes âgés de 55 ans et plus [3], des personnes diabétiques [4], et des personnes traitées par biphosphonates (en cours). Pour chacune de ces études, un groupe de travail associant praticiens de terrain et spécialistes du Système national des données de santé a été mis en place pour élaborer des indicateurs pertinents et directement en lien avec les pratiques des professionnels de santé.

Un recours aux soins dentaires très en deçà des recommandations

Les analyses déjà réalisées sur trois des populations choisies par l'union régionale des professionnels de santé libéraux (enfants, seniors, personnes diabétiques) ont toutes montré un recours très insuffisant au cabinet dentaire, au regard des recommandations de la Haute Autorité de santé (HAS) d'un recours annuel minimum [5].

Le taux de recours annuel est très en deçà des 100 % souhaités : 61 % chez les 6-18 ans [2], 40 % chez les personnes diabétiques [4], et 47 % parmi les plus de 55 ans [3]. Chez ces derniers, le taux annuel de recours au cabinet dentaire décroît de façon continue à partir de 65 ans, et n'est plus que de 25 % au-delà de 90 ans.

De plus, le chaînage des données du Système national des données de santé, qui permet d'analyser l'ensemble des prestations de chaque bénéficiaire, met en évidence l'ampleur du non-recours sur plusieurs années

consécutives. En Pays de la Loire, près d'un enfant sur dix n'a bénéficié d'aucun recours bucco-dentaire préventif (ni examen bucco-dentaire [EBD], ni consultation, ni détartrage) entre 6 ans et 9 ans, et cette proportion atteint 25 % entre 14 ans et 17 ans. Et ce malgré le programme M'T dents de l'Assurance maladie, qui propose un examen bucco-dentaire sans avance de frais aux âges de 6, 9, 12, 15 et 18 ans, et à 3 ans depuis le 1er janvier 2019.

Chez les seniors, la situation est également très défavorable : 25 % des Ligériens âgés de 55 à 70 ans n'ont eu aucun recours au cabinet dentaire sur les années 2016 à 2018, et cette proportion dépasse 50 % au-delà de 90 ans.

Enfin, pour les personnes diabétiques, dont l'état de santé bucco-dentaire est étroitement lié au risque de déséquilibre et de complications du diabète, les résultats sont alarmants : plus d'un tiers des Ligériens pris en charge pour un diabète en 2015 n'ont eu aucun recours au cabinet dentaire au cours des trois années suivantes (2016-2018), et seulement 16 % ont eu un parcours conforme aux recommandations, c'est-à-dire au moins une consultation chacune des trois années.

Ces résultats illustrent bien l'importance des enjeux : malgré des recommandations anciennes, relayées par de nombreux acteurs de santé (Assurance maladie, sociétés savantes et associations d'usagers) et préconisant des soins bien remboursés par l'assurance maladie obligatoire, une part importante de la population a un recours aux soins dentaires très insuffisant, alors que l'impact de la santé bucco-dentaire sur la santé générale est désormais bien établi.

Des publics et territoires encore plus prioritaires que d'autres

Un grand nombre d'études ont montré que la santé bucco-dentaire constitue un excellent marqueur des inégalités sociales de santé. Les indicateurs étudiés ici, issus du Système national des données de santé, confirment ce constat. Chez les enfants, les différences les plus marquées concernent la fréquence du suivi préventif et l'âge du premier recours au cabinet dentaire : 30 % des enfants bénéficiant de la couverture maladie universelle complémentaire [CMU-C] n'ont jamais eu de recours avant 7 ans, contre 16 % de ceux qui n'en bénéficient pas [2]. Chez les personnes âgées de 55 ans et plus, la proportion de celles n'ayant eu aucun recours en trois ans atteint 43 % chez les bénéficiaires de la CMU-C ou de l'aide au paiement d'une complémentaire santé (ACS)[1], contre 29 % chez les personnes qui n'en bénéficient pas, à structure par âge équivalente [3]. Les écarts de recours selon le niveau social sont particulièrement importants pour les poses de prothèse fixe, soins à fort reste à charge pour la période concernée par l'étude, mais également pour les détartrages, pourtant bien remboursés.

Les enfants admis en affection de longue durée (ALD), qui sont le plus souvent atteints de maladies chroniques et sont pour certains en situation de handicap, présentent des indicateurs de recours encore plus dégradés que les autres enfants. Ainsi, 17 % n'ont eu aucune prestation de suivi bucco-dentaire entre 6 et 9 ans (10 % des enfants sans ALD). Lorsqu'un traitement orthodontique est commencé, il l'est plus tardivement chez les enfants en ALD, ce qui peut le rendre moins efficace : 42 % le débutent avant 10 ans et 32 % après 13 ans (contre respectivement 48 % et 25 % chez les enfants sans ALD).

Chez les personnes âgées de 75 ans et plus, pour lesquelles le recours au cabinet dentaire est globalement très insuffisant, le fait de résider ou non en établissement d'hébergement pour personnes âgées dépendantes (Ehpad) est un des principaux facteurs explicatifs d'une augmentation du risque de non-recours [3]. En effet, les analyses multivariées, ajustant sur l'âge, les caractéristiques sociales, l'état de santé et le niveau d'accessibilité potentielle localisée (APL) au chirurgien-dentiste libéral, montrent que l'association entre le non-recours au cabinet dentaire pendant au moins trois ans et le fait de résider en Ehpad est très significative, avec un *odds ratio* de près de 1,5 chez les personnes nouvellement arrivées en Ehpad, et qui s'élève à 2,5 chez celles hébergées depuis au moins deux années, comparées aux personnes vivant à leur domicile [3]. Ce résultat peut en partie être expliqué par un plus grand degré de dépendance des résidents en Ehpad, mais aussi par un éloignement à leur chirurgien-dentiste habituel du fait du déménagement vers l'Ehpad.

La déclinaison des différents indicateurs pour des zonages géographiques fins met par ailleurs en évidence des disparités territoriales de recours très marquées. Selon leur établissement public de coopération intercommunale (EPCI), la part des enfants qui bénéficient d'un parcours préventif bucco-dentaire régulier entre 6 et 9 ans varie de 30 à 60 % [2]. Les enfants domiciliés dans les établissements publics de coopération intercommunale de Sarthe, et plus particulièrement dans les intercommunalités les plus éloignées de la métropole du Mans, ont un recours globalement moins fréquent, moins précoce et moins régulier comparés aux enfants des autres établissements publics de coopération intercommunale de la région. À l'inverse, la Loire-Atlantique et la Vendée (où est né dans les années 1980 le bilan bucco-dentaire auquel a succédé le programme M'T dents) englobent la plupart des établissements publics de coopération intercommunale où les fréquences de parcours préventif chez les enfants sont les plus élevées. La Loire-Atlantique et la Vendée concentrent également la grande majorité des établissements publics de coopération intercommunale où le recours aux soins dentaires est plus satisfaisant au-delà de 55 ans, et chez les personnes prises en charge pour un diabète [3, 4]. Pour ces dernières, les écarts entre territoires sont considérables, avec une part de personnes diabétiques ayant eu un recours satisfaisant entre 2016 et 2018, c'est-à-dire chacune des trois années, qui varie de 5 à 24 %.

Agir, notamment localement dans le cadre des dynamiques interprofessionnelles

Les acteurs susceptibles de se saisir des données du Système national des données de santé pour mettre en place des démarches visant à améliorer ces parcours sont multiples (et non exclusifs les uns des autres).

Au plan national, l'Assurance maladie utilise depuis de nombreuses années le relais que constituent ses caisses locales pour mettre en œuvre et suivre, à partir de son système d'information (qui alimente le SNDS), des programmes de dépistages organisés et de prévention. L'invitation de chaque enfant aux examens bucco-dentaires gratuits du programme M'T dents en fait partie. Cette démarche pourrait également être étendue à des populations adultes cibles (par exemple à l'entrée en Ehpad, comme préconisé par l'Union française pour la santé bucco-dentaire).

Au plan régional, les URPS, qui ont pour mission de contribuer à l'organisation de l'offre de soins et à la politique régionale de santé, aux côtés des agences régionales de santé (ARS), peuvent également s'appuyer sur les données du SNDS pour faire émerger des projets d'actions. Les travaux sur le recours au cabinet dentaire menés

1. Depuis la réalisation de ces études, la CMU-C et l'ACS ont été remplacées par la complémentaire santé solidaire (CSS/C2S).

par l'ORS à la demande et avec le soutien de l'union régionale des professionnels de santé libéraux chirurgiens-dentistes en sont un premier exemple.

Au niveau local, le développement des coopérations interprofessionnelles et la mise en place récente des CPTS[2] (communautés professionnelles territoriales de santé) constituent de formidables opportunités. En effet, la mobilisation coordonnée des différents professionnels de santé, tant pour orienter les patients que pour partager de l'information à leur propos, est l'une des conditions de l'amélioration des parcours. L'étude concernant le recours au cabinet dentaire des personnes diabétiques montre, par exemple, que la proportion de celles ayant un parcours dentaire satisfaisant est sensiblement plus élevée parmi les personnes qui consultent régulièrement leur médecin généraliste, après ajustement sur les autres facteurs [4]. Cette association, bien qu'elle ne démontre pas de rapport de cause à effet, suggère le rôle central que peuvent jouer les généralistes dans l'adhésion de leurs patients aux recommandations de suivi bucco-dentaire. De même, il est important que le chirurgien-dentiste connaisse l'existence du diabète de son patient pour adapter sa prise en charge et lui rappeler l'importance d'un suivi régulier.

Les communautés professionnelles territoriales de santé, auxquelles a été confiée une responsabilité populationnelle, offrent désormais un cadre pertinent à ces dynamiques d'amélioration des parcours basée sur l'élaboration et le suivi d'indicateurs concernant les habitants de leur territoire.

Les indicateurs élaborés avec les professionnels de santé grâce aux données du Système national des données de santé permettent de définir des objectifs d'amélioration des parcours de soins concrets et adaptés au contexte local. Parce qu'ils correspondent à leur pratique au quotidien, ces indicateurs facilitent la mobilisation des professionnels de santé pour participer à des actions en vue d'atteindre ces objectifs. Les indicateurs peuvent cibler finement des populations en fonction de leur lieu d'habitation, de leur âge, de leur parcours de vie (entrée en Ehpad), de leur état de santé (ALD, diabète) ou encore de leur traitement (biphosphonates).

Enfin, la possibilité d'évaluer de façon rapide et fiable l'impact des actions menées en mesurant l'évolution des indicateurs du Système national des données de santé est d'un intérêt majeur. En effet, des résultats positifs permettent à la fois de valider la pertinence des actions, mais aussi de favoriser le maintien de la motivation des professionnels.

Quels enjeux pour les années à venir ?

La connaissance fine des données de santé d'une population ou d'un territoire est nécessaire pour faire émerger des besoins de santé et identifier des leviers d'amélioration. Mais cette connaissance n'est pas suffisante. Encore faut-il ensuite que les acteurs opérationnels, et notamment les professionnels de santé, s'en saisissent pour initier et mener des actions.

En Pays de la Loire et en matière de soins bucco-dentaires, une part du chemin a été parcourue à travers la mobilisation de l'union régionale des professionnels de santé libéraux chirurgiens-dentistes et de l'ORS autour de ces travaux, qui illustrent l'important besoin d'amélioration du recours aux soins. Mais un effort considérable de pédagogie et de communication autour de ces travaux doit encore être accompli, pour convaincre les différentes parties prenantes de la nécessité de mettre en place des actions, et parvenir à mobiliser les moyens humains et financiers que de telles dynamiques impliquent.

Ne pas tirer profit de l'apport des données du SNDS serait un immense gâchis. Ces données, mondialement enviées, constituent un levier majeur pour l'amélioration des parcours de soins, la réduction des inégalités sociales et territoriales de santé, et plus largement l'appropriation des enjeux de santé publique par les professionnels de santé.

2. Les CPTS émanent de l'initiative des acteurs de santé, en particulier des professionnels de santé de ville. Ce sont des équipes projets, s'inscrivant dans une approche populationnelle au sens où les différents acteurs acceptent de s'engager dans une réponse à un besoin de santé de leur territoire, qui peut impliquer pour eux de prendre part à des actions ou d'accueillir des patients sortant de leur exercice et de leur patientèle habituelle (instruction n° DGOS/R5/2016/392 du 2 décembre 2016 relative aux équipes de soins primaires [ESP] et aux communautés professionnelles territoriales de santé).

Références bibliographiques

1. ORS Pays de la Loire. Enquêtes et panels professionnels de santé [en ligne]. https://www.orspaysdelaloire.com/enquetes-et-panels-professionnels-de-sante
2. ORS Pays de la Loire, URPS chirurgiens-dentistes Pays de la Loire. *Recours au cabinet dentaire des enfants et des adolescents. Situation en Pays de la Loire et en France à partir d'une analyse des données du SNDS*. 2018, 76 p.
3. ORS Pays de la Loire, URPS chirurgiens-dentistes Pays de la Loire. *Recours au cabinet dentaire des adultes de 55 ans et plus. Situation en Pays de la Loire et en France à partir d'une analyse des données du SNDS*. 2019, 72 p.
4. ORS Pays de la Loire, URPS chirurgiens-dentistes Pays de la Loire. *Suivi bucco-dentaire des personnes diabétiques en Pays de la Loire à partir d'une analyse des données du SNDS*. À paraître en 2020, 28 p.
5. HAS. *Stratégie de prévention de la carie dentaire. Synthèse et recommandations*. 2010, 26 p.

tribunes

Big data, data reuse en santé : un chemin semé d'embûches nécessitant une approche pluridisciplinaire

Illustration de la complexité d'utilisation de données de santé et de la nécessaire collaboration de plusieurs professions.

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

Nous vous entraînerons avec nous dans un exemple de recherche ayant réussi. L'objet de cette recherche était de mettre en place un système d'intelligence artificielle (IA) permettant de prévenir les effets indésirables du médicament (EIM). Nous verrons que le chemin fut parsemé d'embûches. Ces embûches illustrent l'impossibilité de faire accomplir un tel travail par une machine seule, et l'absolue nécessité d'approches pluridisciplinaires.

Contexte et objectifs

En France, chaque hospitalisation de patient donne lieu à la collecte d'informations codées tels les diagnostics (par exemple « K37 » pour certaines appendicites) et les actes (par exemple « HHFA001 » pour certaines ablations chirurgicales de l'appendice). De plus, dans la plupart des hôpitaux, les médicaments sont prescrits *via* des logiciels et les résultats d'analyses de biologie médicale sont également transmis au service prescripteur par des logiciels spécifiques. Toutes ces données constituent des données massives ou *big data* (voir l'encadré de définitions) [3] : on dénombre en moyenne 100 résultats de biologie médicale par séjour (par exemple le taux d'hémoglobine) et 200 000 séjours par an dans certains CHU. Ces données servent respectivement à facturer le séjour à l'Assurance maladie et à soigner le patient. Elles

Quelques définitions simplifiées

Données massives/*big data*
Données volumineuses, comportant par exemple de nombreux individus et/ou de nombreuses variables, quelles que soient leur origine et leur exploitation.

Réutilisation de données/*data reuse*
Fait d'utiliser (pour de la recherche par exemple) des données qui ont initialement été collectées dans un autre but (le soin par exemple).

Intelligence artificielle (IA)
Fait de doter l'ordinateur de capacités visant à mimer le résultat d'une intelligence. Une IA peut s'appuyer sur un répertoire de règles écrites par un humain, un apprentissage automatisé (*machine learning*), ou un apprentissage par renforcement.

Effet indésirable du médicament (EIM)
Effet secondaire nocif et non voulu lié à la prise, la modification de dose ou l'arrêt d'un médicament (hormis erreurs d'administration et tentatives de suicide).

Emmanuel Chazard
Professeur des universités, faculté de médecine, Cerim ULR 2694, université de Lille, praticien hospitalier, CHU de Lille

peuvent cependant être réutilisées à des fins de recherche, on parle alors de réutilisation de données ou *data reuse*.

Notre objectif ici est de les réutiliser pour détecter automatiquement les circonstances causant des EIM. Une fois ces circonstances identifiées, nous bâtirons un logiciel d'IA [10] qui « surveillera » les prescriptions médicamenteuses et alertera le prescripteur en cas d'élévation du risque d'effet indésirable du médicament. Cet énoncé illustre le lien qu'entretiennent parfois *big data*, *data reuse* et intelligence artificielle. Cela dit, ce lien n'est pas obligatoire. La plupart du temps, le grand public parle de *big data* pour désigner le *data reuse* parce que les jeux de données dont la réutilisation est la plus prometteuse sont généralement de grande taille. Passées ces définitions, lançons-nous dans notre recherche !

Notre recherche, pas à pas

Construction d'un entrepôt de données

La première étape est « d'aspirer » les données disponibles et de les nettoyer (indifféremment de l'objectif). Copiées dans une nouvelle base de données plus simple, elles constituent un « entrepôt de données ». Première surprise, un certain nombre d'enregistrements sont « orphelins » et ne peuvent pas être rattachés à un séjour hospitalier… la complexité des systèmes d'information hospitaliers (SIH) est telle, qu'il est difficile de savoir si le lien existe mais est perdu lors de l'extraction, ou si l'information est d'emblée corrompue. Le chercheur, quant à lui, devra renoncer à tout analyser. Ensuite, les données comprennent de nombreuses erreurs. En voici un exemple : dans un hôpital, le libellé « hématies » est associé à des valeurs comprises entre 4 000 et 5 500/mm³. C'est mille fois trop faible, pourtant ces patients sont tous vivants. L'unité enregistrée est tout simplement fausse : pourtant, le médecin interprète le chiffre précisément mais sans regarder s'il s'agit de milliers ou millions, et parfois sans connaître précisément l'unité attendue. D'autres patients également ont un taux de potassium tournant aux alentours de 30 mmol/l, la norme devant rester inférieure à 4,5. Sont-ce des momies ? Non, simplement des patients dont on a dosé le taux de potassium urinaire, mais sans l'écrire clairement dans la base de données. Ces types d'erreurs n'ont généralement aucun impact sur le soin, car le soignant les corrige mentalement sans même s'en rendre compte. Elles sont critiques lorsqu'on analyse les données.

Leçon apprise

La construction d'un entrepôt de données requiert des informaticiens, mais aussi des médecins capables d'évaluer la qualité des données, et des analystes capables de dire quelles données sont indispensables ou peuvent être sacrifiées.

Analyse de données

Nos données sont enfin prêtes. Mais qu'en faire ? Certes, la construction d'un entrepôt de données nous a permis de passer de plus de mille tables de données d'un SIH à quelques dizaines de tables, mais les données restent complexes : hétérogènes (diagnostics, actes, médicaments, résultats d'analyses…), presque toujours manquantes et presque jamais par hasard (très peu de patients bénéficient d'une IRM de l'hypophyse… parce que leur hypophyse est normale), très complexes (près de 40 000 codes différents pour décrire les maladies, et encore cela ne suffit pas), et plus ou moins fiables !

Et pourtant, observons un médecin parcourir un dossier médical (figure 1) :

« Ce patient fait probablement une hémorragie : son **INR [International Normalized Ratio, l'un des indicateurs de la coagulation sanguine] augmente**, *témoignant d'une activité anticoagulante trop forte, et le lendemain son* **taux d'hémoglobine diminue**, *traduisant sans doute une hémorragie. Un* **antivitamine K** *(anticoagulant) était administré juste avant, et vraisemblablement potentialisé par du* **paracétamol à forte dose**. *Le médecin corrige ensuite cela en* **arrêtant l'anticoagulant** *et en* **introduisant de la vitamine K**, *son antidote. Par la suite, l'*INR **se normalise.** »

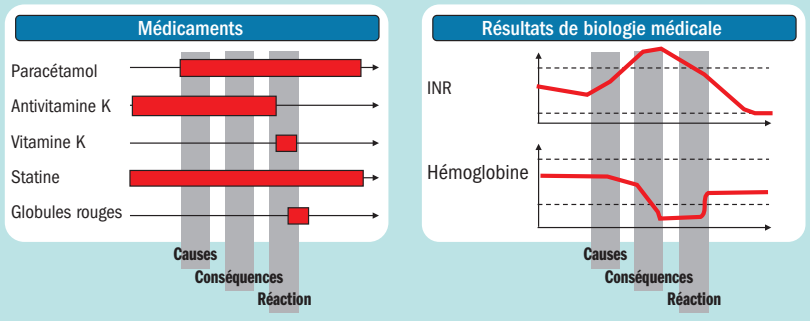
Quoi qu'on en dise, aucun ordinateur n'est capable de tenir un tel raisonnement. L'analyse de ce raisonnement nous permettra néanmoins de définir des « caractéristiques » (ou *features*) qu'il faudra calculer [13]. Il s'agit simplement d'appliquer tout un tas d'algorithmes (décidés par un expert) pour créer autant de variables simplifiées qu'il y avait de constatations **soulignées** ci-dessus, et plus encore par généralisation. Ces caractéristiques se présentent de manière plus simple que les données réelles disponibles, et elles sont « optimisées » pour porter du sens. Ainsi, grâce à l'extraction de caractéristiques, nos algorithmes de *machine learning* ne seront plus comme une poule face à un couteau.

Leçon apprise

L'extraction de caractéristiques nécessite en amont une excellente connaissance des données, une maîtrise de l'algorithmique de base, et en aval une connaissance des formes de données qui « fonctionnent » en *machine learning*.

figure 1

Interprétation de données brutes d'un patient qui présente une hémorragie sous anticoagulants



Lecture : le temps va de gauche à droite ; le commentaire est dans le texte

Prédiction par *machine learning*

L'extraction de caractéristiques nous a permis de construire un grand tableau, comprenant une seule ligne par patient, et des milliers de colonnes, représentant des variables simples. Nous avons supprimé la complexité des données. Il est à présent aisé d'utiliser des techniques d'apprentissage automatisé (*machine learning*) pour prédire automatiquement certaines variables (par exemple présenter une hémorragie) à l'aide de toutes les autres (avoir un anticoagulant, l'âge, le sexe, etc.). À ce jeu-là, les réseaux de neurones sont plutôt doués. Hélas, ils construisent pour ce faire des formules mathématiques de plusieurs pages, incompréhensibles par un humain. Cela leur vaut d'être qualifiés de « boîte noire ». En pratique, dans notre projet, ils s'avèrent inutilisables : comment alerter un médecin sur un risque, sans même être capable de livrer des arguments validés scientifiquement ? Nous préférons donc une technique moins vendeuse d'un point de vue marketing, mais plus lisible, comme les arbres de décision (figure 2). Les premiers arbres nous apprennent que… les patients plus âgés meurent davantage. Ce résultat était peut-être connu dès la Préhistoire ! Après des corrections de variables, les arbres suivants nous apprennent par exemple que le plus fort facteur de risque d'hyperglycémie, c'est d'avoir de l'insuline. Or l'insuline entraîne des hypoglycémies. C'est pourtant évident : association statistique ne signifie pas causalité. Il faut mettre en place des procédures automatisées et expertes de filtrage des associations découvertes par la machine. À la fin, nous tenons le bout (figure 2) : ainsi par exemple, sur sept patients présentant une insuffisance rénale et âgés de plus de 85 ans et ayant de la spironolactone, six (85 %) présentent ensuite une hyperkaliémie. L'association est techniquement, statistiquement et bibliographiquement valide. Pourtant, la relecture des dossiers nous apprend que seulement la moitié des pathologies sont réellement des effets imputables au médicament, car il arrive tout un tas d'autres choses à ces patients, invisibles dans les données.

Leçon apprise

La prédiction par *machine learning* n'est donc pas seulement un jeu de statisticien. Elle nécessite, outre l'extraction de caractéristiques, une forte expertise métier, des résultats analysables et critiquables par un humain, et une validation par retour aux cas réels.

Prévention en vie réelle

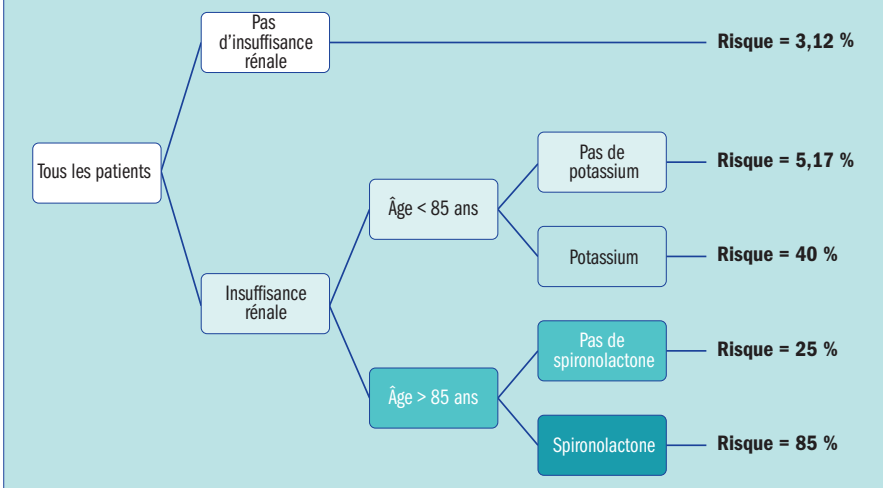
Au bout du processus de *machine learning*, nous tenons enfin un lot de 256 règles permettant de prédire la survenue d'EIM. Il « suffira » de les intégrer dans un système d'aide à la décision connecté au logiciel de prescription. Nous apprenons au passage que le niveau de risque (pourcentages dans la figure 2) varie fortement d'un service à l'autre. Plus généralement, les médecins maîtrisent très bien les médicaments de leur spécialité, et pour eux les alertes sont totalement inutiles : le fait est qu'elles leur pourrissent la vie sans diminuer les erreurs ! Inversement, c'est plutôt sur les situations plus rares qu'ils peuvent se tromper. Ainsi, par exemple, on ne verra jamais d'hyperkaliémie en néphrologie bien que tous les patients soient à risque, mais on peut en voir en pneumologie. Forts de ce constat, nous déployons le premier « SPC-CDSS » [12], c'est-à-dire le premier système d'aide à la décision filtré et contextualisé statistiquement. Bien que novateur et hautement valide, ce système ne sera pas utilisé par les cliniciens. Nous avions tout pris en compte, sauf le facteur humain. Bêtement, comme tant d'autres avant nous, nous avions engendré une monstruosité : un logiciel qui prend du temps à ceux qui en ont le moins, un logiciel qui rajoute des clics, des actions et des alertes à ceux auxquels tout le personnel hospitalier transfère insidieusement son travail : les médecins.

Leçon apprise

On ne peut pas se contenter de dire : « Le logiciel est bon donc ils l'utiliseront. » L'informatisation des hôpitaux a insidieusement transféré les tâches du personnel le moins qualifié vers le plus qualifié. Argumentant des économies, elle a augmenté le coût du travail pour une tâche donnée, et diminué la productivité pour certaines catégories de personnel hautement qualifié. Un bon logiciel d'IA est surtout un logiciel qui s'intégrera dans un *workflow* de manière à alléger la charge cognitive et l'agacement des professionnels qualifiés, et leur faire gagner du temps.

figure 2

Exemple d'arbre de décision (*machine learning*) prédisant automatiquement un risque d'hyperkaliémie



Conclusion

À travers ce cheminement, nous avons illustré les nombreux obstacles rencontrés, et la nécessité de faire collaborer au minimum informaticiens, statisticiens, spécialistes des facteurs humains, et surtout experts des données considérées (ici, des médecins). La vidéo citée en référence présente également ce cheminement [11].

tribunes

La position des institutions publiques françaises dans la promotion et l'utilisation des données en santé publique

La promotion et l'utilisation des données de santé sont organisées pour un service public, par des institutions qui peuvent être en concurrence.

Thomas Lefèvre
Maître de conférences, praticien hospitalier, Iris-UMR 8156-997 CNRS Inserm EHESS université Sorbonne Paris Nord
Sabine Guez
Post-doctorante, Iris-UMR 8156-997 CNRS Inserm EHESS université Sorbonne Paris Nord

Cette tribune n'a pas pour vocation de retracer une sociohistoire précise du rôle des institutions publiques françaises dans la promotion et l'utilisation des données en santé publique, mais d'en donner quelques éléments de repères et de questionnements.

Porter l'attention du privé vers le public en matière de données de santé

S'agissant du *big data*, de l'intelligence artificielle, plus largement de l'irruption du numérique dans notre quotidien, l'attention a surtout été focalisée sur les acteurs privés, comme Google, Amazon, Facebook, Apple ou Microsoft (Gafam), en particulier pour des raisons de souveraineté nationale et de protections juridiques variables et non garanties partout dans le monde de la même façon. Ces acteurs ont proposé, proposent encore régulièrement, de se positionner comme intermédiaires voire comme substituts à ce qui peut être considéré comme des fonctions régaliennes, au minimum des missions de service public. On peut penser à la fonction *safety check* de Facebook, en cas de catastrophe naturelle ou d'événement imprévu de portée collective, menaçant la vie des personnes, ou aux applications smartphone proposées par Google et Apple aux gouvernements dans le cas du *case tracking* dans le contexte de l'épidémie de Covid-19. En comparaison, la position des institutions publiques dans la promotion et l'utilisation des données n'a fait l'objet que de peu d'attention. Or en France l'organisation et l'utilisation des données en santé tiennent essentiellement à des initiatives des institutions publiques, aux niveaux gouvernemental et ministériel, depuis une quarantaine d'années.

Le pilotage médico-économique des établissements de santé basé sur la donnée

Une façon d'aborder le sujet est de partir de la création du PMSI (Programme de médicalisation du système d'information [des établissements de santé]) au début des années 1980, dont l'utilisation a été renforcée en 1996, puis en 2005. L'idée derrière le PMSI est la quantification et la standardisation d'un certain nombre d'informations ayant trait aux hospitalisations : les diagnostics, les durées de séjour. Conjointement, un travail de classification et de nomenclature se met en place, s'institutionnalise, notamment avec la création de l'ATIH (Agence technique de l'information sur l'hospitalisation) en 2000. Il s'agit d'une part d'identifier et de mettre à jour l'ensemble des GHM (les groupes homogènes de malades) à partir des données collectées dans les hôpitaux et, d'autre part, de faire coïncider une nomenclature médico-économique, sur laquelle la tarification à l'activité (la T2A) se base dès 2005. Des données concernant un grand nombre de malades et d'hospitalisations sont collectées ; par des approches algorithmiques, on identifie des groupes de malades qui se « ressemblent » tant en termes médicaux (diagnostics…) que de coûts du séjour en hôpital. Ces groupes permettent de « segmenter » la population hospitalière comme on le fait en marketing pour identifier des sous-populations de clients. La technique n'est en rien novatrice et est importée des pratiques du privé. Cela sert alors à donner

une assiette au budget d'un établissement de santé. Au sein de ces établissements, les départements d'information médicale (DIM) ont, dans beaucoup de cas, la tâche désormais principale de gérer le PMSI local, et de justifier le budget de leur établissement. Au niveau national, les données des PMSI locaux sont fédérées, consolidées.

Les données au-delà du PMSI et du Sniiram : la fédération des données, leur gestion et leur accessibilité

Le PMSI est donc une source de données historique en France. Une autre source de données elles aussi médico-administratives, gérée par une institution publique, est le Sniiram (Système national d'information inter-régimes de l'Assurance maladie). Elle est gérée par la Caisse nationale d'assurance maladie et fait partie de la base du Système national de données de santé (SNDS), regroupant de fait avant tout les données du PMSI et du Sniiram.

Simultanément, d'autres acteurs publics se sont manifestés de façon croissante quant à l'utilisation des données de santé, au-delà des données médico-administratives. On a ainsi assisté à l'émergence des entrepôts de données de santé des hôpitaux, tendant à rassembler l'ensemble des données dérivées de tous les examens et observations effectués lors d'un séjour hospitalier. Le panorama des organismes publics pouvant participer à l'écosystème de la production et de l'utilisation des données de santé en France se complète par les universités (facultés de médecine), les unités Inserm et les DIM déjà cités, ainsi que plusieurs organismes ministériels : Drees, Acoss, CNSA…

Nous avons donc ici le panorama classique d'une multitude de sources de données à réconcilier, néanmoins doublé d'une multitude d'acteurs, tous publics, aux intérêts concurrents. Cette concurrence est fréquemment réduite à l'argument d'une supposée propriété des données : l'hôpital propriétaire des données captées dans son établissement, l'Inserm propriétaire des données générées dans un cadre de recherche, la faculté propriétaire des données recueillis par son personnel hospitalo-universitaire – lui-même fréquemment affilié à une unité Inserm et exerçant en établissement hospitalier. Une solution organisationnelle et technique proposée pour une forme de convergence de ces sources de données, solution dont la teneur réelle reste néanmoins à préciser à l'épreuve de la pratique et du temps, s'incarne depuis le 30 novembre 2019 dans la plateforme de données de santé, le Health Data Hub, constitué en groupement d'intérêt public (GIP). Conçu comme un guichet unique facilitant l'accès et l'utilisation des données pour la santé pour différents acteurs, le Health Data Hub cristallise cependant plusieurs critiques d'autant plus aisément qu'il apparaît comme une personne identifiable plutôt que la multitude des acteurs. Ces critiques sont liées aux aspects de la nouvelle gestion publique (*new public management*). Un exemple est le recours à l'externalisation pour des tâches ou des missions importantes, comme l'hébergement des données attribué à Microsoft. Un autre exemple est celui de la conformité avec le référentiel de sécurité Système national des données de santé, qui par ricochet réglementaire, est applicable ou demandé désormais à des acteurs qui n'en ont pas les moyens, et les privant de données auxquelles ils avaient jusque-là accès, nécessaires à la réalisation de leurs missions.

Parallèlement, ce problème de convergence des sources de données, et de « propriété », aurait pu être au moins techniquement résolu par la création et l'utilisation du dossier médical partagé (DMP), annoncé par la loi du 13 août 2004 relative à l'assurance maladie : il s'agirait simplement d'un carnet de santé électronique, étendu à toute la vie et non plus aux seules premières années. Le projet ne prend pas, certainement pour des raisons de visions divergentes et de concurrences interinstitutionnelles. L'Agence des systèmes d'information partagée en santé (Asip, devenue Agence du numérique en santé en décembre 2019) est créée en 2009, dont une des missions est spécifiquement de relancer le projet du DMP. La Caisse nationale d'assurance maladie reprend le projet à son compte en 2015. Finalement, la dernière itération de ce concept devrait se réaliser dans l'Espace numérique de santé, défini par la loi du 24 juillet 2019, à partir du 1er janvier 2022.

La concurrence n'est pas limitée à une concurrence inter-institutions. Elle existe également au sein même des institutions. Un de ses aspects est une confrontation culturelle et liée à un changement dans les formations des élites françaises, où les écoles de management et de commerce sont depuis plusieurs années plus prisées que les écoles d'ingénieurs. Traditionnellement, au sein de l'institution publique, les personnes en charge des études et des statistiques – la statistique d'État, et ses extensions, *via* le captage et l'automatisation, la fédération des données produites et numérisées – sont majoritairement issues de l'École nationale de la statistique et de l'administration économique (Ensae). On observe, depuis quelques années, le recrutement de personnes plus souvent issues d'écoles privées de management et de commerce (HEC, Essec, Edhec…), voire la constitution interne de départements dédiés et distincts des départements des études et statistiques autour de ces recrutements. Ces nouveaux départements ou directions sont en général en charge de l'innovation ou de la transformation « digitale », et se retrouvent face aux départements historiques des systèmes d'information et des études statistiques. Il existe enfin probablement une concurrence, ou un frein à la convergence, sur le versant sanitaire : les experts en santé publique sont, en France, issus d'au moins trois horizons différents, parfois cumulatifs, mais dont les logiques institutionnelles sous-jacentes sont plutôt concurrentielles là aussi : les médecins de santé publique, les chercheurs en épidémiologie, et les diplômés de l'École des hautes études en santé publique (EHESP).

Globalement, on assiste donc à des efforts publics de concentration des données personnelles, captées à partir de différentes institutions publiques mais aussi depuis le secteur du privé *via* le secteur du travail, définie et imposée par la voie législative. Les efforts d'ouverture de ces données semblent réels, mais loin de reposer ou de favoriser une déconcentration, comme cela aurait pu, ou pourra être, en se basant sur la participation et l'accord de chaque citoyen, *via* un DMP ou tout autre dispositif semblable et acceptable par la société. À ce jour, tout sondage effectué pour mesurer l'acceptation des Français d'un tel dispositif a toujours pointé vers une adhésion massive au principe : quiconque passe de son généraliste aux urgences, puis à l'hospitalisation, puis d'un hôpital à un autre, d'un spécialiste à un autre, sait combien il serait utile de disposer d'un dossier unique.

La donnée pour décider, le fait scientifique et la santé publique

Le protectionnisme de l'État dès qu'il est question de données de santé, envers les autres nations, mais aussi envers une partie du secteur privé, parfois entre institutions publiques et enfin même envers les citoyens, est souvent justifié par le caractère sensible de ces données – on met en avant le caractère individuel, mais d'un point de vue national, elles sont aussi sensibles en cela qu'elles

renseignent sur la population –, le secret professionnel et la valeur des données pour l'industrie et la recherche. Il existe également la dimension que la récente épidémie de Covid-19 a mise en exergue : le besoin de disposer de connaissances scientifiques d'une part, mais également de données pour décider, pour gouverner. Plusieurs commentateurs auront relevé la porosité des champs lexicaux quant à la gestion de l'épidémie en cours, relevant tous des prérogatives régaliennes, à commencer par le langage martial : nous sommes en guerre contre le virus ; un état d'urgence sanitaire peut être déclaré, délivrant des pouvoirs spéciaux aux représentants de l'État. Dans le cadre de l'épidémie, les deux sources évoquées, supports de la décision éclairée, ont été prises en défaut, au sens de leur caractère utile, adapté et précis à un moment donné de l'épidémie : les connaissances scientifiques d'une part, la production de données utiles d'autre part. Ce qui est aux racines historiques de la statistique, dénombrer les morts et les naissances, a été défaillant, et les logiques concurrentielles entre institutions, un problème majeur : le CépiDc, unité Inserm unique dévolue au recueil et à l'étude des causes médicales de décès, en position d'asphyxie depuis plusieurs mois sinon années, a sans doute été un des seuls acteurs à avoir été privé de la manne financière ouverte à destination de la recherche et de l'industrie face à la Covid-19. L'écosystème en place n'a pas pu être mobilisé, renforcé, pour fonctionner et produire les données nécessaires. D'autres institutions publiques ont occupé le terrain de façon peu convaincante ou insuffisante pour documenter efficacement et rigoureusement les connaissances sur la mortalité, comme l'Insee, l'Ined ou Santé publique France. Et pour cause : aucune de ces institutions ne dispose des causes médicales de décès.

Symétriquement, le recueil de données de santé peut être, au-delà de son caractère informatif sur l'état de santé d'une population, le biais par lequel introduire une forme de gouvernement des individus, voire d'une police sanitaire, et la possibilité, parfois inédite, de rapprocher des institutions publiques, et même privées, dans cette optique de contrôle et de régulation des flux de population. Le cas de la santé publique de précision dans le cadre de la gestion de l'épidémie de Covid-19 est éclairant. Prenons l'exemple des applications téléphoniques de *tracking* (traçage des cas). La majorité des pays à hauts revenus s'est dotée d'une application de *tracking*. Aucune n'a été en position d'être sanitairement utile – la couverture nécessaire, c'est-à-dire la proportion de personnes installant et utilisant réellement l'application, est estimée à plus de 60 % pour être utile. Il semble que, dans le meilleur des cas, la proportion des personnes ayant téléchargé une telle application a été de 30 à 40 %. Le cas français (StopCovid) est une illustration classique de nos institutions : un outil propriétaire, développé par les acteurs habituels publics et industriels privilégiés, sans aucune évaluation extérieure. Dans d'autres pays, ce qu'il est important de souligner est alors l'articulation entre institutions derrière le déploiement de l'application, et son insertion dans une politique sanitaire plus large. Ainsi, en Chine, l'application délivrait ni plus ni moins des autorisations – un passeport – de circulation physique selon son niveau individuel de risque. Les algorithmes et les caractéristiques pris en compte pour élaborer le niveau de risque sont inconnus. En Corée du Sud, le système repose sur la circulation d'informations entre les acteurs du système de santé, la police ou encore les aéroports et le gouvernement.

Dans l'épidémie de Covid-19, jusqu'à présent, l'*evidence-based medicine* (EBM, médecine basée sur les preuves) a été mise à l'épreuve face à d'autres conceptions, dont l'école des pragmatiques, et semble ne pas avoir été en position de contribuer significativement à l'adaptation des politiques de santé. Le passage discret à une santé publique de précision, soutenant une *evidence-based policy* (EBP, politique basée sur les preuves), est, d'un point de vue sanitaire, encore moins convaincant.

Les problématiques autour des données de santé révèlent plus que jamais différentes tensions connues, que l'on peine à dénouer : les tensions au cœur de la santé publique, devant concilier intérêt individuel et populationnel ; les tensions inhérentes au *new public management*, et la porosité sélective, inconfortable, entre public et privé dans la répartition et la réalisation de missions de service public ; les tensions, enfin, entre acteurs participant à la production et à l'utilisation des données. En effet, la promotion et l'utilisation des données sont avant tout organisées par et pour un service public, dont les acteurs sont en concurrence. Parmi eux, les producteurs de données ne veulent pas être dépossédés d'une partie de leur travail, qu'ils n'ont pas les moyens de valoriser souvent seuls, la donnée étant présentée comme une valeur en soi, source d'avantage compétitif (entre hôpitaux par exemple).