Statistique et société
mai 2014

Volume 2, Numéro 2

Données de santé : données sensibles

www.statistique-et-societe.fr

Sommaire
Statistique et Société

Volume 2, Numéro 2

Sommaire
Statistique et Société

Volume 2, Numéro 2

Statistique et société

Magazine trimestriel publié par la Société Française de Statistique.
Le but de Statistique et société est de montrer d'une manière attrayante et qui invite à la réflexion l'utilisation pratique de la statistique dans tous les domaines de la vie, et de montrer comment l'usage de la statistique intervient dans la société pour y jouer un rôle souvent inaperçu de transformation, et est en retour influencé par elle. Un autre dessein de Statistique et société est d'informer ses lecteurs avec un souci pédagogique à propos d'applications innovantes, de développements théoriques importants, de problèmes actuels affectant les statisticiens, et d'évolutions dans les rôles joués par les statisticiens et l'usage de statistiques dans la vie de la société.

Design graphique
fastboil.net

ISSN 2269-0271

Éditorial

Emmanuel Didier
Rédacteur en chef

Chers lecteurs,

« Big data » est aujourd'hui partout[1]. A juste titre, puisque notre vie est de plus en plus capturée, orientée ou protégée par des informations chiffrées en nombre toujours plus grand. Cette question intéresse donc, évidemment, l'amateur de statistique.

Pourtant, en se diffusant, l'expression perd de son sens et se met à désigner à peu près tout et n'importe quoi. Ainsi, une conférence organisée par l'institut Max Planck pour l'Histoire des Sciences de Berlin en novembre de l'an dernier, et intitulée « Historiciser les Big Data », au demeurant passionnante, traitait aussi bien des cartes perforées du début du XXème siècle, que des échanges de données entre USA et URSS pendant la guerre froide ou encore des premières expériences de circulation de données sur l'internet. A croire que tout ce qui est statistique et informatique relèverait du « big data ».

Le dossier de ce numéro propose un premier coup de sonde dans cet océan mais, pour échapper à ces travers, s'efforce de se focaliser sur un sujet précis et délimité : l'accès aux données de santé. Nous allons voir en effet que ces données, parmi lesquelles beaucoup émanent d'un système de sécurité sociale ancien et très organisé, se démultiplient de façon impressionnante – elles semblent assurément « big » aujourd'hui - et qu'en outre elles deviennent de plus en plus aisément accessibles. Mais quelle est la nature exacte de ces données ? Qui y a accès et qui ne devrait pas avoir accès ? Pour quoi faire ? Telles sont les questions auxquels nos auteurs se sont attachés à fournir des éléments de réponse.

Ce dossier est accompagné d'un article méthodologique et d'une présentation des débats tenus aux Cafés de la statistique, qui s'avèrent tous les deux liés au sujet du dossier. La méthode multimodale, en plein essor, oblige à évaluer la qualité des données récoltées selon plusieurs modes de collecte, comme le sont très souvent les « big data ». De son côté, le Café de la statistique a traité, sous deux angles différents, des effets sociaux du vieillissement de la population, assurément une question de santé publique.

Nous vous souhaitons bonne lecture, comme d'habitude nous vous invitons à réagir plume à la main, et nous vous donnons rendez vous pour le prochain numéro qui constituera un hommage à Alain Desrosières, disparu en 2013, inspirateur de la pensée sur les liens entre Statistique et société et qui, tant qu'il l'a pu, a siégé dans le comité éditorial de la revue.

1. Les administrations françaises doivent dire : « Données massives »

Données de santé : données sensibles
Présentation du dossier

Jean-Christophe Thalabard[1]
et Marius Fieschi[2]

La capacité à recueillir de manière objective des signes sur un grand nombre de patients et à croiser ces informations (tables de contingence) a marqué au 19ème siècle la naissance d'une médecine scientifique moderne [Foucault, 1963]. La mise en place de registres de maladie au 20ème siècle, l'essor de l'épidémiologie et, dans le champ de la thérapeutique, le développement de la médecine reposant sur les preuves (Evidence Based Medicine - EBM) en représentent des prolongements naturels, que les possibilités techniques permettent d'étendre au delà du champ médical restreint en les délocalisant, les dématérialisant et en ouvrant les possibilités de partages effectifs en temps réel ou presque de cette information. La gestion des données de santé, longtemps laissée aux seuls professionnels sur un mode papier traditionnel, d'intérêt relativement limité et considérée comme une contrainte légale (cf. les archives des dossiers médicaux dans les hôpitaux), s'est transportée sur de nouveaux supports et est considérée de nos jours comme une ressource, objet d'intérêt croissant.

Des actions incitatives, fortement stimulées par la nécessité d'améliorer les conditions d'enregistrement, de suivi et de remboursement des prestations de soins, ont vu les professionnels de santé s'informatiser tant dans le secteur de l'exercice libéral que dans le secteur hospitalier, générant au fur et à mesure des bases de données multiples aux finalités diverses, très hétérogènes dans leur qualité, et dont une partie de la gestion est de plus en plus confiée à des sociétés privées du secteur concurrentiel. Les media se font largement l'écho d'un souhait "d'une ouverture" de ces données de santé au plus grand nombre, propre à anticiper, voire éviter quelques scandales sanitaires, comme plaidé par certains [de Kervasdoué et Sicard 2013, Bar-Hen et Flahault, 2013], point de vue qui fait par ailleurs l'objet de discussions plus nuancées [Cabut, 2013, Larousserie, 2014].

Cette ouverture apparaît répondre à des objectifs variés : classification comparative de l'offre de soins permettant, en théorie, au consommateur de soins d'intervenir dans le choix de son soignant et/ ou de son lieu de prise en charge sur des critères qui lui seraient propres ; analyse des comportements de soins afin de pouvoir investir et au besoin influer sur ces comportements dans une logique industrielle concurrentielle ; analyse des disparités géographiques en termes d'offre de soins et harmonisation des pratiques professionnelles par rapport à un référentiel de recommandations élaboré par des acteurs du soin, tel que

1. Membre du comité de rédaction de Statistique et Société
2. Professeur de statistiques, information médicale, ancien directeur du Laboratoire d'Enseignement et de Recherche sur le Traitement de l'Information Médicale, Université de la Méditerranée, Marseille

proposé par la Haute autorité de santé dans l'optique d'une amélioration de l'efficience des pratiques (rapport efficacité bénéfice/ coût), etc. Enfin, la connaissance de la maladie est une finalité, composante importante de cette demande, que ce soit, en chronique, dans l'étude de l'histoire naturelle de l'évolution des pathologies, avec ou sans traitement, en population dans leur environnement géographique climatique, professionnel, social, ou, en aigu, dans la capacité à détecter des signaux d'émergence de formes particulières de pathologies (détection d'épidémies, malformations, anomalies génétiques, etc.)

Une confusion dans la nature des données disponibles

Avner Bar- Hen, enseignant-chercheur statisticien, ancien enseignant à l'École des Hautes Études en Santé Publique, insiste sur l'ouverture dite "citoyenne" des données publiques quelles que soient leurs natures, dont celles relatives à la santé, tout en rappelant les différents types et soulignant bien la différence entre la fouille de données orientées santé dans des grandes bases d'information croisées (Big data) et l'accès ouvert aux données individuelles d'ordre administratif, pour peu qu'elles aient été convenablement anonymisées au préalable. La question du contrôle d'une part de l'information partageable et d'autre part des personnes habilitées à accéder aux dites informations reste un sujet de débat important. La réponse ne peut être binaire et le débat doit être libéré d'une opposition formelle entre les supposés anciens et les modernes auto-déclarés. Des décisions récentes concernant l'accès prises par la CADA[3] [Daël, 2013] marquent l'évolution de la jurisprudence.
Face à cette position volontiers polémique, propre à ouvrir le débat et susciter des réactions, Statistique et Société a souhaité rassembler des contributions de professionnels des bases de données de santé publiques ou semi- publiques confrontés à cette nouvelle demande.

Les bases de données de la Sécurité sociale sont au centre de bien des controverses

La France bénéficie d'un système de santé original, né sur un principe de solidarité au décours de la seconde guerre mondiale (Sécurité Sociale et assurances complémentaires). Ce système médico-administratif produit de grandes bases de données centralisées. La revendication d'une ouverture large de ces données au plus grand nombre, formulée par certains, témoigne parfois d'une méconnaissance des possibilités existantes. Le point de vue de l'opérateur principal (Caisse nationale d'assurance-maladie des travailleurs salariés - CNAMTS) à travers la contribution de Dominique Polton et Claude Gissot du service des statistiques et de l'évaluation économique de la CNAMTS montre des avancées concrètes, illustrées par des contributions importantes en termes de connaissance, de croisements possibles, tout en soulignant la prudence nécessaire du fait des limites de ces bases.

Spécificité des données de santé

L'existence d'une spécificité des données de santé qui déborderait la réflexion sur le secret statistique est abordée par la contribution de Jean-Pierre Le Gléau. Il rappelle la raison d'être du secret statistique, son évolution et les efforts déployés pour le contrôler sans entraver le travail des chercheurs et des utilisateurs des données, en amenant quelques pistes techniques qui pourraient s'appliquer aux données de santé. Au delà des aspects techniques, l'accès à des données sensibles touchant à la vie privée soulève les questions importantes des garanties apportées à chacun de protection de sa propre intimité. La particularité de notre époque est bien que la donnée est devenue dématérialisée et voyage facilement, exigeant une réflexion et des décisions appropriées. L'accès aux données de santé expose, en cas de divulgation consciente

3. Commission d'accès aux documents administratifs

ou inconsciente, les personnes concernées à des risques de stigmatisation avec toutes les conséquences éventuelles psychologiques et sociétales, potentiellement dramatiques, qui touchent à une dimension particulière des individus, au delà de l'économique, et ce d'autant plus qu'elle concerne le sujet malade, donc plus vulnérable. Il nous a paru important de donner la parole à une juriste qui a travaillé sur ces sujets à la CNIL et est aujourd'hui liée à la structure de coordination dédié à l'e- santé, l'ASIP (Agence des Systèmes d'Information Partagés en Santé), apparue dans le paysage depuis quelques années http://esante.gouv.fr. Jeanne Bossi nous rappelle les principes et règles encadrant les données sensibles dans le cadre français, au sein du concert européen [Directive, 2013]. Est particulièrement importante la distinction entre données directement ou indirectement identifiantes, d'autant que les avancées algorithmiques permanentes rendent de plus en plus complexe cette délimitation [Schadt et al., 2011, Schadt, 2012, White, 2013].

L'intérêt de données de suivi individuel sur le long terme et du croisement des bases d'information

L'histoire de la santé est riche d'avancées liées à l'accès à de telles données, dont l'importance n'est pas forcément liée à la multiplicité des variables recueillies mais surtout à la qualité d'un recueil systématique dans une population bien définie. Dès le 18ème siècle, Spallanzani avait pu mettre en évidence des liens entre occupation professionnelle et pathologies, plus récemment S Barker a pu montrer le lien entre certaines anomalies simples staturo-pondérales à la naissance et l'apparition, plus de 50 ans plus tard, de pathologies cardio-vasculaires et métaboliques [Barker et al., 1989, Barker et al., 1993], posant indirectement la question de l'utilisation de données pour d'autres finalités que celles pour lesquelles elles avaient été initialement recueillies. La contribution de Marie Zins et Marcel Goldberg, avec l'exemple de la cohorte Constances, nous montre l'évolution des pratiques des épidémiologistes dans la constitution de grandes cohortes s'appuyant sur les bases médico-administratives existantes pour les mises à jour et l'amélioration de la qualité des données recueillies au cours du temps. L'objectif ici est bien la qualité des données recueillies dans une finalité très générale de mettre à disposition un outil pour des questionnements précis futurs, véritable entreprise technique ouverte. L'analyse d'une telle masse de données ne se résume plus à des méthodes simples bien maitrisées mais peut nécessiter des techniques d'imputation diagnostique sophistiquées, qui ne sont plus propres au domaine de santé.

Des enjeux et jeux d'acteurs sociaux importants

L'ensemble de ces questionnements est loin de se limiter à de simples aspects techniques, ou aux réticences de certains acteurs à ouvrir leurs informations. La contribution de François Briatte et Samuel Goëta nous éclaire sur les aspects sociologiques et sur les enjeux sous-jacents. Concernant l'atteinte à la vie privée notamment, toute la question est bien de savoir si les avantages d'une fouille de bases de données comme celles de la CNAMTS ayant pour finalité revendiquée une amélioration de la qualité des soins, que ce soit l'efficacité dans des choix de prise en charge au niveau individuel ou l'efficience en terme de dépenses de santé pour rester sur le modèle économique d'un système solidaire, justifie le petit risque assumé d'une rupture de cette confidentialité pour quelques uns ou plus ? Quelle est la perception de cette question par la génération Y, voire la suivante élevée dans la banalité du cloud, de l'i-phone et des réseaux sociaux, où finalement la donnée personnelle se partage sans arrière pensée particulière en la confiant, dans un premier temps assez naïvement, à des prestataires de service ne garantissant pas la maitrise des usages qui pourraient en être faits et le droit à l'oubli ? Quelle est alors la connaissance et la perception de la loi actuelle (loi de 1978 modifiée en août 2004) qui insiste beaucoup sur des règles qui concernent les données identifiantes que ce soit directement mais également indirectement ?

Les évolutions en cours et à venir

Nous vivons une période d'intenses débats et échanges où la volonté d'ouverture et d'ajustement au monde actuel et à ses capacités techniques est tangible, comme le montrent les travaux de commissions parlementaires consultables sur internet et la disponibilité récente du rapport Bras [Bras et Loth, 2014]. Le présent dossier vise à apporter quelques éléments factuels propres à nourrir le débat.

Références

[1] Bar-Hen A. et Flahault A.. Donnons aux citoyens accès aux données de santé. Le Monde, 8 Mars 2013.
[2] Barker D. J., P. D. Winter, C. Osmond, B. Margetts, and S. J. Simmonds. Weight in infancy and death from ischaemic heart disease. Lancet, 2 (8663): 577-580, Sep 1989.
[3] Barker D.J., C. N. Hales, C. H. Fall, C. Osmond, K. Phipps, and P. M. Clark. Type 2 (non-insulin-dependent) diabetes mellitus, hypertension and hyperlipidaemia (syndrome x): relation to reduced fetal growth. Diabetologia, 36 (1): 62-67, Jan 1993.
[4] Bras P.L.et A. Loth. Rapport sur la gouvernance et l'utilisation des données de santé. Ministère des Affaires Sociales et de la Santé, IGAS, 2014.
[5] Cabut Sandrine Santé publique: les leçons du modèle scandinave. Le Monde, 1° Février 2013.
[6] Daël. S. Avis de la Cada 20134348 du 21/11/2013, CADA Novembre 2013.
[7] de Kervasdoué J. et D. Sicard. Plus grave que le débat sur la pilule: l'affaire des données de santé publique. Le Monde, 16 janvier 2013.
[8] Directive 2013/37/EU du parlement européen et du conseil du 26/06/2013 modifiant la directive 2003/98/EC concernant la réutilisation des informations du secteur public. Journal Officiel de l'Union européenne.
[9] Foucault M. Naissance de la Clinique. PUF, 1963. ISBN : 978-2-13-057865-9.
[10] Larousserie David. Données sensibles. l'équation impossible. Le Monde, 9 avril 2014.
[11] Schadt E. E. The changing privacy landscape in the era of big data. Mol Syst Biol, 8: 612, 2012.
[12] Schadt E. E., M. D. Linderman, J. Sorenson, L. Lee, and G. P. Nolan. Cloud and heterogeneous computing solutions exist today for the emerging big data problems in biology. Nat Rev Genet, 12 (3): 224, Mar 2011.
[13] Weitzman E. R., L. Kaci, and K. D. Mandl. Sharing medical data for health research: the early personal health record experience. J Med Internet Res, 12 (2): e14, 2010. doi: 10.2196/jmir.1356. URL http://dx.doi.org/10.2196/jmir.1356.
[14] White S. E.. De-identification and the sharing of big data. J AHIMA, 84 (4): 44-47, Apr 2013.

Données de santé : le temps de la maturité et de l'ouverture ?

Avner Bar-Hen
Professeur de Statistiques, Université Paris Descartes, Pôle de recherche et d'enseignement supérieur Sorbonne Paris Cité

La santé : un bien public

L'industrialisation et l'émergence des préoccupations sanitaires ont favorisé la mise en place d'une protection des travailleurs dès le milieu du 19ème siècle (protection des salariés les plus faibles avec la loi de 1841, inspirée du rapport Villermé, puis lois sur l'inspection du travail, les accidents du travail, les maladies). La protection n'a cessé de se renforcer tout au long du 20ème siècle. De par son étymologie, la statistique a été, dès son origine, un "outil de gouvernement" et il n'est donc pas surprenant que la santé soit l'un des domaines importants en termes de collecte des données.

Le lien entre politique publique et données fiables a été abondamment discuté (voir par exemple A. Desrosières : Histoire de la raison statistique[1]) et l'importance d'une information de qualité en santé est primordiale pour préserver un système de soins juste et efficace. Ceci est d'autant plus important dans un contexte (i) d'espérance de vie en constante augmentation, (ii) d'une spécialisation des traitements et des examens qui augmentent sensiblement les coûts et (iii) du rétrécissement du périmètre de l'Etat et de la puissance publique par des économies budgétaires.

La loi, notamment en France, encourage les citoyens à s'impliquer dans les données publiques. En effet, la loi du 17 juillet 1978 :

reconnaît à toute personne un droit très large d'obtenir communication des documents détenus par une administration, quels que soient leur forme ou leur support.

Ainsi, au nom de la citoyenneté et de la transparence des actions des gouvernements, de l'amélioration de la gestion du territoire, ou encore de la promotion de l'innovation, les données publiques ont donc une vocation naturelle à être partagées. Leur ouverture est d'autant plus facilitée qu'elle peut répondre à des objectifs politiques ce qui inclut évidemment la santé.

Une mutation numérique qui bouleverse les codes et les habitudes

L'arrivée massive du numérique dans notre quotidien tend à créer des nouveaux modes d'implication dans l'espace public. Il en résulte un affaiblissement des corps intermédiaires dans les rapports sociaux et politiques. Les politiques publiques ne sont plus l'apanage des responsables politiques, marquant un changement de paradigme important dans nos démocraties. Beaucoup d'acteurs s'efforcent de construire des propositions sans passer par la "case" parti et revendiquent d'incarner l'intérêt général sans passer par la "case" État. La situation de quasi-monopole de la production de propositions politiques détenue par les

1. DESROSIERES, A., 2010, La politique des grands nombres. Histoire de la raison statistique, La Découverte/Poche, Paris, 3ème édition.

partis s'écorne. En parallèle l'élévation continue du niveau d'étude donne aux citoyens des outils (et crée des attentes) pour appréhender les questions de santé. La suite logique est donc une réappropriation de l'action publique et non plus une délégation aux responsables politiques, surtout lorsque leur capacité d'action est limitée et onéreuse. Il est donc naturel de voir se développer une demande de démocratie scientifique et des politiques volontaires sur l'ouverture des données publiques. A titre d'exemple, l'épisode de pollution de mars 2014 dans les grandes villes a montré la difficulté pour le citoyen de penser qu'il est une partie éventuelle de cette pollution par son comportement et il y a une forme de lâcheté à laisser aux politiques le soin de prendre des mesures très onéreuses pour un résultat limité d'autant plus que la volonté politique se heurte au poids de lobbys parfois puissants. La faiblesse de la mobilisation sur les effets délétères du diesel traduit sans doute la difficulté du positionnement des professionnels de santé et le faible poids de leur ministère de tutelle pour se faire entendre vis à vis d'acteurs économiques publics et privés. A contrario l'émergence de l'épidémie VIH a démontré la capacité de mobilisation importante du mouvement associatif concerné amenant à une forme originale de "partenariat" entre les associations de patients et les professionnels de santé, avec une réelle efficacité dans l'évaluation scientifique des traitements anti- rétro- viraux et le contrôle de la maladie, même si les conditions socio- culturelles du développement de l'épidémie ne sont pas directement extrapolables à d'autres situations morbides, notamment en terme d'égalité d'accès aux traitements.

La santé, un monde à part dans cette évolution ?

L'un des domaines qui rencontre le plus de résistance à ce mouvement est la santé. Les arguments les plus fréquemment avancés sont la complexité des données, la garantie du secret médical et de la protection des données personnelles. Il y aurait donc une spécificité des données de santé.

La protection des données personnelles est un domaine très largement travaillé par la statistique publique et si la question est loin d'être anodine, il y a nombre d'experts et d'expériences au sein de l'administration française. Le but est d'avoir une information de qualité avec une protection stricte et rigoureuse des règles de confidentialité et de respect des libertés individuelles. Pourquoi ne pas s'inspirer des travaux qui permettent d'accéder à la production industrielle, aux revenus des ménages, aux votes ou à nombre d'informations à des niveaux de précision tout à faits pertinents pour l'ensemble des acteurs ? Différents niveaux de recueil existent, et les données ne sont publiées qu'avec un nombre minimal. Ce concept de degré d'agglomération est classique en statistique publique. Le suivi statistique sur une certaine durée soulève, lui, un tout autre problème, qui est celui de pouvoir suivre le parcours d'un certain nombre d'individus sur un certain laps de temps sans avoir à connaître leur identité véritable. Sous réserve d'une extension de ses compétences, la commission du secret statistique pourrait représenter un acteur central dans ce débat.

L'OpenData n'est pas le Big Data

Il y a une confusion classique entre l'OpenData, qui implique de manière très claire l'absence de données personnelles et le respect du secret statistique, d'un côté, et, de l'autre, le "Big Data", ce secteur émergent de "l'économie numérique" qui se nourrit quasi-exclusivement de données à caractère personnel et qui alimente à juste titre d'importantes questions sur le respect par ses entreprises de la vie privée des citoyens. Si cette confusion est logiquement entretenue par les lobbyistes du "Big Data", elle l'est également par les opposants à la transparence démocratique qui abusent du faux prétexte de données personnelles pour refuser d'ouvrir des données absolument non sensibles. En réalité l'Open Data est à l'origine d'une prise de conscience sur un meilleur respect des données personnelles en permettant de détecter le non-respect des

règles en vigueur en termes d'anonymisation et de respect des données personnelles. Ceci a pu être vu dans deux cas récents :
1/ les données du système d'immatriculation des véhicules (cartes grises) ont été vendues en toute illégalité pendant des années et la légalisation de leur revente n'a pas réglé le problème de divulgation des données personnelles[2]
2/ à l'occasion des travaux préparatoires au passage en Open Data, des problèmes d'anonymisation ont pu être identifiés et finalement corrigés[3].

Le secret médical n'est pas que du ressort des bases de données publiques. A côté de la collecte publique se développe un nombre important de nouvelles données soit personnelles[4], soit indirectes (Google trends). Le risque donc est une dévaluation de la donnée publique au profit d'une information pas toujours vérifiée et pas toujours de qualité. Il serait certes inadmissible que l'on puisse identifier le porteur d'une maladie chronique à partir des données médicales mais dans le même temps on ne s'interroge qu'assez peu sur la détection de cette même maladie à partir d'un téléphone portable. Le magasin américain Target a pu identifier des femmes enceintes à partir de leur profil d'achat sans avoir aucunement besoin de données médicales. L'exemple du père se plaignant des publicités reçues et découvrant par ce biais la grossesse de sa fille a plus été vu comme un succès algorithmique que comme une intrusion dans la vie privée.

Ceci illustre une autre source de confusion : la protection de l'anonymat recouvre en réalité la protection de la vie privée des individus. Le consentement éclairé, la possibilité d'exclusion des bases de données devrait donc être tout autant en débat que l'anonymisation. Il est aussi pertinent de se demander qui doit être le garant de la vie privée. Le désengagement de l'Etat, sa versatilité[5] dans la gestion des données peuvent amener à se poser des questions sur son rôle naturel. Doit-on avoir plus confiance dans des organismes de recherche alors, qu'au contraire des services ministériels, la certification morale et technique est inexistante ? L'expertise des organismes de recherche ne se résume-t-elle pas à une auto-affirmation, qui mérite sans doute d'être questionnée régulièrement ?

L'OpenData est utile à la démocratie

Les récentes affaires de santé publique (dont la plus emblématique est celle du Médiator) montrent les limites de la gestion de l'information en santé dans notre pays. Rappelons que la plus grande partie des données recueillies par le système de santé est destinée, au départ, à sa seule gestion financière (en vue du remboursement des actes ou de la tarification de l'activité dans les hôpitaux).

Par ailleurs, ces données ne concernent que la part de la médecine prise en charge par les organismes sociaux, et elles ne sont disponibles que de manière très restreinte et parfois avec un retard de plusieurs années. Enfin, des procédures d'anonymisation apportent déjà un niveau très élevé de garantie de l'anonymat du traitement de l'information, au moins égal à celui mis en place dans nos pays voisins d'Europe du Nord.

Il est aujourd'hui possible de connaître les prescriptions mensuelles de chaque généraliste en Grande-Bretagne avec moins de six mois de délai, alors que les honoraires des professionnels de santé ne sont plus actualisés depuis 2010 en France. L'accès aux données de santé permettra

2. http://www.senat.fr/questions/base/2012/qSEQ120700184.html
3. http://www.autorite-statistique-publique.fr/pdf/Actualites/ASP_PV_18_04_13.pdf
4. Le quantified self: nouvelle forme de partage des données personnelles, nouveaux enjeux ? Sophie Vuilliet-Tavernier CNIL http://www.cnil.fr/fileadmin/documents/La_CNIL/publications/DEIP/LettreIP_5.pdf
5. Pour l'exemple anglais voir : http://www.theguardian.com/commentisfree/2014/jan/20/nhs-patient-care-data-policy-medical-information

peut-être d'anticiper des problèmes, mais offrira aussi une information a posteriori : tel médicament est-il plus prescrit dans telle région ou telle ville ? Plutôt aux personnes âgées ? Aux hommes ? Il s'agit de conceptions socioculturelles différentes selon les pays. Tant que l'activité médicale sera tarifée à l'acte, il n'y aura pas de place pour une médecine sociale qui pourrait par exemple comprendre (comme en Angleterre) une rémunération pour la participation à la collecte des données de santé.

Revenons enfin sur la complexité des données : en quoi les données de santé sont-elles plus compliquées que celles de n'importe quel ministère, les données d'organismes internationaux, d'un moteur de recherche ou d'un réseau social (sans parler des questions de confidentialité dans ce dernier exemple) ?

Aujourd'hui, l'accréditation d'accès aux données de santé nécessite un décret ministériel, au même titre que l'accès à des terrains militaires, en raison des exigences de sécurité. L'accès aux données publiques de santé est restreint même une fois celles-ci totalement anonymisées, et cela doit changer ! Il n'est pas normal que seulement les "bien-pensants" et les "bien-sachants" triés par les organismes sociaux détenteurs de ces informations aient droit d'accéder aux données publiques de santé. Il faudrait que tout citoyen puisse réclamer l'accès permanent à ces données publiques anonymes. La collectivité accepte une restriction des libertés pour le bien commun : c'est le principe de l'obligation de la ceinture au volant ou de l'interdiction des armes dans les avions. A moins de considérer tout utilisateur de données comme un criminel en puissance, il semble plus raisonnable de prévoir des sanctions pour une utilisation abusive de ces données (sinon il faut aussi interdire la vente des fourchettes qui peuvent servir à crever les yeux des passants dans la rue). Les données de santé représentent un bien social qui n'est pas la propriété de d'agence, d'unités de recherche ou de gourous autoproclamés spécialistes universels.

La dimension économique de la santé

La concomitance du débat sur l'ouverture des données de santé avec la diminution des budgets de recherche voire la mise en cause de programmes de prévention et de surveillance est troublante. Il est légitime de relier ce mouvement d'ouverture des données à une tendance profonde qui considère la santé comme un produit économique classique. Les données de santé représentent clairement un bien économique par leur potentielle réutilisation dans un contexte marchand : il y a donc une forme de symétrie entre les données de l'assurance maladie et les informations que détiennent les industriels à travers les données de vente. A titre d'exemple, relier les ventes aux profils de prescription est (entre autres) un enjeu économique Au lieu de refuser toute évolution du système, ne vaut-il mieux pas se demander comment les gains attendus peuvent être partagés entre les différents acteurs, dont les patients et le secteur public (prévention, recherche, etc.) ?

L'OpenData, une source de progrès pour la santé

Le suivi en temps direct du fonctionnement des moteurs d'avion permet une bien meilleure gestion des pannes. De la même manière, un suivi individualisé a de grandes chances d'être plus efficace qu'un profilage approximatif qui autorise des raccourcis plus ou moins douteux. Au lieu de proposer des dépistages pour tous comme le dépistage du cancer colorectal à 50 ans ou le dépistage de la trisomie 21 par amniocentèse à partir de 38 ans, il est peut- être possible d'arriver à une prévention mieux ciblée, plus efficace et moins onéreuse, évitant de stigmatiser telle ou telle sous-population.

Au-delà d'obscurantistes de la dernière heure accrochés à leurs privilèges, n'avons-nous pas un réel problème de compétence ? Lorsqu'une plate-forme comme Etalab ou datagouv ne possède même pas une interface de programmation (API) pour utiliser les données, pouvons-nous réellement parler de données ouvertes ? Vaut-il mieux financer de nouvelles études ou plutôt chercher à développer l'utilisation des données existantes en réunissant des spécialistes de l'utilisation de la donnée publique, de leur maniement et de leur analyse ? Une publication comme ce numéro du journal Statistique et Société de la SFdS est une opportunité pour faire avancer ce débat et la SFdS a une place naturelle sur cette question importante.

Le débat disparaîtra avec les derniers dinosaures des temps anciens. Cacher une difficulté à gérer une situation nouvelle en se réfugiant derrière un maintien de dogme dépassé peut prêter à sourire. Mais si les discours fondés sur ces données avaient pu circuler un peu plus facilement, un temps précieux au bénéfice de la santé publique aurait peut-être pu être gagné, tout en évitant nombre d'affaires qui sapent la confiance des Français en leur système de santé chaque fois que des lanceurs d'alerte, plus ou moins bien intentionnés et/ ou indépendants, expliquent que les données existaient, mais qu'elles n'ont pas été utilisées. Notons qu'ouvrir les données n'est pas suffisant. Dans le cas du Mediator, la Caisse régionale d'assurance-maladie de Bourgogne avait ainsi visiblement pu lancer une alerte mais l'impression d'un mélange constant des genres fait qu'il est difficile de démêler les volontés de suspendre des remboursements pour des questions de santé publique ou pour des questions budgétaires. Le mélange de l'approche médicale, comptable et parfois scientifique est encore une fois délétère.

Il est temps de faire confiance aux citoyens et qu'ils puissent se réapproprier leur propre système de santé, car c'est la condition de sa survie, dans un respect et une confiance mutuels regagnés entre tous ses acteurs. Si, au nom de principe douteux, quelques spécialistes autoproclamés veulent garder le monopole des données, il serait temps de leur demander des comptes !

Les bases de données de l'assurance-maladie : un potentiel pour l'amélioration du système de santé et pour la recherche

Claude Gissot et Dominique Polton
Service des statistiques et de l'évaluation économique
Caisse nationale d'assurance-maladie des travailleurs salariés

Les bases de données de l'assurance-maladie sont rassemblées dans un système national qui ouvre des possibilités remarquables en santé publique. Ce système a en effet des utilisations multiples : il permet d'examiner la dépense de soins selon des critères médicaux, d'analyser des parcours de soins, de connaître l'efficacité de traitements en vie réelle, et peut contribuer à surveiller la sécurité des médicaments. Ce système est de plus en plus utilisé, notamment par les chercheurs, et la Caisse nationale d'assurance-maladie s'est organisée pour faciliter ces utilisations. Un débat a été lancé récemment sur l'intérêt et les enjeux d'une ouverture plus large des données, et les modalités d'accès pourraient être amenées à évoluer.

La création d'un Système National Inter-Régimes d'Assurance Maladie (SNIIRAM) a été prévue par la loi de financement de la sécurité sociale pour 1999. Après plusieurs années de travail technique et l'accord de la CNIL, un entrepôt de données a été constitué à partir de 2003, puis complété et enrichi au fil des années, constituant aujourd'hui une source d'information très riche sur la santé de la population et le fonctionnement du système de soins.

Ce que contient le SNIIRAM

Le SNIIRAM est une base de données qui décrit les soins fournis à chaque individu, examens diagnostiques, interventions médicales, médicaments, soins infirmiers,..., à l'aide des informations contenues dans les feuilles de soins ou les factures des cliniques. Les bénéficiaires des soins sont repérés par un identifiant anonyme pérenne commun avec le PMSI (programme de médicalisation des systèmes d'information) sur les séjours hospitaliers : il est ainsi possible de reconstituer le parcours de soins des patients et la chronologie des événements de santé qu'ils reflètent. Les professionnels et établissements qui délivrent ces soins sont identifiés dans la base de données, qui comporte également des informations concernant la personne bénéficiaire des soins, notamment celles qui sont en rapport avec le remboursement, comme le bénéfice d'une affection de longue durée, ALD (ouvrant droit à l'exonération du ticket modérateur) ou de la CMUC (couverture maladie universelle complémentaire).
On dispose ainsi, pour l'intégralité de la population française, de renseignements sur les pathologies traitées (diagnostics précis en cas d'hospitalisation ou en cas d'ALD) et sur les soins fournis, avec un grand degré de détail (codes des médicaments prescrits, descriptions des actes médicaux réalisés, praticien ayant réalisé l'acte,...). Ces données peuvent donner lieu à de multiples utilisations, que nous illustrerons plus loin par quelques exemples.
La puissance statistique, l'absence de perdus de vue ou de biais d'enregistrement, la précision de la chronologie sont des points forts de cette base de données. Il y a bien sûr aussi des limites pour certaines études, car elles ne comprennent pas de données cliniques (exhaustivité des

diagnostics, données staturo-pondérales et index de masse corporelle (IMC), niveau tensionnel quantifié,...), paracliniques (résultats d'examens), ni d'informations sur les antécédents ou facteurs de risque (tabac, alcool,...), et peu de données sociales. Néanmoins ces limites peuvent être dépassées, notamment par le biais d'appariement avec d'autres sources. Actuellement un appariement exhaustif est réalisé en routine pour intégrer des données sur les séjours hospitaliers (venant du PMSI) et les dates de décès ; un pilote a été réalisé, avec des résultats très positifs, pour l'appariement avec les causes de décès, et de nombreux appariements ponctuels ou réguliers sont réalisés à la demande d'équipes de recherche avec des enquêtes, cohortes, registres,...

Des utilisations multiples

Les usages que l'on peut faire de ces données sont potentiellement très nombreux. Elles peuvent être utilisées pour l'analyse des pratiques de soins, la régulation du système de santé, la surveillance des effets des médicaments, la recherche épidémiologique,....
Quelques types d'utilisations sont illustrés ci-dessous par des exemples issus des travaux que l'assurance maladie réalise elle-même. Ce ne sont pas les seuls, loin de là, car un nombre croissant d'administrations et de chercheurs se sont investis dans les années récentes pour pouvoir les utiliser pour leurs propres besoins, comme nous le verrons en conclusion.

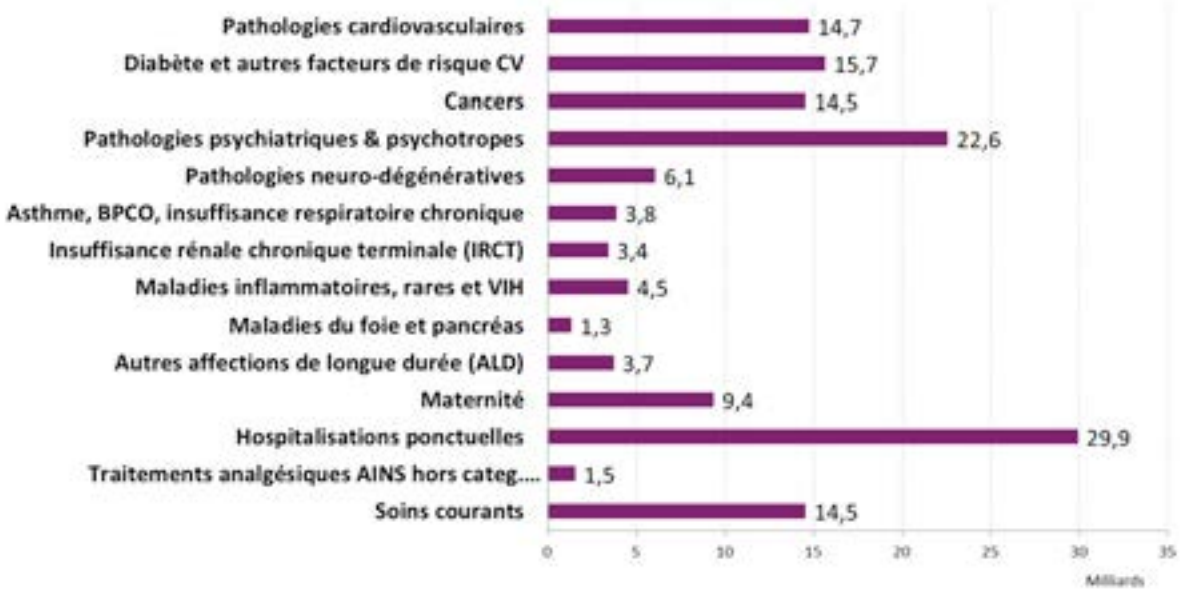
Cartographier les pathologies et médicaliser l'analyse de la dépense

La CNAMTS a entrepris de développer des algorithmes permettant d'identifier les pathologies ou des facteurs de risque dont souffre la population, à partir des traitements qui sont prodigués: diagnostics posés au cours d'une hospitalisation ou enregistrés par les médecins conseil à l'occasion d'une mise en ALD, médicaments traceurs, etc.
Ces données n'ont pas, bien sûr, une valeur épidémiologique identique à celle d'enquêtes ou de registres qui permettent d'évaluer les prévalences de manière précise : elles ne repèrent, par construction, que les patients pris en charge pour ces pathologies, avec les biais que cela peut entraîner ; elles sont dépendantes de règles administratives qui peuvent évoluer. Mais à l'expérience, dans beaucoup de domaines, la photographie de l'état de santé et les tendances d'évolution qu'elle permet d'établir corroborent les données épidémiologiques disponibles, et elles peuvent être utilisées pour beaucoup d'analyses avec un coût de collecte et de traitement très avantageux.
Pour que ces algorithmes puissent être discutés, améliorés et partagés au sein de la communauté des chercheurs, mais aussi des administrations et agences de santé, plusieurs initiatives ont été prises :

- un réseau informel, plutôt piloté par l'Inserm, rassemble les organismes qui ont vocation à utiliser de tels algorithmes pour leurs besoins propres (équipes de recherche, agences, CNAMTS, Haut conseil de santé publique,...), avec l'objectif de mutualiser les expertises et de mettre en commun les requêtes statistiques élaborées ;
- un travail d'analyse critique du premier jeu d'algorithmes construit par la CNAMTS a été demandé à un département d'information médicale (DIM) d'un centre hospitalier, les conclusions de ce travail pouvant constituer une base pour le travail collectif évoqué ci-dessus.

Cette analyse par pathologie permet déjà, par exemple, une approche nouvelle de la dépense de santé et d'assurance maladie, avec une grille de lecture qui renvoie à sa finalité première – prendre en charge les problèmes de santé de la population – plutôt qu'au raisonnement habituel par catégorie d'offres de soins (soins hospitaliers, soins de ville, médecins, dentistes, infirmiers, médicament...). On peut l'illustrer dans le graphique ci-dessous, qui montre la répartition de l'objectif national des dépenses d'assurance maladie[1] par groupe de pathologie ou traitement.

1. Hors dépenses médico-social et dotations forfaitaires



Graphique 1. Répartition des dépenses de l'assurance-maladie en 2011 (146 milliards d'euros)

Au-delà de cette approche très "macro", il est possible d'analyser les parcours de soins de patients pour une pathologie donnée, leur conformité aux recommandations, leur variabilité sur le territoire. Un exemple, parmi beaucoup d'autres, en est donné ci-dessous.

Analyser les parcours de soins : exemple des thyroïdectomies pour cancer ou nodule bénin

Environ 40 000 thyroïdectomies sont réalisées chaque année en France. Il s'agit de femmes dans 80% des cas, âgées en moyenne de 51 ans.
L'ablation de la thyroïde est en principe réalisée devant la découverte d'un cancer, d'un nodule de la thyroïde pour lequel les résultats de la ponction préalable à l'intervention sont douteux (probabilité de cancer faible mais non nulle), devant certaines formes de goître et d'hyperthyroïdie.
Selon les recommandations de la Haute autorité de santé, l'intervention n'est pas justifiée en cas de nodule bénin. Pour les cancers de très petite taille, l'intervention est très discutée, car elle n'est pas sans conséquences, alors que le cancer de la thyroïde est un cancer le plus souvent peu invasif.
Avant une thyroïdectomie pour nodule, les patients devraient idéalement avoir subi une échographie et un examen biologique (TSH), puis dans la plupart des cas une cyto-ponction (afin de porter l'indication opératoire ou de se livrer à de la simple surveillance) en cas de nodule suspect à l'échographie.
Les recommandations s'accordent sur le fait que la cyto-ponction fait partie des examens qui doivent orienter la décision thérapeutique. Les données du SNIIRAM confirment en effet que la chirurgie pour nodule bénin n'est pas anodine :

- tous les patients ayant une thyroïdectomie totale doivent avoir un traitement par hormones thyroïdiennes à vie, et 44% de ceux ayant une thyroïdectomie partielle en ont également un. Ce traitement nécessite un suivi régulier et peut avoir parfois des répercussions désagréables sur la vie quotidienne en cas de déséquilibre hormonal (troubles de l'humeur, fatigue, frilosité...),

- 4% des patients ont des répercussions sur le fonctionnement d'une corde vocale qui peut nécessiter ensuite des séances d'orthophonie et 1% une atteinte définitive des glandes parathyroïdes qui requiert un traitement par calcium à vie,
- sans compter d'autres complications possibles (cicatrices).

Or l'analyse des données de l'assurance maladie montre qu'aujourd'hui 69% des patient opérés n'ont pas subi de cyto-ponction avant intervention chirurgicale. A l'inverse, les dosages hormonaux sont fréquents, mais pas tous adaptés. La pratique de la cyto-ponction, insuffisante en moyenne, varie par ailleurs fortement d'une région à l'autre. Si globalement en France aujourd'hui, pour 4 cancers on opère 5 nodules bénins, ce ratio est également très différent d'une région à l'autre.

Ce type d'analyse est aujourd'hui reproduit sur de nombreux traitement et épisodes de soins, et permet de mettre au point des programmes d'action pour homogénéiser les pratiques et améliorer le respect des recommandations médicales.

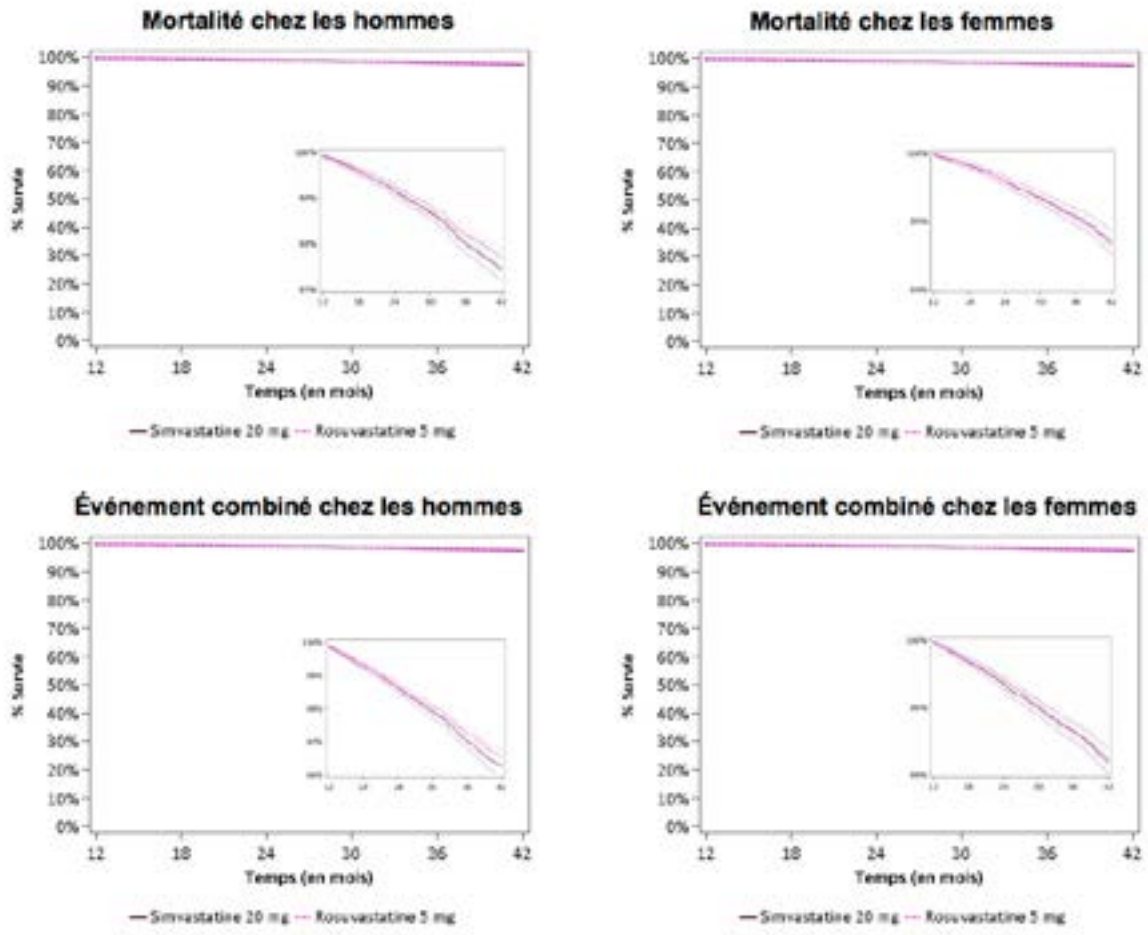
Connaître l'efficacité des traitements en vie réelle : l'exemple des statines[2]

Les statines sont des molécules utilisées pour faire baisser les taux de certains cholestérols. A l'inverse des autres pays européens où les prescriptions de statines se concentrent sur la simvastatine, en France, la rosuvastatine non génériquée occupe une place importante dans les prescriptions.

Dans ce contexte, la CNAMTS a souhaité, à partir des données du SNIIRAM, mesurer en population réelle l'efficacité de la rosuvastatine, comparativement à la simvastatine, pour prévenir, chez des patients indemnes de pathologies cardiaques, la survenue d'événements cardiovasculaires graves et de décès ; les dosages étudiés pour les deux composés (rosuvastatine 5 mg et simvastatine 20 mg), ayant montré une capacité comparable à réduire le LDL cholestérol[3], marqueur de risque démontré de pathologies cardio- vasculaires.

L'étude a porté sur une cohorte de plus de 160 000 patients dont les deux tiers débutaient un traitement par rosuvastatine 5 mg. Dans aucune des analyses réalisées (ventilées par sexe, événement étudié et type d'analyse), le hazard ratio ajusté[4] comparant rosuvastatine 5 mg et simvastatine 20 mg n'était significativement différent de un, ce qui indique une absence de différence dans l'efficacité. Les résultats de cette étude en vie réelle n'étaient donc pas en faveur d'une prescription préférentielle de la rosuvastatine par rapport à la simvastatine en prévention primaire.

2. Référence : Neumann A, Maura G, Weill A, Ricordeau P, Alla F, Allemand H. Comparative effectiveness of rosuvastatin versus simvastatin in primary prevention among new users: a cohort study in the French national health insurance database. Pharmacoepidemiol Drug Saf 2013. Epub 2013 Dec 2.
3. LDL : « low-density lipoprotein »
4. Le « hazard », ou risque, est la probabilité pour un sujet de développer la maladie à un âge donné, sachant qu'il ne l'a pas présentée jusqu'alors. Cette fonction dépend des facteurs de risque et des traitements éventuels pris. Comparer deux traitements revient à comparer le rapport ou « hazard ratio » de ces risques entre sujets comparables (d'où le terme ajusté) soumis à l'un ou l'autre traitement. C'est un indicateur couramment utilisé en épidémiologie clinique.



Graphique 2. Efficacités comparées de deux médicaments anti-cholestérol : la rosuvastatine 5 mg et la simvastatine 20mg.(instaurations en prévention primaire par un généraliste en 2008-2009 chez les 40-79 ans ; prise régulière pendant la première année). Les graphiques du haut représentent les taux de survie de 12 à 42 mois ; les graphiques du bas représentent les taux de survie sans hospitalisation pour cardiopathie ischémique aigüe ou accident vasculaire cérébral aigu. A gauche : graphiques concernant les hommes ; à droite : graphiques concernant les femmes.

Surveiller la sécurité du médicament : l'exemple de la contraception orale combinée[5]

En décembre 2012, une femme de 25 ans exposée à un contraceptif oral combiné (COC) de troisième génération et victime d'accident vasculaire cérébral avec séquelles portait plainte contre la firme pharmaceutique et les autorités sanitaires. Les medias ont fortement relayé cette information.
Le débat qui a suivi a généré de nombreuses questions de la part des utilisatrices de contraceptifs oraux et des réactions de professionnels de la santé. Le 11 janvier 2013, la ministre des affaires sociales et de la santé annonçait plusieurs mesures visant à limiter, en France, la prescription

5. Référence : « Risque d'embolie pulmonaire, d'accident vasculaire cérébral ischémique et d'infarctus du myocarde chez les femmes sous contraceptif oral combiné en France : une étude de cohorte sur 4 millions de femmes de 15 à 49 ans à partir des données du SNIIRAM et du PMSI ». Rapport final du 26 juin 2013. Accessible en ligne sur Ameli.fr

des pilules de troisième et quatrième génération et demandait qu'une étude rétrospective pharmaco-épidémiologique soit réalisée sur des données françaises pour évaluer la sécurité de ces molécules.

Les résultats de cette étude, conduite par la CNAMTS en partenariat avec l'ANSM[6], ont été rendus publics quelques mois plus tard, en juin de la même année. Ils étaient similaires à ceux des études observationnelles internationales les plus récentes et les plus puissantes. L'analyse de la cohorte de plus de 4 millions de femmes, résidant en France, et ayant eu des remboursements de COC confirmait ainsi l'existence d'un doublement du risque d'embolie pulmonaire des COC de troisième génération par rapport à ceux de deuxième génération. Le risque d'embolie pulmonaire entre les pilules de deuxième et de troisième génération passait de 25 à 50 pour 100 000 personnes-années.
Mais cette étude a permis également de quantifier les risques, moins connus, associés aux différents dosages d'éthinylestradiol. Au total, il s'avérait que les progestatifs d'ancienne génération comme le lévonorgestrel, combinés à 20 µg d'éthinylestradiol, étaient associés à un moindre risque thromboembolique veineux et artériel ; l'association de 100 µg de lévonogestrel et de 20 µg d'éthinylestradiol étant commercialisée et remboursée en France depuis avril 2010.

L'accès au SNIIRAM

Le SNIIRAM a été d'emblée conçu comme un système ouvert vers des utilisateurs externes : c'est ce qui a guidé les choix d'architecture (portail unique pour tous, utilisation de certains logiciels du marché facilement manipulables par des non experts, etc.). C'est aussi la raison pour laquelle l'accord de la CNIL a été sollicité non seulement sur la constitution de cette base de données, mais aussi dès le début sur son utilisation par des tiers.
De nombreux organismes ont des accès pérennes : Ministère, agences ou autorités sanitaires publiques comme l'ANSM, l'InVS[7], la HAS[8] et les ARS[9], instituts de recherche comme l'INSERM[10] ou l'IRDES[11], membres de l'Institut des données de santé comme les unions professionnelles des professions de santé, l'union nationale des organismes d'assurance maladie complémentaire…
Ces organismes accèdent aux données à travers une procédure d'habilitation, le périmètre des données accessibles par chaque organisme étant déterminé en fonction de ses missions. Il peut s'agir des bases de données agrégées, de l'échantillon de bénéficiaires au 1/100ème, qui comporte une profondeur historique d'une dizaine d'années, jusqu'à la base de données comportant les données exhaustives détaillées.
Tout autre organisme à but non lucratif de recherche, université, école ou autre structure d'enseignement lié à la recherche a la possibilité, pour effectuer une étude en santé publique, d'accéder pour cette étude aux bases de données agrégées ou à l'EGB[12] après approbation de l'Institut des données de santé (IDS) et autorisation de la Commission nationale de l'informatique et des libertés (CNIL). Au-delà de ces produits préformatés, l'ensemble des organismes de recherche ou d'études à but non lucratif peuvent également solliciter une extraction ad hoc des données exhaustives, qui est réalisée à la demande par la CNAMTS après accord des instances ayant compétence à autoriser ces traitements (CNIS[13] , CCTIRS[14] , IDS, CNIL) et sous certaines conditions juridiques et techniques.

6. Agence nationale de sécurité du médicament
7. Institut national de veille sanitaire
8. Haute autorité de santé
9. Agences régionales de santé
10. Institut national de la santé et de la recherche médicale
11. Institut de recherche et documentation en économie de la santé
12. Echantillon généraliste des bénéficiaires (au 1/100°)
13. Conseil national de l'information statistique
14. Comité consultatif sur le traitement de l'information en matière de recherche dans le domaine de la santé

Il n'y a en revanche pas d'accès prévus par les textes pour les organismes à but lucratif.
Ouvertes à beaucoup d'acteurs du système de santé et de représentants de la société civile, les données du SNIIRAM sont parfois difficiles à exploiter pour les utilisateurs potentiels de par leur technicité et leur complexité. En effet, ces informations ne sont initialement pas recueillies dans un objectif d'études mais dans le but de verser des prestations aux assurés.

Leur architecture et leur contenu intègrent des contraintes de production et évoluent avec la législation. C'est pourquoi la CNAMTS propose des formations aux organismes habilités à accéder au SNIIRAM et met à disposition des utilisateurs des outils d'aide au traitement des données (dictionnaire des concepts et des données, forum d'échanges, centre de support, journées d'information, clubs utilisateurs, requêtes prédéfinies). A titre d'exemple, la CNAMTS a formé environ 160 chercheurs à l'échantillon généraliste des bénéficiaires (EGB) depuis son ouverture en 2007. En 2013, une cinquantaine de chercheurs a interrogé de manière régulière cette base de données. Dans le cas des extractions ad hoc, une demi-journée d'information est organisée par la CNAMTS pour accompagner les équipes de recherche dans l'utilisation des données du SNIIRAM à des fins d'études et de recherche.
L'accès aux données de santé, et notamment aux données du SNIIRAM, est actuellement en cours d'examen dans le cadre de la commission « Open data » mise en place par la Ministre de la santé. Les modalités d'accès pourraient être modifiées suite aux propositions de cette commission, avec l'objectif de faciliter et simplifier l'accès aux données et de développer encore leur utilisation, en particulier pour la mise en œuvre des politiques de santé.

En conclusion

Le SNIIRAM (système national d'information inter-régimes de l'Assurance maladie) est une base de données médico-administratives qui ouvre des possibilités remarquables en santé publique. D'année en année, cette base de données s'est enrichie et améliorée. Pour donner un exemple, ce n'est qu'en 2010 que le chaînage avec le PMSI a pu être opéré en routine, permettant l'analyse des parcours complets de soins des patients. C'est à partir de cette même année que les dates précises d'hospitalisation (au lieu du mois) ont également été disponibles, ce qui est évidemment essentiel pour analyser l'impact de certains traitements. Ce n'est donc que depuis quelques années que cet outil offre un potentiel d'analyse réellement très riche, qui va encore s'améliorer dans l'avenir.
Le SNIIRAM est de plus en plus utilisé, notamment par les chercheurs, et la CNAMTS a accompagné cette dynamique d'ouverture (formations, club utilisateurs, outils d'aide au traitement, soutien financier à une plate-forme pour la recherche,...). Un débat a été lancé récemment sur l'intérêt et les enjeux d'une ouverture plus large des données, et les modalités d'accès pourraient être amenées à évoluer.

L'accès aux données confidentielles de la statistique publique

De la sensibilité des données économiques à la sensibilité des données de santé

Jean-Pierre LE GLÉAU
Inspecteur général honoraire de l'Insee

En France comme dans l'ensemble de l'Europe, les données de la statistique publique font l'objet de dispositions légales particulières, qui viennent s'ajouter aux protections prévues par le droit pour toutes les catégories de données. Le « secret statistique » apparaît ainsi comme une garantie donnée aux répondants en contrepartie de la sincérité des informations qu'ils fournissent à l'administration. La loi prévoit des exceptions à ce secret pour la recherche scientifique : afin de répondre à la demande croissante d'accès à des données confidentielles de la part des chercheurs, des dispositifs spécifiques ont été mis en place ces dernières années. Ces modalités déjà éprouvées peuvent, peut- être, servir de modèle pour faciliter l'accès des chercheurs à d'autres types de données sensibles, telles que les données de santé.

Pourquoi le secret statistique ?

Pour connaître l'économie et la démographie d'un pays, il est nécessaire d'interroger de temps à autre ses habitants, ses entreprises, ses exploitations agricoles, ses administrations. Toutes les informations ainsi recueillies permettent, combinées à d'autres, de dresser un tableau aussi fidèle que possible de la réalité du pays.

Si l'on veut que l'information recueillie soit utile pour dresser un panorama conforme à la réalité, il faut que celui qui est interrogé (particulier ou entreprise) fournisse une réponse sincère. Pour cela, il est indispensable qu'il soit assuré que ses réponses ne seront pas utilisées dans un sens qui puisse lui causer du tort. On pense immédiatement aux impôts, à la police, mais aussi aux concurrents, aux voisins, à la famille, etc. Les réponses aux questions doivent donc rester absolument confidentielles afin de pouvoir donner au répondant l'assurance que sa réponse ne lui portera pas tort.

C'est justement l'objet du secret statistique : protéger les informations recueillies au moyen d'enquêtes statistiques, afin d'obtenir de la part de la personne interrogée des réponses sincères, tout en lui garantissant que ces réponses ne pourront en aucun cas lui porter préjudice.

Comme toutes les informations individuelles détenues par l'administration, celles qui sont recueillies au moyen d'enquêtes statistiques sont protégées par des règles générales de confidentialité qui s'appliquent pour tous les agents ayant à en connaître. Cependant, le secret statistique va plus loin, car il confine les données ainsi recueillies dans une "bulle" très étanche, interdisant par exemple leur communication d'une administration à une autre.

Les dispositions légales

Le secret statistique est présent dans le traité de l'Union Européenne, qui précise dans son article 338 que « L'établissement des statistiques se fait dans le respect (…) de la confidentialité des informations statistiques ».
Il est développé dans le règlement de mars 2009 relatif aux statistiques européennes, qui lui consacre un chapitre (articles 20 à 26) dans lequel sont énumérées les mesures s'appliquant « pour garantir que les données confidentielles sont utilisées exclusivement à des fins statistiques et pour empêcher leur divulgation illicite »
Par ailleurs, la loi statistique française du 7 juin 1951, maintes fois modifiée, définit dans ses articles 6 et 7bis la protection qui doit être apportée aux informations couvertes par le secret statistique.
Mais quelles sont au juste ces informations couvertes par le secret statistique ?
Les informations couvertes par le règlement de l'Union européenne sont celles qui sont utilisées par cette dernière pour produire les résultats demandés par les instances européennes.
Celles qui sont protégées par la loi française sont de deux types :

- d'une part, les informations collectées au moyen d'enquêtes statistiques, ayant reçu un numéro de visa attribué par le ministre chargé de l'économie ; leur liste est publiée chaque année au Journal officiel ;
- d'autre part, les informations recueillies par l'Insee ou des services statistiques ministériels (SSM) auprès d'autres administrations, en vue d'établissement de statistiques.

Pour la France, il y a un très fort recouvrement entre ces deux ensembles.
Le règlement européen et la loi française fixent un principe général, des modalités de mise en œuvre, ainsi que quelques exceptions :

- le principe est celui d'un confinement des informations confidentielles auprès de celui qui les a collectées et traitées pour produire des statistiques anonymes : il ne doit les communiquer à personne, pas même à d'autres administrations ;
- en France, ces informations restent secrètes pendant un temps assez long (25 ans pour les informations d'ordre économique et financier, 75 ans pour celles qui ont trait aux faits et comportements d'ordre privé) et de lourdes sanctions s'appliquent à qui trahirait ce secret (un an de prison et 15 000 € d'amende) ;
- des exceptions sont prévues dans la loi française. Certaines sont d'ordre institutionnel, d'autres visent à favoriser l'emploi des données collectées à des fins de recherche scientifique. Les exceptions institutionnelles visent par exemple à obliger quiconque aurait connaissance d'un crime ou d'un délit grâce à ces données confidentielles, à en informer le procureur ; elles permettent aussi à ce dernier d'avoir connaissance d'informations couvertes par le secret statistique, dans le cadre d'une commission rogatoire. Ces dispositions sont une entorse au secret statistique, peu conformes aux principes européens. Par ailleurs, une disposition permet aux chercheurs d'avoir accès à la masse des informations collectées en vue de l'établissement des statistiques. Ce dispositif est prévu aussi bien dans le cadre européen que dans celui de la loi française.

L'accès des chercheurs aux informations confidentielles

Au moment où la loi statistique française a été adoptée (1951), la question de l'accès des chercheurs aux données individuelles ne se posait guère. Tous les questionnaires étaient sur support papier, et il était hors de question de permettre un accès général, fût-ce à des fins de recherche, à ces documents. Seules des dérogations ponctuelles, très rares, étaient envisageables.
Avec le développement de l'informatique, tant du côté de la recherche que du côté des services producteurs de statistiques, le contexte a changé.

Il est devenu possible de mettre à la disposition des chercheurs des fichiers de données individuelles rendues anonymes. En supprimant ou agrégeant un certain nombre de variables, on pouvait constituer des fichiers présentant un intérêt pour la recherche, mais préservant l'anonymat de ceux qui avaient répondu.

C'est ainsi qu'ont été créés et diffusés des fichiers détail permettant aux chercheurs de travailler sur des données individuelles, sans porter atteinte à la vie privée des personnes qui avaient répondu aux enquêtes.

Malheureusement, ces fichiers individuels ne pouvaient être produits pour les entreprises. En effet, dès lors que l'on donne pour une entreprise son activité économique, sa taille, voire un indice sur sa localisation, il devient souvent très facile de deviner de quelle entreprise il s'agit. On a donc été obligé de constater que la plupart des fichiers individuels d'entreprises ne permettaient pas de sauvegarder le secret statistique. Quel serait, pour un chercheur, l'intérêt d'un fichier individuel d'entreprises, dans lequel on aurait fait disparaître les variables « activité économique », « taille » et « localisation » ?

En 1984, la loi sur le secret a donc été modifiée, afin de permettre, sous condition, l'accès des chercheurs aux informations individuelles sur les entreprises. Cela revient à élargir aux chercheurs la « bulle » dans laquelle étaient jusqu'alors confinées ces informations.
La loi prévoit que cet élargissement peut se faire après avis d'un comité créé à cet effet et appelé « Comité du secret statistique concernant les entreprises ». Il était présidé par un membre du Conseil d'État et comprenait quatre représentants de l'administration (dont un représentant du ministre de la Justice), quatre représentants des entreprises, un représentant des organisations syndicales de salariés et un représentant des utilisateurs régionaux et locaux de la statistique publique.

Après avis de ce comité, les chercheurs recevaient les informations demandées sur support magnétique. Ils avaient auparavant signé un engagement de confidentialité et s'étaient engagés à détruire les données à un certain terme fixé par le comité du secret statistique.

Pour les fichiers de données sur les ménages, la mise à disposition de fichiers individuels a donné satisfaction pendant quelque temps. Mais il est vite apparu deux inconvénients :

- d'une part, avec le développement d'internet, ces fichiers étaient accessibles sur le site de l'Insee et donc n'importe qui pouvait les télécharger, de façon anonyme. Y compris des personnes, éventuellement animées d'intentions malveillantes, et connaissant par exemple certains répondants à l'enquête. Au moyen des informations déjà connues sur ceux-ci, il leur était dans certains cas possible d'identifier l'enregistrement correspondant et donc de prendre connaissance des réponses effectuées par ces personnes. Ces cas étaient très rares, supposaient de la part de l'internaute une démarche volontaire et complexe ainsi qu'une véritable envie de violer la loi, mais ils représentaient un danger trop grand pour la préservation du secret statistique ;
- d'autre part, pour assurer une protection élémentaire du secret, ces fichiers ne comportaient pas tout le détail qui aurait été nécessaire aux chercheurs : profession codée sur deux chiffres et non sur quatre, localisation à la région seulement, etc.

Pour répondre à ces inconvénients, une première démarche a consisté à mettre à disposition des chercheurs, et d'eux seulement, des fichiers un peu plus détaillés, mais préservant encore l'anonymat, pour qui ne tenterait pas systématiquement d'identifier des individus. C'est ce que l'on appelle les « Fichiers de production et de recherche » (FPR), mis à la disposition des chercheurs, via le réseau Quetelet. Le réseau Quetelet s'assure que le demandeur est bien un chercheur et lui donne accès à des fichiers « raisonnablement anonymes », c'est-à-dire où il n'est pas possible d'identifier qui que ce soit, tant que l'on utilise ces fichiers à des fins de

recherche scientifique. Un chercheur qui romprait cet engagement pourrait tenter d'identifier une ou deux personnes dont il saurait qu'elle a participé à l'enquête. Le plus souvent il n'y arriverait pas. Mais, exceptionnellement, il pourrait y parvenir. Le chercheur serait alors en infraction avec la loi et, compte tenu de la traçabilité du réseau Quetelet, il courrait un risque sérieux d'être démasqué.

Mais, avec le développement des nouveaux moyens (logiciels et matériels) de traitement de données, les chercheurs ont eu besoin d'une information encore plus détaillée pour leurs travaux.

C'est pourquoi la loi a été à nouveau modifiée en 2008 pour permettre l'accès aux données les plus détaillées sur les ménages, à des fins de recherche scientifique ou historique, ou de statistique publique. Le comité du secret statistique doit donner son avis pour une telle communication. Dans ce but, il a été renommé et réorganisé, pour permettre l'entrée d'organisations concernées par la transmission de données à caractère personnel. En effet, ces données permettant parfois l'identification des personnes, il est nécessaire, pour y avoir accès, d'accomplir également des formalités auprès de la Cnil, variables selon la nature des informations concernées.

Comment accéder aux données ?

La procédure formelle d'autorisation inclut l'avis du comité du secret statistique, l'accord de l'autorité dont émanent les documents (le plus souvent l'Insee ou un service statistique ministériel) puis une décision de l'administration des archives, puisque les enquêtes statistiques sont considérées comme des archives publiques.

Munis de ces autorisations, les chercheurs peuvent donc avoir accès à l'ensemble des données recueillies ou traitées par la statistique publique. C'est un progrès majeur. Mais cette facilité impose en contrepartie une protection supplémentaire de ces données.

En particulier, il était apparu depuis quelque temps que la procédure qui permettait à des chercheurs d'emporter dans leur labo un CD comportant des données confidentielles ne présentait pas les garanties suffisantes en termes de protection du secret statistique.

C'est pourquoi, l'Insee a développé, grâce au Groupement des écoles nationales d'économie et de statistique (Genes), un centre d'accès sécurisé aux données, le CASD.

Placé sous la responsabilité du Genes, le CASD permet aux chercheurs qui ont été habilités par le comité du secret statistique d'accéder aux données les plus confidentielles. Ils ont signé un engagement de confidentialité et la loi leur interdit toute communication externe, sous peine des sanctions mentionnées précédemment.

Le CASD leur remet alors une boîte appelée « SD Box », qui est un terminal grâce auquel les chercheurs peuvent accéder au serveur, localisé au Genes, sur lequel se trouvent les données confidentielles. Le système leur permet de « voir » les données, de travailler dessus, mais en aucun cas de les imprimer, ou de les recopier sur un autre support (clé USB, disque dur, etc.). Lorsque les chercheurs ont obtenus les résultats qu'ils souhaitaient, ils les placent dans une boîte à lettres virtuelle. Des experts du CASD vérifient que ces résultats ne contreviennent pas aux règles du secret statistique. Si tel est le cas, le fichier de résultat est renvoyé au chercheur par simple messagerie.

Ce système permet une bonne protection du secret statistique, en assurant la traçabilité de toutes les opérations réalisées par les chercheurs. Mis en place depuis 2010, il semble aussi donner satisfaction aux chercheurs qui peuvent enfin avoir accès à l'information la plus détaillée.

Un exemple pour l'accès à d'autres types de données confidentielles ?

Les données statistiques sont cependant loin d'être les seules données publiques couvertes par une obligation de confidentialité.

On peut en citer bien d'autres. Pour certaines d'entre elles, l'exemple de l'accès aux données de la statistique publique peut être riche d'enseignements.

Ainsi, la loi du 22 juillet 2013 permet, sous certaines conditions, l'accès des chercheurs aux données fiscales. Celui-ci devrait se faire, après avis favorable du comité du secret statistique, par l'intermédiaire d'un centre d'accès sécurisé, qui permet de préserver la confidentialité et d'empêcher la dissémination des informations par inadvertance ou malveillance. Le centre d'accès sécurisé aux données de la statistique publique semble avoir toutes les qualités requises pour servir d'instrument de mise à disposition des données fiscales. On aurait ainsi, pour les données fiscales, un cheminement analogue à celui qui existe pour les données de la statistique publique : avis du comité du secret statistique, puis accès par un centre d'accès sécurisé. La seule différence serait que l'accord de l'autorité dont émanent les documents et la décision des archives serait remplacés par une décision du ministre chargé du budget.

On peut imaginer un processus analogue pour l'accès à des données d'ordre médical. Les traitements de données à caractère personnel ayant pour fin la recherche dans le domaine de la santé font l'objet d'un chapitre particulier (Chapitre IX, articles 53 à 61) de la loi du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés. Le législateur a considéré que les données relatives à la santé devaient bénéficier d'une protection particulière : elles sont en effet classées parmi les données sensibles dans cette même loi, au même titre que celles qui font apparaître les origines raciales ou ethniques, les opinions politiques, philosophiques ou religieuses ou l'appartenance syndicale des personnes, ou qui sont relatives à la vie sexuelle de celles-ci.

Il convient donc de distinguer de façon précise les données médicales selon qu'elles permettent ou non d'identifier, directement ou indirectement, les individus auxquels elles se rapportent. Ce travail aboutit généralement à la classification des données médicales en trois catégories :

- celles qui sont totalement anonymes et ne permettent aucune identification ;
- celles qui permettent, de façon directe ou, le plus souvent, indirecte l'identification de certains individus ;
- celles qui se trouvent dans une zone intermédiaire : elles ne permettent en général pas une identification des personnes concernées, mais, certains individus possédant des informations spécifiques pourraient, en se donnant du mal, avoir une probabilité non négligeable d'identifier un petit nombre d'individus.

Ces trois catégories doivent être traitées de façon spécifique.

La première contient un grand nombre de tableaux statistiques, où on s'est assuré qu'il y avait un nombre suffisant d'individus dans chaque case. Elle contient aussi des fichiers de données individuelles, où les variables liées à chaque personne ont été suffisamment agrégées ou floutées pour rendre impossible toute identification. Ces tableaux et ces fichiers peuvent être mis sans inconvénient à disposition du public le plus large (internet, open data,...)

La troisième catégorie constitue ce que l'on appelle parfois la « zone grise ». L'identification d'individus de la base n'est pas strictement impossible, mais elle nécessiterait la disposition d'informations complémentaires faiblement répandues dans le public, le déploiement de moyens importants pour tenter de parvenir à une identification, avec comme résultat une simple probabilité, sans certitude, d'avoir identifié quelques individus (en général peu nombreux). La mise à disposition de tels fichiers doit être rendue possible, pour des personnes dont on aurait vérifié le sérieux et la moralité, après qu'elles auraient détaillé le projet pour lequel elles ont besoin d'avoir accès à ces données. Ces personnes signeraient un engagement de n'utiliser ce fichier que dans le cadre dudit projet et de ne tenter en aucun cas d'identifier un individu précis de la base. Tout manquement à cet engagement a des chances non négligeables d'être repéré et de lourdes sanctions peuvent s'appliquer dans ce cas. Cette catégorie s'apparente à celle qui a été décrite pour l'accès aux données de la statistique publique, via le réseau Quetelet.

La deuxième catégorie nécessite un contrôle beaucoup plus serré, car elle est constituée de fichiers permettant, à un faible coût et sans disposer d'une information rare, d'identifier un grand nombre d'individus. L'accès à ces données pourrait se faire par l'intermédiaire d'un centre d'accès sécurisé, avec toutes les garanties déjà prévues pour les données issues de la statistique publique et, bientôt, pour les données fiscales :

- présentation d'un projet pour lequel il est démontré qu'il est nécessaire d'avoir accès à des informations très détaillées et donc potentiellement identifiantes ;
- garanties sur la personne qui présente la demande : environnement institutionnel, caution d'une personne d'un rang hiérarchique suffisant, accès dans un environnement matériel correctement sécurisé ;
- avis d'une autorité (analogue au comité du secret statistique) sur le respect de ces critères et formalités adaptées auprès de la Cnil ;
- identification forte (biométrique) de la personne ayant accès aux données ;
- consultation et travail rendus possibles sur des données ; mais celles-ci restent sur un serveur et ne peuvent être ni copiées, ni transcrites, ni imprimées sur aucun support (papier, CD, clef USB…) ;
- vérification par des experts que les résultats issus de ces travaux sont strictement anonymes.

Un organisme rendant ce genre de service pourrait être du même type que le centre d'accès sécurisé aux données de la statistique publique (CASD).

Des sanctions pénales restent prévues dans ce cas, mais l'idée est plutôt de rendre quasiment impossible l'accès aux données personnelles et de repérer facilement ceux qui auraient tenté d'enfreindre les interdictions fixées par la loi.

Toutes ces conditions permettent des modalités d'accès adaptées au degré de confidentialité des données, qui préservent la vie personnelle des individus tout en permettant un travail de recherche dans de bonnes conditions.

Technologies de l'information et de la communication et données de santé :
pour un cadre juridique en phase avec les évolutions technologiques et les besoins du système de santé

Jeanne Bossi
Secrétaire générale de l'Agence des systèmes d'information partagés de santé[1]

Les données personnelles de santé sont confidentielles : pour protéger la vie privée des citoyens, la loi édicte des limitations strictes quant à leur gestion et à leur transmission. Cependant, les besoins d'échange et de traitement de ces données se font de plus en plus sentir, que ce soit pour le traitement des malades ou pour des études de santé publique. Des progrès techniques ont été réalisés pour faciliter ces échanges : création de référentiels et d'identifiants communs des acteurs. Pour aller plus loin, il faudra préciser le cadre juridique et définir une nouvelle gouvernance des données de santé.

Le développement des nouvelles technologies de l'information et de la communication dans les domaines sanitaire et médico-social peut constituer l'une des réponses aux problématiques que traverse actuellement notre système de santé : égalité d'accès aux soins dans un contexte économique contraint, conséquences du vieillissement de la population et de la dépendance, coordination du suivi médical tout au long du parcours de soins étendu au domaine médico-social et multidisciplinarité croissante de l'exercice médical.
Ces facteurs accroissent en effet le besoin d'échange de données de santé dans l'intérêt d'une meilleure prise en charge des personnes. Toutefois, les données personnelles de santé qui permettent d'identifier un individu sont également susceptibles de révéler l'intimité de la vie privée. A ce titre, le droit leur reconnaît un statut particulier et impose le respect de règles ayant pour objectif de garantir leur confidentialité.
Comment alors permettre le développement des échanges de données de santé dématérialisés, nécessaires à l'amélioration du système de soins, sans toutefois renier les principes fondamentaux de la protection de la vie privée ? Au-delà de l'impulsion d'une politique publique volontariste, comment encadrer le développement spontané et très rapide des technologies dans le secteur des systèmes d'information de santé ?

Nous aborderons cette problématique en commençant par dresser un état des lieux du cadre légal qui régit actuellement l'échange et le partage de données de santé. Nous verrons ensuite comment ce cadre juridique est désormais complété par des référentiels permettant d'assurer la confidentialité des données. A la lumière de cet état des lieux, nous pourrons enfin déterminer les limites du cadre juridique actuel et proposer des pistes de réflexions pour faire face aux nouveaux enjeux soulevés par les évolutions simultanées des secteurs sanitaire et numérique.

1. L'ASIP-Santé est un groupement d'intérêt public fondé en 2009 pour renforcer la maîtrise d'ouvrage publique des systèmes d'information qui se développent dans le secteur de la santé et accompagner l'émergence de technologies numériques en santé.

Un cadre juridique qui doit s'adapter au partage des données de santé

Les données de santé font l'objet en France d'un encadrement juridique qui vise à protéger leur confidentialité.
Le cadre actuel de l'échange et du partage des données de santé s'articule autour de différents textes de lois qui traitent de leurs conditions d'utilisation et des moyens assurant leur confidentialité.
La gestion et le traitement des données de santé sont protégés par la directive européenne 95/46 du 24 octobre 1995 et la loi Informatique et Libertés du 6 janvier 1978 modifiée relative à l'informatique, aux fichiers et aux libertés.
L'information préalable de la personne sur l'informatisation de ses données et, en particulier, l'information sur ses droits représentent toujours une garantie importante. Dans certains cas, le recueil du consentement peut être une protection supplémentaire de la personne.

a. Les principes de la protection des données personnelles

Ces principes, définis dans la loi « Informatique et Libertés », se concentrent autour de cinq notions clés : une finalité de traitement déterminée et légitime, des données pertinentes (principe de proportionnalité), une durée de conservation déterminée à l'avance et dont la réalité est appréciée au regard de cette finalité (droit à l'oubli), le respect du droit des personnes et de leur information, et enfin la mise en place de mesures de sécurité de nature à garantir la confidentialité des données.
C'est le rôle de l'autorité administrative indépendante qu'est la Commission Nationale de l'Informatique et des Libertés de faire respecter ces principes.
Si l'article 8-1 de la loi Informatique et Libertés pose le principe de l'interdiction de la collecte et du traitement des données de santé à caractère personnel, le deuxième titre du même article procède à l'énumération limitative des cas dans lesquels leur traitement est admis. Les conditions posées diffèrent toutefois selon la finalité poursuivie : recherche médicale, intérêt public, médecine préventive, évaluation des pratiques… etc.
Parmi ces exceptions et cas particuliers, on retrouve les traitements mis en œuvre à des fins de coordination des soins et qui nécessitent le partage de données de santé, dont la CNIL a considéré qu'ils relevaient de l'intérêt public. Il s'agit là par exemple du régime juridique retenu pour le Dossier Médical Personnel (DMP) lancé en 2011 et déployé par l'Agence des systèmes d'information partagés de santé, ASIP Santé.
Le traitement des données personnelles est également admis dans le cas d'une expression du consentement exprès de la personne, sauf si la loi prévoit que l'interdiction ne peut être levée par ce consentement.
Dans certains cas en effet, le consentement de la personne est sans effet parce que la loi elle-même interdit la collecte et le traitement des données de santé. C'est le cas de l'interdiction pour le médecin d'une compagnie d'assurance d'accéder à un dossier médical ou à un employeur d'exiger d'un futur candidat des examens médicaux ou l'accès à son dossier médical.
Enfin, au-delà de l'affirmation de la spécificité des données de santé, la CNIL a toujours recommandé la mise en œuvre de mesures pratiques visant à informer le patient de ses droits et des modalités d'utilisation et de conservation de ses données, conformément aux articles 32, 39 et 40 de la loi informatique et libertés, et articles L.1111-2 et L. 1111-7 du code de la santé publique.
Ainsi, s'est-elle toujours attachée particulièrement aux mesures de sécurité qui doivent être mises en œuvre pour garantir aux données médicales la confidentialité exigée par la loi.[2]

2. Délibération n°01-011 du 8 mars 2001.

C'est ainsi que les traitements de données de santé à caractère personnel que nécessite l'hébergement de ces données doivent être réalisés dans le respect des dispositions de la loi du 6 janvier 1978 et de l'article L1111-8 du code de la santé publique introduit par la loi du 4 mars 2002 sur les droits des malades et la qualité du système de soins, qui disposent que les professionnels de santé ou les établissements de santé ou la personne concernée peuvent déposer des données de santé à caractère personnel, recueillies ou produites à l'occasion des activités de prévention, de diagnostic ou de soins, auprès de personnes physiques ou morales agréées à cet effet.
L'hébergement des données exige le consentement exprès de la personne concernée, doit respecter les dispositions de la loi Informatique et Libertés et le secret professionnel dans les conditions et sous les peines prévues à l'article 226-13 du code pénal.
L'agrément est délivré pour trois ans par le ministre en charge de la santé après avis de la CNIL et du comité d'agrément des hébergeurs (CAH), dont l'ASIP Santé assure le secrétariat, après une évaluation des capacités des candidats, portant sur les aspects financiers, éthiques et de sécurité de leur activité. Il porte sur une prestation particulière, même si une mutualisation des services proposés est possible.
La liste exhaustive des hébergeurs agréés est disponible sur le site de l'ASIP Santé.

Tableau 1 : Bilan chiffré du Comité d'agrément des hébergeurs à mars 2014

- 188 dossiers ont été réceptionnés depuis le 1er juin 2009
- dont 16 dossiers de renouvellement d'agrément
- 82 dossiers ont été agréés
- 57 refus d'agrément ont été prononcés.

b. L'encadrement de l'échange et du partage des données de santé

Le droit commun de l'échange de données de santé à caractère personnel entre professionnels de santé est fixé à l'article L1110-4 du code de la santé publique. Cet article définit actuellement trois régimes d'échange et de partage des données de santé à caractère personnel :

- l'échange de données de santé entre plusieurs professionnels de santé qui prennent en charge un même patient en dehors d'un établissement de santé, sauf opposition ;
- le partage de données de santé entre professionnels de santé exerçant au sein d'un même établissement de santé, sauf opposition ;
- le partage de données de santé au sein d'une maison ou d'un centre de santé soumis au consentement.

En outre, le législateur a prévu le recueil du consentement de la personne concernée, sous différentes formes, dans le cadre des services nationaux de partage de données de santé que sont le dossier pharmaceutique (DP), le dossier médical personnel (DMP) et l'historique de remboursement.
Cette multiplicité de régimes juridiques est source d'incompréhension tant pour les usagers que pour les professionnels et ne garantit pas une protection efficace des droits des personnes concernées et de leurs données. Ils ne répondent pas à une logique de situation mais résultent d'une succession de textes intervenus au fil du temps et qui, souvent par simplicité, ont exigé de façon systématique un consentement.

c. La nécessaire prise en compte du parcours de soins

On l'a vu, le législateur a étendu la notion d'équipe de soins aux maisons et centres de santé : les professionnels peuvent partager les informations concernant les personnes qu'ils prennent en charge, sans toutefois que le législateur ait imposé la condition de l'appartenance à la même équipe de soins.
Or, on peut regretter que cet élargissement se limite à ces seules structures d'exercice regroupé, dans la mesure où la distinction qui a pu être faite entre une donnée de santé et une donnée médicosociale[3] trouve aujourd'hui ses limites dans la nécessité d'une prise en charge globale de la personne, qu'elle fasse appel au secteur sanitaire ou médicosocial.
Les textes de loi intervenus récemment insistent pourtant sur la nécessité d'une coordination des acteurs, en particulier à l'aide de systèmes d'informations. Ils consacrent également des modèles d'exercice collectif de la prise en charge au sein de structures de groupe. On peut citer la loi "Hôpital, patients, santé, territoires" (HPST) de juillet 2009 qui consacre l'exercice collectif au sein de maisons et centres de santé ou la création des maisons pour l'autonomie et l'intégration des malades d'Alzheimer, ou plus récemment l'article 48 de la loi du 17 décembre 2012 de financement de la sécurité sociale pour 2013, qui prévoit des expérimentations de collaboration entre structures médico-sociales et sanitaires dans le cadre de l'optimisation du parcours de santé des personnes âgées en risque de perte d'autonomie (PAERPA).
Plus globalement, aucun système d'information de santé aujourd'hui ne devrait se développer sans prendre en compte cette dimension dès sa conception.
Il faut donc tendre aujourd'hui vers une homogénéité des règles applicables (information préalable, droit d'opposition, consentement exprès) au partage des données de santé autour de la notion de parcours de soins, en élargissant la notion actuelle d'exercice en équipe à l'ensemble des professionnels de santé impliqués dans la prise en charge du patient.
La personne prise en charge doit pouvoir bénéficier d'un suivi utile, documenté et rendu accessible à l'ensemble de la communauté des professionnels qui seront appelés à la prendre en charge.
Si tous ces textes posent des principes forts pour assurer la protection des données personnelles de santé, la difficulté de leur application concrète et en particulier l'inadaptation des moyens prévus pour assurer la sécurité et la confidentialité des données face au développement de systèmes d'information organisés de façon non concertée ont rendu indispensables la normalisation et la sécurisation du partage de l'information.
C'est précisément cette idée qui a mené à la création de l'ASIP Santé, en 2009 : il s'agit d'instaurer un pilotage stratégique et cohérent des systèmes d'information de santé au niveau national, et de définir le cadre fonctionnel et de sécurité pour l'échange et le partage des données de santé.

Des référentiels désormais accessibles

Ce cadre fonctionnel est constitué par un ensemble de dispositifs techniques, normes et référentiels dans les champs de l'interopérabilité (capacité des systèmes à échanger des informations) et de la sécurité.

3. Lire aussi sur le sujet : BOSSI, Jeanne « Le cadre juridique du partage d'informations dans les domaines sanitaire et médicosocial. État des lieux et perspectives » in Médecine et Droit Vol 2013 – Janvier 2013 – Elsevier Masson – P5 – 8
http://www.em-consulte.com/article/785892/article/le-cadre-juridique-du-partage-dinformations-dans-l

Le cadre national d'interopérabilité des systèmes d'information de santé (CI-SIS)

Mis en place par l'ASIP Santé dès sa création et à la suite d'une concertation avec l'ensemble des industriels, ce référentiel spécifie les standards (le plus souvent internationaux) à utiliser dans les échanges et lors du partage de données de santé entre SIS[4], et contraint la mise en œuvre de ces standards par des spécifications d'implémentation.
L'interopérabilité est définie comme la capacité que possède un produit ou un système informatique à fonctionner avec d'autres produits ou systèmes existants ou futurs. C'est la possibilité qu'ont des systèmes à fonctionner en synergie, à « communiquer », ce qui implique d'utiliser un langage (interopérabilité sémantique) et des référentiels techniques (interopérabilité technique) communs.
Les spécifications et outils du CI-SIS sont modulaires et répartis en trois couches : une couche de contenus interopérables, qui concentrent les moyens de l'interopérabilité sémantique – c'est-à-dire, la structuration et la signification de l'information échangée entre les SI de santé – une couche de services d'interopérabilité et une couche de transport qui représentent quant à elles le socle d'interopérabilité technique du référentiel.
Ainsi, le programme « Santé connectée », lancé par la Haute Autorité de Santé et l'ASIP Santé en novembre 2013[5], vise à amener progressivement les professionnels de santé à saisir et utiliser des données standardisées en cours de consultation. Pour ce faire, le projet définit des règles de langage communes (nomenclatures) pour le poste de travail qui seront intégrées dans le CI-SIS.
La version actuelle du référentiel CI-SIS est la version 1.3 publiée le 18 octobre 2012. En outre, le cadre d'interopérabilité est en évolution constante au gré de l'implémentation de nouveaux volets et des commentaires que les utilisateurs peuvent remonter à l'ASIP Santé via son site web, esante.gouv.fr.

Les dispositifs d'identification des acteurs

Ces dispositifs sont majeurs et sans eux, aucun partage ou échange de données de santé à caractère personnel ne peut être réalisé de façon sécurisée.
S'agissant des professionnels de santé, leur identification repose sur la mise en place du Répertoire partagé des professionnels de santé (RPPS).
Créé par l'Arrêté du 6 février 2009, ce répertoire permet, à partir de données d'identification transmises par les autorités d'enregistrements (ordres professionnels, ARS[6]) de certifier les identités et d'attribuer un numéro RPPS à chaque professionnel de santé (Ordonnance n° 2009-1586 du 17 décembre 2009 relative aux conditions d'enregistrement des professions de santé).
Le RPPS contient à ce jour les médecins, pharmaciens, sages-femmes et chirurgiens-dentistes et doit se substituer à terme au répertoire ADELI[7], pour les professionnels de santé réglementés par le code de la santé publique. Les masseurs-kinésithérapeutes et les pédicures-podologues devraient y figurer prochainement.

Attestant de cette identité, le professionnel de santé peut alors devenir titulaire d'une Carte de Professionnel de Santé (CPS) ou d'un dispositif équivalent qui permet de l'identifier lors de l'échange et du partage de données tout en sécurisant leur transfert par des protocoles de sécurité spécifiques.

4. Systèmes d'information de santé
5. Cf « Programme Santé Connectée DataSet de Bonnes Pratiques DSBP », présentation du 21 novembre 2013 : http://fr.slideshare.net/esante_gouv_fr/20131121-jni-hasinterop
6. Agence Régionale de Santé
7. ADELI signifie Automatisation DEs LIstes. C'est un système d'information national sur les professionnels relevant du code de la santé publique.

La Carte CPS est une carte d'identité professionnelle électronique et une clé d'accès pour le professionnel de santé. Elle contient les données d'identification de son porteur (identité, profession, spécialité) mais aussi ses situations d'exercice (cabinet ou établissement).
L'usage de la carte de professionnel de santé (CPS) est en principe rendu obligatoire pour la conservation et la transmission par voie électronique d'informations médicales à caractère personnel. Cette obligation découle des dispositions de l'article L 1110·4 du code de la santé publique et du décret n° 2007-960 du 15 mai 2007 relatif à la confidentialité des informations médicales conservées sur support informatique ou transmises par voie électronique, communément appelé « décret confidentialité» (articles R 1110-1 à R 1110-3 du code de la santé publique).
Initialement cantonnée à la télétransmission des feuilles de soins électroniques, la CPS permet aujourd'hui d'identifier le professionnel de santé lors de la consultation ou de la création du Dossier Médical Personnel (DMP) de ses patients, à la messagerie sécurisée de santé (MS Santé), réaliser des actes de télémédecine… les applications potentielles sont encore nombreuses.
Cependant, la dématérialisation accrue des données de santé qui accompagne les nouveaux modes d'exercice de la médecine amène à rechercher d'autres moyens d'accès aux données de santé qui permettent de conserver le même niveau de sécurité que celui apporté par l'usage de la CPS là où l'usage de celle-ci s'avère impossible ou mal adapté. Le législateur lui-même en a acté le principe puisque l'article L1110-4 a été complété par la loi n° 2009-879 du 21 juillet 2009 dite « loi HPST » qui introduit à côté de l'utilisation de la CPS tout autre « dispositif équivalent agréé par l'organisme chargé d'émettre la carte de professionnel de santé », rendant ainsi caduques les dispositions du décret de 2007.
Des dispositifs sont d'ores et déjà reconnus comme alternatifs : couple login/ mot de passe associé à un mot de passe à usage unique (OTP[8]) ou certificats logiciel par exemple. Ils devraient être consacrés dans le cadre des travaux de la Politique générale de sécurité des systèmes d'information actuellement menés par les pouvoirs publics.
L'identification des patients, qui garantit que ces données sont appariées de façon certaine à une même identité et qu'elles ne sont pas susceptibles d'être mal utilisées, est assurée quant à elle par l'Identifiant National de Santé, (INS) prévu par L'article L.1111-8-1 du code de la santé publique, qui postule qu'un « identifiant de santé des bénéficiaires de l'assurance maladie pris en charge par un professionnel de santé ou un établissement de santé ou dans le cadre d'un réseau de santé défini à l'article L. 6321 est utilisé, dans l'intérêt des personnes concernées et à des fins de coordination et de qualité des soins, pour la conservation, l'hébergement et la transmission des informations de santé. » Un décret fixe le choix de cet identifiant ainsi que ses modalités d'utilisation.
Dans l'attente de la détermination de cet identifiant, l'ASIP Santé a mis en place dès 2009 une première version de l'INS afin de ne pas retarder le déploiement du cadre d'interopérabilité et permettre la sécurisation de l'accès au DMP. Un INS dit « INS-C » est ainsi calculé localement à partir des traits d'identité contenus dans la carte vitale. Il comporte des imperfections tenant essentiellement à l'absence de carte d'assurance maladie individuelle qui permettrait d'identifier de façon autonome les ayant-droits. Néanmoins, il est apparu plus sécurisant pour le patient d'améliorer les conditions de son identification alors que la situation actuelle reste très insatisfaisante pour le déploiement des systèmes d'information partagés de santé.
Le déploiement actuel des SI de santé autour du patient et de la notion de parcours de soins impose en effet de faire le choix d'un identifiant simple, pérenne, fondé sur des outils de certification déjà existants et reconnus et, dans un contexte budgétaire très contraint, de privilégier l'efficacité à des solutions coûteuses. Le numéro de sécurité sociale, ou « NIR »[9], présente ces caractéristiques.

8. One-time password http://fr.wikipedia.org/wiki/Mot_de_passe_unique
9. Numéro d'inscription au répertoire national d'identification des personnes physiques

En outre, l'utilisation du NIR permettrait de disposer de données de santé fiables directement issues des processus de soins et qui viennent nourrir les bases de données médico-administratives permettant ensuite aux chercheurs de distinguer les éléments utiles à la définition d'une politique de santé publique efficace.
Enfin, cette solution réglerait la question d'un identifiant commun aux secteurs sanitaire et social qui reste un frein à la coordination des soins, notamment dans la prise en charge des maladies chroniques et de la dépendance.
En tout état de cause, le NIR est utilisé comme identifiant permettant d'éviter les doublons et collisions, et ne saurait être utilisé, du fait de son caractère prédictible et non confidentiel, comme clef d'accès aux systèmes d'information. La question est donc à nouveau posée[10] et la CNIL devrait être saisie.

Et demain ? Vers une nouvelle conception des Systèmes d'Information de Santé

Ces réflexions et l'expérience désormais acquise en matière d'échange et partage dématérialisés de données de santé permettent de dégager une vision prospective des principes qui devront à l'avenir guider leur développement.
Tout d'abord, au-delà de l'élargissement de l'accès des données de santé au secteur médico-social et de la mise en place d'une loi sur le secret professionnel partagé, il faut aller vers une redéfinition de la notion de donnée de santé, telle qu'elle semble prônée par les instances européennes. La proposition de règlement du parlement européen et du conseil du 5 janvier 2012[11] sur la protection des données définit ainsi la donnée de santé comme « toute information relative à la santé physique ou mentale d'une personne, ou à la prestation de services de santé à cette personne »
Cette définition élargie doit guider une rénovation des textes permettant une communication plus aisée des informations dans une optique d'amélioration de la coordination des soins, sans pour autant déroger aux principes fondamentaux de la protection de la vie privée et de l'intimité des personnes.
Une définition unifiée permettrait également d'aller vers la mise en place d'une véritable gouvernance des données de santé qui régisse tous les aspects de leur utilisation : coordination des soins, comme on l'a vu, mais également utilisation secondaire.
On parle d'utilisation secondaire des données de santé lorsque les données collectées sont utilisées à des fins de statistiques ou pour la recherche et la mise au point d'indicateurs utiles au pilotage des politiques publiques.
L'utilisation secondaire des données personnelles de santé peut prendre diverses formes, et couvrir des domaines très différents : études épidémiologiques, recherches « biomédicales », activités observationnelles (dont la veille sanitaire), bases médico-administratives.
Or, la France dispose de bases de données médico-sociales et économiques nationales centralisées, couvrant de façon exhaustive et permanente l'ensemble de la population dans divers domaines stratégiques pour la santé publique et la recherche, qui « constituent un patrimoine considérable, vraisemblablement sans équivalent au monde. »[12] : Le PMSI[13] et le SNIIRAM[14] en sont les exemples les plus connus.

10. La CNIL, dans un contexte très différent, a estimé dans son avis du 20 février 2007 que le NIR (ou « numéro de Sécurité sociale ») n'apportait pas les garanties de confidentialité suffisantes pour être l'identifiant du secteur de la santé. Cf « Conclusions de la Commission Nationale de l'Informatique et des Libertés sur l'utilisation du NIR comme identifiant de Sante » Avis rendu le 20/02/2007 http://www.cnil.fr/fileadmin/documents/approfondir/dossier/NIR/Rapport%20NIR.pdf
11. Règlement du parlement européen et du conseil relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données (règlement général sur la protection des données) CE – 25 janvier 2012 http://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2012:0011:FIN:FR:PDF
12. Pour une meilleure utilisation des bases de données nationales pour la santé publique et la recherche, rapport du Haut Conseil de la Santé Publique – collection documents, mars 2012
13. Programme de médicalisation des systèmes d'information http://fr.wikipedia.org/wiki/Programme_de_m%C3%A9dicalisation_des_syst%C3%A8mes_d'information
14. Système national d'information inter-régimes de l'assurance maladie http://www.sante.gouv.fr/IMG/pdf/CNAMTS___Le_SNIIRAM_et_les_bases_de_donnees_de_l_assurance_maladie_en_2011.pdf

Construites de façon dédiée, elles portent toutefois en elles leurs propres limites. Soit parce qu'elles ignorent le sujet de la santé publique, soit en raison de leur finalité précise et organisée (cohortes), l'accès à ces bases est aujourd'hui techniquement hétérogène et difficile. Il nécessite plusieurs avis préalables d'instances qui complexifient l'accès et participent ainsi à une absence de transparence réelle.
En outre, la validité scientifique des données auxquelles on accède peut être discutée (conditions de collecte, durée de conservation…).
L'accès aux données contenues dans ces bases et surtout l'utilisation d'autres sources de données plus proches de la réalité de l'administration des soins devraient révolutionner notre approche de l'épidémiologie, de la veille sanitaire et des politiques de santé publique qui en découlent (Fig.1).

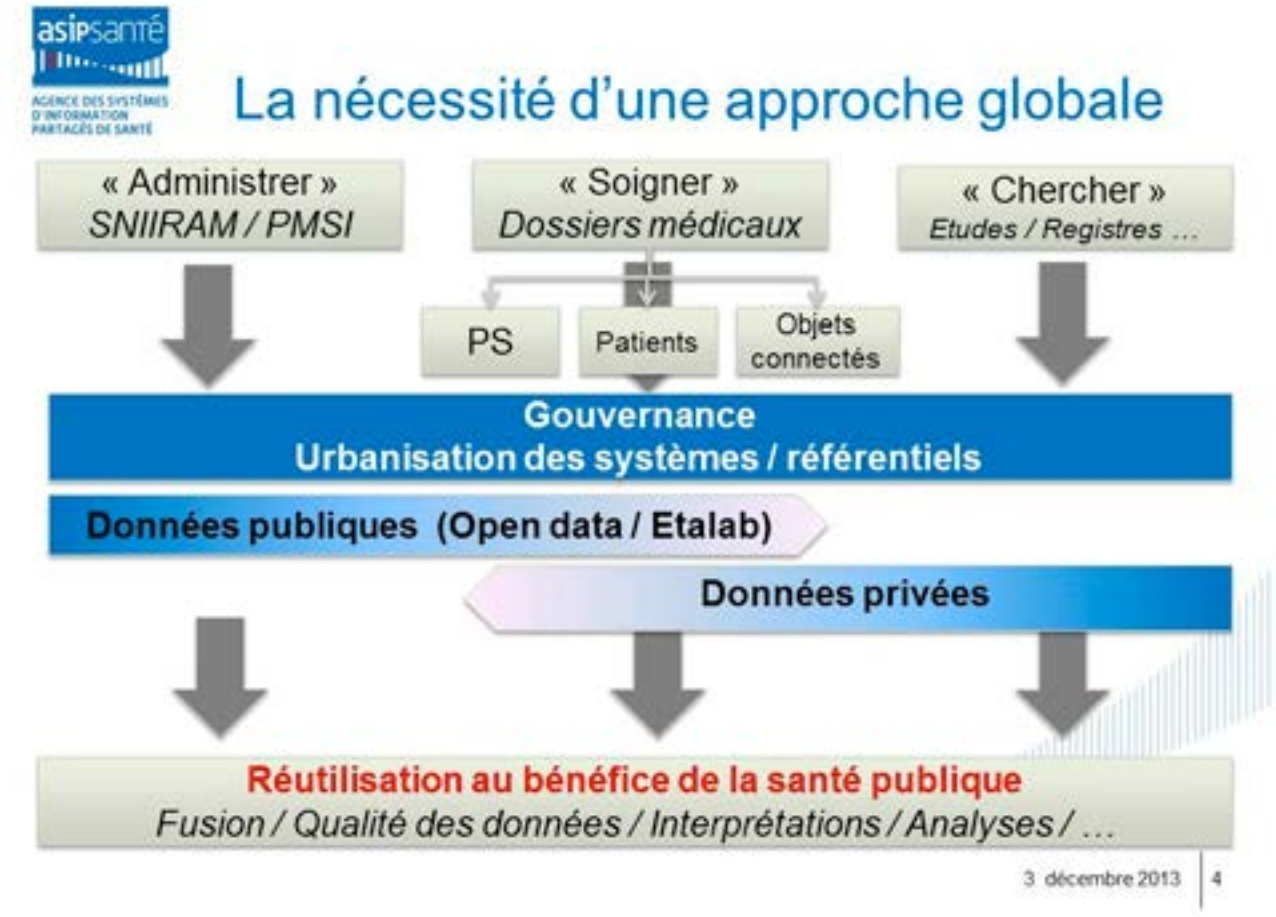


Figure 1. proposition de schéma d'organisation de recueil des données de santé pour leur utilisation secondaire à des fins de santé publique.

Les différentes bases de données de santé, ici médico-sociales avec le SNIIRAM et le PMSI, de soins avec les dossiers médicaux type DMP, et épidémiologiques (recherche) obéissent à des règles de traitement spécifiques et ne partagent pas encore des principes d'urbanisation communs qui pourraient les rendre interopérables (cadre d'interopérabilité). Toutes ces données pourraient être recueillies et compilées, croisées si besoin de façon agile et rapide, dans le respect de principes éthiques communs (anonymisation, finalité, durée de conservation…), afin de servir d'indicateurs en temps quasi réel pour l'évaluation des politiques de santé publique.

Pour faire face à ces problèmes et exploiter enfin tout le potentiel de la « e-santé » pour l'amélioration du système de soins, il sera nécessaire de définir une nouvelle gouvernance des données de santé.
Le Haut conseil pour la santé publique propose deux modèles[15] de gouvernance, en précisant toutefois qu'une entité créée ad hoc ne se substituerait pas aux autorités délivrant les autorisations règlementaires d'accès aux données :

- Gouvernance décentralisée : chaque organisme public gestionnaire de bases de données fixe des règles explicites d'accès (incluant une politique tarifaire et la possibilité de refuser l'accès à ses données) et met en place un « guichet » destiné à traiter les demandes et accompagner les demandeurs.
- Gouvernance centralisée : une structure centrale gère un guichet unique et fait office d'interface entre les demandeurs et les organismes gestionnaires de bases de données, selon des règles homogènes et sous le contrôle d'une instance de gouvernance unique.

Cette gouvernance devra s'accompagner d'une prise en compte de la dimension de santé publique dès la conception des systèmes d'information partagés de santé. Par exemple, en prévoyant dans l'architecture des projets, à côté de la fonction de production de soins, une fonction destinée à produire de la connaissance et une fonction de retour d'information permettant de valoriser les personnes qui les produisent, les professionnels de santé.

Conclusion

Les dispositifs juridiques et techniques mis en place pour prévenir toute utilisation abusive des données personnelles de santé et encadrer leur gestion par un tiers traduisent l'importance attachée par le législateur à la sécurité du traitement de ces données.
Toutefois, l'explosion récente des outils électroniques de partage et d'échange de données, et les usages qui en découlent doivent être pris en compte par les pouvoirs publics afin de fournir un cadre juridique et technique adapté, capable de protéger les droits des personnes mais aussi d'accompagner une vraie révolution technologique qui doit être mise au service de la santé publique.

15. Pour une meilleure utilisation des bases de données nationales pour la santé publique et la recherche, rapport du Haut Conseil de la Santé Publique – collection documents, mars 2012 – P 40

L'apport des bases de données d'origine administrative aux cohortes épidémiologiques : l'exemple de la cohorte Constances

Marie Zins et Marcel Goldberg
Inserm et Université de Versailles-Saint-Quentin

La France, contrairement à d'autres pays développés, ne dispose pas encore d'une cohorte de grande taille pour sa recherche épidémiologique. En revanche, notre pays dispose de bases de données administratives extrêmement riches. Exploiter ces gisements pour constituer à moindre coût de grandes cohortes est une voie de progrès prometteuse : c'est ce qui est entrepris dans la cohorte « Constances ». D'autres expériences pourront suivre, si les problèmes légaux et méthodologiques sont correctement pris en charge.

Les cohortes épidémiologiques en population : un besoin encore méconnu en France

La cohorte épidémiologique est un type d'enquête dont le principe est le suivi longitudinal, à l'échelle individuelle, d'un groupe de sujets. Les cohortes en population générale s'intéressent essentiellement aux causes des maladies, particulièrement les maladies plurifactorielles aux déterminants environnementaux et génétiques multiples. Ces cohortes doivent inclure et suivre, souvent pendant des décennies, de très vastes échantillons pour lesquels sont recueillies de façon prospective des données personnelles, de mode de vie, sociales, professionnelles et environnementales, et qui s'accompagnent de « biobanques »[1]. Globalement, les études de cohorte sont celles qui permettent de proposer les meilleures conditions pour juger en termes de causalité du rôle sur la santé de facteurs de risque (ou d'interventions préventives), en permettant de prendre en compte les évolutions temporelles et les interactions entre facteurs.

Actuellement, l'épidémiologie fait face à la nécessité de développer des études de taille autrefois inimaginable. Qu'il s'agisse de mettre en évidence des risques de faible ampleur associés à l'exposition à des agents potentiellement pathogènes, d'évaluer l'efficacité d'interventions dont on attend des bénéfices d'ampleur modeste, ou de décrire la distribution et l'évolution d'événements peu fréquents, ce sont aujourd'hui des cohortes de centaines de milliers, voire de millions de sujets qui sont suivis de façon prospective pendant des périodes qui s'étendent sur des décennies [1].

1. Une biobanque est une collection d'échantillons biologiques destinés à la recherche scientifique, en biologie (notamment en génomique) et en médecine.

Dans ce paysage, on constate que les cohortes prospectives françaises se caractérisent par leur taille relativement faible, aucune ne dépassant un petit nombre de dizaines de milliers de sujets, alors que certaines cohortes prospectives dans d'autres pays peuvent atteindre plusieurs centaines de milliers de sujets, voire plus. À titre d'illustration, on peut citer en Grande-Bretagne la One Million Women Study [2], le projet UK Biobank [3] qui a mis en place le suivi prospectif de 500 000 personnes, ou la Norvegian Mother and Child Cohort Study qui a inclus 100 000 femmes à la 18ème semaine de grossesse, puis leurs 100 000 nouveau-nés, ainsi que 70 000 pères, soit au total 270 000 personnes [4]. La Nurses'Health Study a été mise en place aux États-Unis dès 1976 et assure le suivi prospectif de près de 250 000 infirmières [5]. Actuellement se mettent en place en Europe de nouvelles très grandes cohortes en Suède, aux Pays-Bas, ou en Allemagne qui doivent inclure et suivre plusieurs centaines de milliers de sujets recrutés en population générale. On peut citer aussi l'exemple des pays scandinaves, qui disposent de multiples registres dans le domaine de la santé, de la protection sociale ou de l'activité économique, couvrant la totalité de la population de ces pays, et qui sont largement ouverts aux chercheurs, permettant par appariement de ces bases de données de constituer des cohortes dont l'effectif se compte en millions de sujets et qui sont à l'origine d'une immense bibliographie scientifique.

La relative modestie des cohortes françaises s'explique par de nombreuses raisons, notamment du fait de difficultés d'ordre financier, organisationnel et technique. Les coûts des cohortes sont élevés, car l'épidémiologie fait essentiellement appel à des données qui sont le plus souvent recueillies auprès des personnes elles-mêmes par des moyens divers : entretiens, auto-questionnaires, examens médicaux, collecte de matériel biologique, etc. Ces coûts sont largement supérieurs aux budgets qu'il est possible de demander aux organismes nationaux de financement de la recherche En effet, contrairement aux autres pays scientifiquement avancés, la France n'a pas mis en place un système de financement adapté, et continue de facto d'ignorer l'importance scientifique de telles plateformes de recherche, malgré des efforts récents (appels à projet « Très grandes infrastructures de recherche - Cohortes » 2009 et « Cohortes » 2010 des Investissements d'avenir). Cependant, les budgets qui ont été distribués sont très loin des coûts véritables, et très largement inférieurs aux financements des cohortes étrangères citées plus haut, montrant bien à quel point les besoins scientifiques sont actuellement sous-estimés par les autorités françaises de la recherche.

D'autres difficultés tiennent à la nécessité de l'implication à long terme des équipes dont la pérennité n'est souvent pas assurée du fait de la quasi impossibilité de disposer de personnel stable et d'un niveau de qualification suffisant en l'absence de statut reconnu pour ce type d'activité dans les organismes publics de recherche. Pourtant la durée des projets est incompatible avec un trop fréquent renouvellement des personnels qualifiés qui doivent assurer la continuité des procédures et des recueils de données.

Or, si on veut que la France se dote d'outils épidémiologiques d'envergure comparable à ce qui existe dans les pays d'un niveau scientifique équivalent, de nouvelles cohortes prospectives sont indispensables, dont l'effectif ne se comptera plus en dizaines, mais en centaines de milliers de sujets.

Les bases de données médico-administratives

Une grande partie des coûts des cohortes prospectives en population vient de la nécessité de « tracer » les sujets et de recueillir pour chacun des données de santé et de situation sociale. Or, de ce point de vue, notre pays dispose d'un atout potentiel d'importance. Il existe en effet en France des systèmes d'information gérés par des organismes de protection médicosociale ou de gestion hospitalière extrêmement puissants, dont peu de pays disposent à l'échelle nationale.

On utilise encore très peu en France les possibilités offertes par ces bases de données, qui offrent pourtant un intérêt potentiel majeur pour la réalisation d'études épidémiologiques. qu'il s'agisse de l'inclusion et du suivi des sujets, ou de l'accès à des données concernant des événements de santé ou de vie socioprofessionnelle d'intérêt. On se restreindra ici à la description des deux principaux systèmes d'information de nature médicale et administrative.

Bases de données concernant des événements de santé

Outre les données de mortalité (statut vital et causes de décès) qui peuvent être obtenus par l'accès au Répertoire national d'identification des personnes physiques (RNIPP) et à la base de données du Centre d'épidémiologie des causes de décès de l'Inserm (CépiDc), il existe différentes bases de données réunissant des données diverses pouvant être utilisées dans des protocoles épidémiologiques.

Le PMSI (Programme de Médicalisation du Système d'Information) a pour objectif de produire des informations à contenu médical sur l'activité hospitalière. Il consiste en un recueil exhaustif d'informations administratives et médicales pour chaque séjour hospitalier (essentiellement diagnostic principal, diagnostics associés et actes pratiqués), qui sont centralisées dans une base de données nationale.

Les systèmes d'informations des différents régimes de l'Assurance maladie enregistrent des données très détaillées sur les consommations de soins remboursés (médicaments, consultations de professionnels de santé, etc.), dont l'objectif premier est la liquidation des prestations d'assurance maladie. Des informations médicales diverses sur les Affections longue durée (ALD), les Accidents du travail (AT) et les Maladies professionnelles (MP), dont l'objectif initial est le contrôle des pathologies ouvrant droit à une prestation, sont également enregistrées. L'ensemble des bases de données concernant les événements de santé est réuni au sein du Système national d'information inter régimes de l'assurance maladie (SNIIRAM) qui concerne aussi bien la médecine de ville que les hospitalisations. Chaque personne est identifiée par un numéro d'anonymat permanent non réversible, qui permet de chaîner toutes les données le concernant dans les différentes sources qui alimentent le SNIIRAM. Au total, le SNIIRAM qui couvre la totalité de la population française, constitue la plus grande base de données de santé au monde.

Bases de données concernant des événements socioprofessionnels

La Caisse nationale d'assurance vieillesse (Cnav) a notamment pour rôle d'assurer le droit au paiement de la retraite. Pour cela, la Cnav a mis en place un système permettant de collecter et traiter les données sociales issues de différents organismes et régimes gestionnaires des prestations sociales pour chaque individu jusqu'à la liquidation de ses droits à la retraite : périodes d'activité professionnelle ou assimilées (chômage, maladie, maternité ou congés parentaux…), incluant les employeurs et la catégorie socioprofessionnelle.

Un apport potentiel majeur pour les cohortes

Dans un contexte épidémiologique, ces bases de données offrent de nombreux avantages : quasi exhaustivité de la population cible (et par conséquent absence de biais de sélection et effectifs immenses pour certaines analyses), quasi absence de perdus de vue pendant le suivi, données parfois plus fiables que celles obtenues par déclaration pour certaines informations (comme les consommations de soins par exemple). Couplées avec des données recueillies auprès des personnes, ces bases de données peuvent apporter des solutions satisfaisantes à

divers problèmes rencontrés par les cohortes : traçage des sujets au cours du suivi, y compris de très longue durée ; acquisition permanente de données d'intérêt, ce qui permet le suivi de nombreux problèmes ; validation de données de déclaration ; analyse des biais de participation à toutes les étapes (inclusion et suivi), allégement des questionnaires.

… malgré certaines limites

Des problèmes de validité des données médicales se posent. Ainsi, l'utilisation du PMSI comme source d'information sur les pathologies s'avère délicate et ne peut reposer uniquement sur le diagnostic principal [5]. De plus, les données de remboursement ne comportent pas d'information sur la nature des maladies traitées, et excluent par définition l'automédication et les prestations non présentées au remboursement. Les ALD ont des limites connues : imprécision des diagnostics, absence d'exhaustivité des cas déclarés, risque de double déclaration [6].

Dans de nombreuses situations, il est donc nécessaire de mettre en place des procédures de validation des diagnostics extraits des bases de données : retour au médecin traitant, confrontation avec des questionnaires remplis par les sujets, croisement avec d'autres sources (données de registre, causes de décès…). Une voie prometteuse est le développement d'algorithmes incluant des données d'ALD, de remboursement de médicaments, de diagnostics et d'actes enregistrés dans le SNIIRAM et le PMSI. Ainsi un travail récent a montré qu'il est possible à partir de ce type de données d'identifier avec d'excellentes sensibilité et spécificité les patients souffrant d'une maladie de Parkinson [7].

L'exemple de la cohorte Constances www.constances.fr

Récemment, grâce à un partenariat avec la Caisse nationale d'assurance maladie des travailleurs salariés (CNAMTS) et un important financement des Investissements d'avenir dans le cadre des Infrastructures nationales en biologie et santé, la cohorte Constances a pu être initialisée [8].

Constances est une importante cohorte épidémiologique destinée à fournir des informations à visée de santé publique et de contribuer au développement de la recherche épidémiologique en constituant une infrastructure largement accessible à la communauté scientifique.

Constances est un échantillon représentatif de la population couverte par le Régime général de Sécurité sociale (plus de 85 % de la population française) âgée de 18 à 69 ans, constitué par tirage au sort. L'effectif total prévu est de 200 000 sujets qui seront inclus sur une période de 5 ans ; le recrutement a commencé courant 2012 et actuellement (février 2014) environ 35 000 sujets sont déjà inclus et les données de 25 000 d'entre eux ont déjà été appariées avec succès avec le SNIIRAM et la Cnav. Sa structure est proportionnelle à la population-cible pour le sexe, l'âge et la catégorie sociale. Les personnes éligibles sont celles qui habitent dans 16 départements dont les Centre d'examens de santé (CES) participent à Constances. L'inclusion des participants se fait dans ces CES : les volontaires complètent un questionnaire concernant leur santé, leurs modes vie et un historique professionnel et bénéficient d'un examen de santé complet ; des prélèvements de sang et d'urine permettent de constituer une biobanque. Le suivi est « actif » : un questionnaire est complété chaque année et une invitation à revenir au CES tous les 5 ans pour un nouvel examen de santé est proposée. Il est également « passif » par appariement annuel avec les bases de données de la Cnav et du SNIIRAM ; le statut vital et les causes de décès sont également suivis dans les bases du CépiDc-Inserm. Les principales données recueillies à l'inclusion et durant le suivi concernent notamment la situation sociale et professionnelle, la santé (morbidité, capacités fonctionnelles physiques et cognitives), le recours aux soins, les comportements, l'exposition à des facteurs de risque professionnels et environnementaux.

L'apport des bases de données d'origine administrative est essentiel à toutes les étapes de la mise en place et du suivi de la cohorte. La constitution de l'échantillon repose sur les bases de données de la Cnav permettant un tirage au sort tenant compte des caractéristiques sociodémographiques et professionnelles ; la Cnav fournit également des données individuelles à l'inclusion et pendant le suivi. Le SNIIRAM fournit des données de santé et de consommation de soins exhaustives et très détaillées.

Outre l'accès à des données nombreuses, les bases de données administratives offrent également d'autres avantages. Elles garantissent la quasi-absence de perdus de vue pendant le suivi de la cohorte, même en l'absence de réponses aux questionnaires, ce qui est essentiel dans le contexte d'études longitudinales. Elles permettent également de contrôler les effets de sélection et les biais potentiels occasionnés par les effets de sélection à l'inclusion comme pendant le suivi (attrition). En effet, les personnes tirées au sort qui ne participent pas ou qui abandonnent la cohorte diffèrent des participants pour de nombreux paramètres liés à la santé et la position sociale. Pour tenir compte de ces différences une « cohorte contrôle» a été constituée, selon une procédure agréée par la CNIL, par tirage au sort d'un échantillon parmi les non-participants. Les mêmes données de la Cnav et du SNIIRAM sont extraites pour les deux cohortes (participants et non-participants), à l'inclusion et durant le suivi ; il est ainsi possible d'identifier les facteurs liés à la non-participation et de produire des estimations de prévalence de maladies et de facteurs de risque redressées pour ces facteurs par des méthodes de repondération.

Les perspectives

L'utilisation de bases de données d'origine administrative peut grandement faciliter les travaux des épidémiologistes, voire améliorer la qualité des études. Il reste cependant de nombreux problèmes à résoudre pour leur utilisation optimale.

Aspects légaux : l'identification des personnes dans les bases de données administratives repose sur le « Numéro d'inscription au répertoire » (NIR). Or la loi Informatique et libertés, qui exige un décret en Conseil d'État, rend pratiquement impossible la collecte de ce numéro dans le cadre d'une étude épidémiologique, ce qui constitue actuellement un obstacle insurmontable pour la plupart des études. Les pouvoirs publics réfléchissent actuellement à une évolution des textes pour assouplir les conditions d'utilisation du NIR.

Par ailleurs, un très important travail méthodologique et technique est nécessaire en raison de la complexité et du volume de ces bases de données. Leur utilisation dans des conditions compatibles avec les contraintes de qualité des études épidémiologiques nécessite des moyens lourds et des compétences spécialisées. Seule une structure de type « plateforme scientifique et technique » pourrait les développer et permettre à la communauté scientifique de bénéficier réellement des bases de données nationales d'origine administrative.

L'exemple d'autres pays montre que tout ceci est faisable, potentiellement très utile et pourrait contribuer au développement en France de grandes cohortes comparables à celles qui existent ailleurs.

Références

[1] Thompson A. Thinking big: large-scale collaborative research in observational epidemiology. Eur J Epidemiol, 2009. 24: 727-31.
[2] Darling GM, Davis SR, Johns JA. Hormone replacement therapy compared with simvastatin for postmenopausal women with hypercholesterolemia. N Eng J Med 1998; 338:64.
[3] Collins, R. and UK Biobank Steering Committee. UK Biobank: Protocol for a large-scale prospective epidemiological resource. 2007, Manchester: UK Biobank Coordinating Centre.
[4] Naess O et al. Cohort profile: cohort of Norway (CONOR). Int J Epidemiol. 2008 Jun;37(3):481-5.
[5] Couris CM, Forêt Dodelin C, Rabilloud M et al. Sensibilité et spécificité de deux méthode d'identification des cancers du sein incidents dans les services spécialisés à partir des données médico-administratives. Rev Epidemiol Sante Publique 2004, 52, 151-60.
[6] Fender, P, Weill, A. Épidémiologie, santé publique et bases de données médico-tarifaires. Rev. Epidemiol. Sante Publique, 2004 ; 52: 113-117.
[7] Moisan F, Gourlet V, Mazurie JL et al. Prediction model of Parkinson's disease based on antiparkinsonian drug claims. Am J Epidemiol 2011;174:354-363.
[8] Zins M, Bonenfant S, Carton M, Coeuret-Pellicer M, Guéguen A, Gourmelen J, Nachtigal M, Ozguler A, Quesnot A, Ribet C, Rodrigues G, Serrano A, Sitta R, Brigand A, Henny J, Goldberg M. The CONSTANCES Cohort: an Open Epidemiological Laboratory. BMC Public Health 2010; 10:479.

Les logiques politiques de l'ouverture des données de santé en France

François Briatte[1] et Samuel Goëta[2]
Doctorants

L'open data semble pleine de promesses pour le système de santé : sécurité sanitaire, transparence des dépenses de santé et des tarifs médicaux… Armée de ces arguments, une coalition hétéroclite d'acteurs fait aujourd'hui pression sur les pouvoirs publics pour obtenir l'ouverture des données de santé. Mais quelles données exigent-ils ? Pour quoi faire ? Et surtout, quel peut être l'impact de l'open data sur la régulation du système de santé français ?

En France, où le scandale du 'Mediator' a incité les pouvoirs publics à réexaminer les moyens de la surveillance pharmaco-épidémiologique, la "libération des données" a récemment pris la forme d'un débat sur l'accès aux données de santé, en particulier aux fichiers anonymisés des caisses d'assurance-maladie, mesure étudiée dans un rapport récent de l'Inspection générale des affaires sociales[3].

Ces données, qui proviennent principalement du système d'information inter-régimes de l'assurance-maladie (SNIIRAM), fournissent une information extrêmement détaillée sur le remboursement des prescriptions de la médecine de ville, ce qui a pour effet de rendre mesurable la consommation médicale à l'échelle d'un territoire donné, et de rendre ainsi visible l'activité (et le coût) des professionnels de santé.

Les revendications d'une multitude d'acteurs exigeant un accès facilité à ces données font allusion directement à la notion "d'open data"[4]. Cette dernière fait référence à une injonction de la société civile et à une obligation légale émergente qui vise à établir la mise à disposition de données publiques, anonymisées et non personnelles, de manière proactive par les administrations. L'open data facilite, techniquement et légalement[5], la réutilisation des données telles que les administrations les détiennent par quiconque, sans exiger de connaitre l'identité ni les objectifs de l'usager[6]. Ce dernier point est crucial dans le domaine de la santé où l'utilisateur de données doit souvent décliner patte blanche avant d'accéder aux données même anonymisées. Dans une démarche mêlant transparence et innovation, les politiques d'open data visent à développer une meilleure connaissance de la société et de l'action publique par les multiples médiations créées par la société civile, journalistes de données et entreprises en tête.

1. Doctorant à l'Institut d'Études Politiques de Grenoble et à l'EuropeanSchool of Political Sciences, Lille ; f.briatte@ed.ac.uk.
2. Doctorant à Telecom ParisTech, département de sciences économiques et sociales, Paris ; samuel.goeta@telecom-paristech.fr.
3. Pierre-Louis Bras, Rapport sur la gouvernance et l'utilisation des données de santé, Paris, Ministère des affaires sociales et de la santé, 2013.
4. Laetitia Clavreul, "Bataille autour de l'accès aux données de santé", Le Monde, 31 janvier 2013.
5. "TenPrinciples for Opening Up Government Information", Sunlight Foundation, Sebastopol (CA), 11 août 2010.
6. Cet élément est mis en exergue dans "l'Open Definition" produite par l'Open KnowledgeFoundation : http://opendefinition.org/.

Or, bien que ce mouvement soit désormais bien ancré suite à la signature d'une charte pour l'open data par les dirigeants du G8 de 2013, l'accès aux données de santé obéit à toute une série de restrictions légales et techniques qui limitent cette réutilisation[7]. De telles réticences ont été exploitées par l'Initiative Transparence Santé, un regroupement d'acteurs hétéroclites dont nous aborderons la composition et les revendications, qui s'interroge dans sa pétition sur les "choses à cacher" que protégeraient les pouvoirs publics ou une éventuelle crainte de voir des dysfonctionnements mis en lumière. Désignant clairement un sens de l'histoire, leur manifeste publié en mars 2013 regrette que la France soit à "contre-courant" à "l'heure de l'open data"[8].

Néanmoins, au-delà de l'accès aux données, la controverse en cours sur les données de santé interroge plus largement la configuration du système de santé français. Historiquement, le terme open data est associé à celui d'open government qui, par un accès plus large à l'information, vise à étendre la participation de la société civile dans l'élaboration des politiques publiques[9]. L'ouverture des données peut alors se comprendre comme une opportunité, pour les acteurs d'un secteur de l'action gouvernementale, de renégocier avec l'État l'étendue de sa régulation, aussi bien en termes de participation de la société civile aux politiques publiques qu'en termes d'accès à un marché comme celui de la santé par des sociétés privées.

La mobilisation de l'Initiative Transparence Santé

L'initiative Transparence Santé (ITS) a été lancée le 25 janvier 2013 et se présente comme un regroupement d'acteurs sans rattachement à une structure formelle. Parmi ses premiers initiateurs, on trouve des personnalités bien insérées dans le milieu de la santé comme Thomas Laurenceau, Rédacteur en chef de 60 millions de consommateurs, Alain Bazot, Président de l'UFC-Que-Choisir ou Christian Saout, ancien Président du Collectif interassociatif sur la Santé (Ciss), un regroupement d'associations de patients. On y trouve aussi des dirigeants d'entreprises de taille moyenne dont le cœur de métier dépend de la diffusion des données de santé.

Parmi celles-ci, Celtipharm, une entreprise qui se décrit comme "spécialiste de la conception et de la vente de programmes marketing-ventes pour l'industrie pharmaceutique" est un de ses membres les plus actifs. La société a exercé en 2013 une intense activité de lobbying auprès du ministère de la santé car l'accès aux données de la santé est son fonds de commerce. En plus de réclamer l'accès aux données publiques, Celtipharm tente de faire valoir l'autorisation accordée par la CNIL en 2011 de collecter des données anonymisées sur le contenu des ordonnances. Il s'agissait de réaliser des "études épidémiologiques à partir de données issues des feuilles de soins électroniques anonymisées". L'entreprise se heurte aux caisses d'assurance maladie et aux mutuelles, qui refusent de mettre à disposition les clefs de déchiffrement indispensables à l'analyse des flux de données chiffrées[10]. La mobilisation de l'entreprise auprès des pouvoirs publics pour exiger de décrypter les télétransmissions des ordonnances a valu à la société d'être nominée aux "Big Brother Awards" de l'ONG Privacy France[11].

Une autre entreprise, Fourmi Santé, mérite d'être citée parmi les acteurs à l'initiative du mouvement. En septembre 2012, l'entreprise a fait l'objet d'une intense couverture médiatique[12]. Elle réutilisait les données mises à disposition par la Caisse nationale de l'assurance maladie des travailleurs salariés (CNAMTS) sur le site ameli.fr. Le site permettait aux utilisateurs de sélectionner les médecins, notamment en fonction des prix moyens des consultations disponibles sur la base de données. La CNAMTS a bloqué l'extraction automatique des données

7. Une récente enquête sur les dépassements d'honoraires à Paris illustre l'important travail à accomplir afin d'obtenir une vue d'ensemble des tarifs des consultations médicales : Jean-Baptiste Chastand, Laetitia Clavreul et Alexandre Léchenet, "Tarifs médicaux : enquête sur les dépassements d'honoraires", Le Monde, 10 avril 2012.
8. Manifeste du collectif Open Data Santé, mars 2013.
9. Harlan Yu et David G. Robinson, "The New Ambiguity of 'Open Government'", UCLA Law ReviewDiscourse, 178, 2012.
10. Guillaume Champot, "CeltiPharm veut déchiffrer les ordonnances transmises à la Sécu", Numerama, 7 mars 2013.
11. "Celtipharm", Big Brother Awards 2013.
12. AFP, "Honoraires médicaux: la "Sécu" interdit à un site de comparer les tarifs", 5 septembre 2012.

qu'effectuait l'entreprise et a émis une mise en demeure exigeant le retrait des liens vers le site ameli.fr[13]. Fourmi Santé a alors répliqué par une intense campagne médiatique puis a joué un rôle moteur dans la création de l'Initiative Transparence Santé.

L'Initiative Transparence Santé est vite parvenue à imposer son message dans l'espace public par une stratégie d'influence efficace. Elle s'appuie sur une intense activité éditoriale à travers la publication du blog opendatasante.com et de tribunes dans le journal Le Monde et le site d'information libéral Contrepoints. Le message de l'Initiative est aussi portée par la société civile avec une pétition signée par près de 2500 personnes. Une rapide analyse sémantique des professions déclarées par les signataires sur le site opendatasante.com montre une très forte proportion de professionnels de santé et environ 30% de retraités. Mobiliser ce dernier public sur une problématique technique comme celle de l'open data témoigne d'une organisation structurée des initiateurs de la démarche et de relais bien établis dans les associations de patients et de consommateurs susceptibles de porter le message auprès de ce public.

La pétition exige une diffusion de l'ensemble des données de santé dans une définition très large de la "manne d'informations relatives à l'utilisation et au fonctionnement de l'offre de soins". Les données sont représentées visuellement comme des détenus à libérer, une image de la Prison de la Santé à Paris illustre la pétition en arrière-plan. La pétition donne quelques exemples de cette "manne" à libérer : "tarifs des professionnels, des produits de santé, des hôpitaux, informations relatives à la qualité et à la consommation des soins, rapports d'activité des caisses primaires d'assurance maladie, efficacité des médicaments, qualité de la prescription"[14]. En somme, le manifeste relaie une exigence d'ouverture de l'ensemble des données publiques relatives au domaine de la santé, dans la lignée du mouvement plus large d'ouverture des données publiques.

13. "La Sécurité sociale censure Fourmi Santé, un site comparateur de tarifs médicaux", Slate.fr, 2013.
14. Manifeste du collectif Open Data Santé, op. cit.

L'objet du litige : la base SNIIRAM

Passé le manifeste qui fait office de vitrine de l'initiative, les tribunes publiées sur opendatasante.com font part d'une demande bien plus ciblée qu'une simple volonté d'ouvrir les données de santé dans leur ensemble. La mission Etalab sous l'autorité du Premier Ministre, en charge de l'ouverture des données publiques de l'Etat, a entrepris un ambitieux recensement des bases de données disponibles en matière de santé produites par l'Etat mais aussi des fondations, des entreprises et même le gouvernement américain[15] ! Près de 280 données produites par 58 organismes sont recensées dans le tableur publié par Etalab. Dans son travail de recensement, Etalab a évalué l'ouverture des données de santé qu'elle a pu identifier selon quatre critères principaux qui correspondent aux exigences les plus communes des militants de l'open data : la liberté d'accès aux données par le grand public plutôt qu'une catégorie de personnes, la gratuité d'accès, la mise à disposition des données dans un format exploitable et enfin la possibilité de réutiliser les données sans restriction. Tous ces critères ont été évalués à la fois pour les données au niveau granulaire défini comme le " niveau le plus fin qu'il est possible d'obtenir en fonction de l'origine de la donnée et du système de collecte" ou agrégé soit le niveau obtenu "en regroupant des données granulaires selon une ou plusieurs caractéristique(s) commune(s)". Le recensement entrepris par Etalab est toujours en cours, de nombreux indicateurs ne sont pas encore renseignés.

Critère	Données agrégées	Données granulaires
Liberté d'accès	Accessible à tous : 95 En accès restreint : 2 Non précisé : 189	En accès restreint : 154 Accessible à tous : 47 En accès fermé : 19 Non précisé : 66
Cout d'accès	Gratuit : 97 Non précisé : 189	Gratuit : 138 Payant : 40 Payant ou gratuit : 10 Non précisé : 98
Format d'accès	Non exploitable : 73 Exploitable : 22 Non précisé : 191	Exploitable : 163 Non exploitable : 25 Non précisé : 98
Condition de réutilisation	Explicité avec restriction : 57 Explicité sans restriction : 22 Non explicité : 9 Non précisé : 198	Explicité avec restriction : 107 Explicité sans restriction : 10 Non explicité : 14 Non précisé : 155

Figure 1. Synthèse de la cartographie des données de santé réalisée à partir du fichier mis à disposition par Etalab (avril 2014)

Comme l'illustre le tableau ci-dessus, les restrictions d'accès se concentrent principalement sur les données granulaires. Ce sont en effet celles-ci qui concentrent les difficultés d'anonymisation, essentielles pour éviter la ré-identification des patients. Ce sont aussi ces données difficilement accessibles qui intéressent le plus les entreprises, les militants et les chercheurs à l'origine de l'Initiative Transparence Santé.
Si l'on cumule les quatre critères d'ouverture des données au niveau le plus proche des revendications des militants de l'open data, on constate que 13 données agrégées et 9 données granulaires sont d'ores et déjà disponibles. La plateforme data.gouv.fr propose ainsi des

15. Etalab, "Cartographie des bases de données publiques en santé", avril 2014.

données comme les honoraires moyens des professionnels de santé, les indicateurs relatifs aux infections nosocomiales, ou encore la base complète des médicaments en vente en France. Ces données déjà disponibles, l'Initiative Transparence Santé ne les évoque pas tout comme elle n'a pas réagi à la cartographie des données de santé en dépit du fait que cette étude serait précieuse à ses revendications.

L'Initiative concentre son action sur un acteur en particulier, la CNAMTS. Depuis mars 2013, peu après la publication de sa tribune fondatrice, l'Initiative a précisé ses exigences dans un billet sur opendatasante.com qui dénonce la mauvaise exploitation de la base SNIIRAM. L'article s'appuie sur un rapport de l'Institut des Données de Santé qui décrit l'absence d'accès direct à la base pour des institutions comme la Haute Autorité de Santé ou l'Agence Nationale de Sureté du Médicament. Le plaidoyer est clair dans ses ambitions : "l'accès à ces données devrait être élargi à l'ensemble de la société civile et non restreint à quelques privilégiés sous la houlette d'un ministère de tutelle décidant qui a le droit d'obtenir de l'information et qui en sera privé."[16]

Que contient cette base tant convoitée ? Selon le rapport de l'inspecteur général des affaires sociales Pierre-Louis BRAS sur la gouvernance et la diffusion des données de santé adressé à la ministre de la Santé Marisol Touraine en septembre 2013, le SNIIRAM tire initialement ses données des feuilles de soins, environ 1,2 milliard d'enregistrements par an qui sont anonymisés selon une procédure stricte. Mise en place par une loi de 1998 dans le Code de Sécurité Sociale et en fonctionnement depuis 2003, la base SNIIRAM se compose de données sur les bénéficiaires, les professionnels et les structures de soin, les demandes de prise en charge et leur remboursement et enfin de données sur les hospitalisations. Les données sont consultables pour une période de 3 ans et l'année en cours mais sont conservées par la CNAMTS pendant 10 ans, en cas de crise nécessitant une exploration des données dans des périodes antérieures.

En juillet 2013, l'Initiative Transparence Santé change de tactique pour enfin mettre la main sur les données de la base SNIIRAM. Elle adresse une requête à la CNAMTS pour obtenir des données sur la consommation de Mediator s'appuyant sur la loi de 1978 qui donne un droit d'accès aux informations publiques : "En juillet, notre collectif a officiellement saisi la CNAMTS d'une demande relative à la consommation de Mediator. Quelles quantités ont été consommées ? Dans quelle mesure les prescriptions étaient médicalement justifiées ? Et surtout, combien la collectivité a-t-elle dépensé afin de rembourser l'empoisonnement de centaines de patients ?"[17]. La CADA (Commission d'Accès aux Documents Administratifs) créée par la loi de 1978 a émis un avis favorable à ce que les données soient communiquées à l'Initiative Transparence Santé : "la communication de ces informations au collectif sous la forme demandée par celui-ci, n'est pas de nature à porter atteinte au secret médical ou au secret en matière commerciale et industrielle"[18]. Une nouvelle requête a donc été formulée à la suite de cet avis et la CNAMTS a adressé sa réponse dans une lettre comportant les données demandées, notamment l'effectif des assurés par département de 1999 à 2009 à qui du Médiator a été prescrit, les spécialités des médecins à l'origine des prescriptions et l'évolution des montants remboursés.

Cette opération a permis à l'Initiative de démontrer les possibilités de l'ouverture des données de santé. Avec les données du Médiator, l'Initiative déploie son argument le plus fort : que la transparence et la plus grande redevabilité (accountability) du système de santé rendues possibles par l'open data permettra d'éviter les prochains scandales sanitaires. L'initiative décrit des scénarios d'usage des données de santé dans lesquels quiconque peut "suivre l'utilisation d'un médicament [...] et mesurer si les recommandations d'usage, émises par les agences sanitaires, n'ont pas été ignorées ou détournées"[19]. Mais, dans les arguments de l'Initiative, il s'agit au delà de reconfigurer la régulation du système de santé en encourageant « le développement de la culture du débat contradictoire et de la contre expertise » reprenant l'expression de l'Institut des Données de Santé. Le comité d'experts de cette institution en

16. ITS, "Accès aux données de santé : l'opposition se précise", 20 mars 2013.
17. ITS, "L'open data, la ministre de la Santé n'en veut pas", Le Monde, 2013.
18. AFP, "Avis favorable pour plus de transparence des données de santé", 30 décembre 2013.
19. ITS, "Débattre d'urgence de l'accès aux données de santé", Le Monde, 30 janvier 2013.

charge du partage et de la diffusion des données entre organismes de santé s'est en effet récemment prononcé en faveur d'une plus large diffusion des données de santé, notamment en conditionnant l'accès non pas à un statut institutionnel mais à la finalité de l'étude conduite avec les données[20].
Ainsi, l'ouverture des données de santé amène plus largement à réfléchir à une reconfiguration du système de santé français. En ouvrant les données de la base SNIIRAM, il s'agit de donner la possibilité à un plus large panel d'acteurs - sociétés privées, chercheurs, citoyens experts - de participer au contrôle du système de santé. Une ouverture qui remet en cause les rapports d'acteurs établis progressivement dans les institutions de santé.

La configuration du système de santé français[21]

Le système français de protection sociale s'est construit sur un rapport de force constitutif : bien que d'inspiration universaliste, sa généralisation n'a pas abouti à la création d'un régime unique de sécurité sociale, ou d'un "service national de santé" à l'image du système de santé britannique. Au contraire, la gestion des prestations sociales y est confiée aux représentants de différentes branches du salariat, réunis au sein de caisses autonomes de l'État et régies par les principes de la gestion paritaire[22].
Dans le secteur de la santé, cette organisation signifie que la régulation de l'assurance-maladie repose primordialement sur un accord entre deux acteurs non étatiques, représentant l'offre médicale et la demande de soins : les professionnels de santé, dont la représentation syndicale en France est très fragmentée, et les dirigeants des caisses d'assurance-maladie, au sein desquelles la CNAMTS, qui assure quatre personnes sur cinq en France et qui finance actuellement les trois quarts des dépenses de santé, est amenée à jouer un rôle central.
Cette organisation explique pourquoi la "libération" des données de santé en France revient en partie à les ouvrir… aux services de l'État : jusqu'à récemment, les bases de données de la Sécurité sociale n'étaient directement accessibles qu'aux agents de l'assurance-maladie. Cet accès, désormais étendu aux services ministériels, permet à l'État de faire intervenir sa propre expertise des dépenses de santé, sans avoir recours à celles de personnels indépendants.
Vu sous cet angle, "libérer les données" de l'assurance-maladie revient ainsi à supprimer une prérogative des organismes chargés, à la création du système de santé, d'autonomiser la régulation des soins des services de l'État. Cette dynamique d'étatisation graduelle prévaut dans les réformes du système de soins français sur les vingt dernières années, se retrouve également dans d'autres branche de la sécurité sociale[23], et a plusieurs équivalents fonctionnels dans d'autres systèmes de santé.
Ces changements d'attribution obéissent à deux principaux impératifs pour les pouvoirs publics: réduire la pression fiscale exercée par les branches maladies des États-providence, et prévenir les scandales sanitaires causés par les crises de santé publique[24]. Il pourrait éventuellement en servir d'autres, comme la réduction des inégalités sociales et territoriales de santé, qui ne dispose toutefois ni du sentiment d'inévitabilité associé à la régulation des dépenses de soins, ni du sentiment d'urgence concomitant aux crises sanitaires.

20. ITS, "Le comité d'experts de l'Institut des données de santé favorable à un très large accès", 4 mai 2013.
21. Cette section a fait l'objet d'une présentation à l'Apéro Science & Web n°33 en février 2014. Merci aux participants pour leurs questions et leurs réactions.
22. Dominique Damamme et Bruno Jobert, "Les paritarismes contre la démocratie sociale", Pouvoirs, n°94, 2000, p. 87–102.
23. Jean-Claude Laborier et Bruno Théret, Le système français de protection sociale, Paris, La Découverte, 2009.
24. Didier Tabuteau, Démocratie sanitaire, Paris, Fayard, 2013, p. 181.

Les implications de l'ouverture des données de santé

La "libération" des données de santé a ainsi plusieurs interprétations possibles, qui vont du simple travestissement d'intérêts industriels à l'expression de revendications citoyennes portées par des représentants d'associations de consommateurs. Les tensions relatives au système de soins français expliquent en partie cette situation, dans la mesure où la lisibilité des données de l'assurance-maladie est la résultante d'un rapport de force institutionnel organisé autour de la régulation des soins.

Pour ces raisons, la limitation de l'accès aux données joue, dans le domaine de la santé, le rôle de barrière à l'entrée d'un champ d'action stratégique, dédié au contrôle des dépenses de santé, et où s'élabore la mesure économique de la maladie et des soins. L'élite décisionnelle de ce champ est largement composée de hauts fonctionnaires qui ont intérêt à l'ouverture de ces données, mais prioritairement voire exclusivement pour équiper leurs propres services.

Que signifierait, à terme, l'accès de tous aux données de santé en France ? L'effet premier d'un tel changement serait de faire baisser drastiquement les coûts d'analyse de l'activité clinique pour les acteurs déjà impliqués dans cette entreprise, au premier rang desquels les acteurs industriels de l'assurance et des produits de santé. L'effet espéré d'une modernisation de la surveillance épidémiologique en France est quant à lui à mettre en perspective avec la réaction des professionnels de santé à l'examen "ouvert" de leur activité par quiconque, plutôt qu'au moyen des dispositifs volontaires existants mis en place au sein des Agences Régionales de Santé[25].

Comprise dans le contexte plus général de la réforme des États-providence européens, la libéralisation de l'accès aux données de santé faciliterait l'évaluation économique de la consommation médicale pour les acteurs publics, tout en rendant les acteurs privés plus à même de participer à la compétition autour de l'offre dans ce domaine. Cette dualisation caractérise de nombreuses réformes de régimes de protection sociale, recommandées par des institutions comme la Banque mondiale et visant à établir des systèmes mixtes public-privé de couverture assurantielle[26].

L'ouverture des données repose toutefois sur une prémisse invitant à définir un enjeu supplémentaire : l'accès généralisé aux données permettrait à de nouveaux entrants d'investir le champ d'action stratégique de la santé en remettant en question la qualité des décisions qui y sont prises, chiffres et données à l'appui. Dans cet ordre d'idée, il est peu surprenant d'apercevoir des sociétés de services aux industries assurantielles et pharmaceutiques au premier rang de ceux appelant à l'émergence d'une contre-expertise en matière de santé.

Le projet d'une contre-expertise citoyenne, "statactiviste"[27], dans le secteur de la santé, ferait quant à lui écho au projet de la démocratie sociale, mais aussi à celui de la démocratie sanitaire, dans la mesure où il viserait à permettre aux intéressés eux-mêmes d'infléchir responsablement sur le déroulement et l'organisation des soins, en plaçant entre leurs mains les ressources nécessaires pour en décider collectivement. Ce projet n'a néanmoins que très peu de prise sur la réalité pratique des demandes auxquelles les pouvoirs publics sont aujourd'hui confrontés.

25. Ces dispositifs désignent les "contrats" passés entre les pouvoirs publics et les médecins libéraux, et qui permettent notamment à l'État d'inciter les praticiens à prescrire des médicaments génériques et à se rendre dans des zones géographiques de faible démographie médicale.
26. Bruno Palier, Gouverner la Sécurité sociale, Paris, Presses Universitaires de France, coll. "Quadrige", 2005.
27. Isabelle Bruno, Julien Prévieux et Emmanuel Didier (dir.), Statactivisme, comment lutter avec des nombres, Paris, La Découverte, coll. "Zones", à paraître.

Les enquêtes multimode :
attention aux effets de mode

Gaël de Peretti[1]
et Tiaray Razafindranovona[2]
Insee, Département des méthodes statistiques

Les organismes producteurs d'information statistique, et en particulier la Statistique publique (« Official statistics » dans le monde anglo-saxon) doivent faire face à une demande d'enquêtes toujours plus exigeante en termes de qualité, qualité au sens large à savoir précision, pertinence, comparabilité, cohérence, clarté, fraîcheur, dans un contexte général de restriction budgétaire. L'utilisation d'internet comme mode privilégié ou complémentaire de recueil des données paraît une bonne solution. En effet, il est a priori peu coûteux, que ce soit dans ses dimensions financière ou de durée. Ainsi, il y a une volonté commune à ces organismes pour recourir de façon plus systématique à la collecte par internet, et plus généralement multimode, pour faire face aux problèmes de couverture, d'échantillonnage, de non-réponse ou de mesure, tout en respectant des contraintes budgétaires. Il s'agirait de limiter conjointement l'erreur d'enquête totale et les coûts d'enquête. Cependant, si la collecte par internet est un mode peu coûteux, elle pose des problèmes méthodologiques non négligeables : couverture, auto-sélection ou biais de sélection, non-réponse et les difficultés de sa correction, « satisficing »[3], etc. Aussi, avant de développer ou généraliser l'utilisation du multimode, les effets de mode de collecte doivent être étudiés pour voir dans quelle mesure la collecte multimode peut être effectivement intégrée dans le processus de production des enquêtes.

Les informations statistiques sont produites à partir de deux grands types de sources : les données administratives et les enquêtes. Ces dernières se sont développées en France à partir des années 1950 suite à la diffusion en France des techniques d'échantillonnage. Depuis deux décennies, le recours aux données administratives se multiplie afin de limiter la charge de réponse qui pèse sur les entreprises et les ménages, mais aussi de profiter d'informations plus ou moins exhaustives que l'on peut ventiler de façon détaillée. Toutefois, la statistique d'enquête reste une source importante de la statistique publique compte tenu de sa capacité à s'adapter à des questionnements complexes, des populations spécifiques, voire les combinaisons des deux.

1. gael.de-peretti@insee.fr
2. tiaray.razafindranovona@insee.fr
3. Voir plus loin

Cependant, si les enquêtes ont été développées à l'origine pour éviter des recensements coûteux, ces opérations, compte tenu de la complexité croissante des protocoles et de la taille importante des échantillons nécessaire pour produire des informations sur des sous-populations plus ou moins fines, sont elles-aussi coûteuses. Or, les instituts nationaux de statistique (INS) et les organismes publics ou parapublics producteurs de données statistiques doivent faire face à la fois à des restrictions budgétaires mais aussi à une baisse tendancielle du taux de réponse, ce qui n'est pas sans incidence sur la qualité des enquêtes.
De ce fait, ces organismes réfléchissent à la mise en place de nouveaux processus de production pour les enquêtes en recourant à une collecte multimode, et plus particulièrement à la collecte par internet afin de réduire les coûts de production sans diminuer la qualité des informations produites, c'est-à-dire en limitant autant que possible l'erreur d'enquête totale. Il faut aussi avoir en tête que la qualité d'une enquête ne se réduit pas à la seule précision des estimations qu'elle permet de produire mais doit aussi prendre en compte d'autres dimensions comme la pertinence, la comparabilité (dans le temps ou l'espace), la cohérence (avec d'autres sources), l'accessibilité et la clarté pour les utilisateurs potentiels, la fraîcheur, etc.

L'erreur d'enquête totale, un paradigme ancien de nouveau à la (multi)mode

Si le terme erreur d'enquête totale est relativement récent (Andersen, Kasper, Frankel et al., Total Survey Error, 1979), le concept est né en même temps que les enquêtes. En effet, dès le début, des travaux portent sur les différentes dimensions possibles de l'erreur d'enquête et de multiples typologies ont été développées pour distinguer les sources d'erreur.
Ainsi, les principales sources d'erreur d'une enquête seraient :

- l'erreur de spécification : conception du questionnaire pas adapté aux objectifs de l'enquête ;
- l'erreur de couverture : base de sondage incomplète ou incluant à tort des unités que l'on ne souhaite pas enquêter ;
- l'erreur d'échantillonnage : erreur due au fait que l'on interroge qu'une fraction de la population visée ; c'est l'erreur la plus documentée et qui possède l'avantage de pouvoir en général être calculée ;
- l'erreur de non-réponse : elle peut être totale si l'unité ne répond à aucune question de l'enquête ou partielle si seule une partie du questionnaire est concernée. Là encore, il est possible dans la plupart des cas de corriger et de quantifier les effets de cette non-réponse ;
- l'erreur de mesure ou d'observation : erreur dont les sources sont multiples. Elle peut provenir de l'enquêté. Dans ce cas, les deux sources les plus étudiées sont relatives à des notions de psychologie : la désirabilité sociale, à savoir la tendance à fournir des réponses conformes aux normes sociales ou donnant une bonne image de l'enquêté, de façon consciente ou pas ; le « satisficing » c'est-à-dire le fait de choisir une réponse « dont on se satisfait » mais qui n'est pas la réponse optimale. Cet effet de satisficing dépend à la fois de la difficulté de la question mais aussi des compétences de l'enquêté et de sa motivation. Par exemple, lors d'une question avec plusieurs modalités, les personnes sujettes au satisficing peuvent avoir tendance à choisir la première des modalités ("primacy") ou la dernière ("recency"). De façon générale, elles sont enclines à répondre de façon machinale. Mais les erreurs d'observation peuvent aussi provenir de l'enquêteur ou de l'enquête ;
- l'erreur de traitement des données : là encore les sources sont multiples (saisie, codification, pondération, tabulation, imputation, etc.).

Parmi ces erreurs, comme nous l'avons écrit précédemment, certaines sont susceptibles d'être quantifiées et sont prises en compte via des techniques statistiques plus ou moins sophistiquées : erreur d'échantillonnage, erreur de non-réponse. Pour d'autres, les résultats sont empiriques et les solutions quand elles existent ne sont pas généralisables mais adaptées à la source analysée. Parfois, il est même possible d'identifier ou de constater des erreurs sans

être en mesure de pouvoir les corriger. Tout au plus, le statisticien pourra préciser les limites de l'information qu'il a produite.

Ainsi, aussi séduisant soit le paradigme de l'erreur d'enquête totale, il n'est pas encore totalement opérationnel pour les producteurs de données. Il est toutefois utile pour identifier les sources potentielles d'erreur, pouvoir apporter des informations sur le poids relatif de ces différentes composantes et ainsi orienter sur quels champs il faut faire des progrès méthodologiques pour une opération donnée et enfin réfléchir a priori au moyen de les éviter.
S'il n'est pas nouveau, ce paradigme est à l'origine de nombreux travaux méthodologiques et est au cœur de nombreux colloques, comme par exemple le séminaire annuel ITSEW (International Total Survey Error Workshop, premier en 2005 puis annuel depuis 2008) du NISS (National Institute for Statistical Science). Par ailleurs, dès lors que l'on souhaite modifier un protocole de collecte, cette approche conceptuelle paraît tout à fait adaptée pour réfléchir aux effets de ces modifications d'un point de vue global. Aussi, le renouveau de ce paradigme n'est-il pas sans lien avec le développement des enquêtes multimode.

« Il y a plus de quarante ans que je fais des enquêtes multimode sans que j'en susse rien »

La notion de mode intervient à plusieurs étapes dans une enquête : contact, collecte, suivi, etc. Par mode, nous entendons façon de faire ou de procéder pour réaliser le contact, la collecte, le suivi, etc., d'une enquête. Le mode de contact recouvre les moyens employés pour solliciter la participation d'un ménage ou d'une personne. Le mode de collecte correspond à la façon dont on va recueillir l'information. Le mode de suivi concerne plutôt les opérations longitudinales, c'est-à-dire avec plusieurs interrogations, et les relations post-enquêtes avec les répondants comme l'envoi de premiers résultats, le suivi d'adresse, etc.
En toute rigueur, une enquête est dite multimode dès lors que le mode de contact, le mode de suivi, le mode de collecte ou les modes de collecte, etc., sont différents, ce qui est quasiment toujours le cas dans les enquêtes de la Statistique publique. En effet, une fois l'échantillon constitué, les personnes sont souvent contactées via un courrier ou lettre-avis leur expliquant les objectifs de l'enquête et l'importance de leur participation, puis ils sont approchés directement ou par téléphone par les enquêteurs afin de fixer un rendez-vous, rendez-vous qui se passera généralement en face à face avec une collecte assistée par un ordinateur. Par la suite les enquêtés pourront être recontactés soit parce qu'ils font partie d'un panel ou d'une enquête en plusieurs vagues, soit parce que l'on souhaite contrôler les informations qu'ils ont fournies, etc.
En fait, le terme enquête multimode fait plutôt référence à la notion de collecte multimode. Généralement, trois cas de collecte multimode sont distingués : le multimode intégré ; le multimode séquentiel ; le multimode concurrent. Dans le premier cas, il s'agit d'interroger la même personne selon différents canaux. Le cas le plus classique consiste à recueillir une partie de l'information en face à face via un enquêteur et une autre à l'aide d'un questionnaire auto-administré. Il peut s'agir d'un carnet recensant les dépenses du ménage sur une période donnée (enquête Budget des familles par exemple) ou les activités d'une personne (enquête Emploi du temps), mais aussi d'un questionnaire spécifique par exemple sur des questions sensibles afin de limiter les biais de désirabilité sociale déjà évoqués précédemment. Dans le deuxième cas, il s'agit au fur et à mesure de la collecte de proposer des modes différents de collecte afin de limiter la non-réponse. Par exemple, l'organisme va contacter une entreprise par internet, puis en cas de non-réponse par téléphone, puis toujours en cas de non réponse envoyer un enquêteur sur place. Enfin, le multimode concurrent consiste à offrir dès le début de la collecte le mode de son choix à l'enquêté, souvent avec le même souci de vouloir limiter la non-réponse.

Les modes de collecte les plus classiques sont : le face à face ; le téléphone ; le questionnaire auto-administré qui peut être un questionnaire papier, un questionnaire en ligne sur internet (plus rarement un questionnaire audio intégré au sein d'une enquête en face à face qui permet d'assurer la confidentialité des réponses malgré la présence d'un enquêteur). Toutefois, on peut aussi parler de collecte passive, quand par exemple, des données d'enquête sont appariées avec des sources externes. C'est le cas de l'enquête de l'Insee sur les Revenus fiscaux et sociaux qui permet, entre autres, de calculer le taux de pauvreté : on réalise un appariement entre les données de l'enquête Emploi, des fichiers fiscaux et des fichiers sociaux afin d'améliorer la qualité des informations recueillies sur les revenus et les aides sociales perçues.
Comme nous l'avons énoncé en introduction, les producteurs de statistique doivent faire face à une baisse tendancielle des taux de réponse, en particulier pour les enquêtes Ménages (par la suite, l'essentiel des résultats exposés portera sur les enquêtes Ménages). Cette baisse des taux de réponse peut s'expliquer à la fois par un accès plus difficile au logement des enquêtés (digicode, interphone, etc.), un moindre sens civique ou plus simplement une multiplication des enquêtes qui peuvent lasser les enquêtés. Ainsi, il devient de plus en plus difficile pour un simple citoyen de distinguer entre une enquête menée par un institut privé souvent à but lucratif et une enquête de la Statistique publique censée produire des informations pour éclairer le débat public. Pour lutter contre cette baisse, le recours à la collecte multimode paraît une solution séduisante.

Un exemple d'enquête multimode

Dans le cadre d'un règlement européen, l'Institut national de la statistique et des études économiques (Insee) réalise chaque année depuis 2007 une enquête téléphonique sur les Technologies de l'information et de la communication (TIC). Compte tenu de l'absence de base de sondage téléphonique exhaustive, l'Insee a lancé plusieurs opérations méthodologiques entre 2009 et 2011 pour limiter l'erreur de couverture avant de mettre en place un nouveau protocole. La collecte est réalisée en partie par téléphone (personnes retrouvées dans l'annuaire téléphonique, échantillon S1), en partie par internet/papier avec deux sous-populations : une sous-population dont les personnes sont présentes dans l'annuaire (échantillon S21) ; une sous-population non-présente dans l'annuaire (échantillon S22). Il est donc possible après standardisation des échantillons S1 et S21 d'identifier a priori un effet de mode, les personnes ayant été affectées aléatoirement à un mode de collecte. Cet effet de mode est corrigé en partant du principe que l'enquête téléphonique est le bon étalon. Cette correction est appliquée aux répondants internet/papier non présents dans l'annuaire. Cette méthodologie a permis d'obtenir des résultats cohérents avec d'autres enquêtes de la Statistique publique réalisées en face à face (ne souffrant donc pas de problème de couverture) sur le taux d'équipement (enquête Statistiques sur les ressources et conditions de vie) et l'utilisation d'internet (enquête Emploi du temps). Cette intégration de la collecte par internet/papier dans le processus de production est possible du fait de la grande simplicité du questionnaire.

La collecte par internet a des avantages indéniables mais il ne faut pas céder sans réflexion à cet effet de mode

Compte tenu de la diffusion des ordinateurs et d'internet au sein de la population, le recours à la collecte par internet est une alternative intéressante à plus d'un titre pour les instituts producteurs d'information statistique. Tout d'abord, c'est le mode de collecte le moins coûteux, le coût marginal d'une enquête étant quasi nul. Ensuite, il est rapide et la gestion de la collecte peut se faire en continu. De même, il peut être interactif et apporter un soutien à l'enquêté via des liens hypertexte, des info-bulles, des Foires aux questions (FAQ), des contrôles intégrés qui peuvent éviter des erreurs de saisie. Enfin, c'est un mode peu intrusif qui permet à l'enquêté de répondre au moment où il le souhaite et peut aussi limiter les biais de désirabilité sociale puisque a priori, il est toujours plus facile de se confier à un ordinateur qu'à un enquêteur.
Mais, il faut aussi avoir à l'esprit qu'une partie de la population n'a pas accès à internet et qu'elle possède des caractéristiques socioéconomiques particulières. Par ailleurs, les choix techniques retenus par les instituts doivent être compatibles avec la plupart des navigateurs web et s'adapter dans une certaine mesure aux nouveaux outils d'accès à internet comme les tablettes, les smartphones, etc. Mais le défaut majeur de la collecte par internet est son faible taux de réponse. Ainsi, dans les expérimentations en cours à l'Insee depuis le début des années 2010, le taux de réponse des enquêtes internet est de 15 à 20 % (on atteint 35 à 40 % en intégrant la collecte papier associée). Cela n'est évidemment pas sans conséquence sur la qualité de l'information produite.

Aussi de nombreux travaux méthodologiques sont réalisés depuis les années 2000 sur le multimode pour voir dans quelle mesure, il est possible de mettre en place de tels protocoles sans détériorer l'erreur totale d'enquête. Car, ce qu'il faut savoir, c'est que lorsque l'on souhaite agir sur une composante de l'erreur d'enquête totale, cela peut avoir des conséquences sur les autres composantes. Ainsi, on peut offrir à des enquêtés difficilement joignables ou récalcitrants à une enquête en face à face la possibilité de répondre par internet afin de diminuer l'erreur de non-réponse. Il faudra s'assurer que les erreurs de mesure liées par exemple à un fort « satisficing » de ces personnes ne viennent pas limiter le gain en erreur totale d'enquête.

À la recherche des effets de mode

De nombreuses études de méthodologie d'enquête s'intéressent à la mesure des effets de mode. Il s'agit d'analyser les effets du mode de collecte sur l'estimation du phénomène étudié. Le principe générique est de comparer les réponses fournies par deux modes différents. Cet effet de mode global est généralement séparé en deux effets : effet de sélection ; effet de mesure souvent appelé effet de mode dans la littérature.
Pour essayer de contrôler le premier, les techniques utilisées sont multiples mais le principe général est identique : il s'agit de rendre les deux échantillons comparables en s'assurant de la similarité sur des caractéristiques sociodémographiques observables dans les deux modes de collecte. L'hypothèse forte sous-jacente est que la prise en compte de ces variables de contrôle suffit à supprimer les effets de sélection. Une fois cet effet de sélection éliminé, les écarts constatés entre les deux modes de collecte sont qualifiés d'effet de mode ou d'effet de mesure (voir exemple en encadré). Évidemment, il est probable que nous ne disposions pas forcément de toutes les variables nécessaires à la bonne prise en compte des effets de sélection. Ceci est d'autant plus préjudiciable si les liens entre une variable omise et la sélection sont forts et que de plus le pouvoir explicatif de cette variable est important pour notre variable d'intérêt.

Un exemple d'effet de mode

Fin 2010, l'Insee a réalisé une enquête expérimentale sur le logement et la mobilité résidentielle sur internet auprès de 10 000 personnes tirées dans les fichiers fiscaux, ce qui permettait, entre autres, d'avoir des informations sur le revenu fiscal et le logement (surface, nature, c'est-à-dire appartement ou maison, statut d'occupation). Ces personnes ont reçu une lettre-avis début octobre. Les non-répondants ont reçu une première lettre de relance deux semaines après contenant un coupon réponse leur permettant d'obtenir une version papier du questionnaire (sans enveloppe préaffranchie), puis une deuxième un mois plus tard. Le taux de réponse est de 24 % (20 % si l'on ne prend pas en compte les répondants papier). La structure des répondants à cette expérimentation, en termes de caractéristiques sociodémographiques et de conditions matérielles du logement, était très différente de celle de l'enquête Logement 2006. Afin de corriger de ces effets de structure (ou de sélection), les données ont été standardisées (on parle de calage sur marges) pour que la structure, en termes d'âge, de sexe, de diplôme, de statut d'occupation, de revenu fiscal, et enfin de tranche d'unité urbaine, surface, nombre de pièces, nombre d'habitants et statut d'occupation du logement, soit identique à celle observée en population générale en 2010. Après cette standardisation, on constate des réponses presque toujours plus négatives sur internet que dans l'enquête Logement sur toutes les questions d'opinion sur le logement, comme les problèmes de bruit, de température (excès de froid ou de chaud), de relation avec le voisinage, de qualité de l'air, etc. Deux explications sont avancées. Certains considèrent que malgré les efforts réalisés pour prendre en compte les effets de sélection, les personnes souhaitant se plaindre de leur logement seraient plus enclines à répondre (diminution de la qualité de l'enquête). D'autres estiment que les biais de désirabilité sociale liés à la présence d'un enquêteur diminueraient. En effet, il peut paraître délicat de reconnaître devant l'enquêteur une privation de confort, ce qui n'est pas le cas par internet.

Mais l'identification des effets de mode n'est pas une fin en soi. Dès lors que l'on souhaite réaliser une enquête multimode, il faut se poser la question de l'agrégation des résultats produits par les différents modes de collecte. Pour certains auteurs, l'objectif principal est de repérer les effets de mode pour pouvoir à terme les réduire. Pour d'autres, il existe un étalon parmi les modes et il s'agit de s'approcher des résultats produits par cet étalon. Par ailleurs, si les estimateurs de paramètres classiques comme les totaux, les moyennes, les quantiles sont identiques, il faut aussi contrôler que les corrélations entre la variable d'intérêt et les covariables sont semblables quel que soit le mode.

Parmi les résultats établis, il est clair qu'il existe une hiérarchie entre les différents modes que sont le face à face, le téléphone, les questionnaires auto-administrés papier puis internet en termes de taux de réponse et de coût : ils sont décroissants du face à face à internet). Par ailleurs, d'autres résultats semblent faire consensus sur les erreurs de mesure. Ainsi, les biais de désirabilité sociale seraient moins forts pour les questionnaires auto-administrés que pour le téléphone puis le face à face. En revanche, le biais de satisficing serait plus important sur internet puis sur papier, au téléphone et enfin en face à face. Ainsi, il y a beaucoup plus de non-réponse partielle dans les questionnaires auto-administrés que dans les enquêtes téléphoniques ou en face à face. De même, lorsqu'il y a des batteries de question (souvent d'opinion), avec des modalités de réponse identiques (par exemple tout à fait d'accord, plutôt d'accord, plutôt pas d'accord ; pas du tout d'accord), les répondants internet ont tendance à

choisir plus systématiquement la même réponse. Enfin, les enquêtés seraient plus pessimistes en l'absence d'enquêteur. Toutefois, de nombreux auteurs posent la question de la difficile généralisation d'un résultat obtenu lors d'une expérimentation à l'ensemble des enquêtes. Dans une certaine mesure, cette prudence incite à construire des expérimentations spécifiques à une enquête donnée pour étudier dans quelle mesure l'introduction d'une collecte multimode, et plus particulièrement de la collecte par internet peut permettre de réduire les coûts sans détériorer l'erreur d'enquête totale, voire en la diminuant. Pour l'instant, ce qui ressort des premières expérimentations d'enquête internet/papier lancées par l'Insee depuis le début des années 2010, c'est que compte tenu des faibles taux de réponse, il est difficile d'affirmer que les effets de mode mesurés sont « nets » des effets de sélection. Aussi à court terme, tant que les taux de réponse ne se seront pas améliorés, il paraît plus sage de voir internet comme un mode complémentaire plutôt que le mode principal des enquêtes Ménages.

Références

[1] Andersen R., Kasper J., Frankel M. R. and associates "Total survey error" San Francisco, Jossey-Bass Publisher, 1979.
[2] Couper M. P. "The Future of Mode Data Collection", Public Opinion Quaterly, 75 (5), 889-908, 2011.
[3] Frippiat, D., Marquis, N. « Les enquêtes par internet en sciences sociales : un état des lieux » Population, 65(2), 309-338, 2010.

Prévoir l'accroissement du nombre des personnes âgées, anticiper ses conséquences

Comptes rendus de deux Cafés de la statistique

Jean-François Royer
SFdS

Chacun sait que le nombre et la part des personnes âgées vont augmenter dans de nombreux pays, dont la France. L'ampleur de cet accroissement est évalué par les projections démographiques de façon convaincante. Pour en anticiper les conséquences, il faut recourir à d'autres hypothèses : hypothèses économiques, lorsqu'il s'agit du financement des retraites ; hypothèses sanitaires et sociales, lorsqu'il s'agit du risque de dépendance. On est alors beaucoup plus loin du consensus.
Deux Cafés de la Statistique se sont tenus sur ce thème en janvier et février 2014 à Paris : l'un sur « L'avenir des retraites », avec Didier Blanchet (Insee) et Antoine Bozio (École d'économie de Paris), l'autre sur « Dépendance et vieillissement », avec Alain Colvez (Inserm). Les comptes rendus détaillés sont disponibles sur le site de la SFdS [1], [2].

En France en 2060, le nombre des personnes de 60 ans ou plus aura été multiplié par 1,8 par rapport à sa valeur de 2007, le nombre des personnes de 75 ans ou plus aura été multiplié par 2,3. Ces chiffres sont tirés des dernières projections démographiques de l'Insee [3]. Ces projections reposent sur plusieurs hypothèses, dont celle selon laquelle le progrès de l'espérance de vie va se poursuivre ; leurs résultats sont assez convaincants pour les prendre comme base de nos raisonnements. Même si l'espérance de vie cessait de croître, l'accroissement du nombre des plus de 60 ans serait encore important.
Pour en apprécier les conséquences sur les régimes de retraite, dont les ressources viennent essentiellement des actifs, on calcule un simple ratio « nombre de personnes de 60 ans ou plus / nombre de personnes de 20 à 59 ans » (figure 1 ci-dessous). Plus ce ratio est élevé, plus les retraites sont difficiles à financer. Un tiers de l'augmentation de ce ratio d'ici 2060 provient de la fin des effets du « baby-boom » que la France a connu après la guerre de 1939-1945, un tiers provient de la montée en régime des hausses d'espérance de vie passées, et c'est seulement le dernier tiers qui dépend des hausses futures de l'espérance de vie.
De telles projections démographiques, jointes à des hypothèses sur la croissance économique, sont à la base des scénarios élaborés régulièrement par le Conseil d'orientation des retraites (COR). Les derniers scénarios publiés par le COR en décembre 2012 ont été considérés comme rassurants par certains : dans le scénario « médian », le déficit des régimes de retraites et le poids des retraites dans le produit intérieur brut seraient en 2060 un peu inférieurs à ce qu'ils sont aujourd'hui. Mais ce résultat relativement optimiste est très sensible aux hypothèses économiques faites sur l'évolution du chômage et de la productivité [4].

Figure 1. Ratio des 60 ans et plus aux 20-59 ans, en %
Lecture : En 2010, il y avait 39 personnes de plus de 60 ans pour 100 personnes âgées de 20 à 59 ans. Dans une population où le vieillissement n'aurait dépendu que de la baisse de la mortalité, ce ratio aurait été de 48,5 %.
Source [1]

Financer les retraites n'est pas le seul défi que l'accroissement du nombre des personnes âgées pose à la société. Il faut aussi prendre en compte leur état de santé. Un certain nombre d'entre elles doivent faire face à des handicaps chroniques, qui dans certains cas les rendent dépendantes. On peut retenir six dimensions pour la reconnaissance des incapacités et des handicaps :

- la mobilité physique dans l'environnement de la personne (du lit à l'espace social) ;
- l'indépendance physique pour les actes élémentaires de la vie courante (s'habiller, se laver sans l'aide d'un tiers) ;
- les occupations, les activités socialement valorisantes ;
- l'intégration sociale ;
- la suffisance économique ;
- l'orientation dans le temps et dans l'espace.

Cette conceptualisation permet de désigner les personnes qui ont besoin d'être aidées. Elle est à la base des grilles utilisées pour affecter les personnes à des groupes homogènes, que ce soit pour attribuer des aides publiques comme l'APA « Allocation personnalisée d'autonomie », ou pour des enquêtes statistiques visant à cerner les effectifs en cause. La dernière enquête en date, en France, est l'enquête « Handicap-Santé » de 2007-2009 : cette enquête complexe associe plusieurs modes de collecte et comprend un appariement avec les données de l'assurance-maladie [5]. Grâce à de telles enquêtes, on peut chiffrer la prévalence de certaines incapacités à divers âges et en déduire une « espérance de vie sans incapacité », ou « espérance de vie en bonne santé », sans doute le meilleur concept pour rendre compte de l'état sanitaire d'une population [6]. Eurostat publie désormais annuellement les espérances de vie en bonne santé pour tous les pays de l'Union Européenne en utilisant une enquête européenne harmonisée [7]. Toutes ces évaluations restent fragiles, et l'évolution temporelle ne fait pas consensus. Les pessimistes soutiennent que les années de vie supplémentaires sont autant d'années d'incapacité ; les optimistes au contraire estiment qu'on va faire reculer les incapacités. Arbitrer entre ces thèses est un enjeu important pour le système d'observation, s'il veut aider les pouvoirs publics à bien anticiper les conséquences du vieillissement.

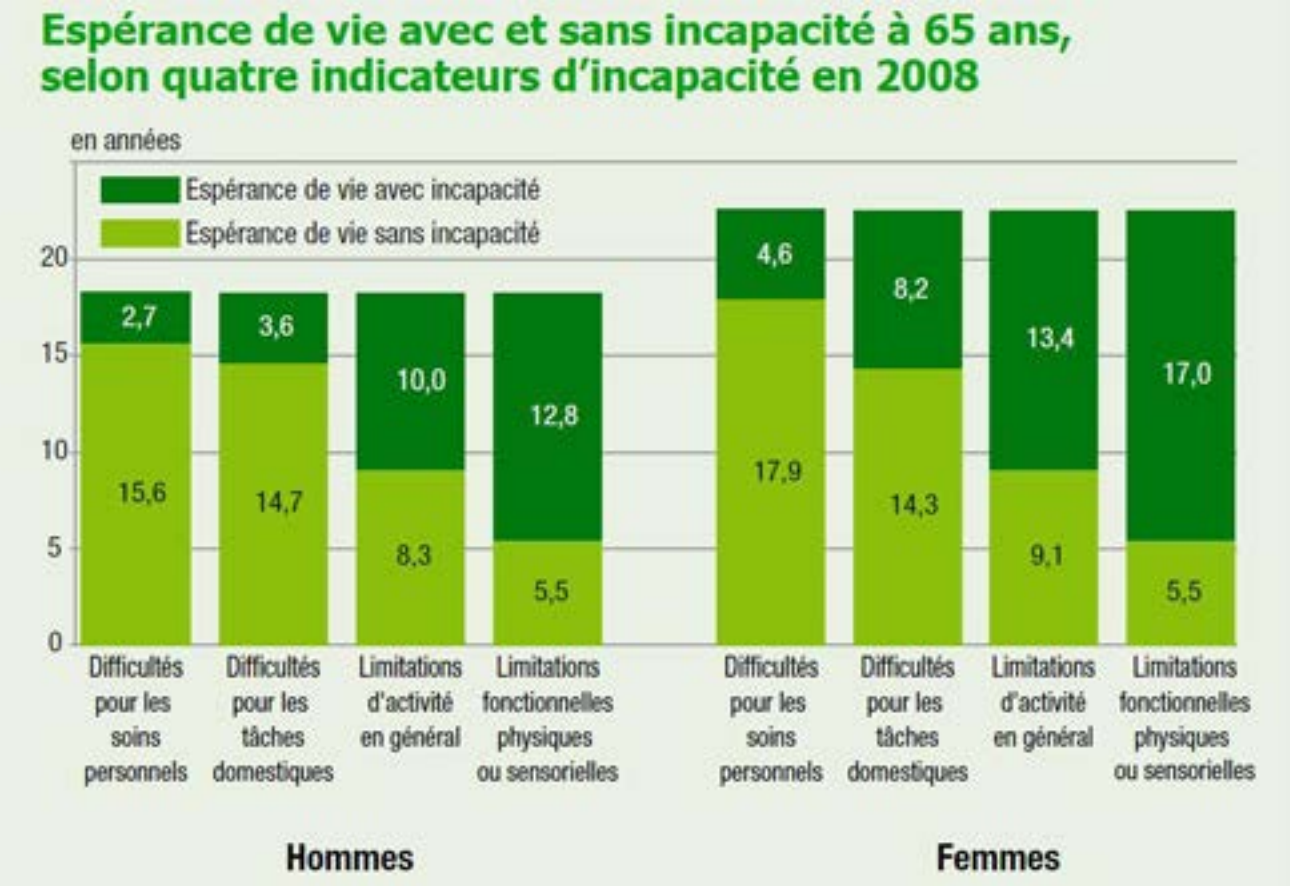


Figure 2. Espérances de vie en France métropolitaine (population des ménages ordinaires)
Lecture : L'espérance de vie des femmes à 65 ans est de 22,5 ans ; leur espérance de vie sans limitation d'activité est de 9,1 ans
Source [6] - Enquête Handicap-Santé 2008

Références

[1] Compte rendu du Café de la statistique du 14 janvier 2014 « L'avenir des retraites » : http://www.sfds.asso.fr/ressource.php?fct=ddoc&i=1673
[2] Compte rendu du Café de la statistique du 12 février 2014 « Dépendance et vieillissement » : http://www.sfds.asso.fr/ressource.php?fct=ddoc&i=1743
[3] « Projections de la population à l'horizon 2060 – Un tiers de la population âgé de plus de 60 ans » Nathalie Blanpain, Olivier Chardon - Insee Première n°1320 Octobre 2010
[4] « Retraites : vers l'équilibre en longue période ? » Didier Blanchet – Les notes de l'Institut des politiques publiques n°3 – Février 2013
[5] « L'enquête Handicap-santé : présentation générale » Gérard Bouvier – Document de travail F1109 – Direction des statistiques démographiques et sociales – Insee
[6] « L'état de santé de la population en France » Sandrine Danet - Études et Résultats n° 805 – juin 2012 – Drees
[7] « Les espérances de vie en bonne santé des Européens » Jean-Marie Robine, Emmanuelle Cambois – Population et Sociétés n°499 – Avril 2013 - Ined