DIRECTION DE LA RECHERCHE, DES ÉTUDES, DE L'ÉVALUATION ET DES STATISTIQUES

Conclusion du groupe de travail sur les risques de ré-identification dans les bases de données médico-administratives

Annexe 9 du rapport de la Commission Open Data en santé

Juillet 2014

A la suite du rapport Bras[1] qui lui a été remis le 3/10/2013, la ministre, Madame Marisol Touraine, a demandé à Franck Von Lennep, directeur de la recherche, des études, de l'évaluation et des statistiques (DREES), de diligenter une expertise technique sur la sécurité des données concernant le risque de ré-identification des personnes à partir de données [supposées] anonymes.

L'étude a été réalisée de novembre 2013 à avril 2014 par un groupe de travail animé par André Loth (DREES) et composé de : Alireza Banaei et Max Bensadon (ATIH), Hélène Caillol (CNAMTS), Nora et Frédéric Cuppens (Institut Telecom Bretagne), Emmanuelle Denis (DSS), Françoise Dupont (INSEE et CASD), Noémie Jess (DREES), J-Pierre Le Gléau (expert), Grégoire Rey (INSERM-CépiDc).

Y ont également contribué (pour le test de jeux de données anonymes) Maxime Bergeat (INSEE), Dominique Blum (expert) et les équipes techniques du CASD.

Le comité de pilotage, présidé par Franck von Lennep, s'est réuni 3 fois. Ont participé à tout ou partie de ces réunions Max Bensadon (ATIH), Philippe Burnel (DSSIS), Anne Coat (ANSSI), Claude Gissot (CNAMTS), Nora et Frédéric Cuppens (Institut Télécom Bretagne), François Godineau (DSS), Philippe Cuneo et Michel Isnard (INSEE), Jean-Pierre Le Gléau (expert), André Loth (DREES), Christian Saout (CISS) ainsi que Marc-André Beaudet (expert de la CNIL, en observateur).

Ce rapport constitue l'annexe n°9 du rapport de la Commission open data en santé du 9 juillet 2014 (disponible sur le site de la DREES).

[1] Rapport sur la gouvernance et l'utilisation des données de santé (sur les risques de ré-identification dans les bases de données médico-administratives, voir notamment pp. 26 à 30 et 56 à 58 du rapport).

1/ L'objet de l'étude

En substance la mission confiée au groupe de travail était

- d'identifier et évaluer les risques de ré-identification dans les bases médico-administratives considérées (SNIIRAM, PMSI…) notamment pour les jeux de données issus de ces bases et ayant donné lieu à une diffusion relativement large (données hospitalières du PMSI, échantillon généraliste des bénéficiaires) ou dont l'anonymat faisait débat (datamart des consommations inter-régimes de l'assurance maladie) ;
- d'établir une ligne de démarcation entre
    - jeux de données anonymes (pouvant donc être mis en open data ou publication sans restriction)
    - jeux de données présentant un risque de ré-identificaton et devant pour cette raison être en accès restreint…
- De recommander des moyens pour élargir, dans cette offre de données, la part et le volume des données en accès libre ;
- De préciser pour les données en accès restreint (parce que comportant des risques de ré-identification) sous quelles conditions techniques les personnes habilitées à accéder à ces données devraient pouvoir le faire afin de limiter le risque.

Le groupe a entendu des chercheurs et parcouru la littérature sur ce sujet. Il a été surpris et déçu de constater qu'il n'existe pas vraiment encore de doctrine établie et que la plupart des responsables de bases de données, publiques et privées mais aussi les responsables des autorités de contrôle (les équivalents de la CNIL) en sont encore à tâtonner. Le groupe des 29 CNIL européennes vient ainsi de définir une liste de trois conditions suffisantes pour qu'un jeu de données puisse être qualifié d'anonyme mais ces trois conditions ne sont pratiquement jamais réunies et on est renvoyé dès lors à une évaluation au cas par cas[2] : la veille technologique et scientifique sur ces questions devra être poursuivie demain, à l'initiative des gestionnaires des bases mais aussi dans les instances de gouvernance et dans les instances de contrôle afin que la doctrine s'affirme (et évolue quand c'est nécessaire).

2/ L'évaluation du risque de ré-identification dans le cas des données médico-administratives de santé

Il est entendu qu'il n'est question ici que des bases de données dé-identifiées ou « pseudonymisées », dites parfois, à tort, anonymes, ou de « jeux » de données issus de ces bases.

Le risque de ré-identification des personnes dans ces bases ne vient pas tant du pseudonyme utilisé[3] que de la possibilité pour des tiers de reconnaître, dans des jeux de données, des personnes sur lesquelles ils disposent d'informations par ailleurs.

Classiquement (cf. le Modèle d'évaluation harmonisée du risque –MEHARI- du Club de la sécurité de l'information français), le risque, dépend à la fois :

- de la probabilité ou potentialité que l'événement se réalise, malgré des mesures dissuasives ou préventives (potentialité généralement évaluée sur une échelle de 1, faible, à 4, très fort) ;
- de l'impact (noté de 1 à 4 également) qu'a sa réalisation, compte tenu d'éventuelles mesures compensatoires ou réparatrices…

[2] http://www.cnil.fr/linstitution/actualite/article/article/le-g-29-publie-un-avis-sur-les-techniques-danonymisation/

[3] Dans les bases SNIIRAM et PMSI, le pseudonyme est le NIR chiffré deux fois de manière irréversible. Une bonne organisation (par exemple des clés de chiffrement secrètes ou une table de correspondance, confiées à plusieurs personnes différentes extérieures à la base de données) évite le risque que le pseudonyme permette de remonter à la vraie identité de la personne, identité qui est parfaitement inutile pour les traitements ordinaires. Dans les cas où remonter à l'identité des personnes est nécessaire (pour rendre possible des appariements avec d'autres jeux de données, d'enquête par exemple), le principe de cloisonnement évite que les gestionnaires de la base connaissent les vraies identités. Ce principe de gestion séparée des données directement identifiantes (nom, NIR, numéro de téléphone…) est celui qui est posé par le projet de règlement européen sur la protection des données personnelles (art. 81 et 83).

Risque = Potentialité x Impact

S'agissant de l'**impact** d'une divulgation non souhaitée de données personnelles de santé, le groupe a considéré qu'il était *très fort*, s'agissant de données classées comme sensibles dans le droit français et européen, et protégées par le droit constitutionnel à la vie privée. Des exemples de conséquences potentiellement très dommageables pour la personne concernée peuvent être cités (portant sur la carrière professionnelle ou l'image ou la vie de famille notamment).

Le fait que certaines personnes puissent rendre publiques elles-mêmes des informations sur leur propre santé ne change rien à l'évaluation de l'impact pour les victimes d'une divulgation non souhaitée. En outre, une fois l'information divulguée, le mal est fait et on ne peut ni remettre l'information dans la boîte ni considérer qu'une compensation financière réparera le préjudice.

Si on s'accorde à évaluer l'impact comme très fort, cela conduit à refuser en l'espèce une conception utilitariste de l'analyse bénéfice/risque où on admettrait que l'utilité économique pour le plus grand nombre compense le préjudice subi par quelques uns (supposés peu nombreux) dont la vie privée serait divulguée. Les membres du groupe ont préféré une autre conception de l'analyse bénéfice-risques, consistant à maximiser le niveau de bénéfice pour un niveau de risque maîtrisé. Cela permet de conclure à l'intérêt :

- de rendre les données personnelles de santé accessibles – notamment aux chercheurs ou à des bureaux d'études –dans des conditions convenables (rapidité, assistance…)- mais avec l'accord de la CNIL et en contrôlant le risque de divulgation ;
- d'ajuster le niveau des précautions au niveau du risque ;
- et de maximiser la production et la diffusion des données anonymes issues -par divers procédés- des données individuelles brutes.

S'agissant de la **probabilité de ré-identification** et de l'efficacité de la dissuasion ou des mesures préventives, le groupe a considéré sans s'y appesantir que la diffusion de données agrégées issues des bases médico-administratives ne présentait pas en général de risque de ré-identification et qu'il n'y avait donc pas lieu d'en limiter l'accès à telle ou telle catégorie de personnes ou d'organismes comme c'est encore actuellement le cas, et qu'il y avait en outre certainement des moyens d'accroître la richesse et la pertinence des statistiques mises en ligne.

Comme on le verra, le groupe a également travaillé sur la possibilité d'étendre la production et l'usage de jeux de données individuelles -jeux exhaustifs ou échantillons- ayant pour caractéristiques que le risque de ré-identification y soit a priori inexistant et que ces données puissent donc elles aussi être publiées et réutilisées par tous.

Cependant, il lui était d'abord demandé de se prononcer sur la politique de diffusion des données médico-administratives individuelles par le ministère de la santé et les agences ou établissements publics placés sous sa tutelle. S'agissant des données individuelles à vocation exhaustives du SNIIRAM, du PMSI et de l'ensemble SNIIRAM-PMSI, il confirme l'évaluation concluant à un niveau élevé du risque :

- le risque que présente une diffusion mal contrôlée des données hospitalières du PMSI exhaustif (dans les conditions et avec les précautions insuffisantes[4] qui caractérisent son mode de diffusion actuel) ;

- le risque que cela entraîne pour l'Échantillon généraliste des bénéficiaires (EGB) au 1/97, accessible à un nombre important d'organismes, où on trouve les parcours de soins de 600 000 personnes sur (potentiellement) 20 ans. Cet échantillon a été malheureusement compromis par la diffusion trop large du PMSI lequel permet comme on l'a vu d'identifier des personnes hospitalisées et de les retrouver, avec les mêmes caractéristiques dans l'EGB ;

- le risque que présenterait un accès trop large au datamart de consommations inter-régimes (DCIR) exhaustif de l'assurance maladie, si par exemple il devenait ouvert à toutes les personnes ayant le statut

[4] Les calculs de D. Blum, confirmés par l'ATIH, donnent les ratios de 89 % et 100 % de séjours où une personne est seule à présenter les caractéristiques cherchées (respectivement sur l'ensemble des séjours de 2008 et sur les séjours des patients hospitalisés au moins 2 fois dans l'année). Ces calculs ont été effectués en 2011sur une base présentant le même type de « floutage » que les bases diffusées au cours des années récentes (imprécision volontaire sur la date de sortie et code géographique correspondant à au moins 1000 habitants). La diffusion à plusieurs centaines d'exemplaires de copies de cette base complète, sous la forme de disques, la diffusion systématique du fichier de chaînage des séjours, l'existence avérée de copies sauvages (et l'intérêt d'en obtenir en raison de leur valeur marchande et pour éviter des procédures et un paiement), l'impossibilité pratique d'empêcher ces copies ou de contrôler le respect des engagements pris par les titulaires des autorisations de la CNIL ont été les principaux arguments cités.

de chercheurs, même sans possibilité de croisement des variables sensibles et même sans croisement avec le PMSI (ce qui lui retirerait d'ailleurs une bonne part de son intérêt).

Dans tous ces cas, le nombre de personnes présentant les mêmes caractéristiques au regard des quasi-identifiants (date en mois des soins, durée d'hospitalisation, arrêt de travail, âge en années…) est très petit, tendant rapidement vers la valeur 1 si les données sont chaînées dans le temps : **tous les parcours de soins sont différents**.

Si on prend l'exemple de l'INSEE, celui-ci ne prend pas le risque de diffuser un fichier exhaustif présentant ce type de risques autrement que via un dispositif sécurisé à des personnes dont le besoin de connaître les données a été préalablement vérifié et évalué (c'est également ce qui est prévu pour les données fiscales).

Dans cette évaluation de la probabilité de ré-identification, il convient de distinguer trois temps :

- ✓ Le *comptage* du nombre de « personnes dans une case » (personnes présentant les mêmes caractéristiques au regard des quasi identifiants que sont l'âge, le sexe, le lieu de résidence, le lieu, la nature et les dates de soins, la date de décès éventuellement). Il présente un caractère objectif. Ce nombre est fortement réduit jusqu'à devenir vite égal à un pour toute la population dès que les données relatives à une même personne sont chaînées[5] ;
- ✓ Le *classement* des quasi-identifiants par degrés de notoriété n'est guère contestable non plus. Les dates de naissance et de décès par exemple sont aisément connues de tous ; les dates et lieu d'hospitalisation ou les dates d'arrêt de travail restent assez faciles à connaître, même si pour une personne donnée, le nombre de personnes susceptibles de les connaître est généralement petit (proches, employeur, assureur). En revanche, les dates de consultations médicales ou de soins par des auxiliaires médicaux sont plutôt mal connues et vite oubliées ;
- ✓ L*'estimation* de la probabilité que ces informations donnent lieu à une divulgation par les personnes qui en ont eu connaissance est plus subjective (« à dire d'experts ») :
    - o cette probabilité augmente évidemment avec le nombre de personnes mises dans le secret, directement ou indirectement, et dépend donc de la procédure de sélection des demandes d'accès ;
    - o elle est réduite avec l'emploi de dispositifs techniques permettant le confinement des données, la traçabilité des accès et des requêtes et l'interdiction des copies, alors qu'au contraire la diffusion des données au moyen de copies et d'extractions augmente la probabilité de divulgation par négligence et/ou malveillance ;
    - o elle est réduite aussi avec le caractère dissuasif des sanctions pénales pour divulgation illicite, que le groupe a jugé
        - ▪ assez efficace dans le cas d'institutions telles que les assurances (secret de pratiques illicites difficile à garder, souci d'image…)
        - ▪ mais peu efficace devant d'autres « attaquants » (par exemple adversaires politiques, malveillance dans un cadre familial ou professionnel, presse à scandale, démonstration qu'on peut le faire ou simple curiosité…) surtout si la probabilité d'être pris est très faible (pas de traçabilité pour les données extraites). On notera que les cas de divulgation des données peuvent alors rester confinés à un petit cercle.

En tout état de cause, l'argument « jusqu'ici tout va bien » ne dispense pas, selon le groupe de travail, de se conformer aux standards professionnels en vigueur et ne peut justifier la continuation du mode actuel de diffusion des données du PMSI et de l'EGB, ni l'ouverture incontrôlée des accès au DCIR, qui sont porteuses de risques pour les personnes concernées et qui seraient de nature, en cas d'incidents publics, à saper la confiance dans le système d'information.

Il appartient à la CNIL de faire savoir quelle sera désormais son attitude concernant la diffusion des données du PMSI, mais le GT recommande au ministère de la santé de revoir sa politique de diffusion des données qui est jugée trop restrictive dans le cas des données agrégées que chacun s'accorde à classer comme anonymes et trop laxiste dans le cas des données du PMSI.

[5] Même sans chaînage des données, certaines configurations facilitent la réidentification des personnes (patients soignés loin de chez eux, petits établissements, décès…).

Pour mémoire, le groupe de travail ne s'est pas penché sur le cas des usages opérationnels des données (par exemple dans les caisses d'assurance maladie pour les feuilles de soins ou dans les ARS pour les données du PMSI) : ces usages sont ceux qui ont justifié initialement le recueil des données et sont donc par hypothèse nécessaires au service public. Le groupe rappelle seulement que dans ces cas aussi l'accès aux données doit être encadré (personnes habilitées individuellement et tenues au secret professionnel[6], tenue de registres, accès sécurisé…).

3/ Une réponse graduée selon le risque de ré-identification

L'offre que le groupe de travail propose de faire à ce stade (travaux à poursuivre, en liaison avec les services de la CNIL et la future instance de gouvernance s'articule autour de trois axes :

a) une offre pour tous les publics en accès direct sur internet

La manière la plus simple de mettre à disposition des données anonymes issues de bases indirectement nominatives est évidemment de produire des tableaux statistiques répondant par des données agrégées aux questions que se posent les différentes catégories d'utilisateurs. Le groupe n'a pas exploré cette piste faute de temps mais elle ne pose pas de problème de principe : seulement de faisabilité technique et de coûts. La limite de l'exercice est que la demande est potentiellement infinie, alors que la production de statistiques agrégées a un coût et répond souvent mal ou trop lentement aux besoins particuliers des uns et des autres, notamment aux besoins des chercheurs ou à des questions d'une grande technicité ou exigeant le recours à plusieurs sources de données.

Une solution est de rendre plus « agile » la production de ces tableaux statistiques à l'aide d'outils techniques par exemple en mettant à la disposition des utilisateurs des bibliothèques de requêtes en libre service ; à la limite cela revient à laisser interroger la base à l'aide d'outils logiciels paramétrables fournissant des réponses anonymes sans que l'auteur de la requête ait lui-même accès aux données. Le même résultat peut être obtenu, plus classiquement en organisant des collaborations (un « guichet ») entre une équipe d'experts attachés à la base de données et les experts « métier » des demandeurs, pour fabriquer les sorties agrégées adaptées aux besoins, sachant que la mobilisation d'agents pour des travaux à façon a un coût et que là aussi on ne répond pas toujours aux besoins les plus pointus.

L'autre solution explorée par le Groupe de travail « risques de ré-identification » consiste en la production d'un grand nombre de « jeux » de données rendus anonymes, tels que pour chacun de ces jeux, le nombre de personnes présentant des caractéristiques identiques soit suffisant (K=10 au moins dans le test) et que parmi ces personnes il y en ait toujours un nombre suffisant (L=3 au moins dans le test) qui aient des maladies différentes. On a démontré que c'est possible et que les critères d'anonymat peuvent être vérifiés… Des précautions doivent être prises cependant pour qu'il ne soit pas possible de reconstituer ainsi des jeux de données identifiants… Une sécurité supplémentaire peut être obtenue en utilisant des techniques d'échantillonnage (si la règle de tirage est secrète).

Les fichiers qui seront mis à disposition de cette façon doivent être absolument anonymisés, pour tout type de consultant qui n'a d'ailleurs pas besoin de se faire connaître (assureur, employeur, voisins, famille…).

Le groupe de travail est parti du fichier PMSI et a construit, par généralisation ou suppression de variables, des jeux répondant aux critères d'anonymisation : pour toute combinaison de variables quasi identifiantes, il y a au moins 10 individus. Parmi ces individus, il y a au moins 3 types de « maladies » différentes.

À titre d'exemple, on a construit un fichier comprenant pour chaque séjour d'hospitalisation : le sexe et l'âge (en 18 groupes d'âge) de la personne hospitalisée, son mode d'entrée et de sortie (en 2 groupes : « domicile » ou « autre ») et la durée d'hospitalisation (12 modalités) et le groupe homogène de malades (indicateur de la maladie).

Un autre fichier donnant un peu moins de précision sur la durée d'hospitalisation (2 modalités), mais comprenant la région de résidence du malade semble aussi répondre aux critères. D'autres exemples sont décrits en annexe mais un grand nombre de jeux répondant aux critères d'anonymisation peut a priori être mis à disposition de la même manière,

Des fichiers analogues pourront, selon la même méthode, être construits à partir du SNIIRAM.

[6] Le groupe rappelle qu'il n'est pas besoin d'être médecin pour respecter et faire respecter le secret professionnel sur des données personnelles de santé.

Le groupe ne donne pas de recommandation définie sur le nombre minimal de personnes présentant les mêmes caractéristiques que doit comporter un jeu de données pour être considéré comme anonyme (k-anonymat)… Certains membres du groupes préconisaient d'examiner les seuils K=3 et L=3 qui préservent en théorie l'anonymat en admettant qu'il serait peut-être nécessaire de se fixer sur un seuil un peu plus élevé (K=5, L=3 par exemple). Les travaux devront être poursuivis à ce sujet, car certains membres du groupe considéraient à l'inverse que ces valeurs sont trop basses pour garantir l'anonymat.

Il est mis à l'étude aussi la proposition que des fichiers plus détaillés sur les caractéristiques médicales du séjour (médicaments en sus, DMI, diagnostics et actes…) soient fournis mais sans indication du nom de l'établissement et sous la forme de sondage (au 1/10 voire 1/3). Certes, les données médicales ne sont pas considérées comme un quasi identifiant (c'est l'information que « l'assaillant » est supposé *rechercher* dans la base : s'il en disposait déjà , il n'aurait pas besoin de la base) mais elles doivent être appauvries elles aussi dans les jeux de données anonymes parce que la partie médicale du fichier de résumé des sorties anonymisés (RSA) est très *discriminante* (chaque RSA est à cet égard différent des autres et on pourrait, si on conservait cette information commune dans des jeux de données par ailleurs appauvris, reconstituer le jeu d'origine).

La technique de sondage confère en tout état de cause une sécurité supplémentaire et peut être appliquée partout où un fichier exhaustif n'est pas indispensable, à la condition bien sûr qu'on ne parte pas d'un fichier où chaque personne a déjà notoirement des caractéristiques uniques et que la règle de tirage soit gardée secrète..

Le caractère anonyme de ces fichiers, notamment l'impossibilité de remonter au fichier d'origine, devra être validé par des instances à désigner (on pense notamment à un avis de la CNIL sur le mode de construction de ces fichiers et non sur chaque fichier pris isolément).

La piste du « bruitage » (ajout ou substitution de données rendant « fausses » les données individuelles mais conservant des valeurs telles que moyenne, médiane, écart type etc.) n'a pas été suivie, notamment parce qu'elle nous a paru supposer une utilisation prédéterminée du jeu de données alors qu'en open data, les usages sont à la main des utilisateurs. Elle suppose également une compétence forte des utilisateurs pour en maitriser les limites d'utilisation.

b) une offre plus limitée, destinée à un public ciblé : un nouvel EGB ?

S'agissant des données à faible risque d'identification, la proposition faite est de constituer au moins un échantillon de données de grande taille tel qu'on ne puisse y reconnaître personne de manière certaine et dont l'accès soit ouvert à des organismes publics ou privés dont les missions et les compétences le justifient. Le principe de l'accès leur étant acquis, une limitation s'impose cependant en raison d'un risque résiduel tenant au caractère très particulier de certains parcours de soins[7] : il conviendra qu'ils souscrivent un certain nombre d'engagements sur le respect des règles d'accès, d'usage des données et de publication des résultats.

Il appartiendrait à la CNIL après avis de l'instance technique décrite au point précédent de valider le niveau de risque. Dès lors, l'accès à tel ou tel organisme (chercheurs, organisations professionnelles, administrations) pourrait relever comme aujourd'hui d'un arrêté ministériel.

Un fichier spécifique pourrait être construit à partir du SNIIRAM, chaîné avec le PMSI. Ce fichier ne serait pas exhaustif, mais issu d'un sondage (par exemple au 1/10[8]). Il est important que le mode de sondage et les paramètres retenus pour celui-ci demeurent strictement confidentiels. Il sera alors impossible à quiconque de savoir si telle ou telle personne est présente dans ce fichier.

Si l'on découvre alors une personne ayant exactement les caractéristiques d'un individu (et d'un seul) présent dans ce fichier, on ne pourra pas en conclure à son identification, car :

- on ne sait pas si la personne visée est ou non présente dans le tirage ;
- il n'y a qu'une certaine probabilité (dans l'exemple, une chance sur dix) pour que la personne repérée dans le fichier soit la personne connue.

[7] Dans l'échantillon, si la règle de tirage est secrète et si certaines données sont rendues suffisamment imprécises, on ne pourra jamais affirmer qu'une personne est reconnaissable même si elle présente (âge, sexe, parcours de soins) des caractéristiques uniques.

[8] Un échantillon au 1/10 a été suggéré dès 2006 dans un rapport des Professeurs Bégaud et Costagliola.

Cette remarque doit cependant être tempérée par le fait qu'une accumulation de ressemblances avec une personne connue pourrait, au fil de l'addition des millésimes diffusés, au fur et à mesure de l'enrichissement annuel de la base, conduire à une probabilité d'identification proche de l'unité.

C'est pourquoi, ce type de fichier devrait être

- rendu imprécis dans certaines de ses dimensions (telles que dates et lieu de soins, adresse, âge, date de décès ?) ce qui le rendrait inutilisable pour le suivi précis des parcours de soins mais très utile pour des études de pharmaco-épidémiologie par exemple…
- mis en accès limité, avec identification de la personne qui y a accès et engagement de sa part à n'utiliser le fichier qu'à des fins déclarées à l'avance, lesquelles n'incluent pas la tentative d'identifier une personne (modèle du réseau Quetelet ou de l'accès actuel des chercheurs INSERM aux données de l'EGB).

c) une offre restreinte pour les fichiers permettant une identification indirecte des malades

Comme cela a été vu, certains fichiers, utiles pour la recherche, sont extrêmement détaillés et permettent une identification indirecte de nombreuses personnes.

Les fichiers PMSI et EGB ayant été assez largement diffusés, ils ne devraient dorénavant être disponibles que moyennant des procédures d'accès très strictes.

Les personnes demandant un accès doivent justifier de la raison qui leur fait effectuer cette demande. Elles doivent démontrer les garanties de sérieux pour elles-mêmes et pour leur environnement professionnel. Au moment de leur accès, une trace des opérations effectuées doit être conservée. Enfin, un organisme doit vérifier que les sorties effectuées à partir de ces fichiers ne mettent pas en danger la protection de la vie privée des personnes.

Des systèmes et des procédures existent qui permettent ce type d'accès à des données confidentielles sans que cela implique un site unique.

Mais si ces conditions sont réunies et qu'on admet que le fichier est indirectement[9] nominatif, il n'y a sans doute pas lieu d'appauvrir les données fournies : la distinction actuelle entre les personnes autorisées ou non à croiser les quatre données sensibles (dates de soins, mois de naissance, commune de résidence et date de décès) n'a a priori plus de raison d'être.

Ce système à trois niveaux devrait permettre de présenter une offre aussi vaste que possible, tout en préservant la confidentialité des données.

[9] Bien sûr, puisque les traitements autorisés visent exclusivement à tirer des conclusions générales à partir de données individuelles, il n'y a pas lieu de mentionner dans les données traitées les données directement identifiantes (nom, NIR, adresse etc.).