SOCIETE FRANÇAISE DE STATISTIQUE

Groupe Statistique et société

Séminaire Appariements sécurisés
du 13 novembre 2008

Compte rendu

L'affluence d'une petite centaine de personnes a d'emblée justifié le déplacement de la séance vers le grand amphi LHERMITTE. Ce succès a compensé le handicap induit en 2007 par la grève des transports et est un heureux présage pour la séance prochaine prévue

le lundi 16 novembre 2009.

Nous recueillons dès aujourd'hui vos propositions pour l'ordre du jour de cette prochaine séance.

Les diaporamas de la présente séance sont disponibles sur le site de la SFdS (sfds.asso.fr) à la page du groupe Statistique et Société

En introduction**,** Gilles Trouessin (OPPIDA, gilles.trouessin@oppida.fr) a révoqué les risques d'identification et rappelé les diverses techniques d'appariements sécurisés, des notions au confluent de la statistique et de la sécurité informatique.

Puis, Catherine Quantin (CHU Dijon, catherine.quantin@chu-dijon.fr) et Gilles Trouessin ont présenté les limites de la stratégie actuelle d'appariements sécurisés en France fondée exclusivement sur le hachage des identifiants. Par principe, ne peuvent être réalisés que les appariements prévus dès l'origine du projet. Cette précaution empêche que, même avec toutes autorisations utiles, il n'est pas possible de faire communiquer des études épidémiologiques distinctes.

Les auteurs proposent une solution recourant à une généralisation des tiers de confiance (Gensbittel 2007), comme lorsqu'une enquête épidémiologique exige le retour au patient à la suite du diagnostic d'une pathologie exigeant des soins. Mais la solution proposée ne romprait pas l'anonymat : la procédure consiste à d'abord hacher l'identifiant personnel de manière irréversible, puis pour chaque étude, on chiffrerait cet identifiant haché avec une clé spécifique à cette étude et conservée secrète par une Instance de Coordination des Identifiants. Suite à une autorisation de la CNIL, cette instance pourrait déchiffrer les identifiants spécifiques aux études concernées et apparier les données sans retour à l'identité des personnes.

Cette proposition, prochainement publiée par la Revue d'Epidémiologie et de Santé Publique renvoie à une difficulté que la recherche épidémiologique appelle à résoudre : la base de données de l'assurance maladie, le SNIIR-AM, n'est pas accessible aux épidémiologistes faute d'une coordination entre la procédure FOIN, à la base de l'anonymisation du SNIIR-AM. A moyen terme, il serait souhaitable que le système d'information médical français prenne en compte ce besoin. Cette nouvelle stratégie de protection des données combinant hachage et chiffrement devrait faciliter cette évolution.

David-Olivier Jacquet-Chiffelle (HES bernoise et Université de Lausanne, david-olivier.jaquet-chiffelle@bfh.ch) a présenté l'expérience des données de santé publique sécurisées en Suisse. Elle se situe dans un contexte bien différant du nôtre où longtemps le chiffrement, considéré

comme une arme de guerre, était réservé à la défense nationale et la diplomatie. Le hachage privilégié en France évitait tout risque de transport et déchiffrement d'un message portant atteinte au pays. Ignorant cette angoisse, le contexte suisse a privilégié le chiffrement dès la fin de la décennie 90. Ainsi très tôt l'Office fédéral de statistique a bénéficié d'une base sécurisée de l'ensemble des dépenses de santé abondée par l'ensemble des acteurs suisses de santé publique. Le diaporama de l'auteur, magnifique de pédagogie et de graphisme, suggère aux lecteurs de s'y reporter sans que l'auteur du compte rendu ne déflore cette belle présentation.

Les systèmes d'information belge et néerlandais sur la santé relèvent d'une logique tout autre : le SNIIR-AM, l'ex-projet de Dossier Médical Personnel français, le système d'information sécurisé suisse conduisent à la constitution d'un immense entrepôt de données, lourd à manier devant être impérativement confiné pour des impératifs de sécurité. Tout à l'opposé se présente le système belge présenté par Stefan Verschuere (Vice-Président de la Commission pour la Vie Privée, Bruxelles). Les données de santé relatives à un patient restent stockées où elles ont été produites. Elles sont simplement repérées. L'appariement des données d »'un patient sont effectuées à la demande de traitement et non de façon préalable, en gérant les droits d'accès avec une grande priorité donnée à la recherche épidémiologique. Le degré d'anonymat des données auxquelles accède un épidémiologiste est défini en fonction du besoin de la recherche. Ainsi au lieu de gérer un stock de données comme le SNIIR-AM, le système gère une multitude de petits flux nécessaire à l'appariement ad hoc.

L'exposé de Marcel Goldberg, Marie Zins et Sébastien Bonenfant (INSERM, U687, Marcel.Goldberg@inserm.fr marie.zins@inserm.fr sebastien.bonenfant@inserm.fr) concenr un projet épidémiologique de grande ampleur, la cohorte Constances et la mise au point d'un système sécurisé et général de gestion des cohortes, le projet de Plate-forme PLASTICO en coopération entre l'INSERM et l'assurance maladie (Goldberg, 2007). Il s'agit de suivre 200.000 personnes tant dans leur carrière professionnelle que leur état de santé. Ainsi sont mobilisées un nombre important de grandes sources :

- le Répertoire National Interrégimes de l'assurance maladie, géré par la CNAV ;

- le système national d'Information sur la Gestion des Carrières de la Caisse Nationale d'Assurance Vieillesse, exploitant notamment les Déclarations Annuelles de Données Sociales fournies par les entreprises à l'UNEDIC.

- le SNIIR-AM de l'assurance maladie ;

- les examens de santé des centres de soins pour l'échantillon Constances tiré dans les fichiers de la CNAV.

L'excellent diaporama détaille cette grande mobilisation de données à la fois médicales, médico-administratives et socio-professionnelle. La mise en place d'un dispositif pérenne de gestion des cohortes est une formidable ambition projetant de faire accomplir un saut durable à la recherche épidémiologique française souvent désespérée de pouvoir accéder aux sources extraordinaires dont dispose la France, mais fortement verrouillées, voire dispersées. En demandant à la CNAV de tirer l'échantillon, o l'équipe Constances bénéficie d'un accès (sécurisé) au NIR et donc, parès hachage par FOIN, aux données médico-administratives du SNIRR-AM. Néanmoins, marcel Goldberg estime que cet accès au NIR par la CNAM n'est pas une solution suffisamment générale pour les épidémiologistes et qu'une réflexion est en cours pour leur permettre d'accéder plus facilement aux données sécurisées du SNIIR-Am come ils accèdent aux données des causes de décès de l'Epidc.

Les avancées socio-médicales sont considérables. Qu'en est-il de l'accès des économistes et des démographes aux sources économico-sociales, elles aussi très largement tributaires du NIR et de ses restrictions d'accès ?

Le projet Constances illustre la source exceptionnelle de la CNAV en matière de carrières et son inexploitation par les chercheurs, en dehors de toutes les projections relatives au système des pensions, finalité fondamentale du système.

Nathalie Picard (Université de Cergy), Benoît Riandey et Anne Solaz (Ined) (nathalie.picard@u-cergy.fr riandey@ined.fr solaz@ined.fr) ont évoqué les besoins des économistes en matière d'appariements de données : une demande données longitudinales longues au niveau des couples et concernant des domaines aussi multiples que les aspects démographiques (matrimoniaux, parentaux, résidentiels), économiques (consommations individualisés, carrières, revenus, investissements immobiliers et mobiliers, éducationnels et patrimoniaux avec les donations et les héritages).

Est-il illusoire de rechercher toutes ces données auprès d'une même source ? Le panel PSID de l'université de Michigan, plus que cinquantenaire, montre que ce fut possible ailleurs et depuis longtemps, mais pas chez nous.

Si une source unique ne peut exister le rapprochement de données de sources multiples par appariement, mais aussi par micro-simulation ouvre une voie. Une difficulté vient de la nécessité d'obtenir des données sur les ménages et non sur les individus, alors même que le ménage n'est pas une unité statistique pérenne donc longitudinale.

Stéfan Lollivier, Directeur à l'INSEE (Stefan.lollivier@insee.fr) a d'abord incité les démographes à modé rer leur enthousiasme pour les panels de longue durée dont l'échantillon final risque d'avoir été bien sélectionné par rapport à la population initiale. Il a apporté une réponse solide à cette demande en présentant le projet d'extension de l'Echantillon Démographique Permanent en taille et en contenu : initialement, échantillon au 1/ 100ème (4 jour par an) puisant dans le recensement et l'état civil, l'EDP quadruple sa taille à 16 jours par an, mais multiplie ses sources. Comme le projet Plastico, il puise aux données des DADS, mais aussi à ceux des déclarations fiscales. Le Recensement Rénové de Population a introduit une collecte par sondage qui potentiellement privait l'EDP d'informations censitaires sur 30 % des ménages (60 % de ceux résidant dans les communes de plus de 10.000 habitants, et un peu plus compte tenu des migrations intérieures). Cette lacune se voit résorbée bien au-delà par l'enrichissement de données fiscales d'une double nature, économique en termes de revenus déclarés, démographique en terme de composition du foyer fiscal. Plus généralement, Stéfan Lollivier a défini une structure en quatre silos de données longitudinales disjoints des données de l'EDP, dont le rapprochement ne serait réalisé qu'occasionnellement

Sans répondre à toute la demande des démo-économistes, le projet en développement de Stéfan Lollivier leur apporte beaucoup de grain ; la mobilisation multi-sources des données administratives progresse enfin en France après quelques décades de blocage. C'est l'objectif que la SFdS cherchait à appuyer depuis son premier séminaire sur les appariements sécurisés en 2001. C'est sur ce point de vue optimiste que Michel-Henri Gensbittel (Université Paris 1, SFdS) Michel-Henri.Gensbittel@univ-paris1.fr a suspendu le débat en nous donnant rendez-vous en novembre 2009.