

Inserm



Institut national
de la santé et de la recherche médicale

Proposition pour la création d'une PLATE-FORME scientifique et technique pluri-organismes pour l'aide à la gestion de cohortes et de grandes enquêtes épidémiologiques

Le projet *Plastico*

Rapport pour l'Institut de recherche en santé publique

Marcel Goldberg, Marie Zins, France Lert

Octobre 2007



SOMMAIRE

1	AVANT-PROPOS : CONTEXTE ET LIMITES DE CE RAPPORT.....	3
2	COHORTES ET GRANDES ETUDES EPIDEMIOLOGIQUES	5
2.1	LES COHORTES PROSPECTIVES	5
2.1.1	<i>Des outils épidémiologiques indispensables.....</i>	5
2.1.2	<i>Difficultés rencontrées pour le fonctionnement des cohortes prospectives.....</i>	6
2.2	LES AUTRES TYPES DE GRANDES ENQUETES	7
2.3	LES COÛTS DES GRANDES ENQUETES EPIDEMIOLOGIQUES	7
2.3.1	<i>Études cas-témoins</i>	7
2.3.2	<i>Cohortes prospectives.....</i>	8
3	POURQUOI LA PLATE-FORME <i>PLASTICO</i> ?	9
4	ACTIVITES DE LA PLATE-FORME.....	10
4.1	PRINCIPALES ACTIVITES « NON SPECIFIQUES » DE LA REALISATION DES ENQUETES EPIDEMIOLOGIQUES	10
4.2	ACTIVITES PROPOSEES POUR LA PLATE-FORME <i>PLASTICO</i>	11
4.2.1	<i>Utilisation des bases de données nationales.....</i>	11
4.2.2	<i>Traçage de sujets inclus dans des enquêtes.....</i>	21
4.2.3	<i>Codage de certains types de données.....</i>	22
4.2.4	<i>Saisie automatisée de questionnaires.....</i>	22
5	ORGANISATION ET FONCTIONNEMENT DE LA PLATE-FORME <i>PLASTICO</i>.....	23
5.1	ORGANISMES ET EQUIPES	23
5.2	RESSOURCES A REUNIR	23
5.2.1	<i>Personnel.....</i>	23
5.2.2	<i>Moyens techniques et locaux.....</i>	24
5.3	PRINCIPES DE FONCTIONNEMENT	24
5.4	ASPECTS INSTITUTIONNELS	25
5.4.1	<i>Insertion.....</i>	25
5.4.2	<i>Modalités de partenariat inter-organismes</i>	25
6	PERSPECTIVES	27
7	ANNEXE : LES PREMIERES COHORTES CONCERNEES PAR <i>PLASTICO</i>	28
7.1	GAZEL	28
7.2	CONSTANCES	28
7.3	COSET.....	29

1 AVANT-PROPOS : CONTEXTE ET LIMITES DE CE RAPPORT

Les aspects opérationnels liés à la réalisation d'enquêtes épidémiologiques de grande ampleur en termes d'effectifs, de sources de données et de durée, font appel à des ressources scientifiques, techniques et organisationnelles complexes et de haut niveau de compétence. Ces ressources dépassent largement les moyens disponibles (ou ne peuvent être mobilisées que pendant des périodes de durée limitée) au sein des équipes françaises d'épidémiologistes, quel que soit leur organisme d'appartenance.

Dans le paysage actuel des institutions qui développent des activités en épidémiologie, seule une coopération inter-organismes permettrait de réunir les moyens humains, financiers et organisationnels suffisants pour mettre en place une structure scientifique et technique stable, pouvant jouer le rôle de support à la réalisation de projets épidémiologiques de grande ampleur. Une telle structure inter-organismes est indispensable si on veut que notre pays puisse développer les études épidémiologiques nécessaires pour les besoins de la santé publique, et participer dans de bonnes conditions à la compétition internationale dans le domaine de la recherche épidémiologique, comme le souligne un récent rapport de l'Académie des sciences¹.

C'est dans ce contexte que l'Institut de recherche en santé publique (IReSP) a souhaité une étude de ce que pourrait être une telle structure en termes d'activités et d'organisation, quels services elle pourrait rendre à la communauté épidémiologique française et dans quelles conditions. C'est l'objet du présent rapport, qui propose la création d'une structure *ad hoc*, la « PLate-forme Scientifique et Technique pour l'aide à la gestion de COhortes et de grandes enquêtes épidémiologiques » (projet *PLASTICO*). Cette plate-forme devrait réunir des organismes qui ont des besoins communs en termes de ressources diverses et qui pourraient mutualiser des moyens à cet effet dans le cadre d'un partenariat stable.

À l'occasion de cette étude, il est apparu que plusieurs organismes sont actuellement amenés à mettre en place divers éléments pouvant constituer cette plate-forme. Les organismes et équipes concernés en première ligne à ce stade sont : l'Inserm, dont plusieurs équipes gèrent des cohortes importantes et/ou en développent actuellement de nouvelles, ainsi que des projets de grande ampleur relevant d'autres modèles d'enquête épidémiologique ; l'InVS qui est impliqué dans la mise en place d'une cohorte d'adultes actifs et d'une cohorte d'enfants, préparées en liaison avec plusieurs équipes ; la CNAMTS, concernée à la fois parce qu'elle recueille et gère de nombreuses données nécessaires aux épidémiologistes, et par l'intermédiaire de l'équipe *Risques Postprofessionnels – Cohortes* du Cetaf, qui fait partie de l'Unité mixte Inserm – CNAMTS 687, et qui a la responsabilité de plusieurs cohortes importantes.

On propose donc dans ce rapport la mise en place d'une structure de préfiguration de la plate-forme *Plastico* associant ces trois organismes, d'autres pouvant par la suite participer au fonctionnement de la plate-forme. On détaillera dans ce qui suit les activités et les modalités d'organisation envisagées, en s'appuyant largement sur l'expérience de la cohorte GAZEL (Inserm Unité 687), ainsi que sur les travaux préparatoires en cours pour la mise en place des nouvelles cohortes CONSTANCES et COSET (Inserm Unité 687 – Cetaf et Département Santé Travail de l'InVS).

Il faut d'emblée souligner deux importantes limites de ce rapport :

Parmi l'ensemble des travaux qui relèvent de l'épidémiologie, *Plastico* ne concerne que les enquêtes à l'échelle individuelle, incluant des sujets dans des enquêtes de type essentiellement « cas-témoins » et « cohortes prospectives ». En pratique, il s'agit des enquêtes qui impliquent le recours à des « données à caractère personnel » selon les termes de la loi du 6 août 2004 relative à l'informatique, aux fichiers et aux libertés.

¹ Valleron AJ et al. *Épidémiologie : conditions de son développement, et rôle des mathématiques. Rapport RST, Paris : Académie des sciences (sous presse).*

Plusieurs des prestations susceptibles d'être proposées par la plate-forme *Plastico* correspondent à des procédures qui n'ont pas encore été mises en œuvre en « vraie grandeur » en France, dans le contexte d'études épidémiologiques, ou seulement très partiellement. Il s'agit essentiellement de faire appel de façon extensive à diverses bases de données existantes, comme cela est une pratique ancienne et usuelle dans certains pays à forte tradition épidémiologique, où les organismes de protection sociale ont compris depuis longtemps l'intérêt de l'épidémiologie pour la santé publique et facilitent l'accès à leurs bases de données. Ces procédures font actuellement l'objet d'études exploratoires et de tests divers à partir de données disponibles, notamment dans la cohorte GAZEL et la cohorte AZF. Les travaux correspondants, qui seront pris en charge par la plate-forme, doivent donc encore faire l'objet de diverses mises au point qui sont en cours au moment de la rédaction de ce rapport. C'est une des raisons qui amènent à proposer une première phase de préfiguration avant la mise en place définitive de *Plastico*.

2 COHORTES ET GRANDES ETUDES EPIDEMIOLOGIQUES

2.1 LES COHORTES PROSPECTIVES

2.1.1 DES OUTILS EPIDEMIOLOGIQUES INDISPENSABLES

La cohorte épidémiologique prospective est un type d'enquête dont le principe est le suivi longitudinal, à l'échelle individuelle, d'un groupe de sujets. Ses objectifs peuvent être étiologiques, de description et de surveillance des expositions à des facteurs de risque divers et des pathologies, d'évaluation de l'efficacité d'interventions de nature préventive ou curative. Sur le plan méthodologique, les avantages principaux des cohortes prospectives sont la possibilité d'analyses statistiques longitudinales permettant de tenir compte au mieux de phénomènes liés au temps (séquence temporelle exposition-effet, effet âge, effet génération, effet période). Globalement, les études de cohorte prospective sont celles qui permettent de proposer les meilleures conditions pour juger du rôle sur la santé des facteurs de risque en terme de causalité, tout en permettant de prendre en compte les évolutions temporelles et les interactions entre facteurs.

Les domaines d'utilisation des cohortes prospectives sont aussi diversifiés que l'épidémiologie elle-même, et concernent tous les aspects de la santé en relation avec des facteurs de risque de type varié. Outils de recherche épidémiologique, les cohortes prospectives sont également le support d'activités de surveillance, d'études et de connaissance statistique intéressant de nombreux organismes de santé. On peut aussi établir, bien que les frontières soient largement arbitraires, une distinction entre cohortes prospectives « généralistes » et cohortes prospectives « spécialisées ». Les premières, établies en population générale et souvent de grande taille, se caractérisent par une couverture large de problèmes de santé et de déterminants et une ouverture vers des utilisateurs diversifiés ; ces caractéristiques expliquent que les données recueillies sur les sujets inclus soient généralement relativement superficielles. Les cohortes spécialisées sont centrées sur un problème spécifique (pathologie et/ou groupe de population), les sujets en nombre souvent plus restreint sont habituellement recrutés sur la base de caractéristiques particulières, et les données recueillies sont très détaillées, incluant notamment des investigations biocliniques approfondies.

La littérature épidémiologique internationale a rendu familières certaines cohortes prospectives « historiques », comme celles de *Framingham*², *Whitehall*³, ou l'*Étude Prospective Parisienne*⁴, qui permettent d'aborder de nombreux domaines. En France, on a vu se développer depuis une dizaine d'années de nombreuses cohortes prospectives aux objectifs divers. Une illustration de l'importance prise par les études de cohorte dans le monde français de l'épidémiologie a été l'organisation en 2000 d'un Colloque scientifique de l'ADELF (*Association des Epidémiologistes de Langue Française*) intitulé « *Cohortes épidémiologiques : aspects méthodologiques, éthiques et pratiques* », suivi de la mise en place du « *Club Cohortes* » de l'ADELF, qui réunit régulièrement toutes les équipes spécialisées en France ; un Appel à propositions « Cohortes » de l'Inserm lancé fin 2003 a reçu 116 projets (la moitié pour des cohortes existantes, et la moitié pour des cohortes nouvelles), concernant des domaines très divers de la santé et des facteurs de risque relevant de domaines variés, certaines étant considérées comme des cohortes généralistes, d'autres étant centrées sur des objectifs scientifiques spécifiques.

Les cohortes prospectives françaises se caractérisent cependant par leur taille relativement faible, aucune ne dépassant un petit nombre de dizaines de milliers de sujets (la plus grande cohorte française, l'étude E3N⁵, a inclus 100 000 femmes), alors que certaines cohortes

² Dawber TR, Meadors GF, Moore FEJ. *Epidemiological approaches to heart disease: the Framingham study*. *Am J Pub Health*, 1951, 41: 279-286

³ Marmot MG, Davey Smith G, Stansfeld S et al. *Health inequalities among British civil servants: the Whitehall II study*. *Lancet*, 1991, 337: 1387-92

⁴ Ducimetière P, Richard J, Claude JR et al. *Les cardiopathies ischémiques : incidence et facteurs de risque. L'Étude Prospective Parisienne*. Paris, Éditions Inserm, 1981

⁵ Clavel-Chapelon F and the E3N-EPIC group. *Secular trends of age at menarche and at onset of regular cycling in a large cohort of French women*. *Human Repr* 2002; 17: 228-232.

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*⁶, le projet *UK Biobank*⁷ qui prévoit le suivi prospectif de 500 000 personnes, ou la *General Practice Research Data Base*⁸ qui gère les données de santé de 4 millions de personnes depuis plus de 15 ans. En Norvège, la *Norwegian Mother and Child Cohort Study* est une « cohorte de naissance », qui a inclut 100 000 femmes à la 18^e semaine de grossesse, puis leurs 100 000 nouveau-nés, ainsi que 70 000 pères, soit au total 270 000 personnes⁹. La *Nurses'Health Study* a été mise en place aux États-Unis dès 1976 et assure le suivi prospectif de 122 000 infirmières ; une deuxième vague a été mise en place en 1989 et a inclut 117 000 femmes¹⁰.

Par comparaison avec ces réalisations, la raison essentielle de la relative modestie des cohortes prospectives françaises, outre les problèmes de financement, tient à l'absence en France de dispositifs destinés à surmonter les difficultés techniques et logistiques inhérentes à la gestion des cohortes épidémiologiques longitudinales.

2.1.2 DIFFICULTES RENCONTREES POUR LE FONCTIONNEMENT DES COHORTES PROSPECTIVES

Même si les cohortes prospectives françaises actuelles ne réunissent qu'un effectif de sujets relativement restreint et n'ont pour la plupart qu'un recul encore faible, l'expérience acquise par plusieurs équipes a permis d'acquérir une expertise logistique et scientifique certaine, mais aussi de mieux en comprendre les difficultés. Celles-ci sont essentiellement liées à la taille des échantillons, pouvant atteindre actuellement plusieurs dizaines de milliers de sujets, et à la durée du suivi de ceux-ci, plusieurs cohortes prospectives françaises ayant déjà plus de 15 de suivi longitudinal, voire plus pour l'*Étude Prospective Parisienne* mise en place dès 1967¹¹.

Ces difficultés sont notamment d'ordre organisationnel, technique et méthodologique. On rencontre en effet un certain nombre de problèmes communs à toutes les cohortes longitudinales, qui nécessitent toujours un lourd travail de mobilisation et de coordination sur de longues périodes :

Effectif souvent trop faible, on l'a souligné, même pour celles qu'on considère à l'échelle française comme de « grandes » cohortes, interdisant par manque de puissance certaines analyses détaillées.

Difficultés de suivi : ceci concerne les « perdus de vue » (sujets dont on a perdu toute trace), ainsi que le suivi des événements d'intérêt (facteurs de risque et incidence des problèmes de santé, notamment).

Accès à des sources de données diversifiées, nécessaire pour un suivi large de données de santé, et de facteurs de risque personnels, professionnels, sociaux.

Coûts élevés (personnel et logistique générale : impression et envoi de questionnaires en grand nombre, codage, saisie, contrôle de qualité...), amplifiés par la durée des suivis de cohorte.

⁶ Darling GM, Davis SR, Johns JA. Hormone replacement therapy compared with simvastatin for postmenopausal women with hypercholesterolemia. *N Eng J Med* 1998; 338:64.

⁷ <http://www.ukbiobank.ac.uk/status.htm>.

⁸ Jick H, Jick SS, Derby LE. Validation of information recorded on general practitioner based computerised data resource in the United Kingdom. *Br Med J* (1991); 302: 766-768.

⁹ Stoltenberg C. *The Norwegian Network of Human Research Biobanks and Health Studies*. Norwegian institute of public health, Division of epidemiology, Oslo, Jan 2003.

¹⁰ Zhang SM, Willett WC, Hernan MA, Olek MJ, Ascherio A. Dietary fat in relation to risk of multiple sclerosis among two large cohorts of women. *Am J Epidemiol* 2000; 152: 1056-64.

¹¹ Ducimetière P, Richard J, Claude JR et al. *Les cardiopathies ischémiques : incidence et facteurs de risque*. L'Étude Prospective Parisienne. Paris, Éditions Inserm, 1981.

Implication à long terme des équipes, qui manquent de moyens suffisants et adaptés à la durée des projets, et dont la pérennité n'est souvent pas assurée.

Difficultés pour disposer de personnel spécialisé stable et d'un niveau de qualification suffisant, notamment du fait de l'absence de statut reconnu pour ce type d'activité encore peu développé en France, alors que la durée des suivis de cohorte est incompatible avec un trop fort renouvellement des personnels techniques qui doivent assurer la continuité des procédures et des recueils de données.

En janvier 2001, la DARES et l'Inspection médicale du travail avaient organisé une journée d'étude sur les cohortes épidémiologiques dans le domaine de la santé au travail. Les principales conclusions ont été que le fonctionnement sur le long terme d'une cohorte épidémiologique exige d'importants moyens financiers et surtout logistiques, et qu'il importait donc de « professionnaliser » la gestion des cohortes, en dégagant des moyens spécifiques et stabilisés pour mettre sur pied des dispositifs plus sophistiqués et dotés d'une pérennité suffisante. À l'évidence, ces conclusions s'appliquent à toutes les cohortes prospectives, quel que soit leur domaine d'étude, comme l'ont montré les discussions récentes au sein du Comité de pilotage de l'appel à propositions « Cohortes » de l'Inserm.

2.2 LES AUTRES TYPES DE GRANDES ENQUETES

La plupart des difficultés opérationnelles citées pour la gestion de cohortes prospectives se retrouvent également pour d'autres types d'enquêtes de grande taille. Ainsi, certaines études cas-témoins en population actuelles incluent plusieurs milliers de sujets et le recueil de données se déroule sur plusieurs années, comme c'est le cas de l'étude Icare, qui vise à inclure environ 9 000 sujets en population générale. Elles nécessitent également des procédures de validation de diagnostics, de codage et de saisie parfois très lourdes, ont des coûts élevés et nécessitent un personnel spécialisé stable et qualifié.

2.3 LES COUTS DES GRANDES ENQUETES EPIDEMIOLOGIQUES

Le coût d'une étude épidémiologique peut être extrêmement variable, non seulement évidemment en fonction de l'effectif des sujets inclus et de la durée du suivi pour les études longitudinales, mais aussi selon les investigations menées auprès des sujets, ainsi que des données qui peuvent éventuellement être disponibles par ailleurs. Si on prend pas en compte le coût des examens et des analyses biologiques ou génétiques (IRM, dosage de métabolites sanguins, puces à ADN, etc.), qui sont spécifiques de certaines recherches et qu'on ne peut facilement généraliser, l'essentiel des coûts d'une étude épidémiologique concerne le recueil de données concernant les sujets inclus. On peut donner des ordres de grandeur à partir de quelques exemples.

2.3.1 ÉTUDES CAS-TEMOINS

Le coût d'une étude cas-témoins en population générale est estimé au minimum à 230 € par sujet inclus. A titre d'exemple, dans l'étude ICARE déjà citée¹², qui porte sur les expositions aux facteurs de risque professionnels des cancers du poumon et des voies aérodigestives supérieures et qui inclut 9 000 sujets (6 000 cas et 3 000 témoins), la structure des coûts concerne essentiellement les salaires (épidémiologistes, enquêteurs, personnels réalisant le codage, la saisie et le contrôle de qualité des données, statisticiens, secrétariat) et les frais de fonctionnement (déplacements des enquêteurs, missions de coordination, réunions régulières d'enquêteurs, frais généraux). Ce sont les salaires qui représentent la part de loin la plus importante du coût des études (environ 90 %), notamment les salaires d'enquêteur, car dans une étude en population générale, il faut compter environ une journée de travail d'enquêteur pour l'inclusion d'un sujet (recherche du sujet, contact pour rendez-vous, déplacement, interrogatoire, remplissage des questionnaires et autres recueils de données). Les études cas-témoins en milieu hospitalier sont habituellement moins coûteuses, car les patients peuvent être enquêtés pendant leur séjour à l'hôpital, ce qui réduit sensiblement certains frais.

¹² D. Luce, I. Stucker. *Communication personnelle.*

2.3.2 COHORTES PROSPECTIVES

On peut citer quelques exemples qui donnent des éléments de coût.

La cohorte GAZEL permet le suivi de plus de 20 000 agents EDF GDF depuis maintenant plus de 15 ans à partir de différentes sources déjà existantes (fichiers du personnel, mutuelles, services médicaux de l'entreprise qui ont construit des systèmes d'information épidémiologique, incluant un registre des cancers et un registre des pathologies cardiaques ischémiques) et par un autoquestionnaire postal annuel. Le coût total annuel par sujet (hors personnel permanent) est de 25 Euros environ, constitué essentiellement des salaires des gestionnaires de données, des codeurs et du personnel de saisie, des dépenses d'impression, de frais postaux, de saisie et codage, d'informatique. GAZEL a été conçue comme un laboratoire épidémiologique ouvert à la communauté scientifique et plus de 30 études sont actuellement en cours dans la cohorte ; elles concernent en moyenne 10 000 sujets par étude, soit l'équivalent d'au moins 300 000 sujets inclus, pour lesquels les coûts peuvent être très faibles, car de nombreuses données sont déjà recueillies et la logistique de suivi est prise en charge par le fonctionnement courant de la cohorte.

La cohorte SUVIMAX a assuré le suivi de 15 000 sujets en population générale pendant 8 ans. Le coût total (hors personnel permanent) a été estimé à 10 millions d'Euros. Le coût annuel par sujet inclus est de 83 Euros environ.

Dans l'Enquête Santé et Protection Sociale (IRDES - CNAMTS) le recueil de données pour 10 000 sujets, sous-traité à une société de service, coûte 1,2 million d'Euros par an (hors personnel permanent). Le coût annuel par sujet inclus est de 120 Euros environ.

La cohorte EDEN inclue 2 000 enfants suivis pendant 5 ans, soit 10 000 enfants/années. Le coût total (hors personnel permanent) est estimé à 2,5 millions d'Euros, soit 250 Euros/enfant/an.

En Angleterre, la cohorte WHITEHALL II suit 10 000 fonctionnaires depuis 1985 ; son budget annuel (incluant les salaires de toute l'équipe, y compris celui de l'investigateur principal) est 1,8 millions de livres (soit environ 2,6 millions d'Euros par an). Le coût annuel par sujet inclus est de 260 Euros environ.

On voit que le coût du suivi d'une cohorte prospective peut aisément varier d'un facteur de 1 à 10 selon le contexte. Ceci s'explique par diverses raisons : possibilité ou pas d'utiliser des dispositifs de recueil de données déjà en place (GAZEL bénéficie de données provenant des services médicaux d'EDF-GDF, déjà validées par les épidémiologistes de l'entreprise, des fichiers du personnel incluant l'adresse postale à jour des sujets, des caisses de sécurité sociale, de l'infrastructure des Centres d'examen de santé de la sécurité sociale, etc.) ; type de données recueillies ; salaires des personnels déjà pris ou non en charge par les organismes de recherche.

Au total, si on intègre les salaires des membres permanents des équipes, les coûts essentiels des études de cohorte prospective sont liés au recueil des données dans ses diverses composantes : traçage et contact des sujets, recueil proprement dit, codage, validation, gestion des données. Quand il est possible d'utiliser des sources de données validées préexistantes (cas de GAZEL), le coût est très sensiblement inférieur à celui des suivis de cohortes qui doivent assurer intégralement l'ensemble des procédures de recueil et de suivi des sujets.

3 POURQUOI LA PLATE-FORME *PLASTICO* ?

Le projet de création de la plate-forme *PLASTICO*, ouverte à la communauté scientifique des épidémiologistes, vise à faciliter la réalisation de certaines activités liées à la réalisation de grandes études épidémiologiques. Il est clair que cette plate-forme ne résoudra ni les problèmes de financement, ni ceux liés au statut du personnel nécessaire à de tels projets de longue durée au sein des équipes de recherche. Elle peut cependant apporter une aide importante à leur réalisation.

En effet, les études de cohorte vont continuer de se développer dans la plupart des domaines de l'épidémiologie, et de nouvelles cohortes prospectives sont actuellement en préparation. L'effectif envisagé de certaines de ces cohortes ne se compte plus en dizaines, mais en centaines de milliers de sujets, permettant à la France de se doter d'outils épidémiologiques d'envergure comparable à ce qui existe dans plusieurs pays. Or, le fonctionnement de telles cohortes prospectives va poser les problèmes et les difficultés évoqués ci-dessus, mais amplifiés par l'échelle des projets. Il en est de même pour de très grandes études cas-témoins en population générale.

Il semble donc opportun d'envisager une mutualisation de moyens pour la gestion de cohortes prospectives et d'autres types de grandes enquêtes sous la forme d'une plate-forme scientifique et technique. En effet, malgré la diversité de leurs objectifs scientifiques, des méthodes et des populations observées, la plupart ont de nombreux besoins communs, qu'il s'agisse de la collecte de certains types de données, de leur validation, de leur gestion ou du suivi des sujets inclus. Une telle plate-forme devrait permettre la mise en commun de certaines ressources méthodologiques et d'outils de recueil de données, la mutualisation de compétences de provenance diverse, et favoriser le développement d'activités partagées et de synergies scientifiques.

Répondant à la nécessité de structures pérennes pour des opérations de très longue durée, *Plastico* devrait offrir à la collectivité épidémiologique française des prestations diverses, dans des conditions de fonctionnement apportant des solutions à certaines difficultés évoquées plus haut, tout en induisant d'importantes économies d'échelle.

4 ACTIVITES DE LA PLATE-FORME

Qu'il s'agisse des objectifs scientifiques, de la conception, du suivi opérationnel des travaux, de l'analyse des données recueillies et de la diffusion des résultats, chaque cohorte et chaque enquête épidémiologique a ses particularités et des aspects originaux. On peut cependant identifier certaines tâches qui sont très fréquemment présentes, qu'on peut qualifier de « non spécifiques », et qui mobilisent des ressources qui peuvent être largement partagées. Ce sont ces tâches qui peuvent être prises en charge, au moins partiellement, par une plate-forme qui pourrait proposer certaines prestations de nature scientifique et technique aux équipes concernées.

4.1 PRINCIPALES ACTIVITES « NON SPECIFIQUES » DE LA REALISATION DES ENQUETES EPIDEMIOLOGIQUES

On peut dresser la liste des principales activités de base liées au fonctionnement des cohortes et des grandes enquêtes épidémiologiques. Parmi celles-ci, on examinera celles qui peuvent raisonnablement relever d'une plate-forme inter-organismes.

Développement et maintenance de logiciels de gestion d'enquêtes, nécessairement complexes du fait de la variété des opérations qu'il faut monitorer, du niveau élevé de qualité et de sécurité qu'il faut assurer, et des très lourdes contraintes apportées par le respect de la confidentialité des données individuelles des sujets. Ce dernier aspect est particulièrement fondamental pour les enquêtes longitudinales qui impliquent un recueil de données répété auprès des mêmes sujets, ainsi que dans les études qui font appel à plusieurs sources différentes de données à caractère personnel pour les mêmes sujets.

Accès à des grandes bases de données nationales : diverses bases de données peuvent être sollicitées pour l'inclusion et le suivi des sujets et d'événements d'intérêt, qu'il s'agisse d'événements de santé ou de vie socioprofessionnelle : fichiers de la Cnav, SNIIR-AM, RNIAM, DADS, PMSI, ALD, CépiDc, etc. Les procédures d'accès, de transmission sécurisée, de vérification de cohérence et de complétude, de maintien de l'intégrité des données sont complexes et nécessitent des moyens lourds et des compétences spécialisées.

Appariement de données individuelles en provenance de bases de données nationales : l'intégration de données gérées par des organismes divers concernant les sujets inclus dans des cohortes pose de très importants problèmes techniques et est particulièrement sensible en termes de confidentialité. Cette activité, essentielle dans le suivi à long terme de cohortes qui peuvent être de très grande taille, doit être soigneusement encadrée sur le plan méthodologique et déontologique.

Vérification et validation des diagnostics : de nombreuses cohortes et enquêtes utilisent des données de morbidité extraites de bases de données nationales, comme le PMSI ou les ALD, ou provenant d'auto-questionnaires remplis par les sujets. De telles sources ne permettent habituellement pas directement d'obtenir des diagnostics suffisamment fiables et précis par référence aux contraintes épidémiologiques, et doivent être complétées par des procédures de validation adéquates.

« Traçage » de sujets inclus dans des cohortes prospectives : le traçage de sujets en population ouverte est un problème particulièrement difficile, dont les résultats sont souvent médiocres, générant un nombre de « perdus de vue » qui peut être élevé. Il est cependant souvent possible de retrouver une personne par l'intermédiaire des organismes servant les prestations d'assurance maladie et les prestations sociales auxquels sont rattachés les individus, ou par d'autres sources de traçage.

Codage : la plupart des cohortes et des enquêtes épidémiologiques génèrent une importante activité de codage de données selon des nomenclatures diverses dans le domaine de la santé, du contexte professionnel et social, etc. Cette activité requiert une bonne connaissance des nomenclatures utilisées et une forte expérience pour garantir une qualité suffisante du codage.

Saisie : pratiquement toutes les enquêtes reposent au moins en partie sur l'utilisation de questionnaires ; le volume parfois très important des questionnaires recueillis nécessite des moyens importants pour la saisie, qui peuvent bénéficier d'une large automatisation utilisant des techniques de lecture automatisée de documents (LAD).

Impression, envois en nombre : un volume parfois très important de questionnaires est envoyé régulièrement aux participants des enquêtes ; des lettres d'information, journaux, etc. également. Une logistique adéquate, impliquant des moyens lourds, est mobilisée pour la fabrication, les envois en nombre, le routage, la réception des questionnaires.

Collections biologiques : de nombreuses cohortes et enquêtes cas-témoins incluent un recueil de matériel biologique (sérum, ADN, cellules, tissus divers, etc.), qui posent aux équipes d'épidémiologistes de nombreux problèmes matériels et logistiques pour leur conservation et leur gestion¹³.

4.2 ACTIVITES PROPOSEES POUR LA PLATE-FORME PLASTICO

Toutes les activités listées ci-dessus pourraient largement bénéficier des prestations d'une plate-forme d'aide à la gestion de cohortes prospectives et de grandes enquêtes. Cependant, la « valeur ajoutée » d'une structure scientifique et technique spécialisée en épidémiologie n'est pas identique pour toutes ces activités. En première analyse, on a donc écarté celles qui impliquent des compétences de nature non épidémiologique, ou qui peuvent facilement être sous-traitées à des organismes existants.

Ainsi, il n'a pas semblé réaliste de proposer que *Plastico* prenne en charge le développement et la maintenance de logiciels de gestion d'enquêtes, alors que plusieurs firmes privées proposent de tels logiciels, et que les compétences informatiques requises nécessitent des moyens particulièrement importants pour atteindre le niveau de qualité, de fiabilité requis et l'assurance d'une maintenance à long terme. De même, pour les activités d'impression et d'envois en nombre, il existe de nombreux prestataires privés susceptibles de prendre en charge ces travaux de façon satisfaisante. Par contre, les travaux de saisie automatisée de questionnaires nous semblent relever des activités d'une plate-forme épidémiologique, en raison de la nécessaire proximité qui existe dans de nombreuses situations entre techniciens de la saisie et responsables des études épidémiologiques, impliquant des allers-retours nombreux (mise au point des masques, procédures de contrôle et de validation en cours de saisie, etc.).

La gestion de collections biologiques, quant à elle, requiert des compétences et des équipements spécifiques, et pour la plupart d'une nature très différentes des autres composantes de la plate-forme. C'est pourquoi il ne semble *a priori* pas opportun de proposer que *Plastico* prenne en charge des prestations de gestion de collections biologiques.

Finalement, les principales fonctions de *Plastico* pourraient être les suivantes : accès à des grandes bases de données nationales et appariement de données individuelles ; validation de diagnostics ; traçage de sujets inclus dans des enquêtes ; codage de certains types de données ; saisie automatisée de questionnaires. On précisera pour chaque activité la nature des prestations que *Plastico* devrait offrir aux équipes d'épidémiologie. Rappelons que certaines des fonctions envisagées ne sont, au moment de la rédaction de ce rapport, qu'au stade des études préparatoires.

4.2.1 UTILISATION DES BASES DE DONNEES NATIONALES

La plupart des enquêtes épidémiologiques (études cas-témoins, études de cohorte ou autres) impliquent l'utilisation de données individuelles à des stades divers de leur réalisation : inclusion de sujets présentant certaines caractéristiques d'état de santé, de statut socioéconomique, etc. ; recueil de données diverses à l'inclusion et lors du suivi pour les protocoles longitudinaux ; vérification des diagnostics, etc. Ces opérations suivent des procédures très variables selon les cas : recrutement de sujets dans des structures médicales,

¹³ Cohortes et Banque de Données Biologiques. *Revue d'Epidémiologie et de Santé Publique (Numéro spécial)*, 2003, 51.

tirage au sort dans des listes diverses, examens de santé, accès à des documents médicaux, entretiens d'enquêteurs avec les sujets, etc.

Pour des raisons diverses, on utilise cependant très peu en France les possibilités offertes par les bases de données alimentées par les organismes de protection sociale et médicale, qui offrent pourtant un intérêt potentiel majeur pour la réalisation d'études épidémiologiques à l'échelle individuelle, qu'il s'agisse de l'inclusion et du suivi des sujets, ou de l'accès à des données concernant des événements d'intérêt, de santé ou de vie socioprofessionnelle. L'utilisation à des fins épidémiologiques de ces bases de données offrirait en théorie des avantages très importants, en termes méthodologiques (exhaustivité, absence de certains biais de sélection et d'information), et opérationnels (données déjà recueillies, effectifs immenses, etc.). Bien évidemment, un très important travail épidémiologique est nécessaire pour définir les procédures d'accès, de transmission sécurisée, de vérification de cohérence et de complétude, de maintien de l'intégrité des données. Celles-ci sont complexes et nécessitent des moyens lourds et des compétences spécialisées pour permettre l'utilisation de ces bases de données dans des conditions compatibles avec les contraintes de qualité des études épidémiologiques ; mais leur disponibilité peut faciliter les travaux des épidémiologistes dans des proportions très importantes.

Avant d'envisager les prestations que pourraient fournir *Plastico*, on décrira les principales bases de données nationales françaises concernant des données socioprofessionnelles et des données de santé. Un rapport détaillé, décrivant les bases de données concernées, les données disponibles et les modalités d'accès et de transmission, est en cours de finalisation au moment de la rédaction de ce rapport¹⁴. Ce qui suit en résume les éléments essentiels.

4.2.1.1 Événements socioprofessionnels

Les bases de données

Les bases de données de la Caisse nationale d'assurance vieillesse (Cnav) sont un élément essentiel, à la fois pour l'accès aux données socioprofessionnelles et pour le traçage des sujets. Le rôle de cet organisme est notamment d'assurer le droit au paiement de la retraite pour toute personne ayant appartenu au moins une fois au Régime général de Sécurité sociale (RGSS) durant sa vie. 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 (aux niveaux national, régional et local). La Cnav exerce la mission de collecte, de contrôle et de traitement des données sociales pour l'ensemble de ces partenaires, chacun d'entre eux étant ensuite rendu destinataire des informations qui le concernent.

Pour remplir son rôle, la Cnav a mis en place et gère plusieurs bases de données, qu'on présente succinctement, ainsi que l'origine des données qui les alimentent.

Le SNGI (*Système national de gestion des identités*) qui contient l'ensemble des données (Numéro d'inscription au répertoire (NIR), état civil, statut vital) pour toute personne née en France métropolitaine ou dans les DOM, ainsi que les données d'identification des personnes nées à l'étranger ou dans les TOM et résidant sur le territoire français ; il a pour finalité de certifier l'identité d'une personne. L'INSEE a en charge l'immatriculation de toute personne née en France métropolitaine ou dans les DOM ; ces informations sont contenues dans le RNIPP. Il incombe à la CNAV depuis 1981, dans le cadre de sa mission déléguée par l'INSEE, de procéder à l'immatriculation des ayants droit nés à l'étranger ou dans les TOM et résidant sur le territoire français.

Le SNGI contient l'ensemble de des éléments d'identification des personnes (NIR, nom patronymique, prénom(s), sexe, date et lieu de naissance, date et lieu de décès, numéros d'acte de naissance et d'actes de décès), soit reçus de l'INSEE, soit intégrés par la CNAV elle-

¹⁴ M. Cœuret-Pellicer, C. Ribet, M. Zins. *Communication personnelle*. Le travail correspondant est réalisé en coopération avec le Département Santé Travail de l'InVS.

même. Quotidiennement, l'INSEE et la CNAVTS se transmettent mutuellement une copie de leur fichier.

Le SNGC (*Système national de gestion des carrières*) qui permet de retracer pour chaque individu dès l'âge de 16 ans et jusqu'à la liquidation de ses droits à la retraite, ses différentes périodes d'activité : périodes d'activité professionnelle ou assimilées (chômage, maladie, maternité ou congés parentaux, ...). Le SNGC contient donc l'ensemble des données inhérentes à la carrière des assurés du Régime Général, y compris les données concernant d'éventuelles périodes effectuées dans d'autres régimes de base (MSA, Cancava, Organic), ainsi que dans certains régimes particuliers ou spéciaux (SNCF, EDF-GDF, CNRACL, Mines).

Le SNGD (*Système national de gestion des demandes de retraites en cours d'instruction ou de paiement*).

De plus, pour le compte et sous le contrôle des organismes d'assurance maladie, la Cnav met en œuvre le *Répertoire national inter-régimes des bénéficiaires de l'assurance maladie* (RNIAM), qui est constitué pour chaque bénéficiaire, en plus de son NIR et de son état civil, des informations de rattachement à l'organisme lui servant les prestations d'assurance maladie.

Pour la constitution et l'enrichissement de ces bases de données, la Cnav reçoit régulièrement des données en provenance de différentes sources. Les *Déclarations Annuelles des Données Sociales* (DADS), sont transmises chaque année par les employeurs ayant un numéro SIRET. Les *Données Nominatives Trimestrielles* (DNT) sont transmises par les employeurs de personnel de maison. Les informations de périodes d'activité / non activité des individus relevant de l'UNEDIC (chômage), de la CNAMTS (maladie), de la Cnaf (maternité, ...), des régimes particuliers ou spéciaux (SNCF, EDF, RATP, ...), sont également transmises à la Cnav. Il en est de même pour certains autres régimes, et il est prévu que d'ici quatre à cinq ans la Cnav reçoive les données de tous les régimes. Cet ensemble de données est recueilli de façon prospective depuis 1995. Cependant, les données des autres régimes ne sont actuellement collectées que lorsque les sujets atteignent l'âge de 55 ans : c'est à partir de cet âge que la vérification et le remplissage d'éventuelles absences d'informations débutent.

Après avoir reçu des différents organismes gestionnaires des prestations sociales les informations relatives à l'activité des individus, la Cnav procède à des opérations de consolidation : validation des données ; envoi à chaque partenaire (INSEE, services fiscaux, ...) des données le concernant ; recodage et intégration dans le SNGC de la partie des données nécessaires pour le traitement des retraites ; destruction des données initialement transmises par les différents organismes gestionnaires des prestations sociales.

Intérêt pour l'épidémiologie

Les principales caractéristiques des données issues de la Cnav sont leur exhaustivité et leur qualité : pour des raisons évidentes (elles servent de base au calcul des retraites), ces données sont complètes et particulièrement bien validées, notamment pour les périodes les plus récentes, et leur qualité (complétude et exactitude) s'améliore régulièrement au fil des années avec l'informatisation du recueil à la source.

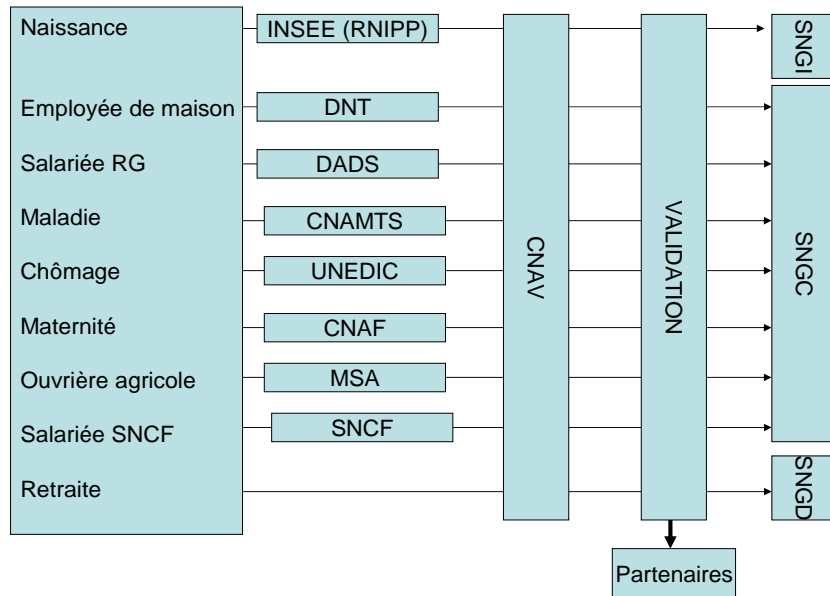
Les bases de données de la Cnav peuvent grandement faciliter des opérations particulièrement lourdes et complexes, dont les résultats sont souvent médiocres, et qui sont courantes dans de nombreuses enquêtes épidémiologiques.

Pour l'essentiel, ces opérations concernent :

le suivi et le traçage des sujets : tous les épisodes socioprofessionnels de la quasi-totalité des personnes vivant en France sont enregistrés de façon prospective et détaillée ; seules les personnes très désocialisées et ne bénéficiant d'aucun salaire et d'aucune prestation sociale échappent à cet enregistrement. Il est donc théoriquement possible de suivre les personnes incluses dans un protocole longitudinal tout au long de leur vie, et de minimiser ainsi les

perdus de vue ; le point spécifique du traçage des adresses des sujets sera développé plus loin.

À titre d'illustration, la figure suivante schématise les flux d'information et les relations entre ces différentes bases de données tout au long de la vie professionnelle d'un sujet fictif, qui aurait commencé à travailler comme employée de maison, puis aurait successivement occupé un emploi salarié, été malade, au chômage, en congé maternité, ouvrière agricole, agent SNCF, puis retraitée.



l'accès aux données socioprofessionnelles : certains domaines de l'épidémiologie, notamment l'épidémiologie sociale et l'épidémiologie des risques professionnels, s'intéressent particulièrement au statut social et professionnel, et à son évolution dans le temps. Les données enregistrées dans les bases de la Cnav sont particulièrement riches de ce point de vue, d'une excellente qualité, et susceptibles d'intéresser différentes équipes d'épidémiologistes, aussi bien pour sélectionner sur des critères socioprofessionnels des sujets à inclure dans des études de méthodologie variée (cas-témoins, cohorte, etc.), que pour avoir accès aux données socioprofessionnelles les concernant tout au long de suivis de longue durée.

4.2.1.2 Événements de santé

Données de mortalité

Le statut vital et les causes de décès peuvent actuellement être obtenus selon la procédure décrite dans le Décret 98-37 autorisant 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). Dans le cadre de cette procédure, le Centre de ressources informatiques (CRI) de l'IFR 69 joue un rôle central, du type « plate-forme » pour cette prestation spécifique d'accès au statut vital et aux causes de décès.

Pour les autres événements de santé, il existe différentes bases de données réunissant des données diverses pouvant être utilisées dans des protocoles épidémiologiques. Les principales bases de données nationales sont décrites brièvement ici.

Données d'hospitalisation : le PMSI

Le PMSI (Programme de Médicalisation du Système d'Information) a pour objectif de produire des informations à contenu médical sur les fonctions hospitalières et de permettre une allocation de ressources dépendante de l'activité hospitalière. Il consiste en un recueil

exhaustif systématique et un traitement automatisé d'informations administratives et médicales. Chaque séjour est ensuite classé dans l'un des 560 GHM (Groupes Homogènes de Malades), économiquement et médicalement considérés comme « homogènes ».

Au sein des établissements hospitaliers, les DIM jouent un rôle central. Le médecin responsable de l'information médicale a un rôle de conseil pour la production des informations et il veille à leur qualité. Les données recueillies sont soumises au secret médical et sont sous la responsabilité du médecin responsable du DIM. Les établissements transmettent trimestriellement (en théorie) les fichiers anonymisés à l'Agence Régionale d'Hospitalisation (ARH), et celles-ci les transmettent à l'Agence Technique de l'Information sur l'Hospitalisation (ATIH), en vue de la constitution des bases de données nationales.

Cette transmission se fait sous la forme de Résumés de sortie anonymisés (RSA), qui contiennent les informations suivantes :

Identification du séjour : Modes d'entrée et de sortie de l'établissement – Nombre d'unités médicales fréquentées – Mois et année de sortie - Durée de séjour de la totalité de l'hospitalisation– Numéro FINESS de l'établissement.

Identification du patient : Sexe - Age en année ou en jours pour les enfants < 1 an - Numéro d'anonymat, construit par l'anonymisation irréversible du numéro de Sécurité sociale, de la date de naissance et du sexe du patient.

Données médicales : Poids de naissance - Diagnostic principal et ensemble des diagnostics associés et des actes pratiqués. Les diagnostics sont codés selon la CIM 10. Jusqu'à une période récente, le codage des actes se faisait selon le Catalogue des actes médicaux (CdAM) ; à partir de juin 2005, une nouvelle classification, la Classification Commune des Actes Médicaux (CCAM), qui harmonise la codification des actes entre médecine de ville et médecine hospitalière, doit être utilisée pour le PMSI.

Les données de l'Assurance Maladie - Échelon loco-régional du RGSS

Parmi les données enregistrées par les systèmes d'information de l'Assurance maladie du RGSS, on distingue les données dites « de production », portant principalement sur les consommations de soins, et dont l'objectif premier est la liquidation des prestations d'assurance maladie, et les données « de référentiels », qui concernent les informations sur les assurés, les établissements de santé et les professionnels de santé. Par ailleurs, les Services Médicaux des Caisses primaires d'Assurance maladie (CPAM) disposent de leurs propres fichiers comportant des informations médicales sur les Affections Longue Durée (ALD), les Accidents du Travail (AT) et les Maladies Professionnelles (MP), et dont l'objectif initial est le contrôle, par les médecins conseil, des pathologies ouvrant droit à une prestation.

Toutes ces données sont rassemblées au niveau des Centres de Traitement Informatique régionaux (CTI), qui jouent ainsi un rôle central dans la gestion des données de l'Assurance maladie. Il existe huit CTI régionaux en France, chacun rassemblant les données d'un groupe de CPAM. Une validation des données est faite à ce niveau. Ces données sont regroupées au sein de deux principales bases : ERASME (données de production essentiellement) et HIPPOCRATE (données médicales).

La base ERASME (*Extraction Recherches Analyses pour un Suivi Médico-Economique*) enregistre les consommations de soins et consommables pharmaceutiques de façon précise (médicaments, actes de biologie), des personnes affiliées au régime général et aux sections locales, incluant l'identification des professionnels de santé (prescripteurs et exécutants) et des établissements sanitaires et sociaux prestataires de soins. ERASME est gérée au niveau des CTI. Il ne s'agit pas d'une base anonyme, les bénéficiaires étant identifiés par le NIR de l'assuré, leur date et leur rang de naissance ; elle contient par ailleurs les nom, prénom, date de naissance, sexe, adresse et qualité des bénéficiaires (assuré, ayant droit conjoint ou enfant).

La base HIPPOCRATE constitue le système d'information du service médical de l'Assurance maladie ; elle est hébergée et administrée par les CTI. Elle enregistre les données médicales (diagnostics codés en CIM10) des patients en ALD, AT et MP. Les ALD, qui concernent les affections susceptibles d'ouvrir droit à une exonération du ticket modérateur sont d'un intérêt particulier pour l'épidémiologie. Il s'agit des affections de la liste ALD 30 (30 affections comportant un traitement prolongé et une thérapeutique particulièrement coûteuse, inscrites sur une liste établie par décret) ; des affections dites « hors liste » (maladies graves de forme évolutive ou invalidante, non inscrites sur la liste des ALD 30, comportant un traitement prolongé d'une durée prévisible supérieure à 6 mois et une thérapeutique particulièrement coûteuse) ; des polyopathologies (patient atteint de plusieurs affections caractérisées entraînant un état pathologique invalidant et nécessitant des soins continus d'une durée prévisible supérieure à 6 mois). Les médecins de l'Échelon local ont accès à l'identité des patients (nom, prénom, adresse) ; les médecins de l'Échelon régional n'ont qu'une version de la base où les patients sont identifiés par un numéro d'anonymat. La liste des ALD 30 est actuellement la suivante :

- Accident vasculaire cérébral invalidant.
- Insuffisances médullaires et autres cytopénies chroniques.
- Artériopathies chroniques avec manifestations ischémiques.
- Bilharziose compliquée.
- Insuffisance cardiaque grave, troubles du rythme graves, cardiopathies valvulaires graves, cardiopathies congénitales graves.
- Maladies chroniques actives du foie et cirrhoses.
- Déficit immunitaire primitif grave nécessitant un traitement prolongé, infection par le VIH.
- Diabète de type 1 et diabète de type 2.
- Formes graves des affections neurologiques et musculaires (dont myopathie), épilepsie grave.
- Hémoglobinopathies, hémolyses, chroniques constitutionnelles et acquises sévères.
- Hémophilies et affections constitutionnelles de l'hémostase graves.
- Hypertension artérielle sévère.
- Maladie coronaire.
- Insuffisance respiratoire chronique grave.
- Maladie d'Alzheimer et autres démences.
- Maladie de Parkinson.
- Maladies métaboliques héréditaires nécessitant un traitement prolongé spécialisé.
- Mucoviscidose.
- Néphropathie chronique grave et syndrome néphrotique primitif.
- Paraplégie.
- Périartérite noueuse, lupus érythémateux aigu disséminé, sclérodermie généralisée évolutive.
- Polyarthrite rhumatoïde évolutive grave.
- Affections psychiatriques de longue durée.
- Rectocolite hémorragique et maladie de Crohn évolutives.
- Sclérose en plaques.

- Scoliose structurale évolutive.
- Spondylarthrite ankylosante grave.
- Suites de transplantation d'organe.
- Tuberculose active, lèpre.
- Tumeur maligne, affection maligne du tissu lymphatique ou hématopoïétique.

Les données de l'Assurance Maladie – Autres régimes

Les autres régimes d'Assurance maladie ont des bases de données qui contiennent pour l'essentiel des données de même nature que le RGSS.

Le Système national d'information inter régimes de l'assurance maladie

L'ensemble des bases de données concernant les événements de santé est désormais réuni au sein du *Système national d'information inter régimes de l'assurance maladie* (SNIIR-AM). Les données du SNIIR-AM incluent tous les régimes de l'assurance maladie : CNAMTS, MSA, CANAM et les 16 autres régimes spéciaux¹⁵, et concernent aussi bien la médecine de ville que les hospitalisations. Les objectifs du SNIIR-AM sont la connaissance des dépenses de l'ensemble des régimes de l'Assurance Maladie, le retour de ces informations auprès des professionnels de santé (informations pertinentes relatives à leur activité, leurs recettes, et s'il y a lieu, à leurs prescriptions), la définition, le suivi et l'évaluation des politiques de santé publique (loi de santé publique du 13 août 2004).

Le SNIIR-AM peut constituer une solution particulièrement efficace pour l'accès à des données individuelles pour les enquêtes épidémiologiques, en évitant le passage par les échelons locaux et régionaux des différents régimes qui rendent complexes et lourdes les procédures d'accès. La base SNIIR-AM est en effet alimentée par les fichiers des bases de données citées ci-dessus ; elle est gérée par le *Centre National de Traitement Informatique* (CENTI) de la CNAMTS.

Le SNIIR-AM est une base de données individuelles mais anonymes qui rassemble les données décrites plus haut : les données de remboursement avec le détail du codage des actes et des médicaments ; les identifiants des professionnels de santé et des établissements de santé qui ont participé aux soins du patient ; les informations sur la pathologie traitée pour les patients en ALD et en AT-MP ; les données issues du PMSI.

Les données individuelles concernant les bénéficiaires sont conservées pendant deux ans au-delà de l'année en cours.

L'anonymisation des variables identifiantes est réalisée par le module FOIN (*Fonction d'Occultation des Informations Nominatives*). Cette fonction repose sur le NIR de l'ouvrant droit, la date de naissance et le sexe du bénéficiaire¹⁶. Les données sont anonymisées en deux étapes : au niveau locorégional (FOIN-1) ; au niveau national (FOIN-2). L'application des algorithmes FOIN construit un identifiant anonyme non réversible : à partir d'un identifiant, on ne peut pas retrouver les données nominatives qui ont servi au calcul.

Intérêt pour l'épidémiologie

Par nature, l'épidémiologie s'intéresse à la santé des sujets inclus dans des études de méthodologie diverse. Les bases de données du PMSI et de l'Assurance maladie ont donc un intérêt *a priori* majeur pour les épidémiologistes. Bien évidemment, elles ne contiennent pas de nombreuses données qui peuvent être essentielles pour des études particulières, mais elles peuvent apporter une aide considérable à la réalisation de très nombreuses enquêtes

¹⁵ Le SNIIR-AM est actuellement en cours de mise en place et ne comporte pas encore la totalité des données décrites ici pour l'ensemble des régimes.

¹⁶ Un fichier de correspondance entre NIR du bénéficiaire et NIR de l'ouvrant droit est géré au niveau locorégional.

épidémiologiques. Pour que cela soit le cas, différentes conditions doivent être réunies. Outre les problèmes techniques et légaux d'accès aux données à caractère personnel et d'appariement de données de sources différentes, sur lesquels on reviendra plus loin, se pose particulièrement le problème de la validité des données de santé de ces bases.

Bien que l'ensemble des bases de données citées n'ait pas fait l'objet d'analyses systématiques de validité, quelques études plus ou moins ponctuelles ont porté sur les données issues des différents fichiers.

L'utilisation du PMSI comme source d'information sur les pathologies s'avère délicate et ne peut reposer uniquement sur le diagnostic principal^{17,18}. Il est nécessaire de développer des algorithmes plus complexes alliant les codes diagnostic aux codes actes spécifiques de la pathologie étudiée. Par ailleurs, lorsque la base du PMSI est utilisée pour estimer l'incidence d'une pathologie, il faut exclure les cas prévalents par la recherche de la pathologie dans les bases PMSI les années antérieures ; cependant, ce problème ne se pose pas lorsque le PMSI est utilisé pour sélectionner des sujets à inclure en fonction d'une pathologie¹⁹, ou pour connaître l'occurrence des pathologies dans le cadre du suivi longitudinal d'une cohorte^{20,21}.

L'intérêt potentiel des bases de données de l'Assurance maladie dans une optique épidémiologique apparaît clairement dans la mesure où elles fournissent des données individuelles médicalisées, structurées et codées de manière standardisée²². Leur utilisation dans une optique épidémiologique nécessite cependant un important travail de réflexion méthodologique, de contrôle et de validation de données.

Ainsi, la base des ALD codés par des médecins reste une base de données à vocation médico-sociale²³, et ses limites sont connues : imprécision des diagnostics, absence d'exhaustivité des cas déclarés, risque de double déclaration²⁴. La prévalence des affections classées en ALD est systématiquement inférieure à la prévalence réelle, car le patient peut être atteint de l'une de ces maladies, mais ne pas répondre aux critères de sévérité ou d'évolutivité exigés, il peut déjà être exonéré du ticket modérateur à un autre titre (autre ALD, invalidité), ou il peut ne pas demander à être exonéré pour des raisons personnelles (assurance complémentaire satisfaisante, souci de confidentialité...). Par ailleurs, la qualité du codage n'a jamais été évaluée, à notre connaissance.

La base de données de remboursements de l'Assurance maladie est adaptée aux objectifs d'analyse des pratiques de prescription²⁵, d'évaluation de l'impact de campagne d'information²⁶. Par contre, elle ne comporte pas d'information sur la nature des maladies traitées, et exclut par définition l'automédication et les prestations non présentées au remboursement.

¹⁷ Couris CM, Forêt Dodelin C, Rabilloud M, Colin C, Bobin JY, Dargent D, Raudran D, Schott AM Sensibilité et spécificité de deux méthodes 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.

¹⁸ Couris CM et al. Method of correction to assess the number of hospitalized incident breast cancer cases based on claims databases. *Journal of Clinical Epidemiology*, 2002, 55 : 386-391.

¹⁹ Geoffroy-Perez B. Confrontation des données du Programme national de surveillance du mésothéliome et du PMSI. Rapport d'étude. Septembre 2004. InVS.

²⁰ Borella L et al. Un essai d'exploitation de la base PMSI nationale pour évaluer le volume et les modes de prise en charge du cancer en secteur hospitalier non lucratif. *Rev Epidemiol Sante Publique*, 2000, 48 : 53-70.

²¹ Laroche ML et al. Qualité des données PMSI au CHU de Limoges : application de la méthode L.Q.A.S. *Rev Epidemiol Sante Publique*, 2002, 50 : 433-439.

²² Fender P, Weill A. Epidémiologie, santé publique et bases de données médico-tarifaire. (Éditorial) *Rev Epidemiol Santé Publique*, 2004, 52, 113-117.

²³ Incidence médico-sociale des ALD30 en 1999. CNAMTS-DSM-Mission des Soins de ville-Mission Statistique. Avril 2004. Disponible sur le site www.ameli.fr/245/doc/1391/article_pdf.html.

²⁴ Chinaud F, Weill A, Ricordeau Ph, Fender P, Allemand H. Incidence du cancer du poumon en France métropolitaine de 1997 à 2002 : les données du régime général de l'assurance maladie. *Revue Médicale de l'Assurance Maladie* Avril-juin 2004, 35 (2), 69-79.

²⁵ Deprez Ph-H, Chinaud F, Clech S, Germanaud J, Weill A, Cornille JL, Fender P. La population traitée par médicaments de la classe des antihistaminiques en France Métropolitaine : données du régime général de l'assurance maladie, 2000. *Revue Médicale de l'Assurance Maladie* Avril-juin 2004, 35 (1), 3-11.

²⁶ Lecadet J, Vialaret K, Vidal P, Baris B, Fender P. Mesure à l'échelle d'une région des effets d'un programme national d'information sur le bon usage des antibiotiques. *Revue Médicale de l'Assurance Maladie* Avril-juin 2004, 35 (2), 81-91.

4.2.1.3 Rôle de Plastico pour l'utilisation des bases de données nationales

Accès aux données individuelles et appariement

Dans une optique épidémiologique, les bases de données citées peuvent l'objet d'utilisations très diversifiées. Cependant, le projet *Plastico* a pour objectif de constituer une plate-forme limitée à l'aide à la réalisation d'enquêtes concernant des individus et impliquant l'utilisation de données à caractère personnel. C'est donc uniquement dans ce contexte qu'on envisagera le rôle de Plastico pour l'utilisation des bases de données nationales.

L'accès à des données individuelles peut concerner deux grands types de procédures épidémiologiques : la sélection de sujets en vue de leur inclusion dans des enquêtes ; l'extraction de données concernant des sujets sélectionnés (y compris par d'autres moyens). On envisagera successivement ces deux aspects.

Sélection de sujets en vue de leur inclusion dans des enquêtes : en théorie, cette opération ne présente pas de difficultés techniques particulières, une fois définis des critères de sélection pertinents correspondant à des données disponibles dans les bases. Deux problèmes peuvent néanmoins se poser : (i) identifier les sujets sélectionnés ; (ii) retrouver les sujets eux-mêmes pour les contacter si le protocole de l'enquête prévoit une participation directe des personnes incluses (entretiens, examens médicaux, etc.) ; on traitera ce point plus loin (« *Traçage de sujets inclus dans des enquêtes* »).

L'identification des sujets sélectionnés est une nécessité si le protocole de l'étude implique l'appariement des données issues d'une des bases concernées avec des données en provenance d'autres sources (questionnaires individuels, autre base de données, etc.), car il faut disposer pour chaque sujet d'un identifiant sans « collision », disponible dans chaque source à appairer (ou créé à partir de données disponibles dans chacune d'elle). Or, le seul identifiant stable qui correspond à ces qualités est le NIR, dont l'utilisation à des fins épidémiologiques est le plus souvent impossible pour des raisons légales.

Les possibilités d'identification dépendent de fait des bases de données, pour lesquelles on trouve des situations différentes de ce point de vue. Ainsi, dans le SNIIR-AM, l'application des algorithmes FOIN construit un identifiant anonyme non réversible (à partir d'un identifiant FOIN, on ne peut pas retrouver l'identité du sujet correspondant) : il est donc impossible d'utiliser le SNIIR-AM pour la fonction de sélection de sujets en vue de leur inclusion dans des enquêtes (sauf si celles-ci ne prévoient que l'utilisation de données incluses dans le SNIIR-AM, à l'exclusion de toute autre source). Sous réserve des autorisations nécessaires, cette opération est cependant techniquement possible pour chacune des bases de données locales ou régionales qui alimentent le SNIIR-AM, ainsi que pour les fichiers des bases de la Cnav, qui disposent de données identifiantes.

Extraction de données concernant des sujets sélectionnés : cette opération ne pose pas en théorie de problèmes techniques si l'on dispose pour chaque sujet des éléments permettant de reconstituer son identifiant dans les différentes bases de données ; il est donc possible de retrouver pour une personne donnée les enregistrements de données le concernant, y compris dans le SNIIR-AM. En pratique, cette fonction implique que dans les protocoles d'enquêtes, on ait prévu le recueil des données individuelles indispensables pour l'appariement avec les identifiants utilisés dans les bases de données sollicitées. Ceci est bien entendu spécifique de chaque enquête, en fonction notamment de la façon dont les sujets ont été inclus et des données disponibles pour ceux-ci. La transmission elle-même des données extraites ne présente pas de difficulté particulière et peut se faire sans difficulté sous une forme anonymisée par des procédures garantissant la confidentialité des données individuelles à caractère personnel, selon des modalités éprouvées depuis longtemps (par exemple, en utilisant un « tiers de confiance » qui génère à chaque transmission une correspondance entre des numéros d'anonymat).

Bien entendu, toutes ces opérations, techniquement réalisables, ne sont possibles que sous réserve que les enquêtes qui en bénéficieraient soient munies préalablement des autorisations légales indispensables²⁷.

Le rôle de la plate-forme *Plastico* concernant l'accès aux données individuelles des bases de données nationales sera de servir d'interface entre les enquêtes épidémiologiques et les bases de données : mise au point avec les responsables concernés des modalités opérationnelles d'accès et de transfert de données ; réception des résultats des opérations informatiques de recherche des sujets et d'extraction des données individuelles ; contrôles de premier niveau (complétude, absence de doublons...) ; transmission sécurisée. La plate-forme pourra prendre en charge cette activité dans des conditions de qualité et de sécurité difficiles à réunir au sein de chaque équipe concernée, en raison des ressources nécessaires et de la compétence et de l'expérience du personnel spécialisé dans ces opérations.

Une autre fonction de *Plastico* devrait être, pour les enquêtes qui font appel à plusieurs sources informatisées de données, d'assurer l'appariement de données individuelles issues de bases de données différentes. En effet, du fait des contraintes très fortes de sécurité et de confidentialité attachées au traitement des données à caractère personnel, et des restrictions pour l'utilisation du NIR, le croisement de données individuelles provenant de plusieurs bases de données est une opération complexe et particulièrement lourde. *Plastico*, structure inter-organismes, qui sera en relation permanente avec les organismes gérant les bases de données nationales, qui devrait disposer de moyens techniques importants et de personnel spécialisé de haut niveau, pourrait, comme cela a été proposé par un récent rapport de l'INSEE²⁸, jouer le rôle d'un « *Centre d'Appariement Sécurisé* ». Les aspects légaux et réglementaires et les modalités de fonctionnement d'une telle fonction devront être définies en relation avec la CNIL.

Validation de diagnostics

L'utilisation des données de morbidité extraites de bases de données nationales, comme le PMSI ou les ALD, ne permet pas d'obtenir des diagnostics suffisamment fiables et précis par référence aux contraintes épidémiologiques, comme on l'a vu plus haut, et dans de nombreuses situations, il est nécessaire de mettre en place des procédures de validation des diagnostics extraits des bases de données. Celles-ci peuvent reposer sur des méthodes très variées : retour au médecin traitant, confrontation avec des questionnaires remplis par les sujets, croisement avec d'autres sources (données de registre, causes de décès...). Ces procédures, fortement consommatrices de ressources en personnel spécialisé, sont prises en charge par les équipes responsables des enquêtes épidémiologiques.

Plastico pourrait remplir une fonction d'aide à la validation pour des diagnostics de pathologie issus des bases de données médicales nationales. Il n'existe cependant pas de méthode « générale » pour un tel objectif. Une voie prometteuse est le développement d'algorithmes alliant un diagnostic à des actes médico-techniques, à des consommations de médicaments plus ou moins spécifiques de la pathologie concernée, etc. Des travaux dans ce sens ont déjà donné des résultats satisfaisants, notamment à partir de diagnostics issus du PMSI ou de consommations de certains médicaments²⁹. L'accès aux bases de données médicalisées citées ici rend possible de telles approches, dans la mesure où elles contiennent des données à la fois sur les diagnostics et sur les actes médico-techniques et les médicaments prescrits.

Bien entendu, de tels algorithmes sont spécifiques des pathologies ; ils doivent de plus être constamment mis à jour en fonction de l'évolution des techniques médicales et de l'introduction de nouveaux médicaments. À l'heure actuelle, on ne dispose de pratiquement aucun algorithme validé qui pourrait être utilisé en routine. *Plastico* serait bien placé, du fait

²⁷ On peut se rapporter à : ADELFI, AEEMA, ADEREST, EPITER. *Recommandations - Déontologie et Bonnes Pratiques en Epidémiologie*. *Rev Epidém et Santé Publ*, 1999, 47 : 1S1-1S32.

²⁸ Chaleix M, Lollivier S. *Outils de suivi des trajectoires des personnes en matière sociale et d'emploi*. INSEE, N° 98/B010, Juin 2004

²⁹ Ils nécessitent cependant une nouvelle validation suite à la mise en place de la classification CCAM.

de la position qu'occupera la plate-forme, pour contribuer à développer, en partenariat avec des équipes d'épidémiologistes spécialisées dans différentes pathologies, un réseau de recherche spécialisé dans le développement d'algorithmes de validation diagnostique, et utiliser ceux-ci au bénéfice d'enquêtes épidémiologiques. De ce point de vue, l'insertion de *Plastico* dans un IFR largement consacré à l'épidémiologie sera un important avantage (cf. plus loin).

4.2.2 TRAÇAGE DE SUJETS INCLUS DANS DES ENQUETES

Qu'il s'agisse de personnes sélectionnées dans des bases de données, ou au cours de suivis longitudinaux, il convient souvent de pouvoir à des fins diverses (envoi de questionnaires, invitation à recevoir un enquêteur ou à se rendre dans une structure d'examen, etc.), contacter les sujets inclus. Or, ceux-ci peuvent déménager, changer d'emploi, sans signaler leur changement d'adresse, et devenir ainsi des sujets non contactables ou des « perdus de vue » des cohortes prospectives, d'autant plus nombreux que le suivi est de longue durée. Ce phénomène est quantitativement important, puisque environ 3 millions de foyers, soit plus de 5,6 millions d'individus, déménagent chaque année en France. Il est donc essentiel de pouvoir retrouver les coordonnées des sujets inclus dans des cohortes longitudinales.

Les méthodes traditionnellement utilisées (recherche par Minitel, contact avec l'entourage, etc.) ne sont pas applicables de façon réaliste à grande échelle. Il faut donc recourir à d'autres moyens, et essayer de retrouver les personnes perdues de vue par l'intermédiaire des organismes servant les prestations d'assurance maladie et les prestations sociales auxquels sont rattachés les individus, ou par d'autres sources de traçage.

Un travail exploratoire a été réalisé, permettant de définir les grandes lignes des procédures à mettre en œuvre. Afin de pouvoir disposer de la mise à jour régulière de l'adresse postale des participants, tout en automatisant fortement les procédures à mettre en œuvre, plusieurs sources sont utilisables. Il s'agit de diverses bases de données alimentées par des organismes qui, pour remplir leur rôle, doivent disposer de l'adresse des personnes. Les organismes concernés sont les suivants :

les Caisses d'allocations familiales (CAF) qui versent les prestations sociales ;

pour les sujets salariés, les Déclarations annuelles des données sociales (DADS) qui contiennent l'adresse connue par les employeurs ;

les organismes servant les prestations d'assurance maladie, *via* le RNIAM et les CPAM de rattachement ;

pour les personnes retraitées, les CRAM qui assurent le versement des pensions de vieillesse.

Chacun de ces organismes enregistre l'adresse des personnes relevant de lui, mais cette adresse n'est toujours mise à jour de façon systématique. Ainsi, si une personne a déménagé et n'a pas effectué son changement d'adresse auprès de sa caisse d'assurance maladie et n'a pas eu de remboursement de soins depuis, sa CPAM ne connaît pas sa nouvelle adresse. Cependant, le recoupement des différentes sources d'informations doit permettre de pallier les limites de chacune d'entre elles.

Il est également possible de retrouver l'adresse des personnes par l'intermédiaire de La Poste. En effet, à l'occasion de leur déménagement, 80 % des personnes s'adressent à La Poste pour faire suivre leur courrier. La Poste détient ainsi leur ancienne et leur nouvelle adresse, qu'elle historise pendant 3 ans glissants dans son « *Fichier des Déménagés* », et elle propose des services issus de ce fichier permettant d'identifier les déménagés et de les suivre à leur nouvelle adresse.

Les procédures permettant de tracer les adresses seront mises au point dans le contexte de l'accès aux bases de données qui ont été décrites plus haut, et des tests seront faits pour optimiser la séquence des opérations de traçage.

4.2.3 CODAGE DE CERTAINS TYPES DE DONNEES

La plupart des enquêtes épidémiologiques s'accompagnent d'une importante activité de codage de données selon des nomenclatures diverses dans le domaine de la santé, du statut socioprofessionnel, etc. Cette activité requiert une bonne connaissance des nomenclatures utilisées et une forte expérience pour garantir une qualité suffisante du codage. Dans certains cas, ce codage peut être au moins partiellement automatisé, comme c'est le cas par exemple pour le statut socioprofessionnel, grâce au logiciel SICORE³⁰, ce qui présente un avantage important en terme de rapidité, de coût et de qualité en supprimant une grande part de la variabilité intra- et inter-codeurs.

Cependant, la mise en œuvre de logiciels de ce type requiert un investissement important et le développement de compétences très spécialisées, qui ne peuvent raisonnablement être réunies dans chaque équipe potentiellement concernée. Il en est de même pour le codage à grande échelle de pathologies, et on sait que la qualité du codage dépend largement de l'expérience des codeurs, qui elle-même est liée au volume des opérations.

C'est pourquoi il semble utile de développer au sein de *Plastico* une fonction de codage, spécialisée dans certains domaines, à laquelle les équipes concernées pourront faire appel.

4.2.4 SAISIE AUTOMATISEE DE QUESTIONNAIRES

De nombreuses enquêtes épidémiologiques reposent au moins en partie sur l'utilisation de questionnaires, qui doivent faire l'objet d'une saisie informatique. Différentes techniques existent pour automatiser cette opération. Ainsi, lorsque le questionnaire est administré par un enquêteur, on peut utiliser la méthode CAPI de saisie en temps réel sur l'ordinateur portable de l'enquêteur. On peut aussi utiliser des techniques de lecture automatisée de documents (LAD).

Cependant, ces techniques sont lourdes : elles nécessitent le développement de masques adaptés, opération spécialisée qui peut être longue pour des questionnaires importants à structure complexe, et pour ce qui concerne la LAD, l'utilisation de coûteux lecteurs optiques et d'une importante chaîne informatique. C'est pourquoi elles ne sont habituellement mises en œuvre que lorsque le volume des données et le nombre des questionnaires recueillis le justifient.

Bien que les opérations de saisie ne soient pas considérées comme particulièrement spécifiques de l'épidémiologie, il semble intéressant de mettre en place un atelier de LAD au sein de la plate-forme *Plastico*. En effet, l'acquisition des appareillages et des logiciels de LAD est coûteuse, et leur mise en œuvre nécessite des compétences techniques particulières qui ne peuvent habituellement pas être réunies dans chaque équipe, ce qui limite de fait l'utilisation de cette méthode pourtant très efficace pour les grandes enquêtes qui nécessitent de nombreux questionnaires ; un atelier de LAD permettrait donc des économies d'échelle qui peuvent être importantes.

Par ailleurs, lorsque les questionnaires sont volumineux et de structure complexe, de nombreuses et fréquentes interactions sont nécessaires entre, d'une part les épidémiologistes qui les ont mis au point et qui sont chargés de la validation des données, et d'autre part les techniciens informatiques qui développent les masques de lecture et les procédures de contrôle et de validation, ainsi que les personnels chargés de la saisie. Un atelier de LAD au sein d'une structure vouée à l'épidémiologie est alors un avantage en termes de productivité et de qualité.

³⁰ Meyer E, Rivière P. SICORE, un outil et une méthode pour le chiffrement automatique à l'INSEE. Direction des Statistiques Démographiques et Sociales, Unité Méthodes Statistiques, INSEE, 2004.

5 ORGANISATION ET FONCTIONNEMENT DE LA PLATE-FORME *PLASTICO*

5.1 ORGANISMES ET EQUIPES

Comme on l'a signalé plus haut, plusieurs organismes mettent actuellement en place les principaux éléments scientifiques et techniques constitutifs d'une plate-forme pour l'aide à la gestion de cohortes et de grandes enquêtes épidémiologiques. Il semble donc réaliste de constituer initialement la plate-forme autour des projets des équipes concernées de ces organismes, pour lesquelles la constitution des éléments de la plate-forme est de toutes façons indispensable pour leurs propres besoins, et qui ont déjà des collaborations actives en cours. Il s'agit, dans ce premier temps, de l'Unité 687, structure mixte entre l'Inserm et la CNAMTS (incluant une équipe de statut Cetaf, organisme dépendant de la CNAMTS, fortement impliquée dans la réalisation d'importantes cohortes prospectives), et du Département Santé Travail de l'InVS. On trouve en annexe une brève description des projets de cohortes prospectives de ces équipes, qui formeront le noyau initial de la plate-forme.

Une fois la plate-forme *Plastico* opérationnelle (et dotée d'un statut stabilisé : cf. plus loin), elle pourra être ouverte à d'autres équipes travaillant dans divers domaines de l'épidémiologie. Pour la plate-forme *Plastico*, une telle ouverture permettra à terme de rentabiliser l'investissement des organismes partenaires du projet sur le plan scientifique, voire financier si les prestations proposées font l'objet d'une rémunération. *Plastico* serait ainsi inscrit dans le paysage de l'épidémiologie et de la santé publique de façon opérationnelle, par des liens de travail étroits avec des équipes travaillant dans des domaines divers de l'épidémiologie. De ce point de vue, l'IReSP devrait contribuer à ouvrir la plate-forme à des équipes diverses appartenant à ses partenaires, comme c'est sa vocation.

Pour les équipes extérieures qui bénéficieraient du support de *Plastico*, une aide considérable pourrait ainsi être mise à disposition, leur permettant ainsi de développer dans de bonnes conditions techniques et matérielles des projets qui ne seraient peut-être pas réalisables de façon autonome, en raison des moyens importants qu'il faut réunir chaque fois sur des durées qui peuvent être très longues. Il faut d'ailleurs souligner que nombre des équipes potentiellement concernées appartiennent ou sont soutenues par les organismes associés au sein de *Plastico*, qui en tireraient alors un retour sur investissement qui peut ne pas être négligeable.

5.2 RESSOURCES A REUNIR

Le volume et la diversité des données que la plate-forme devra gérer, la variété des sources d'information sollicitées, la taille très importante et la très longue durée de certaines des cohortes et autres grandes enquêtes, imposent un type de fonctionnement quasi « industriel », qui implique de réunir des moyens importants, des compétences solides dans des domaines diversifiés et des outils techniques lourds.

Les ressources à réunir, largement à partir des équipes citées plus haut, concernent le personnel, les moyens techniques et les locaux.

5.2.1 PERSONNEL

Comme cela a été rappelé, les enquêtes épidémiologiques de grande envergure nécessitent avant tout avant tout des moyens humains de haut niveau de technicité. C'est donc essentiellement une équipe compétente et expérimentée qu'il faut réunir.

Les qualifications requises sont principalement les suivantes :

- Epidémiologistes
- Statisticiens
- Médecins nosologistes
- Informaticiens spécialistes (gestion des bases de données, gestion des réseaux et de la sécurité informatiques, analyse et programmation)

- Techniciens de saisie et de vidéo codage
- Personnel administratif de soutien

Les effectifs ne peuvent être définis précisément à ce stade, car ils sont largement proportionnels au nombre d'enquêtes qui feront appel à la plate-forme, et de la nature des prestations demandées. Cependant, un minimum d'une douzaine de personnes est nécessaire pour que la plate-forme soit opérationnelle. Il est essentiel que les personnes qui formeront le noyau permanent de la plate-forme soient dotées d'un statut stable.

5.2.2 MOYENS TECHNIQUES ET LOCAUX

Les moyens techniques de la plate-forme sont essentiellement des moyens informatiques et des équipements de LAD.

Hormis des équipements spécialisés de LAD (scanners haut débit), les moyens informatiques (serveurs de bases de données, micro-ordinateurs, systèmes de sauvegarde automatique, etc.) sont standards. Ils requièrent un haut niveau de sécurisation, avec une architecture garantissant l'intégrité des données et l'étanchéité des réseaux internes. Des accès sécurisés aux réseaux de la CNAMTS et de la Cnav sont également indispensables.

Les locaux également doivent être hautement sécurisés. L'atelier de LAD requière des espaces et un aménagement permettant la manutention de grandes quantités de questionnaires.

5.3 PRINCIPES DE FONCTIONNEMENT

Les modalités de fonctionnement d'une plate-forme d'aide à la gestion de cohortes et de grandes enquêtes doivent tenir compte de diverses contraintes, notamment : caractère « sensible » des données susceptibles d'être gérées par la plate-forme ; problèmes de responsabilité scientifique et juridique ; association de plusieurs organismes à la gestion de la plate-forme.

Les points essentiels à examiner concernant les modalités de fonctionnement de *Plastico* sont les suivants :

Modalités d'accès à la plate-forme : qui pourra bénéficier des prestations qui seront proposées ? Selon quelles modalités (financières, responsabilités en termes de qualité, de sécurité, etc.) ?

Propriété des données, propriété intellectuelle, confidentialité : il conviendra de préciser des modalités de fonctionnement de *Plastico* conformes aux textes législatifs et réglementaires en vigueur concernant les données à caractère personnel, et fixer des règles respectueuses de la responsabilité scientifique des équipes faisant appel à la plate-forme (propriété des données, propriété intellectuelle, confidentialité des résultats, etc.).

Une « charte » définissant les règles à observer pour chacun de ces points doit être préparée par les organismes associés à la plate-forme. Concernant les modalités d'accès, elles doivent nécessairement prévoir une structure de nature scientifique chargée d'examiner les demandes en fonction de l'intérêt des enquêtes et de leur qualité (on reviendra plus loin sur les structures de pilotage de *Plastico*). Concernant les aspects financiers, il semble légitime que les demandeurs contribuent aux frais de fonctionnement de la plate-forme (en tenant compte, pour les organismes « fondateurs », de leurs apports récurrents). Cette charte doit par ailleurs être conforme aux Recommandations de Déontologie et de Bonnes pratiques Epidémiologiques³¹.

³¹ ADELFI, AEEMA, ADEREST, EPITER. *Recommandations - Déontologie et Bonnes Pratiques en Epidémiologie*. Rev Epidém et Santé Publ, 1999, 47 : 1S1-1S32.

5.4 ASPECTS INSTITUTIONNELS

5.4.1 *INSERTION*

Plastico associera plusieurs organismes dont les équipes sont susceptibles de contribuer à la plate-forme et/ou de bénéficier des prestations qui seront proposées. Il convient donc d'envisager la structure la plus appropriée pour héberger et gérer la plate-forme.

Une contrainte est la nécessité absolue de l'insertion de la plate-forme dans le milieu de l'épidémiologie, car il n'est pas envisageable de séparer les aspects techniques de la pratique épidémiologique : ils se nourrissent quotidiennement l'un de l'autre, les contraintes méthodologiques des études épidémiologiques appelant des solutions techniques adaptées, souvent originales, et les possibilités techniques de leur côté pouvant influencer les choix méthodologiques des épidémiologistes. Ainsi, les compétences réunies au sein de *Plastico* (connaissance approfondie des bases de données nationales, capacité des systèmes de LAD, des modalités de traçage des sujets, etc.) seront très utiles pour les équipes d'épidémiologie qui développent des enquêtes et qui ne sont pas toujours familières de ces aspects. Symétriquement, les compétences spécifiques dans les domaines de la santé couverts par ces équipes permettront d'améliorer le traitement des données gérées par la plate-forme.

Un autre aspect essentiel est l'insertion de la plate-forme dans le dispositif national de la recherche, de la surveillance et des études en épidémiologie, qui comprend des organismes publics ayant des missions diversifiées. Il faut certainement éviter l'écueil de rattacher cette structure à un organisme ayant une vocation trop limitée.

5.4.2 *MODALITES DE PARTENARIAT INTER-ORGANISMES*

5.4.2.1 Aspects généraux

Les modalités potentielles de partenariat sont nombreuses, allant d'une simple convention de coopération, à une structure plus formelle de type GIS ou GIP. Ces points doivent être examinés du point de vue de la contribution des organismes partenaires aux ressources humaines, matérielles et financières de la plate-forme en fonction des besoins nécessaires au bon fonctionnement de celle-ci, et des modalités de pilotage et de gestion des ressources humaines, matérielles et financières mises en commun.

L'IReSP, dont les membres représentent la grande majorité des acteurs concernés en France, est sans doute, du fait de sa composition et à de ses missions, l'organisme le mieux à même de proposer des modalités de partenariat inter-organismes appropriées.

Il est important de souligner que le projet *Plastico* n'implique ni création d'un organisme nouveau, ni coûts supplémentaires. Au contraire, d'importantes économies d'échelle seront immédiatement obtenues. En termes de faisabilité, l'opération envisagée doit être facilitée par le fait qu'il existe déjà des accords-cadres bilatéraux de coopération scientifique entre tous les partenaires cités : Inserm, InVS et CNAMTS (dont dépend le Cetaf), qui sont par ailleurs des partenaires nationaux de l'IReSP.

5.4.2.2 Structure de préfiguration

Comme on l'a souligné, diverses activités essentielles de la plate-forme sont nouvelles en France, et font actuellement l'objet d'études exploratoires et de tests, sans avoir encore véritablement été mises en œuvre dans des études épidémiologiques dans notre pays. D'autre part, les aspects institutionnels évoqués ci-dessus sont complexes, ainsi que les modalités de gestion qu'il faudra mettre en œuvre ; là aussi, il serait bon d'accumuler une certaine expérience en « vraie grandeur » avant de définir des modalités optimales.

C'est pourquoi on propose dans un premier temps la mise en place d'une structure de préfiguration de la plate-forme *Plastico*. Cette structure de préfiguration temporaire pourrait faire l'objet d'une simple convention de partenariat limitée aux trois organismes envisagés. Son objectif doit explicitement être la mise en place de *Plastico*, sous une forme pérenne. Sa durée devrait être de 2 ans, ce qui permettrait d'accumuler l'expérience suffisante, à la fois sur le plan scientifique, technique et des ressources à réunir, et sur le plan des modalités

institutionnelles et de gestion. Des structures de direction et de pilotage adéquates pourront alors être définies, et la plate-forme configurée selon ses véritables besoins.

6 PERSPECTIVES

Le projet *Plastico* vise à doter la communauté épidémiologique française d'une structure de soutien scientifique et technique pour la réalisation de grandes enquêtes dans certains de leurs aspects les plus lourds qui impliquent une forte logistique, stable et de haut niveau.

Il apparaît que les conditions d'opportunité et de faisabilité d'un tel projet sont actuellement réunies du fait du développement en cours, dans différentes équipes appartenant à plusieurs organismes qui ont déjà établi entre eux des partenariats divers, de projets ambitieux de taille plus importante qu'il n'était usuel jusqu'ici. Ces équipes sont actuellement amenées à mettre en place, pour les besoins de leurs propres projets, les infrastructures scientifiques et techniques nécessaires à la prise en charge de différentes fonctions nécessaires pour la réalisation des enquêtes dont elles ont la responsabilité.

C'est le constat qu'une partie importante de ces fonctions (particulièrement coûteuse et consommatrice de moyens humains et techniques importants) n'est pas spécifique de chaque enquête et peut être réalisée de façon « banalisée », qui ont amené à envisager une mutualisation de ressources sous la forme d'une plate-forme inter-organismes. Il serait en effet particulièrement contreproductif que différentes « micro plates-formes » se mettent en place, chacune disposant de moyens limités ne permettant pas d'atteindre une masse critique de moyens et de compétences suffisante pour développer des activités complexes dont le niveau de qualité doit nécessairement être très élevé. La plate-forme *Plastico* offrirait ainsi l'avantage de permettre d'importantes économies d'échelle pour chaque organisme participant.

Mais *Plastico* doit surtout permettre de réunir, dans le cadre d'une structure stable et pérenne, les compétences scientifiques et techniques de haut niveau susceptibles de répondre à des besoins aujourd'hui mal pris en charge, et dont aucune équipe ne peut actuellement disposer de façon isolée. La situation actuelle limite ainsi fortement l'utilisation pour les enquêtes épidémiologiques des possibilités offertes par les dispositifs gérés par de grands organismes nationaux, qui réunissent des données particulièrement pertinentes pour l'épidémiologie et qui ne sont pas suffisamment exploitées à cette fin, en grande partie du fait de l'absence d'une structure spécifiquement vouée à une telle activité scientifique. *Plastico* a comme objectif, de ce point de vue, de doter la communauté épidémiologique française des moyens dont disposent depuis longtemps les épidémiologistes des pays scandinaves en termes d'accès à des bases de données nationales. Ceux-ci ont permis à ces « petits » pays la mise en place, depuis parfois très longtemps, de dispositifs épidémiologiques d'ampleur bien supérieure à ce qui existe jusqu'à présent en France, occupant ainsi une place prééminente dans la compétition scientifique internationale³².

³² Ainsi, quand on tient compte de la taille de la population des deux pays, la proportion d'articles publiés par le Danemark dans les journaux scientifiques internationaux est environ 15 fois plus élevée que celle de la France (voir : Valleron AJ et al. *Épidémiologie : conditions de son développement, et rôle des mathématiques. Rapport RST, Paris : Académie des sciences - Sous presse*).

7 ANNEXE : LES PREMIERES COHORTES CONCERNEES PAR PLASTICO

Plusieurs projets de mise en place de nouvelles cohortes de « seconde génération » sont en préparation dans divers organismes. Ce sont ces projets qui pourraient être les premiers bénéficiaires de *Plastico*, ceci d'autant plus que divers éléments de la plate-forme sont actuellement en cours de constitution pour la mise en œuvre de certains d'entre eux.

Il s'agit des cohortes *CONSTANCES* (équipe *Risques Postprofessionnels – Cohortes*, Inserm Unité 687, Cetaf) et *COSET* (Département Santé Travail de l'InVS), qui sont à un stade préparatoire avancé d'élaboration des protocoles et de tests préliminaires. Il faut y ajouter la cohorte *GAZEL* (équipe *Risques Postprofessionnels – Cohortes*, Inserm Unité 687, Cetaf), qui est une cohorte ancienne, mais qui est utilisée comme « banc d'essai » de diverses procédures de *CONSTANCES* qui seront intégrées dans la plate-forme (accès aux bases de données nationales, validation de diagnostics, etc.).

On décrira brièvement les principales caractéristiques de ces projets de cohorte³³.

7.1 GAZEL

L'Unité 687 (ex-Unité 88) de l'Inserm a mis en place depuis 1989 un suivi épidémiologique d'une cohorte de 20 624 volontaires (15 010 hommes et 5 614 femmes) composée d'agents d'EDF-GDF âgés de 35 à 50 ans lors du lancement, et qui seront suivis de façon prospective jusqu'à leur décès. Les données qui font l'objet d'un recueil systématique pour toute la cohorte concernent diverses dimensions et sont recueillies auprès de différentes sources : autoquestionnaire annuel (morbidité, modes de vie) ; service du personnel d'EDF GDF (carrière professionnelle) ; Régime particulier de Sécurité Sociale d'EDF GDF (absence pour raisons de santé, Registre des cancers et des cardiopathies ischémiques en activité), médecine du travail (expositions professionnelles, conditions de travail), Caisses Mutuelles Complémentaires et d'Action Sociale (recours aux soins), PMSI (causes d'hospitalisation, en cours de mise au point), Centres d'Examens de Santé (CES) de la Sécurité sociale (bilan de santé, banque de matériel biologique), CépiDc Inserm (causes médicales de décès). Le suivi est particulièrement efficace : au 31/12/2003 (15 premières années de suivi), le nombre de perdus de vue était infime (101 sujets, soit environ 0,5 %). La participation active par autoquestionnaire est particulièrement élevée : au bout de quatorze ans, seuls 4 % des sujets n'ont jamais renvoyé leur questionnaire annuel après avoir participé en 1989. Une trentaine de projets de recherche épidémiologiques portant sur des thèmes très diversifiés ont été mis en place dans cette cohorte par plus d'une vingtaine d'équipes différentes appartenant à des structures diverses, dont une douzaine d'Unités Inserm. Par son ampleur, sa durée, la multiplicité des sources de données et des travaux de recherches qui y sont associés, *GAZEL* permet la réalisation d'études épidémiologiques portant sur des thèmes variés, en offrant aux équipes de recherche françaises et étrangères un accès à des données nombreuses et à une logistique particulièrement complète. Elle peut donc servir, pour divers aspects, comme un « modèle » pour certaines fonctions de *Plastico*, et sont utilisées actuellement pour tester certaines des procédures décrites dans ce rapport.

7.2 CONSTANCES

La cohorte *CONSTANCES* (*CONSULTANTS* des *CES*) a essentiellement un double but : (i) aide à la prévention et à l'évaluation : les Centres d'Examens de Santé (CES) du Régime général de Sécurité sociale orientent prioritairement leurs activités vers des programmes de prévention concernant divers thèmes, dont les risques post-professionnels, la précarité et les inégalités de santé, ainsi que le vieillissement. Dans ce contexte, la cohorte *CONSTANCES* doit fournir un support pour l'aide à ces actions de prévention, et notamment leur évaluation ; (ii) connaissance épidémiologique : il s'agit d'étudier le rôle de très nombreux facteurs personnels

³³ On peut signaler qu'un intérêt pour la plate-forme a également été exprimé par les responsables d'autres cohortes, existantes ou en préparation : cohortes « Enfants » (Inserm Unité 569, InVS-DSE), 3C et EVA (consortium d'Unités Inserm), SUVIMAX (Inserm U557/InVS/INRA).

et professionnels et les interactions entre ces facteurs sur les problèmes de santé, et de permettre le suivi à long terme des sujets présentant des pathologies. *CONSTANCES* constituera un échantillon de structure proportionnelle à celle de la population générale française relevant du Régime général pour les variables d'âge, de sexe, de PCS et de secteur d'activité et de statut d'emploi ; son effectif est fixé à environ 200 000 sujets recrutés sur une période de 5 ans à partir de 2006. Le suivi des sujets repose largement sur l'utilisation des bases de données nationales citées dans ce rapport. Une étude pilote, commune avec la cohorte *COSET*, est en cours à Toulouse et concerne notamment l'exploration de ces grandes bases de données (cf. plus loin).

7.3 COSET

Conformément aux missions du Département Santé Travail de l'InVS, la cohorte *COSET* (Cohorte pour la Surveillance Épidémiologique en milieu de Travail), est conçue comme un outil de base pour la surveillance épidémiologique des risques professionnels. Il s'agit de constituer une cohorte « multirisques et multi-secteurs » constituant un « *Observatoire Epidémiologique* » des risques pour la santé au travail, dont les principaux objectifs sont l'analyse en continu, par un suivi longitudinal à l'échelle individuelle, de l'évolution de paramètres variés (caractéristiques professionnelles et personnelles, état de santé, morbidité et mortalité), permettant notamment des analyses selon les secteurs économiques et le type d'entreprise. *COSET*, dont le protocole est largement harmonisé avec celui de *CONSTANCES*, doit réunir un effectif très important (plusieurs centaines de milliers de sujets), dont une grande partie sera commune avec la cohorte *CONSTANCES* (population active relevant du Régime général de Sécurité sociale), les autres participants (Régime agricole, artisans, professions libérales, notamment), provenant d'autres circuits de recrutement. Comme *CONSTANCES*, le suivi des sujets reposera en grande partie sur des bases de données nationales. En coopération avec l'équipe *Risques Postprofessionnels – Cohortes* (Cetaf - Inserm Unité 687), une étude pilote est en cours avec le Centre d'Examens de Santé de Toulouse dans le cadre du suivi épidémiologique des conséquences de la catastrophe d'AZF.