LA COHORTE CONSTANCES

PROTOCOLE GÉNÉRAL

Décembre 2023

www.constances.fr

UMS 011 « Cohortes épidémiologiques en population »

SIGLES UTILISÉS

ALD : Affection de longue durée
BMB : Banque de matériel biologique
Camieg : Caisse d'assurance maladie des Industries Électrique et Gazière
CCAM : Classification commune des actes médicaux
CépiDC : Centre d'épidémiologie des causes de décès (Inserm)
CES : Centre d'examens de santé de la Sécurité sociale
Cetaf : Centre technique d'appui et de formation des Centres d'examens de santé
CIM : Classification internationale des maladies de l'OMS
CNAM : Caisse nationale d'assurance maladie
Cnav : Caisse nationale d'assurance vieillesse
Cnil : Commission nationale de l'informatique et des libertés
Cnis *:* Conseil national de l'information statistique
***CONSTANCES* :** cohorte des CONSulTANts des CES
COSET : COhorte pour la Surveillance Épidémiologique en milieu de Travail (InVS)
CPAM : Caisse primaire d'assurance maladie
DADS : Déclarations annuelles des données sociales
DST-InVS : Département Santé Travail de l'Santé publique France
EPS : Examen périodique de santé
FOIN : Fonction d'occultation des informations nominatives
HID : enquête Handicap, Incapacités, Dépendances (Insee)
Insee : Institut national de la statistique et des études économiques
Inserm : Institut national de la santé et de la recherche médicale
InVS : Santé publique France
MSA : Mutualité sociale agricole
NAF : Nomenclature d'activités française (Insee)
NIR : Numéro d'inscription au répertoire
PCS : Profession et catégorie socioprofessionnelle (nomenclature Insee)
PMSI : Programme de médicalisation des systèmes d'information des hôpitaux
PND Pli non distribuable
POS : procédures opératoires standardisées de recueil de données dans les CES
RGSS : Régime général de sécurité sociale
RNIAM : Répertoire national inter-régimes des bénéficiaires de l'Assurance maladie (Cnav)
RNIPP : Répertoire national d'identification des personnes physiques (Insee)
RSI : Régime social des indépendants
SLM : Section locale mutualiste
SNGC : Système national de gestion des carrières (Cnav)
SNGI : Système national de gestion des individus (Cnav)
SNDS : Système national d'information inter-régimes de l'Assurance maladie (Cnamts)
TAP : Typologie d'activité professionnelle
TMS : Troubles musculo-squelettiques
U 1018 : Épidémiologie des déterminants professionnels et sociaux de la santé - Centre de recherche en Épidémiologie et Santé des Populations - Unité 1018 Inserm - Université de Versailles-Saint Quentin
Unedic : Union nationale pour l'emploi dans l'industrie et le commerce

SOMMAIRE

1 PRÉPARATION DU PROTOCOLE DE CONSTANCES

L'UMS 011 « Cohortes épidémiologiques en population » est responsable sur le plan scientifique et technique du projet *CONSTANCES*. Du fait de l'étendue des compétences scientifiques nécessaires pour la préparation du protocole, il a été fait appel à de nombreux scientifiques experts des différents aspects traités : épidémiologistes spécialistes des principaux domaines couverts, spécialistes des bases de données de santé, biostatisticiens, gestionnaires de bases de données.

Le protocole scientifique de *CONSTANCES* a été préparé par un Comité de pilotage scientifique auquel ont été associés des groupes de travail. Le tableau ci-dessous donne la liste des personnes ayant participé à l'élaboration du protocole.

Nom	Appartenance	Domaine d'expertise
Berr C	Inserm Unité 888	Vieillissement
Bonenfant S	Équipe Cohortes - Unité 1018	Gestionnaire bases de données
Carton M	Équipe Cohortes - Unité 1018	Risques professionnels
Cœuret-Pellicer M	Équipe Cohortes - Unité 1018	Santé des femmes /base de données Cnamts
Goldberg M	Inserm Unité 1018	Risques professionnels – Inégalités
Guéguen A	Inserm Unité 1018	Biostatistique (longitudinal)
Leclerc A	Inserm Unité 1018	Risques professionnels – Inégalités
Lert F	Inserm Unité 1018	Déterminants sociaux
Luce D	Inserm Unité 1018	Risques professionnels
Ribet C	Équipe Cohortes - Unité 1018	Déterminants sociaux / base de données Cnav
Ringa V	Inserm Unité 822	Santé des femmes
Singh-Manoux A	London University - Unité 1018	Vieillissement - Déterminants sociaux
Zins M	Équipe Cohortes - Unité 1018	Comportements - Vieillissement

Les méthodes d'échantillonnage, d'analyse de la non-participation et de calcul des pondérations ont été préparées par A. Guéguen et R. Sitta (UMS 011), en collaboration avec S. Hallépée (UMS – Insee), L. Bénézet et G. Santin (DST-InVS).

La coordination des groupes de travail pour la réalisation des questionnaires et du bilan « Tests fonctionnels » a été assurée par M. Carton, A. Ozguler, A. Quesnot, C. Ribet et M. Cœuret-Pellicer (UMS 011).

Le questionnaire concernant les expositions professionnelles a été préparé en collaboration avec le Département Santé Travail de l'InVS, aujourd'hui Santé publique France (B. Geoffroy-Perez, C. Cohidon, L. Bénézet, G. Santin), dans le cadre de la collaboration entre les projets *CONSTANCES* et COSET.

L'autoquestionnaire Santé des femmes a été préparé par V. Ringa (Inserm U1018), en collaboration avec N. Bajos (Inserm U1018), P. Dargent (Inserm U149), X. Fritel (CHD Félix Guyon), J. Bouyer (Inserm U1018), , A. Fauconnier (Inserm U1018), G. Plu-Bureau (CHU Necker).

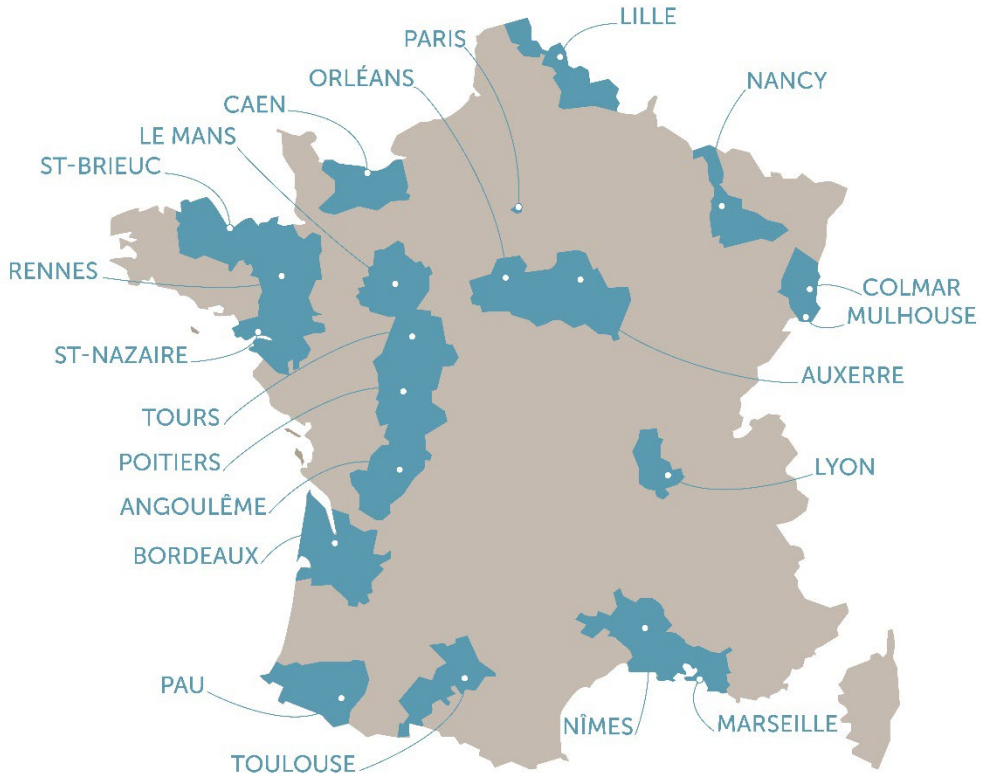
Les protocoles opératoires standardisés des examens paracliniques ont été préparés par A. Brigand (CES Angoulême et équipe Cohortes) et C. Kuntz (Cetaf). Ils ont été expertisés par S. Czernichow (UMR U557 Inserm/Inra/Cnam), C. Delcourt (Inserm U593), N. Roche (Hôtel-Dieu Paris), T. Lang (Inserm U558), E. Couraud (CES de Pau). Les pilotes réalisés dans les CES participants ont été coordonnés par A. Brigand, A. Quesnot, et M. Nachtigal en collaboration avec la société ClinSearch.

Les protocoles opératoires des examens biologiques ont été préparés par JF. Meyer (CES Saint Brieuc) et J. Henny (CES Vandoeuvre les Nancy) avec la participation de l'association Asqualab.

Les circuits et flux de données ont été mis au point par C. Ribet et S. Bonenfant en collaboration avec C. Albert (Cnav), C. Albouy-Cossard, Y. Merlière et L. Duchet (CNAMTS), F. Robergeau (Plateau technique du CESP, Inserm).

Des experts extérieurs ont été également sollicités pour des aspects ponctuels dans leur domaine de spécialité : H. Amieva (Inserm U897), S. Andrieu (Inserm U558), J. Ankri (Inserm U1018), B. Blondel (Inserm U149), A. Colvez (Cetaf), MC. Delmas (InVS-DMCT), JF. Dartigues (Inserm U897), E. de La Rochebrochard (Ined), R. Dray-Spira (Inserm U1018), A. Elbaz (Inserm U708), D. Hassoun (Inserm U1018), F. Kauffmann (Inserm U780), S. Legrain (Gériatrie, CHU Bichat), B. Leynaert (Inserm U700), M. Melchior (Inserm U1018), JJ. Moulin (Cetaf), Y. Roquelaure (CHU Angers), MJ. Saurel (Inserm U149), R. Slama (Inserm U1018), J. Tichet (IRSA, Tours), J. Touchon (CHU de Montpellier et Inserm U888), M. Zureik (Inserm U700).

Avec la participation des CES *CONSTANCES*



2 CONTEXTE – LES COHORTES ÉPIDÉMIOLOGIQUES

2.1 PRINCIPE GÉNÉRAL DES COHORTES ÉPIDÉMIOLOGIQUES

La cohorte épidémiologique est un type d'enquête dont le principe est le suivi longitudinal, à l'échelle individuelle, d'un groupe de sujets. Sa particularité est de pouvoir répondre à des objectifs multiples : description, suivi de l'évolution et surveillance des pathologies et de l'exposition à des facteurs de risque à l'échelle individuelle ; étude des effets de l'exposition à des facteurs de risque sur des problèmes de santé divers ; évaluation de l'efficacité à court, moyen et long terme d'interventions de nature préventive ou réparatrice.

Sur le plan méthodologique, les avantages principaux des cohortes sont la possibilité de tenir compte au mieux de phénomènes liés au temps (séquence temporelle exposition - effet, effet génération, effet période). Il est ainsi possible de modéliser l'enchaînement et les interactions des différents facteurs relatifs aux conditions de vie (alimentation, habitat, accès aux soins, réseau social, ...), à l'environnement (conditions de travail, expositions professionnelles et environnementales, ...), et à l'état de santé (chronologie des phénomènes pathologiques). Par ailleurs, les données d'exposition étant recueillies avant la survenue des effets analysés, on

évite certains biais potentiels des études rétrospectives. Globalement, les études de cohorte sont celles qui permettent de proposer les meilleures conditions pour juger en termes de causalité du rôle sur la santé de facteurs de risque ou d'interventions préventives, en permettant de prendre en compte les évolutions temporelles et les interactions entre facteurs.

Le caractère prospectif des cohortes permet de planifier le recueil de données concernant les expositions à des facteurs de risque nombreux ou à des interventions préventives, et de prendre en compte des problèmes de santé très divers appréhendés en termes de morbidité, voire d'états précliniques, et de mortalité.

Les domaines d'utilisation des cohortes 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 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 « généralistes » et cohortes « 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. Selon la définition de l'ANRS, une cohorte « *doit être conçue pour répondre à plusieurs questions de recherche épidémiologique, clinique, biologique ou de santé publique même si certaines ne sont pas encore formulées de façon précise au démarrage de la cohorte* ». De telles cohortes visent donc à constituer de véritables « plateformes » permettant d'étudier de nombreuses questions.

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. On rencontre également des cohortes « mixtes », où des sous cohortes spécialisées sont « nichées » au sein d'une cohorte généraliste ; des données détaillées sont alors recueillies pour un sous ensemble de sujets, sélectionnés selon des critères variés (caractéristiques personnelles, de santé, d'exposition à un facteur de risque), en complément des données de base communes à tous les sujets.

2.2 LES COHORTES ÉPIDÉMIOLOGIQUES DANS LE MONDE

Il est évidemment impossible de recenser toutes les cohortes existantes à l'échelle internationale. Ainsi, la monographie récemment consacrée par le Centre International de Recherche sur le Cancer aux effets cancérigènes du tabac a dénombré pas moins de 47 cohortes épidémiologiques (dont aucune française), ayant contribué de façon significative à l'établissement des risques de divers cancers dus au tabac, c'est-à-dire ayant un recul et une puissance suffisante pour mettre en évidence de tels risques, y compris pour des cancers peu fréquents (IARC, 2004). On ne décrira donc ici que quelques cohortes particulièrement illustratives des apports de cet outil épidémiologique, et dont certains objectifs se rapprochent de ceux de *CONSTANCES*.

Aux **États-Unis**, certaines grandes cohortes existent depuis des décennies. La plus ancienne et la plus célèbre est sans doute *Framingham*, qui assure depuis sa mise en œuvre le suivi de trois générations d'un village nord-américain de l'État du Massachusetts sur le plan cardiovasculaire (Dawber, 1951 ; Oppenheimer, 2005). Depuis 1948, 5 209 hommes et femmes âgés de 30 à 62 ans à l'inclusion ont été suivis tous les deux ans ; à partir de 1971, leurs enfants (5 124 garçons et filles) ont été suivis tous les quatre ans, et depuis 2002, 3 500 petits enfants sont inclus dans l'étude. Cette cohorte a permis d'identifier pour la première fois et de mesurer la plupart des risques cardiovasculaires reconnus aujourd'hui. La cohorte *Framingham* développe également des recherches sur la démence sénile, les fonctions cognitives et l'ostéoporose, et plus de 1 200 publications scientifiques en sont issues.

La *Nurses' Health Study* a été mise en place aux États-Unis en 1976 chez des femmes âgées de 30 à 55 ans dans le but initial d'étudier les effets à long terme de la contraception orale sur la santé. Cette étude assure le suivi prospectif de 122 000 infirmières par autoquestionnaire. Dès 1989, 33 000 échantillons de sang ont été collectés et conservés en vue d'études ultérieures ; une deuxième vague de collecte de sang s'est effectuée en 2000. La *Nurses' Health Study II* a été mise en place en 1989 dans le but de répondre à des questions soulevées par la première étude et que la méthodologie d'origine n'avait pas permis d'aborder : effets de la contraception orale en tenant compte de l'âge de début, des habitudes alimentaires et d'autres facteurs de risque. Cette deuxième vague a inclus 117 000 femmes âgées de 25 à 42 ans. Les retombées scientifiques concernent de nombreux domaines de la santé, avec plusieurs centaines d'articles publiés dans des revues spécialisées de médecine générale, nutrition, cancérologie, endocrinologie, etc. (Egan *et al.,* 2002).

All of Us est une cohorte mise en place à partir de 2015, encore en construction, visant à recruter un million de participants [N Engl J Med 2019; 381:668-676. DOI: 10.1056/NEJMsr1809937]

En **Grande-Bretagne,** de nombreuses cohortes sont en place parfois depuis très longtemps. Parmi les plus connues, la *British Doctors' Study* a été mise en place dès 1951. Un questionnaire a été envoyé alors à tous les médecins britanniques, et 34 440 hommes et 6 194 femmes ont participé. Des questionnaires de suivi ont été envoyés en 1957, 1966, 1972, 1978 et 1990, et à chaque fois, au moins 94 % des médecins sollicités ont répondu. Des analyses concernant la mortalité par cause ont été publiées après 10, 20 et 40 ans de suivi. Cette étude est certainement celle qui a le plus contribué à la connaissance des différents effets du tabagisme sur la santé, y compris sur l'évolution des risques après cessation du tabagisme (Doll & Hill, 1964, Doll *et al.,* 1994).

La cohorte *Whitehall I* mise en place en 1967-69, incluait un total de 19 018 hommes fonctionnaires britanniques âgés de 40 à 69 ans ; ces sujets ont fait l'objet d'un examen clinique et leur mortalité a été suivie jusqu'en 1987. En 1985, une nouvelle vague incluant 10 308 sujets, également fonctionnaires britanniques, a été lancée (*Whitehall II*) ; ils ont bénéficié d'un examen clinique à l'inclusion, répété régulièrement, et reçoivent également de façon périodique un autoquestionnaire postal concernant de nombreux domaines (Marmot *et al.,* 1991). La cohorte *Whitehall* contribue de façon majeure à la recherche en épidémiologie sociale, et est internationalement considérée comme une des principales sources de connaissance scientifique sur les déterminants sociaux de la santé.

La *General Practitioners Research Database* (GPRD), est une base de données rassemblant depuis 1987 les informations recueillies par un réseau de médecins généralistes du Royaume-Uni (Jick *et al.,* 1991). Elle couvre environ 5 % de la population britannique, soit 3 millions de personnes, ou encore 35 millions de personne-années de suivi longitudinal). Le système est financé et hébergé par l'agence chargée du médicament au Royaume-Uni, et les données recueillies sont accessibles aux chercheurs. Cette base a été par exemple mise à contribution récemment pour apprécier la relation entre l'exposition au vaccin contre l'hépatite B et le risque de sclérose en plaque (Hernan *et al.,* 2004).

On peut aussi citer la *One Million Women Study* (Darling *et al.,* 1998), ou *UK Biobank*, qui assure le suivi prospectif de 500 000 personnes, incluant le recueil et l'analyse de matériel génétique et un important programme d'imagerie (Barbour, 2003, http://www.ukbiobank.ac.uk/status.htm).

En Allemagne la cohorte NAKO a inclut plus de 200 000 volontaires âgés de 19 à 74 ans. Une importante biobanque a été mise en place ainsi qu'un programme d'imagerie[Peters *et al*, 2022].

En **Norvège**, pays de 4,5 millions d'habitants (soit 13 fois moins peuplé que la France), le projet *CONOR* (*Cohort of Norway*) suit environ 200 000 personnes sur le plan de leurs habitudes de vie, pratiquant des examens cliniques et des prélèvements biologiques (Næss *et al*., 2008).

2.3 LES COHORTES ÉPIDÉMIOLOGIQUES EN FRANCE

En France, l'*Étude Prospective Parisienne* mise en place dès 1967 (Ducimetière *et al.,* 1981) a joué un rôle pionnier pour les cohortes épidémiologiques en population ; elle continue, presque 40 ans après, à apporter régulièrement d'importants résultats originaux (voir par exemple Jouven *et al*., 2005). Il a cependant fallu attendre la fin des années 80 pour voir se développer dans notre pays des grandes cohortes épidémiologiques aux objectifs divers, comme ESTEV (Derriennic *et al.,* 1996), SUVIMAX (Hercberg *et al.,* 1995), ou Gazel (Goldberg *et al.,* 2007, Zins *et al.,* 2009), et il existe aujourd'hui de nombreuses cohortes prospectives en population aux objectifs divers, comme NutriNet Santé orientée vers l'étude de la nutrition [https://etude-nutrinet-sante.fr/].

L'exemple de la cohorte Gazel, sur laquelle le projet *CONSTANCES* s'appuie largement, comme on le verra, en donne une illustration. Mise en place en 1989 par l'Unité 88 de l'Inserm, cette cohorte est composée de 20 625 agents d'EDF-GDF volontaires (15 011 hommes et 5 614 femmes), âgés de 35 à 50 ans à l'origine 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 (absences 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), Centres d'Examens de Santé (CES) de la Sécurité sociale (bilan de santé, banque de matériel biologique), CépiDC de l'Inserm (causes médicales de décès).

Le suivi est particulièrement efficace : fin 2008 (20 premières années de suivi), le nombre de perdus de vue était infime (107 sujets, soit environ 0,5 %). La participation active par autoquestionnaire est particulièrement élevée : au bout de plus de 25 ans, environ 75 % des questionnaires sont complétés chaque année, et seuls 2,9 % des sujets n'ont jamais renvoyé leur questionnaire annuel après avoir participé en 1989. Plus d'une quarantaine de projets de recherche épidémiologique portant sur des thèmes très diversifiés ont été mis en place dans cette cohorte par une vingtaine d'équipes françaises et étrangères. Des problèmes de santé aussi différents que la migraine, l'ostéoporose post-ménopausique, la pathologie cardiovasculaire ischémique, la dépression, les troubles musculo-squelettiques, les accidents de circulation font l'objet de projets de recherche au sein de la cohorte. Des facteurs de risque comportementaux (alcool, tabac, par exemple), sociaux (soutien social, soutien familial et professionnel aux malades chroniques), psychologiques, professionnels (gestes professionnels, organisation du travail, facteurs psychosociaux au travail), médicaux (consommations de médicaments et traitements) habituellement en interaction, sont pris en compte par ces recherches (Zins *et al., 2009* ; http://www.gazel.inserm.fr).

3 OBJECTIFS DU PROJET *CONSTANCES*

L'objectif du projet *CONSTANCES* est de mettre en œuvre une importante cohorte épidémiologique destinée à fournir des informations à visée de santé publique et de contribuer au développement de la recherche épidémiologique. Réalisé dans le cadre d'un partenariat avec la Caisse nationale d'assurance maladie (CNAM), le projet concerne la population affiliée au Régime général de sécurité sociale qui couvre environ 85 % de la population française. Le RGSS inclue les Sections locales mutualistes (SLM) ainsi que la Caisse d'assurance maladie des Industries Électrique et Gazière (Camieg).

CONSTANCES constitue une infrastructure de recherche largement accessible à la communauté de la santé publique et de la recherche épidémiologique.

La cohorte *CONSTANCES* s'appuie largement sur deux dispositifs existants à l'échelle nationale : (*i*) les CES présents sur l'ensemble du territoire national, bénéficiant d'un important plateau technique permettant le recueil de données biomédicales dans des conditions particulièrement intéressantes ; (*ii*) les bases de données nationales de l'Assurance maladie

et de la Caisse nationale d'assurance vieillesse, permettant l'accès à des données de santé et socioprofessionnelles de façon longitudinale.

3.1 UN OUTIL POUR LA SANTÉ PUBLIQUE

Dans un contexte où les orientations de la politique de gestion du risque et de la politique santé sont largement renouvelées par les instances les plus élevées de la santé publique de notre pays, il est indispensable pour celles-ci de pouvoir disposer de sources d'informations diversifiées sur la santé de la population, la distribution de facteurs de risque de toute origine, le recours au système de soins et de prévention, la trajectoire médicale, professionnelle et sociale des personnes. De telles informations sont en effet nécessaires pour orienter de façon concrète les politiques de santé définies et les cibler au mieux, et pour en évaluer les résultats à court et long terme.

De nombreux dispositifs de recueil et d'analyse de données pertinentes existent à cet effet afin de renseigner les responsables de la santé publique. Ils ont cependant tous des limites tenant à leur champ de couverture, ainsi qu'à la nature et à la qualité des données. Parmi ces limites, l'insuffisance des études longitudinales de dimension suffisante, couvrant un champ large de la santé et de déterminants, et susceptibles de permettre la compréhension de phénomènes complexes du fait de l'interaction de nombreux facteurs et de l'évolution temporelle des personnes et de l'environnement socio-économique, a souvent été soulignée (Valleron, 2006). Ce manque de systèmes d'observation longitudinaux en France a été également mis en évidence par l'Insee pour l'ensemble du dispositif de suivi des trajectoires des personnes en matière sociale et d'emploi (Chaleix & Lollivier, 2004).

C'est dans ce contexte qu'a été préparé le protocole scientifique d'une cohorte longitudinale représentative des adultes relevant du RGSS, nés entre 1944 et 2000 au moment de l'inclusion, dont les participants seront recrutés au sein des CES. Intitulée *CONSTANCES (**CONS**ul**TAN**ts des **CES***), cette cohorte a été largement conçue comme un outil venant en appui des objectifs de santé publique de la CNAM et de l'État, et de l'évaluation de leur atteinte, par le caractère particulièrement complet du dispositif de suivi et de recueil d'informations très diversifiées, grâce à des méthodes diverses et complémentaires faisant appel à plusieurs sources de données, auprès d'un large échantillon représentatif de la population adulte couverte par le RGSS.

3.2 UN OUTIL POUR LA RECHERCHE ÉPIDÉMIOLOGIQUE

La cohorte *CONSTANCES* doit également contribuer à la production de connaissances sur la santé des populations, à la recherche et à la surveillance épidémiologique. L'opportunité de mettre en place à partir des CES une vaste cohorte dont l'effectif, la qualité et la diversité des données, les modalités de suivi, se comparent aux plus importantes cohortes existant à l'échelle internationale, permet de constituer un puissant outil pour la recherche épidémiologique en France.

Les objectifs scientifiques *CONSTANCES* sont largement centrés sur l'épidémiologie des déterminants professionnels et sociaux de la santé, l'environnement, le vieillissement et la santé des femmes. La cohorte doit également permettre la réalisation de projets concernant des thèmes épidémiologiques variés, grâce à un accès largement ouvert à la communauté des chercheurs en santé. *CONSTANCES* est en effet conçu comme une infrastructure de recherche, à l'instar de la cohorte Gazel, déjà évoquée et sur laquelle les objectifs et le protocole de *CONSTANCES* s'appuient largement. De ce point de vue, *CONSTANCES* devrait apporter une contribution majeure à la recherche en santé publique en France.

3.3 UN OUTIL POUR LA SURVEILLANCE ÉPIDÉMIOLOGIQUE

Une coopération concernant le domaine de la surveillance épidémiologique des risques professionnels a été établie avec Santé publique France autour de la cohorte COSET (***CO**horte pour la **S**urveillance **É**pidémiologique en milieu de **T**ravail*). COSET a pour objectif la surveillance épidémiologique des risques d'origine professionnelle dans la population générale française. *CONSTANCES* n'incluant que des affiliés du RGSS, COSET inclue de son

côté les sujets affiliés aux deux autres principaux régimes de couverture sociale (RSI et MSA). Divers aspects des deux projets se superposant, les protocoles de *CONSTANCES* et de COSET ont été préparés en concertation et prévoient l'inclusion de données identiques, de telle sorte qu'elles puissent être mises en commun pour les analyses épidémiologiques à visée de surveillance qui relèvent de la mission de Santé publique France, que soit le régime de couverture sociale des sujets.

3.4 Orientations générales : une infrastructure de recherche

La cohorte *CONSTANCES*, représentative de la population générale et d'effectif important, se caractérisant par une couverture large de problèmes de santé et de déterminants et une ouverture vers des utilisateurs diversifiés, s'inscrit dans le cadre des cohortes « généralistes ». *CONSTANCES* est une infrastructure de recherche, à l'instar des grands instruments scientifiques (comme un télescope ou un accélérateur de particules, par exemple, ou un laboratoire de génotypage équipé de séquenceurs), qui ne sont pas construits pour répondre à une question spécifique, mais qui sont conçus pour aider à analyser une large gamme de problèmes scientifiques, et qui sont accessibles à la communauté des chercheurs spécialisés.

C'est pourquoi *CONSTANCES* n'a pas d'objectifs spécifiques en termes d'hypothèses concernant des pathologies et/ou des facteurs de risque précis, et aussi pourquoi la durée du projet *CONSTANCES* n'est pas définie : la cohorte qui sera constituée a vocation à faire l'objet d'un suivi longitudinal sans limite de temps, à la fois pour pouvoir étudier les effets de facteurs de risque à très long terme (comme dans les exemples cités ci-dessus des cohortes les plus anciennes, qui continuent de produire des résultats plusieurs décennies après leur mise en place), et pour tenir compte de l'évolution des connaissances et des techniques, qui suscitent constamment de nouvelles questions scientifiques auxquelles *CONSTANCES* permettra d'apporter un éclairage.

3.5 Thématiques scientifiques

Bien que conçue comme une cohorte généraliste à vocation très large, un intérêt particulier lors de la conception de *CONSTANCES* concernait l'étude des déterminants professionnels et sociaux de la santé, le vieillissement et les maladies chroniques, la santé des femmes. Depuis les thèmes couverts se sont élargis, notamment aux effets de l'environnement sur la santé et au recours et aux filières de soins. Ces thèmes constituent des champs de recherche en santé publique et en épidémiologie d'une grande importance, particulièrement actuels et couvrant de nombreux problèmes de santé et des populations diversifiées.

4 MÉTHODES : ÉLÉMENTS ESSENTIELS DU PROTOCOLE

4.1 Phase pilote

Le projet *CONSTANCES* a inclu une phase pilote, afin de préciser divers points concrets destinés à finaliser le protocole opérationnel. Elle s'est décomposée en deux parties principales : pilotes partiels des procédures (questionnaires, examens médicaux et paracliniques, etc.), et test en vraie grandeur portant sur environ 3 500 sujets, depuis le tirage au sort des sujets invités à participer jusqu'au recueil des données pendant environ trois mois dans chacun des sept CES pilotes (Bordeaux, Lille, Pau, Rennes, Saint Brieuc, Toulouse et Tours) et à l'appariement avec les bases de données nationales.

La phase pilote a été réalisée en 2009-2010 pour une campagne de trois mois. Comme prévu, certains aménagements au protocole initial ont été apportés en vue du lancement des inclusions véritables. Le protocole présenté dans ce document tient compte de ces modifications qui sont dans l'ensemble mineures.

4.2 Mise en place et suivi de la cohorte : vue d'ensemble

CONSTANCES est une cohorte épidémiologique prospective, dont la durée de suivi n'est pas déterminée. L'échantillon vise à être représentatif de la population couverte par le RGSS (incluant les principales Sections locales mutualistes et la Camieg) nés entre 1944 et 2000 au moment de l'inclusion ; l'effectif total est de d'environ 220 000 sujets, et sa structure est

proportionnelle à la population du RGSS pour le l'âge, le sexe, le statut d'activité et la PCS à un chiffre. L'inclusion des sujets se fait par courrier (auto-questionnaires remplis à domicile), puis par leur venue dans un CES où le recueil initial des données est complété. La constitution de la cohorte est progressive, sur une période de cinq ans, et implique 21 CES répartis dans toute la France (dont deux à Paris). Le suivi est à la fois actif (autoquestionnaire annuel postal ou par Internet à domicile, retour régulier dans le CES), et passif (extraction de données dans les bases nationales de la Caisse nationale d'assurance vieillesse (Cnav) et de l'Assurance maladie).

Les principales étapes de la mise en place et du suivi de la cohorte, qui sont détaillées dans la suite de ce document, sont brièvement résumées ici.

Sélection des sujets éligibles : Les sujets éligibles sont tirés au sort par sondage stratifié avec probabilités inégales, en surreprésentant les individus ayant une probabilité de non-participation plus forte en fonction des variables : Régime d'affiliation (Régime général, SLM, Camieg), âge, sexe, statut d'activité et PCS à un chiffre. Le tirage au sort est fait dans le Répertoire national inter-régimes des bénéficiaires de l'Assurance maladie (RNIAM) apparié au Système National de Gestion des Carrières (SNGC) de la Cnav.

Invitation à participer : dans un premier courrier, les personnes tirées au sort reçoivent une lettre présentant le projet *CONSTANCES*, ainsi qu'un coupon-réponse.

Inclusion des sujets volontaires dans la cohorte : les personnes ayant donné leur accord pour participer à *CONSTANCES* sont convoquées dans le CES dont elles dépendent par un courrier précisant le jour et lieu de l'examen, accompagné d'auto-questionnaires à compléter à domicile (Modes de vie et Santé, Calendrier professionnel) ; à l'accueil au CES, une brochure d'information qui détaille les procédures mises en place pour l'inclusion et le suivi, ainsi que les droits des personnes, est remise aux volontaires.

Recueil des données à l'inclusion : outre les auto-questionnaires complétés à domicile, les sujets bénéficient d'un Examen de santé périodique (EPS) permettant le recueil de données de santé : examen clinique, analyse de sang, mesure de la tension artérielle, du poids, de la taille et du rapport tour de taille/tour de hanches, électrocardiogramme et spirométrie, examen de la vue et de l'audition, bilan des fonctions physiques et cognitives (à partir de 45 ans), et remplissent des questionnaires complémentaires (questionnaire sur les expositions professionnelles vie entière, autoquestionnaire sur leur santé pour les femmes). Un consentement éclairé est proposé en fin de parcours aux participants, qu'ils doivent signer s'ils acceptent de participer.

Suivi actif : un autoquestionnaire postal est envoyé chaque année au domicile des sujets (questionnaire papier ou Internet au choix du sujet) ; une invitation à venir au CES tous les quatre ans est prévue pour l'ensemble des sujets de la cohorte ; ceux qui ont déménagé hors de leur département d'inclusion entre deux vagues d'invitation peuvent bénéficier d'un examen dans le CES de la nouvelle CPAM dont ils dépendent.

Suivi passif d'événements socioprofessionnels et de données de santé : les principaux événements socioprofessionnels sont régulièrement extraits du Système national de gestion des carrières (SNGC) de la Cnav, qui regroupe des informations issues des Déclarations annuelles des données sociales (DADS), Données nominatives trimestrielles, chômage, absences pour maladie, maternité. Les données de santé sont extraites du Système national des données de santé (SNDS) : données de remboursement des consommations de soins, données issues des services médicaux des caisses (Affections de longue durée (ALD), accidents du travail et maladies professionnelles), PMSI (diagnostic principal et diagnostics associés, actes diagnostiques et thérapeutiques pour chacun des séjours hospitaliers).

4.3 COMPOSITION DE LA COHORTE

4.3.1 Population source et structure de la cohorte

La population source est celle des assurés sociaux du RGSS. Ce régime concerne environ 85 % de la population française. Le RGSS géré par la Caisse nationale d'assurance maladie

(CNAM) concerne les salariés (ainsi que leurs ayants droit) de l'industrie, du commerce, des services et certaines catégories d'emploi assimilées à des salariés.

La diversité des objectifs de *CONSTANCES*, son caractère « généraliste » et ouvert, la variété des utilisateurs potentiels, impliquent que la structure de la cohorte soit suffisamment diversifiée selon les principaux critères démographiques, sociaux et professionnels. On a donc décidé de constituer la cohorte *CONSTANCES* sous la forme d'un échantillon :

- Visant à être représentatif de la population générale française des adultes relevant du RGSS pour les variables : âge, sexe, statut d'activité et PCS à un chiffre ;
- d'effectifs proportionnels à la population générale pour ces mêmes variables.

Un point particulier concerne le choix des classes d'âge. Les objectifs définis s'appliquent à des populations adultes, incluant des personnes âgées. Seuls des sujets dont l'âge minimal est de 18 ans sont donc inclus. Pour les personnes âgées, diverses considérations doivent être prises en compte. Si le choix d'inclure d'emblée des personnes de très grand âge permettrait d'observer dans un délai relativement bref la survenue de divers problèmes de santé spécifiques (démences, notamment), il pose des problèmes pratiques considérables, en raison du nombre relativement restreint de sujets de 75 ans et plus qui fréquentent les CES. On a donc fait le choix d'une limite d'âge supérieure à l'inclusion de 69 ans.

Pour des raisons qui sont détaillées plus loin, l'inclusion des sujets se fait dans 21 CES répartis dans toute la France, et l'échantillon *CONSTANCES* est constitué par sondage stratifié avec probabilités inégales et proportionnelles au volume d'activité des CES participants. De façon pratique, les strates sont constituées par les CES participants, et les bases de sondage sont celles des affiliés des Caisses primaires d'assurance maladie (CPAM) dont relèvent ces CES, des SLM et de la Camieg. Pour les personnes à inviter, on réalise un tirage au sort à probabilités inégales selon les variables : régime d'assurance maladie (Régime général, SLM et Camieg), âge, sexe, statut d'activité et PCS à un chiffre. En effet, comme le montrent les résultats détaillés d'une étude réalisée auprès des consultants des CES, en se contentant de sélectionner les sujets sur ces quelques variables simples, la diversité des caractéristiques personnelles des consultants suffit à garantir une très large palette de métiers, de secteurs d'activité et de statuts d'emploi. Ainsi, on dispose d'une cohorte reflétant la structure de la population générale couverte par le RGSS pour les principales variables démographiques et socioprofessionnelles. On décrira plus loin le protocole d'échantillonnage et les méthodes prévues pour assurer la représentativité de *CONSTANCES* à l'inclusion et au cours du suivi.

4.3.2 Structures-ressource : les Centres d'examens de santé

Chaque bénéficiaire du RGSS est rattaché à l'une des CPAM qui sont réparties sur le territoire, généralement à raison d'une par département, sauf quelques exceptions. Tous les assurés sociaux et leurs ayants droit ont la possibilité de bénéficier d'examens périodiques de santé (EPS), dont le coût est pris en charge par la CPAM de rattachement. La périodicité de cet examen est en général de cinq ans et le contenu fixé par un référentiel national. Ces examens de santé sont réalisés au sein d'environ 110 CES ou antennes de ces CES. En pratique, les assurés reçoivent un courrier de la CPAM ou du CES dont ils dépendent les invitant à venir se présenter au CES pour un examen de santé gratuit.

Le choix des CES pour l'inclusion des sujets de la cohorte pour la réalisation d'examens médicaux standardisés à grande échelle s'explique par leur expérience pour le recrutement d'un grand nombre de personnes, la qualité de leur plateau technique, et leur implantation sur l'ensemble du territoire national. Il n'existe aucune structure équivalente en France, ce qui explique que diverses études à visée nationale se sont déjà réalisées dans ce cadre : enquête de prévalence des marqueurs du virus de l'hépatite B et de l'hépatite C (Santé publique France, 2005), Enquête Nationale Nutrition Santé (Santé publique France, 2006), étude ESTEBAN (Santé publique France,2014-2015), etc.

Une étude détaillée de la population des consultants des CES a été réalisée en 2002 afin de décrire avec précision les principales caractéristiques sociodémographiques et professionnelles des consultants habituels des CES, et de comparer cette population à celle

de la population générale. En 2002, la grande majorité des CES invitaient les consultants à partir de mailings adressés par les CPAM et le taux de venue s'élève à environ 10 %. L'étude a reposé sur une enquête d'une semaine dans tous les CES auprès de l'ensemble des sujets âgés de 18 ans ou plus. Les principales données recueillies auprès de l'échantillon (10 260 sujets) sont les suivantes : caractéristiques démographiques, région de résidence, diplômes, activité actuelle, temps de travail, type d'activité, profession et catégorie sociale (PCS), activité ou fonction principale, statut de l'emploi, ancienneté dans l'entreprise, position professionnelle, activité économique principale de l'employeur (NAF), nombre de salariés dans l'établissement. Les résultats montraient que les consultants habituels constituent une population d'une grande diversité d'âge, de genre, de groupes professionnels et de secteurs d'activité couverts par le RGSS, incluant des retraités, des personnes en situation de précarité professionnelle ou sociale, etc. Les effectifs sont importants, quels que soient les critères considérés. Les CES accueillent donc une population adaptée aux objectifs de *CONSTANCES*.

4.3.3 Effectif – Puissance

CONSTANCES doit nécessairement être de vaste dimension, pour pouvoir répondre aux multiples questions qui sont posées dans des domaines très variés. Mais cette variété même ne permet pas de procéder aux classiques calculs de puissance et d'effectifs nécessaires, qui précèdent la mise en place d'études aux objectifs spécifiques.

Il est cependant important de tenter d'évaluer le potentiel de *CONSTANCES* en termes de capacité à mener des études épidémiologiques susceptibles de bénéficier d'une bonne puissance. Pour cela on a estimé le nombre d'événements de santé attendus dans la cohorte *CONSTANCES* à plus ou moins long terme, sous l'hypothèse d'une cohorte ayant une structure d'âge et de sexe identique à celle de la population générale française de 18 à 69 ans au recensement de 1999. Les calculs ont été faits pour un effectif de 200 000 sujets. On a calculé à terme de 5, 10 et 15 ans, le nombre attendu d'événements pour les problèmes pour lesquels on dispose de données nationales de référence fiables : décès, incidence des cancers, des cardiopathies ischémiques et de la maladie d'Alzheimer ; on a fait l'hypothèse que les taux étaient identiques à ceux de la population générale française, et stables sur toutes les périodes envisagées. Les résultats sont présentés dans le tableau suivant pour une cohorte de 200 000 personnes de même distribution d'âge et sexe que *CONSTANCES*.

Nombre attendu d'événements de santé

(Taux de référence : CépiDC, 1999 ; Remontet *et al.*, 2002 ; ARME, 2007)

	Suivi 5 ans			Suivi 10 ans			Suivi 15 ans		
	H	F	Total	H	F	Total	H	F	Total
Décès toutes causes	4 131	2 133	6 264	9 727	5 502	15 229	16 983	10 736	27 719
Cancers incidents	3 162	2 220	5 381	7 036	4 855	11 892	11 444	7 823	19 267
Cardiopathies ischémiques (35-64 ans)1	681	138	819	1 418	290	1 708	2 178	452	2 630
Maladies d'Alzheimer[2]	265	240	505	793	1 007	1 800	1 548	2 469	4 018

[1] À partir des estimations issues de MONICA

[2] À partir des estimations issues de Paquid

On constate que le nombre de ces événements graves est élevé et permettra de nombreuses études dotées d'une bonne puissance. Des simulations ont montré que pour les maladies dont le nombre de cas est compris entre 500 et 5 000 il est possible de détecter des odds ratios compris entre 1,1 et 1,4 pour des comparaisons entre les quartiles supérieur et inférieur d'une variable quantitative avec une puissance de 0,8 et un niveau de significativité de 0,05. Il est donc possible de réaliser de nombreuses études pour les problèmes fréquents avec une puissance satisfaisante. Pour des problèmes moins fréquents, la puissance est souvent insuffisante : c'est une des raisons des collaborations mises en place au sein de consortiums de cohortes en Europe, afin de pouvoir partager des données.

Un autre problème important conditionnant la puissance d'une étude épidémiologique longitudinale est celui des « perdus de vue ». En effet, les calculs précédents reposent sur

l'hypothèse que la totalité de la cohorte sera suivie pendant les périodes envisagées. Or, on sait qu'il existe inévitablement tout au long du suivi d'une cohorte un certain taux d'attrition, du fait des perdus de vue, qui est susceptible d'induire des biais divers, et d'affaiblir la puissance si son ampleur est élevée (Goldberg & Luce, 2001). Il est donc important de chercher à minimiser le nombre des perdus de vue de la cohorte tout au long d'un suivi qui doit être de très longue durée. Il est impossible d'estimer précisément le nombre de perdus de vue dans *CONSTANCES* à long terme. Actuellement (2023), le taux de participation aux questionnaires annuels est d'environ 60 à 70% selon les années, et seulement 0,5 % des participants ont souhaité quitter la cohorte. Quant au suivi passif, le dispositif d'appariement aux bases administratives nationales garantit que le nombre de véritables perdus de vue devrait être très faible.

En conclusion, il apparaît donc qu'on observera rapidement des nombres de cas importants pour les événements de santé les moins rares, et que la proportion de perdus de vue devrait être faible. Ainsi, de nombreuses études sont possibles, dans des conditions de puissance satisfaisante, pour peu qu'elles concernent des événements de santé et des facteurs de risque relativement fréquents, ce qui sera souvent le cas dans les domaines d'intérêt principal.

Par ailleurs, nous avons établi des collaborations étroites avec d'autres cohortes, notamment avec les cohortes d'affiliés à la Mutualité sociale agricole et au Régime social des indépendants dans le cadre du projet COSET coordonné par Santé publique France, qui ont inclut au total environ 60 000 personnes. Ces cohortes portent sur les mêmes classes d'âge que *CONSTANCES* et de nombreuses données comparables sont recueillies : il est ainsi possible, grâce à la mise en commun de données, d'améliorer la puissance de diverses études épidémiologiques. *CONSTANCES* fait également partie de consortiums associant les principales cohortes en population européennes, dont le but est de faciliter l'accès transnational à de grandes cohortes prospectives.

4.3.4 Durée du suivi et renouvellement de la cohorte

Du fait des objectifs très larges de la cohorte, la durée de suivi doit être la plus longue possible, car le suivi longitudinal prend d'autant plus d'intérêt que le recul est important. La durée est donc indéfinie.

Dans la mesure où la population incluse à l'origine vieillira avec le suivi de la cohorte, il sera envisagé de renouveler les participants, afin de conserver avec le temps une structure sociodémographique restant comparable à celle de la population source.

4.4 MODALITÉS D'INCLUSION

4.4.1 CES participants

La centaine de CES et antennes qui existent actuellement sont de taille très variable, disposent de ressources humaines et techniques également diversifiées et se caractérisent par une grande hétérogénéité des pratiques. Il a donc semblé préférable de ne pas inclure l'ensemble des CES dans le projet *CONSTANCES*, et de se limiter à un nombre plus restreint sur la base du volontariat, afin de ne pas complexifier des procédures d'inclusion et de suivi nécessairement lourdes et contraignantes par nature, et nécessitant un très haut degré de standardisation. Ce sont actuellement 21 CES qui ont souhaité participer au projet. Ces CES sont de taille importante, disposent d'un personnel suffisant et motivé pour l'épidémiologie, d'un plateau technique de qualité. Ce sont uniquement les affiliés au RGSS habitant dans les départements où sont implantés les CES participants qui constituent la population cible de *CONSTANCES* (voir carte).

On constate que la répartition géographique des CES *CONSTANCES* permet de représenter les principales régions françaises. La structure de la population de l'ensemble des départements où sont situés ces CES est pratiquement identique à celle de la France entière pour les principales caractéristiques démographiques et socioprofessionnelles (*cf*. plus loin).

4.4.2 Durée de l'inclusion

L'essentiel du recrutement des volontaires a eu lieu de façon graduelle de 2012 à 2019, avec quelques inclusions résiduelles jusqu'au début 2021.

4.4.3 Procédures

Les personnes éligibles tirées au sort ont été invitées à participer à *CONSTANCES*. Les méthodes de tirage au sort et les procédures pratiques utilisées pour l'inclusion sont précisées plus loin.

4.5 MODALITÉS DE SUIVI LONGITUDINAL

Une des difficultés majeures est le suivi longitudinal individuel des sujets, afin de permettre le recueil en continu des données les concernant et minimiser le nombre des perdus de vue. Trois problèmes distincts doivent être pris en compte : (i) le « traçage » des sujets ; (ii) la participation « active » des sujets au suivi ; (iii) le recueil « passif » de données de santé et d'événements socioprofessionnels à partir de diverses bases de données.

4.5.1 Traçage des sujets

Le traçage de sujets en population ouverte en vue d'un contact direct (postal, téléphonique…) est un problème particulièrement difficile, nécessitant des moyens lourds, et dont les résultats sont souvent médiocres. L'étude de diverses bases de données nationales montre cependant qu'il est possible de minimiser de façon efficace le nombre de perdus de vue dans un suivi longitudinal, tout en automatisant très largement les procédures à mettre en œuvre. Il est ainsi possible de disposer de la mise à jour régulière de l'adresse postale des participants grâce aux services que La Poste a mis en place : procédures de gestion des PND (procédures Optimis et Alliage). Les volontaires ont aussi la possibilité de déclarer une nouvelle adresse par mail ou en appelant le Numéro Vert.

4.5.2 Participation active des sujets au suivi

Un autoquestionnaire postal annuel est envoyé au domicile des sujets, et il est crucial de maximiser le taux de participation personnelle au suivi de la cohorte. Un contact régulier avec les participants est un élément important pour assurer la fidélisation. Ce contact prend la forme d'un « Journal de la cohorte *CONSTANCES* » présentant les résultats acquis, les projets associés, etc., adressé régulièrement aux participants, qui peuvent en outre s'abonner à la Newsletter trimestrielle. Un Numéro Vert est mis à disposition. Le site Internet a des fonctions voisines, et permet un contact direct avec les participants.

4.5.3 Recueil « passif »

Les sujets inclus dans *CONSTANCES* font l'objet d'un suivi dit « passif » (car n'impliquant pas une intervention des sujets eux-mêmes) d'événements socioprofessionnels et de données de santé par appariement régulier avec des bases de données nationales.

1.1.1.1 Événements socioprofessionnels

Les bases de données de la Cnav sont un élément essentiel du dispositif envisagé, à la fois pour le traçage des sujets de *CONSTANCES* et pour l'accès aux données socioprofessionnelles les concernant. Le rôle de cet organisme est notamment d'assurer le droit au paiement de la retraite pour toute personne ayant appartenu au RGSS au moins une fois au cours de sa vie. Pour cela, la Cnav a mis en place des systèmes d'information 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. Pour la constitution et l'enrichissement de ces bases de données, la Cnav reçoit régulièrement les Déclarations Annuelles des Données Sociales (DADS) transmises chaque année par les employeurs ayant un numéro SIRET ; les Données nominatives trimestrielles par les employeurs de personnel de maison ; les informations de périodes d'activité / non activité des individus relevant de l'Unedic (chômage), de la CNAM (absences pour maladie, incluant les dates de début et fin d'arrêt de travail), de la Cnaf (maternité, …), des régimes particuliers ou spéciaux (SNCF, EDF-GDF, RATP, …), et de certains autres régimes.

Cet ensemble de données est recueilli de façon prospective pour toute personne affiliée à un moment ou un autre de sa vie au RGSS. Pour celles ayant des parcours socioprofessionnels dans différents régimes, les données des autres régimes peuvent, dans certains cas, n'être collectées par la Cnav que lorsque les personnes atteignent un âge proche du passage à la retraite, lors de la vérification d'éventuelles absences d'informations.

1.1.1.2 Données de santé

Pour les événements de santé, l'accès au SNDS constitue une solution efficace. Son intérêt pour l'épidémiologie est lié au fait qu'il contient des données individuelles médicalisées, structurées et codées de manière standardisée. Leur utilisation dans une optique épidémiologique nécessite cependant un important travail méthodologique, de contrôle et de validation de données.

La base de données de remboursement de l'Assurance maladie est adaptée aux objectifs d'analyse des pratiques de prescription, mais 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. La base de données des ALD concerne tous les affiliés exonérés du ticket modérateur, après avis du service médical de l'Assurance maladie qui code l'affection exonérante selon la Classification internationale des maladies (CIM-10). La base de données du PMSI (Programme de médicalisation des systèmes d'information des hôpitaux) comprend pour chacun des séjours hospitaliers le recueil du diagnostic principal, du ou des diagnostics associés, de l'âge, du sexe et des actes diagnostiques et thérapeutiques. Les diagnostics sont codés selon la CIM-10 et les actes selon la Classification commune des actes médicaux (CCAM).

L'anonymisation des variables identifiantes est réalisée par le module FOIN (Fonction d'occultation des informations nominatives), qui repose sur le NIR de l'assuré ouvreur de droit (= NIR individuel si l'assuré est son propre ouvreur de droit ou NIR individuel de l'ouvreur de droit si l'assuré est un ayant droit), la date de naissance et le sexe du bénéficiaire. 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. L'utilisation du SNDS dans le cadre de *CONSTANCES* nécessite donc un passage dans la chaîne d'anonymisation FOIN afin de retrouver pour une personne donnée les enregistrements la concernant.

On précise plus loin les procédures prévues pour l'accès sécurisé aux bases de données que sont le SNGC et le SNDS (*Cf*. Tirage au sort et constitution des échantillons).

4.6 PRINCIPALES DONNÉES RECUEILLIES AUX DIFFÉRENTES SOURCES

4.6.1 Domaines couverts

Les données recueillies de façon systématique pour l'ensemble des participants sont destinées à fournir un corpus permettant de décrire et de suivre dans le temps l'évolution de phénomènes constituant une base d'information et une référence sur l'état de santé, la morbidité générale et la mortalité, le statut socio-économique et professionnel, l'environnement familial et social et du lieu de vie, les facteurs de risque personnels et environnementaux.

4.6.2 Choix des données

De façon générale, on s'est efforcé de choisir des variables déjà utilisées dans d'autres enquêtes, à la fois parce qu'il s'agit de mesures validées et du fait qu'il est ainsi possible de disposer de données de référence pour certaines analyses. On a également, chaque fois que cela a été possible et pertinent, utilisé des échelles publiées dans la littérature, dont les qualités métrologiques sont établies. Dans ce qui suit, on présente la nature des données qui seront recueillies, ainsi que la justification des choix effectués.

1.1.1.3 Caractéristiques sociodémographiques, statut et situation sociale

Plusieurs mesures du statut et de la situation sociale des sujets ont été introduites : situation et activité professionnelle, niveau d'études, niveau de revenus, situation matrimoniale, composition du ménage, statut socio-économique des parents et du conjoint, conditions de vie matérielles, notamment. En effet, chacun de ces éléments reflète des aspects différents du statut et de la situation sociale (Ribet et al., 2006).

Les évènements de vie, les trajectoires de vie sont des éléments particulièrement importants pour comprendre l'état de santé à l'âge adulte (Kuh & Ben Shlomo, 1997). Les réseaux de sociabilité, le soutien psychologique ou l'aide matérielle dont on peut bénéficier ont une forte influence sur la santé (Berkman et al., 2004). Enfin, le mode de couverture sociale, l'existence d'un handicap reconnu sont des éléments importants.

1.1.1.4 Localisation territoriale

Les adresses à l'inclusion et successives lors du suivi des participants font l'objet d'un géocodage, qui consiste à transformer les adresses des personnes en coordonnées spatiales, et permet de situer de façon plus ou moins précise une adresse sur une carte. De plus, environ 80 000 volontaires ont accepté de compléter leur calendrier résidentiel vie entière depuis la naissance.

L'intérêt du géocodage des lieux de résidence pour l'épidémiologie n'est plus à démontrer. Deux domaines principaux de recherche, particulièrement actifs, s'intéressent aux effets de l'environnement résidentiel sur la santé : l'épidémiologie environnementale et l'épidémiologie sociale.

Il faut souligner que selon les objectifs scientifiques des recherches concernant le contexte résidentiel, le niveau nécessaire de finesse de la localisation est variable : elle peut être relativement grossière quand il s'agit d'étudier des caractéristiques socioéconomiques

générales du quartier de vie (niveau de l'îlot IRIS, par exemple), alors que dans d'autres études, la localisation doit être beaucoup plus précise quand on s'intéresse à certaines expositions environnementales.

Pour chaque adresse, les coordonnées spatiales correspondantes (latitude et longitude avec une précision de quelques mètres correspondant à l'adresse exacte) sont enregistrées dans la base de données, en complément des adresses. On détaille plus loin les procédures utilisées pour le géocodage des adresses résidentielles et pour garantir leur confidentialité (5.11.3 - Gestion et traitement des adresses).

Remarque importante : les projets de recherche qui demandent la transmission dont la granularité est suffisamment fine pour identifier les sujets sont soumis à l'autorisation préalable de la Cnil.

1.1.1.5 Mortalité

Le statut vital est fourni par la Cnav et les causes de décès sont obtenus via le SNDS.

1.1.1.6 Données de santé communes

CONSTANCES ayant vocation à être le support d'études épidémiologiques concernant des domaines très diversifiés, le recueil de données de santé doit couvrir un large spectre. Les données qui sont recueillies par questionnaire et extraites des bases de données concernent les principaux antécédents personnels et familiaux, des échelles de santé autodéclarée (santé perçue, qualité de vie, santé mentale, échelles spécifiques : cardiovasculaire, troubles musculosquelettiques, respiratoire), les pathologies : liste des pathologies déclarées prévalentes et incidentes, diagnostic des ALD et des hospitalisations, absence au travail, handicaps, limitations, incapacités et traumatismes, recours aux soins et prise en charge de ceux-ci, décès (date et cause médicale).

Lors de l'examen de santé standardisé réalisé dans les CES, les principales données recueillies sont issues de l'examen clinique (poids, taille, rapport taille-hanches, tension artérielle, fréquence cardiaque, évaluation des fonctions physiques et cognitives), de l'examen paraclinique (vision, audition, spirométrie), et des investigations biologiques (notamment régulation glycémique, bilan lipidique, bilan hépatique, créatininémie, numération formule sanguine, examen urinaire). Les modalités de passation de cet examen et la description des données recueillies sont présentées de façon détaillée plus loin (5.6).

1.1.1.7 Origine géographique

L'autoquestionnaire d'inclusion comporte des questions permettant de caractériser l'origine géographique des participants. La recherche de cette information est justifiée par diverses raisons scientifiques. Ainsi, certaines maladies ont une prévalence nettement différenciée selon les zones géographiques (hémopathies, maladies cardiovasculaires, rhumatismales, etc.). Les caractéristiques génétiques sont également très variables selon la répartition géographique des populations, et les recherches s'intéressant à la susceptibilité génétique ou aux interactions gène-environnement doivent pouvoir prendre en compte l'origine géographique des sujets.

Nous avons donc introduit les deux questions suivantes (une pour le père, l'autre pour la mère) :

De quelle zone géographique votre père (mère) est-il (elle) originaire ? : 1) France métropolitaine ; 2) DOM-TOM ; 3) Europe ; 4) Afrique du nord ; 5) Afrique noire (ou subsaharienne) ; 6) Asie ; 7) Ne peut pas ou ne souhaite pas répondre ; 8) Autre

Le recueil se fait par auto-déclaration, et la possibilité de ne pas répondre est proposée aux sujets.

1.1.1.8 Aspects spécifiques du vieillissement

Les recherches concernant le vieillissement se sont essentiellement restreintes à des catégories d'âge au-delà de 60 ou 65 ans ; or, le vieillissement est un processus continu

commençant tôt dans la vie. Des données spécifiques sont donc recueillies lors de l'inclusion dans *CONSTANCES*, et font l'objet d'un suivi longitudinal régulier. L'étape d'inclusion permet un bilan de symptômes et de capacités explorant les caractéristiques physiques, psychiques, cognitives et fonctionnelles des sujets. Grâce au suivi longitudinal, l'avancée en âge permettra progressivement la prise en compte du rôle de facteurs de risque de nature diverse au cours de la vie sur le vieillissement. Le recueil de ces données spécifiques concerne les sujets à partir de l'âge de 45 ans, considérant que les processus de vieillissement sont largement engagés à cet âge.

Pour certains problèmes de santé généraux, une attention particulière est portée à la formulation des questions pour les plus âgés : troubles sphinctériens et prostatiques, pathologie rhumatismale et ostéoarticulaire, troubles sensoriels et bucco-dentaires.

L'évaluation des capacités fonctionnelles chez les personnes âgées est réalisée à l'aide de l'échelle IADL (*Instrumental Activities of Daily Living*) (Lawton *et al*, 1969). On utilise également d'autres outils issus de l'enquête HID de l'Insee (DREES, 2001), permettant d'étudier les changements dans les activités. On y intègre également la capacité à utiliser les nouvelles technologies.

On a fait le choix de certains tests couvrant les dimensions cognitives et le fonctionnement physique en considérant des critères psychométriques (validation) et de faisabilité, et en tenant compte des caractéristiques de la population (encore « jeune » et valide à l'inclusion). Les outils ont été choisis pour être suffisamment sensibles pour détecter des modifications mineures, même chez des sujets encore relativement jeunes. Ils doivent permettre d'évaluer le niveau cognitif des sujets afin de définir une situation de référence pour analyser les phénomènes de déclin. Ces instruments ont été sélectionnés en tenant compte de l'expérience acquise dans la population âgée entre 60 et 69 ans de l'étude EVA (Dufouil *et al.,* 2003), dans l'étude PAQUID (Dartigues *et al.,* 1991), dans celle de l'étude 3C (*Three-Cities Study Group*, 2003), de l'étude européenne SHARE, ainsi que dans la cohorte Whitehall II (Singh-Manoux *et al*., 2005).

Pour les fonctions cognitives, il s'agit des tests suivants : MMSE (Mini mental state examination) (Folstein *et al*., 1985) ; Trail Making Test A - B (Boll & Reitan, 1973 Miner *et al*, 1998) ; Code de Wechsler (Wechsler, 1981) ; Digital Finger Tapping Test (Mitrushina *et al*., 1999) ; Évocation lexicale sémantique et alphabétique *(*Borkowski *et al*., 1967, Cardebat *et al*., 1990*)* ; Free and Cued Selective Reminding Test with Immediate Recall (FCSRT-IR) (Grober *et al*. 1998 *;* Van der Linden *et al*., 2004).

Pour le fonctionnement physique, les épreuves choisies sont les suivantes : Test de vitesse de marche (Shkuratova *et al*., 2004) ; Test de l'équilibre : station unipodale de 10 secondes (Horak *et al*., 1989) ; Hand Grip Test (Giampaoli *et al*., 1999).

1.1.1.9 Problèmes de santé spécifiques des femmes

Les données concernant spécifiquement ce domaine concernent les événements gynécologiques pouvant survenir au cours de la vie : cycles menstruels, contraception, fertilité, grossesses, maladies des seins, dépistage de pathologies gynécologiques, ménopause. Elles s'intéressent également aux fractures ostéoporotiques, aux troubles sphinctériens et troubles de la statique périnéale. Une grande partie des données est recueillie par autoquestionnaire ; d'autres données sont recueillies directement par le médecin lors de l'interrogatoire d'inclusion.

1.1.1.10 Biobanque

L'intérêt scientifique d'un tel outil est considérable, et offre des potentialités à long terme dans de nombreux domaines de recherche. En effet, une biobanque permet la mise en œuvre d'analyses biologiques centralisées, et permet, sous condition d'une standardisation stricte des procédures pré analytiques, de s'affranchir du difficile problème de la variabilité inter-laboratoires. Elle offre aussi l'intérêt de conserver les échantillons à très long terme, ouvrant ainsi la possibilité de tester des hypothèses de recherche nouvelles au fur et à mesure de

l'évolution des connaissances scientifiques, tout en permettant d'utiliser à chaque fois des techniques analytiques en constante amélioration (Zins et *al.*, 2003).

la biobanque contient des échantillons de biologiques de 58 000 volontaires prélevés dans les Centres d'Examens de Santé lors des bilans de santé entre 2018 et 2021. À ce jour (2023), la biobanque inclut pour chaque volontaire, 26 aliquots de sang (sérum, plasma hépariné au lithium, plasma EDTA et buffy coat) et d'urine, totalisant plus de 1 400 000 aliquots

1.1.1.11 Comportements

Les principaux types de comportements liés à la santé concernent les consommations de tabac et d'alcool (consommations historiques et actuelles), les habitudes alimentaires et l'activité physique, ainsi que l'usage du cannabis, l'âge des premiers rapports sexuels et l'orientation sexuelle.

1.1.1.12 Facteurs professionnels

À l'inclusion, les volontaires ont complété deux questionnaires.

- Calendrier professionnel : il consigne tous les emplois successifs occupés pendant au moins 6 mois durant la carrière. Une mise à jour du calendrier professionnel depuis l'inclusion a été effectuée dans le questionnaire de suivi annuel 2023. Les emplois indiqués par les volontaires ont été codés selon les classifications françaises PCS (Professions et catégories sociales) et NAF (Nomenclature d'activités française) de l'INSEE. Fin 2023, le codage des calendriers professionnels de 198 255 volontaires a été réalisé, représentant au total 637 149 épisodes professionnels. Ce codage permet de lier les données aux matrices-emplois-expositions françaises structurées en PCS-NAF. Cependant, il n'est pas directement compatible avec les d'autres MEE, notamment internationales qui utilisent d'autres systèmes de codage, essentiellement l'ISCO (International Standard Classification of Occupations) qui existe en deux versions : ISCO-68 et ISCO-88. Pour résoudre ce problème d'incompatibilité des classifications entre elles et pouvoir apparier Constances avec diverses matrices codées selon ces différents systèmes, deux passerelles de transcodage PCS-NAF vers ISCO-68 et ISCO-88ont été développées.
- Questionnaire Expositions professionnelles vie en entière : il concerne les expositions à des produits chimiques, bruits et températures extrêmes, les contraintes organisationnelles, le stress au travail (échelle « déséquilibre efforts-récompenses » (Siegrist, 2002)) ; les contraintes posturales et gestuelles. Pour chaque nuisance, on recueille l'année de début et de fin d'exposition.

4.6.3 Périodicité du recueil

La périodicité du suivi est variable selon les sources.

L'autoquestionnaire postal est envoyé annuellement. Cette périodicité annuelle se justifie pour plusieurs raisons : elle permet un suivi « serré » de l'évolution de nombreux paramètres ; elle permet de recueillir des données très nombreuses sans demander aux sujets un effort trop important, car on peut étaler le recueil sur plusieurs années en proposant chaque fois un questionnaire de taille raisonnable ; elle permet de réagir rapidement lorsqu'on souhaite mettre en place le recueil de données non initialement prévues (nouvelles études) ; elle fidélise les participants, car un trop long délai entre deux questionnaires est un facteur d'abandon.

Parmi les données recueillies par autoquestionnaire, certaines le sont chaque année (état de santé et morbidité déclarée, événements de vie et caractéristiques du lieu de résidence, tabac, alcool, etc.), et d'autres font l'objet d'un recueil dont la périodicité est plus longue, selon un calendrier préétabli : échelles de santé, questionnaires spécifiques d'un domaine de santé ou de facteurs de risque spécifiques (échelles de santé globale, de dépressivité, d'évaluation de facteurs psychosociaux, de santé respiratoire, etc.). Ceci permet de conserver une taille de questionnaire relativement modeste compatible avec une forte participation.

Le suivi dans les bases de données nationales est permanent, puisque celles-ci enregistrent pour l'essentiel des événements en continu.

Il est également proposé aux participants de revenir tous les 4 ans au CES pour un nouvel examen de santé. Les volontaires ayant déménagé hors de leur département d'inclusion peuvent être revus dans les CES dépendant de la nouvelle CPAM dont ils dépendent.

Remarque : Les questionnaires d'inclusion, les instructions pour la passation des tests, les procédures opératoires standardisées (POS) et les cahiers de notation pour le recueil des données dans les CES, ainsi que la liste des variables et des sources où il est prévu de les recueillir au cours du suivi prospectif peuvent être téléchargés (www.constances.fr).

4.7 CONTRÔLE DE QUALITÉ ET VALIDATION DES ÉVÉNEMENTS DE SANTÉ

Les **auto-questionnaires** font l'objet des contrôles habituels : pourcentages de non réponse, de données manquantes, délai de retour, etc.

Pour les **données recueillies au cours des visites dans les CES**, un contrôle de qualité systématique et permanent est mis en place afin d'évaluer la précision, la reproductibilité, la concordance, la validité interne et externe des données recueillies, et d'étudier les facteurs de variabilité. Ce contrôle utilise diverses techniques, incluant des visites régulières sur sites d'attachés de recherche épidémiologique.

Pour les **données de suivi socioprofessionnel** issues du SNGC, elles sont croisées avec les déclarations des sujets.

Pour les **données de santé extraites des bases de données nationales** se pose un important problème de validation. On sait en effet que, pour des raisons diverses, les données concernant des problèmes de santé réunies au sein du SNDS (codage des actes de biologie, médicaments, données de l'Échelon Médical de l'Assurance Maladie, données du PMSI), ne sont pas toutes fiables. Ceci est notamment le cas pour les diagnostics médicaux précis ; or les contraintes d'un suivi épidémiologique de qualité nécessitent une validation rigoureuse. Une attention particulière est donc portée à la validation des diagnostics extraits des bases de données médico-administratives, qui font l'objet d'un contrôle systématique. La procédure repose sur l'utilisation de plusieurs sources : déclarations des individus (via les questionnaires), données collectées par les CES, données fournies par les bases médico-administratives. Les données collectées aux diverses sources citées sont considérées comme des signalements de pathologies potentielles, qui doivent être confirmées par d'autres moyens, dans le cadre d'une validation directe individuelle.

Une validation directe individuelle des événements de santé est déclenchée à partir des signalements trouvés dans les sources citées, notamment dans le SNDS qui offre l'avantage de l'exhaustivité et qui est indépendante des sujets, donc insensible aux biais de déclaration. Les données extraites régulièrement des bases du SNDS sont les suivantes :

- Affections de longue durée (ALD), codées selon la CIM-10 ;
- Données d'hospitalisation : les Résumés de sortie anonymisés (RSA) permettent d'avoir accès aux données 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), données médicales (diagnostic principal et ensemble des diagnostics associés) et des actes pratiqués. Les diagnostics sont codés selon la CIM-10 ; la Classification commune des actes médicaux (CCAM), qui harmonise la codification des actes entre médecine de ville et médecine hospitalière, est utilisée pour le codage des actes ;
- Maladies professionnelles (MP), codées selon la CIM-10 ;
- Données de consommation de soins : identification et spécialité des professionnels consultés, médicaments présentés au remboursement et dispositifs médicaux, actes de biologie codés.

Aucune de ces sources ne permet, ni de confirmer, ni de dater avec une validité suffisante des événements de santé incidents durant le suivi. Il est donc indispensable de pouvoir accéder à des documents médicaux complémentaires permettant à des spécialistes indépendants, sur la base des documents collectés, d'aboutir à un diagnostic précis.

Pour cela, les participants repérés à partir des sources régulières, sont contactés afin de récupérer les documents médicaux nécessaires pour la certification des évènements. Une fois les comptes-rendus médicaux et autres documents médicaux collectés, ceux-ci sont mis en forme et soumis à des Comités d'experts indépendants pour adjudication à partir d'une grille préétablie de définition des pathologies et des critères de définition. Jusqu'à présent, trois types de pathologies ont été investiguées : infarctus du myocarde, accident vasculaire cérébral et cancers ; ultérieurement, ces procédures seront étendues à d'autres pathologies selon les demandes des chercheurs.

Une autre approche pour obtenir des données médicalisées valides est l'utilisation d'algorithmes multi sources mobilisant des données du SNDS et des questionnaires pour identifier des cas de maladies spécifiques. Actuellement, une centaine de tels algorithmes ont été implémentés.

L'organisation pratique du processus de validation des événements de santé est décrite plus loin (*cf. Aspects opérationnels de l'inclusion et du suivi – 5.12*)

Remarque importante : les procédures de validation des événements nécessitent le recours à des informations complémentaires fournies soit par les participants eux-mêmes, soit par des tiers (médecins traitants, services hospitaliers). Le formulaire de consentement proposé aux sujets de *CONSTANCES* à l'inclusion précise donc explicitement que l'équipe pourra avoir accès aux bases de données médicales, et en cas de besoin contacter directement les sujets, ou les professionnels de santé et les hôpitaux les ayant pris en charge. Le formulaire de consentement permet aux sujets de refuser tout ou partie de ces procédures.

4.8 TIRAGE AU SORT ET CONSTITUTION DES ÉCHANTILLONS : PROBLÈMES MÉTHODOLOGIQUES

On envisage ici les principaux problèmes méthodologiques dus aux effets de sélection à l'inclusion et au suivi de la cohorte, ainsi que les solutions retenues.

4.8.1 Principaux types d'effets de sélection

Une des sources majeures de biais dans les enquêtes épidémiologiques provient des effets de sélection, qui surviennent lorsque la population observée diffère de la population cible en raison de phénomènes liés au recrutement ou au suivi des sujets. Les biais susceptibles de se produire peuvent concerner : (*i*) l'estimation de la prévalence ou de l'incidence de la maladie (ou de la prévalence de l'exposition à un facteur de risque) ; (*ii*) l'estimation de l'association entre exposition et maladie. L'estimation de la prévalence de la maladie (ou de l'exposition) ou de l'association exposition - maladie est biaisée si la probabilité d'être malade (ou exposé) n'est pas indépendante de la probabilité d'être inclus dans l'étude, ou si la relation exposition - maladie est différente chez les sujets inclus et ceux qui ne sont pas inclus. Dans les cohortes longitudinales, des effets de sélection peuvent se produire lors de l'inclusion, et tout au long du suivi du fait de l'attrition de la cohorte (Goldberg & Luce, 2001). Ces effets peuvent être dus à divers phénomènes qui permettent de classer les sujets éligibles selon plusieurs catégories :

- non retrouvés : personnes éligibles sélectionnées dans la base de sondage, mais non retrouvées lors de l'inclusion ;
- non volontaires : personnes sélectionnées et retrouvées, mais n'acceptant pas de participer lors de l'inclusion ;
- non répondants : sujets participants à l'inclusion, mais ne participant plus de façon active pendant le suivi ;
- « perdus de vue » : sujets participants, mais dont on n'a plus aucune nouvelle à partir d'un point du suivi, qu'il s'agisse de participation active ou passive.

Le terme « non participant » regroupe les deux premières catégories (non retrouvés et non volontaires) ; le terme « attrition » concerne le suivi de la cohorte après inclusion, et correspond aux deux dernières catégories (sujets ayant participé à l'inclusion, et ne répondant plus et/ou perdus de vue). **Dans ce qui suit, on considère les « non retrouvés » et les « non volontaires », comme une seule catégorie appelée « non participants** », car si les phénomènes qui sont à l'origine de ces deux types de non-participation peuvent être différents, les conséquences en termes de méthodes de correction sont identiques pour les deux situations.

La littérature montre amplement que ces différents types de sélection ne sont pas aléatoires, et qu'ils sont au contraire associés à de nombreux phénomènes qui peuvent biaiser les résultats obtenus dans une cohorte épidémiologique (Goldberg & Luce, 2001). *CONSTANCES* a l'ambition d'être un outil ayant à la fois des objectifs descriptifs d'information en santé publique (estimation de paramètres variés concernant l'ensemble de la population cible), et de recherche épidémiologique visant à mieux comprendre les relations entre expositions à des facteurs de risque de nature diverse et survenue de maladies (objectif analytique). Or, le problème des biais potentiels lié aux divers effets de sélection est très différent selon qu'il s'agit d'objectifs analytiques ou descriptifs.

4.8.2 Étude analytique des relations entre expositions et maladies

Au sein d'une cohorte dont les procédures d'inclusion ont été les mêmes pour tous les sujets (cas de *CONSTANCES*), la relation exposition - maladie n'est *a priori* pas différente entre les sujets inclus et ceux qui ne le sont pas (Criqui, 1979 ; Austin *et al*., 1981). Une des raisons est que ceux qui participent effectivement sont indemnes des maladies qui seront analysées au moment de l'inclusion, seuls les cas incidents pendant la période de suivi étant pris en compte dans les études de cohorte. Un biais peut théoriquement se produire s'il existe un facteur directement impliqué dans la relation ayant joué sur la participation, et non pris en compte dans l'analyse ; il est établi de façon empirique qu'habituellement ce type de situation, quand il existe, ne joue que de façon marginale. Ainsi, pour analyser les effets du tabac sur le risque de cancer, il n'est pas nécessaire d'observer un échantillon représentatif de la population, mais de disposer d'effectifs suffisants de non-fumeurs et de fumeurs parmi lesquels le niveau d'exposition est contrasté : en effet, sur la base des connaissances actuelles, il est très vraisemblable que parmi les inclus et non inclus, les mécanismes de la cancérogenèse liée au tabac sont identiques à ceux qui concernent l'ensemble de la population. Les modalités de sélection des participants de *CONSTANCES* ne génèreront donc *a priori* pas de biais, ou seulement des biais minimes, lorsqu'il s'agit de comprendre comment les expositions à des facteurs de risque, les caractéristiques professionnelles et sociales, etc., influencent l'état de santé et peuvent être à l'origine de pathologies, ou au contraire protectrices.

Ceci n'est cependant pas toujours vrai. Par exemple, l'existence d'une pathologie grave parmi les ascendants d'une personne et connue de celle-ci peut influencer sa participation à une étude concernant cette maladie ; or, cette personne peut partager avec l'ascendant atteint des caractéristiques associées au risque de cette maladie. Dans un tel cas, des biais sur le rôle de facteurs de risque peuvent survenir si la participation est liée à l'existence d'antécédents familiaux de la maladie. Cependant, ce type de configuration est en pratique peu courant.

Le problème de l'attrition au cours du suivi peut par contre être à l'origine de biais importants si la probabilité de ne plus être suivi diffère chez les exposés et non exposés, et/ou chez ceux qui sont ou ne sont pas devenus malades, ce qui est souvent le cas.

On rappelle que toutes les cohortes épidémiologiques longitudinales basées sur la participation volontaire (c'est-à-dire toutes celles qui impliquent la participation active des sujets pour fournir des données qui ne sont pas disponibles autrement) présentent les mêmes phénomènes de sélection à l'inclusion et pendant le suivi. Certaines, célèbres, ont néanmoins contribué de façon décisive à l'avancement des connaissances scientifiques. On peut citer à titre d'exemple l'étude de Framingham (petite communauté de la banlieue de Boston, qu'on ne peut en aucune façon considérer comme représentative de la population mondiale), qui dans le domaine cardiovasculaire a apporté depuis des décennies des données essentielles

d'utilisation universelle (Oppenheimer, 2005), celle des médecins anglais qui a fondé une grande partie des connaissances actuelles sur les effets du tabac (Doll *et al*., 1994), ou de la cohorte Whitehall composée exclusivement de fonctionnaires britanniques et qui constitue la référence internationale en matière de résultats sur les déterminants sociaux des inégalités de santé (Marmot *et al*., 1991).

4.8.3 Étude descriptive de la fréquence des problèmes de santé et des expositions

Ici, les paramètres d'intérêt doivent être estimés sur un échantillon représentatif de la population cible. Or, la cohorte *CONSTANCES* présente deux caractéristiques particulières : (*i*) elle est constituée à partir d'un recrutement dans un nombre limité de CES dont la répartition géographique ne couvre pas l'ensemble de la France ; (*ii*) elle est composée uniquement de sujets sélectionnés, volontaires parmi les personnes effectivement retrouvées dans les fichiers de la base de sondage.

4.8.3.1 Couverture géographique

On a vérifié que la structure de la population de l'ensemble des départements où sont situés les CES *CONSTANCES* est pratiquement identique à celle de la France entière pour les principales caractéristiques démographiques et socioprofessionnelles.

Comparaison entre France et « départements CES *CONSTANCES* » (Source : Insee, Recensement)

	France	CONSTANCES
Âge		
<20 ans	24,57	23,96
20-29	13,48	14,97
30-39	14,66	14,78
40-49	14,47	14,19
50-59	11,49	11,42
60-69	9,37	9,01
70-79	7,96	7,74
80 +	4,0	3,92
Sexe		
Hommes	48,56	48,16
Femmes	51,44	51,84
PCS (actifs uniquement)		
1- Agriculteurs exploitants	2,72	2,02
2- Artisans, comm. et chefs d'ent.	6,62	6,44
3- Cadres et prof. intell. sup.	13,12	16,67
4- Professions intermédiaires	23,07	24,07
5- Employés	28,83	28,30
6- Ouvriers	25,64	22,49
Secteur d'activité (actifs uniquement)		
1- Agriculture, chasse, sylviculture	4,00	3,19
2- Pêche, aquaculture	0,12	0,09
3- Industries extractives	0,20	0,16
4- Industrie manufacturière	17,26	15,20
5- Prod Distrib électricité, gaz, eau	0,90	0,94
6- Construction	5,83	5,23
7- Commerce - réparations	13,24	13,07
8- Hôtels et restaurants	3,54	3,57
9- Transports et communications	6,44	6,61
10- Activités financières et assurances	3,00	3,17
11- Immobilier, location et services entrep	11,20	12,93
12- Administration publique	9,84	9,81
13- Éducation	7,30	7,98
14- Santé et action sociale	11,56	11,91

15- Services collectifs, sociaux et personnels	4,35	4,90
16- Services domestiques	1,12	1,10

4.8.3.2 Effets de sélection liés à la non-participation et à l'attrition

Le mode d'inclusion faisant appel au volontariat entraîne inévitablement des effets de sélection, même lorsqu'on procède à un tirage au sort aléatoire d'un échantillon dans une base de sondage appropriée. On rencontre en effet à l'inclusion, comme on l'a vu, des non participants, qui sont une source potentielle de biais. Pour y remédier, on s'efforce habituellement de recueillir un minimum de données sur les non participants (âge, sexe, PCS essentiellement), afin de procéder ultérieurement à des redressements pour estimer les paramètres d'intérêt. Cette approche est utilisée dans de nombreuses enquêtes de santé, mais connaît cependant certaines limites. Ainsi, il n'est pas toujours possible de recueillir les données de redressement (âge, sexe, PCS) pour les sujets non participants, ou on en dispose seulement pour une partie d'entre eux. Il n'est pas toujours clair de savoir si ces données sont suffisantes pour contrôler les biais potentiels, car on sait par exemple qu'au sein de la même catégorie socio-économique existent de larges différences à bien des égards, notamment en termes de santé, de comportements, de modes de vie, de réseaux sociaux, etc. La comparaison des volontaires de la cohorte Gazel à leurs collègues d'EDF-GDF non participants et de même catégorie socioprofessionnelle illustre ce point de façon détaillée (Goldberg *et al*., 2001). Finalement, on est rarement en situation de contrôler complètement les biais de sélection potentiels, faute de disposer de données pertinentes recueillies à la fois pour les participants et l'ensemble des non participants.

Afin d'obtenir un échantillon représentatif de la population cible et de minimiser les biais liés aux effets de sélection à l'inclusion et durant le suivi, **on a procédé pour *CONSTANCES* de la façon suivante.**

Une pondération individuelle est calculée et permet, à partir de la cohorte, de produire des estimations de structures ou de niveaux de certaines variables généralisables à la population des affiliés au RGSS des « départements *CONSTANCES »*. Chaque année pendant la période d'inclusion, une pondération a été calculée. Le principe général de calcul de ces pondérations consiste à appliquer successivement au poids de tirage initial attribué à chaque individu deux coefficients : un coefficient de correction de « non-participation » puis un coefficient de correction pour « attrition ». Ces deux coefficients seront déterminés à partir de l'analyse des variables liées d'une part à la participation à la cohorte, et d'autre part à l'attrition. Ayant ainsi obtenu des poids avant calage, la dernière étape consiste à calculer une pondération finale par calage sur les marges, décrivant de façon pertinente la population à laquelle on souhaite généraliser les résultats. Les années pendant lesquelles a eu lieu le recrutement des volontaires a été sont traitée de façon indépendante.

Base de sondage

La base de sondage est constituée des affiliés des CPAM correspondants aux CES *CONSTANCES*, appartenant au Régime général, aux SLM ou à la Camieg, et âgés de 18 à 69 ans à l'inclusion.

Tirage au sort

Le tirage au sort est stratifié avec probabilités d'inclusion inégales ; pour chaque CES, les strates ont été constituées par les variables : type de régime d'assurance maladie (Régime général, SLM et Camieg), âge, sexe, statut d'activité et PCS à un chiffre. La probabilité d'inclusion d'une personne affiliée dépend, d'une part du volume annuel d'examens de santé, et d'autre part de la probabilité *a priori* qu'elle a d'accepter ou non de participer à *CONSTANCES*.

En pratique, les effectifs d'affiliés à inviter dans chaque strate de chaque CES sont calculés de manière à ce que les effectifs attendus des participants à *CONSTANCES* soient proportionnels aux effectifs des strates pour un CES donné ; ces effectifs, rapportés aux

nombres d'affiliés permettent d'obtenir les probabilités d'inclusion à partir desquelles on déduit les pondérations initiales.

Les probabilités *a priori* ont été établies à partir des données du pilote, de l'Enquête décennale santé 2002-2003 et de l'Enquête prévalence des hépatites B et C en France en 2004 (InVS, 2005). Les résultats des premières inclusions ont permis d'obtenir des probabilités *a priori* plus précises pour les vagues suivantes.

Redressement pour non-participation

Un redressement permet de corriger, au moins partiellement, les biais de sélection induits par la non-participation. Le principe du redressement repose sur l'utilisation d'informations auxiliaires disponibles pour la population cible ; sa qualité dépend de la pertinence et de la richesse de ces informations.

Calcul des coefficients de redressement pour non participation. On a constitué une « cohorte contrôle » tirée au sort parmi les non participants pour lesquels on dispose de caractéristiques sociodémographiques, grâce aux données du SNGC (âge, sexe, statut d'activité et PCS à un chiffre), ainsi que de nombreuses informations de santé et de recours aux soins grâce aux données du SNDS. Ainsi, on dispose pour les participants et l'échantillon de non participants de données issues du SNDS et du SNGC. Ces données permettent d'estimer les probabilités de participation à *CONSTANCES* à l'aide de modèles de prédiction (*cf.* ci-dessous : *Modèles de prédiction*) ; l'inverse de la probabilité de participation fournit ensuite le coefficient de redressement propre à chaque participant.

Actualisation des coefficients de redressement pour non participation. L'échantillon de non participants est suivi longitudinalement ; en effet la validité du redressement décrit précédemment est basée sur l'hypothèse que la probabilité de participation ne dépende que de variables disponibles au moment où l'analyse de la participation est réalisée. Il est vraisemblable que cette hypothèse ne soit pas vérifiée, et que la participation à *CONSTANCES* dépende en fait d'autres variables inconnues ou non disponibles (et de fait non mesurées) avant la participation ou non à la cohorte *CONSTANCES*. Les données recueillies ultérieurement sur les non participants (et les participants) peuvent refléter partiellement ces variables et permettent de réactualiser les coefficients de redressement grâce à une meilleure estimation de la probabilité de participation.

Modèles de prédiction

Deux méthodes ont été utilisées afin d'estimer les probabilités de participation. La méthode de référence est la régression logistique, suivie de la méthode des scores pour constituer des groupes homogènes de réponse (Eltinge & Yansaneh 1997). Ensuite, ont été mises en œuvre des méthodes de régression récemment développées, comme l'algorithme du « superlearner » (Polley &, van der Lann 2010). À partir des données et d'un certain nombre de modèles candidats (qui incluent les méthodes de segmentation), cet algorithme détermine en utilisant la validation croisée et une fonction de perte appropriée quel est, parmi les modèles candidats, le modèle le plus adapté pour modéliser la variable d'intérêt. Les modèles sont classés en fonction de leur performance et un nouvel algorithme hybride et original est construit sous la forme d'une combinaison pondérée des meilleurs modèles candidats (Pirracchio 2014).

L'intérêt est de comparer les prévalences obtenues par les deux systèmes de pondération, le premier issu de la méthode de référence, et le deuxième obtenu grâce aux méthodes de régression récentes, afin de préciser la méthode retenue selon l'efficacité et la facilité de mise en œuvre ces deux types de modèle.

Redressement pour attrition

On peut considérer que, hormis les personnes ayant refusé l'appariement, celles qui demanderont à sortir de la cohorte, les erreurs d'enregistrement et les sujets quittant la France de façon définitive, aucune des personnes incluses dans *CONSTANCES* ne sera définitivement perdue de vue, puisque les participants seront suivis passivement dans le

SNDS et le SNGC. Il y aura cependant une attrition due au non renvoi de l'autoquestionnaire annuel. Ainsi, pour que les analyses portant sur les variables de l'autoquestionnaire puissent être menées dans de bonnes conditions de validité, il convient d'effectuer un redressement pour attrition ; ce dernier est calculé à partir des données des participants à *CONSTANCES*.

Calcul des coefficients de redressement pour attrition. Pour l'année suivant l'année d'inclusion, on distingue les participants ayant répondu à l'autoquestionnaire des participants n'y ayant pas répondu. On dispose pour tous les participants des données recueillies lors de l'inclusion dans *CONSTANCES* ainsi que des données issues du SNDS et du SNGC. Les coefficients de redressement pour attrition peuvent ainsi être calculés par une méthode semblable à celle du calcul des coefficients de redressement pour non-participation.

Définition de l'attrition. Il est probable qu'un certain nombre de participants à *CONSTANCES* soit non répondants certaines années et répondants d'autres années. Une solution facile à mettre en œuvre consisterait à considérer qu'un participant à *CONSTANCES* abandonne dès qu'il y a non renvoi d'un autoquestionnaire (ou un autre nombre arbitraire d'années sans réponse) : chaque année, l'analyse de l'attrition serait effectuée, et des coefficients de redressement calculés pour les participants n'ayant pas abandonné. Cependant, cette solution n'est pas satisfaisante : en effet, avec cette définition stricte, le nombre de participants qui abandonneraient la cohorte serait supérieur au nombre d'abandons réels. D'autres solutions dérivées de la méthodologie des non réponses partielles sont utilisées : exclusion des non répondants, utilisation de variables indicatrices de données manquantes, imputations simples ou multiples, ou repondérations.

Actualisation des coefficients de redressement pour attrition. Le suivi passif et longitudinal des participants dans le SNIIR-AM et le SNGC, que ceux-ci abandonnent ou non la cohorte pour non renvoi d'autoquestionnaire, permet d'actualiser les coefficients de redressement pour attrition.

Calage sur marges

Chaque année, on cale la population de la cohorte sur les marges de la population de référence. La première année, on a tiré au sort dans les bases de données de la Cnav (RNIAM et SNGC) un échantillon d'affiliés au Régime général, aux SLM et à la Camieg des départements *CONSTANCES,* âgés de 18 à 69 ans. La deuxième année, et respectivement la troisième, quatrième et cinquième année, l'échantillon d'affiliés tirés au sort a été constitué de personnes âgées de 18 à 70 ans, puis de 18 à 71 ans, de 18 à 72 ans et de 18 à 73 ans. Chacun de ces échantillons est deux fois plus important que celui de la population des participants à *CONSTANCES* de sorte que la population de calage soit significativement plus importante que la population à caler.

Après appariement de ce fichier de données avec le SNDS, on a calculé les marges pertinentes qui, outre les caractéristiques sociodémographiques et économiques, intègrent des variables relatives à la santé et aux caractéristiques du recours aux soins (toute information également disponible pour les participants à l'enquête). La qualité du calage est donc sensiblement améliorée par le calcul de marges ayant un rapport spécifique avec la finalité santé de la cohorte.

5 ASPECTS OPÉRATIONNELS DE L'INCLUSION ET DU SUIVI

5.1 INFORMATION PRÉALABLE À L'INVITATION À PARTICIPER À *CONSTANCES*

Préalablement à l'envoi des invitations à la suite du tirage au sort dans le RNIAM et le SNGC (*cf*. ci-dessous), une information générale a été faite dans les départements concernés par *CONSTANCES*, *via* les médias locaux, le site de la CNAM (site Améli), ceux des SLM participantes et de la Camieg.

5.2 CONSTITUTION DES COHORTES (PARTICIPANTS ET NON PARTICIPANTS)

Dans ce qui suit, les sigles suivants sont utilisés.

Camieg : Caisse d'assurance maladie des Industries Électrique et Gazière

CEN : Centre d'Exploitation National
CENTI : Centre national de traitement informatique
Code Alliage : Combinaison d'un numéro expéditeur et d'un numéro destinataire (dispositif de La Poste pour la gestion des PND)
Code RNIAM Régime obligatoire : Code Grand Régime + Centre Gestionnaire (ou CPAM d'affiliation) + Centre de paiement (ou codification Nationale des Mutuelles dans le cas des Sections Locales Mutualistes et de la Camieg)
CESP : Centre de recherche en Épidémiologie et Santé des Populations
CTI : Centres de traitement informatique de la CNAM
DSI (Cnav) : Direction des systèmes informatiques
État civil : Civilité, nom de famille, nom d'usage, prénom, prénom d'usage, date de naissance, lieu de naissance
LAD : Lecture Automatique de Documents
NIR : Numéro d'inscription au répertoire
PND : N'habite pas à l'adresse indiquée
PCS : Profession et catégorie socioprofessionnelle
RNIAM : Répertoire national inter-régime d'assurance maladie
RNIPP : Répertoire national d'identification des personnes physiques
SNGC : Système national de gestion des carrières
SNGI : Système national de gestion des identités
SNDS : Système national d'information inter-régimes de l'Assurance Maladie
Typologie d'activité professionnelle ou code TAP : Croisement des informations : statut activité (actif/sans activité) et PCS à un chiffre (cadre, profession intermédiaire, employé, ouvrier)

5.2.1 Tirage au sort et création des identifiants nécessaires

Comme on l'a indiqué, *CONSTANCES* est constituée par sondage stratifié avec probabilités inégales selon les variables : régime d'affiliation (Régime général, SLM, Camieg), âge, sexe, typologie d'activité professionnelle (code TAP correspondant au croisement de statut d'activité et de PCS à un chiffre). De façon pratique, les strates sont constituées par les CES participants, et les bases de sondage sont celles des assurés du Régime général et des SLM participantes affiliés à une « CPAM *CONSTANCES »* et de la Camieg. Le nombre d'invités d'un CES dépend du volume d'examens de santé réalisé par ce CES et du nombre de bénéficiaires dans les strates. Comme l'inclusion a duré plusieurs années, un tirage au sort annuel sans remise été effectué pendant les chaque année de la phase d'inclusion. Les modalités concrètes, qui impliquent plusieurs organismes, sont les suivantes.

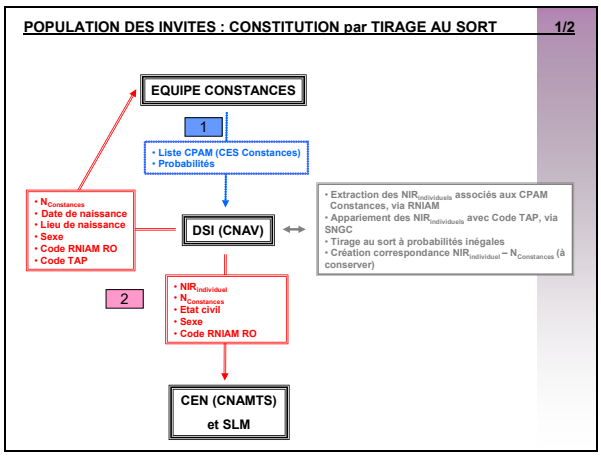
1) L'équipe Constances fournit à la DSI (Cnav) la liste des CPAM participant à *CONSTANCES* ainsi que les probabilités de participation en fonction des variables : régime d'affiliation (Régime général, SLM, Camieg), âge, sexe, typologie d'activité professionnelle, empiriquement définies sur la base d'enquêtes déjà réalisées à partir d'un échantillon aléatoire invité à se rendre dans un CES.

2) La DSI (Cnav) extrait, via le RNIAM, les $NIR_{individuel}$ des bénéficiaires affiliés aux CPAM "*CONSTANCES*", qu'ils soient affiliés au Régime général, à une SLM ou à la Camieg, ainsi que leur code RNIAM Régime obligatoire. Puis, la DSI (Cnav) apparie ces $NIR_{individuel}$ à la PCS actuelle pour les actifs et à la dernière PCS connue pour les autres (informations issues du SNGC). Enfin, le tirage au sort basé sur les probabilités fournies par l'équipe *CONSTANCES* est réalisé. Chaque sujet tiré au sort se voit attribuer un $N_{Constances}$. La correspondance $NIR_{individuel}$–$N_{Constances}$ est conservée par la DSI (Cnav) pour les échanges de données ultérieurs.

Pour chaque sujet tiré au sort, la DSI (Cnav) adresse les données suivantes :

- à l'équipe *CONSTANCES* : $N_{Constances}$, date et lieu de naissance, sexe, code RNIAM Régime obligatoire et code TAP ;

- au CEN (CNAM), à chaque SLM et à la Camieg uniquement pour ses assurés : la correspondance $NIR_{individuel}$-$N_{Constances}$, état civil, sexe et code RNIAM Régime obligatoire.



3) Le Centre d'Exploitation National (CEN), les SLM et la Camieg :

- recherchent la correspondance entre le $NIR_{individuel}$ et le $NIR_{ouvreur\ de\ droit}$ de chaque bénéficiaire tiré au sort ;
- générèrent $N_{FOIN1\ SNDS}$ (sur la base du $NIR_{ouvreur\ de\ droit}$, date de naissance et sexe du bénéficiaire) et conservent une correspondance $N_{Constances}$-$N_{FOIN1\ SNDS}$ pour les échanges ultérieurs avec le CENTI (CNAM). Dans le cas de bénéficiaires ayants droit, une génération du $N_{FOIN1\ SNDS}$ est faite d'une part sur le $NIR_{individuel}$ et d'autre part sur leur $NIR_{ayant\ droit}$ (= $NIR_{individuel}$ de leur ouvreur de droit) (on a donc, associés au $N_{Constances}$, deux $N_{FOIN1\ SNDS}$, ce qui permet d'obtenir les données du SNDS de façon pérenne quelle que soit l'évolution du statut d'ouvreur ou d'ayant droit des assurés) ;
- interrogent chacun leur base de données pour l'obtention des adresses ;
- adressent à un Tiers de confiance : $N_{Constances}$, état civil, sexe, adresse et code RNIAM Régime obligatoire.

4) Le Tiers de confiance adresse : $N_{Constances}$, date et lieu de naissance, sexe, code RNIAM Régime obligatoire et code postal du domicile à l'équipe *CONSTANCES*.

Remarque : les bénéficiaires « ayants droit » à l'inclusion sont des personnes au foyer sous le NIR de leur conjoint, des enfants de moins de 20 ans qui poursuivent leurs études, ou sont atteints d'une infirmité ou d'une maladie chronique les mettant dans l'impossibilité permanente de se livrer à un travail salarié. Ces bénéficiaires sont connus à la CNAM, par conséquent dans les CES, non par leur $NIR_{individuel}$ mais par leur $NIR_{ayant\ droit}$ (= $NIR_{individuel}$ de leur ouvreur de droit), d'où la nécessité de faire intervenir le CEN, les SLM et la Camieg pour passer du $NIR_{individuel}$ utilisé par la Cnav au $NIR_{ayant\ droit}$ utilisé par la CNAM, les SLM et la Camieg.

5.3 INVITATIONS

L'invitation dans un CES s'effectue en deux temps : un premier courrier informe les affiliés de leur droit à bénéficier d'un examen de santé sans avance de frais avec un coupon-réponse, et pour les personnes ayant répondu favorablement, un deuxième courrier de convocation précise le lieu et la date de l'examen. Dans le cadre de *CONSTANCES*, les procédures suivantes ont été utilisées.

L'équipe *CONSTANCES* envoie au Tiers de confiance la correspondance entre code RNIAM Régime obligatoire et code Alliage pour chaque $N_{Constances}$ à inviter (invitations lancées vague par vague).

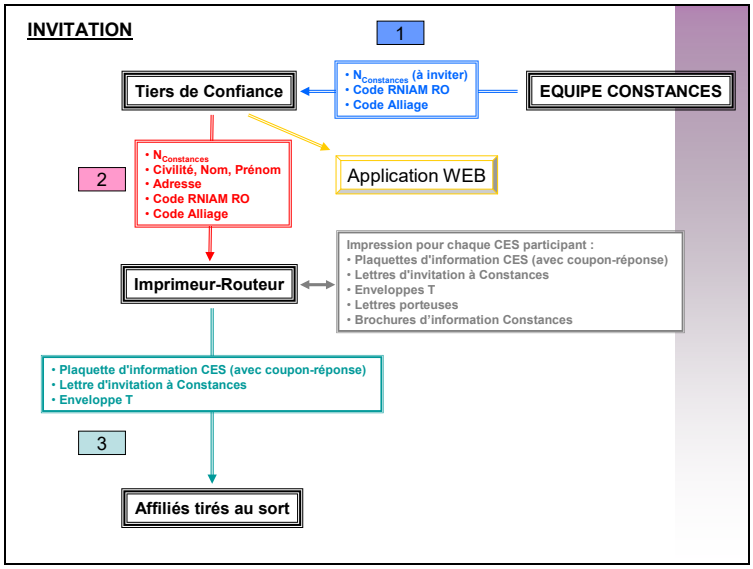
Le Tiers de confiance, pour les $N_{Constances}$ fournis :

- alimente une application Web destinée au suivi de la participation avec les données suivantes : $N_{Constances}$, civilité, nom, prénom, adresse, sexe et code RNIAM Régime obligatoire. Ainsi chaque CES peut accéder aux données des invités le concernant et enregistrer les réponses aux coupons-réponse.
- envoie à un imprimeur-routeur les données suivantes : $N_{Constances}$, civilité, nom, prénom, adresse, code RNIAM Régime obligatoire et code Alliage (voir gestion des adresses plus loin).

À la suite, l'imprimeur-routeur :

- imprime les plaquettes d'information CES (avec coupon-réponse), les lettres d'invitation à participer à *CONSTANCES*, les enveloppes T et les lettres porteuses ;
- personnalise les lettres d'invitation en indiquant : civilité, nom, prénom, adresse, ainsi que le code Alliage sous forme de code à barres ;
- personnalise les plaquettes d'information CES contenant les coupons-réponse en pré-imprimant le numéro non signifiant $N_{Constances}$;
- adresse ces documents (par lettre porteuse) à chaque bénéficiaire tiré au sort.

Par l'intermédiaire du coupon-réponse, les personnes peuvent donner leur accord pour participer à la cohorte ou indiquer leur refus que ce soit pour participer de façon active ou pour être tirées au sort pour faire partie de l'échantillon de non participants (suivis passivement dans les bases de données nationales : cf. ci-dessus : *Effets de sélection liés à la non-participation et à l'attrition*).



5.4 CONVOCATIONS

Les personnes qui ont retourné leur coupon-réponse en indiquant leur accord pour participer à la cohorte, ont reçu en retour une convocation au CES, les auto-questionnaires (Questionnaire « Mode de vie et Santé », Calendrier professionnel) à remplir à domicile et à amener lors de leur venue au CES.

Dans tous ces documents, sont précisées les finalités et les modalités de la cohorte. Concernant l'information aux personnes sur le traitement de leurs données, l'ensemble des documents précise que cette cohorte fait l'objet d'une autorisation de la CNIL, qu'elles disposent d'un droit d'accès, de modification, de rectification, et de suppression des données personnelles les concernant. Une brochure d'information détaillée qui détaille les procédures mises en place pour l'inclusion et le suivi, ainsi que les droits des personnes est également remise aux volontaires dès leur arrivée au CES.

Il a également été rendu possible d'inclure dans la cohorte les personnes qui en font spontanément la demande. Dans ce cas, ces personnes se voient attribuer un $N_{Constances}$ directement par les CES. Ce $N_{Constances}$ et les informations utiles (état civil, sexe, adresse, Code RNIAM Régime obligatoire) pour la mise en place des suivis passif et actif sont renseignées dans l'application Web.

Chaque CES assure l'enregistrement dans l'application Web des réponses et des informations relatives aux candidatures spontanées. Ces données sont ainsi accessibles par le Tiers de confiance et l'équipe *CONSTANCES* selon leurs droits d'accès respectifs.

5.5 INCLUSION DES PARTICIPANTS PAR LES CES

5.5.1 Rappel : l'examen périodique de santé des CES

Tous les assurés du RGSS (incluant les SLM) ont le droit tous les cinq ans de bénéficier d'un examen périodique de santé (EPS) entièrement pris en charge par la Sécurité sociale. Pour les consultants habituels des CES, le contenu de l'EPS peut être modulé selon l'âge, le sexe, les risques et le suivi médical habituel.

Un premier temps de l'EPS est destiné à l'exploration de l'état de santé. Il comprend entre autres tests : une analyse de sang et d'urine pour détecter d'éventuels troubles métaboliques, cardiovasculaires, hépatiques, rénaux, etc. ; la mesure de la tension artérielle, du poids, de la taille et du rapport taille/hanches ; si nécessaire un électrocardiogramme et une spirométrie, des examens de la vue et de l'audition, un questionnaire de perception de la santé, un repérage des risques vis-à-vis du tabac et de l'alcool.

Un second temps est réservé à l'examen clinique : le médecin s'entretient avec le patient sur les modes de vie, les antécédents personnels et familiaux et les problèmes révélés par les tests de la première partie de l'examen.

Une copie de l'ensemble des résultats de l'examen de santé est envoyée au médecin traitant ou à tout autre médecin désigné par le bénéficiaire.

5.6 L'EXAMEN CONSTANCES

L'examen d'inclusion de *CONSTANCES* s'appuie sur le bilan de santé existant, mais effectué selon des procédures standardisées très strictes. À l'accueil, le volontaire doit pouvoir être repéré parmi les autres consultants du CES. À cette fin, différentes procédures ont été mises en place pour la gestion des données des sujets éligibles et le suivi des dossiers individuels dans le CES.

5.6.1 Modalités de suivi du circuit du volontaire dans le CES

- L'opérateur accueille le volontaire, lui remet la brochure d'information détaillée sur CONSTANCES, lui demande de confirmer qu'il veut participer à l'étude, avec possibilité de changer d'avis lors de la signature du consentement ;
- L'opérateur remet au volontaire une pochette contenant un dossier de suivi, sur lequel il colle une étiquette identifiante. Cette pochette comprend différents documents :
    - Une planche d'étiquettes comprenant le numéro Constances non identifiant du volontaire (avec un graphisme spécifique afin de bien les reconaitre) ;
    - Le consentement éclairé ;
    - Les questionnaires des examens paracliniques (voir plus loin) ;
    - Le questionnaire Expositions professionnelles ;
    - Le questionnaire Médical ;
    - Le questionnaire Santé des femmes ;
    - Le cahier Tests fonctionnels pour les 45 ans et plus ;
    - Ce dossier est à compléter par les questionnaires remplis à domicile remis à l'accueil par les volontaires :
        - L'autoquestionnaire Modes de vie et Santé ;
        - Le Calendrier professionnel ;
    - L'opérateur indique les différentes étapes du « circuit Constances » et propose

une première lecture du consentement. Par ce consentement éclairé les participants peuvent exprimer l'acceptation ou le refus de transmission de tout ou partie des données les concernant (données de l'examen de santé effectué lors de l'inclusion ; dans le cadre du suivi : données de santé du SNDS (consommations de médicaments, actes de laboratoires, hospitalisations, consultations de généralistes, consultations de spécialistes), données socioprofessionnelles de la Cnav, adresses postales de La Poste, et l'acceptation d'être contacté par téléphone par un médecin de l'équipe Constances. Le consultant est informé que le médecin pourra répondre à ses questions lors de la consultation médicale, avant la signature du consentement ;

- À la fin de l'examen, l'opérateur récupère le dossier de suivi, vérifie que les documents sont bien identifiés.

5.6.2 Données recueillies dans les CES

L'ensemble des instruments de recueil et des procédures d'examen utilisés ont été mis au point et testés durant les pilotes. Des « Procédures opératoires standardisées » (POS) précisent de façon stricte les modalités opératoires qui doivent être respectées tout au long du circuit dans le CES afin d'obtenir des données de qualité et rigoureusement standardisées. Un programme de contrôle continu de qualité a été mis en place ; des attachés de recherche épidémiologique vérifient de façon régulière l'exhaustivité et la cohérence des dossiers individuels, et que les consentements sont bien remplis.

Questionnaires

Les questionnaires d'inclusion sont au nombre de cinq :

1 - Autoquestionnaire Modes de vie et Santé (domicile). Ce questionnaire est adressé avec la convocation aux CES aux personnes s'étant portées volontaires, et doit être remis complété lors de la venue au CES. Il comprend plusieurs parties :

- Modes de vie : alcool, tabac, cannabis, activité physique, alimentation.
- Caractéristiques sociales et démographiques : nationalité, langue maternelle, statut marital, niveau d'éducation (volontaire et conjoint), PCS (volontaire et conjoint), revenus, …
- État de santé : dépressivité (échelle CES-D), capacité visuelle, état respiratoire (échelle ECRHS), auto perception de la santé, diabète, troubles musculosquelettiques (questionnaire « *Nordic* ») ;
- Santé sexuelle et vie de couple : orientation sexuelle, utilisation de préservatif, satisfaction, etc.

2 – Calendrier professionnel. Il est lui aussi adressé à domicile avec la convocation aux CES aux personnes s'étant portées volontaires ; il permet de reconstituer toute la carrière professionnelle (épisodes de carrière de plus de six mois), y compris les périodes d'inactivité.

3 - Questionnaire Santé des femmes. Distribué à l'accueil dans le CES aux femmes volontaires, il est à remplir au CES. Il comporte des questions sur les événements gynécologiques pouvant survenir au cours de la vie d'une femme : cycles menstruels, contraception, fertilité, grossesses, maladies des seins, dépistage de pathologies gynécologiques, ménopause.

4- Questionnaire Expositions professionnelles. Il est composé de deux parties :

- La première partie concerne l'ensemble de la carrière professionnelle. Elle comprend des questions sur les contraintes organisationnelles et les expositions professionnelles à différentes nuisances (bruits, produits chimiques, biologiques, radiations …).

- La deuxième partie concerne seulement l'emploi actuel. Elle comprend une description de l'emploi, les contraintes posturales ainsi que l'exposition à des températures extrêmes.

5 - Questionnaire Médecin. Il est rempli par le médecin à l'issue de l'examen clinique, et permet de faire le point sur les principales pathologies appareil par appareil, avec recherche de la date d'incidence (année). Le poids de naissance est reporté à partir du carnet de santé s'il est disponible ; les antécédents de fractures ostéoporotiques et les antécédents familiaux (père, mère) sont également notés à partir d'une liste de pathologies préétablie ; cet interrogatoire fait partie de l'examen usuel pratiqué dans les CES.

Examens paracliniques

Les examens paracliniques font partie du bilan usuel pratiqué dans un CES. Lors des visites de suivi quadriennales, ils sont identiques à ceux réalisés pendant la visite d'incusion.

Les modalités de passation de ces examens sont fixées par des Protocoles opératoires standardisé (POS). Un cahier de maintenance est exigé pour chaque matériel. Un questionnaire recense pour chacun des POS des informations sur les conditions de passation (contre-indication, données manquantes, etc.) et permet de recueillir les données. Les examens réalisés sont les suivants :

- **Biométrie** : taille, poids, tour de taille, tour de hanches ;
- **Vision** : de près (Parinaud) et de loin (Monoyer) ;
- **Audition** : audiométrie tonale liminaire à seuil ascendant en conduction aérienne en cabine insonorisée ; fréquences testées : 1000, 2000, 4000, 6000, 500 Hertz ;
- **Spirométrie** : VEMS, CVF (après trois courbes) ;
- **ECG** : enregistrement des tracés ;
- **Tension artérielle :** deux mesures espacées de 2 mn après 5 mn de repos et mesure au bras de référence après une minute de repos ; recherche d'une hypotension orthostatique chez les plus de 65 ans et le diabétiques ; les tensiomètres électroniques sont fournis.

Examens fonctionnels physiques et cognitifs (pour les 45 ans et plus)

Administrés par des neuropsychologues formés, les examens physiques comprennent une batterie de tests des fonctions cognitives, ainsi que vitesse de marche, équilibre unipodal 10 secondes et force de préhension.

Biologie

Bilan usuel, mais répondant à des procédures pré analytiques et analytiques strictes avec un programme Assurance qualité mis en place en coopération avec l'Association Asqualab. Il comprend les éléments suivants :

- Sang : numération formule sanguine, plaquettes, glucose, hémoglobine, hématocrite, volume corpusculaire moyen, cholestérol total, HDL Cholestérol, ALAT, créatinine, triglycérides, gamma GT.
- Urine : protéines, glucose, nitrites, micro albuminurie, créatinine.

5.7 CIRCUITS DE TRANSMISSION DES DONNÉES

Deux types de données sont recueillis : des données nominatives et des données non identifiantes pour lesquelles des flux indépendants ont été mis en place.

5.7.1 Circuit pour les données nominatives

Durant la phase d'inclusion, les données suivantes (éventuellement mises à jour par les CES) : état civil, sexe, adresse, code RNIAM Régime obligatoire et statut de participation, associées aux $N_{Constances}$ ont été rendues accessibles au Tiers de confiance *via* l'application Web.

5.7.2 Circuit pour les données non identifiantes

Durant la phase d'inclusion, les données de participation de tous les bénéficiaires tirés au sort, ont été accessibles pour l'équipe *CONSTANCES* via l'application Web (selon ses droits d'accès spécifiques).

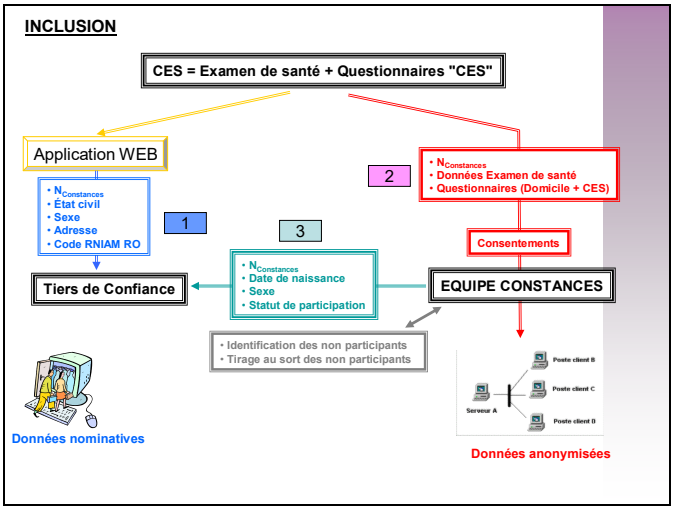
Pour les participants, chaque CES adresse à l'équipe *CONSTANCES*, les données ou documents suivants : résultats de l'examen de santé, questionnaires (remplis à domicile et au CES) identifiées par le $N_{Constances}$.

Enfin, chaque CES envoie à l'équipe *CONSTANCES* les consentements, dans un envoi séparé du précédent.

L'équipe *CONSTANCES* est en mesure d'identifier les non participants par différentiel. Dans cet ensemble de bénéficiaires (n'ayant pas retourné le coupon-réponse), un échantillon de non participants a été tiré au sort et fait l'objet d'un suivi dans les bases de données nationales (SNGC, SNDS).

A la suite, l'équipe *CONSTANCES* adresse au Tiers de confiance les données suivantes : $N_{Constances}$, date de naissance, sexe et l'information quant à l'appartenance à l'échantillon de participants ou l'échantillon de non participants de chaque personne.

Rappel : les sujets ayant exprimé leur refus de participer à la cohorte témoin dans le coupon-réponse du courrier d'invitation sont exclus de cette procédure.



5.8 GESTION DES CANDIDATURES SPONTANÉES

Le Tiers de confiance, pour les candidatures spontanées identifiées, adresse les données suivantes : état civil, sexe, Code RNIAM Régime obligatoire, associées au $N_{Constances}$, à la DSI (Cnav) afin que cette dernière puisse effectuer une recherche du $NIR_{individuel}$ de ces personnes. À la suite, la DSI (Cnav) intègre la correspondance $NIR_{individuel}$-$N_{Constances}$ à la base de correspondance existante.

Ensuite, la DSI (Cnav) effectue :

- un envoi de données vers l'équipe *CONSTANCES* comprenant les informations : $N_{Constances}$, date et lieu de naissance, sexe et Code RNIAM Régime obligatoire ;
- un envoi de données vers le CEN (CNAM) et les SLM comprenant : $NIR_{individuel}$, $N_{Constances}$, état civil, sexe et Code RNIAM Régime obligatoire. Ces derniers :
    - recherchent la correspondance entre le $NIR_{individuel}$ et le $NIR_{ouvreur\ de\ droit}$ pour ces personnes ;
    - ajoutent la correspondance $N_{Constances}$-Identifiants bénéficiaire ($NIR_{ouvreur\ de\ droit}$, date de naissance et sexe du bénéficiaire) à leur base de correspondance existante.

5.9 SAISIE DES DONNÉES

Plusieurs formats de saisie ont été utilisés en fonction du type de données : mise en place d'une application Web pour le suivi de la participation ; création d'adaptations dans le système informatique usuel des CES pour les données recueillies dans le cadre de l'examen de santé ; transmission des questionnaires sous format papier à l'équipe *CONSTANCES* pour traitement en LAD.

5.10 SUIVI PASSIF (INTERROGATION DES BASES DE DONNÉES NATIONALES)

En dehors des données recueillies lors de la phase d'inclusion, un suivi passif est mis en place au travers des bases de données nationales : le SNGC géré par la DSI (Cnav), où les données individuelles d'un sujet sont chaînées sur la base de son $NIR_{individuel}$ (quel que soit son statut d'ouvreur ou d'ayant droit) ; le SNDS géré par le CENTI (CNAM), où les données individuelles sont chaînées sur la base des variables suivantes : $NIR_{ouvreur\ de\ droit}$ + date de naissance et sexe du bénéficiaire (qu'il soit ouvreur de droit ou ayant droit). Elles sont anonymisées en deux étapes par l'algorithme FOIN. La première étape (FOIN 1) se fait par les CTI au niveau régional, la seconde (FOIN 2) par le CENTI au niveau national.

Dans le cas des ouvreurs de droit, le $NIR_{ouvreur\ de\ droit}$ est leur $NIR_{individuel}$. Dans le cas des ayants droit, il existe une correspondance entre leur $NIR_{individuel}$ et leur $NIR_{ayant\ droit}$ (= $NIR_{individuel}$ de l'ouvreur de droit).

5.10.1 Circuit des données CONSTANCES avec la Cnav

À la suite de chaque vague d'inclusion, l'équipe *CONSTANCES* a adressé à la DSI (Cnav) la liste des $N_{Constances}$, et le statut de participation correspondant, pour lesquels un transfert de données doit être effectué. La DSI (Cnav) extrait de son système d'information, en plus du statut vital (date de décès issue du SNGI), les données socioprofessionnelles (issues du SNGC et telles que définies par ailleurs) pour chaque $N_{Constances}$ (quel que soit le statut de participation) et les adresse à l'équipe *CONSTANCES* qui les intègre à la base de données.

Finalement, après chaque vague d'inclusion, la DSI (Cnav) a détruit toutes les données ne concernant ni la cohorte de participants ni la cohorte de non participants.

Chaque année, la DSI (Cnav) adresse à l'équipe *CONSTANCES* les données socioprofessionnelles et statuts vitaux des participants et non participants en utilisant le $N_{Constances}$ comme identifiant.

5.10.2 Circuit des données avec la CNAM

À la suite de chaque vague d'inclusion, l'équipe *CONSTANCES* a adressé au CEN (CNAM) la liste des $N_{Constances}$, et le statut de participation correspondant, pour lesquels un transfert de données doit être effectué.

Le CEN (CNAM) détruit, après chaque vague d'inclusion, toutes les données (nécessaires à la génération de $N_{FOIN1\ SNDS}$) ne concernant ni la cohorte de participants ni la cohorte de non participants qui lui avaient été transmises par la DSI (Cnav) après chaque tirage au sort.

Puis, à partir de la table de correspondance $N_{Constances}$-Identifiants bénéficiaire ($NIR_{ouvreur\ de\ droit}$, date de naissance et sexe du bénéficiaire) mise en place à l'inclusion, le CEN (CNAM) adresse la correspondance $N_{Constances}$-$N_{FOIN1\ SNDS}$ au CENTI (CNAM). A la suite, ce dernier applique l'anonymisation FOIN2 pour chacun des participants (obtention d'un $N_{FOIN2\ SNDS}$).

Enfin, le CENTI (CNAM) extrait de son système d'information les données médico-administratives (telles que définies par ailleurs) pour chaque $N_{Constances}$ (quel que soit le statut de participation) et les adresse chaque année à l'équipe *CONSTANCES* qui les intègre à la base de données.

5.10.3 Circuit des données avec les SLM et la Camieg

À la suite de chaque vague d'inclusion : l'équipe *CONSTANCES* a adressé à chaque SLM et à la Camieg la liste des $N_{Constances,}$ et le statut de participation correspondant, pour lesquels un transfert de données doit être effectué.

Chaque SLM et la Camieg détruisent, après chaque vague d'inclusion, toutes les données (nécessaires à la génération de $N_{FOIN1\ SNDS}$) ne concernant ni la cohorte de participants ni la cohorte de non participants qui lui avaient été transmises par la DSI (Cnav) après chaque tirage au sort.

Puis, à partir de leur table de correspondance mise en place à l'inclusion, les SLM et la Camieg adressent la correspondance $N_{Constances}$-Identifiants bénéficiaire ($NIR_{ouvreur\ de\ droit}$, date de naissance et sexe du bénéficiaire) à la MGEN qui a accepté de centraliser les fichiers. Cette dernière concatène en un fichier unique l'ensemble des informations la concernant avec celles transmises par les SLM et la Camieg, puis transmet ce fichier au CEN (CNAM).

Le CEN (CNAM), après avoir appliqué l'anonymisation FOIN1, adresse $N_{Constances}$-$N_{FOIN1\ SNDS}$ au CENTI (CNAM). À la suite, ce dernier applique l'anonymisation FOIN2 pour chacun des participants (obtention d'un $N_{FOIN2\ SNDS}$).

Enfin, le CENTI (CNAM) extrait de son système d'information, les données médico-administratives (telles que définies par ailleurs) pour chaque $N_{Constances}$ (quel que soit le statut de participation) et les adresse à l'équipe *CONSTANCES* qui les intègre à la base de données.

Chaque année, la CNAM adresse à l'équipe *CONSTANCES* les données médico-administratives des participants et non participants en utilisant le $N_{Constances}$ comme identifiant.

5.10.4 Suppression de données nominatives

Après l'inclusion, le Tiers de confiance a détruit toutes les données ne concernant pas les participants (y compris celles des personnes appartenant à la cohorte de non participants, dont le suivi doit être strictement anonyme).

Ainsi, après l'inclusion, le Tiers de confiance dispose uniquement des données concernant les participants (ceci afin d'assurer le suivi de ces personnes).

5.11 SUIVI ACTIF – INTERROGATION DES PARTICIPANTS PAR AUTOQUESTIONNAIRE

Chaque participant (sauf ceux ayant émis le souhait d'être exclus de la cohorte ou décédés) reçoit tous les ans un autoquestionnaire de suivi lui permettant également de mettre à jour les données nominatives (changement de nom d'usage, d'adresse, de numéro de téléphone). Les volontaires pourront aussi indiquer ces changements *via* l'application Web ou le Numéro Vert.

5.11.1 Envoi des auto-questionnaires

L'équipe *CONSTANCES* adresse au Tiers de confiance la liste des $N_{Constances}$ à contacter (associés au code Alliage correspondant).

À partir de la base de données nominatives, le Tiers de confiance adresse à un imprimeur-routeur les données des participants correspondants : $N_{Constances}$, civilité, nom, prénom, adresse et code Alliage.

À la suite, l'imprimeur-routeur :

- imprime les lettres d'accompagnement, les fiches de suivi, les auto-questionnaires de suivi, les journaux d'information, les enveloppes T et les lettres porteuses ;
- personnalise les lettres d'accompagnement en indiquant : civilité, nom, prénom, adresse, ainsi que le code Alliage sous forme de code à barres ;
- personnalise les fiches de suivi et les autoquestionnaires de suivi en pré-imprimant le numéro non signifiant $N_{Constances}$;
- adresse ces documents (par lettre porteuse) à chaque participant.

5.11.2 Traitement des retours de l'autoquestionnaire et de la fiche de suivi

L'Imprimerie nationale est destinataire de l'ensemble des auto-questionnaires de suivi. Au fur et à mesure de leur réception, ils sont préparés (ouverture enveloppe, massicotage, préparation des plis…) puis mis dans la chaîne de traitement LAD (numérisation,

reconnaissance, vidéocodage, archivage…). En fin de traitement, les données sont intégrées à la base de données.

Les versions papiers des documents seront détruites après un laps de temps suffisant.

5.11.3 Gestion et traitement des adresses

A l'inclusion : après l'inclusion au CES, le Tiers de confiance est détenteur, pour chaque $N_{Constances}$, des données suivantes : état civil, sexe, adresse, Code RNIAM Régime obligatoire et statut de participation.

Pour les participants, l'équipe *CONSTANCES* génère un N_{POSTE} (Numéro de transfert spécifique La Poste) et assure la correspondance entre ce numéro et le $N_{Constances}$. Elle adresse un fichier N_{POSTE}–$N_{Constances}$ au Tiers de confiance.

Le Tiers de confiance envoie un fichier constitué de : nom, prénom et adresse (associés à N_{POSTE}) à Mediapost (La Poste) afin que cette dernière effectue les trois traitements suivants : Optimis 1 (normalisation des adresses), Optimis 2 (correction des PND), Optimis 3 (enrichissement du fichier, dont géocodage = X, Y, code Iris). Après cette opération, Mediapost (La Poste) adresse d'une part, le résultat d'Optimis 1 et d'Optimis 2 (associé au N_{POSTE}) au Tiers de confiance et d'autre part, le résultat d'Optimis 3 (associé au N_{POSTE}) à l'équipe *CONSTANCES*.

Au cours du suivi – Gestion des PND : l'envoi de l'autoquestionnaire de suivi est effectué en utilisant le dispositif Alliage mis en place par La Poste. Ce dispositif permet d'identifier rapidement les PND et d'éliminer les retours physiques (recyclage des plis PND).

Concrètement, l'identification des PND s'opère par le biais d'envois effectués grâce à un logo apposé sur les enveloppes qui permet au facteur d'identifier au cours de sa tournée, les courriers bénéficiant d'un suivi des PND par Alliage ; un code à barres (combinaison d'un numéro expéditeur et d'un numéro destinataire) est apposé sur les lettres d'accompagnement. Pour chaque PND retourné au dépôt, le code à barres est « flashé » et les données suivantes : code à barres, date et établissement de flashage, sont retournées à l'expéditeur (équipe *CONSTANCES*) sous format informatique sécurisé par le Service National de l'adresse (SNA) de La Poste.

A la réception du fichier informatisé des PND, l'équipe *CONSTANCES* adresse au Tiers de confiance la concordance $N_{Constances}$–N_{POSTE} correspondante.

Puis, le Tiers de confiance envoie un fichier constitué de : nom, prénom et adresse erronée (associés à N_{POSTE}) à Mediapost (La Poste) afin que ce dernier effectue les trois traitements Optimis. Après cette opération, Mediapost adresse d'une part, le résultat d'Optimis 1 et d'Optimis 2 (associé au N_{POSTE}) au Tiers de confiance et d'autre part, le résultat d'Optimis 3 (associé au N_{POSTE}) à l'équipe *CONSTANCES*.

5.12 VALIDATION DES ÉVÉNEMENTS DE SANTÉ - ASPECTS OPÉRATIONNELS

5.12.1 Données du consentement

Le formulaire de consentement proposé aux sujets de *CONSTANCES* à l'inclusion précisait explicitement que l'équipe pourra avoir accès aux bases de données médicales, et en cas de besoin contacter directement les sujets volontaires, ou les professionnels de santé et les hôpitaux les ayant pris en charge. Le formulaire de consentement permet aux sujets de refuser tout ou partie de ces procédures. Il comporte les numéros de téléphone de contact.

Ce formulaire de consentement a été saisi en traitement LAD. Les informations nominatives nécessaires à la procédure de validation des événements de santé sont conservées dans une base de données indépendante. Les droits d'accès à cette base sont strictement restreints aux nécessités de la procédure de validation des événements de santé.

5.12.2 Repérage des événements de santé

La validation directe individuelle des événements est déclenchée soit à partir des déclarations des individus (via les auto-questionnaires) ou des données collectées par les CES, soit à partir

des données fournies par le SNDS. Le croisement systématique, à l'aide d'algorithmes spécifiques à chaque pathologie, des informations provenant de ces différentes sources de données amène à repérer les pathologies d'intérêt et à déclencher une enquête.

5.12.3 Recueil d'informations auprès des volontaires et des professionnels de santé

Les volontaires qui ont accepté d'être contactés par téléphone ont appelés au numéro qu'ils ont indiqué sur le consentement initial (ou ultérieurement lors d'une déclaration de changement d'adresse). Dans le cas où les volontaires n'ont pas accepté d'être contactés par téléphone, mais ont accepté le recueil de leurs données issues des Caisses d'assurance maladie, des hôpitaux et des professionnels de santé, les hôpitaux ou professionnels de santé ayant participé aux soins du volontaire (données issues des remboursements de soins) sont contactés directement.

Une plateforme téléphonique est chargée de recueillir des informations complémentaires sur la pathologie repérée du volontaire, ainsi que tous types de documents (résultats anatomo-pathologiques, comptes-rendus opératoires ou d'hospitalisation…) permettant de confirmer cette pathologie repérée et de la coder le plus précisément possible selon les classifications en vigueur (CIM10, type histologique ou grades spécifiques pour les cancers, etc.).

5.12.4 Validation des événements de santé

Les données médicales recueillies, regroupées sous forme de Dossiers médicaux anonymisés, sont soumises à des Comités d'experts indépendants pour validation des événements (Comités Externes de Validation), composés de spécialistes du domaine pour chaque pathologie d'intérêt.

Récapitulatif des numéros non identifiants et données nominatives utilisés

	Tiers de confiance (1)	Équipe CONSTANCES	CES	DSI (Cnav) (2)	CEN (CNAM) (2)	SLM & Camieg (2)	CENTI (CNAM)	Mediapost (La Poste)
Numéros non signifiants								
$N_{Constances}$	X	X	X	X	X	X	X	
$N_{FOIN1\ SNDS}$					X		X	
$N_{FOIN2\ SNDS}$							X	
$N_{CépiDc}$		X		X				
N_{POSTE}	X	X						X
$N_{DIAGNOSTIC}$		X						
Données nominatives								
NIR (individuel ou ayant droit)			X	X	X	X		
Nom patronymique, prénom	X	X (3)	X	X	X	X		X
Date de naissance	X	X	X	X	X	X		
Sexe	X	X	X	X	X	X		
Lieu de naissance	X	X	X	X	X	X		
Adresses	X		X		X	X		X

(1) : Après l'inclusion, tous les numéros non signifiants et données nominatives des non participants seront détruits

(2) : Après l'inclusion, toutes les données ne concernant ni l'échantillon de participants, ni l'échantillon de non participants seront détruites
(3) : Uniquement dans le cadre des activités de la plateforme de validation des diagnostics
☐ : Organisme ou équipe en charge de la création

6 RÉFÉRENCES

ANAES-Agence Nationale d'Accréditation et d'Évaluation en Santé. Diagnostic de l'insuffisance rénale chronique de l'adulte. Recommandations et références professionnelles. Septembre 2002.

ARME. Incidence et prévalence de différentes maladies. Bordeaux, ARME Pharmacovigilance – 2007.

Austin MA, Criqui MH, Barrett *et al*. The effect of response bias on the odds-ratio. Am J Epidemiol, 1981; 114, 137-143.

Berkman LF, Melchior M, Chastang JF, Niedhammer I, Leclerc A, Goldberg M. Social integration and mortality: A prospective study of French men and women employees of Electricity of France-Gas of France, the Gazel Cohort. Am J Epidemiol 2004; 159: 167-74.

Berr C, Derriennic F, Zins M. Cohortes et banques de données biologiques (Éditorial). Rev Epidemiol Santé Publ 2003; 51: 97-98.

Boll TJ, Reitan RM. Effect of age on performance of the Trail Making Test. Percept Mot Skills 1973; 36: 691-4.

Borkowski JG, Benton AL, Spreen O Word fluency and brain damage. Neuropsychologica 1967; 5: 135-140.

Bowling A, Dieppe P. What is successful ageing and who should define it? BMJ 2005; 331: 24-31.

Cardebat D, Doyon B, Puel M, Goulet P, Joanette Y. Formal and semantic lexical evocation in normal subjects. Performance and dynamics of production as a function of sex, age and educational level. Acta Neurol Belg. 1990; 90: 207 – 17.

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

Criqui MH. Response bias and risk ratios in epidemiologic studies. Am J Epidemiol, 1979; 109, 394-399.

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

Dartigues JF, Alpérovitch A. Épidémiologie et vieillissement. In : Valleron AJ (Ed.) : « *Épidémiologie : conditions de son développement, et rôle des mathématiques* ». Rapport sur la Science et la Technologie n° 23, Comité RST de l'Académie des sciences. Éditions EDP Sciences, 2006.

Dartigues JF, Gagnon M, Michel P, *et al*. Le programme de recherche Paquid sur l'épidémiologie de la démence. Méthodes et résultats initiaux. Rev Neurol 1991; 147: 225-230.

Diez Roux AV. Places, people and health. Am J Epidemiol 2002; 155: 516-19.

Doll R, Peto R, Wheatley K, Gray R, Sutherland I. Mortality in relation to smoking: 40 years' observations on male British doctors. Br Med J 1994; 309 : 901–11.

DREES « Coordination des exploitations de l'enquête HID », Dourdan, 18-19 octobre 2001.

DST. Programme MATGÉNÉ - État d'avancement. Saint Maurice, InVS, Mars 2004.

Dufouil C, Alpérovitch A, Ducros V, Tzourio C. Homocysteine, white matter hyperintensities, and cognition in healthy elderly people. Ann Neurol 2003; 53: 214-21.

Eltinge JL, Yansaneh IS. Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the U.S. consumer expenditure survey. Survey Methodol, 1997, 23: 33-40.

Fender P, Weill A. Épidémiologie, santé publique et bases de données médico-tarifaires. (Éditorial) Rev Epidemiol Santé Publique 2004 ; 52 : 113-117.

Folstein M, Anthony JC *et al*. The meaning of cognitive impairment in the elderly. J Am Geriatr Soc 1985; 33: 228-35.

Froissart M, Rossert J, Jacquot C, Paillard M, Houillier P. Predictive performance of the modification of diet in renal disease and Cockcroft-Gault equations for estimating renal function. J Am Soc Nephrol. 2005; 16(3): 763-73.

Gazel: www.gazel.inserm.fr.

Giampaoli S, Ferrucci L, Cecchi F, *et al*. Hand-grip strength predicts incident disability in non-disabled older men. Age and Ageing 1999; 28: 283-288.

Goldberg M, Chastang JF, Leclerc A, Zins M, Bonenfant S, Bugel I, Kaniewski N, Schmaus A, Niedhammer I, Piciotti M, Chevalier A, Godard C, Imbernon E. Socioeconomic, demographic, occupational and health factors associated with participation in a long-term epidemiologic survey. A prospective study of the French Gazel cohort and its target population. Am J Epidemiol 2001; 154: 373-84.

Goldberg M, Chastang JF, Zins M, Niedhammer I, Leclerc A. Attrition during follow-up: health problems are the strongest predictors. A Study of the Gazel Cohort. J Clin Epid. 2006b; 59: 1213-1221.

Goldberg M, Imbernon E. The use of job exposure matrices for cancer epidemiology research and surveillance. Arch Public Health 2002; 60: 173-85.

Goldberg M, Leclerc A *et al*. La cohorte Gazel, laboratoire épidémiologique. Bilan des cinq premières années (1989-1993) de suivi des 20 000 volontaires d'Électricité de France - Gaz de France. Paris : Éditions Inserm – Collection Grandes Enquêtes. 1994.

Goldberg M, Leclerc A, Bonenfant S, Chastang JF, Schmaus A, Kaniewski N, Zins M. Cohort profile: the GAZEL Cohort Study. Int J Epid. 2007; 36: 32-39.

Goldberg M, Luce D. Les effets de sélection dans les cohortes épidémiologiques. Nature, causes et conséquences. Rev Epidemiol Santé Publique 2001; 49: 477-92.

Goldberg M. Les bases de données d'origine administrative peuvent-elles être utiles pour l'épidémiologie ? (Éditorial) Rev Epid Santé Publ, 2006a, 54: 297-303.

Greenland S. Response and follow-up bias in cohort studies. Am J Epidemiol, 1977; 106, 184-187.

Grober E, Buschke H, Cystal H et al. Screening for dementia by memory testing. Neurol., 1998; 38: 900-903.

Guralnik JM, Simonsick EM, Ferrucci L, *et al*. A short physical performance battery assessing lower extremity function: association with self reported disability and prediction of mortality and nursing home admission. J Gerontol 1994; 49: M85-94.

Horak FB, Shupert CL, Mirka A. Components of postural dyscontrol in the elderly: a review. Neurobiol Aging. 1989; 10: 727-38.

Hyde M, Wiggins RD, Higgs P, Blane D. A measure of quality of life in early old age: The theory, development and properties of a needs satisfaction model (CASP-19). Ageing & Mental Health 2003; 7: 186-194.

Santé publique France. Estimation des taux de prévalence des anticorps anti-VHC et des marqueurs du virus de l'hépatite B chez les assurés sociaux du régime général de France métropolitaine, 2003-2004. Analyse descriptive. Saint Maurice, Santé publique France, 2005.

Santé publique France. Rapport annuel 2006, Saint Maurice, Santé publique France, 2007.

Jouven X, Empana JP, Schwartz PJ, Desnos M, Courbon D, Ducimetiere P. Heart-rate profile during exercise as a predictor of sudden death. N Engl J Med 2005; 352: 1951-8.

Kaufman AS, Reynolds CR, McLean JE. Age and WAIS-R intelligence in a national sample of adults in the 20- to 74-year age range: A cross-sectional analysis with educational level controlled. Intelligence 1989; 13: 235–253.

Kuh D, Ben Schlomo Y (Eds.). A lifecourse approach to chronic disease epidemiology. Oxford:Oxford University Press, 1997.

Lawton MP, Brody EM. Assessment of older people: self-maintaining and instrumental activities of daily living. Gerontologist 1969 ; 9: 179-86.

Leclerc A, Landre MF, Chastang JF, Niedhammer I, Roquelaure Y, and the Study Group on Repetitive Work. Upper-limb disorders in repetitive work. Scand J Work and Environmental Health 2001; 27: 268-278.

Levey AS, Eckardt KU, Tsukamoto Y, Levin A, Coresh J, Rossert J, Zeeuw D, Hostetter TH, Lameire N, Eknoyan G. Definition and classification of chronic kidney disease: a position statement from Kidney Disease: Improving Global Outcomes (KDIGO). Kidney Int. 2005 Jun; 67(6): 2089-100.

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

Mathias S, Nayak U, Isaacs B. Balance in elderly patients: the "Get-up and Go" test. Arch Phys Med Rehabil 1986; 67: 387-9.

Miner T, Ferraro FR. The role of speed of processing, inhibitory mechanisms, and presentation order in Trail-Making test Performance. *Brain and cognition*. 1998; 38: 246-53.

Mitrushina MN, Boone KB, D'Elia LF. Handbook of Normative Data for Neuropsychological Assessment. New York: Oxford University Press. 1999.

National Kidney Foundation. K/DOQI Clinical practice guidelines for chronic kidney disease : Evaluation, Classification and Stratification. Am J Kidney Dis 2002, 39 :S1-S266 (suppl 1).

Niedhammer I, Tek ML, Starke D, Siegrist J. Effort–reward imbalance model and self-reported health: cross-sectional and prospective findings from the Gazel cohort Social Science & Medicine. 2004; 58: 1531-1541.

Nohr EA, Frydenberg M, Henriksen TB, Olsen J. Does low participation in cohort studies induces bias? Epidemiology. 2006; 17: 413-8.

Oppenheimer GM. Becoming the Framingham Study. Am J Pub Health 2005; 95: 602-610.

Pirracchio R. () Nouveautés en modélisation non paramétrique - Apports du Super Learner. Revue d'épidémiologie et de santé publique, 2014:s171-s172.

Polley EC, van der Lann MJ. Super learner in prediction. U.C. Berkeley Division of Biostatistics Working Paper Series n°266, 2010.

Remontet L, Buemi A, Velten M, Jougla E, Estève J. Évolution de l'incidence et de la mortalité par cancer en France de 1978 à 2000. Saint Maurice, InVS, 2002.

Ribet C, Melchior M, Lang T, Zins M, Goldberg M, Leclerc A. Caractérisation et mesure de la situation sociale dans les études épidémiologiques. Rev Epid Santé Publ, 2007; 55: 285–295.

Roquelaure Y, Ha C, Touranchet A, Imbernon E, Goldberg M. Surveillance des TMS dans les entreprises des Pays de la Loire : Synthèse des résultats en 2002. Saint Maurice, InVS, 2005.

Share : http://www.share-project.org

Sheikh K, Mattingly S. Investigating non-response bias in mail surveys. J Epidemiol Comm Health, 1981; 35, 293-296.

Shkuratova N, Morris ME, Huxham F. Effects of age on balance control during walking. Arch Phys Med Rehabil. 2004; 85(4): 582-8.

Siegrist J, Pollack CE, Knesebeck OVD. Social productivity and well-being of older people: a sociological exploration. Social Theory & Health 2004; 2: 1-17.

Siegrist J. Effort-reward Imbalance at Work and Health. In: Research in Occupational Stress and Well Being, Historical and Current Perspectives on Stress and Health. In: P. Perrewe (Eds). JAI Elsevier, London. Vol. 2, pp 261-291, 2002.

Singh-Manoux A, Ferrie JE, Lynch JW, Marmot M. The role of cognitive ability (intelligence) in explaining the association between socioeconomic position and health: Evidence from Whitehall II prospective cohort study. Am J Epidemiol 2005; 161: 831-9.

Société de Néphrologie. Évaluation de la fonction rénale et de la protéinurie pour le diagnostic de la maladie rénale chronique. Recommandations pour la pratique clinique. Groupe de travail de la Société de Néphrologie. Néphrologie et Thérapeutique 2009; 5(4): 302-5.

Three-Cities study group. Vascular factors and risk of dementia: design of the Three-City Study and baseline characteristics of the study population. Neuroepidemiology 2003; 22: 316-325.

UK Presidency of the EU. Tackling health inequalities: Governing for Health Summit. Londres, 17-18 octobre 2005.

Valleron AJ (Ed.) : « *Épidémiologie : conditions de son développement, et rôle des mathématiques* ». Rapport sur la Science et la Technologie n° 23, Comité RST de l'Académie des sciences. Éditions EDP Sciences, 2006.

Van der Linden M, Coyette F, Poitrenaud J, Kalafat M, Calicis F, Wyns C, Adam S et les membres du GREMEM. L'épreuve de rappel libre/rappel indicé à 16 items (RL/RI-16. L'évaluation des troubles de la mémoire. Van der Linden M et les membres du GREMEM (eds). Solal Éditeur, Marseille - 2004.

Wechsler D. Manual for the Wechsler Adulte Intelligence Scale-Revised. New-York: Psychological Corporation, 1981.

Zins M, Leclerc A, Goldberg M. The French GAZEL Cohort Study: 20 years of epidemiologic research. Advances in Life Course Research 2009; 14: 135-146.