# Cohorte Constances :

# Document de synthèse méthodologique pour la correction de la non-réponse

# Pondérations 2013-2017

***Version – 22/02/2022***

Ce document est la synthèse d'une documentation plus large transmise au moment de la mise à disposition des pondérations par l'Equipe Constances. Cette synthèse explique la construction des pondérations pour

les années 2013, 2014, 2015, 2016 et 2017 dans la cohorte Constances et fournit quelques éléments sur leur utilisation.

# 1 Introduction

La cohorte Constances est une cohorte épidémiologique généraliste (1) dont l'inclusion s'étale sur plusieurs années. Elle a été construite pour constituer une infrastructure de recherche afin de faciliter des travaux d'épidémiologie analytique, et permettre des études de santé publique et de surveillance épidémiologique. C'est dans ce cadre que de nombreux programmes de santé publique vont s'appuyer sur les données de Constances (2). Constances doit donc aussi permettre d'inférer des prévalences à sa population cible. C'est pour cet objectif que des pondérations sont calculées et mises à disposition.

Néanmoins, Constances ne s'inscrit pas dans les standards des enquêtes descriptives à visée représentative pour deux principales raisons :

1- chaque année, le plan de sondage diffère en raison de contraintes logistiques diverses selon les Centres d'Examen de Santé (CES) et les régimes d'affiliation
   ➔ une pondération par année est donc calculée et il n'est pas possible de combiner plusieurs années de pondérations sans hypothèses supplémentaires.
2- le recueil de données de santé mesurées implique que les personnes se rendent dans un CES, ce qui conduit à un taux de participation très faible (environ 6-7%) généralement observé pour ce type d'enquête (3) ; en effet, les personnes participantes doivent accepter de se déplacer dans un CES qui peut être éloigné de leur domicile et difficile d'accès
   ➔ la faible participation attendue peut entrainer des biais importants qu'il faut minimiser autant que possible ; d'où la nécessité de modéliser les probabilités de participation.

Avant toute utilisation des pondérations et interprétation des résultats pondérés il convient :

- d'avoir une certaine confiance dans le modèle pour considérer que les pondérations mises à disposition permettent de calculer des estimations représentatives
- de considérer que le faible taux de participation entraîne une plus grande dispersion pour les pondérations finales par rapport aux poids de sondage initiaux.

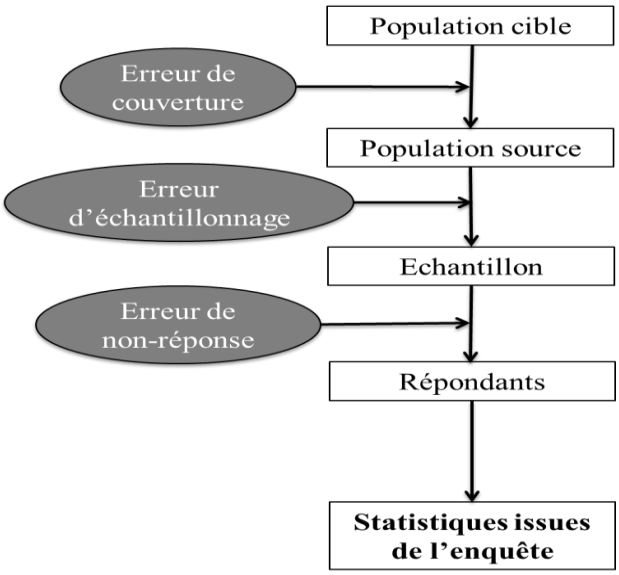
# 2 L'essentiel sur les pondérations issues des enquêtes par sondage avec traitement de la non-réponse totale par repondération

La bonne compréhension et l'utilisation de ces pondérations nécessite une certaine familiarité avec les enquêtes par sondage. Si ce n'est pas le cas, il faudra se référer à des cours ou des livres de référence (5-7). Cette partie propose quelques points d'ancrage pour les néophytes. Il est tiré du document dont la référence suit (8).

## 2.1 De la population cible à l'échantillon de répondants

Le processus qui conduit de la population d'intérêt (population cible) à l'échantillon final de répondants sur lequel des statistiques seront calculées est représenté dans la figure 3-1.

***Figure 2-1 : De la population cible à l'échantillon de répondants***




Dans la plupart des cas, il génère des erreurs, excepté lorsque l'ensemble de la population cible constitue l'échantillon final : on est alors dans le cadre d'un recensement complet et les statistiques estimées ne sont pas entachées par les erreurs liées au fait que seule une partie de la population cible est enquêtée.

Dans le cas où on cherche à inférer des statistiques à une population cible à partir d'une partie de la population, ou échantillon, les estimations comprennent nécessairement des erreurs d'échantillonnage et sont aussi souvent entachées par des erreurs de couverture et des erreurs de non-réponse.

La constitution de l'échantillon est facilitée lorsqu'il existe une base de sondage c'est-à-dire une liste qui permet d'identifier tous les individus de la population source.

## 2.2 Les différents types d'erreur et les biais introduits

Le tableau 3-2 présente les différentes erreurs précédemment évoquées ainsi que les potentiels biais engendrés sur les estimations.

***Figure 2-2 : Typologie des erreurs***

| Type d'erreur | Définition | Conséquences sur les estimations |
|---|---|---|
| **Erreur de couverture** | La population source ne correspond pas exactement à la population cible. Sur-couverture : la base de sondage couvre plus de personnes que la population cible ; Sous-couverture : la base de sondage n'inclut pas l'ensemble de la population cible. | Si les unités non-incluses (sous-couverture) ou incluses à tort (sur-couverture) diffèrent de la population cible, les erreurs de couverture peuvent entrainer un biais dans les estimations de prévalence. En général, la sur-couverture est plus facile à traiter que la sous-couverture. |
| **Erreur d'échantillonnage** | Dans un échantillonnage probabiliste (notre cas), les unités constituant la population source sont sélectionnées selon un processus aléatoire connu : les unités de la base de sondage (ici, l'individu) sont sélectionnées par tirage au sort pour constituer l'échantillon.<br>L'erreur d'échantillonnage correspond à l'écart entre la valeur du paramètre estimé sur l'échantillon et sa valeur sur la population source. | Biais d'un estimateur : différence entre la moyenne de toutes les estimations obtenues sur tous les échantillons qui pourraient être sélectionnés selon un certain plan de sondage et la valeur de ce paramètre dans la population source.<br>Soit $\theta$ le paramètre qu'on cherche à estimer et $\hat{\theta}$ un estimateur de $\theta$. Le biais de $\hat{\theta}$ lié à un plan de sondage *(ps)* s'exprime par : $Biais_{ps}(\hat{\theta}) = E_{ps}(\hat{\theta}) - \theta$<br>L'échantillonnage probabiliste conduit à des estimations (asymptotiquement) sans biais quand le paramètre d'intérêt est une prévalence ou une moyenne.<br>L'erreur d'échantillonnage ne comprend donc pas de biais ; en revanche, elle est composée de variance, qu'on appelle fluctuation d'échantillonnage. |
| **Erreur de non-réponse** | Ecart entre la valeur du paramètre estimé sur l'échantillon de répondants et la valeur qui serait estimée sur l'échantillon des individus tirés au sort par le plan de sondage. L'erreur de non-réponse est conditionnelle à l'échantillon tiré au sort[1] et est due au fait que les individus constituant l'échantillon sélectionné selon un certain plan de sondage ne répondent pas tous à l'enquête. La non-réponse est *totale* lorsque la personne enquêtée ne répond à aucune question de l'enquête. On a alors recours à la repondération pour traiter la non-réponse totale. | Biais de non-réponse : différence entre la moyenne des estimations obtenues sur tous les échantillons de répondants obtenus avec un processus de non-réponse répété un nombre infini de fois et l'estimation calculée sur l'échantillon complet.<br>Dans le cas de non-réponse, le processus qui conduit de l'échantillon complet à l'échantillon de répondants est inconnu. Le biais de non-réponse dépend de la variable d'intérêt étudiée : c'est le produit de l'inverse de la probabilité de réponse moyenne et de la covariance entre la probabilité de réponse et la variable d'intérêt dans la population source.<br>En présence de non-réponse, la taille de l'échantillon étant plus petite que prévue, la variance des estimateurs est en général plus élevée qu'en l'absence de non-réponse. |

[1] L'écart entre la valeur du paramètre estimé sur l'échantillon de répondants et sa valeur sur la population source est également utilisé comme définition de l'erreur de non-réponse dans la littérature.

## 2.3 Lien avec les pondérations

Nous nous plaçons par la suite dans le cas où l'erreur de couverture est nulle.
Une pondération correspond à un poids qu'on affecte à chaque individu de l'échantillon de répondants dont on dispose afin que l'ensemble des couples (pondération, individu) permette de représenter la population cible ; autrement dit, afin qu'il permette d'estimer des paramètres d'intérêt (par exemple une prévalence) sans biais extrapolables à la population cible. En corollaire, la somme des pondérations de l'échantillon correspond à la taille de la population cible.

1- Si l'échantillon disponible correspond exactement à l'échantillon sélectionné par tirage au sort (absence de non-réponse), la pondération affectée à un individu correspond à l'inverse de sa probabilité d'inclusion et permet d'obtenir des estimations sans biais.

2- Si l'échantillon disponible correspond à un échantillon tiré au sort affecté de non-réponse, la pondération affectée à un individu correspond au produit de l'inverse de sa probabilité d'inclusion et de sa probabilité de réponse. La probabilité de réponse étant inconnue, il est nécessaire de l'estimer ; son estimation dépend de l'hypothèse retenue sur le mécanisme de non-réponse (MCAR, MAR, MNAR).

### 2.3.1 Traitement de la non-réponse par repondération

La classification de Rubin[2] (10) propose une typologie des mécanismes de non-réponse pour n'importe quel paramètre concernant la distribution entière de *Y*. Elle fait intervenir la notion d'information auxiliaire.
Nous considérons ici deux hypothèses :

1- Hypothèse MCAR : la probabilité de réponse est égale pour chaque individu, plus précisément elle est égale au taux de réponse observé dans l'enquête.
2- Hypothèse MAR : la probabilité de réponse doit être modélisée en fonction des variables auxiliaires X, causes communes de la probabilité de réponse et des variables d'intérêt.

Avant de modéliser la participation, il faut donc :

- Identifier les causes communes X liées à la fois à la probabilité de réponse et aux variables d'intérêt (pour Constances, les variables d'intérêt sont des variables socioprofessionnelles ou relatives à la santé) ;
- Recueillir les variables auxiliaires X (causes communes) chez les répondants et les non-répondants (pour Constances, les variables auxiliaires sont les données du Sniiram et de la Cnav) ;
- Utiliser une méthode de repondération. Dans le cadre de Constances, deux méthodes ont été utilisées :

[2] Les non-réponses peuvent être classées entre trois types :

- la non-réponse complètement aléatoire (Missing Completely At Random ou MCAR) : il y a indépendance entre la réponse *R* et la variable d'intérêt *Y* ;
- la non-réponse aléatoire (Missing At Random ou MAR) : il y a indépendance entre la réponse *R* et la variable d'intérêt *Y* conditionnellement à *X* ;
- la non-réponse non aléatoire (Missing Not At Random ou MNAR) : il n'y a pas indépendance entre *R* et *Y* conditionnellement à *X*.

Cette typologie est détaillée dans la documentation complète sur le calcul des pondérations dans Constances.

1- La méthode des scores par quantile égaux (11-13) : elle se base sur les probabilités de réponse prédites par un modèle (régression logistique par exemple) expliquant la réponse par les variables auxiliaires pour constituer des groupes homogènes de réponse (GHR). Elle consiste à trier, pour tous les individus échantillonnés, les valeurs prédites par le modèle de non-réponse et de les grouper en k groupes de taille égale puis de calculer, dans chacun de ces groupes, des taux de réponse observés. Il est recommandé de constituer entre 5 et 30 groupes de taille égale. $\hat{\delta}(X)$ est ensuite estimé par le taux de réponse observé dans chaque groupe. Le facteur correctif de non-réponse est alors égal à l'inverse du taux de réponse observé dans chaque groupe. Un article explique en détail l'application de cette méthode (14).
2- Le calage : il consiste à modifier le plus légèrement possible les poids de sondage des individus de manière à ce que la distribution des variables auxiliaires X de l'échantillon de répondants coïncide avec la distribution des variables auxiliaires X dans l'échantillon tiré au sort ou dans la population source.

### 2.3.2 Pondération finale

Le poids final est égal au produit du poids de sondage et du facteur correctif final pour la non-réponse.

# 3 Démarche générale pour le calcul des pondérations dans Constances

La population cible de Constances pour l'année A correspond aux personnes résidant dans un département Constances, affiliées au Régime Général de la Sécurité Sociale (au sens large) et âgées de 18 à 69 ans au moment de leur invitation. Le périmètre de la population cible de Constances varie d'une année à l'autre : année de naissance, CPAM, département de résidence et/ou régime d'affiliation différents.

## 3.1 Plan de sondage

La base de sondage pour l'année A correspond aux personnes du Répertoire national inter-régimes des bénéficiaires de l'assurance maladie (Rniam) géré par la CNAV, âgées de 18 à 69 ans, appartenant au Régime général (au sens strict) ou à l'une des SLM ayant signé une convention avec Constances (Camieg, LMDE, MFPs et MGEN), affiliées à l'une des CPAM couvrant un département Constances et résidant dans l'un d'entre eux. On peut considérer que la base de sondage est de très bonne qualité et que les erreurs de couverture sont minimes.
L'inclusion s'étalant sur plusieurs années, les clés NIR[3] ont été partitionnées en plusieurs groupes distincts et chaque année une base de sondage est constituée en sélectionnant les personnes appartenant au groupe de l'année en cours. Elle comprend l'âge et le sexe des personnes et est appariée avec des données du Système National de Gestion des Carrières (SNGC) géré par la CNAV afin de disposer de la dernière PCS connue et de la notion d'activité/inactivité de la personne (ces

[3] Le Numéro d'Identification au répertoire (NIR) communément appelé numéro de sécurité sociale est suivi d'une clé à deux chiffres qui permet de vérifier la validité des 13 chiffres du NIR. Ces deux chiffres sont non informatifs et peuvent être considérés comme générés aléatoirement.

deux dernières informations combinées seront appelées par la suite typologie d'activité professionnelle ou TAP).

Chaque année, un tirage au sort stratifié à probabilités inégales est réalisé selon les strates suivantes : CPAM, affiliation, âge, sexe, typologie d'activité professionnelle. Ce tirage au sort prend en compte des contraintes logistiques (pour un CES donné, nombre de personnes pouvant bénéficier d'un EPS ainsi que le nombre d'EPS à réaliser fixé par la CNAM), des contraintes liées au partenariat (chaque SLM souhaite disposer d'un échantillon suffisamment large) et des contraintes scientifiques (effectifs suffisants selon la tranche d'âge, le sexe et la Tap).

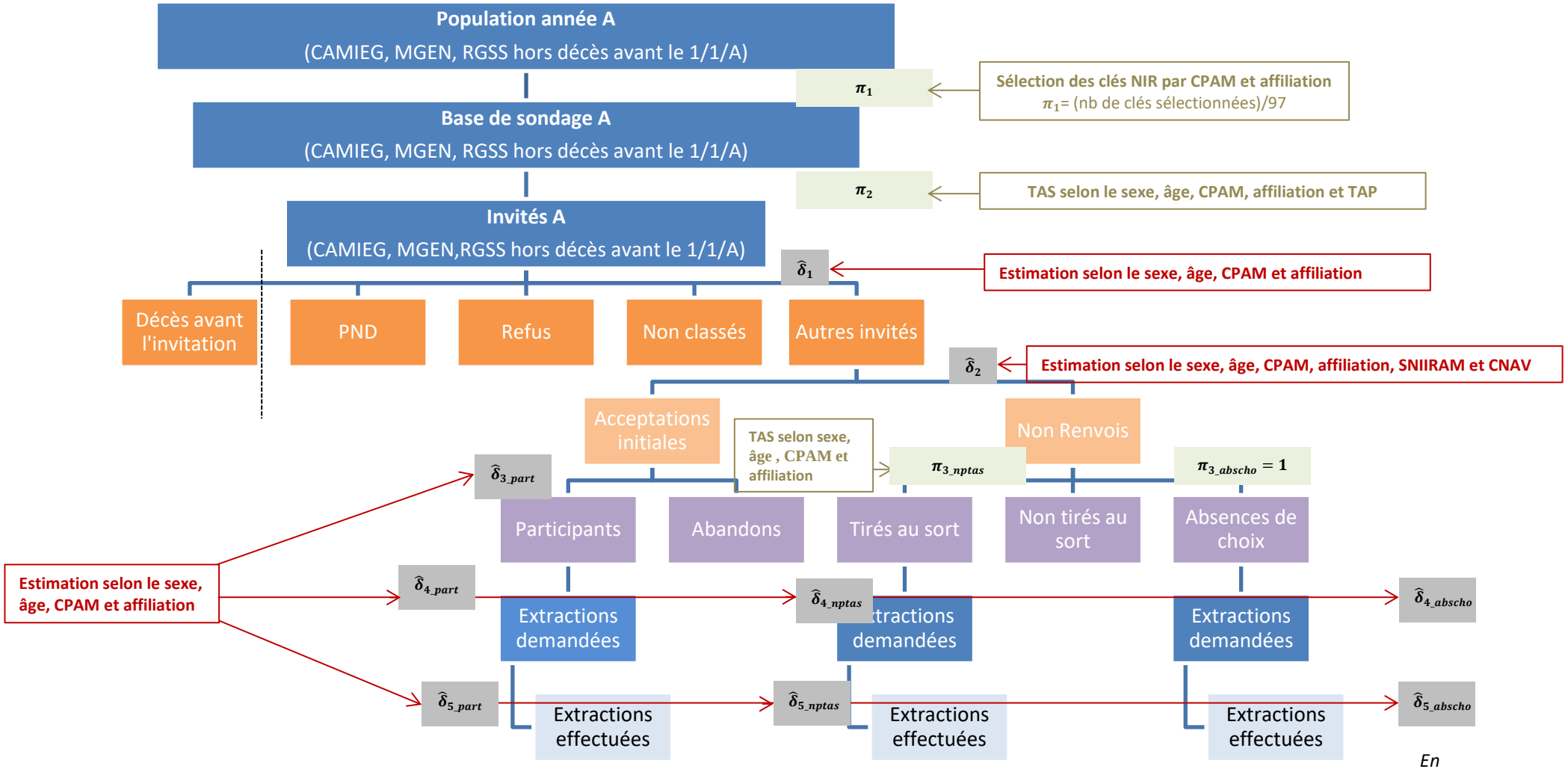
## 3.2 Flux des données

Un courrier d'invitation est envoyé à chaque personne tirée au sort. Certains courriers sont retournés par la Poste avec la mention « Pli Non Distribuable » *PND*. Apres réception du courrier d'invitation, la personne qui désire participer renvoie un coupon-réponse mentionnant son acceptation au CES correspondant à sa CPAM. Elle peut également signifier son *refus* au moyen du coupon réponse. Les personnes invitées n'ayant pas renvoyé de coupon réponse ne reçoivent pas de courrier de relance.
Les personnes ayant renvoyé un coupon réponse mentionnant leur acceptation sont classées en *acceptation initiale*. Elles sont convoquées par leur CES afin d'y passer un examen de santé. Elles sont incluses ensuite dans Constances comme *participants* si elles ont signé un consentement autorisant l'utilisation de leurs données cliniques recueillies au CES. Les personnes qui ne se présentent pas à la convocation, abandonnent au cours de l'examen ou ne signent pas de consentement sont regroupées dans la case *Abandon.*
Les personnes n'ayant pas renvoyé de coupon réponse sont représentées dans la case *Non renvoi* ainsi que les personnes ayant renvoyé un coupon réponse ne spécifiant ni l'acceptation ni le refus. Les personnes ayant des informations non consolidées sont représentés en *non classés*.

A l'issue de la collecte des données de questionnaires et des données issues des CES, il existe quatre principaux motifs de non-participation : courrier d'invitation non distribuable (8%) ; refus exprimé de participer (1%) ; abandon (1%) ; non renvoi de coupon-réponse (90%).
On constitue ensuite un échantillon de non-participants (cohorte témoin) en juin de l'année suivant l'année d'invitation. Pour les participants ayant donné leur autorisation et la cohorte témoin, un appariement avec des données du SNIIRAM et de la CNAV est réalisé. La cohorte témoin ne couvre que la quatrième catégorie de sujets pour des raisons légales (exclusion des personnes non informées et des refus). En raison du volume de données très important généré par le suivi annuel de la cohorte témoin cette dernière est constituée d'un échantillon des personnes n'ayant pas retourné de coupon réponse obtenu après un tirage au sort à probabilités inégales et stratifié sur le sexe, l'âge, la CPAM et l'affiliation (*Tiré au sort*) ; la cohorte témoin comprend également les personnes n'ayant pas précisé de choix sur le coupon réponse (Absence de choix). Pour une année donnée, la taille de la cohorte témoin est environ deux fois plus grande que celle des participants.

Le flux de données est représenté dans la figure 4-1.

***Figure 3-1 : Flux de données de la population cible à l'échantillon de participants pondérables***




*En vert : probabilité d'inclusion connue ; En brun : données de stratification utilisées pour le tirage au sort ; En gris : probabilité de réponse à estimer ; En rouge : données auxiliaires disponibles pour estimer les probabilités de réponse*

## 3.3 Pondérations

Une pondération est calculée pour chaque année A (ici 2013, 2014, 2015, 2016 et 2017). Un participant pondérable est un participant pour lequel un examen périodique de santé (ou EPS) a été réalisé et une extraction des données passives (SNIIRAM et CNAV) a pu être effectuée.
La pondération dépend de la probabilité d'inclusion (donc du poids de sondage) et de la probabilité de participation estimée pour chaque participant pondérable (facteur correctif de la non-réponse).

### 3.3.1 Poids de sondage

A partir de la population cible d'une année A, on sélectionne pour chaque couple CPAMxAffiliation un nombre de clés NIR dont la proportion est notée $\boldsymbol{\pi_1}$ ; on obtient ainsi la population source de l'année A, pour laquelle on dispose de la base de sondage de l'année A.
A partir de la base de sondage de l'année A, on sélectionne un échantillon d'invités après un tirage au sort stratifié selon le sexe, l'âge, la CPAM, l'affiliation et la TAP selon des probabilités d'inclusion notées $\boldsymbol{\pi_2}$.

On parlera par la suite uniquement du poids de sondage pour inférer à la population cible de l'année A : inverse du produit de la proportion de clés utilisées pour constituer la base de sondage ($\boldsymbol{\pi_1}$) et de la probabilité d'inclusion des invités ($\boldsymbol{\pi_2}$) qui dépend des variables de stratification (sexe, âge, affiliation, CPAM, et TAP).

### 3.3.2 Facteur correctif de non-participation

Dans ce qui suit, on considère que les variables sociodémographiques sont le sexe, l'âge, la CPAM, l'affiliation ; les données passives sont les données du SNIIRAM et de CNAV.
Pour comprendre le cheminement suivi pour calculer le facteur correctif de non-réponse, il faut partir des participants pour lesquels une extraction des données passives a pu être effectuée.
Cinq pondérations pour non-participation ou non-accès aux données passives ont été calculées en suivant un ordre précis.

***Figure 3-2 : Facteurs de pondération calculés pour la pondération finale***

| | |
|---|---|
| $\widehat{\boldsymbol{\delta}}_{\boldsymbol{5\_part}}$ probabilité que l'extraction des données passives des participants pour lesquels on dispose d'une autorisation d'accès à leurs données passives ait été effectuée. | Estimée par régression logistique en utilisant les variables sociodémographiques. A partir des probabilités estimées par le modèle, des GHR ont été constitués par la méthode des scores et un premier facteur correctif de non-réponse, noté $\boldsymbol{w_{5\_part}}$ , a été estimé. |
| $\widehat{\boldsymbol{\delta}}_{\boldsymbol{4\_part}}$ probabilité que les participants autorisent l'accès à leurs données passives sachant qu'ils étaient effectivement participants. | Estimée par régression logistique en utilisant les variables sociodémographiques et en pondérant les participants avec une extraction effectuée par le facteur correctif $w_{5\_part}$ (pour représenter les participants autorisant l'accès à leurs données passives). A partir des probabilités estimées par le modèle, des GHR ont été constitués par la méthode des scores et un deuxième facteur correctif de non-réponse, noté $w_{4\_part}$ , a été estimé. |
| $\widehat{\boldsymbol{\delta}}_{\boldsymbol{3\_part}}$ probabilité d'être participant sachant que la personne avait signifié une acceptation initiale. | Estimée par régression logistique en utilisant les variables sociodémographiques et en pondérant les participants avec une extraction effectuée par le facteur correctif $w_{4\_part}$ (afin de représenter les participants). A partir des probabilités estimées par le modèle, des GHR été constitués par la méthode des scores et un troisième facteur correctif de non-réponse, noté $w_{3\_part}$ , a été estimé. |

| | |
|---|---|
| $\widehat{\boldsymbol{\delta}}_{\boldsymbol{2}}$ probabilité qu'une personne soit classée en acceptation initiale sachant que cette personne avait été invitée, non décédée, non PND, et n'avait pas signifié de refus<br>C'est à ce niveau central que se joue la non-participation. Elle demande donc une attention particulière. | Estimée par régression logistique, en utilisant les variables sociodémographiques et les données passives et en pondérant les participants avec une extraction effectuée par le facteur correctif $w_{3\_part}$ (afin que ces derniers représentent les acceptations initiales).[4]<br>Variables explicatives (une centaine initialement) : sociodémographiques, SNIIRAM (consommation de soins, hospitalisation, ALD, CMU ; les données sur 3 années glissantes précédant l'invitation) et CNAV (emploi, revenus…) de l'année d'invitation :<br>1-Régressions pas à pas descendantes par groupe de variables (sociodémographiques, CNAV, consommation de soins, remboursements, hospitalisation et ALD).<br>2- Régression incluant toutes les variables associées à la participation (seuil de 20%) pour construire le modèle final.<br>A partir des probabilités estimées par le modèle, des GHR ont été constitués par la méthode des scores et un quatrième facteur correctif de non-réponse, noté $\boldsymbol{w_2}$, a été estimé. |

**Le facteur correctif $\mathbf{w_2}$ de la non-participation est calé sur les variables sociodémographiques. La distribution de référence est estimée grâce aux invités de l'année n. On obtient alors le facteur correctif final pour la non-participation noté $\mathbf{w_1}$.**

### 3.3.3 Pondération finale

Pour inférer à la population cible de Constances, le poids final a été divisé par $\boldsymbol{\pi_1}$ ($\boldsymbol{\pi_1}$=nombre de clés NIR de l'année A/ nombre de clés NIR).
Etant donné la forte disparité attendue dans la distribution de ces poids complets, et des potentiels problèmes d'explosion de variance qu'elle peut induire, un poids robuste a été également construit en suivant ces étapes :

1- On affecte aux personnes ayant un poids supérieur au 99ème percentile au poids complet la valeur au 99ème percentile. On obtient un poids auxiliaire.
2- On calcule le total des poids manquants en soustrayant la somme des poids à la somme des poids auxiliaires.
3- On répartit uniformément le total des poids manquants à tous les individus de l'échantillon.
4- On obtient alors le « poids final robuste ».

Ce poids robuste est un poids tronqué ; il permet de diminuer la variance des estimations, mais peut entrainer des biais. Le nombre de poids tronqué doit alors être choisi sur un principe de parcimonie biais/variance. Nous avons par ailleurs considéré comme critère acceptable un rapport de poids robuste max/min équivalent au rapport de poids de sondage max/min.

**La pondération finale mise à disposition est donc le poids final robuste ; il permet d'inférer les estimations à la population cible de Constances de l'année A vivante au moment de l'invitation.**

[4] Dans cette régression logistique, les personnes classées en « Acceptation initiale » sont opposées aux personnes classées en « Non-renvoi de coupon-réponse ». Les personnes classées en « Non-renvoi de coupon-réponse » sont des personnes qui, soit n'ont pas renvoyé de coupon-réponse, soit ont renvoyé un coupon-réponse non renseigné (« absence de choix sur le coupon-réponse »). La catégorie « Non-renvoi de coupon réponse » fait l'objet d'un traitement particulier développé dans la documentation globale (estimation de $\widehat{\boldsymbol{\delta}}_{\boldsymbol{4\_abscho}}$ et de $\widehat{\boldsymbol{\delta}}_{\boldsymbol{5\_nptas}}$ ).

### 3.3.4 Qualité de la démarche

#### *3.3.4.1 Correction de la non-participation*

Afin d'évaluer la qualité de la correction de la non-participation, une prévalence gold standard a été estimée ainsi qu'une prévalence sur l'échantillon de participants en utilisant plusieurs jeux de pondérations : poids de sondage et poids corrigé pour la non-participation.
Nous avons utilisé comme gold standard des données du SNIIRAM non incluses dans le modèle de non-réponse, les variables relatives à la prescription de médicaments.
A partir des poids de sondage initiaux et des données des participants et de la cohorte témoin de non participants, des prévalences sans biais, considérées comme gold standard, ont pu être calculées.
A partir des poids de sondage initiaux et des poids corrigés pour la non-réponse, des prévalences sur l'échantillon de répondants ont été calculées.
Les prévalences ont été comparées en utilisant comme critère l'erreur relative. Les tableaux de résultats complets figurent dans la documentation complète.

***Figure 4-3 : Exemples de prévalences des prescriptions de médicaments gold standard et estimées à partir des participants - 2017***

| | Gold standard | | Sur les participants | | | Comparaison au gold standard | |
|---|---|---|---|---|---|---|---|
| | | | | Sans correction NR | Correction NR | Sans correction NR | Correction NR |
| | n | % | n | % | % | ER | ER |
| **SNIIRAM : Nombre total de boîtes délivrées (2015-2017)** | | | | | | | |
| **D "Dermatologie"** | | | | | | | |
| Aucune | 35097 | 38,7 | 7869 | 33,8 | 35,7 | -11,8 | -5,7 |
| 1-inf6 | 47202 | 49,6 | 12939 | 56,1 | 53,3 | 13 | 6,6 |
| Au moins 6 | 10098 | 11,7 | 2316 | 10,1 | 11,0 | -16,4 | -9,5 |
| **R "Système respiratoire"** | | | | | | | |
| Aucune | 33117 | 36,2 | 7760 | 33,2 | 34,8 | -7,4 | -5,5 |
| 1-inf5 | 41378 | 43,9 | 10978 | 47,7 | 45,1 | 9,9 | 6,8 |
| 5-inf10 | 9719 | 10,7 | 2471 | 10,8 | 11,2 | -3,7 | -3 |
| Au moins 10 | 8183 | 9,2 | 1915 | 8,2 | 8,9 | -13 | -6,6 |
| **CNAV** | | | | | | | |
| **Typologie d'activité professionnelle (TAP)** | | | | | | | |
| Aucune info | 28657 | 23,2 | 6207 | 26,0 | 22,0 | 25,5 | -4,9 |
| Actif CS NR | 3390 | 4,0 | 1047 | 4,6 | 4,5 | 13,7 | 14,6 |
| Actif CS3 | 6545 | 6,6 | 3016 | 13,1 | 7,1 | 102,2 | 11,2 |
| Actif CS4 | 6661 | 7,1 | 2550 | 11,3 | 7,4 | 57,8 | 5,5 |
| Actif CS5 | 16390 | 20,3 | 4290 | 19,3 | 20,9 | -8,2 | 4,4 |
| Actif CS6 | 15440 | 19,3 | 2607 | 11,2 | 18,7 | -46,1 | -4,3 |
| Inactif CS3 | 2600 | 3,1 | 867 | 3,7 | 2,6 | 10,8 | -25,7 |
| Inactif CS4 | 1668 | 1,9 | 530 | 2,3 | 2,0 | 20,3 | -15,2 |
| Inactif CS5 | 5791 | 7,6 | 1249 | 5,4 | 7,1 | -32,9 | -2,5 |
| Inactif CS6 | 5255 | 6,9 | 761 | 3,1 | 7,6 | -56,6 | 2,4 |
| **Allocation maladie** | | | | | | | |
| 0 | 68078 | 70,3 | 17274 | 74,2 | 71,9 | 3,9 | -0,1 |
| 1 | 24319 | 29,7 | 5850 | 25,8 | 28,1 | -39,2 | 1,2 |

Note : hypothèses sur le processus de non-réponse : MCAR/sans correction de la non-réponse et MAR/variables sociodémographiques, SNIIRAM et CNAV)

#### *3.3.4.2 Pondération robuste proposée*

Les poids finaux robustes peuvent entrainer une augmentation du biais et une diminution de la variance. Afin d'évaluer l'impact de la troncature des poids, nous avons estimé les prévalences obtenues sur l'échantillon de participants en utilisant le poids final complet et le poids final robuste et nous avons comparé ces prévalences ainsi que leurs écarts-types (cf. documentation complète).

# 4 Les pondérations Constances en chiffres

Les pondérations été calculées indépendamment pour chaque année d'invitation.

***Figure 5-1 : Description des données passives demandées ou extraites - 2017***

| Année de pondération | Bases de données Constances consolidées | Extrapolation des résultats |
|---|---|---|
| **2013** | Juin 2017 | Population cible de Constances 2013 |
| **2014** | Juin 2017 | Population cible de Constances 2014 |
| **2015** | Juillet 2019 | Population cible de Constances 2015 |
| **2016** | Mai 2020 | Population cible de Constances 2016 |
| **2017** | Septembre 2021 | Population cible de Constances 2017 |

Pour chaque année, les poids ne sont pas disponibles pour l'ensemble des individus du fait que les données auxiliaires n'étaient pas disponibles pour ces individus au moment du calcul des pondérations.

***Figure 5-2 : Description des données passives demandées ou extraites - 2017***

| | Effectif | % |
|---|---|---|
| **Demande d'extraction de données passives** | | |
| Cohorte témoin | 82 200 | 100,0 |
| Participants | 29 339 | 93,5 |
| **Extractions de données passives** | | |
| Cohorte témoin | 74 317 | 90,4 |
| Participants | 27 379 | 93,6 |

Les pondérations visant à redresser les estimations de la non-réponse sont disponibles sur les périodes 2013-2017 avec les effectifs indiqués dans la figue 5-3.

***Figure 5-3 : Effectifs des poids sur la période 2013-2017***

| Année | Poids disponibles | % participants avec un poids |
|---|---|---|
| **2013** | 14 521 | 88,9% |
| **2014** | 19 717 | 87,5% |
| **2015** | 25 278 | 87,0% |
| **2016** | 23 113 | 80,6% |
| **2017** | 27 379 | 87,8% |
| **Total** | **110 008** | **86,1%** |

Sur les 5 années représentées, les pondérations sont disponibles pour 86,1% des participants à la cohorte Constances.

Les effectifs concernant le nombre de participants pondérables peuvent différer de ceux mis à disposition, les participants pondérables pouvant changer de statut au cours du temps (par exemple suite à un déconsentement).

# 5 En pratique

## 5.1 Informations mises à disposition

Le fichier mis à disposition pour les pondérations contient l'identifiant projet, l'année de la vague (2013, 2014, 2015, 2016, 2017), la pondération robuste de 2013, 2014, 2015, 2016 et 2017.

***Tableau 7-1 : Représentation graphique simplifiée de la base de données***

| nconstances | Annee_vague | W | Variable d'intérêt Y |
|---|---|---|---|
| **2XXXXXXXX** | 2013 | $W\_2013_i$ | $Y\_2013_i$ |
| **2XXXXXXXX** | 2013 | $W\_2013_i$ | $Y\_2013_i$ |
| **3XXXXXXXX** | 2014 | $W\_2014_i$ | $Y\_2014_i$ |
| **4XXXXXXXX** | 2015 | $W\_2015_i$ | $Y\_2015_i$ |
| **4XXXXXXXX** | 2015 | $W\_2015_i$ | $Y\_2015_i$ |
| **5XXXXXXXX** | 2016 | $W\_2016_i$ | $Y\_2016_i$ |
| **5XXXXXXXX** | 2016 | $W\_2016_i$ | $Y\_2016_i$ |

Ainsi, supposons que l'on souhaite estimer le total de la variable y en 2013.

- Si on souhaite calculer le total de l'échantillon, il sera estimé par $\hat{y}_{tot,ech} = \sum_{k=1}^{i} y\_2013_k$
- Si on souhaite calculer le total de la population, il sera estimé par $\hat{y}_{tot,pop} = \sum_{k=1}^{i} w\_2013_k y\_2013_{\_k}$

## 5.2 Variance

Dans les enquêtes par sondage, la variance est liée au plan de sondage et à la non-réponse. En général, les logiciels classiques ne permettent pas d'estimer correctement la variance en présence de non-réponse. Dans Constances, compte tenu du fort taux de non-réponse et des 5 phases de non-réponse, la variance estimée par les logiciels classiques est potentiellement biaisée.

## 5.3 Logiciels

Pour exploiter les données sous un logiciel classique, **le seul paramètre à prendre en compte est le poids final**.

### 5.3.1 Sous SAS

Sous SAS, il faut utiliser les procédures SURVEY (surveyfreq, surveymeans, surveyreg, surveylogistic).

***Exemple pour l'estimation d'une prévalence :***

```
proc surveyfreq data= base_de_données;
        tables variable /row cl;
        weight poids;
run;
```

### 5.3.2 Sous Stata

Sous Stata, il faut d'abord déclarer le plan de sondage par la commande SVYSET puis utiliser les procédures SVY (débuter chaque commande par svy : commandes classiques).

***Déclaration du plan de sondage (et description) :***

svyset [pweight= poids]
svydes

***Exemple pour l'estimation d'une prévalence :***

svy: prop variable

## 5.4 Comment présenter les résultats ?

Classiquement, pour les enquêtes par sondage, sont présentés les effectifs observés dans l'échantillon et les prévalences pondérées, avec leurs intervalles de confiance estimés par SAS ou Stata. Il est conseillé de l'indiquer dans la partie « méthodes » des publications.

***Exemple :*** « Toutes les estimations présentées prennent en compte le plan de sondage et la correction de la non-réponse. En revanche, les effectifs présentés sont les effectifs observés dans l'échantillon ».

| | n | % | IC 95% | |
|---|---|---|---|---|
| **Etat de santé général perçu** | | | | |
| Très bon | 9981 | 54,1 | 53,1 | 55,1 |
| Moyennement/pas bon | 8858 | 45,9 | 44,9 | 46,9 |

## 5.5 Combinaison de plusieurs années

En cas d'événements rares, il peut être tentant de combiner les pondérations de 2013 et 2014. C'est déconseillé avec les pondérations mises à disposition aujourd'hui. En effet, les populations cibles de Constances en 2013 et 2014 ne sont pas exactement les mêmes et il aurait fallu vérifier la stabilité des estimations d'une année à l'autre.

En effet, la validité de la combinaison des deux années 2013 et 2014, repose sur des hypothèses supplémentaires. D'une part, la population source de 2013 doit avoir la même structure que celle de 2014 et les changements de strates d'une année à l'autre doivent être négligeables. D'autre part, la variable d'intérêt ne doit pas être affectée par un effet période entre 2013 et 2014 ; autrement dit, en adoptant une approche contrefactuelle (17) un sujet invité en 2013 né une année donnée aurait répondu en moyenne comme un sujet invité en 2014 né un an plus tard et ayant exactement les mêmes caractéristiques influant sur la variable d'intérêt que le sujet précédent.

# 6 Références


(1) Zins M, Goldberg M, Carton M, Guéguen A, Henny J, Le Got S, et al. La cohorte CONSTANCES : une infrastructure pour la recherche et la santé publique. Bull Epidémiol Hebd 2016.

(2) Equipe Constances. Projets validés par le Conseil scientifique international. 1-3-2016. Ref Type: Online Source

(3) Meffre C. Prévalence des hépatites B et C en France en 2004. Saint Maurice: Institut de veille sanitaire; 2016.

(4) Ardilly P. Présentation des plans de sondage classiques. Les tehcniques de sondage.Paris: Editions Technip; 1994. p. 47-93.

(5) Ardilly P. Les techniques de sondage. Paris: 1994.

(6) Lohr SL. Sampling: Design and analysis. 1999.

(7) Särndal CE, Swensson B, Wretman J. Model assisted survey sampling. New York: Springer-Verlag; 1992.

(8) Santin G. Non-réponse totale dans les enquêtes de surveillance épidémiologique 2015.

(9) Bethlehem J.G. Weighting nonresponse adjustments based on auxiliary information. 2013.

(10) Rubin DB. Inference and missing data. Biometrika 1976;63:581-90.

(11) 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.

(12) Haziza D, Beaumont JF. On the construction of imputation classes in surveys. Int Stat Rev 2007;75(1):25-43.

(13) Little RJA. Survey nonresponse adjustments for estimates of means. Int Stat Rev 1986;54:139-57.

(14) Santin G, Geoffroy B, Benezet L, Delezire P, Chatelot J, Sitta R, et al. In an occupational health surveillance study, auxiliary data from administrative health and occupational databases effectively corrected for nonresponse. J Clin Epidemiol 2014 Jun;67(6):722-30.

(15) Martikainen P, Laaksonen M, Piha K, Lallukka T. Does survey non-response bias the association between occupational social class and health? Scand J Public Health 2007;35(2):212-5.

(16) Vercambre MN, Gilbert F. Respondents in an epidemiologic survey had fewer psychotropic prescriptions than nonrespondents: An insight into health-related selection bias using routine health insurance data. J Clin Epidemiol 2012;65(11):1181-9.

(17) Hernan MA, Robins JM. Instruments for causal inference: an epidemiologist's dream? Epidemiology 2006 Jul;17(4):360-72.