ESTIMATIONS DE PRÉVALENCES PAR COMBINAISON DES ÉCHANTILLONS ANNUELS DE LA COHORTE CONSTANCES

Laetitia Bénézet, Adeline Renuy, Marie-Christine Delmas, Yuriko Iwatsubo

ESTIMATIONS DE PRÉVALENCES PAR COMBINAISON DES ÉCHANTILLONS ANNUELS DE LA COHORTE CONSTANCES

Laetitia Bénézet[1], Adeline Renuy[2], Yuriko Iwatsubo[3], Marie-Christine Delmas[4]

[1] *Santé publique France, Saint-Maurice, France*
laetitia.benezet@santepubliquefrance.fr
[2] *UMS 011 Inserm UVSQ "Cohortes épidémiologiques en population", Villejuif, France*
adeline.renuy@inserm.fr
[3] *Santé publique France, Saint-Maurice, France*
yuriko.iwatsubo@santepubliquefrance.fr
[4] *Santé publique France, Saint-Maurice, France*
marie-christine.delmas@santepubliquefrance.fr

Résumé. L'objectif de cette communication est de présenter la méthode utilisée pour combiner les échantillons annuels de la cohorte Constances. La constitution de la cohorte Constances s'est étalée sur plusieurs années. Chaque année un échantillon aléatoire a été tiré au sort et des pondérations annuelles robustes ont été calculées et mises à disposition. Nous présenterons les éléments à prendre en compte avant de combiner les échantillons ainsi que la méthode utilisée. Pour illustrer la méthode, dans le cadre de la surveillance épidémiologique des maladies respiratoires à Santé publique France, les échantillons 2013 et 2014 ont été combinés. Les prévalences estimées à partir des échantillons annuels et de l'échantillon combiné pour certaines variables d'intérêt seront comparées.

Mots-clés. Combinaison d'enquêtes périodiques, enquête de santé, estimation de prévalence, cohorte.

Abstract. The aim of this communication is to present the method of combining the annual samples of the Constances cohort. The Constances cohort participants were included over several years. A random sample was drawn each year and robust weights were calculated. We will present the method used and main elements to take into account before combining the samples. To illustrate the method in the epidemiologic surveillance of respiratory health conducted by Santé publique France, the 2013 and 2014 samples were combined. The estimated prevalences from the annual samples and the combined sample for certain variables of interest will be compared.

Keywords. Combining periodic surveys, health surveys, prevalence estimate, cohort.

1 Contexte

La cohorte Constances est une cohorte épidémiologique en population générale (Zins 2015). Elle constitue une infrastructure de recherche permettant des études d'épidémiologie analytique, de santé publique et de surveillance épidémiologique. Elle doit permettre de produire des estimations au niveau de sa population cible.
Le recueil de données implique que les personnes se rendent dans un centre d'examen de santé (CES). De ce fait l'inclusion s'étale sur plusieurs années. Chaque année, un échantillon aléatoire est tiré au sort et des pondérations annuelles sont calculées et mises à disposition.

L'utilisation des échantillons annuels ne permet pas toujours d'étudier des sous-groupes d'intérêt ou des évènements de santé rares. De plus, des fluctuations d'échantillonnage d'une année sur l'autre peuvent être observées, notamment dans des sous-groupes d'effectifs restreints. C'est pourquoi il peut être intéressant de combiner les échantillons annuels pour pallier ces inconvénients.

L'objectif de cette communication est de présenter les éléments à prendre en compte avant de combiner les échantillons et de décrire la méthode utilisée. Pour illustrer la méthode dans le cadre de la surveillance épidémiologique des maladies respiratoires à Santé publique France, les échantillons 2013 et 2014 de Constances ont été combinés. Les prévalences estimées, pour certaines variables d'intérêt, à partir des échantillons annuels et de l'échantillon combiné seront comparées.

2 Matériel et méthode

2.1 Les échantillons annuels de Constances

En raison des contraintes logistiques, la population cible de Constances et la taille des échantillons tirés au sort diffèrent chaque année.

La population cible, pour une année donnée, correspond aux personnes affiliées au Régime Général de la Sécurité Sociale ou à l'une des sections locales mutualistes (SLM) ayant signé une convention avec Constances[1], âgées de 18 à 69 ans au moment de leur invitation et résidant dans un département couvert par un CES d'une caisse primaire d'assurance maladie (CPAM) participant à Constances[2].

Chaque année, une base de sondage est constituée à partir du Répertoire national inter-régimes des bénéficiaires de l'assurance maladie (Rniam) géré par la caisse nationale d'assurance vieillesse en sélectionnant un certain nombre de clés du NIR (numéro d'inscription au répertoire)[3]. L'inclusion s'étalant sur plusieurs années, les clés NIR ont été préalablement partitionnées afin de constituer des bases de sondage annuelles indépendantes. Chaque année,

[1] Une SLM est une mutuelle complémentaire qui gère pour le compte du régime général la part obligatoire remboursée par la sécurité sociale pour ses adhérents. En 2013 et 2014, les SLM incluses dans la population cible de Constances étaient la Camieg (Industries électriques et gazières) et la MGEN (Éducation nationale).
[2] En 2013, 17 CPAM participaient (Angoulême, Bordeaux, Lille, Lyon, Marseille, Nancy, Nîmes, Paris, Pau, Poitiers, Rennes, Saint-Brieuc, Saint-Nazaire, Toulouse, Tours, Lens, Orléans (hors affiliés Camieg pour ces 2 CPAM)), la CPAM de Bayonne et les affiliés Camieg de toutes les CPAM ont été ajoutés en 2014.
[3] Le NIR est constitué de 13 chiffres suivi d'une clé de contrôle à 2 chiffres qui peuvent être considéré comme générés aléatoirement.

une base de sondage est ainsi constituée en sélectionnant les personnes appartenant aux clés NIR de l'année en cours. Un tirage au sort à probabilités inégales stratifié selon le régime d'affiliation, le CES, la classe d'âge, le sexe et la typologie d'activité professionnelle est ensuite réalisé. Les personnes sont alors invitées à se rendre dans leur CES afin de bénéficier d'un examen de santé et à répondre à différents questionnaires auto-administrés.
Des pondérations annuelles robustes sont calculées et mises à disposition des équipes qui exploitent les données de Constances. Elles prennent en compte la probabilité d'inclusion, un facteur correctif de la non-réponse totale et sont tronquées afin de limiter la dispersion des poids (Constances 2017).

2.2 Conditions pour combiner les échantillons

Avant de pouvoir combiner les échantillons annuels de Constances, il est nécessaire de vérifier certaines conditions (Schenker 2002, Thomas 2009) :

1. S'assurer que les échantillons sont indépendants. C'est le cas pour Constances puisque des bases de sondage annuelles indépendantes sont constituées à partir d'une partition des clés NIR.
2. S'assurer de la stabilité de la population cible. Combiner des échantillons annuels revient à considérer qu'ils sont issus d'une même population et l'échantillon combiné est considéré comme un échantillon plus important issu de cette population. La population cible de Constances a été légèrement élargie au cours du temps mais sa structure en termes de sexe et d'âge (variables issues de la base de sondage sur lesquelles les poids annuels robustes sont calés) reste relativement stable.
3. S'assurer de la comparabilité des plans de sondage. Mis à part les ajustements des fractions de sondage pour s'adapter à certaines contraintes (nombre d'examens de santé réalisables dans chaque CES, sur-représentation des SLM, effectif suffisant dans chaque strate), le plan de sondage reste le même d'une année sur l'autre.
4. S'assurer que la collecte des données et que les variables mesurées sont homogènes. Le protocole de recueil des données et les questionnaires étant strictement identiques d'une année sur l'autre, il n'y aura pas d'effet de mode.
5. S'assurer que l'effet période est limité sur les indicateurs de santé d'intérêt ou les déterminants en lien avec la santé respiratoire. Les périodes cumulées sont deux années consécutives et on observe peu de variation entre les estimations annuelles pour la plupart des variables.

2.3 Méthode de combinaison

L'échantillon combiné étant considéré comme un échantillon issu d'une seule population, les poids annuels doivent être rééchelonnés en calculant des coefficients de combinaison. Dans le cas contraire, l'estimation du nombre de cas prévalent d'une maladie d'intérêt serait surestimée, d'un facteur deux si l'on combine deux échantillons annuels (Korn et Graubard 1999).

Plusieurs méthodes ont été proposées pour le calcul des coefficients de combinaison. (Kish 1999, Korn et Graubard 1999, Friedman 2002).
Certaines méthodes proposent de choisir des coefficients dont le calcul est basé sur l'effet du plan de sondage sur les variables d'intérêt ou sur les effectifs des sous-groupes d'intérêt mais il est alors nécessaire de calculer autant de coefficients que de variables ou de sous-groupes d'intérêt.

D'autres méthodes proposent un coefficient unique pour chaque échantillon. Le poids robuste (ω_n^i) d'un individu i inclus l'année n est alors multiplié par le coefficient de combinaison de l'année (α_n).

$$\omega_{comb}^i = \omega_n^i \times \alpha_n$$

Les coefficients de combinaison peuvent être déterminés a priori, en choisissant par exemple des coefficients identiques pour chaque année ou bien accorder plus d'importance à l'année la plus récente. Leur calcul peut également se baser sur les effectifs, en choisissant des coefficients proportionnels à la taille des échantillons ou à celle des populations cibles. Enfin les coefficients peuvent être proportionnels à l'effet du plan de sondage sur les poids robustes (Westat 2006). L'avantage de cette dernière méthode est de donner plus de poids à l'enquête dont la dispersion des poids est la plus faible. C'est celle que nous avons choisie d'utiliser.

Dans le cas présent, pour la combinaison des échantillons 2013 et 2014, les coefficients de combinaison sont alors calculés selon la formule :

$$\alpha_{2013} = 1 - \frac{deff(\omega_{2013})}{deff(\omega_{2013}) + deff(\omega_{2014})}$$

$$\alpha_{2014} = 1 - \alpha_{2013} = 1 - \frac{deff(\omega_{2014})}{deff(\omega_{2013}) + deff(\omega_{2014})}$$

où $deff(\omega_n) = 1 + \left(\frac{CV(\omega_n)}{100}\right)^2$

et $CV(\omega_n)$ correspond au coefficient de variation des pondérations de l'année *n*.

Les poids combinés ont ensuite été calés sur les effectifs par régime d'affiliation, CES, sexe et classe d'âge de la population cible de l'année 2014.

3 Résultats

Entre 2013 et 2014, la taille de la population cible a augmenté de 0,6 %, en lien avec une augmentation de la couverture géographique et l'ajout de SLM en 2014. Le nombre d'individus inclus dans la cohorte Constances a été supérieur en 2014. Les pondérations 2013 ayant une variabilité plus importante, le coefficient de combinaison de l'année 2014 est supérieur (Tableau 1).

Tableau 1 : Effectifs de la population cible, de l'échantillon de répondants, et coefficient de combinaison

	2013	**2014**	**Δ**
Effectif population cible*	8 901 649	8 958 234	+ 0,6 %
Effectif échantillon de répondants	14 521	19 717	+ 36 %
Coefficient de combinaison α_n	0,4669	0,5331	

* Estimé à partir des effectifs dans la population source issue de la base de sondage de l'année n et du nombre de clé NIR sélectionnées cette année.

Le tableau 2 décrit les principales caractéristiques sociodémographiques estimées dans la population cible de Constances pour chacune des deux années. La répartition par sexe et niveau de diplôme ne varie pas d'une année sur l'autre. La population 2014 compte plus de personnes de moins de 50 ans, plus de personnes en recherche d'emploi et un peu moins de retraités.

Tableau 2 : Caractéristiques sociodémographiques de la population cible de Constances en 2013 et 2014 estimées à partir des échantillons annuels 2013 et 2014

	2013		**2014**		
	%	**[IC 95 %]**	**%**	**[IC 95 %]**	**p**
Sexe					NS
Homme	49,3	[48,1-50,6]	48,6	[47,6-49,7]	
Femme	50,7	[49,4-51,9]	51,4	[50,3-52,4]	
Classe d'âge					0,04
≤ 29 ans	17,0	[15,9-18,1]	18,3	[17,4-19,1]	
30-39 ans	22,5	[21,4-23,6]	22,8	[21,9-23,7]	
40-49 ans	21,8	[20,8-22,8]	22,6	[21,8-23,5]	
50-59 ans	20,3	[19,3-21,3]	19,2	[18,4-20,0]	
≥ 60 ans	18,5	[17,6-19,4]	17,2	[16,4-17,9]	
Niveau de diplôme*					NS
Niveau 0 à 2	11,9	[11,0-12,8]	11,6	[10,9-12,3]	
Niveau 3 et 4	37,7	[36,4-38,9]	38,4	[37,4-39,4]	
Niveau 5 et 6	30,9	[29,8-32,1]	30,3	[29,3-31,2]	
Niveau 7 et 8	19,5	[18,6-20,5]	19,8	[19,0-20,6]	
Situation vis-à-vis de l'emploi					0,01
En emploi	68,1	[66,9-69,3]	68,5	[67,5-69,5]	
Demandeur d'emploi	10,5	[9,6-11,5]	12,0	[11,3-12,8]	
Retraité	16,0	[15,1-16,9]	14,5	[13,9-15,2]	
Autre sans activité	5,4	[4,7-6,1]	4,9	[4,4-5,5]	

% : pourcentage pondéré par les poids annuels robustes
[IC 95 %] : intervalle de confiance à 95%
* Niveaux de la classification internationale type de l'éducation
p : degré de signification du test du chi² modifié de Rao et Scott

Les prévalences estimées pour les indicateurs en lien avec la santé respiratoire sont stables d'une année sur l'autre (test du chi² de Rao et Scott non significatif pour toutes les variables présentées) et les estimations combinées sont donc très proches des estimations annuelles (Tableau 3). Les coefficients de variation des estimations combinées sont cependant légèrement inférieurs aux coefficients de variations des estimations annuelles, notamment lorsque les évènements sont moins fréquents.

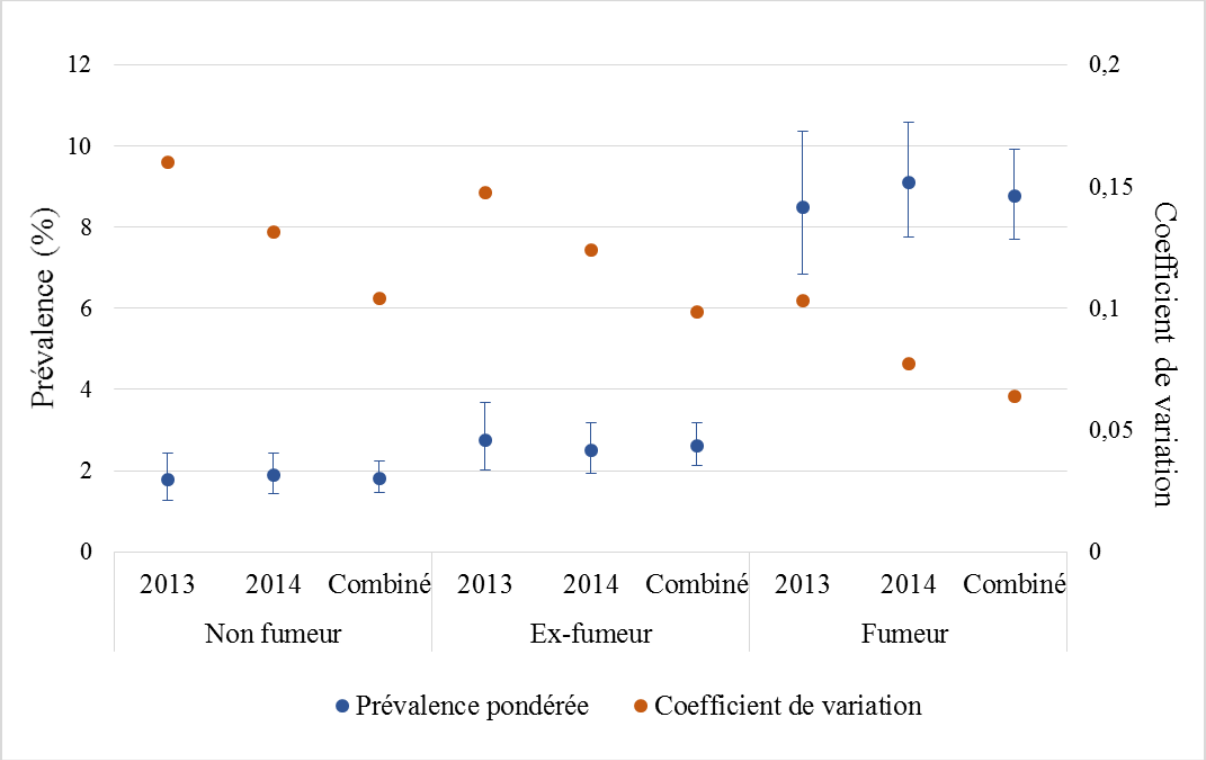
Tableau 3 : Estimation des prévalences annuelles et combinées pour les indicateurs en lien avec la santé respiratoire dans la population cible de Constances

	2013			**2014**			**Combiné**	
	n	**% [IC 95 %]**	**CV**	**n**	**% [IC 95 %]**	**CV**	**% [IC 95 %]**	**CV**
Sifflements	1926	16,6 [15,6-17,6]	0,03	2724	16,8 [16,0-17,6]	0,02	16,8 [16,1-17,4]	0,02
Asthme actuel	670	5,6 [5,0-6,3]	0,06	1062	5,9 [5,4-6,4]	0,04	5,8 [5,4-6,3]	0,04
Bronchite chronique	411	3,7 [3,2-4,3]	0,07	572	3,9 [3,4-4,3]	0,06	3,8 [3,4-4,2]	0,05
Allergies nasales	4916	35,6 [34,4-36,9]	0,01	6673	35,8 [34,8-36,8]	0,01	35,8 [35,0-36,6]	0,01
Non-fumeur	6156	44,5 [43,2-45,8]	0,01	8592	45,0 [44,0-46,1]	0,01	44,9 [44,1-45,8]	0,01
Fumeur	2677	24,2 [23,0-25,4]	0,02	3801	24,3 [23,4-25,3]	0,02	24,4 [23,6-25,2]	0,02
Ex-fumeur	4956	31,3 [30,1-32,4]	0,02	6337	30,6 [29,7-31,6]	0,02	30,7 [29,9-31,4]	0,01
Minceur -maigreur	325	2,9 [2,5-3,5]	0,08	472	2,7 [2,4-3,0]	0,06	2,9 [2,6-3,2]	0,05
Corpulence normale	7682	53,0 [51,7-54,3]	0,01	10345	52,4 [51,4-53,4]	0,01	52,9 [52,1-53,8]	0,01
Surpoids	4426	30,2 [29,0-31,4]	0,02	6020	30,3 [29,3-31,2]	0,02	29,9 [29,2-30,7]	0,01
Obésité	1725	13,8 [12,9-14,8]	0,03	2570	14,7 [13,9-15,5]	0,03	14,3 [13,7-14,9]	0,02

n : effectif dans l'échantillon
% [IC 95 %] : prévalence pondérée et intervalle de confiance à 95%
CV : coefficient de variation

Si l'on s'intéresse à des évènements de santé peu fréquents dans des sous-groupes de population comme la prévalence de la bronchite chronique selon le statut tabagique, les effectifs des échantillons annuels sont limités. Dans la figure 1, on observe que la prévalence de bronchite chronique selon le statut tabagique est stable d'une année sur l'autre mais avec des intervalles de confiance importants, notamment chez les fumeurs. Les estimations à partir de l'échantillon combiné sont du même ordre de grandeur mais le coefficient de variation est plus faible et les intervalles de confiance sont plus restreints.

Figure 1 : Estimation des prévalences annuelles et combinée de bronchite chronique selon le statut tabagique dans la population cible de Constances



4 Discussion

Les échantillons annuels 2013 et 2014 de la cohorte Constances ont été combinés afin d'estimer des prévalences d'évènements en lien avec la santé respiratoire. Cette combinaison permet d'augmenter les effectifs et d'étudier des sous-groupes ou des évènements peu fréquents de façon plus robuste.

La méthode de rééchelonnage des pondérations que nous avons retenue prend en compte l'effet du plan de sondage sur les pondérations et donne plus d'importance à l'année dont les poids ont le moins de variabilité (dans le cas présent l'année 2014) mais de nombreuses autres méthodes de calcul des coefficients ont été décrites. Dans le cas présent, la plupart des méthodes donnerait plus d'importance à l'année 2014 puisque qu'elle cible une population plus large, se base sur un échantillon plus important et présente un effet du plan de sondage plus faible pour la plupart des variables d'intérêt. Une analyse de sensibilité en utilisant d'autres méthodes de calcul des coefficients de combinaison pourra être réalisée pour mesurer l'impact sur les prévalences estimées.

L'inclusion dans la cohorte Constances s'est prolongée jusqu'en 2020 et des échantillons pondérés seront disponibles jusqu'en 2017. Il sera intéressant de les combiner pour augmenter la puissance statistique des analyses et estimer des prévalences avec une meilleure précision. Les éléments discutés en 2.2 devront alors être réexaminés afin de s'assurer que les conditions nécessaires pour pouvoir combiner les échantillons annuels sont réunies. En effet, plus la période s'allonge, plus la stabilité de la population et des évènements d'intérêt est discutable. Cependant, lorsque l'on combine des échantillons, il est possible d'inclure la période comme variable d'ajustement et donc de prendre en compte l'évolution des indicateurs de santé au cours du temps.

Bibliographie

Constances (2017), Constances : Pondérations 2013 et 2014. Document de travail méthodologique sur la construction des pondérations

Friedman EM, Jang D, Williams VT. (2002), Combined Estimates from Four Quarterly Survey Data Sets. *Proceedings from the Joint Statistical Meetings – Section on Survey Research Methods*. pp. 1064-69.

Kish L (1999). Cumulating/combining population surveys. Survey Methodology, 25, pp.: 129-138.

Korn, E. L. Graubard B. I. (1999), Analyses Using Multiple Surveys, *Analysis of Health Surveys*. Wiley series in probability and statistics, New York.

Schenker N, Gentleman JF, Rose D, Hing E, Shimizu IM. (2002), Combining estimates from complementary surveys: a case study using prevalence estimates from national health surveys of households and nursing homes. *Public Health Reports,* 117, pp. 393–407.

Thomas S, Wannell B. (2009), Combining cycles of the Canadian Community Health Survey. Health Reports, 20, pp. 53-58.

Westat (2006), Data on Health and Well-being of American Indians, Alaska Natives And Other Native Americans. Data Catalog. U.S. Department of Health and Human Services. Washington, D.C.

Zins M, Goldberg M, and Constances Team. (2015), The French CONSTANCES population-based cohort: design, inclusion and follow-up. *European Journal of Epidemiology,* 30, pp. 1317-1328.