CORRECTION DE LA NON-RÉPONSE ET ESTIMATION DE PRÉVALENCES : RÉSULTATS ISSUS DE TROIS COHORTES ÉPIDÉMIOLOGIQUES CIBLANT LES RISQUES PROFESSIONNELS

N. SOULLIER[1], B. GEOFFROY-PEREZ[2], A. GUEGUEN[3], L. BENEZET[2], J. CHATELOT[2], M. ZINS[3], G. SANTIN[3]

[1] *Santé publique France, 12 rue du Val d'Osne, 94415 Saint-Maurice Cedex France. noemie.soullier@santepubliquefrance.fr*

[2] *Santé publique France, F-94415 Saint-Maurice Cedex France*

[3] *UMS011 INSERM-UVSQ – Unité « Cohortes épidémiologiques en population », F-94807, Villejuif, France*

Résumé. Nous présentons la méthode utilisée dans trois cohortes françaises pour corriger de la non-réponse totale à l'inclusion, ainsi que l'impact de la correction de la non-réponse sur l'estimation des prévalences de variables d'intérêt issues des questionnaires.

Ce travail s'appuie sur des données des cohortes Constances, Coset-MSA et Coset-RSI, soit des échantillons d'environ 440 000, 10 000 et 20 000 personnes tirées au sort et invitées en 2012-2013, 2010 et 2012 respectivement. Des données administratives ont été collectées pour les trois cohortes par appariement avec les systèmes d'information des différents régimes de Sécurité Sociale et de retraite. Ces données ont été utilisées pour corriger de la non-réponse.

Nos résultats montrent qu'une position sociale élevée et un emploi stable favorisent la réponse aux enquêtes. Les personnes qui prennent soin de leur santé sont également plus enclines à répondre à nos enquêtes, dont un des sujets est la santé. A l'inverse, les personnes qui ont des problèmes de santé plus graves sont moins enclines à répondre, probablement en lien avec leur état de santé. Un résultat marquant est que l'association entre le sexe et la réponse est captée, dans les modèles multivariables, par des covariables reflétant des consommations de soins liées au sexe. Enfin, en étudiant l'impact de la correction de la non-réponse sur les prévalences estimées, il ressort que les fumeurs et les personnes avec des symptômes dépressifs avaient moins répondu aux enquêtes. Ces résultats seront très utiles pour d'autres enquêtes qui nécessitent de corriger de la non-réponse totale.

Mots-clés. Non-réponse, cohorte, santé, travail, prévalences, données administratives.

Abstract. In the context of declining response rates in surveys, developing an efficient methodology to correct for non-response is crucial. This paper presents the modeling of non-response correction for three French cohorts focusing on occupational epidemiology and evaluates its impact on prevalence estimates for outcomes collected using questionnaires.

The data from the Constances, Coset-MSA and Coset-RSI cohorts studied here are based on stratified random samples of approximatively 440 000, 10 000 and 20 000 people, invited in 2012-2013, 2010 and 2012 respectively. Each cohort investigated a different target population. We collected administrative data for all three cohorts from social security databases and insurance fund databases by linkage and used them to correct for non-response.

We found that a high socio-professional status and job stability were positive determinants of survey response. Standard health care utilization was also associated with a higher probability of response, which suggests that people who care more about their health are more likely to answer a survey about health. In contrast, people with serious diseases were less likely to respond, probably due to their medical condition. One important result is that the effect of gender on non-response was absorbed in the multivariable models by covariates reflecting gender-specific health consumptions. Finally, looking at the impact of the correction for non-response on outcome prevalence estimates, smokers and people with depressive symptoms were less likely to respond to the cohort surveys. These results could be very useful for other cohorts that need to correct for non-response.

Keywords. non-response, cohort, health, occupation, prevalence, administrative data

1. Introduction

Les enquêtes sont encore aujourd'hui un outil essentiel de collecte de données épidémiologiques, et ce malgré des taux de réponse sur le déclin (Tolonen et al., 2006). Les faibles taux de réponse sont problématiques pour deux raisons principales. Premièrement, les estimations, basées sur un nombre de répondants plus faible, sont alors moins précises. Deuxièmement, la non-réponse peut être sélective et produire des estimations biaisées, en particulier pour des proportions (Van Loon et al., 2003).

Des méthodes adaptées pour corriger de la non-réponse peuvent permettre de réduire ce biais. Pour cela, il est nécessaire de disposer de données associées à la fois à la non-réponse et aux variables d'intérêt. Idéalement, ces données doivent être disponibles à la fois pour les répondants et pour les non-répondants.

Nous présentons ici la correction de la non-réponse à l'inclusion pour trois cohortes françaises s'intéressant à la santé et aux risques professionnels : Constances, Coset-MSA et Coset-RSI.

2. Matériel

2.1. Cohortes étudiées

La cohorte Constances inclut les personnes âgées de 18 à 69 ans, résidant dans une des régions participantes et affiliées à la Cnam-TS, qu'elles soient étudiantes, inactives, retraitées, actives en emploi ou au chômage (Zins et al., 2015). Ici, la population a été restreinte aux personnes âgées de 30 à 69 ans, invitées en 2012-2013 et en vie au 31 janvier 2014, soit 439 472 personnes. Parmi celles-ci, 31 642 ont répondu au questionnaire d'inclusion (7,2 %).

La cohorte Coset-MSA inclut les actifs âgés de 18 à 65 ans affiliés à la Mutualité Sociale Agricole (MSA) (Santin et al., 2014). Ici, nous avons utilisé les données de l'enquête pilote menée en 2010. Un échantillon de 10 000 actifs affiliés à la MSA pendant au moins 90 jours en 2008 et résidant dans un des cinq départements participants a été tiré au sort ; 2 363 ont répondu au questionnaire d'inclusion (23,6 %).

La cohorte Coset-RSI inclut les actifs âgés de 18 à 65 ans affiliés au Régime Social des Indépendants (RSI). Ici, nous avons utilisé les données de l'enquête pilote menée en 2012. Un échantillon de 20 000 actifs affiliés en février 2012 à une des trois caisses régionales participantes pendant au moins 6 mois a été tiré au sort ; 2 661 ont répondu au questionnaire d'inclusion (13,3 %).

2.2. Données administratives

Des données administratives ont été obtenues par appariement pour chacune des trois cohortes. Ces données étaient disponibles pour toutes les personnes informées et n'ayant pas refusé la collecte de ces données, qu'elles aient ou non répondu au questionnaire.

Pour les trois cohortes, des données socio-démographiques (âge, sexe, région de résidence et caisse d'affiliation) ont été obtenues dans chaque base de sondage respective. De même, pour chaque cohorte ont été recueillies des données issues du Système d'Information Inter-Régimes de l'Assurance Maladie (Sniiram).

Par ailleurs, pour chaque cohorte ont été recueillies des données professionnelles dans les bases des Régimes respectifs. Ces données sont différentes selon les cohortes : elles dépendent à la fois de l'ampleur des bases de données et des spécificités propres à chaque Régime.

Les données issues du Sniiram et des bases professionnelles des Régimes sont vastes et ne présentent pas une information par individu. Des variables indicatrices ont donc été créées afin de synthétiser l'information au niveau de l'individu.

2.3. Données de questionnaire

Nous avons choisi d'estimer les prévalences pour trois variables d'intérêt communes aux trois cohortes et qui couvrent les différentes thématiques abordées par les questionnaires. Ces variables d'intérêt sont les suivantes :

- Le statut tabagique,
- La symptomatologie dépressive, mesurée par l'échelle CES-D (Radloff, 1977; Fuhrer and Rouillon, 1989),
- L'item « Vu tous mes efforts, je reçois le respect et l'estime que je mérite à mon travail » du questionnaire dit de Siegrist (Siegrist, 1996; Siegrist et al., 2009).

3. Méthodes

3.1. Correction de la non-réponse

La non-réponse totale a été traitée par repondération, en utilisant la méthode des scores (Haziza and Beaumont, 2007; Little, 1986; Eltinge and Yansaneh, 1997). Cette méthode consiste à augmenter les poids de sondage des répondants, afin qu'ils représentent les non-répondants qui ont des caractéristiques communes.

Etant donné la quantité importante de données auxiliaires, le modèle logistique qui permet d'estimer la probabilité de réponse a été construit par étapes, en sélectionnant dans un premier temps les variables associées à la non-réponse au sein de chaque type de données (socio-démographiques, médicales, professionnelles), puis en déterminant le modèle final avec les variables sélectionnées.

3.2. Estimation des prévalences

L'impact de la correction de la non-réponse est étudié en comparant les prévalences pondérées par les poids corrigés de la non-réponse aux prévalences pondérées par les poids de sondage.

4. Résultats et discussion

Nos résultats montrent qu'une position sociale élevée et un emploi stable favorisent la réponse aux enquêtes. Les personnes qui prennent soin de leur santé sont également plus enclines à répondre à nos enquêtes, dont un des sujets est la santé. A l'inverse, les personnes qui ont des problèmes de santé plus graves sont moins enclines à répondre, probablement en lien avec leur état de santé. Par ailleurs, l'âge est positivement associé à la réponse à nos enquêtes. Ces résultats sont concordants avec d'autres travaux (Goldberg et al., 2001; Reijneveld and Stronks, 1999; Goyder et al., 2002).

Un résultat marquant est que l'association entre le sexe et la réponse est captée, dans les modèles multivariables, par des covariables reflétant des consommations de soins liées au sexe.

En étudiant l'impact de la correction de la non-réponse sur les prévalences estimées, il ressort que les fumeurs et les personnes avec des symptômes dépressifs avaient moins répondu aux trois enquêtes. Ce résultat est concordant avec des études précédentes (Vercambre and Gilbert, 2012; Shahar et al., 1996). En revanche, la correction de la non-réponse n'a pas d'impact sur l'estimation du respect reçu au travail dans les trois cohortes.

Ces résultats seront très utiles pour d'autres enquêtes qui nécessitent de corriger de la non-réponse totale.

Bibliographie

Eltinge JL and Yansaneh IS. (1997) Diagnostics for Formation of Nonresponse Adjustment Cells, With an Application to Income Nonresponse in the U.S. Consumer Expenditure Survey. *Survey Methodology* 23: 33-40.

Fuhrer R and Rouillon F. (1989) La version française de l'échelle CES-D (Center for Epidemiologic Studies-Depression Scale). Description et traduction de l'échelle d'autoévaluation. / The French version of the CES-D (Center for Epidemiologic Studies-Depression Scale). *Psychiatrie et Psychobiologie* 4: 163-166.

Goldberg M, Chastang JF, Leclerc A, et al. (2001) 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. *American Journal of Epidemiology* 154: 373-384.

Goyder J, Warriner K and Miller S. (2002) Evaluating socio-economic status (SES) bias in survey nonresponse. *Journal of Official Statistics* 18: 1-11.

Haziza D and Beaumont J-F. (2007) On the Construction of Imputation Classes in Surveys. *International Statistical Review* 75: 25-43.

Little RJA. (1986) Survey Nonresponse Adjustments for Estimates of Means. *International Statistical Review* 54: 139-157.

Radloff LS. (1977) The CES-D Scale: A Self-Report Depression Scale for Research in the General Population. *Applied Psychological Measurement* 1: 385-401.

Reijneveld SA and Stronks K. (1999) The impact of response bias on estimates of health care utilization in a metropolitan area: the use of administrative data. *International Journal of Epidemiology* 28: 1134-1140.

Santin G, Geoffroy B, Benezet L, et al. (2014) In an occupational health surveillance study, auxiliary data from administrative health and occupational databases effectively corrected for nonresponse. *Journal of Clinical Epidemiology* 67: 722-730.

Shahar E, Folsom AR and Jackson R. (1996) The effect of nonresponse on prevalence estimates for a referent population: insights from a population-based cohort study. Atherosclerosis Risk in Communities (ARIC) Study Investigators. *Annals of Epidemiology* 6: 498-506.

Siegrist J. (1996) Adverse health effects of high-effort/low-reward conditions. *Journal of Occupational Health Psychology* 1: 27-41.

Siegrist J, Wege N, Puhlhofer F, et al. (2009) A short generic measure of work stress in the era of globalization: effort-reward imbalance. *International Archives of Occupational and Environmental Health* 82: 1005-1013.

Tolonen H, Helakorpi S, Talala K, et al. (2006) 25-year Trends and Socio-demographic Differences in Response Rates: Finnish Adult Health Behaviour Survey. *European Journal of Epidemiology* 21: 409-415.

Van Loon AJ, Tijhuis M, Picavet HS, et al. (2003) Survey non-response in the Netherlands: effects on prevalence estimates and associations. *Annals of Epidemiology* 13: 105-110.

Vercambre MN and Gilbert F. (2012) Respondents in an epidemiologic survey had fewer psychotropic prescriptions than nonrespondents: an insight into health-related selection bias using routine health insurance data. *Journal of Clinical Epidemiology* 65: 1181-1189.

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