

**English version following the French version.**

**TITRE DU PROJET : PREVALEST - Estimation de prévalences utilisant des données d'enquête et de données médico-administratives**

**RESPONSABLE :** Alice Guéguen, Inserm, UMS 11 Cohortes épidémiologiques en population, Villejuif

**RESUME**

**Contexte**

Conduire et évaluer des actions de prévention primaire nécessite de connaître la prévalence des maladies d'intérêt et de leurs facteurs de risque dans la population-cible. L'approche usuelle pour estimer de telles prévalences est la réalisation d'enquêtes auprès d'échantillons représentatifs. Cependant, même si les sujets sont sélectionnés par tirage au sort, leur participation à une enquête faisant appel au volontariat s'accompagne inévitablement d'effets de sélection à l'origine de biais potentiels. Quand la non réponse peut être expliquée par un ensemble de variables connues, ces biais peuvent être corrigés par des méthodes de repondération. Ces méthodes impliquent qu'une partie des données soit connues pour les répondants et les non répondants (« données auxiliaires »). L'appariement des données individuelles d'enquête avec des bases de données médico-administratives contenant des données socioéconomiques, démographiques et de santé couvrant l'ensemble de la population cible, peut permettre de modéliser avec les données auxiliaires la probabilité de participation et de corriger les estimations pour les biais dus à la non réponse.

**Objectifs**

L'objectif principal du projet PREVALEST est le développement d'une méthodologie pour la correction des biais de non réponse dans les études de prévalence de maladies et de facteurs de risque à partir des données recueillies auprès échantillons représentatifs de participants des cohortes Constances, Coset-MSA et Coset-RSI. Ces méthodes reposent sur la disponibilité de cohortes de sujets tirés au sort parmi les non participants, et à l'appariement de ces cohortes avec les données du SNIIRAM et de la Cnav permettant l'extraction de données auxiliaires pertinentes. Ces méthodes seront développées à partir de données spécifiques des cohortes Constances et Coset-MSA, mais pourront être appliquées pour la correction de biais de non-participation de toute enquête auprès d'échantillons de la population française.

**Méthodes**

Données : certaines données sont disponibles uniquement pour les participants de la cohorte à partir de questionnaires et d'examens de santé (différentes données de santé et de facteurs de risque personnels ; les données d'examens de santé sont disponibles uniquement pour Constances) ; celles qui sont extraites des bases de données nationales sont disponibles à la fois pour les participants et les non participants (données socioéconomiques, ALD, hospitalisations, absentéisme, utilisation du système de soins) et seront utilisées comme variables auxiliaires pour les repondérations afin d'estimer des prévalences.

Stratégie d'analyse statistique : les analyses seront conduites en 3 étapes, pour les cohortes Constances, Coset-MSA et Coset-RSI : (1) détermination des variables permettant de modéliser la probabilité de participation à la cohorte par régression logistique ; (2) estimations de la prévalence de facteurs de risque et de pathologies après correction pour la non réponse avec des méthodes de

repondération utilisant les données auxiliaires ; (3) analyses de sensibilité pour étudier la fiabilité des méthodes.

### **Perspectives**

La méthode la plus usuelle pour estimer la prévalence de maladies et de facteurs de risque dans une population est la réalisation d'enquêtes transversales dans des échantillons représentatifs, éventuellement répétées pour étudier les changements au cours du temps. Du fait des biais potentiels dus aux effets de sélection liés au volontariat des participants, des méthodes utilisant conjointement des données recueillies auprès des participants et des données issues des bases de données médico-administratives nationales pour produire des estimations corrigées des effets de sélection peuvent grandement contribuer à améliorer la qualité des résultats d'enquête.

## **TITLE OF THE PROJECT: PREVALEST - Prevalence estimation using data from individual surveys and administrative databases**

**HEAD OF THE TEAM:** Alice Guéguen, Inserm (French National Institute of Health and Medical Research), UMS 11 Population-based Epidemiological Cohorts Unit, Villejuif

### **SUMMARY**

#### **Background**

The definition and evaluation of primary prevention of diseases programs implies the knowledge and monitoring of the prevalence of the diseases of interest and of their risk factors in the target population. The main methodological approach for estimating prevalence is to perform surveys on representative samples of the target population. However, even if a random sample of subjects is initially selected, their voluntary participation inevitably produces selection effects and potential nonresponse biases. This can be corrected when non-response can be explained by a known set of variables by reweighting methods to compensate selection effects by subsequent adjustments for estimating the relevant parameters. These methods require that part of the data is known both on respondents and non-respondents (“auxiliary data”). Linkage of individual survey data to administrative databases covering the whole target population with health and healthcare, occupation and sociodemographic information, may allow for modelling more accurately the probability of response with available auxiliary variables to correct the estimates for nonresponse bias.

#### **Objectives**

The main objective of the PREVALEST project is to develop a methodology for correcting for nonresponse bias the estimated prevalence of risk factors and health conditions calculated from the data collected among randomly selected volunteers participating in the Constances, Coset-MSA and Coset-RSI cohorts. The methods rely on the availability of “control cohorts” of subjects randomly selected among non-participants, and the linkage of both cohorts (Constances, Coset-MSA and Coset-RSI volunteer participants and the non-participants cohorts) to the SNIIRAM and Cnav national databases for collecting pertinent auxiliary data. While the methodology will be developed using specific datasets, it could be generalized to be used on any sample of volunteer participants in France where estimates have to be corrected for nonresponse bias.

#### **Methods**

Data: some are collected for Constances, Coset-MSA and Coset-RSI participants only through questionnaires and health examinations (various health parameters and personal risk factors); those extracted from the national databases are available for both participants and non-participants samples (socioeconomic data, longterm diseases (ALD), hospital discharge, sick leaves, healthcare utilization) and will be used as auxiliary data for post-survey reweighting of prevalence estimates.

Statistical strategy: analyses will be conducted in 3 steps: (1) modeling of the probability of participation by using auxiliary data; (2) estimations of the prevalence of risk factors and diseases after correction for nonresponse; (3) sensitivity analyses to study the reliability of the method.

#### **Perspectives**

The most usual way for estimating the prevalence of diseases and risk factors in the population is to conduct cross-sectional surveys of representative samples, eventually repeated over time for

studying temporal evolutions. Considering the potential biases in prevalence estimates due to the voluntary participation of persons in health surveys, methods mixing data from specific surveys and from the national health and economic administrative databases to produce estimates corrected for selection effects can greatly improve the quality of survey results.