No 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2006 : Enjeux méthodologiques reliés à la mesure de la santé des populations

2006

Statistics Canada Statistique Canada

Canada

Recueil du Symposium 2006 de Statistique Canada
Enjeux méthodologiques reliés à la mesure de la santé des populations

Défis posés par la conception de la National Health and Nutrition Examination Survey

Leyla Mohadjer[1] et Lester R. Curtin[2]

Résumé

La National Health and Nutrition Examination Survey (NHANES) fait partie d'une série de programmes ayant trait à la santé parrainés par le National Center for Health Statistics des États-Unis. Une caractéristique unique de la NHANES est que tous les répondants de l'échantillon sont soumis à un examen médial complet. Afin de normaliser la façon dont ils sont effectués, ces examens se déroulent dans des centres d'examen mobiles (CEM). L'examen comprend des mesures physiques, des tests tels que l'examen de la vue et des dents, ainsi que le prélèvement d'échantillons de sang et d'urine pour des analyses biologiques. La NHANES est une enquête annuelle continue sur la santé effectuées auprès de la population civile des États-Unis ne résidant pas en établissement. Les principaux objectifs analytiques comprennent l'estimation du nombre et du pourcentage de personnes dans la population des États-Unis et dans des sous-groupes désignés qui présentent certaines maladies et certains facteurs de risque. Le plan d'échantillonnage de la NHANES doit permettre d'établir un juste équilibre entre les exigences liées à l'obtention d'échantillons annuels et pluriannuels efficaces et la souplesse requise pour pouvoir modifier les paramètres essentiels du plan afin de mieux adapter l'enquête au besoin des chercheurs et des décideurs qui élaborent les politiques en matière de santé. Le présent article décrit les défis associés à la conception et à la mise en œuvre d'un processus d'échantillonnage permettant d'atteindre les objectifs de la NHANES.

MOTS-CLÉS : échantillonnage à plusieurs degrés, échantillonnage par domaine, mesure pondérée de la taille, centres d'examen mobiles

1. Introduction

La National Health and Nutrition Examination Survey (NHANES) fait partie d'une série de programmes liés à la santé parrainés par les Centers for Disease Control and Prevention (CDC) des États-Unis par l'entremise du National Center for Health Statistics (NCHS). La NHANES, qui est conçue pour évaluer l'état de santé et l'état nutritionnel de la population des États-Unis ne résidant pas en établissement, est réalisée depuis plus de 45 ans. Les données recueillies sont utilisées pour estimer la prévalence des principales maladies et des principaux facteurs de risque de maladie. Les données sur la nutrition permettent une surveillance temporelle de la population nationale sur des facteurs tels que le régime alimentaire, le taux de cholestérol, l'hypertension, la carence en fer, l'anémie et l'obésité. La NHANES est également conçue pour évaluer la relation entre le régime alimentaire, la santé et l'environnement, afin de pouvoir établir le lien entre les évaluations nutritionnelles et des affections comme la maladie cardiovasculaire, le diabète, l'hypertension et l'ostéoporose.

La collecte des données de la NHANES comprend au moins trois étapes : un questionnaire de sélection des ménages, une interview et un examen médical. L'objectif principal du questionnaire de sélection est de déterminer si un membre du ménage est admissible à l'interview et à l'examen médical. Le questionnaire de sélection vise à recueillir des renseignements de base sur la composition et les caractéristiques démographiques du ménage. L'interview est conçue pour recueillir des données sur le plan du ménage, de la famille et de la personne sur les caractéristiques démographiques et socioéconomiques, la santé et les caractéristiques nutritionnelles. À la fin de l'interview, on demande au répondant de participer à un examen médical. Afin de normaliser la façon dont ils sont administrés et les protocoles, ces examens se déroulent dans un centre d'examen mobile (CEM) spécialement conçu et équipé. L'examen comprend des mesures physiques, des tests comme un examen des yeux et des dents, des mesures physiologiques et le prélèvement d'échantillons de sang et d'urine pour des analyses biologiques. Le site

[1] Leyla Mohadjer, Westat, 1650 Research Blvd., Rockville, Maryland, U.S.A. 20850, LeylaMohadjer@Westat.com

[2] Dr. Lester R. Curtin, National Center for Health Statistics, 3311 Toledo Road, Hyattsville, Maryland, U.S.A. 20782, lrc2@cdc.gov

Web de la NHANES http://www.cdc.gov/nchs/nhanes.htm fournit des renseignements détaillés sur les composantes médicales de l'enquête.

L'élaboration d'un plan d'échantillonnage efficace oblige à résoudre plusieurs questions de conception particulières à la NHANES en plus de celles qui se posent habituellement en échantillonnage. Le présent article traite des aspects uniques et compliqués du plan d'échantillonnage de la NHANES. Néanmoins, nous estimons qu'il est utile de commencer par un résumé général de ce plan d'échantillonnage, ce que nous faisons ci-après, avant de discuter de ses caractéristiques uniques.

L'échantillon de la NHANES représente l'ensemble de la population civile des États-Unis ne résidant pas en établissement. Les militaires actifs et les personnes placées en établissement ne font pas partie de la population de référence. Un plan d'échantillonnage à quatre degrés est utilisé. Les unités primaires d'échantillonnage (UPE), souvent appelées emplacements de collecte, sont sélectionnées à partir d'une base de sondage comprenant tous les comtés des États-Unis. Les UPE sont principalement des comtés uniques; dans quelques cas, des comtés adjacents sont fusionnés afin que la taille des UPE demeure supérieure à une taille minimale fixée. Les UPE de la NHANES sont sélectionnées avec probabilités proportionnelles à la taille (PPT). Chaque échantillon annuel comprend 15 emplacements de collecte.

Le deuxième degré d'échantillonnage correspond à la sélection de segments de régions (*area segments*) constitués d'îlots de recensement ou de combinaison d'îlots. En moyenne, 24 segments sont échantillonnés dans chaque UPE. L'échantillonnage est conçu de façon à ce que la taille d'échantillon soit approximativement la même dans chaque UPE et la plupart des UPE comptent exactement 24 segments. Les segments sont également sélectionnés avec probabilités proportionnelles à la taille. Les mesures de taille des segments, lorsqu'elles sont combinées aux taux de sous-échantillonnage utilisés dans les segments, fournissent des nombres approximativement égaux de personnes échantillonnées (PE) par segment, quoique la variation relative de la charge de travail soit plus grande dans les segments que dans les UPE.

Le troisième degré d'échantillonnage consiste à sélectionner les ménages et les logements collectifs non institutionnels, comme les dortoirs. Dans une UPE donnée, après la sélection de segments, toutes les unités de logement (UL) comprises dans les segments échantillonnés sont répertoriées et un sous-échantillon de ménages et de logements collectifs compris dans les UL sont désignés pour une présélection afin de repérer les PE éventuelles pour l'interview et l'examen médical. Les PE comprises dans les ménages ou les logements collectifs constituent le quatrième degré d'échantillonnage. Tous les membres admissibles d'un ménage sont répertoriés et un sous-échantillon de personnes est sélectionné. Chaque échantillon annuel comprend environ 5 000 PE ayant répondu à l'interview et subi les examens. Les examens de la NHANES requièrent du personnel hautement spécialisés, ainsi que l'analyse en laboratoire des échantillons prélevés. Par conséquent, la mise en œuvre des composantes de l'examen médical peut être très coûteuse. Afin de limiter les coûts et de réduire le fardeau de réponse, certaines composantes de l'examen médical ne sont administrées qu'à un sous-échantillon des répondants qui se présentent au CEM. Un seul algorithme de sous-échantillonnage sert à contrôler le degré de chevauchement entre les divers sous-échantillons afin qu'il soit possible d'analyser les corrélations entre les divers examens et composantes de laboratoire. L'affectation des PE aux sous-échantillons est déterminée entièrement avant l'arrivée des PE au CEM.

Les données recueillies dans le cadre des enquêtes de la NHANES ont joué un rôle extrêmement important dans l'obtention des renseignements nécessaires sur la santé et l'état nutritionnel de la population des États-Unis. Par conséquent, à partir de 1999, la NHANES est devenue une enquête annuelle permanente (Montaquila et coll., 1998). Il est essentiel d'accorder beaucoup d'attention à l'élaboration et à la tenue à jour d'un plan d'échantillonnage efficace dans le cas d'une enquête aussi importante et complexe. Le présent article décrit les défis posés par la conception et la mise en œuvre d'un processus d'échantillonnage permettant d'atteindre les multiples objectifs de la NHANES. Il porte sur le plan d'échantillonnage utilisé jusqu'en 2006 (afin de répondre aux nouvelles exigences analytiques, certains aspects du plan seront modifiés à partir de 2007).

La section 2 décrit les principaux objectifs de l'enquête et la section 3 donne un aperçu des facteurs les plus importants ayant une incidence sur le plan d'échantillonnage. La section 4 décrit les caractéristiques uniques du plan d'échantillonnage de la NHANES. Enfin, la section 5 résume brièvement l'article.

2. Principaux objectifs de la NHANES

La NHANES est une enquête annuelle permanente sur la santé réalisée auprès de la population civile des États-Unis ne résidant pas en établissement. Ses principaux objectifs sont : 1) estimer la prévalence nationale de certaines maladies et de certains facteurs de risque, 2) estimer les distributions de référence dans la population nationale de certains paramètres de la santé et contaminants présents dans l'environnement, 3) décrire et étudier les raisons des tendances séculaires de certaines maladies et de certains facteurs de risque, 4) contribuer à la compréhension des causes des maladies, 5) étudier l'évolution naturelle de certaines maladies, 6) étudier la relation entre le régime alimentaire, la nutrition, l'environnement, la génétique et la santé et 7) explorer les questions de santé publique naissantes.

3. Principaux facteurs ayant une incidence sur le plan d'échantillonnage

Comme nous l'avons mentionné plus haut, une caractéristique unique de la NHANES est qu'un examen médical complet est effectué dans les centres d'examen mobiles. En outre, le plan doit produire des tailles d'échantillon efficaces pour un grand nombre de sous-domaines de la population générale. De nombreuses caractéristiques de la santé et de la nutrition diffèrent considérablement selon l'âge, le sexe et la race ou ethnicité, et varient selon la situation de revenu. Par conséquent, la plupart des analyses des données de la NHANES sont effectuées pour des groupes d'âge particuliers dans divers sous-groupes socioéconomiques de la population. L'enquête est donc conçue afin de produire des tailles d'échantillon efficaces pour un très grand nombre de sous-domaines de la population des États-Unis.

En général, le plan d'échantillonnage de la NHANES doit permettre d'établir un juste équilibre entre les exigences liées à l'obtention d'échantillons de sous-domaine efficaces, d'une part, et d'une charge de travail pouvant être gérée par le personnel du CEM, d'autre part, tout en maintenant les taux de réponse aussi élevés que possible. Plus précisément, le plan d'échantillonnage de la NHANES vise à 1) obtenir des tailles préspécifiées d'échantillons autopondérés pour un ensemble d'environ 75 sous-domaines prédésignés, 2) produire une taille d'échantillon par UPE donnant lieu à une charge de travail gérable pour les intervieweurs et le personnel du CEM, 3) obtenir des échantillons susceptibles de produire des taux de réponse élevés, 4) être aussi rentable que possible, 5) produire des échantillons annuels efficaces, 6) permettre le cumul d'échantillons au cours du temps, surtout pour les sous-domaines ou les maladies rares et 7) être souple afin de permettre la modification des paramètres clés, y compris les domaines d'échantillonnage et les taux d'échantillonnage en vue de répondre aux questions d'actualité en matière de santé.

La suite de la section est consacrée à un bref résumé de l'incidence de chacun de ces objectifs sur la conception et la mise en œuvre de la NHANES.

Sous-domaines de la NHANES – Le plan d'échantillonnage de la NHANES permet d'atteindre un niveau préspécifié de précision pour les données transversales et les comparaisons au cours du temps pour un ensemble de sous-domaines prédésignés. Plus précisément, 77 domaines d'échantillonnage (dans l'échantillon de 2006) sont définis en fonction de la race/ethnicité, du sexe, de l'âge, du revenu et de la situation de grossesse. Les Noirs, les Mexicains, les très jeunes enfants, les adolescents, les personnes âgées, les femmes enceintes et les personnes à faible revenu sont suréchantillonnés.

Lorsqu'on évalue que les estimations de totaux d'univers pour l'ensemble de population sont de la plus haute importance, la meilleure estimation disponible du total de population est utilisée comme mesure de taille dans le processus d'échantillonnage. Dans le cas de la NHANES, où l'on s'intéresse à des sous-domaines de la population totale, une autre mesure de taille est nécessaire pour améliorer l'exactitude des estimations et permettre de mieux contrôler la taille de l'échantillon. La section 4 décrit les mesures de taille utilisées pour l'échantillonnage des UPE et des segments dans le cadre de la NHANES.

L'objectif du suréchantillonnage (en utilisant des probabilités de sélection différentielles) est d'obtenir un échantillon contenant des nombres proportionnellement plus élevés de membres de certains sous-domaines de population que n'en contient la population. Le but est d'obtenir des tailles d'échantillon adéquates pour faire des inférences sur des sous-domaines représentant une proportion relativement faible de l'univers d'intérêt total et de le

faire de façon à réduire au minimum les variances compte tenu du budget de l'enquête. Diverses stratégies de suréchantillonnage sont utilisées, selon le domaine d'intérêt. Par exemple, le suréchantillonnage des sous-populations minoritaires est réalisé par stratification des régions géographiques selon la concentration de ces groupes minoritaires et par sélection des segments à un taux plus élevé dans les régions à forte concentration. Par ailleurs, un grand échantillon de sélection peut être nécessaire pour le suréchantillonnage des personnes appartenant à des groupes d'âge particuliers. La sous-section suivante portant sur les ratios des coûts décrit pourquoi les procédures de suréchantillonnage appliquées dans le cas de la NHANES diffèrent de celles de nombreuses enquêtes par sondage à base aréolaire commune.

Centres d'examen mobiles (CEM) – Les CEM sont constitués de quatre remorques spécialement conçues et équipées et contient tout l'équipement médical. Chaque remorque mesure environ 45 pieds de long et 10 pieds de large. Un camion tracteur mobile conduit les remorques d'un emplacement à un autre. Les CEM se rendent aux divers emplacements de collecte à travers le pays. Les remorques sont installées côte à côte et jointes par des passerelles fermées. L'espace à l'intérieur du CEM est divisé en salles pour permettre le respect de la vie privée durant les examens et les interviews. L'examen comprend diverses évaluations et mesures physiques et dentaires, des analyses de laboratoire et d'autres interviews sur la santé.

Étant donné les difficultés logistiques associées à l'utilisation des CEM, pour chaque emplacement échantillonné, la taille d'échantillon doit être déterminée d'avance et considérée comme étant fixe, afin qu'il soit possible de planifier les opérations sur le terrain de manière efficace et pratique. En outre, il est nécessaire d'établir un calendrier ferme pour chaque emplacement de collecte, afin que les rendez-vous puissent être pris pour les examens. Il est impossible de modifier le calendrier, car celui-ci doit être coordonné avec la visite du CEM à d'autres emplacements dont le calendrier est également préétabli.

Taux de réponse – Obtenir des taux de réponse élevés est une préoccupation dans le cas de presque toutes les enquêtes par sondage. Pour la NHANES, le défi est particulièrement grand, étant donné la portée des interviews et des examens. D'ailleurs, on a offert une rémunération pour améliorer les taux de réponse. En outre, la NHANES comprend un programme de relations communautaires important notamment des contacts avec les organismes locaux et les personnes dont il faut obtenir la coopération et la couverture dans les médias locaux pour joindre le plus grand nombre possible de PE. Parmi les questions soulevées par le plan d'échantillonnage, la sélection d'échantillons de plus grande taille dans les ménages échantillonnés est une approche qui a eu un effet favorable sur les taux de réponse. L'un des facteurs que l'on croit à l'origine de l'accroissement des taux de réponse dans les ménages comptant plusieurs PE est le dédommagement reçu par chaque personne pour son temps et sa participation et qu'il est généralement plus commode pour les membres du ménage de se rendre ensemble au CEM.

Par conséquent, la NHANES est conçue en vue de maximiser le nombre de PE par ménage. Cette approche est faisable dans le cas d'études de ce genre où l'échantillon est constitué d'un grand nombre de sous-domaines. Autrement dit, l'effet de la mise en grappes dans les ménages n'est pas très préoccupant, parce que la plupart des analyses sont faites dans des sous-domaines âge-sexe particuliers (ou dans certains groupes limités de sous-domaines) et que la mise en grappes dans les ménages est généralement faible au niveau du sous-domaine.

Ratio des coûts – Dans les échantillons d'enquête aréolaire, le coût des opérations de collecte des données sur le terrain comprend le coût d'établissement des listes d'unités de logements, la présélection des ménages pour repérer les répondants admissibles et la réalisation de l'interview. Dans le cas de la NHANES, la composante de l'interview comprend l'interview auprès du ménage et l'examen au CEM. La NHANES requiert de l'équipement médical, du personnel et des processus de laboratoire hautement spécialisés. Par conséquent, le coût d'un examen est très élevé comparativement à d'autres coûts d'enquête. En fait, le coût de l'établissement des listes et de la présélection ne représente que de 3 % à 4 % du coût de l'interview et de l'examen. Ce ratio des coûts (le coût de l'interview et de l'examen relativement au coût de l'établissement des listes et de la présélection) a une incidence très importante sur le plan d'échantillonnage de la NHANES.

Comme nous l'avons souligné plus haut, une méthode de suréchantillonnage doit être appliquée à un grand nombre de sous-domaines prédésignés de l'enquête afin d'atteindre les tailles requises d'échantillons. Pour les populations minoritaires, il est possible de réduire considérablement la présélection en suréchantillonnant les régions à forte concentration de groupes minoritaires. En général, un plan optimal est élaboré en déterminant l'effet, sur les coûts et la variance, de diverses procédures d'échantillonnage et en choisissant celle qui réduit la variance au minimum pour

un coût fixé. Lors de l'évaluation des compromis entre les coûts et la variance, supposons qu'une stratégie de suréchantillonnage particulière réduise le nombre de ménages dont il faut dresser la liste des membres et qui doivent être soumis à la présélection, tout en accroissant la variance de la plupart des statistiques. Les économies réalisées grâce à la réduction des coûts d'établissement des listes et de présélection pourraient être utilisées pour accroître la taille de l'échantillon et, donc, réduire la variance. Cependant, dans la NHANES, établir la liste des membres et présélectionner des ménages représentent une très faible fraction du coût, de sorte que les économies qui y sont liées devraient être très importantes pour justifier un accroissement modéré de la variance. Par conséquent, les procédures de suréchantillonnage établies pour l'enquête reflètent le ratio des coûts de la NHANES et diffèrent de celles appliquées dans le cas d'enquêtes aréolaires typiques.

Plan d'échantillonnage souple – Un objectif clé de la NHANES est d'explorer les questions d'actualité en matière de santé publique. L'enquête doit être souple afin que l'on puisse l'adapter à l'évolution des exigences et aux nouveaux défis. Donc, le plan d'échantillonnage doit tenir compte à la fois du besoin d'échantillons de sous-domaines efficaces et de souplesse afin de pouvoir modifier les paramètres clés de l'enquête. Jusqu'à présent, le plan d'échantillonnage existant de la NHANES a permis d'intégrer de petites modifications des définitions des sous-domaines et des taux d'échantillonnage, lorsque ces modifications ont été apportées après la sélection des UPE. Toutefois, dans des circonstances extrêmes, des changements importants dans les définitions des sous-domaines ou les exigences relatives aux tailles d'échantillons nécessiteraient la sélection d'un nouvel échantillon d'UPE.

Échantillons annuels et pluriannuels – Afin de faciliter les couplages éventuels avec d'autres enquêtes de grande portée, de maintenir la souplesse du plan d'échantillonnage et de permettre la production d'estimations annuelles pour de grands sous-domaines, la NHANES est devenue une enquête annuelle permanente à partir de 1999. Les déplacements requis aux États-Unis pour obtenir des échantillons annuels représentatifs de la population nationale posent un défi de taille. Trois CEM, dont, en tout temps, deux sont stationnés dans des UPE et le troisième se déplace, fonctionnent selon un calendrier minutieusement établi afin de répondre aux exigences du plan de l'étude.

Dans le cas de toute enquête, la capacité de faire des inférences significatives dépend à la fois de la précision des estimations proprement dites et de la précision des estimations de la variance des estimations utilisées dans l'analyse. L'une des principales limites de l'échantillon annuel de la NHANES est le petit nombre d'UPE (15 par année), qui donne un petit nombre de degrés de liberté pour l'estimation et l'analyse, si bien que les estimations de la variance par rapport au plan sont relativement imprécises. En outre, les tailles effectives d'échantillons pour la plupart des sous-domaines sont trop faibles. La plupart des analyses pour le sous-domaine devront être faites sur un certain nombre d'échantillons annuels cumulés afin d'obtenir la précision et la puissance statistique suffisantes pour les comparaisons. Les procédures suivies pour combiner les échantillons annuels doivent être relativement simples et adaptées aux progiciels du commerce afin de maximiser l'utilité pour une grande variété d'utilisateurs des données de la NHANES. Par conséquent, il est essentiel de concevoir un plan d'échantillonnage qui permette de cumuler efficacement des échantillons annuels au cours des années.

4. Caractéristiques uniques du plan d'échantillonnage de la NHANES

Les facteurs décrits à la section 3 jouent un rôle important dans l'élaboration du plan d'échantillonnage et créent certaines caractéristiques qui sont uniques au plan de la NHANES. Ces caractéristiques uniques sont les suivantes : 1) mesure pondérée de la taille des UPE et des segments, 2) très grand échantillon de présélection, 3) échantillons annuels et pluriannuels efficaces, 4) nombre maximisé de personnes échantillonnées par ménage, 5) tailles d'échantillon contrôlées pour les UPE, 6) attribution progressive de l'échantillon de l'UPE, 7) méthodes spéciales pour tenir compte de la détérioration de l'efficacité du plan de sondage optimal au cours du temps et 8) méthodes spéciales afin de réduire le risque de divulgation de données par identification géographique.

Suit une brève description des caractéristiques uniques du plan d'échantillonnage de la NHANES.

Mesures de la taille – Dans la NHANES, la taille d'échantillon doit être suffisamment grande pour produire une charge de travail efficace pour chaque UPE, compte tenu du temps et du coût de déplacement d'un CEM entre deux emplacements de collecte et du temps nécessaire pour monter le CEM et le démonter pour le déplacement. Certaines études ont montré qu'un nombre moyen d'environ 340 PE examinées est un nombre approximativement optimal qui

produit le nombre maximal d'UPE, tout en maintenant la taille d'échantillon suffisamment grande dans chaque région pour justifier les coûts du déménagement du CEM. En outre, dans le cas de la NHANES, les UPE sont habituellement définies comme des comtés individuels pour réduire le temps de déplacement des répondants qui se rendent au CEM et donc accroître la probabilité d'obtenir des taux de réponse élevés.

L'échantillon de la NHANES est conçu pour produire un échantillon autopondéré pour chaque sous-domaine échantillonné, tout en créant une charge de travail efficace pour chaque UPE. Les UPE et les segments sont sélectionnés avec une probabilité proportionnelle à la taille pondérée qui reflète la population de l'UPE dans les sous-domaines d'intérêt. La probabilité de sélection d'une UPE détermine le taux maximal auquel chaque personne résidant dans cette UPE particulière peut être sélectionnée. Voir le document *Vital and Health Statistics, Series 2, No. 113, September 1992, CDC/NCHS*, consultable à http://www.cdc.gov/nchs/products/pubs/pubd/series/sr02/120-101/120-101.htm pour une description des mesures de taille utilisées dans le cadre de la NHANES.

Échantillons annuels et pluriannuels – Un moyen de réaliser des échantillons annuels représentatifs de la population nationale consiste à sélectionner un échantillon indépendant d'UPE chaque année. Pour la NHANES, étant donné le nombre limité d'UPE et le fait que ces dernières sont sélectionnées avec probabilité proportionnelle à la taille, cette approche donnerait vraisemblablement lieu à un chevauchement important des UPE d'une année à l'autre. Le chevauchement des échantillons, même sur le plan des UPE, pourrait entraîner une perte de précision des estimations calculées d'après les données de l'enquête si les échantillons de plusieurs années de référence sont combinés (à cause de l'accroissement de la mise en grappes dans l'échantillon). Donc, au lieu d'échantillonner les UPE indépendamment chaque année, il a été décidé de sélectionner pour la NHANES un échantillon de six ans, à partir d'une structure hiérarchique de strates principales et secondaires (décrites plus loin), puis à affecter une UPE provenant de chaque strate principale chaque année. Cette structure hiérarchique de l'échantillon de six ans évite le chevauchement des UPE non autoreprésentatives durant les six années.

Le plan de stratification de l'échantillon de six ans de la NHANES est élaboré en ayant pour objectif principal l'efficacité de l'échantillon de six ans, ainsi que celle des échantillons annuels et des échantillons mobiles de deux ans et de trois ans. Le plan de stratification est conçu de façon que les UPE formant les échantillons annuels et pluriannuels soient réparties uniformément en fonction de certaines caractéristiques géographiques et démographiques.

Pour l'enquête courante, 12 strates principales ont été définies d'après les caractéristiques géographiques et la situation de région statistique métropolitaine (RSM) des UPE. Soixante-douze strates secondaires ont été définies d'après les caractéristiques démographiques des UPE. Chaque strate principale comprenait six strates secondaires et une UPE a été sélectionnée dans chacune de ces strates finales. Dans chaque strate principale, les strates secondaires ont été appariées. Chaque paire a été assignée aléatoirement aux années de référence de l'étude avec un intervalle de trois ans. L'affectation des paires aux ensembles particuliers d'années de référence et l'affectation des années de référence dans les paires a été faite aléatoirement dans la première strate principale et le même schéma a été suivi dans toutes les autres.

Ce plan de stratification permet de produire des estimations annuelles et pluriannuelles efficaces sans compromettre l'efficacité des estimations sur six ans. L'échantillon de six ans est établi selon un plan d'échantillonnage d'une UPE par strate secondaire et chaque échantillon annuel, selon un plan d'échantillonnage d'une UPE par strate principale.

Les strates secondaires ont été construites de façon qu'elles soient de taille égale dans la mesure du possible (en fonction de la mesure de taille totale). Une procédure aléatoire est utilisée pour affecter les UPE aux échantillons annuels. Les UPE autoreprésentatives sont triées en fonction de la mesure de taille et les UPE non autoreprésentatives sont triées selon l'ordre de sélection. Les UPE sont ensuite appariées et l'année est attribuée selon un schéma déterminé aléatoirement. Les UPE de chaque paire sont affectées avec un intervalle de trois ans.

Nombre maximisé de personnes échantillonnées par ménage – Après avoir obtenu l'échantillon de ménages présélectionnés, on sélectionne un échantillon de personnes qui seront interviewées et soumises à l'examen médical. La liste de tous les membres admissibles d'un ménage est dressée et un sous-échantillon de personnes est sélectionné en fonction du sexe, de l'âge, de la race ou de l'ethnicité et du revenu (toutes les femmes enceintes sont sélectionnées avec certitude). Les personnes échantillonnées sont sélectionnées à des taux établis pour faire en sorte que les tailles d'échantillons cibles par sous-domaine soient atteintes.

L'échantillon de PE est sélectionné de façon à maximiser le nombre moyen de personnes échantillonnées par ménage afin d'accroître le taux global de réponse à l'enquête. Si l'on recourait à des sélections aléatoires indépendantes pour les sous-domaines, dans la plupart des cas, une seule personne serait sélectionnée par ménage et la taille moyenne d'échantillon par ménage serait assez faible, à peine supérieure à 1. Par conséquent, au lieu d'une randomisation non limitée, on utilise une procédure pseudo-aléatoire afin de maximiser le nombre de PE par ménage. Consulter Waksberg, Mohadjer (1991) pour une description de l'approche.

Tailles contrôlées d'échantillons par UPE – La taille d'échantillon dans chaque UPE (emplacement de collecte) qui est effectivement générée d'après un échantillon autopondéré dans chaque domaine est basée sur un certain nombre d'hypothèses, dont la distribution par âge et par race/ethnicité de la population de l'UPE. Ces hypothèses ne sont vérifiées qu'approximativement. Une fois calculées, les tailles d'échantillons, elles sont traitées comme des quotas et le nombre de PE dans chaque emplacement de collecte est forcé de correspondre étroitement au quota. On procède ainsi afin de s'assurer que les opérations sur le terrain soient gérables et efficaces. Il est nécessaire d'établir un calendrier ferme pour chaque emplacement de collecte afin que des rendez-vous puissent être donnés pour les examens des PE. Le calendrier tient naturellement compte du nombre prévu de PE à chaque emplacement de collecte. Comme nous l'avons signalé plus haut, il est difficile de modifier le calendrier d'un emplacement, puisqu'il doit être coordonné avec les visites du CEM à d'autres emplacements dont le calendrier est également préétabli.

Il n'y a aucun moyen de savoir d'avance si le quota assigné pour un emplacement particulier est plus faible ou plus élevé que celui que l'on obtiendrait à partir d'échantillons autopondérés dans les divers domaines. L'incertitude tient en partie au fait que la mesure de taille utilisée pour sélectionner l'échantillon est basée sur le recensement décennal le plus récent et n'est donc peut-être pas à jour. Le problème est compliqué davantage par les variations des taux de réponse d'un emplacement de collecte à l'autre, ainsi que par la variation d'échantillonnage du nombre de PE identifiées. Par conséquent, il est nécessaire d'utiliser une méthode d'échantillonnage qui peut produire des échantillons un peu plus grands ou un peu plus petits que ceux résultant de l'application des taux d'échantillonnage autopondérés.

Affectation séquentielle de l'échantillon dans chaque emplacement de collecte – Afin de réaliser l'objectif susmentionné, on sélectionne dans chaque emplacement de collecte un échantillon initial en appliquant des taux d'échantillonnage 50 % plus élevés que ceux requis pour obtenir les tailles d'échantillon cibles dans chaque domaine. Chaque échantillon initial d'emplacement de collecte est ensuite divisé en un groupe de sous-échantillons. Chaque sous-échantillon est un sous-échantillon systématique de l'échantillon initial, où les ménages sont ordonnés selon le numéro de segment et un numéro de série provisoire, basé sur les caractéristiques géographiques, avant le sous-échantillonnage. Donc, chaque sous-échantillon recoupe l'ensemble des segments, sauf si l'on est limité par la taille d'échantillon.

En règle générale, le sous-échantillon de 50 % (c.-à-d. le sous-échantillon A) est le premier qui est attribué aux intervieweurs. Le rendement pour ce sous-échantillon est surveillé et utilisé pour projeter les estimations du nombre total prévu de PE lorsque la sélection de ce sous-échantillon sera achevée. D'après ces chiffres, des sous-échantillons supplémentaires sont attribués au besoin. L'échantillon est surveillé quotidiennement afin de déterminer si l'attribution de sous-échantillons supplémentaires est nécessaire.

Le problème opérationnel que pose la procédure de surveillance du rendement de l'échantillon tient au fait qu'elle ne permet pas de contrôler entièrement les tailles d'échantillon des sous-domaines. La distribution des sous-domaines diffère, dans une certaine mesure, des nombres prévus d'après les données de recensement les plus récentes (utilisées pour calculer les taux d'échantillonnage). Les leçons tirées de la NHANES indiquent qu'il faut s'attendre à certains changements de population qui ont une incidence sur les tailles d'échantillon. D'autres facteurs qui influent sur le rendement d'échantillon de sous-domaine sont les profils de non-réponse et de sous-dénombrement dans les emplacements de collecte. Une option en vue de corriger le déficit (ou l'excédent) dans

les tailles d'échantillon de sous-domaine consiste à modifier les taux d'échantillonnage pour les futurs emplacements de collecte. Cependant, de tels changements augmenteront l'hétérogénéité des poids d'échantillonnage, donc auront un effet indésirable sur la précision des estimations pour le sous-domaine et ne sont pas conseillés, sauf dans des circonstances extrêmes.

Traitement de la détérioration de l'efficacité du plan optimal au cours du temps dans un échantillon étroitement contrôlé – Dans le cas des échantillons aréolaires, la pratique habituelle consiste à dresser la liste de tous les ménages dans les segments d'échantillon et à appliquer un taux d'échantillonnage préspécifié aux ménages énumérés. Cette approche donne à tous les ménages la probabilité de sélection souhaitée. Par exemple, si le taux d'échantillonnage est de 50 %, alors la moitié des unités de logement énumérées dans les segments seront incluses dans l'échantillon. Si le nombre d'unités de logement a triplé en raison de nouvelles constructions (c.-à-d. des unités de logement construites depuis le recensement décennal le plus récent), le même taux d'échantillonnage produira un nombre trois fois plus grand d'interviews et d'examens médicaux que le nombre prévu au départ. Des changements aussi spectaculaires de la taille des segments n'est pas surprenante lorsque la période de collecte des données est ultérieure de plusieurs années au recensement décennal le plus récent pour lequel des fichiers de données sont disponibles.

Dans le cas de la NHANES, les tailles d'échantillons ne peuvent pas varier fortement, en raison des calendriers établis pour les CEM. Le sous-échantillonnage effectué dans les UPE pour essayer d'obtenir des échantillons de même taille dans toutes les UPE n'est pas recommandé non plus, car il introduirait des facteurs de pondération inégaux qui réduiraient l'efficacité de l'échantillon.

Le programme de la NHANES a utilisé deux méthodes pour mettre à jour les mesures de taille des segments, à savoir 1) la création de segments de constructions neuves et 2) l'échantillonnage à deux phases pour mettre à jour la mesure de taille. À l'heure actuelle, une troisième approche consistant à acheter des listes d'adresses commerciales pour mettre à jour la mesure de taille dans un plan d'échantillonnage à deux phases est à l'étude.

Selon l'approche des nouvelles constructions (Bell et coll., 1999), les unités qui viennent d'être construites sont exclues des segments de région. De nouveaux segments sont créés en se basant sur l'information disponible d'après le recensement sur les permis émis pour des nouvelles constructions depuis le recensement décennal le plus récent. Les segments de nouvelles constructions comprennent des grappes de permis de construire émis durant un ou plusieurs mois contigus par un bureau d'octroi de permis de construire. Les fichiers de l'enquête sur ces permis menée par le Census Bureau sont utilisés comme sources de données sur le nombre de permis de construire résidentiels émis par les bureaux d'octroi de ces permis.

L'échantillonnage à deux phases est utilisé dans un certain nombre d'applications statistiques. L'une de ces applications est la mise à jour d'une base de sondage lorsque l'échantillon doit être sélectionné en fonction d'une mesure de taille, mais qu'une estimation fiable de la mesure de taille n'est pas disponible. Selon cette approche, un échantillon plus grand d'unités (dans le cas de la NHANES, les unités sont les segments) est sélectionné. Une valeur mise à jour de la mesure de taille est alors recueillie pour cet échantillon plus grand (également appelé échantillon de première phase). L'échantillon final d'unités (segments) est sélectionné à partir de l'échantillon de première phase en utilisant la mesure de taille mise à jour.

À compter de 2000, la mesure de taille des segments de la NHANES a été mise à jour (pour les emplacements de collecte pour lesquels cette mise à jour semblait nécessaire) en utilisant une méthode d'échantillonnage à deux phases (Montaquila et coll., 1999). Dans ces cas, les personnes chargées de dresser les listes de logements se rendent dans l'emplacement de collecte pour obtenir un dénombrement des unités de logement (UL) dans chaque segment de l'échantillon de première phase. Partant de ces dénombrements, une mesure de taille mise à jour reflétant le ratio du nombre actuel d'UL au nombre prévu d'UL est calculée pour chaque segment de première phase. L'échantillon final de segments est alors sélectionné par sous-échantillonnage des segments de première phase en utilisant la mesure de taille mise à jour.

Risque de divulgation des données par identification géographique – Dans le monde d'aujourd'hui, les questions de confidentialité et le risque de divulgation des données posent de réels défis aux organismes qui parrainent les enquêtes. La capacité d'identifier les répondants à une enquête, d'après des combinaisons uniques de variables disponibles dans un seul fichier de données ou en couplant diverses bases de données est un grave sujet de

préoccupation. Il en est particulièrement ainsi de la NHANES, étant donné la grande quantité de renseignements de nature délicate recueillis sur chaque personne échantillonnée et le petit nombre d'UPE dans l'échantillon. Par conséquent, le risque de divulgation est évalué sur deux fronts, à savoir la divulgation géographique et la divulgation d'après les caractéristiques individuelles. Diverses méthodes (diffusion limitée ou suppression de données) sont utilisées par la NHANES pour masquer les caractéristiques individuelles qui ont un grand risque d'être à l'origine de l'identification des personnes faisant partie de l'échantillon. Les éléments de données délicats, à diffusion limitée ou non diffusés sont disponibles dans les centres de données de recherche (CDR). À l'heure actuelle, seules des estimations nationales peuvent être produites d'après les fichiers de microdonnées à grande diffusion, et les analyses géographiques détaillées doivent être faites dans le CDR.

Bien que l'on ne puisse produire que des estimations nationales, l'estimation directe des erreurs d'échantillonnage pour ces estimations nécessite la diffusion des variables de plan, comme les identificateurs de strate et d'UPE. Habituellement, ces variables indiquent qu'un groupe de personnes échantillonnées vivent toutes dans le même comté, mais n'identifie pas le comté. La divulgation géographique est une cause de souci particulier, car, dans le cas de la NHANES : 1) le nombre d'UPE est petit, 2) les UPE sont limitées géographiquement à un comté et 3) une grande quantité de démarches de relations communautaires sont menées dans chaque UPE pour améliorer les taux de réponse. Le programme de relations communautaires comprend la prise de contact avec divers organismes et personnes à chaque emplacement de collecte pour essayer d'obtenir leur appui et l'utilisation des médias (journaux, télévision et radio) afin de rejoindre le plus grand nombre possible de PE. Il est, par conséquent, relativement facile de déterminer les comtés et les années de référence où les périodes de collecte de la NHANES ont eu lieu. La composition raciale ou ethnique d'un comté, ainsi que la situation de région statistique métropolitaine ou non métropolitaine fournissent des renseignements suffisants pour apparier correctement une liste de comtés connus à des groupes identifiés comme représentant une grappe de comtés dans le fichier de données à grande diffusion. Pour limiter la divulgation géographique, on recourt à des méthodes de permutation probabiliste des enregistrements au deuxième degré d'échantillonnage (permutation des segments) afin de créer des unités à variance masquée. L'objectif est de réduire le risque d'identifier des individus en masquant leur emplacement. Consulter Park et coll. (2006) pour une description des procédures de permutation appliquées à l'échantillon de la NHANES.

5. Sommaire et conclusion

Une caractéristique unique de la NHANES est l'examen médical complet effectué dans les CEM. En outre, l'enquête est conçue afin de produire des tailles d'échantillon efficaces pour un grand nombre de sous-domaines de la population des États-Unis, puisque la plupart des analyses des données de la NHANES sont faites pour des groupes d'âge définis dans divers sous-groupes socioéconomiques de la population. Donc, le plan d'échantillonnage de la NHANES doit permettre d'établir, d'une part un équilibre entre les exigences liées à l'obtention d'échantillons de sous-domaine efficaces et, d'autre part d'une charge de travail efficace pour le personnel d'examen du CEM, tout en maintenant les taux de réponse aussi élevés que possible. En outre, le plan doit être aussi rentable que possible, produire des échantillons annuels efficaces et permettre le cumul des échantillons au cours du temps pour les sous-domaines ou les maladies rares. De surcroît, le plan doit être souple pour permettre de modifier les paramètres clés, y compris les domaines d'échantillonnage, et les taux d'échantillonnage pour répondre aux questions d'actualité en matière de santé.

Les exigences susmentionnées se traduisent par un plan d'échantillonnage très complexe dont certaines caractéristiques sont propres à la NHANES. En particulier, l'échantillon courant est conçu afin de produire des échantillons efficaces annuels et mobiles de deux ans et de trois ans. La NHANES utilise des UPE pondérées et des mesures de taille de segment pour produire des échantillons autopondérés pour chaque sous-domaine, tout en produisant une charge de travail efficace dans chaque UPE. Une fois que les tailles d'échantillon sont calculées, elles sont traitées comme des quotas. Les tailles d'échantillon sont strictement contrôlées dans chaque UPS afin que les opérations sur le terrain soient gérables et efficaces. Un très grand échantillon de sélection est utilisé afin de suréchantillonner la plupart des sous-domaines d'âge et de revenu, et le suréchantillonnage des régions à forte concentration est utilisé pour certains sous-domaines minoritaires très rares. L'échantillon de PE est sélectionné selon une méthode pseudo-aléatoire afin de maximiser le nombre moyen de personnes échantillonnées par ménage, car cela a semblé accroître le taux global de réponse au cours des enquêtes précédentes.

Références

Bell, B., Mohadjer, L., Montaquila, J., et Rizzo, L. (1999). "Creating a frame of newly constructed units for household surveys". *Proceedings of the Survey Research Methods Section, American Statistical Association*.

Park, H. Dohrmann, S., Montaquila, J., Mohadjer, L., et Curtin, R.L. (2006). "Reducing the risk of data disclosure through area masking: Limiting biases in variance estimation". *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

Montaquila, J., Bell, B., Mohadjer, L., et Rizzo, L. (1999). "A methodology for sampling households late in a decade". *Proceedings of the Survey Research Methods Section, American Statistical Association*.

Montaquila, J., Mohadjer, L., et Khare, M. (1998). "The enhanced sample design of the future National Health and Nutrition Examination Survey (NHANES)". *Proceedings of the Survey Research Methods Section, American Statistical Association*.

Waksberg, J., et Mohadjer, L. (1991). "Within household sampling". *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 350-355.