Composante du produit nº 12-001-X
Division des méthodes d'enquêtes auprès des entreprises

Article

Trouver l'équilibre entre les divers objectifs du plan d'échantillonnage de la National Health and Nutrition Examination Survey

par Leyla Mohadjer et Lester R. Curtin

Juin 2008

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, juin 2008
Vol. 34, N° 1, pp. 131-140
Statistique Canada, N° 12-001-X au catalogue

Trouver l'équilibre entre les divers objectifs du plan d'échantillonnage de la National Health and Nutrition Examination Survey

Leyla Mohadjer et Lester R. Curtin [1]

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 ont lieu 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 réalisée 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 des États-Unis par l'entremise du National Center for Health Statistics (NCHS). La NHANES est utilisée depuis plus de 45 ans pour évaluer l'état de santé et l'état nutritionnel de la population civile des États-Unis ne résidant pas en établissement. Les données recueillies servent à 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 en ce qui a trait à 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 en vue d'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, à savoir 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 au niveau 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, il est demandé 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 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 a nécessité la résolution de 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 ne résidant pas en établissement des États-Unis. Les militaires d'active et les personnes placées

1. Leyla Mohadjer, Westat, 1650 Research Blvd., Rockville, Maryland, États-Unis 20850. Courriel : LeylaMohadjer@Westat.com; Lester R. Curtin, National Center for Health Statistics, 3311 Toledo Road, Hyattsville, Maryland, États-Unis 20782. Courriel : lrc2@cdc.gov.

en établissement ne font pas partie de la population d'inférence. L'échantillonnage de la NHANES ne se fait pas selon un plan à probabilité égale; les fractions d'échantillonnage sont établies de manière à « suréchantillonner » les Mexicano-Américains (et les Hispaniques dans les échantillons de 2007 et au-delà), les Américains de race noire, les Américains de race blanche et autres à faible revenu, les personnes de moins de 20 ans et les personnes de plus de 60 ans. Un plan d'échantillonnage à quatre degrés est utilisé. Les unités primaires d'échantillonnage (UPE), souvent appelées emplacements de collecte *(stands)*, 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. La base de sondage de la NHANES contient près de 3 000 UPE qui sont échantillonnées avec probabilité proportionnelle à une mesure de 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. Comme les UPE ne sont pas toutes de même taille, la taille des segments et le nombre de segments par UPE présentent une certaine variabilité. Chaque segment est formé d'environ 150 ménages (ou unités de logements) en moyenne, environ 5 000 segments sont créés dans chaque UPE et 24 segments sont échantillonnés en moyenne. 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 probabilité proportionnelle à la taille. Les mesures de taille des segments, lorsqu'elles sont combinées aux taux de sous-échantillonnage 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 d'identifier 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. Les taux de sous-échantillonnage des ménages dans les segments et de personnes dans les ménages sont prédéterminés. La combinaison de la présélection et des taux d'échantillonnage différentiels fournit l'accroissement de taille d'échantillon pour les sous-domaines démographiques d'intérêt (âge, sexe, race/ethnicité et revenu). Ainsi, dans les 30 UPE dans lesquelles les données ont été recueillies durant le cycle de collecte de deux ans de 2005-2006, 716 segments ont été sélectionnés et 26 529 ménages ont été tirés pour la présélection. Après la présélection pour déterminer la composition selon l'âge, le sexe et la race/ethnicité et la situation de faible revenu, une ou plusieurs personnes ont été sélectionnées dans 6 372 ménages pour faire partie de l'échantillon. En tout, 12 862 personnes ont été sélectionnées, et parmi celles-ci, 9 950 ont répondu à l'interview et subi les examens.

Les examens de la NHANES requièrent du personnel hautement spécialisé, 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 qu'elles se présentent 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 (Mohadjer et Khare 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 ont été 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 ne résidant pas en établissement des États-Unis. Ses principaux objectifs sont : 1) estimer la prévalence nationale de certaines maladies et 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 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. Beaucoup de caractéristiques de la santé et de la nutrition diffèrent considérablement selon l'âge, le sexe et la race ou ethnicité, et dépendent aussi de 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 socio-économiques 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 des tailles 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 nouvelles questions en matière de santé.

La suite de la section est consacrée à un bref résumé de l'incidence de chacun de ces sept 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, le sexe, l'âge, le revenu et la situation de grossesse. Les Noirs, les Mexicains, les enfants très jeunes, les adolescents, les personnes âgées, les femmes enceintes et les personnes à faible revenu sont suréchantillonnés.

Quand les estimations de totaux d'univers pour l'ensemble de la population sont considérées comme étant 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 la mesure de taille utilisée 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 de tirer 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 pour 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 sur les ratios des coûts qui figure plus bas décrit pourquoi les procédures de suréchantillonnage appliquées dans le cas de la NHANES diffèrent de celles habituellement utilisées pour de nombreuses enquêtes par sondage à base aréolaire.

Charge de travail des centres d'examen mobiles (*CEM*) - Les CEM sont constitués de quatre remorques spécialement conçues et équipées et contiennent tout l'équipement médical. Chaque remorque mesure environ 45 pieds de long et 10 pieds de large. Un camion tracteur détachable 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 des 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 fixe, afin que la planification des opérations sur le terrain soit 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 les visites du CEM à d'autres emplacements dont le calendrier est également préétabli.

Taux de réponse - L'obtention de taux de réponse élevés est une préoccupation dans le cas de presque toutes les enquêtes par sondage. Dans celui de la NHANES, le défi est particulièrement grand, étant donné la portée des interviews et des examens. L'offre d'une rémunération a été utilisée comme moyen d'améliorer les taux de réponse. En outre, la NHANES comprend un programme de relations communautaires important englobant des contacts avec les organismes locaux et les personnes dont il faut obtenir la coopération, ainsi que la couverture dans les médias locaux afin de joindre un aussi grand nombre que possible de PE. Dans le contexte des questions soulevées par le plan d'échantillonnage, une approche qui s'est avérée avoir un effet favorable sur les taux de réponse a été la sélection d'échantillons de plus grande taille dans les ménages échantillonnés. L'un des facteurs soupçonnés d'être à l'origine de l'accroissement des taux de réponse dans les ménages comptant plusieurs PE est que chaque personne reçoit un dédommagement pour son temps et sa participation et qu'il est généralement plus commode pour les membres du ménage de se rendre ensemble au CEM. Le tableau 1 donne le taux de réponse à la composante de l'examen des PE provenant des ménages dans lesquels une seule personne a été sélectionnée comparativement au taux pour les PE provenant des ménages où plusieurs personnes ont été sélectionnées. Comme le montre le tableau, les taux de réponse augmentent d'environ 4 à 7 % selon le type de ménage.

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. Le nombre moyen de PE sélectionnées par ménage (dans les ménages où au moins une PE a été sélectionnée) dans les domaines définis d'échantillonnage varie de 1 à 1,24 dans l'échantillon pour 1999 à 2006. Le regroupement des domaines, afin de réduire leur nombre entre 12 et 15, en fonction de l'âge et/ou de la race/ethnicité donne des nombres moyens variant de 1,01 à 1,37 PE par ménage. Par conséquent, un certain niveau de mise en grappes existe dans la mesure où des domaines groupés sont utilisés pour l'analyse. Il convient de souligner que l'échantillon de PE est utilisé fondamentalement pour l'analyse au niveau des PE (par exemple, statistiques sur la santé et la nutrition). La mise en grappes des PE est évidemment plus importante aux niveaux de la famille et du ménage. Cependant, des variables au niveau du ménage ou de la famille sont utilisées pour ce genre d'analyse (par exemple, niveau de poussière dans le ménage, revenu familial ou assurance familiale). Consulter Curtin et Mohadjer (2008) pour une discussion de l'effet de la mise en grappes et des probabilités inégales de sélection des sous-domaines sur les niveaux de précision des diverses estimations.

Tableau 1
Examen des taux de réponse selon le nombre de PE dans le ménage, par type de ménage, dans l'échantillon de la NHANES pour 1999 à 2006

Type de ménage	Nombre de PE sélectionnées par ménage		Taux de réponse (%)	
	Une PE	Deux PE ou plus	Une PE	Deux PE ou plus
Noir/mexicain	4 892	20 222	76,5	82,3
Autre à faible revenu[1]	1 362	3 349	77,6	84,5
Autre pas à faible revenu	5 597	15 508	68,8	72,6

[1] Le groupe Autre comprend toutes les PE qui ne sont ni noires ni mexicaines. Le seuil de faible revenu est fixé à 130 % du seuil de pauvreté.

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 afin de repérer les répondants admissibles et la réalisation de l'interview pour recueillir les données. Dans le cas de la NHANES, la phase de collecte des données comprend

l'interview auprès du ménage ainsi que 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 mentionné 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'échantillon. 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 à inscrire sur la liste et à soumettre à la présélection, tout en accroissant la variance de la plupart des statistiques. Les économies réalisées grâce à la réduction des coûts d'établissement de la liste et de présélection pourraient être utilisées pour accroître la taille de l'échantillon et, donc, réduire la variance. Cependant, dans la NHANES, le listage et la présélection des ménages représentent une très faible fraction du coût, de sorte que les économies relatives à ces opérations doivent 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.

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 ainsi que l'analyse, si bien que les estimations de la variance par rapport au plan sont relativement imprécises. En outre, les tailles effectives d'échantillon pour la plupart des sous-domaines dans les échantillons annuels sont trop faibles. La plupart des analyses au niveau du sous-domaine devront être faites en cumulant un certain nombre d'échantillons annuels afin d'obtenir une précision et une 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 permet de cumuler efficacement les échantillons annuels au cours des années.

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-domaine 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 certaines 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.

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

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

Les paragraphes qui suivent décrivent brièvement les caractéristiques uniques du plan d'échantillonnage de la NHANES.

Mise en grappes et mesures de taille - Dans la NHANES, la taille d'échantillon doit être suffisamment grande pour produire une charge de travail efficace dans 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. Selon l'expérience acquise lors de réalisations antérieures de la NHANES, un nombre moyen de 340 PE examinées est un nombre approximativement optimal qui produit le nombre maximal d'UPE, tout en maintenant la taille d'échantillon dans chaque région suffisamment grande 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 afin de réduire le temps de déplacement des répondants pour se rendre 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 dans chaque UPE. Les UPE et les segments sont sélectionnés avec une probabilité proportionnelle à une mesure de 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 les personnes résidant dans cette UPE particulière peuvent être sélectionnées. 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 et stratification - Un moyen de réaliser des échantillons annuels représentatifs de la population nationale consiste à sélectionner un échantillon indépendant d'UPE. Étant donné le nombre limité d'UPE de la NHANES 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 au niveau 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 (décrite plus loin), puis à affecter une UPE de chacune des strates principales à chaque année. Cette structure hiérarchique de l'échantillon de six ans évite le chevauchement des UPE non autoreprésentatives durant les six années.

L'échantillon de six ans de la NHANES est sélectionné selon un plan stratifié à deux UPE par strate qui a été élaboré en ayant pour objectif principal l'efficacité de l'échantillon de six ans, ainsi que celle des échantillons pluriannuels. 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.

Le plan d'échantillonnage de la NHANES comprenait (jusqu'à 2006 inclusivement) 18 UPE autoreprésentatives. Ces UPE variaient de celles qui étaient autoreprésentatives pour les échantillons annuels à celles qui l'étaient pour les échantillons de trois ans ou de six ans. Ces UPE ont été affectées de telle façon que le nombre d'UPE autoreprésentatives soit le même chaque année, les UPE autoreprésentatives des échantillons de trois ans étant espacées de trois ans. 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 pour créer des pseudo-strates. Chaque paire a été assignée aléatoirement à 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 a donné un échantillon de 72 UPE non autoreprésentatives qui permettent de produire des estimations annuelles et pluriannuelles efficaces sans compromettre l'efficacité des estimations sur six ans. L'échantillon de six ans est obtenu selon un plan d'échantillonnage d'une UPE par strate secondaire (ou un plan d'échantillonnage de deux UPE par pseudo-strate) et chaque échantillon annuel, selon un plan d'échantillonnage d'une UPE par strate principale. En plus, ce plan offre la souplesse nécessaire pour répondre aux changements d'exigences d'échantillonnage (si un nouvel échantillon doit être sélectionné), puisque pendant les trois premières années, l'échantillon correspond à un plan d'échantillonnage d'une UPE par pseudo-strate.

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 PE sont sélectionnées à des taux établis

afin d'être certain que soient réalisées les tailles d'échantillons cibles par sous-domaine.

L'échantillon de PE est sélectionné de façon à maximiser le nombre moyen de personnes sélectionnées 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 de pseudorandomisation afin de maximiser le nombre de PE par ménage. Consulter Waksberg et Mohadjer (1991) pour une description de l'approche.

Tailles contrôlées d'échantillon 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'échantillon 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 mentionné 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 supérieurs de 50 % à 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'est-à-dire 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 auront 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 au niveau du 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 à cause de

constructions neuves (c'est-à-dire 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, à cause des calendriers établis pour les CEM. Le sous-échantillonnage 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. 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.

Sous l'approche des constructions neuves (Bell, Mohadjer, Montaquila et Rizzo 1999), les unités qui viennent d'être construites sont exclues des segments de région et de nouveaux segments sont créés en se basant sur l'information du U.S. Census Bureau sur les permis émis pour des constructions neuves depuis le recensement décennal le plus récent. Les segments de constructions neuves comprennent des grappes de permis de bâtir émis durant un ou plusieurs mois contigus par un bureau d'octroi de permis de bâtir. Les fichiers de la Building Permits Survey menée par le Census Bureau servent de sources de données sur le nombre de permis de bâtir résidentiels émis par les bureaux d'octroi de permis de bâtir.

L'échantillonnage à deux phases est utilisé dans un certain nombre d'applications statistiques. L'une d'elles 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 (de segments dans le cas de la NHANES) 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 (de segments) est sélectionné à partir de l'échantillon de première phase en utilisant la mesure de taille mise à jour.

À partir 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, Bell, Mohadjer et Rizzo 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 - De nos jours, 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 par couplage de 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 posant un grand risque de permettre d'identifier des personnes qui font partie de l'échantillon. Les éléments de données délicats, à diffusion limitée ou non diffusés, sont consultables dans un centre de données de recherche. 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 Centre de données de recherche.

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 particulière, 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) de nombreuses démarches de relations communautaires sont menées dans chaque UPE en vue d'améliorer les taux de réponse. Le programme de relations communautaires comprend la prise de contact avec divers organismes et

individus à chaque emplacement de collecte pour essayer d'obtenir leur appui et l'utilisation des médias (journaux, télévision et radio) afin de rejoindre un nombre aussi grand que possible de PE. Il est, par conséquent, relativement facile de déterminer les comtés compris dans l'échantillon de la NHANES. 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 formant une grappe dans un comté 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, Dohrmann, Montaquila, Mohadjer et Curtin (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 de manière à 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 particuliers, dans divers sous-groupes socioéconomiques de la population. Donc, le plan d'échantillonnage de la NHANES doit établir un équilibre entre les exigences liées à l'obtention d'échantillons de sous-domaine efficaces, d'une part, et d'une charge de travail efficace pour le personnel d'examen du CEM, d'autre part, 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 afin de répondre aux nouvelles questions 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 annuels et pluriannuels efficaces. 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 présélection 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 sélectionnées par ménage, parce que cela a semblé accroître le taux global de réponse lors des enquêtes précédentes.

Les défis décrits dans le présent article ont trait aux principaux aspects de la NHANES. Cette dernière possède de nombreux autres caractéristiques uniques dont il faut tenir compte lors de l'analyse des données. Par exemple, chaque échantillon annuel ne contient qu'un très petit nombre d'UPE, mais les données recueillies dans ces UPE ne le sont pas aléatoirement d'une saison à l'autre. En particulier, s'il existe une interaction entre la saison et la région géographique pour une variable d'intérêt, le plan d'échantillonnage actuel de la NHANES ne permettra pas de l'estimer. Étant donné le petit nombre d'UPE dans chaque cycle de diffusion des données, tout couplage contextuel des données au niveau géographique doit se faire au centre de données de recherche du NCHS. Comme le nombre de sous-échantillons de la NHANES est élevé, il convient de prendre tout spécialement soin d'utiliser le poids de sous-échantillon approprié; ainsi, les estimations des cas de diabète non diagnostiqués doivent être calculées en utilisant le poids spécial pour le test effectué à jeun.

Afin de faciliter la bonne utilisation des CEM pour la collecte des données, aucun effort n'a été fait en vue de répartir aléatoirement l'échantillon d'UPE entre les périodes dans les échantillons annuels. Cependant, la dimension temporelle joue un rôle important dans certains indicateurs de la santé, telle que la nutrition. De surcroît, l'analyse des données sur la nutrition peut aussi être influencée par la nature complexe du plan d'échantillonnage et de la collecte des données. Des poids de sondage spéciaux construits pour les deux jours de collecte des données du rappel alimentaire de 24 heures tiennent compte de la variation du nombre d'examens selon le jour de la semaine. Un tutoriel sur Internet est développé à l'heure actuelle afin de faciliter l'analyse des données sur la nutrition de la NHANES. Un tutoriel général concernant l'analyse fondée sur le plan de sondage des données de la NHANES peut être obtenu à http://www.cdc.gov/nchs/tutorials/.

Remerciements

Les auteurs remercient le rédacteur associé et les examinateurs de leurs suggestions et commentaires constructifs qui leur ont permis d'améliorer considérablement l'article.

Bibliographie

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.

Curtin, L.R., et Mohadjer, L. (2008). Design trade-offs for the National Health and Nutrition Examination Survey. *Proceedings of the Ninth Conference on Health Surveys Research Methods*, à paraître.

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.

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

Waksberg, J., et Mohadjer, L. (1991). Automation of within-household sampling. *Proceedings of the Survey Research Methods Section*, American Statistical Association.