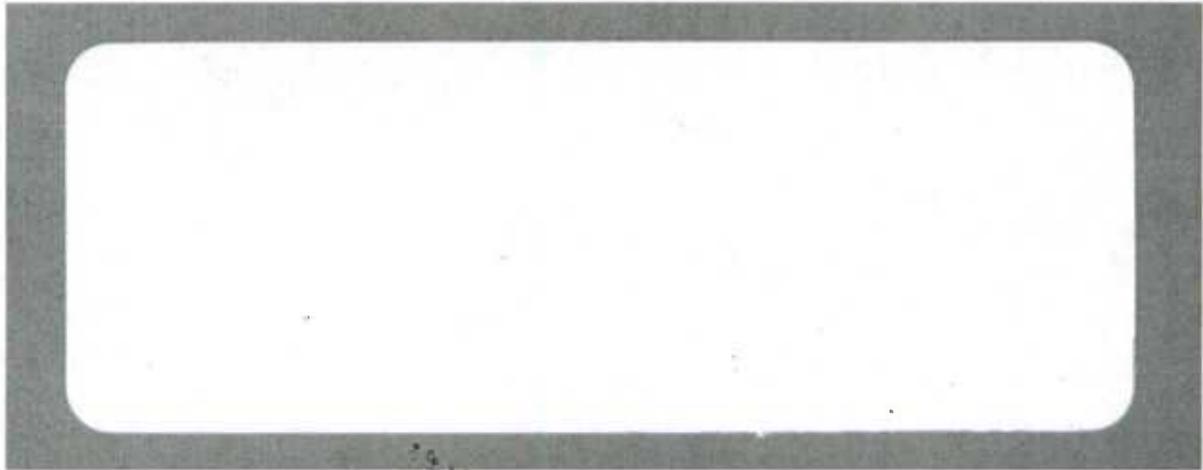


# Methodology Branch

# Direction de la méthodologie



Household Survey  
Methods Division

Division des méthodes  
d'enquêtes auprès des ménages



**TROIS PREMIERS CYCLES DE  
L'ENQUÊTE NATIONALE SUR LA SANTÉ DE LA POPULATION  
VOLET MÉNAGES**

DMEM – 2008-006F

Mylène Lavigne,

STATISTICS STATISTIQUE  
CANADA CANADA

NOV 26 2008

LIBRARY  
BIBLIOTHEQUE

Division des méthodes d'enquêtes auprès des ménages  
Statistique Canada

9 octobre 2008

---

Les travaux présentés sont diffusés sous la responsabilité des auteurs et ne reflètent pas nécessairement les opinions ni les politiques de Statistique Canada

**LES CINQ CYCLES (1994-1995 À 2002-2003) DE  
L'ENQUÊTE NATIONALE SUR LA SANTÉ DE LA POPULATION  
VOLET ÉTABLISSEMENTS DE SOINS DE SANTÉ**

**MYLÈNE LAVIGNE,**

**RÉSUMÉ**

L'Enquête nationale sur la santé de la population (ENSP) est une enquête biennale qui a vu le jour à Statistique Canada en 1994. Cette enquête vise principalement à mesurer l'état de santé des Canadiens et à favoriser une meilleure compréhension des facteurs déterminants pour la santé. D'ailleurs, pour dresser un portrait tout entier de la population canadienne, l'ENSP comporte trois volets : une enquête auprès des ménages, une enquête auprès des résidents à long terme des établissements de soins de santé et une enquête nordique (qui couvre les ménages dans les territoires). Dans ce document, on s'intéresse uniquement aux cinq cycles de l'enquête auprès des résidents à long terme (plus de 6 mois) des établissements de soins de santé des dix provinces canadiennes. Toutes les autres sections portent spécifiquement sur le volet établissements de soins de santé de l'ENSP, en décrivant toutes ses étapes soient la couverture, la base de sondage, le plan d'échantillonnage, le contenu, le questionnaire, la collecte, le traitement, la non-réponse, la pondération, la qualité des données, la confidentialité et la diffusion.

**THE FIVE CYCLES (1994/1995 TO 2002/2003) OF THE  
NATIONAL POPULATION HEALTH SURVEY  
HEALTH INSTITUTIONS COMPONENT**

**Mylène Lavigne**

**ABSTRACT**

The National Population Health survey (NPHS) is a biennial survey that originated at Statistics Canada in 1994. The main goal of this survey is to measure the Canadian health status and to promote a better understanding of determinant health factors. To provide a complete portrait of the Canadian population, the NPHS has three components: a household survey, a survey of long-term residents (more than 6 months) of health care institution, and a northern survey (which covers households in the territories). This paper covers only the five cycles of the Health Institutions survey in the ten Canadian provinces. This paper firstly presents a background of the NPHS in general. All the others sections covers only the Health Institutions Component, by describing all its various aspects, including coverage, sampling frame, sample design, content, questionnaire, nonresponse, weighting, data quality, confidentiality and dissemination.

## TABLE DES MATIÈRES

<b>1. INTRODUCTION</b> .....	<b>7</b>
<b>2. CONTEXTE</b> .....	<b>8</b>
<b>3. COUVERTURE ET BASE DE SONDAGE</b> .....	<b>9</b>
<b>4. PLAN D'ÉCHANTILLONNAGE</b> .....	<b>11</b>
4.1 CYCLE 1 .....	11
4.1.1 Stratification .....	11
4.1.2 Taille d'échantillon .....	12
4.1.3 Répartition de l'échantillon .....	13
4.1.4 Sélection de l'échantillon .....	17
4.2 ÉCHANTILLON LONGITUDINAL .....	17
4.3 ÉCHANTILLON TRANSVERSAL DU CYCLE 2 .....	17
4.3.1 Suivi du panel longitudinal .....	18
4.3.2 L'échantillon supplémentaire .....	18
<b>5. QUESTIONNAIRES ET CONTENU</b> .....	<b>21</b>
<b>6. COLLECTE DES DONNÉES</b> .....	<b>22</b>
6.1 MODE ET PÉRIODE DE COLLECTE .....	22
6.2 PARTICULARITÉS DE LA COLLECTE .....	22
6.3 INTERVIEWS PAR PROCURATION .....	23
<b>7. TRAITEMENT DES DONNÉES</b> .....	<b>25</b>
7.1 COLLECTE DES DONNÉES ET ENVOI POSTAL .....	25
7.2 SAISIE DES DONNÉES .....	26
7.3 VALIDATION .....	26
7.4 REFORMATAGE .....	26
7.5 CODAGE .....	27
7.6 VÉRIFICATION .....	28
7.7 VARIABLES DÉRIVÉES .....	28
7.8 PONDÉRATION .....	28
7.9 CRÉATION DES FICHIERS FINAUX .....	28
<b>8. NON-RÉPONSE</b> .....	<b>30</b>
8.1 NON-RÉPONSE TOTALE .....	30
8.1.1 Échantillons transversaux .....	30
8.1.2 Échantillon longitudinal .....	30
8.1.3 Taux de réponse .....	31
8.2 NON-RÉPONSE D'ITEM .....	32
8.3 IMPUTATION .....	32
<b>9. DIFFÉRENTS PRODUITS</b> .....	<b>34</b>
<b>10. PONDÉRATION ET ESTIMATION</b> .....	<b>36</b>
10.1 PONDÉRATION TRANSVERSALE DU CYCLE 1 ET PONDÉRATION LONGITUDINALE DU SOUS-ENSEMBLE CARRÉ .....	36
10.1.1 Probabilité de sélection des établissements au cycle 1 (1994-1995) .....	37
10.1.2 Poids de base, ajustement pour la non-réponse et poids final de l'établissement .....	37
10.1.4 Poids de base, ajustement pour la non-réponse et poids final du résident .....	38
10.2 PONDÉRATION TRANSVERSALE DU CYCLE 2 (1996-1997) .....	38
10.2.1 Poids final d'un établissement sélectionné au cycle 1 (1994-1995) .....	38
10.2.3 Probabilité de sélectionner un résident dans les établissements sélectionnés au cycle 1 (1994-1995) .....	39

10.2.4	<i>Probabilité de sélectionner un résident dans les établissements sélectionnés au cycle 2 (1996-1997)</i> .....	40
10.2.5	<i>Poids de base du résident au cycle 2 (1996-1997)</i> .....	40
10.2.6	<i>Ajustement pour les décès</i> .....	40
10.2.7	<i>Ajustement pour la non-réponse des résidents</i> .....	41
10.2.8	<i>Poids final du résident associé aux fichiers maître et FMGD transversaux du cycle 2</i> 42	
10.2.9	<i>Poids final du résident associé au fichier partagé transversal du cycle 2</i> .....	42
10.3	PONDÉRATION LONGITUDINALE .....	43
10.3.1	<i>Pondération pour les sous-ensembles complets des cycles 2 à 5</i> .....	43
10.3.2	<i>Pondération pour les sous-ensembles complet partagé des cycles 2 à 5</i> .....	45
<b>11.</b>	<b>QUALITÉ DES DONNÉES</b> .....	<b>47</b>
11.1	ERREUR DUE À L'ÉCHANTILLONNAGE .....	47
11.1.1	<i>Calcul de l'estimation de l'erreur due à l'échantillonnage</i> .....	47
11.2	ERREUR NON DUE À L'ÉCHANTILLONNAGE .....	50
<b>12.</b>	<b>CONFIDENTIALITÉ ET FICHIERS DE MICRODONNÉES À GRANDE DIFFUSION</b> .....	<b>51</b>
<b>13.</b>	<b>DIFFUSION DES DONNÉES</b> .....	<b>53</b>
	<b>REMERCIEMENTS</b> .....	<b>54</b>
	<b>BIBLIOGRAPHIE</b> .....	<b>55</b>

## 1. Introduction

L'Enquête nationale sur la santé de la population (ENSP) est une enquête qui a vu le jour à Statistique Canada en 1994. Cette enquête vise principalement à mesurer l'état de santé des Canadiens et à favoriser une meilleure compréhension des facteurs déterminants pour la santé. D'ailleurs, pour dresser un portrait tout entier de la population canadienne, l'ENSP comporte trois volets : une enquête auprès des ménages (se référer à Brisebois, Dufour, Kelly, Lavigne, Mathieu et Tolusso 2005 et Statistique Canada 2006), une enquête auprès des résidents à long terme des établissements de soins de santé et une enquête Nord (qui couvre les ménages dans les Territoires). Dans ce document, on s'intéresse uniquement aux cinq cycles de l'enquête auprès des résidents à long terme (plus de 6 mois) des établissements de soins de santé des dix provinces canadiennes.

Le premier cycle de collecte de données de l'ENSP débute en 1994. Les deux premiers cycles du volet établissements de soins de santé de l'ENSP permettent de diffuser des estimations non seulement longitudinales mais également transversales. À partir du cycle 3, le volet devient strictement longitudinal. On recueille tout d'abord des renseignements de base sur tous les établissements échantillonnés. Les renseignements touchant la santé sont recueillis lors d'une interview en personne avec les résidents choisis au hasard dans les établissements sélectionnés. L'*échantillon longitudinal de résidents*, choisis au hasard lors du premier cycle, fera l'objet d'un rappel à tous les deux ans pour une période de 8 ans. Après cinq cycles de collecte (1994-1995 à 2002-2003), le volet établissements de soins de santé de l'ENSP se termine à cause du grand nombre de décès dans l'échantillon.

Le présent document comporte treize sections. La section 2 présente tout d'abord une mise en contexte de l'ENSP en général. Toutes les autres sections portent spécifiquement sur le volet établissements de soins de santé de l'ENSP. Plus spécifiquement, la section 3 traite de la population cible et des bases de sondage utilisées. À la section 4, on retrouve une description du plan d'échantillonnage. La section 5 se penche sur les questionnaires et le contenu. La collecte des données est ensuite discutée à la section 6 pour poursuivre avec le traitement des données à la section 7. La section 8 présente les défis imposés par la non-réponse partielle et totale. Suit ensuite, à la section 9, une brève description des différents produits mis à la disposition des utilisateurs. La section 10 donne un aperçu de la méthodologie de pondération et d'estimation. La qualité des données est discutée à la section 11. La section 12 décrit les mesures prises en matière de confidentialité en accord avec les différents produits offerts aux utilisateurs. Finalement, la diffusion des données est ensuite discutée à la section 13.

## 2. Contexte

Avant 1994, les enquêtes sur la santé de la population sont peu fréquentes à Statistique Canada (Catlin et Will, 1992). La dernière enquête d'importance sur la santé de la population, soit l'Enquête santé Canada, est réalisée en 1978. Prévues à l'origine sur une base permanente, cette enquête cesse ses activités en raison de restrictions budgétaires affligées à l'ensemble de la Fonction publique fédérale au deuxième trimestre de 1978. Depuis, diverses enquêtes comblent certaines des lacunes observées en ce qui a trait à la santé grâce à des sujets particuliers ou auprès de populations précises. Par exemple, en 1983, Statistique Canada réalise l'Enquête sur la santé et les invalidités au Canada et, par la suite, l'Enquête sur la santé et les limitations d'activités en 1986 (Dolson, McClean, Morin et Théberge, 1987). Ces deux enquêtes visent principalement à recueillir de l'information particulière sur la nature des problèmes éprouvés par la population canadienne ayant une incapacité. La dernière de ces deux enquêtes comptait un volet établissements.

Par ailleurs, en 1990, Statistique Canada réalise l'Enquête promotion santé Canada (Santé et Bien-être Canada, 1993) et, en 1991, la thématique de l'Enquête sociale générale porte sur la santé. Toutefois, l'étendue de ces deux enquêtes est très limitée (Yeo, 1999). Aucune de ces enquêtes ne permet de dresser un bilan complet de l'état de la santé de la population et de la multitude de facteurs qui ont une incidence sur la santé.

À l'automne 1991, le Conseil national d'information sur la santé recommande la réalisation d'une enquête nationale permanente sur la santé de la population. Cette recommandation découle des pressions économiques et fiscales exercées sur le système de santé ainsi que des besoins en information afin d'améliorer l'état de santé de la population canadienne. En avril 1992, suite à cette recommandation, Statistique Canada reçoit le mandat et le financement pour développer une enquête longitudinale sur la santé qui doit être souple et capable de produire des résultats valables, fiables et actuels. Cette enquête doit également pouvoir s'adapter facilement aux besoins, aux politiques et aux intérêts changeants. En 1994, l'Enquête nationale sur la santé de la population (ENSP) est sur le terrain pour la première fois.

Les objectifs de l'ENSP sont très nombreux. Il s'agit de :

- i) Faciliter l'élaboration de politiques gouvernementales conçues pour améliorer l'état de santé;
- ii) Fournir des données pour réaliser des études analytiques permettant de mieux comprendre les déterminants de la santé;
- iii) Réunir des données sur les facteurs économiques, sociaux, démographiques, professionnels et environnementaux qui sont en corrélation avec l'état de santé;
- iv) Aider à mieux comprendre la relation entre l'état de santé et l'utilisation des services de santé;
- v) Offrir des données sur un échantillon constant qui seront le reflet de la dynamique de la santé et de la maladie, et produire des estimations transversales périodiques;
- vi) Mettre à la disposition des provinces, des territoires et des autres utilisateurs un véhicule de collecte permettant de compléter le contenu ou l'échantillon;
- vii) Permettre le couplage des données avec des sources administratives.

À l'enquête, s'ajoute un volet spécial, qui couvre les résidents des établissements de soins de santé, parce que cette population fait rarement partie du champ d'observation des enquêtes nationales et que les caractéristiques de leur état de santé diffèrent vraisemblablement de celles de la population générale. Le volet établissements de soins de santé de l'ENSP, est la première enquête nationale longitudinale menée auprès des résidents des établissements de soins de santé au Canada.

### 3. Couverture et base de sondage

La population cible du volet établissements de soins de santé de l'ENSP comprend les résidents des établissements de soins de santé de longue durée (plus de 6 mois) de toutes les provinces, sauf ceux des territoires, des réserves indiennes et des bases des Forces canadiennes. Les établissements considérés par l'enquête doivent avoir un minimum de quatre lits et les résidents ne doivent pas être autonomes<sup>1</sup>. Les établissements qui ne font pas partie du système de santé, c'est-à-dire les établissements correctionnels, les prisons, les établissements pour jeunes délinquants, ainsi que les foyers collectifs pour enfants (orphelinats) et les établissements religieux, ne sont pas couverts par l'enquête.

La base de sondage a été formée à partir de deux listes d'établissements de soins de santé soient la liste de l'Enquête sur les établissements de soins pour bénéficiaires internes (EÉSBI) établie par l'Institut canadien d'information sur la santé (ICIS) et la liste de l'Enquête annuelle sur les hôpitaux (EAH) et le Registre des hôpitaux (RH) tenue à jour par la Division de la statistique sur la santé (DSS) de Statistique Canada (voir Mohl 1994 et Parsons 1993). Le tableau 3.1 résume les différents types d'établissements couverts par chacune d'elle. La liste de l'EÉSBI comprend tous les établissements de santé (d'au moins quatre lits et excluant les hôpitaux) qui sont approuvés, financés ou autorisés par les ministères de la santé provinciaux. Cette liste est mise à jour annuellement. Pour sa part, la liste annuelle du EAH/RH donne de l'information sur les hôpitaux tels que leur taille et le type de soins prodigués.

**Tableau 3.1 : Types d'établissement de soins de santé des listes EÉSBI et EAH/RH.**

<i>Enquête sur les établissements de soins pour bénéficiaires internes (EÉSBI)</i>	<i>Enquête annuelle sur les hôpitaux et le registre des hôpitaux (EAH/RH)</i>
Traitement d'alcoolisme et de toxicomanie	Général (pas d'unité de long terme)
Personnes âgées	Général (avec unité(s) de long terme)
Enfants avec problèmes émotifs	Soins prolongés, problèmes chroniques
Personnes avec déficience intellectuelle	Réhabilitation
Personnes handicapées physiquement	Pédiatrique
Personnes avec problèmes psychiatriques	Maternité
Délinquants	Orthopédique
Transition	Hôpital d'avant-poste (outpost hospital)
Autres (ex. : femmes battues et leurs enfants)	Maison de convalescence
	Autre spécialité
	Traitement d'alcoolisme et de toxicomanie
	Personnes avec problèmes psychiatriques (long terme)
	Personnes avec problèmes psychiatriques (court terme)

Afin de respecter la couverture de l'enquête, certains types d'établissements parmi la liste du tableau 3.1 sont exclus de la base sondage. De l'EÉSBI, on a seulement retenu les établissements

<sup>1</sup> Par autonome, on entend les établissements où le niveau de soins est inférieur au type 1. Les soins de type 1 sont les soins qu'exige une personne ambulatoire et (ou) capable de se déplacer seule, dont les facultés physiques et (ou) mentales sont affaiblies et qui exige une supervision et (ou) de l'aide principalement pour vaquer aux activités quotidiennes. Il faut noter qu'au Québec, la variable *niveau de soins* n'est pas disponible. Cependant, la liste des établissements pour bénéficiaires internes ne comprend généralement pas d'établissements dont les résidents sont autonomes.

de soins de longue durée pour les personnes âgées, les établissements de soins de longue durée pour les enfants présentant des troubles affectifs et les établissements de soins de longue durée pour les personnes présentant des troubles affectifs, des troubles du développement, un handicap physique ou des troubles mentaux. De la liste des hôpitaux, on a retenu uniquement les hôpitaux généraux comprenant des unités de soins de longue durée, les établissements de soins de santé prolongés, les centres de réadaptation et les hôpitaux spécialisés comprenant des soins de longue durée (ex. : hôpitaux pédiatriques ou psychiatriques).

## 4. Plan d'échantillonnage

Le volet établissements de soins de santé de l'ENSP avait pour objectif de produire uniquement des estimations nationales. Des estimations transversales, en plus des estimations longitudinales, ont été produites pour les deux premiers cycles de l'enquête. À partir du cycle 3 (1998-1999), le volet établissements de soins de santé de l'ENSP est devenue uniquement une enquête longitudinale.

### 4.1 Cycle 1

Le plan d'échantillonnage utilisé pour le volet établissements de soins de santé de l'ENSP est un plan stratifié à deux degrés. L'unité primaire d'échantillonnage (UPE) est formée d'établissements et l'unité secondaire d'échantillonnage (USE) est le résident.

#### 4.1.1 Stratification

Afin de s'assurer que tous les types d'établissements sont bien représentés dans l'échantillon, il était nécessaire de regrouper les établissements qui avaient des caractéristiques similaires et de sélectionner l'échantillon dans chacun de ces groupes appelés *strates*. Au départ, la liste d'établissements de soins de santé (la base de sondage) a été stratifiée selon la région géographique (strates géographiques), puis selon la catégorie d'établissements (strates de catégorie) et finalement selon le nombre de lits (strates de taille).

Les strates géographiques correspondent à cinq régions : les provinces de l'Atlantique, le Québec, l'Ontario, les provinces des Prairies et la Colombie-Britannique. Étant donné que l'objectif était de produire uniquement des estimations nationales, il n'était pas nécessaire de raffiner davantage les strates géographiques. Cette stratification permettait de s'assurer que toutes les régions du pays étaient représentées.

Ensuite, dans chaque strate géographique, il fallait trouver une façon logique de regrouper les différentes catégories d'établissements qui se retrouvaient sur la base de sondage, car certaines catégories contenaient très peu de personnes et/ou pouvaient être différemment identifiées selon la province. Par exemple, au Québec, il n'y avait pas de distinction entre les établissements pour personnes avec un problème psychiatrique et les établissements pour personnes avec une déficience intellectuelle. Pour cette raison, ces deux catégories ont été groupées et trois catégories d'établissements ont été définies :

*Établissements pour personnes âgées* : les établissements de soins pour personnes âgées et les hôpitaux de soins prolongés (environ 50 % des établissements et 74 % des lits de la base de sondage);

*Établissements de soins cognitifs* : les établissements de soins pour enfants présentant des troubles affectifs et des personnes présentant des troubles mentaux ou des troubles de retard du développement et les hôpitaux psychiatriques (environ 38 % des établissements et 16 % des lits de la base de sondage);

*Autres établissements de réadaptation* : les centres de réadaptation, les hôpitaux pédiatriques et d'autres hôpitaux spécialisés, les hôpitaux généraux comprenant des unités de soins de longue durée, ainsi que les établissements de soins accueillant des personnes physiquement handicapées (environ 11 %

des établissements et 10 % des lits de la base de sondage).

La dernière étape de la stratification était de créer des strates de taille en regroupant les établissements comptant un nombre comparable de lit. Plus un établissement contenait de lits, plus il était possible de sélectionner des résidents de cet établissement. De cette façon, le nombre d'établissements que l'on devait visiter était réduit et conséquemment les coûts.

Pour créer ces strates de taille, on devait premièrement déterminer le nombre de strates de taille dans chaque strate, géographique-catégorie et ensuite déterminer les bornes pour délimiter les strates de taille. Le nombre de strates de taille était basé sur la taille d'échantillon dans chacune des strates, géographique-catégorie (se référer à la section 4.1.3). À l'aide de ces tailles d'échantillon, le nombre minimum de strates de taille a été fixé à 2 et le nombre maximum à 5. Seulement quatre résidents pouvaient être choisis dans la strate de taille des plus petits établissements parce que le plus petit établissement de cette strate pouvait contenir uniquement quatre lits. Dans la deuxième strate de taille, environ dix résidents étaient sélectionnés. Si plus de deux strates de taille étaient nécessaires alors l'échantillon était d'environ 20, 25 et 30 résidents pour les troisième, quatrième et cinquième strates de taille. Une fois le nombre de strates de taille déterminé, les bornes étaient définies en utilisant la méthode  $\sqrt{f(y)}$  décrite dans Cochran (1977), où  $f(y)$  représente le nombre de lits. Une seule restriction, la première borne devait être d'au moins 15. Le tableau 4.1 donne les strates de taille (nombre de lits) pour chaque strate géographique-catégorie.

**Tableau 4.1 : Strates de taille pour chaque strate géographique-catégorie au cycle 1**

Strates géographiques	Strates de catégorie		
	Établissements pour personnes âgées	Établissements de soins cognitifs	Autres établissements de réadaptation
Atlantique	4 à 20, 21 à 50, 51 à 105, 106 et plus	4 à 15, 16 et plus	4 à 25, 26 et plus
Québec	4 à 45, 46 à 75, 76 à 120, 121 à 220, 221 et plus	4 à 25, 26 à 85, 86 et plus	4 à 50, 51 et plus
Ontario	4 à 55, 56 à 95, 96 à 155, 156 à 250, 251 et plus	4 à 15, 16 à 60, 61 et plus	4 à 45, 46 et plus
Prairies	4 à 35, 36 à 65, 66 à 110, 111 à 180, 181 et plus	4 à 20, 21 et plus	4 à 25, 26 et plus
Colombie-Britannique	4 à 40, 41 à 80, 81 à 140, 141 et plus	4 à 15, 16 et plus	4 à 60, 61 et plus

#### 4.1.2 Taille d'échantillon

La taille d'échantillon est liée à la précision voulue des estimations. La taille d'échantillon du volet établissements de soins de santé de l'ENSP devait être assez grande pour produire des estimations à l'échelle nationale dont le coefficient de variation (CV) serait au plus 16,5 % pour les caractéristiques présentées par au moins 10 % de la population. Pour calculer cette taille d'échantillon, un taux de réponse hypothétique de 85 % et un effet de plan de 2,5 ont été fixés. En se basant sur les résultats de l'Enquête sur la santé et les limitations d'activité de 1991 (ESLA), les estimations du taux de réponse et de l'effet de plan étaient conservateurs. En effet, l'ESLA 1991 a obtenu un taux de réponse pour les établissements de 95 % et un taux de réponse individuel de 90 %. Les effets de plan obtenus par l'ESLA 1991 variaient entre 2,0 et 2,5 à l'échelle provinciale.

De plus, la taille d'échantillon devait être assez grande afin d'obtenir des résultats (à l'interne) à l'échelle des établissements pour personnes âgées. Ainsi, la taille d'échantillon  $n$  requise pour estimer une proportion  $p$ , étant donné une variable avec deux catégories (cas de réponse binomiale), sous un plan d'échantillonnage aléatoire simple est calculée comme suit :

$$n = \frac{p \times (1 - p)}{[cv \times p]^2}$$

De plus, il a été démontré par Tortora (1978) que la taille d'échantillon nécessaire pour une distribution multinomiale (variable avec plus d'une catégorie) est entre 1,5 et 2 fois plus grande que la taille d'échantillon requise pour une distribution binomiale. Le tableau 4.2 présente quelques tailles d'échantillon calculées à l'aide de cette formule et selon différents paramètres. On constate qu'une taille de 2 250 est suffisante pour donner des résultats selon une binomiale avec un CV de 10 % et selon une multinomiale avec un CV inférieur à 16,5 % (2 250 est plus grand que 1 650).

**Tableau 4.2 : Tailles d'échantillon préliminaires selon différents paramètres**

CV	Tailles d'échantillon requises pour une proportion minimale de 10 %		
	Selon une binomiale		Selon une multinomiale
	plan aléatoire simple	effet de plan de 2,5	effet de plan de 2,5
10,0 %	900	<b>2 250</b> (900 x 2,5)	4 500 (2 250 x 2)
16,5 %	330	825 (330 x 2,5)	<b>1 650</b> ( 825 x 2)

Selon la base de sondage, environ 74 % des résidents des établissements sont dans la strate des établissements pour personnes âgées. On pouvait donc supposer qu'environ 1 665 résidents (74 % x 2 250) seraient sélectionnés dans cette strate. Pour cette strate, les estimations pour les distributions binomiale et multinomiale auraient un CV dont la limite serait de 16,5 %.

Finalement, en corrigeant le nombre voulu de résidents suivant le taux de réponse prévu (85 %), la taille de l'échantillon est de 2 600 résidents.

#### 4.1.3 Répartition de l'échantillon

Pour maximiser l'efficacité de l'échantillonnage stratifié, il était important de considérer comment la taille totale de l'échantillon  $n$ , serait répartie dans chaque strate. Au cycle 1, deux méthodes ont été examinées, la répartition proportionnelle à  $N$  et la répartition proportionnelle à  $\sqrt{N}$  où  $N$  est le nombre de lits dans l'établissement. La répartition proportionnelle à  $N$  a été utilisée parce qu'elle favorise l'autopondération du plan d'échantillonnage, ce qui facilite l'analyse. L'avantage de la répartition proportionnelle à  $\sqrt{N}$  est que les tailles d'échantillon des strates sont distribuées plus uniformément. C'est avantageux lorsqu'on veut estimer des petits domaines, mais pour ce volet, uniquement des résultats à l'échelle nationale étaient désirés.

##### 4.1.3.1 Répartition de l'échantillon dans les strates géographique-catégorie

Dans la répartition proportionnelle à  $N$  (nombre de lits total), la taille de l'échantillon  $n_g$  (nombre de résidents) de chaque strate géographique-catégorie ( $g$ ) est proportionnelle à la taille de la population  $N_g$  (nombre de lits) de la strate géographique-catégorie. Une part plus importante de l'échantillon est donc attribuée à une strate plus grande qu'à une strate plus petite. La répartition de l'échantillon est faite à l'aide de l'équation suivante :

$$n_g = \frac{N_g}{N} n.$$

Le tableau 4.3 donne les résultats de la répartition de l'échantillon dans chaque strate géographique-catégorie.

**Tableau 4.3 : Répartition de l'échantillon dans chaque strate géographique-catégorie selon la répartition proportionnelle à N.**

Strate géographique	Population et échantillon	Strate de catégorie			Total
		Établissements pour personnes âgées	Établissements de soins cognitifs	Autres établissements de réadaptation	
Atlantique	Population - Établissements	484	249	50	783
	Population - Nombre de lits	18 031	3 570	1 643	23 244
	Échantillon - Résidents	187	37	30	254
Québec	Population - Établissements	495	246	124	865
	Population - Nombre de lits	42 590	11 281	6 076	59 950
	Échantillon - Résidents	440	116	63	619
Ontario	Population - Établissements	646	566	193	1 405
	Population - Nombre de lits	71 877	14 064	8 469	94 410
	Échantillon - Résidents	745	146	87	978
Prairies	Population - Établissements	432	326	64	822
	Population - Nombre de lits	32 818	5 933	1 886	40 637
	Échantillon - Résidents	339	61	30	430
Colombie-Britannique	Population - Établissements	284	389	91	764
	Population - Nombre de lits	21 199	4 937	6 941	33 077
	Échantillon - Résidents	220	51	72	343
<b>Total</b>	Population - Établissements	2 341	1 776	522	<b>4 639</b>
	Population - Nombre de lits	186 515	39 785	25 018	<b>251 318</b>
	Échantillon - Résidents	1 931	411	282	<b>2 624</b>

En utilisant la répartition proportionnelle à N, des tailles d'échantillon inférieures à 30 ont été obtenues dans deux cas (autres établissements de réadaptation, Atlantique et Prairies). Pour ces deux cas, la taille d'échantillon a été augmentée à 30 afin de créer deux strates de taille.

#### 4.1.3.2 Répartition de l'échantillon dans les strates de taille

Pour répartir l'échantillon dans les strates de taille à l'intérieur de chaque strate géographique-catégorie, on a une autre fois utilisée la répartition proportionnelle à N. Dans la répartition proportionnelle à  $N_g$  (nombre de lits total de la strate géographique-catégorie g), la taille de l'échantillon  $n_h$  (nombre de résidents de chaque strate géographique-catégorie-taille h) est proportionnelle à la taille de la population  $N_h$  (nombre de lits de la strate géographique-catégorie-taille h.). Une part plus importante de l'échantillon est donc attribuée à une strate d'établissements de plus grande taille qu'à une strate d'établissements de plus petite taille. La répartition de l'échantillon est faite à l'aide de l'équation suivante :

$$n_h = \frac{N_h}{N_g} n.$$

Ensuite, la taille d'échantillon de résidents a été ajustée afin de déterminer la taille d'échantillon d'établissements. Le nombre de résidents choisi dans chaque établissement dépend de la strate de taille où se trouve l'établissement. Dans la 1<sup>re</sup> strate de taille (strate d'établissements avec les plus petites tailles), on sélectionne 4 résidents, ensuite on sélectionne environ 10, 20, 25 et 30 résidents dans les 2<sup>e</sup>, 3<sup>e</sup>, 4<sup>e</sup> et 5<sup>e</sup> strates de taille respectivement. La répartition finale est présentée dans le tableau 4.4. Au total, 232 établissements et 2 623 résidents devaient être sélectionnés.

**Tableau 4.4 : Tableau de la répartition finale**

Strates géographiques	Strates de catégorie										Total
	Personnes âgées					Soins cognitifs			Réadaptation		
Atlantique	strate 1 4-20	strate 2 21-50	strate 3 51-105	Strate 4 106 +		strate 1 4-15	strate 2 16 +		strate 1 4-25	strate 2 26 +	
Population – Établissements	258	123	62	41		212	37		38	12	783
Population – Nombre de lits	2 533	4 071	4 608	6 819		1 394	2 176		456	1 187	23 244
Échantillon – Résidents	26	42	48	71		14	23		8	22	254
Échantillon ajusté - Résidents	28	40	48	72		12	24		8	22	254
Échantillon - Établissements x Résidents	7 x 4	4 x 10	3 x 16	3 x 24		3 x 4	3 x 8		2 x 4	2 x 11	27 établi.
Québec	strate 1 4-45	strate 2 46-75	strate 3 76-120	Strate 4 121-220	strate 5 221 +	strate 1 4-25	strate 2 26-85	strate 3 86 +	strate 1 4-25	strate 2 26 +	Total
Population – Établissements	168	144	87	60	36	170	53	23	82	42	865
Population – Nombre de lits	5 261	8 401	8 056	9 568	11 304	1 474	3 019	6 788	1 904	4 175	59 950
Échantillon – Résidents	54	87	83	99	117	15	31	70	20	43	619
Échantillon ajusté - Résidents	52	90	84	100	116	16	30	69	20	44	621
Échantillon - Établissements x Résidents	13 x 4	9 x 10	4 x 21	4 x 25	4 x 29	4 x 4	3 x 10	3 x 23	5 x 4	4 x 11	53 établi.
Ontario	strate 1 4-55	strate 2 56-95	strate 3 96-155	Strate 4 156-250	strate 5 251 +	strate 1 4-15	strate 2 16-60	strate 3 61 +	strate 1 4-25	strate 2 26 +	Total
Population – Établissements	160	212	133	91	50	456	77	33	138	55	1 405
Population – Nombre de lits	5 560	15 244	16 109	17 856	17 108	2 952	2 144	8 968	8 968	5 631	94 410
Échantillon – Résidents	58	158	167	185	177	31	22	93	29	58	978
Échantillon ajusté - Résidents	56	160	168	182	180	27	24	90	28	60	975
Échantillon - Établissements x Résidents	14 x 4	16 x 10	8 x 21	7 x 26	6 x 30	7 x 3	3 x 8	5 x 18	7 x 4	6 x 10	79 établi.
Prairies	strate 1 4-35	strate 2 36-65	strate 3 66-110	Strate 4 111-180	strate 5 181 +	strate 1 4-20	strate 2 21 +		strate 1 4-25	strate 2 26 +	Total
Population – Établissements	143	116	84	55	34	295	31		49	15	822
Population – Nombre de lits	3 277	5 649	7 334	7 751	8 807	1 963	3 970		657	1 229	40 637
Échantillon – Résidents	34	58	76	80	91	20	41		10	20	430
Échantillon ajusté - Résidents	32	60	76	81	90	20	40		12	18	429
Échantillon - Établissements x Résidents	8 x 4	6 x 10	4 x 19	3 x 27	3 x 30	5 x 4	4 x 10		3 x 4	2 x 9	38 établi.
Colombie-Britannique	strate 1 4-40	strate 2 41-80	strate 3 81-140	Strate 4 141 +		strate 1 4-15	strate 2 16 +		strate 1 4-60	strate 2 60 +	Total
Population – Établissements	92	100	56	36		331	58		62	29	764
Population – Nombre de lits	2 179	6 247	5 747	7 026		2 021	2 916		1 147	5 794	33 077
Échantillon – Résidents	23	65	59	73		21	30		12	60	343
Échantillon ajusté - Résidents	24	66	60	72		20	30		12	60	344
Échantillon - Établissements x Résidents	6 x 4	6 x 11	3 x 20	3 x 24		5 x 4	3 x 10		3 x 4	6 x 10	35 établi.

#### 4.1.4 Sélection de l'échantillon

Dans chacune des strates de taille, à l'intérieur de chaque strate géographique-catégorie (géographique-catégorie-taille), la sélection de l'échantillon a été effectuée en deux étapes. La première étape permet de sélectionner un échantillon d'établissements à partir de la base de sondage. La deuxième étape requiert une liste de résidents pour chaque établissement choisi et à l'aide de cette liste, on sélectionne des résidents.

##### 4.1.4.1 Sélection de l'échantillon des établissements

Dans chaque strate géographique-catégorie-taille, les établissements ont été triés aléatoirement. Ensuite, le nombre requis d'établissements (voir tableau 4.4) a été sélectionné dans chaque strate géographique-catégorie-taille, en utilisant la méthode d'échantillonnage systématique avec probabilité proportionnelle à la taille (PPT). Pour cette sélection, la caractéristique de taille est le nombre de lits.

Un siège social regroupant plusieurs petits établissements pouvait se trouver sur la base de sondage. Le cas échéant, on a obtenu la liste de tous les établissements relevant de ce siège social et on en a sélectionné deux, à savoir le plus grand (d'après le nombre de lits) et un second sélectionné au hasard selon la méthode d'échantillonnage PPT. Dans presque tous les cas, on obtient un nombre suffisant de lits en sélectionnant deux établissements. Toutefois, dans de rares occasions, si le nombre de lits était insuffisant, on sélectionnait un troisième établissement.

##### 4.1.4.2 Sélection de l'échantillon des résidents

Lors du premier contact, l'intervieweur demande que l'établissement lui fournisse une liste des résidents de longue durée lors de sa visite. Cette liste est alors utilisée pour sélectionner l'échantillon des résidents soit 2 623 résidents à l'aide de la méthode d'échantillonnage systématique circulaire (se référer à Särndal, Sewensson et Wretman 1991). À la suite de la collecte du cycle 1, l'échantillon longitudinal et transversal du cycle 1 comprenait 2 287 répondants qui résidaient dans un établissement de soins de santé en 1994-1995.

#### 4.2 Échantillon longitudinal

L'échantillon longitudinal, aussi appelé *panel longitudinal* ou simplement *panel*, est composé des 2 287 personnes sélectionnées au cycle 1 qui ont répondu complètement ou partiellement au questionnaire pour le répondant (QR) du cycle 1. C'est cet échantillon longitudinal qui a été suivi aux cycles 2, 3, 4 et 5.

L'échantillon longitudinal n'a pas été renouvelé au fil du temps. La taille de l'échantillon longitudinal est donc demeurée égale à 2 287 pour tous les cycles. Le panel de 2 287 personnes défini au cycle 1 a été tout simplement contacté à nouveau à tous les cycles afin de recueillir les données longitudinales. Il faut se rappeler qu'à compter du cycle 3, le volet établissements de soins de santé de l'ENSP est devenue uniquement une enquête longitudinale.

#### 4.3 Échantillon transversal du cycle 2

Puisque l'ENSP est établie avant tout comme étant une enquête longitudinale, le cycle 2 consistait principalement à recueillir les données auprès du panel de 2 287 personnes, tel que défini à la section 4.2. Pour les fins transversales, un échantillon de remise à niveau est ajouté afin de pouvoir représenter les sous-populations non couvertes par le panel. La population couverte au

cycle 2 par le panel n'inclut pas les personnes nouvellement admises dans les établissements qui ont participé ou non au premier cycle de l'enquête. De plus, l'érosion de l'échantillon de résidents des établissements de soins de santé est due principalement au décès d'une partie des membres du panel longitudinal. Comme la réduction de l'effectif de l'échantillon est considérable, la variance des estimations pourrait augmenter fortement si on ne sélectionnait aucun échantillon supplémentaire. Au cycle 2, pour la partie transversale de l'enquête, on a donc ajouté des établissements et des résidents à l'échantillon longitudinal afin d'ajuster l'échantillon transversal. L'échantillon transversal du cycle 2 comprend les membres de l'échantillon longitudinal suivis au deuxième cycle, ainsi que la sélection d'un échantillon transversal supplémentaire.

#### **4.3.1 Suivi du panel longitudinal**

Au deuxième cycle, les données ont été recueillies en deux phases. Lors de la première phase, les intervieweurs se sont de nouveau rendus dans les établissements de santé sélectionnés au premier cycle afin de déterminer si les personnes choisies pour faire partie du panel longitudinal vivaient encore dans l'établissement ou si elles avaient déménagé ou étaient décédées. Les administrateurs ou les personnes-ressources de tous les établissements de santé sélectionnés dans l'échantillon du premier cycle ont été contactés en octobre et en novembre 1996. L'intervieweur leur a demandé de remplir un questionnaire papier sur la situation des membres du panel longitudinal (pour obtenir la date de décès des personnes décédées ou des renseignements permettant de dépister celles qui avaient déménagé) et de dresser la liste de tous les résidents de l'établissement bénéficiant de soins de longue durée en indiquant la date de leur admission. Puis, l'intervieweur a rempli un formulaire de renseignements de sélection donnant le nombre de résidents recevant des soins de longue durée admis avant le 1<sup>er</sup> avril 1995, le nombre admis le 1<sup>er</sup> avril 1995 ou après cette date, le nombre de membres du panel longitudinal qui vivait encore dans l'établissement et le nombre qui avait quitté l'établissement. Cette date limite a été choisie parce que le 31 mars 1995 était le dernier jour de la collecte des données du premier cycle. Ces renseignements ont permis d'identifier les résidents nouvellement admis. Les résultats de la première phase ont montré que plus du tiers des personnes sélectionnées au départ pour faire partie du panel longitudinal ne vivaient plus dans l'établissement où elles résidaient au premier cycle. De celles-ci, presque les trois quarts étaient décédés et le reste avait déménagé (en fait, on s'est aperçu plus tard, durant le dépistage, que certaines personnes qui avaient déménagé étaient aussi des membres du panel qui étaient décédés). Tous les renseignements ont été envoyés au bureau central pour permettre la sélection de l'échantillon supplémentaire pour le fichier transversal de 1996-1997.

Il est à noter que tous les membres du panel longitudinal qui vivaient encore dans l'établissement où ils résidaient au premier cycle ont été interviewés et sont inclus dans le fichier transversal. Tous les autres membres du panel étaient exclus de l'échantillon transversal.

#### **4.3.2 L'échantillon supplémentaire**

La création de la base de sondage de 1996 se fonde sur les mêmes critères que ceux appliqués pour la base de sondage du cycle 1 (voir section 3). L'EÉSBI établie par ICIS et l'EAH/RH tenu à jour par la DSS de Statistique Canada ont de nouveau été utilisées et les ministères provinciaux de la santé ont vérifié les listes et les ont mises à jour au besoin afin de s'assurer de leur exactitude.

Les bases de sondage de 1994 et de 1996 d'établissements provinciaux de soins de longue durée ont été appariées afin d'identifier les nouveaux établissements. Il convient de souligner que l'expression « nouvel établissement » ne désigne pas nécessairement un établissement nouvellement établi depuis le cycle 1, mais plutôt tous les établissements inclus dans la base de sondage de 1996 qui ne figurent pas dans celle de 1994. Certains établissements sont aujourd'hui

dans le champ de l'enquête parce qu'ils offrent un niveau plus élevé de soins en 1996 (type 1 en 1996 et non en 1994).

La répartition de l'échantillon adoptée pour le deuxième cycle dépend en grande partie de celle adoptée pour le premier cycle. La section 4.1.3 décrit la stratification et la répartition de l'échantillon au premier cycle. Le plan de stratification appliqué au premier cycle (région géographique, catégorie d'établissements et nombre de lits) a été utilisé pour classer les nouveaux établissements de santé. D'après les renseignements recueillis à l'étape du suivi des membres du panel longitudinal, on a décidé du nombre supplémentaire de résidents qu'il fallait interviewer dans chaque établissement du premier cycle et on a sélectionné au hasard un échantillon supplémentaire de résidents. La taille d'échantillon supplémentaire de chaque établissement ayant participé au premier cycle dépend de l'érosion de l'échantillon en question, du nombre de résidents recevant des soins de longue durée en 1996 et de la fraction de sondage correspondante en 1994.

Un échantillon de nouveaux établissements (8 établissements) a aussi été choisi. On a d'abord fixé la taille de l'échantillon comme étant égale au nombre de résidents sélectionnés par établissement en 1994 pour la strate observée puis, on a rajusté la taille de l'échantillon de façon à tenir compte de la non-réponse personne au premier cycle, c'est-à-dire 93,6 %. Le tableau 4.5 montre la répartition finale de l'échantillon de 1996-1997 selon la catégorie d'établissement.

**Tableau 4.5 : Taille d'échantillon pour l'enquête transversale de 1996-1997**

<b>Taille de l'échantillon selon la catégorie d'établissement</b>			
<i>Personnes âgées</i>	<i>Soins cognitifs</i>	<i>Autres établissements de réadaptation</i>	<i>Total</i>
1 816	339	238	2 393

#### **4.3.2.1 Sélection de l'échantillon supplémentaire à partir des établissements du cycle 1**

D'après la liste des résidents bénéficiant de soins de longue durée du deuxième cycle, on a produit pour chaque établissement, deux sous listes de résidents en fonction de la date d'admission. La première comprend tous les résidents qui vivaient dans l'établissement au moment du premier cycle de l'enquête (c.-à-d. admis avant le 1<sup>er</sup> avril 1995). Les résidents longitudinaux qui vivaient encore dans l'établissement au moment du deuxième cycle ont été supprimés de la liste avant cette étape. La deuxième liste comprend les noms de toutes les personnes nouvellement admises (c.-à-d. admises le 1<sup>er</sup> avril 1995 ou après cette date). Un échantillon systématique de résidents a été sélectionné à partir de la liste des personnes nouvellement admises. Pour certains établissements, on a sélectionné un autre échantillon systématique de résidents d'après la liste des personnes qui vivaient dans l'établissement au moment du premier cycle de l'enquête. Puis, l'échantillon supplémentaire sélectionné par le bureau central a été transmis aux bureaux régionaux afin qu'ils procèdent aux interviews.

#### **4.3.2.2 Sélection de l'échantillon supplémentaire à partir des nouveaux établissements du cycle 2**

Comme on l'a mentionné à la section 4.1.3.1, au premier cycle, on a tiré un échantillon systématique d'établissements dans chaque strate géographique-catégorie-taille par échantillonnage avec probabilité proportionnelle au nombre de lits. Étant donné que les établissements ont été sélectionnés systématiquement, il était possible de sélectionner un

échantillon de nouveaux établissements en ne modifiant ni le plan de sondage original ni la méthode d'estimation de la variance. Pour cela, on a trié les nouveaux établissements de façon aléatoire et on les a ajoutés au bas de la liste de la strate géographique-catégorie-taille appropriée de la base de sondage de 1994 comme s'ils faisaient partie de la liste originale. Puis, de l'origine choisie au hasard et du pas de sondage de chaque strate en 1994, on a pu échantillonner les nouveaux établissements comme s'ils avaient été sélectionnés dans l'échantillon original. Cependant, par cette méthode, on sélectionne un moins grand nombre d'établissements que si les nouveaux établissements étaient stratifiés différemment. On a jugé que les avantages d'une modification de la stratification, donc de la sélection d'un plus grand nombre d'établissements, étaient faibles et que le changement de méthode aurait nécessité des modifications importantes au calcul de la variance.

Pour les nouveaux établissements sélectionnés lors du deuxième cycle, on a tiré un échantillon systématique de résidents d'après la liste de tous les résidents recevant des soins de longue durée. Suite à la collecte, le fichier transversal du cycle 2 comptait 2 118 résidents répondants et 213 établissements.

## 5. Questionnaires et contenu

Deux questionnaires papiers sont utilisés lors de la collecte du volet établissements de soins de santé de l'ENSP. Le premier est appelé, *formulaire de contrôle pour l'établissement (FCE)* et le deuxième, *questionnaire pour le répondant (QR)*.

Le FCE permet de recueillir de l'information sur les politiques de l'établissement concernant par exemple l'usage du tabac, la consommation d'alcool et les activités organisées. De plus, ce questionnaire permet de recueillir l'information nécessaire pour effectuer l'entrevue avec le répondant c'est-à-dire l'information sur le répondant (numéro, nom et interview par personne interposée requise ou non), sur le proche parent (nom, numéro de téléphone et s'il demande une personne déléguée par l'établissement ou non), sur la personne interposée déléguée par l'établissement (consentement donné et nom), sur la prise d'un rendez-vous (date, heure, lieu et langue) et l'état du QR. Un seul FCE est rempli par établissement.

Chaque répondant (ou personne désignée) remplit un QR. Le QR est divisé en trois parties. La première partie permet de vérifier si le répondant longitudinal réside toujours au même endroit (pour les cycles 2 à 5). Si le répondant a déménagé, les nouvelles informations pour localiser le répondant y sont recueillies lorsque disponibles. Si le résident est décédé, la date du décès, la ville et la province où résidait la personne lors du décès sont recueillies. La deuxième partie recueille l'information sociodémographique et sur la santé du résident uniquement pour les répondants vivant dans un établissement. Finalement, la troisième partie recueille aussi de l'information sociodémographique et sur la santé, mais cette fois uniquement pour les répondants déménagés dans un ménage. Le contenu de ces deux dernières parties est très similaire, il y a toutefois moins de questions pour les répondants vivant dans un ménage. Les différentes sections de ces deux parties sont : consentement du plus proche parent (répondants en établissement seulement), information sur le répondant (état matrimonial, interview par procuration ou non, etc.), l'état de santé général (autoévaluation), l'état de santé (vision, ouïe, élocution, mobilité, agilité, main et doigts, sentiments, mémoire, pensée, douleur et malaises), les problèmes de santé chroniques, les limitations des activités, l'équilibre, l'usage du tabac, la consommation d'alcool, le soutien social (répondants en établissement seulement), quelques renseignements généraux (sur le revenu), les personnes à contacter (pour le dépistage), les consentements (jumelage et partage des données), la consommation de médicaments et finalement l'utilisation des soins de santé. Pour plus de détails concernant les questionnaires, se référer à Statistique Canada (2002).

Puisque l'ENSP partage les données recueillies avec ses partenaires, il est impératif de demander aux répondants s'ils acceptent de partager leur information. Pour ce faire, une *question de partage* est utilisée et se lit comme suit :

### Pour les cycles 1, 2 et 3

« Afin d'éviter les chevauchements, Statistique Canada a l'intention de partager les données recueillies grâce à ce sondage avec les ministères provinciaux de la santé, Santé Canada et Ressources humaines et Développement social Canada (cycle 1 et 3 seulement). Ces organismes ont accepté de garder les données confidentielles et de les utiliser uniquement à des fins statistiques.

Acceptez-vous que les données que vous avez fournies soient partagées? »

### Pour les cycles 4 et 5

« Statistique Canada aimerait que vous nous donniez la permission de partager l'information recueillie lors de toutes les interviews effectuées dans le cadre de cette enquête avec les ministères provinciaux de la santé et Santé Canada. Tous les renseignements demeureront strictement confidentiels et ne seront utilisés qu'à des fins statistiques.

Acceptez-vous que les données que vous avez fournies soient partagées? »

## **6. Collecte des données**

La particularité du volet établissements de soins de santé de l'ENSP implique des procédés de collecte bien établis et fixes dans le temps. La période de collecte d'un cycle à l'autre doit être très similaire, le mode de collecte doit demeurer sensiblement le même, les questionnaires doivent être modifiés le moins possible, etc. De plus, pour assurer une collecte de qualité, les intervieweurs doivent être formés adéquatement et doivent suivre les instructions dictées par l'enquête.

### **6.1 Mode et période de collecte**

Les questions du volet établissements de soins de santé de l'ENSP sont conçues pour que l'interview soit complétée en personne par la méthode papier et crayon. L'interview téléphonique est permise dans les rares cas où une personne répondant par procuration ne peut être interrogée sur place.

Chaque cycle de collecte de l'ENSP s'échelonne sur une période d'environ quatre mois. Au cycle 1, la collecte des données a eu lieu de janvier à mars 1995. Au cycle 2, la collecte des données s'est déroulée de décembre 1996 à avril 1997 et le dépistage des personnes qui avaient déménagé s'est poursuivi jusqu'en juin 1997. Au cycle 3, les interviews ont été menées de janvier à juin 1999. Aux cycles 4 et 5, les périodes de collecte ont eu lieu de février à mai 2001 et d'avril à juin 2003 respectivement.

### **6.2 Particularités de la collecte**

Les interviews sont menées par des intervieweurs de Statistique Canada. Tous les intervieweurs travaillent sous la supervision d'une équipe d'intervieweurs principaux. Ces derniers sont responsables de veiller à ce que les intervieweurs connaissent bien les concepts et les procédures de l'enquête. Ils contrôlent périodiquement le travail des intervieweurs et revoient les documents que ceux-ci remplissent. Les intervieweurs principaux sont, à leur tour, sous la surveillance des gestionnaires de programme postés dans les divers bureaux régionaux de Statistique Canada. À chaque cycle, toute l'équipe d'intervieweurs reçoit la formation nécessaire afin de réaliser les interviews de façon adéquate.

Un des points importants d'une enquête longitudinale est de rester en contact avec les membres du panel. S'ils déménagent dans un nouvel établissement ou un logement, ils sont toujours éligibles à être interviewés. Ainsi, lorsqu'un membre longitudinal déménage, tous les efforts sont mis en place pour tenter de le localiser à nouveau. Le travail de dépistage est en fait une composante importante du travail des intervieweurs.

Pour faciliter le dépistage, on demande aux répondants de l'information pour aider à les contacter lors du prochain cycle. Ils fournissent alors le nom, l'adresse et le numéro de téléphone de deux personnes que l'ENSP pourrait contacter afin de retrouver le répondant longitudinal dans l'éventualité d'un déménagement. Cette information permet à l'ENSP de poursuivre la collecte auprès des répondants longitudinaux. Il est très important, tout au fil de l'enquête, d'être capable de contacter les membres du panel.

Des efforts sont aussi déployés afin de minimiser la non-réponse. Les intervieweurs sont spécialement formés et des intervieweurs spécialisés ou expérimentés s'occupent des cas difficiles. Les refus de la part des établissements font aussi l'objet d'un suivi effectué par des intervieweurs principaux, des chargés de projet ou d'autres intervieweurs pour tenter de convaincre ces établissements de participer à l'enquête. Les intervieweurs reçoivent pour

instructions de faire tous les efforts raisonnables en vue d'interviewer les résidents sélectionnés. Toutefois, il faut noter que le nombre de refus est faible pour le volet établissements de soins de santé de l'ENSP (voir section 8.1.3).

En début de collecte, les intervieweurs principaux communiquent avec chaque établissement pour organiser une rencontre entre un intervieweur et l'administrateur ou une personne-ressource de l'établissement. Si le répondant a déménagé dans une nouvelle institution, l'intervieweur crée un nouveau FCE. Durant cette visite de liaison, l'intervieweur fait remplir un court questionnaire sur les politiques de l'établissement (FCE). La liste des résidents pour lesquels les réponses doivent être recueillies par une personne interposée est également dressée à ce moment-là et l'intervieweur obtient pour chacun de ces résidents le nom et le numéro de téléphone d'un proche parent. Puis, l'intervieweur téléphone aux proches parents auxquels on offre d'être interviewés eux-mêmes ou d'autoriser un membre informé du personnel de l'établissement à répondre aux questions de l'enquête. Aux cycles 1 et 2, on a aussi dressé la liste de tous les résidents habitant dans l'établissement et on a sélectionné l'échantillon lors de cette visite. Ensuite, le questionnaire des répondants (QR) est rempli avec le répondant ou avec une personne interposée. Lors de cette interview, la permission du répondant ou du plus proche parent est demandée afin que l'intervieweur recueille, auprès d'un membre du personnel de l'établissement, des renseignements sur le répondant concernant sa consommation de médicaments et ses consultations avec des professionnels de la santé. À la fin de chaque interview, une première vérification des données recueillies est effectuée sur place avant de quitter le répondant. L'intervieweur doit s'assurer que toutes les questions qui s'appliquent ont été posées et qu'une réponse a été obtenue, que toutes les inscriptions sont complètes et lisibles, que les instructions «Passez à» ont été bien suivies, que l'intervieweur a indiqué si l'interview s'est fait par personnes interposées ou non et qu'un code d'état final a été assigné.

Comme mentionné à la section 6.1, la plupart des interviews ont lieu sur place. La durée moyenne d'une interview varie un peu d'un cycle à l'autre. Toutefois, l'interview complète dure en moyenne autour de 40 minutes dans le cas de l'interview sans procuration et 30 minutes dans le cas de l'interview par procuration.

Finalement, il faut noter que le volet établissements de soins de santé sert aussi de véhicule pour la collecte des répondants du volet ménages qui sont déménagés dans un établissement de soins de santé.

### **6.3 Interviews par procuration**

Pour chaque établissement, l'administrateur ou une personne-ressource a déterminé pour lesquels des résidents sélectionnés les réponses devaient être recueillies par procuration parce qu'ils étaient malades ou souffraient d'une incapacité. La personne répondant par procuration pouvait être un parent, un membre du personnel ou un bénévole de l'établissement. Les taux pour les interviews par procuration sont donnés au tableau 6.1. Il faut aussi mentionner qu'un membre du personnel de l'établissement a fourni pour chaque résident sélectionné, des renseignements sur la consommation de médicaments, le poids et la taille (cycle 2 seulement) et les contacts avec les professionnels de la santé.

**Tableau 6.1 : Taux d'interviews par procuration par cycle et distribution selon le type**

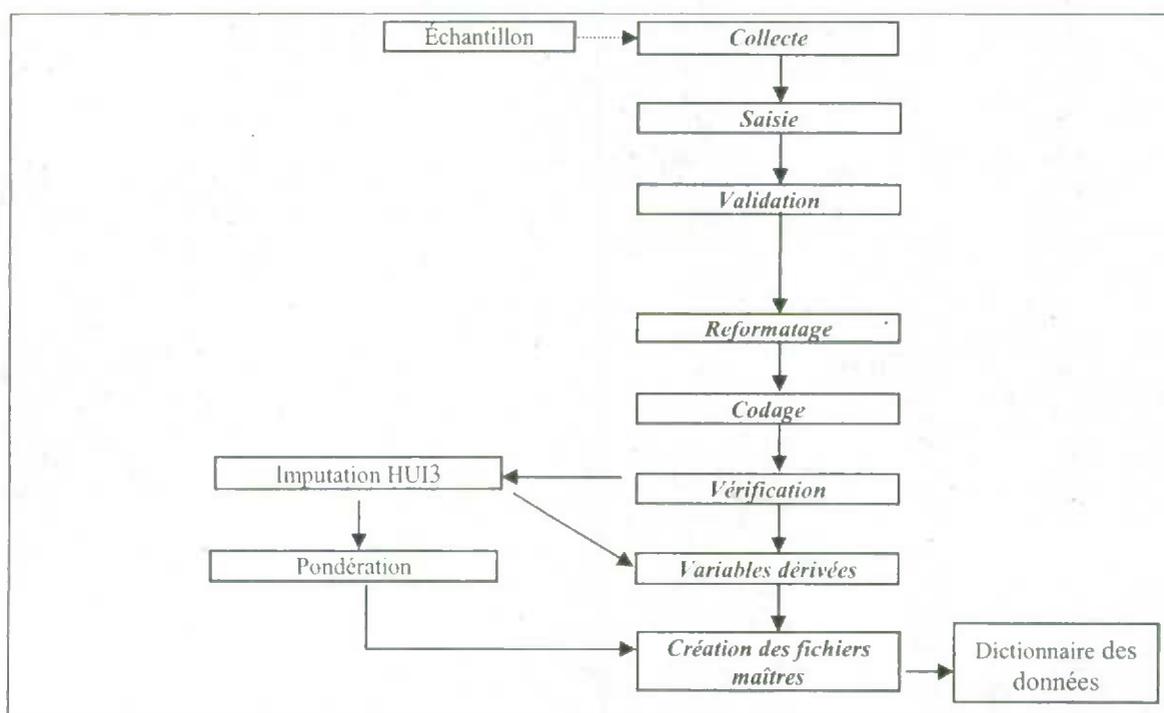
Cycle	Échantillon	Interviews par procuration	Type de personne interposée	
			<i>Plus proche parent</i>	<i>Membre du personnel ou bénévole de l'établissement</i>
1	Longitudinal/Transversal	57 %	71 %	29 %
2	Longitudinal	63 %	67 %	33 %
	Transversal	59 %	73 %	27 %
3	Longitudinal	70 %	47 %	53 %
4	Longitudinal	71 %	30 %	70 %
5	Longitudinal	72 %	34 %	66 %

## 7. Traitement des données

La figure 7.1 présente les différentes étapes du traitement permettant le passage des données collectées à la création des fichiers finaux (ceux-ci seront discutés à la section 9). Il s'agit, en ordre chronologique, de la saisie, de la validation, du reformatage, du codage, de la vérification, de la production des variables dérivées, de l'imputation de l'indice de l'état de santé (HUI3), de la pondération et de la création des fichiers maîtres.

La sous-section 7.1 présente d'abord un aperçu du processus entourant la collecte de données et de leurs envois au bureau central où les étapes du traitement sont exécutées. Les sous-sections qui suivent présentent à tour de rôle le détail de chacune des étapes du traitement tel qu'illustré à la figure 7.1.

Figure 7.1 : Traitement des données



### 7.1 Collecte des données et envoi postal

Pour recueillir les données de l'enquête, chaque intervieweur est muni de formulaires de contrôle des établissements (FCE) et de questionnaires pour les répondants (QR) en format papier (voir section 6). Tel que discuté à la sous-section 6.2, une première vérification des données recueillies est effectuée sur place avant de quitter le répondant. Ensuite, quand tous les QR pour un établissement sont remplis, l'intervieweur s'assure qu'il y a un QR pour chacune des personnes dont le nom figure sur le FCE, que le QR a été rempli pour chaque répondant, que le code d'état final est indiqué sur le FCE et que le code d'état final indiqué sur le QR est le même que celui qui figure sur le FCE. Après ces premières vérifications, les formulaires de contrôle pour les établissements et les QR dûment remplis sont envoyés dans les bureaux régionaux, pour être ensuite renvoyés au bureau central à Ottawa où le traitement des données est effectué. Pour les envois postaux, les intervieweurs doivent suivre la procédure indiquée dans le manuel Transmission de renseignements confidentiels par les intervieweurs ou les intervieweurs principaux.

## 7.2 Saisie des données

Tel que mentionné en 7.1, à la fin de chaque interview, l'intervieweur a examiné le QR pour s'assurer d'avoir suivi correctement l'enchaînement des questions. Les bureaux régionaux ont fait une vérification supplémentaire des questionnaires pour s'assurer que les réponses étaient complètes, lisibles et cohérentes. La première étape de traitement au bureau central est la saisie des données. Toutefois, avant la saisie, une vérification manuelle du FCE et du QR est faite au bureau central afin de minimiser les erreurs lors la saisie.

Le QR ainsi que le FCE ont été saisis dans les bureaux régionaux aux cycles 1, 2 et 3 et au bureau central aux cycles 4 et 5 en utilisant différents systèmes au cours des cycles. La méthode de reconnaissance de textes par codage optique (RTCO) ou la saisie manuelle à l'aide de différents programmes tels que DC2 et EP90 (« Entry Point 90 ») sont des outils utilisés pour la saisie de données. Les programmes de saisie de données des questionnaires sont conçus pour éviter la saisie de valeurs aberrantes. En général, l'information saisie a été vérifiée à 100 % dû au petit nombre de répondants.

Après la saisie des données, on s'est assuré que l'enchaînement des questions était correct et on a procédé à des contrôles de cohérence sur certaines zones des enregistrements.

## 7.3 Validation

Cette étape de traitement permet d'identifier différentes incohérences qui ont besoin d'être corrigées. La plupart des incohérences concernent la première page du FCE, la première page et la première section du QR. Par exemple, pour le FCE, il faut s'assurer que les numéros d'identification pour les établissements et les numéros séquentiels sont uniques. Il faut aussi s'assurer que les changements des numéros séquentiels sur les QR reflètent les changements sur les FCE. Toutes les erreurs et les incohérences sont corrigées.

Un appariement initial du FCE du QR est aussi fait à cette étape. Cet appariement sert également à identifier et corriger toutes les erreurs possibles à cette étape.

Le numéro d'identification pour les établissements est aussi mis à jour durant cette étape en utilisant la plus récente information. Cette variable est presque uniquement utilisée pour le fichier échantillon et la collecte. Toutefois, cette variable est aussi utilisée pour dériver les trois variables suivantes qui apparaissent sur le fichier des données : bureau régional, strate géographique-catégorie et numéro d'institution qui est arbitraire et unique dans chaque strate géographique-catégorie.

## 7.4 Reformatage

Le reformatage des données est un processus itératif appliqué aux variables jusqu'à l'obtention du résultat désiré. On parle de processus itératif entre l'équipe de traitement du volet établissements de soins de santé l'ENSP qui vérifie les résultats obtenus et celle des programmeurs qui apporte les corrections nécessaires aux programmes jusqu'à ce que l'on obtienne les résultats escomptés. Les activités principales sont les suivantes :

- reformater les valeurs saisies lors de la collecte selon des catégories de non-réponse standard. Les codes de non-réponse standard utilisés varient selon la longueur du champ

de la variable et sont : sans objet (6, 96, 996, ...), ne sait pas (7, 97, 997, ...), refus (8, 98, 998, ...) et non déclaré (9, 99, 999, ...);

- convertir en variables à caractère dichotomique “ oui (1) ou non (2)”, les catégories de toutes les réponses des questions avec la mention “cochez toutes les réponses appropriées”;
- reformater à sans objet certaines variables afin de s’assurer de bien respecter les valeurs plausibles selon les règles de branchement du questionnaire;
- préparer les fichiers nécessaires pour l’étape suivante soit le codage des réponses descriptives;
- assembler la partie du QR posée uniquement aux répondants vivant dans les établissements et celle posée uniquement aux répondants déménagés dans un ménage.

En parallèle avec cette étape de traitement, les résidents décédés de l’échantillon sont appariés à la Base canadienne de données de l’état civil (Décès) afin de confirmer le décès et, obtenir la date et la cause du décès.

## 7.5 Codage

Dans une enquête, on entend par codage l’opération par laquelle on attribue un code aux réponses descriptives. Cette étape facilite par la suite l’analyse des résultats de l’enquête. Dans le cadre du volet établissements de soins de santé de l’ENSP, ce processus est utilisé dans le cas de plusieurs questions pour lesquelles il était possible de donner une réponse descriptive. Les variables concernant les limitations d’activité, les médicaments et la cause de décès ont été codées à l’aide de systèmes de classification internationaux.

Pour tous les cycles, les problèmes de santé causant une limitation des activités ont été codés conformément à la Classification internationale des maladies, 9<sup>e</sup> édition (CIM-9) ou au Système de codage supplémentaire pour les incapacités ostéo-articulaires et musculaires établi au Canada aux fins de l’Enquête sur la santé et les limitations d’activités (ESLA). Au cycle 5, les problèmes de santé causant une limitation des activités ont aussi été codés conformément à la Classification statistique internationale des maladies et des problèmes de santé connexes, 10<sup>e</sup> révision (CIM-10), et ce, pour tous les cycles. Les deux codes sont disponibles sur le fichier longitudinal pour les cycles 1 à 5. La cause de décès a aussi été codée à l’aide des CIM-9 et CIM-10.

La classification des médicaments se fonde sur la Classification anatomique, thérapeutique chimique (ATC) développée par l’Organisation mondiale de la santé telle que contenue dans la Base de données sur les produits pharmaceutiques (BDPP) de Santé Canada en date de septembre 2003. Une révision complète des codes de médicaments a été faite pour tous les répondants longitudinaux de l’ENSP pour le cycle 5 (2002-2003) et pour tous les cycles précédents.

Finalement, pour toutes les autres variables acceptant des réponses autres que celles prédéterminées dans l’application, ces réponses sont soumises au codage afin de former de nouvelles catégories distinctes ou de les classer dans une catégorie existante.

## 7.6 Vérification

Une première vérification des données recueillies à l'aide du QR est effectuée sur place avant de quitter le répondant afin de relever les erreurs évidentes et d'effectuer certaines corrections (sous-section 6.2). Malheureusement, toutes les erreurs ne sont pas identifiées lors de cette première vérification et certaines règles de vérification doivent être appliquées au bureau central à ce moment-ci du traitement (règles de vérification du bureau central). On distingue deux types de règles de vérification du bureau central : les règles de validité et les règles de cohérences. Contrairement à la première vérification, très peu de corrections sont faites si une règle de vérification du bureau central n'est pas respectée. Aucun suivi auprès du répondant n'est fait. En général, les incohérences sont laissées comme telles, mais comptabilisées. Dans quelques cas, les incohérences sont corrigées en indiquant non déclaré à la question et dans de rares cas en prenant la valeur la plus plausible.

## 7.7 Variables dérivées

L'ENSP fournit à ses utilisateurs toute une panoplie de variables dérivées. Ces variables, créées à partir des éléments du QR de l'ENSP, ont plusieurs attraits : i) elles facilitent l'analyse des données, ii) elles permettent de protéger la confidentialité des répondants et iii) elles favorisent la cohérence entre différentes analyses. Dans certains cas, une variable dérivée est simplement obtenue en groupant des catégories de réponse d'une variable existante. Dans d'autres cas, plusieurs variables sont combinées pour en former une nouvelle.

Les variables dérivées peuvent être définies d'une façon différente d'un cycle à l'autre, mais en général on tente d'assurer une stabilité. Si la définition d'une variable dérivée est modifiée, on dérive à nouveau cette variable pour les cycles précédents dans le fichier longitudinal. L'ENSP a des variables dérivées concernant l'indice de l'état de santé (Health Utility index – HUI3), les problèmes de santé chroniques (le nombre de problèmes), la consommation d'alcool (type de buveur) et sur plusieurs autres sujets. Les renseignements concernant la dérivation de ces variables sont fournis en référence avec les fichiers de microdonnées (Statistique Canada, 2007).

C'est aussi à cette étape où on a eu recours à l'imputation pour calculer les valeurs manquantes de la variable, indice de l'état de santé (Health Utility index – HUI3). La méthode d'imputation utilisée par l'ENSP est décrite en détail à la sous-section 8.3.

## 7.8 Pondération

Le principe sur lequel repose l'estimation pour un échantillon probabiliste comme celui de l'ENSP veut que chacune des personnes faisant partie de l'échantillon représente, en plus d'elle-même, plusieurs autres personnes de la population. La phase de pondération est l'étape où l'on détermine ce poids pour chaque unité de l'échantillon. Pour le volet établissements de soins de santé de l'ENSP, des poids sont créés pour chacun des sous-ensembles de données. La méthode de pondération utilisée est décrite en détail à la section 10.

## 7.9 Création des fichiers finaux

La résultante de toutes les étapes du traitement est en fait la création des différents fichiers et des dictionnaires de données. Cette étape consiste à combiner les données traitées, les variables dérivées et les poids d'échantillonnage. Pour plus de détails sur la liste des produits, voir la section 9.

## **8. Non-réponse**

Le volet établissements de soins de santé de l'ENSP, comme toute enquête de nature volontaire, fait face à de la non-réponse. On distingue habituellement deux types de non-réponse; la *non-réponse totale* et la *non-réponse d'item* (aussi communément appelée *non-réponse partielle*). L'effet de la non-réponse sur les résultats de l'enquête constitue une source d'erreurs non dues à l'échantillonnage dans les enquêtes; il est donc important de s'y attarder. La sous-section 8.1 traite de la non-réponse totale, la sous-section 8.2 de la non-réponse d'item et la sous-section 8.3 traite de l'imputation.

### **8.1 Non-réponse totale**

Puisque le volet établissements de soins de santé l'ENSP était à la fois transversale et longitudinale au cours de ses deux premiers cycles, et que le concept de non-réponse n'est pas exactement défini de la même façon dans les deux cas, on examine d'abord la non-réponse totale pour les échantillons transversaux de l'ENSP dans la sous-section 8.1.1, puis la non-réponse totale d'un point de vue longitudinal dans la sous-section 8.1.2.

#### **8.1.1 Échantillons transversaux**

Il est important de distinguer les deux composantes de la collecte pour définir la non-réponse des échantillons transversaux : la composante établissements et la composante résidents (voir section 5 pour plus de détails). Pour la composante établissements, chaque établissement interviewé se voit attribuer un profil de réponse : accepte de participer, refuse de participer, aucun contact ou autre. Si l'établissement n'est pas hors du champ de l'enquête, il est considéré répondant s'il a accepté de participer à l'enquête sinon l'établissement est considéré non-répondant.

Pour la composante résidents, le profil de réponse attribué dépend du niveau de complétion du questionnaire. Des points critiques sont établis a priori dans le questionnaire pour définir les réponses partielles ou complètes. Un résident sera considéré répondant complet au questionnaire pour le répondant (QR) s'il a répondu à toutes les questions. Bien entendu, il n'a pas à répondre aux questions qui ne le concernent pas. Il sera considéré répondant partiel si le questionnaire a été complété jusqu'à la section E (État de santé) inclusivement. La personne sera considérée non-répondante si elle ne dépasse pas cette section.

Il faut noter que les fichiers de microdonnées transversales pour les utilisateurs et les partenaires contiennent uniquement les enregistrements des répondants complets ou partiels

Finalement, pour compenser cette perte d'échantillon due à la non-réponse totale, et contrôler le biais potentiel pouvant exister du fait que le profil des personnes répondantes peut être différent de celui des personnes non-répondantes, on ajuste le poids d'échantillonnage des répondants. La section 10 traite de la pondération et aborde les techniques utilisées pour faire cet ajustement.

#### **8.1.2 Échantillon longitudinal**

Tel que décrit à la section 4, l'échantillon longitudinal est composé des 2 287 personnes sélectionnées ayant répondu de façon complète ou partielle au QR du cycle 1 (voir définition en 8.1.1).

À chaque cycle, chacun des membres du panel se voit assigner un profil de réponse relatif au cycle en question. Suivant le résultat de l'interview, une personne se voit assigner un des quatre profils suivants : non-réponse, institutionnalisé, décédé ou en ménage. Similairement à ce qui est fait pour les échantillons transversaux (voir 8.1.1), il existe une série de règles pour distinguer ces profils. Si le membre longitudinal n'a pas au moins une réponse partielle au QR, il est, longitudinalement, considéré un non-répondant. La personne ayant un statut complet ou partiel est considérée un répondant complet et a le profil « institutionnalisé ». Finalement, on doit noter que le profil « en ménage » indique qu'une réponse complète ou partielle a été obtenue, mais que le répondant a déménagé dans un ménage.

Au fil des cycles, les profils de réponse d'un même répondant sont concaténés en une seule variable appelé le « patron de réponse longitudinal – LONGPAT ». Cette variable est disponible sur les fichiers longitudinaux et permet de connaître rapidement le profil de réponse d'un membre du panel. Cette variable est aussi utilisée pour définir les différents sous-ensembles analytiques tels que décrits à la section 9.

Dans une enquête longitudinale, la non-réponse à un ou plusieurs cycles, est très coûteuse. Elle rompt d'une certaine façon la séquence temporelle d'information disponible auprès d'un répondant, ce qui peut complexifier les analyses faites sur ces données. Pire encore, une non-réponse chronique à l'enquête réduit la taille de l'échantillon disponible pour l'analyse, diminuant ainsi le potentiel d'observer des résultats statistiquement significatifs. Par conséquent, de nombreux efforts doivent être déployés lors de la collecte afin de minimiser la non-réponse (voir sous-section 6.2 pour plus de détails).

### 8.1.3 Taux de réponse

Les taux de réponse pour l'échantillon transversal et l'échantillon longitudinal selon les cycles sont donnés au tableau 8.1. Pour chacun de ces échantillons, deux taux de réponse distincts peuvent être calculés, le taux de réponse à l'échelle des établissements et le taux de réponse à l'échelle de l'individu. Rappelons qu'à compter du cycle 3 (1998-1999), le volet établissements de soins de santé l'ENSP, est devenu uniquement une enquête longitudinale.

Pour l'échantillon longitudinal, les taux de réponse à l'échelle de l'individu rapportés pour les cycles 2, 3, 4 et 5 représentent la proportion des 2 287 répondants ayant un profil répondant complet, en ménage ou décédé. Le taux du cycle 1 représente quant à lui la proportion de réponses complètes ou partielles au QR parmi les personnes sélectionnées originalement à ce cycle.

Pour plus de détails concernant les taux de réponse, veuillez consulter les guides d'utilisateur fournis avec les données à chaque cycle (Statistique Canada, 1994, 1996, 1998, 2000 et 2002a)

**Tableau 8.1 : Taux de réponse par cycle selon l'échantillon longitudinal et l'échantillon transversal**

Cycle	Longitudinal		Transversal	
	Établissement	Individuel	Établissement	Individuel
Cycle 1	95,5 %	93,6 %	95,5 %	93,6 %
Cycle 2	100,0 %	95,9 %	100,0 %	88,9 %
Cycle 3	100,0 %	98,4 %		
Cycle 4	99,3 %	96,9 %		
Cycle 5	100,0 %	97,7 %		

## 8.2 Non-réponse d'item

Il y a non-réponse d'item lorsque l'information est disponible pour certaines questions seulement, par exemple, lorsque la personne répond à une partie seulement du QR. En termes de contenu dans les fichiers de données, on réfère souvent à ces non-réponses comme étant des valeurs manquantes. Au volet établissements de soins de santé de l'ENSP, on distingue quatre catégories de valeurs manquantes :

- Sans objet : Indique que la question ne s'appliquait pas au répondant.
- Ne sait pas : Indique que le répondant n'a pu fournir de réponse à la question.
- Refus : Indique que le répondant a refusé de répondre à la question.
- Non déclaré : Indique que l'information n'a pu être rapportée du fait que la question n'a pas été posée alors qu'elle aurait dû l'être. Typiquement, il s'agit de questions associées à un filtre, où la question filtre n'a pas été répondue, empêchant ainsi l'obtention de réponses aux questions conditionnelles à ce filtre. Il est à noter que toutes les variables des personnes décédées sont mises à « non déclaré » à partir du cycle auquel elles sont identifiées comme étant décédées.

Il faut noter que les valeurs manquantes sont recodées dans les fichiers de données, utilisant un système de codification répandu dans les enquêtes auprès des ménages à Statistique Canada (voir sous-section 7.4).

Une approche populaire dans le domaine des enquêtes est d'imputer les données manquantes, c'est-à-dire d'attribuer une valeur de remplacement pour une valeur manquante. Dans le cas du volet établissements de soins de santé de l'ENSP, il n'y a pas d'imputation, sauf pour une variable (voir section 8.3), laissant ainsi aux analystes le soin de traiter eux-mêmes les valeurs manquantes.

## 8.3 Imputation

On a eu recours à l'imputation pour calculer les valeurs manquantes de la variable, indice de l'état de santé (Health Utility index – HUI3) (pour plus de détails sur cette variable, voir la documentation sur les variables dérivées, Statistique Canada 2007). Cet indice de l'état de santé global englobe l'évaluation de la vue, de l'ouïe, de l'élocution, de la mobilité (capacité à se déplacer), de la dextérité (mouvement des mains et des doigts), des sentiments, de la cognition (mémoire et pensée) et de la douleur. La cote globale de l'HUI3, qui peut varier de -0,360 à 1,000, est calculée d'après les réponses à une série de questions sur l'état de santé. Or, il est

impossible de calculer cette cote globale si la réponse concernant l'une ou plusieurs de ces composantes est manquante, situation qui s'est produite pour environ 10 % des répondants. Il a été décidé de procéder à l'imputation des valeurs manquantes pour le calcul de l'HUI3 pour les cinq cycles du volet établissements de soins de santé. On a donc recouru à une forme d'imputation par la méthode « hot deck » pour imputer des valeurs aux composantes manquantes, afin de pouvoir calculer l'HUI3 global pour les personnes concernées.

L'HUI3 a été calculé d'après les réponses aux questions de huit attributs de la section sur l'état de santé du QR. On a calculé une cote partielle pour chacun des attributs, puis on a poursuivi les calculs sur ces cotes partielles pour obtenir la cote globale de l'HUI3. L'imputation a été exécutée pour les huit cotes partielles plutôt que pour les questions. Après l'imputation, on a modifié légèrement le programme pour le calcul de la variable dérivée de l'HUI3 afin qu'il choisisse comme données d'entrée les huit valeurs imputées pour la vue, l'ouïe, l'élocution, la mobilité, les sentiments, la cognition, la dextérité et la douleur.

L'imputation a été effectuée en deux étapes. La première étape correspond à une imputation déterministe. Dans certains cas, même si la personne n'avait pas répondu à la question fournissant la cote partielle, on possédait suffisamment de renseignements pour déduire cette cote avec certitude. On a donc attribué la cote partielle d'après ces renseignements partiels dans tous les cas où il a été jugé approprié de le faire.

La deuxième étape correspond à une imputation « hot deck » par donneur pour attribuer les cotes partielles manquantes. La méthode du plus proche voisin a été utilisée pour identifier les donneurs. Pour identifier un plus proche voisin, un HUI3 temporaire a été calculé en ne tenant compte que des cotes partielles ne contenant pas de données manquantes.

## 9. Différents produits

Plusieurs types de fichiers de données sont produits avec les données recueillies par le volet établissements de soins de santé de l'ENSP. En plus des aspects transversal et longitudinal, les différences entre les produits se situent principalement dans le nombre de variables ou d'observations inclus dans ces fichiers.

Des fichiers transversaux ont été produits aux cycles 1 et 2. Dès le cycle 2 de l'ENSP, des fichiers longitudinaux ont pu être produits. On distingue deux types de fichiers longitudinaux : *carré* et *complet*. Le *fichier carré* contient les données pour tous les 2 287 membres du panel longitudinal, peu importe le profil de réponse de leur QR (non-réponse, complet, décédé ou en ménage). Le nombre d'enregistrements inclus dans le *fichier carré* est donc identique d'un cycle à l'autre, mais le nombre de variables quant à lui augmente après chaque cycle. En effet, après chaque cycle de l'enquête, les données recueillies sont ajoutées à celles des cycles précédents. Pour sa part, le *fichier complet* ne contient que les répondants avec un profil de réponse « complet », « partiel », « décédé » ou « en ménage » pour le cycle en cours ainsi que pour tous les cycles précédents. Par conséquent, le nombre d'enregistrements contenu dans le *fichier complet* ne fait que décroître dans le temps dû à la non-réponse.

Puisque les données sont utilisées par des utilisateurs de diverses provenances, différentes versions de ces fichiers doivent être produites afin de permettre le partage des données avec d'autres organismes ou partenaires d'enquête et de s'assurer de préserver la confidentialité des données. Trois versions de fichiers existent à l'ENSP : les *fichiers maîtres*, les *fichiers de partage* et les *fichiers de microdonnées à grande diffusion (FMGD)*.

Les *fichiers maîtres* contiennent l'ensemble des répondants avec toutes les variables de l'enquête. Seules les variables qui permettent d'identifier directement le répondant (nom, adresse complète, etc.) ont été enlevées. Ces fichiers sont utilisés par les employés assermentés de Statistique Canada seulement.

Les *fichiers de partage* contiennent les mêmes variables que les fichiers maîtres, mais excluent les répondants qui ont refusé que leurs données soient partagées avec d'autres organismes qui sont soumis à la Loi sur la statistique (c'est-à-dire qui assurent la confidentialité des données). Pour obtenir la liste des organismes ou des partenaires avec lesquels l'information recueillie est partagée, se référer aux questionnaires (voir section 5). À titre indicatif, environ 95 % des répondants acceptent habituellement de partager leurs données.

Finalement, les *FMGD* contiennent, tout comme les fichiers maîtres, l'ensemble des répondants, mais excluent les variables qui permettraient d'identifier directement ou indirectement les répondants. Ces fichiers sont accessibles à tous ceux qui en font la demande moyennant des frais. La section 12 décrit plus en détail les aspects relatifs à la création de ces FMGD.

Le croisement de toutes ces caractéristiques génère un nombre considérable de fichiers produits à chaque cycle. De plus, chaque cycle possède ses particularités et présente un certain nombre d'exceptions. Le tableau 9.1 résume l'ensemble des fichiers de données qui ont été produits pour chacun des cycles. Les fichiers de poids bootstrap sont également présentés dans le tableau et sont décrits à la sous-section 11.1.1.2.

**Tableau 9.1 : Fichiers de données produits (et taille correspondante) pour chacun des cycles**

Fichiers	Cycle 1		Cycle 2		Cycle 3		Cycle 4		Cycle 5	
	Fichier de données pondérées	Fichier de poids bootstrap	Fichier de données pondérées	Fichier de poids bootstrap	Fichier de données pondérés	Fichier de poids bootstrap	Fichier de données pondérées	Fichier de Poids bootstrap	Fichier de données pondérées	Fichier de poids bootstrap
	Nombre d'obs.	Nombre de poids	Nombre d'obs.	Nombre de poids	Nombre d'obs.	Nombre de poids	Nombre d'obs.	Nombre de poids	Nombre d'obs.	Nombre de poids
<b>Fichiers maîtres</b>										
<b>Transversal :</b>	2 287	n.d.	2 118	2 000	s.o.	s.o.	s.o.	s.o.	s.o.	s.o.
<b>Longitudinal :</b>										
LNGF – complet	s.o.	s.o.	2 192	n.d.	2 178	2 000	2 143	2 000	2 131	2 000
LONG – carré	n.d.	n.d.	2 287	n.d.	2 287	n.d.	2 287	n.d.	2 287	n.d.
<b>Fichiers de partage</b>										
<b>Transversal :</b>	n.d.	n.d.	2 055	n.d.	s.o.	s.o.	s.o.	s.o.	s.o.	s.o.
<b>Longitudinal :</b>										
Complet	n.d.	n.d.	n.d.	n.d.	2 121	2 000	2 080	2 000	2 067	2 000
<b>Fichiers FMGD</b>										
<b>Transversal</b>	2 287	n.d.	2 118	2 000	s.o.	s.o.	s.o.	s.o.	s.o.	s.o.

Note : n.d. : non disponible  
s.o. : sans objet

## 10. Pondération et estimation

Cette section présente les détails de la pondération des cinq cycles du volet établissements de soins de santé de l'ENSP. La sous-section 10.1 traite de la pondération du cycle 1 des fichiers transversaux maître et à grande diffusion (FMGD) et des sous-ensembles longitudinaux carrés des cycles 1 à 5. La sous-section 10.2 se penche sur la pondération transversale du cycle 2 alors que la sous-section 10.3 concerne la pondération longitudinale des cycles 2 à 5. Le tableau 10.1 énumère les ensembles de poids du volet établissements de soins de santé de l'ENSP.

Tableau 10.1 – Ensembles de poids de l'ENSP

	Transversal	Longitudinal	
		Carré	Complet
<b>Cycle 1</b>			
Maître	WTI4WGHT	WTI4LS	s.o.
Partagé	s.o.	s.o.	s.o.
MGD	WTI4LS	s.o.	s.o.
<b>Cycle 2</b>			
Maître	WTI6WGHT	WTI4LS	WTI6LF
Partagé	WTI6 S	s.o.	s.o.
MGD	WTI6WGHT	s.o.	s.o.
<b>Cycle 3</b>			
Maître	s.o.	WTI4LS	WTI8LF
Partagé	s.o.	s.o.	WTI8SLF
<b>Cycle 4</b>			
Maître	s.o.	WTI4LS	WTI0LF
Partagé	s.o.	s.o.	WTI0SLF
<b>Cycle 5</b>			
Maître	s.o.	WTI4LS	WTI2LF
Partagé	s.o.	s.o.	WTI2SLF

s.o. : sans objet

### 10.1 Pondération transversale du cycle 1 et pondération longitudinale du sous-ensemble carré

Les poids attribués aux unités répondantes du volet établissements de soins de santé de l'ENSP sont fondés sur les probabilités de sélection des établissements et des résidents (voir le plan d'échantillonnage section 4). Ensuite, des ajustements pour la non-réponse des établissements et des résidents sont apportés.

Pour faciliter la description du processus de pondération, la notation suivante est utilisée :

- $M_h$  : nombre de lits dans la strate  $h$  (d'après la liste des hôpitaux et des établissements de soins)
- $M_{h,i}$  : nombre de lits dans la strate  $h$ , établissement  $i$  (d'après la liste des hôpitaux et des établissements de soins)
- $n_h$  : nombre d'établissements à sélectionner dans la strate de taille  $h$
- $L_{h,i}$  : nombre réel de résidents de longue durée dans la strate  $h$ , établissement  $i$  (obtenu au moment de la première visite)
- $r_{h,i}$  : nombre de résidents à sélectionner dans la strate  $h$ , établissement  $i$

### 10.1.1 Probabilité de sélection des établissements au cycle 1 (1994-1995)

Les établissements ont été sélectionnés à partir de la base de sondage, la probabilité de sélection étant proportionnelle au nombre de lits. Par conséquent, dans la plupart des cas, la probabilité de sélection d'un établissement  $i$  est la suivante :

$$a_i = n_h \times \frac{M_{h,i}}{M_h}$$

Dans les cas où un siège social a été sélectionné (pour plus de détails, voir la section 4.1.4.1), la probabilité de sélection est la suivante :

$$a_i = n_h \times \frac{M_{h,i}}{M_h} \times P_{h,i,j}$$

où  $P_{h,i,j}$  est la probabilité qu'un établissement  $j$  relevant du siège social  $i$  soit sélectionné. Dans le cas du plus grand établissement relevant de  $i$ ,  $P_{h,i,j}=1$ . Dans le cas des autres établissements  $j$  :

$$P_{h,i,j} = \frac{M_{h,i,j}}{\sum_{j \in i'} M_{h,i,j}}$$

où  $i'$  représente tous les établissements relevant du siège social  $i$ , à l'exclusion du plus grand.

### 10.1.2 Poids de base, ajustement pour la non-réponse et poids final de l'établissement

Le poids d'un établissement correspond au nombre d'établissements que représente l'établissement échantillonné. Le **poids de base de l'établissement** ( $W_{Ebi}$ ) est donc égal à l'inverse de la probabilité de sélectionner cet établissement ( $W_{Ebi} = 1/a_i$ ).

La non-réponse étant possible, un ajustement est donc nécessaire pour tenir compte des établissements qui refusent de participer. Dans les cas où des interviews n'ont pu être réalisées dans un établissement sélectionné faisant partie du champ de l'enquête, une correction est apportée aux poids des établissements répondants appartenant à la même strate de taille pour tenir compte de l'établissement non-répondant. Cet ajustement correspond à :

$$C_E = \frac{\text{nombre d'établissements répondants et non-répondants}}{\text{nombre d'établissements répondants}}$$

La multiplication du poids de base de l'établissement ( $W_{Ebi}$ ) par ce facteur d'ajustement ( $C_E$ ) donne le **poids final de l'établissement au cycle 1**

$$W_{Efi} = C_E \times W_{Ebi}$$

### 10.1.3 Probabilité de sélection des résidents au cycle 1 (1994-1995)

Une fois un établissement sélectionné, chaque résident de cet établissement avait une probabilité égale d'être sélectionné, la probabilité est définie par :

$$d = \begin{cases} \frac{r_{h,i}}{L_{h,i}} & \text{si } L_{h,i} \geq r_{h,i} \\ 1 & \text{si } L_{h,i} < r_{h,i} \end{cases}$$

### 10.1.4 Poids de base, ajustement pour la non-réponse et poids final du résident

Le **poids de base du résident** ( $W_{Rb1}$ ) s'obtient en multipliant le poids final de l'établissement par l'inverse de la probabilité de sélection d'un résident à l'intérieur de l'établissement ( $W_{Rb1} = W_{Ef1} / d$ ).

Il est possible que certains résidents sélectionnés n'aient pas répondu au questionnaire pour le répondant (QR). Un ajustement supplémentaire doit alors être apporté aux poids de base des résidents pour tenir compte de la non-réponse des résidents. Cet ajustement correspond à :

$$C_R = \frac{\text{nombre de résidents sélectionnés répondants et non-répondants}}{\text{nombre de résidents sélectionnés répondants}}$$

Le **poids final du résident** (WTI4LS) s'obtient en multipliant le poids de base du résident par l'ajustement pour la non-réponse des résidents

$$WTI4LS = W_{Rb1} \times C_R .$$

Le poids final du résident (WTI4LS) est le poids associé aux fichiers transversaux (maître et FMDG) du cycle 1 et des sous-ensembles longitudinaux carrés des cycles 1 à 5.

## 10.2 Pondération transversale du cycle 2 (1996-1997)

La méthode de pondération du cycle 2 (1995-1996) est fondée en grande partie sur celle du cycle 1 (1994-1995). Les sous-sections 10.2.1 à 10.2.7 décrivent toutes les étapes requises pour la création des ensembles de poids transversaux du cycle 2 du volet établissements de soins de santé de l'ENSP.

### 10.2.1 Poids final d'un établissement sélectionné au cycle 1 (1994-1995)

Pour les établissements qui ont participé au premier cycle de l'enquête, on prend comme point de départ le poids final de l'établissement au cycle 1 ( $W_{Ef1}$ ). Pour plus de détails sur les étapes pour arriver au poids final de l'établissement au cycle 1, veuillez consulter la section 10.1.1 et 10.1.2.

### 10.2.2 Poids final d'un établissement sélectionné au cycle 2 (1996-1997)

Pour les établissements sélectionnés au cycle 2, il faut d'abord calculer leur probabilité de sélection. Comme on l'a décrit à la section 4.3.2.2, les nouveaux établissements ont été sélectionnés conformément au plan de sondage original. Par conséquent, la probabilité de sélection d'un nouvel établissement correspond à :

$$a_2 = n_h \times \frac{M_{h,i}}{M_h}$$

où, ici,  $M_{h,i}$  représente le nombre de lits dans l'établissement  $i$  de la strate  $h$  provenant de la base de sondage de 1996. Huit nouveaux établissements ont été sélectionnés au cycle 2 conformément à cette méthode de sélection. De ces établissements, deux ont été jugés hors du champ de l'enquête (parce qu'ils ne fournissaient que des soins de courte durée); les autres ont accepté de participer à l'enquête. Le **poids final de l'établissement sélectionné au cycle 2** ( $W_{Ef_2}$ ) est donc l'inverse de la probabilité de sélection de ce nouvel établissement

$$W_{Ef_2} = 1 / a_2 .$$

### 10.2.3 Probabilité de sélectionner un résident dans les établissements sélectionnés au cycle 1 (1994-1995)

La sélection du nouvel échantillon de résidents doit répondre à deux grands objectifs, à savoir l'obtention d'un échantillon représentatif et l'utilisation d'une méthode d'échantillonnage efficace. Par représentativité, on entend que la probabilité de sélection doit être non nulle pour chaque résident, y compris les nouveaux résidents. Pour que la méthode d'échantillonnage soit efficace, autrement dit que la variance des estimations soit faible, il est souhaitable que les poids d'échantillonnage appliqués aux résidents sélectionnés dans un établissement soient aussi semblables que possible. De surcroît, l'échantillon devrait être représentatif de la distribution de la population de l'établissement en ce qui concerne les résidents nouvellement admis par opposition aux résidents qui vivaient déjà dans l'établissement au moment de l'enquête du premier cycle. Remplacer simplement les résidents sélectionnés au premier cycle qui ne vivent plus dans l'établissement par un échantillon de résidents nouvellement admis peut produire des écarts importants entre les poids d'échantillonnage. En outre, si le taux d'érosion de la population d'un établissement diffère considérablement de celui observé pour l'échantillon prélevé dans cet établissement, il se peut que l'on doive sélectionner certains résidents qui vivaient déjà dans cet établissement au moment du premier cycle. Pour chaque établissement, on a appliqué la fraction de sondage de 1994 aux deux sous-populations pour déterminer combien de résidents supplémentaires il fallait interviewer. De cette façon, on a rendu les poids d'échantillonnage des résidents sélectionnés dans un établissement aussi semblables que possible. Pour faciliter l'écriture de la probabilité de sélection, la notation suivante est utilisée :

- $R_{h,i,b}$  = nombre de résidents recevant des soins de longue durée dans l'établissement  $i$  de la strate  $h$  admis avant le 1<sup>er</sup> avril 1995;
- $r_{h,i,b}$  = nombre de résidents de l'établissement  $i$  de la strate  $h$  admis avant le 1<sup>er</sup> avril 1995 et désignés pour être sélectionnés (sauf les membres du panel longitudinal);
- $R_{h,i,a}$  = nombre de résidents recevant des soins de longue durée dans l'établissement  $i$  de la strate  $h$  admis le 1<sup>er</sup> avril 1995 ou après cette date;
- $r_{h,i,a}$  = nombre de résidents de l'établissement  $i$  de la strate  $h$  admis le 1<sup>er</sup> avril 1995 ou après cette date et désignés pour être sélectionnés; et
- $r_{\text{panel}}$  = nombre de membres du panel longitudinal vivant encore dans l'établissement  $i$  de la strate  $h$ .

Dans un établissement sélectionné au cycle 1, la probabilité qu'un résident soit sélectionné est définie par :

$$d_1 = \begin{cases} \frac{r_{h,i,b} + r_{panel}}{R_{h,i,b}}, & \text{si résident admis avant le 1<sup>er</sup> avril 1995 ou membre du panel longitudinal} \\ \frac{r_{h,i,a}}{R_{h,i,a}}, & \text{si résident admis le 1<sup>er</sup> avril 1995 ou après cette date.} \end{cases}$$

#### 10.2.4 Probabilité de sélectionner un résident dans les établissements sélectionnés au cycle 2 (1996-1997)

Une fois un établissement sélectionné, chaque résident de cet établissement avait une probabilité égale d'être sélectionné, la probabilité est définie par :

$$d_2 = \frac{r_{h,i}}{R_{h,i}},$$

où  $R_{h,i}$  est le nombre réel de résidents recevant des soins de longue durée dans l'établissement  $i$  de la strate  $h$ ; et  $r_{h,i}$  est le nombre de résidents de l'établissement  $i$  de la strate  $h$  désignés pour être sélectionnés.

#### 10.2.5 Poids de base du résident au cycle 2 (1996-1997)

Pour calculer le poids de base applicable à un résident, on multiplie le poids final de l'établissement par l'inverse de la probabilité de sélection d'un résident dans l'établissement:

$$W_{Rb2} = W_{Ef1} / d_1, \text{ si établissement du cycle 1 ou} \\ W_{Rb2} = W_{Ef2} / d_2, \text{ si établissement du cycle 2.}$$

#### 10.2.6 Ajustement pour les décès

Comme, dans certains cas, il y a eu un décalage important entre le moment de la sélection de l'échantillon et le début de l'enquête, le nombre de décès observés durant l'enquête a été beaucoup plus élevé que le nombre prévu. Puisque le profil des décès chez les résidents des établissements de santé n'est pas aléatoire (les personnes les moins bien portantes étant plus susceptibles que les autres de mourir), les estimations risquent d'être biaisées si aucune correction n'est faite pour remédier à ce problème. Pour tenir compte du fait que certains résidents sélectionnés sont décédés durant la période de collecte, on a calculé un facteur d'ajustement pour les décès que l'on a appliqué aux poids de base des résidents.

Le calcul de cet ajustement pour les décès comprend de nombreuses étapes. En premier lieu, en utilisant les données recueillies au cycle 1, des classes de pondération sont définies en regardant les caractéristiques des résidents qui sont ou non décédés dans les six mois qui ont suivi l'interview du premier cycle. Les classes sont formées en regroupant les résidents qui affichent le même risque d'être décédés dans les six mois suivant l'interview du cycle 1. Ces classes ont été définies à l'aide d'un algorithme de segmentation qui permet de répartir les résidents selon certaines caractéristiques. La version améliorée de l'algorithme CHAID (Chi-Square Automatic Interaction Detector) offerte par le logiciel Knowledge Seeker IV pour Windows (Angoss

Software, 1995), a été utilisée pour produire la structure arborescente. À l'aide des données du cycle 1, pour chaque classe, un ratio est défini par :

$$R = \frac{\text{somme des poids de 1994 pour les membres du panel décédés six mois après l' interview du premier cycle dans la classe}}{\text{somme des poids de 1994 des membres du panel non décédés six mois après l' interview du premier cycle dans la classe}}$$

Les classes ont été définies d'après des caractéristiques observées au premier cycle, toutefois, les données sur les caractéristiques des résidents du deuxième cycle sont utilisées pour définir l'appartenance à une classe. En fait, l'hypothèse étant que les classes et le ratio R pour les données du deuxième cycle devraient être fort semblables à ceux définis d'après les données du premier cycle.

Idéalement, la somme des poids des résidents sélectionnés qui sont décédés durant la période de collecte devrait être redistribuée dans chaque strate/classe. Cependant, cette méthode produirait des corrections instables étant donné le petit nombre d'enregistrements dans certaines combinaisons strate-classe. Pour respecter autant que possible la stratification originale (voir section 4.1.1) et produire des corrections plus stables, on a subdivisé l'échantillon en six poststrates, à savoir quatre pour les établissements pour personnes âgées, une pour les établissements de soins cognitifs et une pour les autres établissements de réadaptation. Dans le cas des établissements pour les personnes âgées, chaque région comprend quatre à cinq strates de taille. Les quatre poststrates ont été formées de façon à ce qu'elles soient à peu près de même taille (en ce qui a trait au total des poids et au nombre d'enregistrements) et correspondent, dans l'ensemble, aux établissements de petite, moyenne, grande et très grande taille. On notera que la majorité des résidents sélectionnés qui sont décédés durant la période de collecte vivaient dans des établissements pour personnes âgées.

Ensuite, les résidents répondants du cycle 2 sont divisés dans les groupes « poststrate-classe ». Puis, la somme des  $W_{Rb_2}$  (poids de base des résidents) est calculée pour chaque groupe « poststrate-classe ». En multipliant cette somme par le ratio R (tel que défini plus haut) déterminé pour la classe en question, on obtient la somme  $W_{Rb_2}$  (poids de base des résidents) sélectionnés décédés pour chaque groupe « poststrate-classe ».

Finalement, l'ajustement pour les décès est donné par :

$$C_d = \frac{\text{Somme des } W_{Rb_2} \text{ répondants} + \text{Somme des } W_{Rb_2} \text{ sélectionnés décédés dans le groupe « poststrate - classe »}}{\text{Somme des } W_{Rb_2} \text{ répondants dans le groupe « poststrate - classe »}}$$

La multiplication du poids de base du résident ( $W_{Rb_2}$ ) répondant par ce facteur d'ajustement ( $C_d$ ) donne le **poids de base du résident répondant après l'ajustement pour les décès** :

$$W_{Rb_d} = C_d \times W_{Rb_2} .$$

### 10.2.7 Ajustement pour la non-réponse des résidents

Comme certains résidents sélectionnés n'ont pas répondu au QR, un ajustement supplémentaire doit être apporté au poids de base des résidents répondants après l'ajustement pour les décès ( $W_{Rb_d}$ ) pour tenir compte de la non-réponse. Comme on l'a décrit à la section 4.2.2, l'échantillon

du deuxième cycle comprend les membres du panel longitudinal ainsi que des résidents nouvellement sélectionnés. Pour les membres du panel longitudinal (représentant environ 60 % de la taille de l'échantillon), l'âge et le sexe sont connus. Pour les résidents nouvellement sélectionnés lors du deuxième cycle qui ont refusé de répondre, seulement le sexe est connu pour pratiquement l'ensemble des résidents. L'ajustement est donc calculé pour chaque strate catégorie, âge et sexe pour les membres du panel et pour chaque strate catégorie et sexe pour les nouveaux résidents sélectionnés au cycle 2.

Pour les membres du panel, l'ajustement pour la non-réponse des résidents est donné par :

$$C_R = \frac{\text{somme des } W_{Rb1} \text{ répondants et } W_{Rb2} \text{ non - répondants dans la strate catégorie/âge/sexe}}{\text{somme des } W_{Rd1} \text{ répondants dans la strate catégorie/âge/sexe}}$$

Il convient de souligner que certains groupes originaux âge-sexe ont été regroupés pour les établissements de soins cognitifs et pour les autres établissements de réadaptation. L'objet du regroupement était de produire une correction plus stable.

Pour les résidents sélectionnés au cycle 2, l'ajustement pour la non-réponse des résidents est donné par :

$$C_R = \frac{\text{somme des } W_{Rb1} \text{ répondants et } W_{Rb2} \text{ non - répondants dans la strate catégorie/ sexe}}{\text{somme des } W_{Rdb1} \text{ répondants dans la strate catégorie/ sexe}}$$

### 10.2.8 Poids final du résident associé aux fichiers maître et FMGD transversaux du cycle 2

La multiplication du poids après l'ajustement pour les décès ( $W_{Rbd}$ ) par l'ajustement pour la non-réponse des résidents donne le poids final du résident :

$$WTI6WGHT = C_R \times W_{Rbd}$$

Le poids final du résident (WTI6WGHT) est le poids associé aux fichiers transversaux (maître et FMGD) du cycle 2.

### 10.2.9 Poids final du résident associé au fichier partagé transversal du cycle 2

Le fichier partagé transversal du cycle 2 inclut tous les résidents qui ont accepté de partager l'information recueillie lors des interviews menées dans le cadre de l'ENSP avec les ministères provinciaux de la santé et Santé Canada. Étant donné que ces partenaires reçoivent seulement les enregistrements des résidents qui ont accepté de partager leur information (soit 97 % des résidents inclus dans le fichier maître), un poids spécial doit être créé de telle sorte que les estimations calculées à partir de ce sous-ensemble représentent correctement la population totale. Un seul ajustement simple est apporté au poids du fichier maître transversal (WTI6WGHT) pour créer les poids du fichier partagé (WTI6\_S). Cet ajustement est donné par :

$$C_s = \frac{\text{somme des WTI6WGHT dans la strate catégorie/âge/sexe}}{\text{somme des WTI6WGHT ayant accepté de partager dans la strate catégorie/âge/sexe}}$$

Il est à noter que cet ajustement est apporté à l'intérieur de chacun des groupes, strate catégorie/âge/sexe, tels que définis lors de l'ajustement de la non-réponse des résidents (voir 10.2.7). Le poids final pour le fichier partagé (WTI6\_S) est obtenu en multipliant le poids du fichier maître transversal (WTI6WGHT) par cet ajustement :

$$\text{WTI6\_S} = \text{WTI6WGHT} * C_s.$$

### 10.3 Pondération longitudinale

La présente section décrit le processus de pondération longitudinale pour chacun des sous-ensembles de répondants décrits à la section 9. Le processus de pondération longitudinale est nécessairement différent de celui de la pondération transversale pour plusieurs raisons. Premièrement, les poids longitudinaux doivent représenter la probabilité de sélection de l'unité d'analyse au moment de la sélection de l'échantillon. Puisque l'échantillon longitudinal a été sélectionné en 1994-1995, les poids doivent représenter la probabilité de sélectionner l'individu au cycle 1, et non dans les cycles qui suivent. De plus, la définition d'une réponse longitudinale est différente d'une réponse transversale (voir 8.1), ce qui nécessite des ajustements différents particuliers à chaque type de non-réponse. Malgré le fait que les membres du panel font partie à la fois des fichiers transversaux et longitudinaux, leurs poids ne sont pas identiques dans ces deux types de fichier, mais plutôt ajustés afin de bien représenter la population d'intérêt. Les sous-sections suivantes décrivent la procédure utilisée pour la pondération longitudinale pour les sous-ensembles complet et complet partagé. La pondération longitudinale du sous-ensemble carré est décrite à la sous-section 10.1.

#### 10.3.1 Pondération pour les sous-ensembles complets des cycles 2 à 5

Le sous-ensemble longitudinal complet inclut seulement les membres du panel qui ont fourni une réponse complète, c'est-à-dire les membres qui ont eu un profil de réponse « institutionnalisé », « décédé » ou « en ménage », à chaque cycle (voir 8.1.2). Les membres du panel exclus de ce sous-ensemble ont donc été non-répondants, c'est-à-dire qu'ils ont eu un profil « non-répondant », à un certain point durant les cycles de l'enquête, et leur poids doit donc être redistribué pour compenser la non-réponse.

Le point de départ est le poids longitudinal carré (WTI4LS) (voir section 10.1) et des ajustements pour la non-réponse sont faits. Un ajustement différent est effectué pour la non-réponse à chacun des cycles et ces ajustements sont cumulatifs d'un cycle à l'autre. Par exemple, pour obtenir le poids du cycle 5, les ajustements pour la non-réponse aux cycles 2, 3, 4 et 5 sont appliqués successivement au poids WTI4LS.

Les ajustements nécessaires en vue d'obtenir le poids longitudinal complet du cycle 5 sont décrits ci-dessous.

#### Ajustement 1 : Ajustement pour la non-réponse au cycle 2 (1996-1997)

Afin de corriger pour les membres longitudinaux qui n'ont pas répondu au cycle 2, l'ajustement qui suit est apporté aux poids des résidents répondants :

$$a_1 = \frac{\text{somme des poids des résidents répondants et non - répondants}}{\text{somme des poids des résidents répondants}}$$

Cet ajustement est fait au niveau des groupes de non-réponse, à l'échelle du Canada. Les classes sont déterminées à l'aide de l'algorithme CHAID.

#### Ajustement 2 : Ajustement pour la non-réponse au cycle 3 (1998-1999)

Afin de corriger pour les membres longitudinaux qui n'ont pas répondu au cycle 3, l'ajustement qui suit est apporté aux poids des résidents répondants :

$$a_2 = \frac{\text{somme des poids des résidents répondants et non - répondants}}{\text{somme des poids des résidents répondants}}$$

#### Ajustement 3 : Ajustement pour la non-réponse au cycle 4 (2000-2001)

Afin de corriger pour les membres longitudinaux qui n'ont pas répondu au cycle 4, l'ajustement qui suit est apporté aux poids des résidents répondants :

$$a_3 = \frac{\text{somme des poids des résidents répondants et non - répondants}}{\text{somme des poids des résidents répondants}}$$

#### Ajustement 4 : Ajustement pour la non-réponse au cycle 5 (2002-2003)

Afin de corriger pour les membres longitudinaux qui n'ont pas répondu au cycle 5, l'ajustement qui suit est apporté aux poids des résidents répondants :

$$a_4 = \frac{\text{somme des poids des résidents répondants et non - répondants}}{\text{somme des poids des résidents répondants}}$$

Ces quatre ajustements sont appliqués pour chaque patron de réponse longitudinal (voir section 8.1.2 pour la définition du patron de réponse longitudinal).

#### Ajustement 5 : Poststratification

Puisque le nombre total de personnes au Canada vivant en établissements de soins de santé est inconnu (selon la définition d'établissement de l'ENSP), il est impossible d'effectuer une poststratification en se basant sur ces totaux. Cependant, une poststratification est faite en utilisant les poids longitudinaux du sous-ensemble carré (WTI4LS).

$$a_5 = \frac{\text{somme des poids WTI4LS}}{\text{somme des poids des résidents répondants}}$$

Cette poststratification est faite en deux étapes : premièrement à l'échelle de cinq régions, puis pour chaque catégoric d'établissement/âge-scxe.

**Le poids final pour le sous-ensemble longitudinal complet du cycle 2 (WTI6LF)**, est calculé en prenant d'abord le poids longitudinal carré (WTI4LS) et en le multipliant ensuite par les ajustements 1 et 5 :

$$WTI6LF = WTI4LS * a_1 * a_5.$$

**Le poids final pour le sous-ensemble longitudinal complet du cycle 3 (WTI8LF)**, est calculé en prenant d'abord le poids longitudinal carré (WTI4LS) et en le multipliant ensuite par les ajustements 1, 2 et 5 :

$$WTI8LF = WTI4LS * a_1 * a_2 * a_5.$$

**Le poids final pour le sous-ensemble longitudinal complet du cycle 4 (WTI0LF)**, est calculé en prenant d'abord le poids longitudinal carré (WTI4LS) et en le multipliant ensuite par les ajustements 1, 2, 3 et 5 :

$$WTI0LF = WTI4LS * a_1 * a_2 * a_3 * a_5.$$

**Le poids final pour le sous-ensemble longitudinal complet du cycle 5 (WTI2LF)**, est calculé en prenant d'abord le poids longitudinal carré (WTI4LS) et en le multipliant ensuite par les ajustements 1 à 5 :

$$WTI2LF = WTI4LS * a_1 * a_2 * a_3 * a_4 * a_5.$$

### 10.3.2 Pondération pour les sous-ensembles complet partagé des cycles 2 à 5

Le sous-ensemble complet partagé inclut les membres du panel qui ont accepté de partager l'information recueillie lors de toutes les interviews menées dans le cadre de l'ENSP avec les ministères provinciaux de la santé et Santé Canada et ont fourni une réponse au cycle n (où n = 2, 2, 3, 4 ou 5). Étant donné que ces partenaires reçoivent seulement les enregistrements de ceux qui ont accepté de partager leur information, un poids ajusté doit être créé de telle sorte que les estimations calculées à partir de ce sous-ensemble représentent correctement la population totale en établissement de soins de santé.

Un seul ajustement simple est apporté au poids longitudinal complet pour créer le poids du sous-ensemble complet partagé à chaque cycle. Cet ajustement est donné par :

$$a_{sn} = \frac{\text{somme des poids des résidents répondants au cycle } nx}{\text{somme des poids des résidents répondants au cycle } n \text{ et ayant accepté de partager}}$$

Cet ajustement est appliqué pour chaque strate catégoric, patron de réponse longitudinal et âge-scxe possible, où n = 2, 3, 4 ou 5.

Il est à noter que quelques profils de réponse longitudinaux ont été regroupés afin de produire des ajustements plus stables. Le regroupement a été fait pour quelques strates catégoric/âge-scxe qui n'avaient que quelques observations dans certains profils de réponse. Dans chaque cas, le patron de réponse problématique a été regroupé avec un autre patron de la même strate catégoric/âge-scxe de telle sorte que la somme des poids corresponde toujours, sauf exception, aux totaux de la

population. Le regroupement a aussi été fait en respectant le plus possible le patron de réponse du cycle  $n$ . Par exemple, si le cas problématique avait un profil de réponse « décédé » au cycle  $x$ , il a été regroupé avec un patron de réponse longitudinal qui indiquait également un profil de réponse « décédé » au cycle  $n$ . Pour quelques cas, un tel regroupement n'a pu être fait alors, plutôt que de regrouper des patrons de réponse, on a regroupé des groupes d'âge-sexe. En général, puisque cet ajustement est fait en respectant les classes utilisées lors de l'ajustement pour la poststratification, aucune poststratification supplémentaire n'est nécessaire.

Le poids final pour le sous-ensemble complet partagé du cycle 2, WT16SLF, est donc obtenu en multipliant le poids longitudinal complet, WT16LF, par l'ajustement ( $a_{s2}$ ). De la même façon pour les autres cycles, le poids final pour le sous-ensemble complet partagé du cycle 3, WT18SLF, est obtenu en multipliant le poids longitudinal complet, WT18LF, par l'ajustement ( $a_{s3}$ ), le poids final pour le sous-ensemble complet partagé du cycle 4, WT10SLF, est obtenu en multipliant le poids longitudinal complet, WT10LF, par l'ajustement ( $a_{s4}$ ) et finalement le poids final pour le sous-ensemble complet partagé du cycle 5, WT12SLF, est obtenu en multipliant le poids longitudinal complet, WT12LF, par l'ajustement ( $a_{s5}$ ).

## **11. Qualité des données**

L'évaluation de la qualité des données est un aspect important à toute enquête. Elle permet de vérifier plusieurs éléments dont en particulier l'exactitude ou la fiabilité de l'information recueillie. De plus, il est possible de s'en servir pour améliorer la qualité de la prochaine réalisation de l'enquête. À Statistique Canada, cette évaluation de la qualité doit être conçue de manière à respecter les exigences obligatoires et minimales de la politique visant à informer les utilisateurs de la qualité des données et de la méthodologie (Statistique Canada, 1998b). Ces exigences minimales incluent la production de mesures de l'erreur de couverture, de taux de réponse puis de mesures de l'erreur due à l'échantillonnage. Cette section couvre uniquement les mesures relatives à l'erreur due à l'échantillonnage.

### **11.1 Erreur due à l'échantillonnage**

Comme toute enquête par échantillon, le volet établissements de santé de l'ENSP est sujette à l'erreur due à l'échantillonnage. Cette erreur dépend de plusieurs facteurs tels que les tailles d'échantillon et de population, le plan de sondage, puis de la variabilité de la population. La mesure de cette erreur dépend aussi de la méthode d'estimation utilisée. Toute enquête de Statistique Canada se doit de fournir à ses utilisateurs de données, une façon de connaître l'ampleur de cette erreur due à l'échantillonnage et ainsi avoir une meilleure idée de la précision des estimations produites. Le calcul de l'estimation demande un peu de travail et la technique de calcul utilisée a changé au cours des cycles de l'enquête. La prochaine sous-section donne plus de détails à ce sujet.

#### **11.1.1 Calcul de l'estimation de l'erreur due à l'échantillonnage**

Au cours des cycles du volet établissements de soins de santé de l'ENSP, deux méthodes ont été utilisées pour estimer la variance; une formule simple et bien connue, puis le bootstrap. Le bootstrap est une méthode d'estimation de la variance dite de rééchantillonnage. Carlson (1998) fait un survol intéressant de la problématique ainsi que des différentes méthodes de rééchantillonnage possibles. De plus, il examine différents logiciels conçus pour l'analyse de données d'enquête, et plus spécifiquement leurs façons de traiter les données provenant d'un plan d'enquête complexe.

L'utilisation d'une formule simple au cours des deux premiers cycles de l'enquête est discutée dans la sous-section 11.1.1.1. Suit ensuite une sous-section sur la méthode du bootstrap.

### 11.1.1.1 Une formule simple

Aux cycles 1 et 2, une formule simple et bien connue est utilisée pour calculer la variance et le coefficient de variation (CV) des estimations. Cette formule s'appuie sur l'hypothèse selon laquelle les établissements sont sélectionnés avec probabilité inégale et avec remise. En réalité, les établissements ont été sélectionnés sans remise; autrement dit, une fois sélectionné, un établissement ne pouvait être choisi de nouveau. Cette mesure a été prise pour des raisons d'ordre opérationnel plutôt que pour améliorer la variance, si bien que l'hypothèse d'un échantillon avec remise devrait avoir un effet négligeable.

Un programme de calcul de la variance, rédigé pour les systèmes SAS<sup>®</sup> et SPSS<sup>®</sup>, est fourni avec les fichiers de microdonnées des cycles 1 et 2. Les deux programmes (SAS<sup>®</sup> et SPSS<sup>®</sup>) qui accompagnent les fichiers de microdonnées permettent à l'utilisateur de calculer des totaux et des rapports sans trop de difficulté. Les formules utilisées pour calculer la variance d'un total Y ou d'un rapport  $R=Y/X$  sont :

$$V(\hat{Y}_{total}) = \sum_{h=1}^N \frac{\sum_{i=1}^{n_h} (\hat{Y}_{h,i} - \hat{Y}_h)^2}{n_h(n_h - 1)} \quad \text{et} \quad V(\hat{R}) = \frac{1}{\hat{X}^2} \sum_{h=1}^N \frac{\sum_{i=1}^{n_h} (\hat{Y}_{h,i} - \hat{R}_h \hat{X}_{h,i})^2}{n_h(n_h - 1)}$$

où :

$\hat{Y}_h$  est l'estimation, pour la strate h, de la variable de réponse Y basée sur tous les répondants dans la strate h;

$\hat{Y}_{h,i}$  est l'estimation, pour la strate h, de la variable de réponse Y basée sur l'ensemble des répondants de l'établissement i de la strate h;

N est le nombre de strates;

$n_h$  est le nombre d'établissements échantillonnés dans la strate h;

$\hat{X}_h$  est l'estimation, pour la strate h, de la variable X du dénominateur du rapport basée sur l'ensemble des répondants dans la strate h;

$\hat{X}_{h,i}$  est l'estimation, pour la strate h, de la variable X du dénominateur du rapport basée sur l'ensemble des répondants de l'établissement i de la strate h;

$\hat{X}$  est l'estimation globale de la variable X du dénominateur du rapport; et

$\hat{R}_h$  est le rapport  $\hat{Y}_h / \hat{X}_h$ .

### 11.1.1.2 Poids bootstrap

À partir du cycle 3, le volet établissements de soins de santé de l'ENSP a utilisé la méthode du bootstrap; une méthode qui présente des avantages intéressants (voir Rao et Wu (1988)) pour une description du bootstrap dans un contexte général d'enquête à plan complexe). Une description de la méthode de même qu'un examen de ses principaux avantages versus d'autres méthodes sont donnés dans l'article de Yeo, Mantel et Liu (1999).

En bref, la méthode du bootstrap consiste à rééchantillonner un grand nombre de fois l'échantillon total pour créer un ensemble de répliques. Chaque réplique représente un sous-échantillon où les poids de sondage de ce sous-échantillon sont ajustés de façon à ce que celui-ci représente la population totale. La réplique est en fait représentée dans le fichier final sous forme d'une variable de poids, communément appelé le *poids bootstrap*. L'estimation de la variance d'un certain paramètre peut alors être obtenue en calculant la simple variance des estimations de ce paramètre obtenue avec chaque poids bootstrap généré.

Du côté technique, le nombre de répliques à utiliser a fait l'objet de plusieurs travaux de recherche. Yeo, Mantel et Liu (1999) observent à la suite de simulations sur les données du volet ménages de l'ENSP que la variance se stabilise lorsqu'on utilise au-delà de 200 répliques pour des estimations de totaux et de rapports, et fait de même autour de 400 répliques pour des régressions.

S'inspirant de ces résultats, l'ENSP a fixé à 500 le nombre de répliques à distribuer avec chaque poids de sondage calculé pour le volet ménages. Pour le volet établissements de soins de santé, étant donné que la taille d'échantillon de ce volet est beaucoup plus petite que celle du volet ménages, 2 000 répliques bootstrap ont été produites à chacun des cycles,

L'estimation de la variance est donc faite selon la méthode du bootstrap à partir du cycle 3 (puis rétrospectivement appliquée aux cycles 1 et 2). Un fichier contenant les poids bootstrap (indépendant de celui contenant les données) est créé pour chaque poids d'échantillonnage. Les fichiers de poids bootstrap sont distribués à tous les utilisateurs, à l'exception des utilisateurs de FMGD.

Un programme est aussi fourni aux utilisateurs pour calculer l'estimation de la variance. Ce programme, appelé *Bootvar*, est composé de plusieurs macros auxquels l'utilisateur n'a qu'à faire appel en spécifiant le nom des variables pour lesquelles il désire calculer la variance. Malgré qu'il ne puisse répondre à l'utilisation de toutes les méthodes d'analyses possibles, *Bootvar* supporte le calcul de la variance pour une vaste gamme de statistiques : totaux, rapports, différences de rapports, régressions (linéaire et logistique), et modèles linéaires généralisés. Une documentation complète (dans les deux langues officielles) accompagne le tout et fournit des exemples d'utilisation du programme. *Bootvar* a d'abord été développé en SAS<sup>®</sup> mais une version bêta du *Bootvar* a aussi été développée en SPSS<sup>®</sup> et n'a été rendue disponible que sur demande lors des trois premiers cycles de l'ENSP. Une version améliorée du *Bootvar* (version 2.0 en SAS<sup>®</sup> et SPSS<sup>®</sup>) a éventuellement été produite et diffusée à partir du cycle 4 de l'ENSP. Cette version est beaucoup plus conviviale et s'utilise avec les données de tous les cycles de l'enquête. Pour plus de détails sur *Bootvar*, veuillez consulter les sites web suivants :

SAS: [http://www.statcan.ca/francais/rdc/bootvar\\_sas\\_f.htm](http://www.statcan.ca/francais/rdc/bootvar_sas_f.htm) et

SPSS: [http://www.statcan.ca/francais/rdc/bootvar\\_spss\\_f.htm](http://www.statcan.ca/francais/rdc/bootvar_spss_f.htm).

## **11.2 Erreur non due à l'échantillonnage**

L'erreur non due à l'échantillonnage est reliée à un grand nombre d'étapes d'une enquête, telles que la base de sondage, le questionnaire, la collecte, le traitement, etc. Bref, presque toutes les étapes sont sujettes à introduire une forme d'erreur dans les données. Malgré de bonnes procédures de contrôle à chaque étape de l'enquête, il demeure impossible d'éliminer complètement l'erreur non due à l'échantillonnage. En plus d'être inévitable, cette erreur est souvent difficile à mesurer. Le volet établissements de soins de santé de l'ENSP n'échappe pas à cette difficulté. Le volet établissements de soins de santé de l'ENSP produit quelques indicateurs pour quantifier cette erreur comme par exemple des taux de réponse. Les taux de réponse sont présentés à la section 8.

## 12. Confidentialité et fichiers de microdonnées à grande diffusion

La confidentialité est un aspect important à Statistique Canada que ce soit la confidentialité des fichiers de données diffusés, ou encore des tabulations publiées ou produites sous forme de demandes spéciales. L'ENSP investit beaucoup d'énergie et de travail pour respecter la confidentialité.

Pour chacun de ses cinq cycles, le volet établissements de soins de santé de l'ENSP a produit différents fichiers maîtres : les fichiers transversaux (aux cycles 1 et 2), puis une gamme de fichiers longitudinaux. Afin de rendre ces données disponibles au public, des FMGD transversaux ont été créés aux cycles 1 et 2. Seuls des fichiers FMGD transversaux ont pu être diffusés; la création d'un fichier longitudinal public respectant les règles de confidentialité n'étant pas possible. En effet, une étude de faisabilité a montré que la diffusion d'un FMGD longitudinal conjointe à celle des FMGD transversaux risquerait de révéler de l'information additionnelle pouvant compromettre la confidentialité des données. L'appariement du FMGD longitudinal avec un des FMGD transversaux fournirait aux utilisateurs un niveau plus approfondi de détails que celui permis sur ces fichiers publics. Une réduction des variables sur le FMGD longitudinal pourrait réduire les risques d'appariement, toutefois cette réduction devrait être tellement imposante qu'elle rendrait le fichier sans valeur aux fins d'analyse. Béland (1999) discute plus en détail des problèmes entourant la diffusion d'un FMGD longitudinal dans le cadre de l'ENSP.

La création d'un FMGD nécessite plusieurs analyses et une certaine stratégie s'impose. Cette stratégie doit en fait être présentée et approuvée au préalable par le Comité de diffusion de microdonnées de Statistique Canada. Une fois la stratégie présentée, les analyses examinant les risques de divulgation sont entreprises, permettant d'identifier les mesures à prendre pour protéger l'anonymat des répondants à l'enquête, et finalement mener à la création d'un FMGD final. Avant d'en faire sa diffusion, les résultats des analyses faites doivent toutefois être présentés et approuvés par le Comité afin de s'assurer que la politique sur la diffusion de microdonnées soit respectée.

Les analyses menées pour créer, valider et s'assurer de la confidentialité du FMGD sont nombreuses. Le fichier maître est le point de départ puisqu'il contient toutes les variables recueillies durant l'enquête, de même que celles dérivées de celles-ci. À noter que les identificateurs directs tels que le nom, l'adresse et le numéro de santé du répondant ne figurent pas sur le FMGD.

Le premier aspect habituellement examiné est le détail de géographie pouvant être rapporté sur le FMGD. L'idée est de s'assurer d'avoir un échantillon de taille considérable pour chaque région géographique rapportée, en tenant compte évidemment de la taille de population de cette région. Pour le volet établissements de soins de santé l'ENSP, aucune variable géographique n'a été retenue étant donné que l'échantillon était représentatif à l'échelle nationale seulement.

Chacune des variables est ensuite examinée en relation avec plusieurs critères tels que sa capacité analytique, son niveau de sensibilité et de détail. Selon le cas, la variable sera soit éliminée, catégorisée à nouveau ou encore laissée intacte.

On portera aussi un intérêt particulier aux variables considérées sensibles ou discriminantes; on réfère souvent à celles-ci comme étant des identificateurs indirects, c'est-à-dire des variables ne pouvant pas identifier un répondant sans avoir une certaine connaissance de celui-ci. Pour ces variables, on s'assurera qu'elles ne permettent pas d'identifier uniquement un répondant à elles seules, ou en combinaison avec d'autres variables sensibles.

La variable représentant le poids de sondage fait aussi face à un examen détaillé. On vérifiera entre autres que l'ordre de grandeur du poids de sondage ne permet pas de recréer avec certitude des régions géographiques non rapportées dans le fichier public.

Finalement, étant donné que l'échantillon transversal du volet établissements de soins de santé de l'ENSP a été en grande partie composé du panel longitudinal à chacun ces deux premiers cycles, on a dû s'assurer que les FMGD créés à chaque cycle ne pouvaient être appariés de façon à recréer des enregistrements longitudinaux. De tels appariements augmenteraient considérablement l'information disponible pour une personne donnée, et mettraient à risque la confidentialité de son information recueillie.

Ceci illustre donc de façon succincte le travail nécessaire à la création d'un FMGD. Les détails des analyses ne peuvent évidemment pas être rapportés afin de préserver totalement la confidentialité des données publiées. Encore une fois, Béland (1999) fournit plus de détails concernant le travail entourant la création des FMGD pour l'ENSP.

### 13. Diffusion des données

L'ENSP, tout comme plusieurs enquêtes de Statistique Canada, doit relever le défi de diffuser autant de données que possible tout en respectant le caractère confidentiel des renseignements.

Pour des questions de confidentialité, l'accès aux fichiers maîtres est restreint aux employés de Statistique Canada. Toutefois, conjointement avec certaines universités du pays, Statistique Canada a ouvert des Centres de données de recherche (CDR), donnant ainsi l'accès aux fichiers maîtres aux chercheurs intéressés. Un chercheur doit préalablement soumettre une proposition de recherche, qui est minutieusement étudiée par un comité, avant d'avoir accès aux données en tant qu'employé réputé de Statistique Canada. Pour plus de détails sur le programme Centres de données de recherche, veuillez consulter le site web [http://www.statecan.ca/francais/rdc/index\\_f.htm](http://www.statecan.ca/francais/rdc/index_f.htm).

Un autre service offert, uniquement pour les fichiers transversaux du volet établissements de soins de santé de l'ENSP, est le service à la clientèle à frais recouvrables. Les utilisateurs peuvent soumettre des demandes spéciales comme par exemple la production de différents tableaux. Le service à la clientèle effectue alors le travail. Après vérification, les résultats considérés non confidentiels sont ensuite envoyés aux utilisateurs.

Bien que les fichiers maîtres soient accessibles seulement à un groupe restreint de personnes, et sous des conditions relativement strictes, les FMGD sont quant à eux accessibles à tous, moyennant un certain coût. Il faut toutefois noter que la clientèle académique universitaire a accès facilement et sans frais aux FMGD via l'Initiative de démocratisation des données; consulter le site web [http://www.statcan.ca/francais/Dli/dli\\_f.htm](http://www.statcan.ca/francais/Dli/dli_f.htm) pour plus de détails sur le sujet.

Finalement, les fichiers de partage représentent un autre moyen de fournir l'accès aux données à une clientèle très spécifique. Ces fichiers de partage contiennent uniquement les enregistrements des répondants ayant accepté de partager leurs données avec les partenaires de l'enquête. Les fichiers sont donc destinés aux ministères provinciaux de la santé, Santé Canada et Ressources humaines et Développement social Canada (pour les cycles 1 et 3 seulement).

## **Remerciements**

L'auteur tient tout d'abord à remercier tous les méthodologistes qui ont travaillé de près ou de loin à la conception de l'ENSP. L'auteur de ce document ne figurait pas parmi les pionniers de l'ENSP. Tout au long de la rédaction de ce document, l'auteur s'est grandement inspiré des travaux et documentation laissés par ses prédécesseurs. L'auteur tient donc à souligner le travail et l'aide apportée par l'équipe initiale de l'ENSP. Elle tient également à remercier France Bilocq, Johane Dufour, Patrice Mathieu, Janet Pantalone et Michelle Simard pour leurs précieux commentaires qui ont permis d'améliorer la qualité de ce rapport.

## Bibliographie

ANGOSS Software (1995). Knowledge Seeker IV for Windows – User's Guide. ANGOSS Software International Limited.

Béland, Y. (1999). Release of Public Use Microdata Files for NPHS? Mission ... partially accomplished! American Statistical Association – Proceedings of the Survey Research Methods Section.

Carlson, B.L. (1998). Software for Sample Survey Data, *Encyclopedia of Biostatistics*, Volume 5, John Wiley & Sons, New York. 4160-4167 (voir aussi article sur Internet à [http://www.fas.harvard.edu/~stats/survey-soft/blc\\_cob.html](http://www.fas.harvard.edu/~stats/survey-soft/blc_cob.html))

Catlin, G. et Will, P. (1992). Enquête nationale sur la santé de la population : premiers faits saillants. Rapport sur la santé, Vol. 4, No. 3, Catalogue 82-003.

Cochran, W. G. (1977). *Sampling Techniques*. 3rd Ed. John Wiley & Sons, New York.

Dolson, D., McClean, K., Morin, J.-P. et Théberge, A. (1987). Plan d'échantillonnage pour l'Enquête sur la santé et les limitations d'activités. *Techniques d'enquête*, Vol.13, No.1, 101-117.

Mohl, C. (1994). National Population Health Survey, Institutional Sample. Statistique Canada, Division des méthodes d'enquêtes auprès des ménages, Document interne.

Parsons, G. (1993). National Population Health Surveys. Sampling from an Institutional Component. Statistique Canada, Division des méthodes d'enquêtes auprès des ménages, Document interne.

Rao, J.N.K., et Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.

Santé et Bien-être social Canada (1993). Rapport technique : Enquête promotion santé Canada. Publié sous la direction de T. Stephens et G.D. Fowler, Ottawa, Ministère des approvisionnement et services Canada.

Särndal, C.-E., Sewenon B. et Wretman, J. (1991). *Model Assisted Survey Sampling*. Springer-Verlag.

Statistique Canada (1994). Enquête nationale sur la santé de la population 1994-1995 : volet établissements de soins de santé - Fichier de microdonnées à grande diffusion – Documentation, Statistique Canada, No 82M0010XCB au catalogue.

Statistique Canada (1996). Enquête nationale sur la santé de la population 1995-1997 : volet établissements de soins de santé - Fichier de microdonnées à grande diffusion – Documentation, Statistique Canada, No 82M0010GPF au catalogue.

Statistique Canada (1998). Enquête nationale sur la santé de la population, Volet établissements de soins de santé, Cycle 3 (1998-1999) : Documentation longitudinale accompagnant l'ensemble de données longitudinales, Février 2002.

Statistique Canada (1998b). Statistique Canada – Lignes directrices concernant la qualité : Troisième édition – Octobre 1998. Statistique Canada, N° 12-539-XIF au catalogue.

Statistique Canada (2000). Enquête nationale sur la santé de la population, Volet établissements de soins de santé, Cycle 4 (2000-2001) : Documentation longitudinale accompagnant l'ensemble de données longitudinales, Avril 2004.

Statistique Canada (2002). [http://www.statcan.ca/francais/concepts/nphs/nphs2\\_f.htm](http://www.statcan.ca/francais/concepts/nphs/nphs2_f.htm).

Statistique Canada (2002a). Enquête nationale sur la santé de la population, Volet établissements de soins de santé, Cycle 5 (2002-2003) : Documentation longitudinale accompagnant l'ensemble de données longitudinales, Janvier 2007.

Brisebois, F., Dufour, J., Kelly, D., Lavigne, M., Mathieu, P., et Tolusso, S.(2005). Trois premiers cycles de l'enquête nationale sur la santé de la population, Volet ménages, Document de travail, DMEM -2005-017F/E.

Statistique Canada (2006). Enquête nationale sur la santé de la population, Volet ménages, Cycle 6 (2004-2005) : Documentation longitudinale accompagnant l'ensemble de données longitudinales, Novembre 2006.

Statistique Canada (2007). Enquête nationale sur la santé de la population, Volet établissements de soins de santé, Cycles 1 à 5 (1994-1995 à 2002-2003). Documentation des variables dérivées, Janvier 2007.

Tambay, J.-L., Goldmann, G. et White, P. (2001). Providing Greater Access to Survey Data for Analysis at Statistics Canada. American Statistical Association – 2001 Proceedings of the Section on Survey Research Methodology.

Tortora, R.D. (1978). A Note on Sample Size Estimation for Multinomial Populations, American Statistician, Vol. 32, No.3.

Yeo, D. (1999). After the First Steps : The Evolution of a Longitudinal Survey, Statistique Canada, paper presented at the Workshop on Longitudinal Research in Social Science – A Canadian Focus, Population Studies Centre, University of Western Ontario.

Yeo, D., Mantel, H. et Liu, T.P. (1999). Bootstrap Variance Estimation for the National Population Health Survey, 1999 Proceedings of the Survey Research Methods Section, American Statistical Association, pp. 778-783.

STATISTICS CANADA LIBRARY  
BIBLIOTHÈQUE STATISTIQUE CANADA



1010449361

*C3*  
*2005*