Project on Matching Census 1986 Database and Manitoba Health Care Files: Private Households Component

by Christian Houle,[1] Jean-Marie Berthelot,[1] Pierre David, [2] Cam Mustard,[3] D.Sc., Roos L,[3] PhD, M.C. Wolfson,[4] PhD

No. 91

11F0019MPE No. 91

ISBN: 0-660-16422-1

March 1996

Aussi disponible en français

[1] Health Analysis and Modelling Group, Statistics Canada, Ottawa, K1A 0T6
[2] Social Survey Methods Division, Statistics Canada, Ottawa, K1A 0T6
[3] Manitoba Centre for Health Policy and Evaluation, Department of Community Health Sciences, Faculty of Medicine, University of Manitoba, Winnipeg, Manitoba, R2H 2A6
[4] Institutions and Social Statistics Branch, Statistics Canada, Ottawa, K1A 0T6.

ABSTRACT

Introduction: In the current economic context, all partners in health care delivery systems, be they public or private, are obliged to identify the factors that influence the utilization of health care services. To improve our understanding of the phenomena that underlie these relationships, Statistics Canada and the Manitoba Centre for Health Policy and Evaluation have just set up a new database. For a representative sample of the population of the province of Manitoba, cross-sectional microdata on individuals' health and socio-economic characteristics were linked with detailed longitudinal data on utilization of health care services.

Data and methods: The 1986-87 Health and Activity Limitation Survey, the 1986 Census and the files of Manitoba Health were matched (without using names or addresses) by means of the CANLINK software. In the pilot project, 20,000 units were selected from the Census according to modern sampling techniques. Before the files were matched, consultations were held and an agreement was signed by all parties in order to establish a framework for protecting privacy and preserving the confidentiality of the data.

Results: A matching rate of 74% was obtained for private households. A quality evaluation based on the comparisons of names and addresses over a small subsample established that the overall concordance rate among matched pairs was 95.5%. The match rates and concordance rates varied according to age and household composition. Estimates produced from the sample accurately reflected the socio-demographic profile, mortality, hospitalization rate, health care costs and consumption of health care by Manitoba residents.

Discussion: The matching rate of 74% was satisfactory in comparison with the response rates reported in most population surveys. Because of the excellent concordance rate and the accuracy of the estimates obtained from the sample, this database will provide an adequate basis for studying the association between socio-demographic characteristics, health and health care utilization in province of Manitoba.

KEY WORDS : Probabilistic matching; confidentiality; sampling; longitudinal data; health; socio-economic status.

0.0 INTRODUCTION

A number of studies have clearly shown that there is a link between an individual's socio-economic status and the probability of his or her death during a given period [1,2,3]. Other studies have shown that the prevalence of certain diseases varies greatly depending on the socio-economic characteristics of the area in which an individual resides [4,5,6]. In addition, several Canadian surveys have already provided cross-sectional data on individuals' health status and socio-economic status, along with self-reported information on the use of health services (General Social Survey of 1991[7], Ontario Health Survey of 1990 [8], Enquête Santé Québec of 1987 and 1992-93 [9], Health and Activity Limitation Survey of 1986 and 1991[10], Canadian Health and Disability Survey of 1983-84 [11], Canada Health Survey of 1978-79 [12]). However, to our knowledge, there is no Canadian longitudinal database that combines exhaustive information on health, use of health services and socio-economic characteristics. In an effort to meet this information need, Statistics Canada and the Manitoba Centre for Health Policy and Evaluation (MCHPE) jointly set up a pilot project to evaluate the possibility of creating such a database using existing data.

The primary objective of this pilot project was to evaluate the feasibility of combining the following three data sources: the 1986 Census of Population, the 1986-1987 Health and Activity Limitation Survey (HALS) , and the Manitoba Health (MH) longitudinal file on health care service utilization. The database resulting from this combination is intended to enable researchers to explore new directions with respect to health determinants. In this article, we describe the matching of files, the selection of the sample for analysis purposes and the results, which show the representativeness of the database created and also validate the techniques employed.

0.1 Confidentiality and right to privacy

When creating a database from both administrative and survey data, it is essential to ensure the confidentiality of the data and prevent any unwarranted intrusion into individuals' privacy. In accordance with the policies of the agencies that collaborated on this project, certain procedures were undertaken prior to matching these data sets. They include consultations with the Privacy Commissioner of Canada, the Faculty Committee on the Use of Human Subjects in Research at the University of Manitoba, and Statistics Canada's Confidentiality and Legislation Committee. In addition, Manitoba Health's Committee on Access and Confidentiality was informed of the project.

Following these consultations, and in accordance with the formal policies of Statistics Canada, the Minister responsible for Statistics Canada authorized the matching as proposed: this was to be a pilot project for evaluating the feasibility and utility of the data matching; individuals' names and addresses would not be used for matching purposes, nor would they appear in the database; the matching would be done entirely on the premises of Statistics Canada by persons sworn in under the Statistics Act; only a sample of 20,000 matched units would be used for purposes of research and analysis; and access to the final data would be strictly controlled in accordance with the provisions of the Statistics Act. In addition, all activities with the linked data set are covered by a memorandum of understanding among Statistics Canada, the University of Manitoba and the Manitoba Ministry of Health.

1.0 DATA

The detailed questionnaire (questionnaire 2B) of the 1986 Census of Population contains detailed socio-economic information including variables such as family composition, dwelling characteristics, tenure, ethnic origin and mother tongue, as well as a number of variables relating to income and educational attainment [13,14]. This questionnaire was filled by persons residing in Manitoba on June 3, 1986 in a proportion of approximately one household in five. The other households completed a short form designed solely for enumerating the population. Thus the file used for matching purposes consisted of 261,861 records. The individuals represented by these records lived in two types of dwellings: private or collective. Examples of collective dwellings are hospitals, hospices, nursing homes, institutions for the physically handicapped, orphanages, psychiatric institutions, hotels/motels, work camps, jails, Hutterite colonies, military residences, religious institutions, student residences and YMCAs. In 1986 there were 26,161 persons in Manitoba living in a collective dwelling according to the Census. This article focuses primarily on the private household component.

The 1986 HALS was a postcensal survey that sought to identify individuals who because of their health were limited by the type or amount of daily activities that they could perform. By postcensal, we mean that a question from the Census (in this case, Question 20 on disabilities) served to enrich the survey sample by identifying a high proportion of the target population. An appropriate questionnaire was then completed for each person sampled. For HALS, the Manitoba population living in private households and having disabilities was studied on the basis of a sample of 5,480 persons representing a population of 150,857 persons having at least one disability. The data set thus created contained information on individuals' health and functional limitations as well as on type of employment, educational level, transportation, housing and recreation. Since the survey was of the self-reporting type, the data represent the situation of respondents from their viewpoint rather than from an administrative or clinical viewpoint.

The MH longitudinal file, for its part, contains information on visits to physicians, stays in hospital, diagnoses, surgical procedures, admission to personal care (nursing) home, health care received at home, the date and cause of death, and other data on health care utilization. A number of innovative studies in health care research have used this file [15,16,17]. For this pilot project, a register of persons covered by Manitoba health insurance was identified from June 1986, using the date of commencement of health insurance coverage and the date of cancellation of coverage. The register contained 1,047,443 records.

2.0 METHODS

The matching project was divided into three main stages. The first stage consisted of pairing individuals belonging to three distinct data sources. The second stage consisted of assessing what proportions of the pairs formed represented the same individual. The third stage consisted of selecting a sample of 20,000 matched units in order to create the database for analysis purposes. In this section, we shall deal with the methodology used in each of these stages in turn.

2.1 Matching

The Canlink system [18,19] developed at Statistics Canada was used for the pairing stage. Canlink is a probabilistic matching software that pairs records from two sets of data by using the discriminatory power of the common variables available. The system weights the pairs of records according to the degree of concordance of the values observed and also takes account of the probability of random concordances. The files paired were that of the 2B sample from the 1986 Census covering the province of Manitoba and the file of persons registered with MH in June 1986, containing only a subset of the variables available. Only these two files were involved in the probabilistic matching, since the 1986-1987 HALS sample was drawn from the Census 2B sample [20], all the HALS records[5] were already paired to those in the Census by a single key.

The individual records which were paired thus come from two files, one containing the records of 261,861 individuals living in Manitoba (derived from the 2B file of the 1986 Census), and the other containing the records of 1,047,443 persons (a derivative of the Manitoba Health file). The strategy adopted for identifying pairs representing the same individual (good pairs) consisted in dividing up the two data sets into blocks and forming only pairs of individuals belonging to the same block.

The pairs of records were compared only if all the blocking variables already agreed. It was therefore necessary to choose carefully so as not to eliminate at this early stage a great number of good pairs. It will be recalled that the most discriminant variables, namely surnames, given names and addresses, were not used in this study. Because of this constraint, we were forced to choose other combinations of variables that had more limited discriminatory power and to back up these choices by applying innovative techniques.

Two matching phases were carried out. First, after examining various possible definitions, we defined a block as a set of four individual characteristics, namely a person's sex, year of birth, month of birth and postal code. In the second matching phase, the definition was relaxed in order to form more pairs of individuals. The exact year and month of birth were replaced by the person's age, which made it possible to compare an individual with a greater number of candidates. In addition, the area covered by the geographic variable in urban settings was expanded by a factor of approximately three, with the postal code being replaced by the census enumeration area.

Through these matchings the census file was divided into three subsets: records which had clearly matched (definites), those which had matched but for which the discriminatory power of the available variables raised a doubt according to the CANLINK criteria (possibles), and those which had never matched.

[5] Except for collective households of the institutional type.

While the information on family structure was used in the matching process, the CANLINK system compares only two individuals at a time, without taking account of matches obtained for other family members. We had to define a series of rules in order to ensure the consistency of matchings within a given family and between two matching phases. A detailed account of this stage of the methodology has already appeared in a publication [21].

2.2 Evaluation of the concordance of the pairs formed

In our evaluation we pursued two objectives. First, it was important to determine the degree of accuracy with which we had associated the Manitoba Health data with the Census data (definite matches only). Then it was necessary to assess whether the rules that had been developed for rejecting certain "possible" matches were adequate.

A sample of 1,000 families was drawn, representing 2,102 matched individuals. As stratification variables, we used urban/rural area as determined by Census, family composition (person living alone, couple with child, couple without child, multiple family) and matching status (definite or possible). MH then proceeded to extract the names and addresses of all these individuals and their family members. It should be understood that this identifying information was not used to determine the validity of specific matches, but only to estimate actual matching rates at aggregated levels. These names and addresses were compared manually with those on the microfilmed 2B questionnaires kept at Statistics Canada.

Only 17 cases could not be compared, either because the microfilm was illegible or because the name and address had not been found in the MH file. The weight of these units was redistributed within the strata in which the nonresponse was identified.

2.3 Sample selection

Since the project involved three databases, it should be specified that our sampling frame was the 2B file from the 1986 Census. Even before a sample design was developed, there were a number of structural constraints to be dealt with. It will be recalled that the sample size was already set at a maximum of 20,000 units and that the database created had to combine information from the census files and the MH file, as well as information from the HALS file. The HALS sample used the individual as its sampling unit, whereas our analysis of the overall population of Manitoba used the household as defined by the Census. Several options were accordingly considered in order to try to construct a single database, but the complexity of the analysis that would have resulted practically negated any gains in accuracy. To ensure that there was a balance between simplicity of analysis and an effective design, it was decided that the selection process would consist of constructing two independent databases: the first to study the link between disability, socio-economic status and health and the second to analyse the general population of Manitoba. To maximize the use of the 20,000 units, it was also necessary to take account of the overlap between these two databases.

Owing to the complex sample design of HALS, the relatively small number of individuals sampled for HALS and the importance of this database from an analytical standpoint, matched individuals with disabilities were <u>all</u> selected. These accounted for 4,434 basic units. This sample formed the first database, used for the analysis of persons with at least one functional disability. The weighting used in the analysis would be the HALS weighting, adjusted to take account of the matching rate specific to certain subgroups.

There were therefore 15,566 units left to form the general population database plus the expected number of units overlapping the two databases. Still pursuing the objective of optimizing the sample design, we evaluated that stratification was appropriate. The reasoning here was that stratification is an especially effective tool for the purposes of three aspects of a sample survey. First, it serves to reduce the overall variance of the estimates. Second, it ensures a standard of quality for estimates relating to subgroups of interest in the population. Third, stratification can result in improved accuracy in certain cases in which non-sampling error can be taken into account. Lastly, stratification is especially effective when the stratification variables are correlated with the target variable.

Since a number of studies have established links between socio-economic status and health, it was natural to use socio-economic variables to construct the strata. In addition, there was no disadvantage to using the household as the sampling unit, since socio-economic status is generally the same for all members of a given household.

It should be kept in mind that the sample had to represent the household population of Manitoba in general. Since it was the 1986 Census file that entirely determines the composition of this population, all the stratification variables[6] were either taken directly from that file or derived from it. The following list describes how the stratification cells were formed:

1. Type of enumeration area;
2. Family structure of household;
3. Household with at least one yes to Question 20;
4. Age group of person representing household;
5. Sex of person representing household;
6. Urban or rural area;
7. Educational tercile of person representing household;
8. Income tercile adjusted for household structure.

To construct some of these variables, it was necessary to define a person representing the household. Where there were children, considering the correlation between the age of the mother and the ages of the children, it seemed to us to be logical to choose the oldest adult female in the family to represent it. Otherwise we chose the oldest adult. The choice of these variables represents a judicious balance between making matching rates uniform within a given stratum and using socio-economic variables.

While no oversampling was done in advance, it was then necessary to make several groupings in order to meet the criteria of minimal size of sampling units per stratum. The final number of strata for private households was 611. The total number of units drawn was 16,387. These represented 46,670 persons.

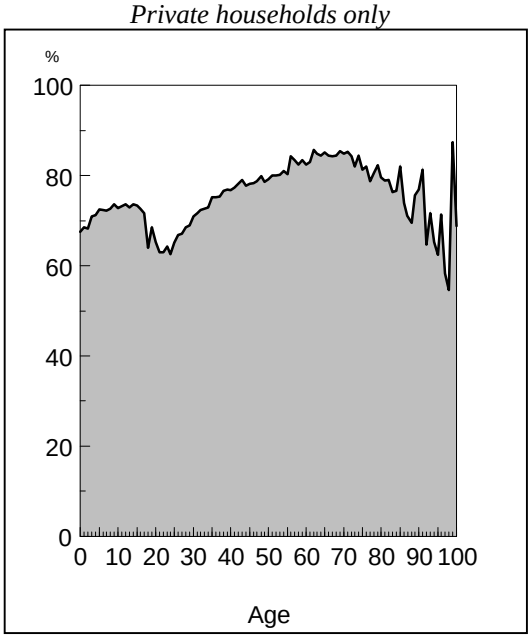
Lastly, it is common practice to adjust sampling weights so that the totals estimated by the sample will reflect as accurately as possible the counts of the population studied. With post-stratification, the counts can be adjusted for categories for which the number of units was insufficient to create a real stratum but which were of sufficient analytical importance to justify the use of special techniques. These techniques changed the initial weight subject to the constraints of minimum change [32]. For private households, the counts by age group, rural or urban geographic area, marital status and sex were used to adjust the weights at the level of individuals, while rural or urban geographic area, household size and tenure played the same role for adjusting at the level of households.

[6] These variables are described in detail in the appendix.

3.0 RESULTS

3.1 Matching

Figure 1. Match rate by age
Private households only



Despite all the restrictions and exclusions applied to the initial matches, overall, 74% (174,476 out of 235,700) of individuals from the census living in a private household were matched with an individual in the Manitoba file. This rate varied according to geographic mobility, age, marital status and family size.

The factors that had the greatest influence on the match rate were all related to individuals' geographic mobility. Hence the following groups of individuals were more difficult to match: young adults (between 20 and 25 years of age: Figure 1), persons who had changed their place of residence between the 1981 and 1986 censuses (Table 1), and divorced or separated persons (Table 2). Among these groups, frequent changes of address and family structure made concordance between the two data sources more difficult than among less mobile groups. The reasons for this is that since the Census figures date from June 3, 1986 precisely, and some MH data are dated December 31, 1986, there was more likely to be an information lag with respect to mobile individuals.

Table 1. Match rate according to mobility
Private households only

Mobility	Match rate %
Same household	81.7
Same CD	65.8
Other CD	62.5

CD : Census Division, a geographic unit used by the Census. Manitoba is made up of twenty-three Census Divisions.

Table 2. Match rate according to marital status
Private households only

Marital status	Match rate %
Married	78.5
Widowed	74.5
Single	71.2
Divorced	61.4
Separated	43.4

As may be observed, there was a low match rate among separated persons. This is explained by the mobility inherent in the phenomenon of separation and also by the information lag between the two data sources.

The effect of age on the match rate was not surprising. Children under fifteen years of age and adults between thirty and sixty years of age had better rates, owing to their more stable situation. Among individuals over 85 years of age, there was greater variability in the data because of the phenomenon of institutionalization and the small number of cases.

Among individuals who did not move between the 1981 and 1986 censuses (same household), one might have expected an even better match rate. The rate of 81.7% is perhaps an indication that using the methodology described thus far, there is a ceiling of around 80% on matches, given that the files are not totally free of errors.

Intuitively, family size is correlated in two opposite ways with the match rate. While a large family has an intrinsic constraint on the mobility of the family nucleus, on the other hand some members of the family will periodically attach themselves to this nucleus or leave it. And indeed, as Table 3 shows, the match rate dropped off significantly as family size increased.

Table 3. Match rate according to family size
Private households only

Family size	Match rate %
1	66.4
2	78.5
3	74.7
4	79.8
5	77.1
6	70.0
7	55.9
8	51.4
9	39.2
10+	46.7

3.2 Evaluation of concordance of pairs formed

Table 4 shows that overall, more than 95% of the definite matches retained represented the same individual. Since the sample of 20,000 units was drawn from definite matches only, this means that the matching was of exceptional quality. Also, the fact that the rate of concordance of names among possible pairs was only 40% indicates that our strict rules were justified and prudent, since they prevented us from accepting a large proportion of bad matches.

Household size was closely related to the concordance rate. Persons living alone and those living in households of eight or more persons exhibited a lower concordance rate, namely 86.8% and 90.8% respectively. These results would seem to be due to the small number of discriminant variables available

for persons living alone and the fact that in the case of large households there was often more than one family within the household.

Table 4. Rate of concordance of names according to various groupings
Private households only

	Concordance on names %	Standard error* %
"Possible" match"	40.1	2.4
"Definite" match"	95.5	0.5
Indian reserves	95.3	1.5
Household of 1 person	86.8	2.5
Household of 2 to 7 persons	96.5	0.5
Household of 8 or more persons	90.8	2.0

** The design effect is ignored in calculating the standard error*

A final point to be observed is that matched inhabitants of Indian reserves[7] had a concordance rate equivalent to that of persons living off reserve.

3.3 Sample selection

Overall, we were pursuing two specific objectives in designing the stratification. First, it was necessary to come as close as possible to having a self-weighting design so as to allow for the use of existing computer tools. The costs of specific tools and the time required to develop them would have been a major handicap for any subsequent analysis. This objective was attained by avoiding oversampling of strata to the extent possible and by maintaining a certain uniformity of weights within each stratum formed.

Our second objective was to use socio-economic variables in the process of forming strata. Therefore, when stratum sizes permitted, we used variables derived from income, education level, family structure, age and geography.

Since we were creating a new database drawing on administrative files from various sources, we compared the estimates produced by our sample with the estimates from the original sources. Tables 5 and 6 show how accurately we managed to reproduce the 1986 2B census counts, after post-stratification.

It should be noted that before post-stratification, four categories were far enough from the target to justify immediate action in and of themselves. These were: males and females who were divorced (difference of +7.8%), separated (difference of -17.6%) or widowed (difference of +6.1%) as well as females 0 to 4 years of age (difference of -4.6%). These categories represented a small number of individuals each, and it was not possible to treat them separately in the stratification at the time of sample selection. As a result of post-stratification, all the differences were reduced to under 1.5%.

[7] It should, however, be kept in mind that the match rate on Indian reserves was only 44.5%, considerably lower than the average rate of 74%. This could lead to a bias, since the matched individuals may have had very different characteristics from the reserve population as a whole.

Table 5. Accuracy of the sample by age group

Private households only

	Males Census 2B	Difference from sample %	Females Census 2B	Difference from sample %	Total Census 2B	Difference from sample %
0 to 4 years	39,929	-0.05	37,709	-0.06	77,638	-0.05
5 to 14 years	78,855	-0.13	76,074	-0.15	154,929	-0.14
15 to 24 years	84,897	-0.20	84,250	-0.12	169,147	-0.16
25 to 44 years	155,160	-0.11	157,430	-0.05	312,590	-0.08
45 to 64 years	93,145	-0.06	98,851	-0.05	191,996	-0.05
65 years and +	52,829	-0.01	68,309	-0.02	121,138	-0.02
Total	504,815	-0.10	522,623	-0.07	1,027,438	-0.09

Table 6. Accuracy of the sample by marital status

Private households only

	Males Census 2B	Difference from sample %	Females Census 2B	Difference from sample %	Total Census 2B	Difference from sample %
Previously married	26,649	-0.51	69,037	-0.10	95,686	-0.21
Married	245,443	-0.03	245,799	-0.03	491,242	-0.03
Single	232,723	-0.13	207,787	-0.12	440,510	-0.12
Total	504,815	-0.10	522,623	-0.07	1,027,438	-0.09
Divorced	9,146	-0.08	14,388	-0.15	23,534	-0.12
Separated	9,137	-1.40	11,242	-0.04	20,379	-0.65
Widowed	8,366	-0.01	43,407	-0.09	51,773	-0.08

Lastly there were major conceptual differences with respect to the definition of the populations represented by the Census and by the MH file. Only persons having a usual place of residence in Manitoba on June 3, 1986 were enumerated in that province. Census operations do not succeed in having all persons who were residents fill out a questionnaire, a situation that resulted in undercoverage. On the other hand, the MH file was made up of all persons who were covered by the health insurance plan. Some of these persons no longer lived in Manitoba or may not have indicated a change in their status, resulting in some overcoverage. The MH file contained no information on residents of several categories of collective dwellings for which medical services were provided by the federal government, such as military camps and some Indian reserves, whereas the Census considered these persons to be residents of Manitoba.

For purposes of comparison, we excluded persons living in nursing homes from the MH counts in the tables that follow, since they were then living in an institutional collective dwelling. Furthermore, according to the census definition, persons who had stayed for 180 days or more in a health care institution were considered institutionalized, and therefore we excluded them from the counts in the following tables. Despite these efforts to make the populations uniform, the fact remains that we managed only to approximate the counts of persons living in institutions. On the basis of complementarity, the populations compared represent approximately those persons living in a private household or a non-institutional collective household.

As Table 7 shows, despite major conceptual differences, the sizes of these two populations by age group were quite comparable. Overall, the estimated total sizes of the two populations differed by only 0.1%, although males were underestimated by 1.2% and females were overestimated by 1.4%. It may also be observed that the greatest differences were found among younger individuals.

Table 7. Accuracy of the sample by age group versus MH
Private and non-institutional collective households

	Males MH	Difference from sample %	Females MH	Difference from sample %	Total MH	Difference from sample %
0 to 4 years	32,743	2.57	31,105	1.98	63,848	2.28
5 to 14 years	78,076	1.47	73,912	2.87	151,988	2.15
15 to 24 years	86,722	-1.24	82,971	1.61	169,693	0.15
25 to 44 years	165,783	-2.84	159,458	1.92	325,241	-0.51
45 to 64 years	96,989	-1.71	98,997	0.92	195,986	-0.38
65 years and +	57,904	-1.34	74,129	-1.10	132,033	-1.20
Total	518,217	-1.20	520,572	1.39	1,038,789	0.10

Tables 8, 9 and 10 compare the mortality rate, medical care utilization and hospital care utilization according to whether they were estimated from our sample or from the MH file. It should be noted that the death rates reported in the literature (in particular [22]) were slightly higher than those presented in Table 8, with the difference increasing with age. This may be explained by the fact that our files exclude individuals living an institutional collective dwellings, who exhibit a higher mortality rate than persons living in private households.

Table 8. Annualized mortality rates based on the period from June 1986 to May 1989.
Private and non-institutional collective households

Age	Annual mortality rate* (x 1,000) MH	Annual mortality rate* (x 1,000) Sample	95% confidence interval for the sample
0-4	0.51	0.02	(0 – 0.18)
5-9	0.16	0.20	(0 – 0.66)
10-14	0.24	0.27	(0 – 0.80)
15-19	0.78	0.60	(0 – 1.39)
20-24	0.94	0.70	(0 – 1.52)
25-29	0.83	0.40	(0 – 1.02)
30-34	0.86	0.44	(0 – 1.11)
35-39	1.12	1.07	(0 – 2.19)
40-44	1.85	2.07	(0.37 – 3.77)
45-49	3.23	2.78	(0.57 – 4.99)
50-54	5.12	3.68	(1.03 – 6.33)
55-59	8.42	8.91	(4.80 – 13.02)
60-64	12.43	10.48	(6.01 – 14.95)
65-69	18.96	19.03	(12.69 – 25.37)
70-74	28.63	24.59	(16.76 – 32.42)
75-79	42.77	43.09	(30.85 – 55.33)
80 and over	76.98	73.67	(57.41 – 89.93)
Total	6.71	6.11	(5.39 – 6.83)

* Number of estimated deaths (over the three years period) divided by estimated population total times 3.

Overall, the mortality rate estimated by our sample (6.11) was lower than the one derived from the MH file (6.71), but this difference was not statistically significant at a 95% confidence level. Note that the confidence intervals derived from our sample contained the value calculated by MH for all age groups except children aged 0 to 4. While the number of deaths in this category was relatively small, it appears that the difficulty matching children under one year of age may be related to this underestimate. This also indicates that any analysis specific to children from 0 to 4 years of age will have to be conducted with caution, especially where the prevalence of a disease or condition was low.

Table 9. Number and costs of medical services, 1986-87 fiscal year
Private and non-institutional collective households

	Number of services			Costs of services ($)		
Type of practice	MH	Sample	Relative difference (%)	MH	Sample	Relative difference (%)
Internal medicine	699,542	702,735	0.46	19,060,658	18,904,922	-0.82
Paediatrics	416,157	449,122	7.92	6,932,217	7,359,290	6.16
Psychiatry	154,279	146,704	-4.91	8,970,489	8,468,584	-5.60
Surgery	406,907	409,097	0.54	21,772,057	21,230,743	-2.49
Ophthalmology	339,334	357,273	5.29	10,017,371	10,506,461	4.88
Otorhinolaryngology	100,859	110,164	9.23	2,952,711	3,181,710	7.76
Dermatology	115,516	126,086	9.15	1,955,985	2,114,772	8.12
Radiology	623,712	653,850	4.83	9,564,330	9,970,191	4.24
Pathology	2,941,244	3,126,365	6.29	21,369,502	22,489,399	5.24
Obstetrics and Gynaecology	294,288	328,728	11.70	8,774,785	9,151,231	4.29
Anaesthesiology	74,624	74,442	-0.24	5,762,857	5,572,964	-3.30
General practice	4,762,316	4,858,641	2.02	75,806,649	76,545,594	0.97
Physical Medicine	9,325	8,333	-10.64	447,187	413,127	-7.62
Totals	10,938,103	11,351,540	3.78	193,386,798	195,908,988	1.30

For most categories of medical practice, the estimates drawn from the sample were fairly close to those presented by MH, both for the number of services and for the costs generated in providing these services. The accuracy achieved was all the more remarkable since no post-stratification was carried out at any level in order to adjust the consumption of health care services to the MH figure.

Table 10. Number and duration of hospital stays, 1986-87 fiscal year
Private and non-institutional collective households

	Number of stays			Duration of stays		
Age group	MH	Sample	Relative difference (%)	MH	Sample	Relative difference (%)
0-64	100,127	96,303	-3.82	538,616	499,665	-7.23
65 and over	43,226	41,318	-4.41	452,172	414,555	-8.32
Total	143,353	137,621	-4.00	990,788	914,220	-7.73

Table 10 shows the results of the comparison of the number and duration of hospital stays. Once again taking the conceptual differences between the two data sources into account, it is satisfactory that the accuracy of the estimates was within 10%. A larger underestimate for the duration of stays than for the number of stays indicates that longer stays were more underestimated than short stays. This situation may be explained by the difficulty in identifying residents of institutional collective dwellings on the MH files.

4.0 DISCUSSION

With the methodology presented in this article, approximately 74% of the part of the census file corresponding to private households could be matched with the MH file, using mainly age, sex, postal code, family size and family structure. This rate of 74% is satisfactory when compared to the response rate reported in a number of surveys. For example, the response rates for the Nova Scotia Nutrition Survey were 79.7% among located respondents and 60.0% for the total sample[23]. The Manitoba Heart Health Survey registered response rates of 77.1% among located respondents and 60.8% for the total sample [24].

Obviously, considering the various types of errors that can occur in matching on a large scale, it is not realistic to expect a matching rate of 100%. It is inevitable that the success rate of any probabilistic matching operation will be affected by erroneous data, lags in the collection or updating of the information, as well as conceptual differences between the data sets to be matched. Furthermore, while non-matched individuals exhibited relatively different characteristics from matched individuals, very rich socio-demographic information concerning this non-matched population was available from the Census. This information was used to select a sample of matches representative of the entire population.

In 95.5% of cases, the pairs formed did indeed associate the data on an individual's health care utilization with the socio-economic data on the same individual in the 1986 Census. This rate of accuracy is exceptional, considering that surnames, given names and birth dates were not used in the matching process.

The accuracy obtained in estimating various indicators associated with the consumption of health care (such as mortality, number and costs of medical services, number and duration of hospital stays) justifies the care with which the matching and sampling methods were developed.

In light of the match rates obtained, the rates of concordance of names and the accuracy of the estimates, it can be said not only that the new database created is unique in Canada but also that the quality of the data coded in it greatly exceeds that of many surveys based on interviews.

At a time when health expenditures exceed 10% of the GDP in Canada and 13 % in the United States [25], substantial efforts are being made to identify the relationships between health care utilization and health itself. While it is suspected that the level of health perceived by the patient explains a sizeable portion of consumption, many studies have focused on consumption by a specific client group, such as the elderly, or on consumption of health care in the years prior to death [16,26,27,28].

Along these lines, the newly constructed microdatabase opens the door to various studies that have never been undertaken in Canada. For example, one of the projects proposed by the MCHPE consists of analysing morbidity according to the individual's occupation by examining the extent to which the health care utilization for a class of illnesses is related to the basic occupational group. The census data can be used to classify individuals according to the occupation that they have reported, or according to whether or not they are employed and whether or not they are in the labour force. Using the 9th revision of the International Classification of Diseases (ICD-9), the medical conditions primarily studied will be musculo-skeletal disorders, cardio-vascular diseases, mental disorders, gastrointestinal illnesses and injuries.

From an even more general viewpoint, there are plans to study differences in the level of health care utilization according to socio-economic status at different stages in life. On the one hand, it is well-documented that the greatest consumption of health services occurs toward the end of one's life [16,27,28]. On the other hand, a major decline in infant mortality between 1960 and 1990 has also been observed [29a,

29b]. These two phenomena alone are justification for undertaking more thorough comparisons at all age levels. Using data on visits to doctors, health care at home and hospital admissions, it will be possible to compare health care utilization by different age groups according to socio-economic status for the classes of illnesses mentioned above. In addition, several studies [3,30,31] suggest that differences in health conditions according to socio-economic status are greater among persons between 35 and 64 years of age than for the other age groups. Analyses by age group obtained by using this new database could confirm these hypotheses or shed new light on these matters.

Even though the links between socio-economic status and health are the object of intensive research, one of the most frequently encountered problems in this type of research is that it is impossible to have precise information on socio-economic status at the individual level. Some researchers have no other choice but to use an indicator obtained through the aggregation of taxation or census data for an area of a given size, such as the census enumeration area or the postal code area [33]. Little research has been done to verify the impact and the validity of this methodology. This tends to reduce the capacity of such models to detect more subtle but theoretically quite important determinants. On the basis of a study of respiratory illnesses among Manitoba children aged 0 to 4, it should be possible to analyse the impact of the use of aggregated socio-economic data as compared to microdata. The reason for this is that the utilization of health care services for acute respiratory episodes exhibits major differences in behaviour according to socio-economic status. Children from poor urban backgrounds have a four times greater chance of being hospitalized for acute respiratory episodes than those from wealthier backgrounds (MCHPE). This type of episode and pneumonia together account for fully 25% of all hospital admissions without surgery for Manitoba children under 15 years of age. Similarly, the rates of coronary bypass surgery among elderly persons exhibit characteristics that also lend themselves to an analysis of the effect of using aggregated socio-economic data.

The HALS file, combined with administrative data from health care utilization records, opens the door to comparisons which until now have been difficult if not impossible to make. HALS offers us a clear and detailed image of individuals suffering from disability. Whether by age group or sex, by type of disability (mobility, sight, hearing, dexterity, cognition, etc.) or severity, the Manitoba population suffering from a disability can be compared to the general population by means of the census file. Specific analyses of these data will focus on mortality and on health care utilization.

ACKNOWLEDGEMENTS

The authors wish to thank the following persons for their significant and generous contribution to this study: Shelley Derksen, J. Patrick Nicol and Leonard McWilliam, Manitoba Centre for Health Policy and Evaluation.

BIBLIOGRAPHY

[1] WOLFSON, M.C., ROWE, G., GENTLEMAN, J.F., and TOMIAK, M. (1993). Career earnings and death: a longitudinal analysis of older Canadian men. *Journal of Gerontology: Social Sciences.*

[2] MARMOT, M.G. (1986). Social inequalities in mortality: the social environment. In *Class and Health, Research and Longitudinal Data*, (Ed. R.G. Wilkinson). London: Tavistock Publications.

[3] WILKINS, R., ADAMS, O., and BRANCKER, A. (1991). Changes in mortality by income in urban Canada from 1971 to 1986. *Health Reports*, 1(2), 137-174.

[4] ANDERSON, G., GRUMBACH, K., LUTT, H., ROOS, L.L., and MUSTARD, C. (1993). Use of coronary artery bypass surgery in the United States and Canada: influence of age and income. *Journal of the American Medical Association*, 269, 1661-1666.

[5] DOUGHERTY, G., PLESS, I.B., and WILKINS, R. (1990). Social class and the occurrence of traffic injuries and death in urban children. *Canadian Journal Of Public Health*, 81, 204-209.

[6] GENTLEMAN, J.F., WILKINS, R., NAIR, C., and BEAULIEU, S. (1991). An analysis of frequencies of surgical procedures in Canada. *Health Reports*, 3(4), 291-309.

[7] STATISTICS CANADA (1994). Health status of Canadians : Report of the 1991 General Social Survey, Cat. No. 11-612E, No. 8.

[8] ONTARIO MINISTRY OF HEALTH (1992). Ontario Health Survey, 1990, highlights and working papers, Toronto.

[9] Ministère de la Santé et des Services sociaux (1988). Et la santé ça va ? Rapport de l'enquête Santé Québec 1987, Tome I.

[10] STATISTICS CANADA (1988). The Health and Activity Limitation Survey, Selected data for Canada, provinces and territories, Cat. No. 41034.

[11] STATISTICS CANADA (1986). Report of the Canadian Health and Disability Survey 1983-1984, Cat. No. 82-555E

[12] HEALTH AND WELFARE CANADA(1981). The Health of Canadians: Report of the Canada Health Survey, Cat. No. 82-538E.

[13] STATISTICS CANADA (1986). *Census Handbook*, Cat. No. 99-104.

[14] STATISTICS CANADA (1986). *General review of the Census*, Cat.No. 99-137.

[15] ROOS, L.L., NICOL, J.P., and CAGEORGE, S.M. (1987). Using administrative data for longitudinal research: comparisons with primary data collection. *Journal of Chronic Diseases*, 40(1), 41-49.

[16] ROOS, N.P., MONTGOMERY, P., and ROOS, L.L. (1987). Health care utilization in the years prior to death. *The Milbank Quarterly*, 65(2), 231-254.

[17] SHAPIRO, E. and ROOS, L.L. (1984). Using health care: rural/urban differences among the Manitoba elderly. *The Gerontologist*, 24(3), 270-274.

[18] SMITH, MARTHA E., (1981). Generalized Iterative Record Linkage System, Health Division, Statistics Canada.

[19] FELLEGI, I.P. and SUNTER, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

[20] DOLSON, D., McCLEAN, K., MORIN, J.-P., and THÉBERGE, A. (1987). Sample Design for the Health and Activity Limitation Survey. *Survey Methodology*, 13(1), 93-108

[21] DAVID, P. et al. (1993). Linking Survey and Administrative Data to Study the Determinants of Health. *Proceedings of the American Statistical Association*, San Francisco

[22] STATISTICS CANADA (1994). *Causes of death 1992*, Cat. No. 84-208

[23] MACLEAN, D.R. (1993). Report of the Nova Scotia Nutrition Survey. Nova Scotia Heart Health Program, Department of Health, Government of Nova Scotia.

[24] YOUNG, T.K., GELSKEY, D.E., MACDONALD, S.M., HOOK, E., and HAMILTON, S (1991). The Manitoba heart health survey: technical report.

[25] HEALTH AND WELFARE CANADA (1993). Health expenditures in Canada - 1991, Policy, Planning and Information Branch.

[26] ROOS, N.P (1989). How a universal health care system responds to an aging population, *J. Aging. Health.*

[27] SHAPIRO, E. and ROOS, L.L. (1981). Preliminary Findings on Health Care Utilization by the Elderly. Medical Care, vol. 19, No 6.

[28] BARER ML et al. (1987). New evidence onold fallacies. *Soc. Sci. Med.* 24(10), 851-862.

[29a] PAPPAS G, QUEEN S, et al. (1993). The increasing disparity in mortality between socioeconomic groups in the United States, 1960 and 1986. *N. Engl. J. Med.* 329, 103-109.

[29b] MARMOT MG and McDOWALL ME. (1986). Mortality decline and widening social inequalities. *Lanctet*. I, 274-276.

[30] UGNAT AM ET Mark E (1987). Life expectancy by sex, age and income level. *Chronic Disease in Canada*.

[31] WILLIAMS DR (1990). Socio-economic differentials in health : a review and redirection. *Social Psychology Quarterly* 53(2):81-99.

[32] Kovacevic M. (1995). The weight adjustment for the sample from the 'whole population database' (Private Household component), Technical note, Statistics Canada, March 24, (not published).

[33] Geronimus AT, Bound J, Neidert LJ (1994). On the validity of using Census geocode characteristics to proxy socioeconomic status, Annual meetings of the Population Association of America.

[34] Wilkins R (1993). Use of postal codes and addresses in the analysis of health data, *Health Reports*, 5(2): 157-177, Cat. No. 82-003.

APPENDIX

To ensure a certain homogeneity in the weights that would be used in the analysis, it was necessary to create a variable called type of enumeration. It indicates whether Census Operations distributed 2B questionnaires to all households in the enumeration area (EA) or only to one household in five. The distinction is between "take all" and "take some" EAs. Furthermore, the importance of using this variable for stratification was brought home by a difference of more than 27% in the match rates of "take all" EAs (51.8%) and "take some" EAs (78.8%). It should be noted that "take all" EAs are mainly located in remote areas, which explains the poorer match rates.

The second variable constructed is called household structure It is defined on the basis of a hierarchical classification of the census families that constitute the household. This classification is based on the structure of the family according to the following categories: household consisting of more than one family with one of them containing children; household consisting of more than one family with none containing children; household consisting of one couple with at least one child; household consisting of a single-parent family; household consisting of only a couple; household consisting of a single person, never married; and lastly, household consisting of one person, previously married. As the match rate varied considerably according to household structure, this variable could not be ignored in the stratification process.

The third variable is derived from Question 20 of the 2B census questionnaire. That question invited respondents to state whether they were limited in their activities. If at least one member of the household had indicated a limitation, the household was categorized as "with limitation." Otherwise it was classified as "without limitation." Since we wanted to study the association between health care utilization and activity limitation, this variable was indispensable.

The fourth variable constructed was the age group representing the household. In order to maintain a consistent definition, it was necessary to designate the individual who would represent the household. First the family that defined the household structure was identified. If it was a family with children, the age of the oldest female was selected, since the age of the mother should be correlated with the number of children and their age. Where there were no females, the oldest male was selected. If the family identified had no children, the oldest adult was chosen. The age limits of the groups were 0 to 44, 45 to 64 and 65 and over. There were two reasons for using this variable. First, age is a major determinant of health care utilization, but also the match rate varies considerably from one age group to another.

The variable gender of the household was constructed solely for single-parent families and families consisting of only one person. The person who served to define the age group also designated the sex of the household. The use of this variable was justified by the need to compare single-parent families with two-parent families.

The geography variable was used to distinguish households in rural areas from those in urban areas, on the basis of the census definition.

The education variable is defined as the highest level of education attained by an adult in the family defining the household structure. Terciles were calculated from the cells obtained by cross-tabulating geographic variables and age group variables, and categories 1, 2 and 3 were derived.

Lastly, the low income measure was constructed on the basis of the total household income divided by the number of individuals adjusted for consumption. The first individual has a value of 1, the second a value of 0.37, the third a value of 0.36, the fourth a value of 0.26, and the fifth and each additional individual has a value of 0.18. As with the case of the preceding variable, it was terciles derived from the same cross-tabulations that were retained.