Misclassification of Income Quintiles Derived from Area-based Measures

A Comparison of Enumeration Area and Forward Sortation Area

Danielle A. Southern, MSc[1,2]
William A. Ghali, MD, MPH[1-3]
Peter D. Faris, PhD[1,2]
Colleen M. Norris, PhD[4]
P. Diane Galbraith, BN[2,3]
Michelle M. Graham, MD[5]
Merril L. Knudtson, MD[2]
for the APPROACH Investigators

ABSTRACT

Background: Census-based methods are often used to estimate socioeconomic status. We assessed the agreement between Forward Sortation Area (FSA) and Enumeration Area (EA) derived income levels for all patients undergoing cardiac catheterization in Alberta, Canada, from 1995-1998.

Methods: Income quintiles were calculated from census data for FSA and EA level. FSA- and EA-derived income measures were compared for misclassification. Both methods were then applied to the data to determine 4-year survival by income grouping in 21,446 patients following catheterization.

Results: The variability in EA-derived incomes for any given FSA-derived income is large. Only 40% of income quintiles are in agreement between the methods. For EA-based analyses, there is a linear relationship between higher income and lower mortality across all quintiles, while for FSA-based analyses, only the lowest income quintile had significantly higher mortality.

Discussion: Assuming that FSA-based methods are more likely to misclassify income compared to EA-based measures, the results for the FSA-based analyses are more likely to be erroneous. EA-derived measures should therefore be used when individual data are not available.

La traduction du résumé se trouve à la fin de l'article.

1. Department of Community Health Sciences, University of Calgary, Calgary, AB
2. Centre for Health and Policy Studies, University of Calgary
3. Department of Medicine, University of Calgary
4. Department of Public Health Sciences, University of Alberta, Edmonton, AB
5. Department of Medicine, University of Alberta

Correspondence and reprint requests: Dr. William Ghali, HSG239, 3330 Hospital Dr. NW, Calgary, AB T2N 4N1, Tel: 403-210-9317, Fax: 403-210-3818, E-mail: wghali@ucalgary.ca

Acknowledgements: APPROACH Clinical Steering Committee: Drs. Merril Knudtson (Principal Investigator), Vladimir Dzavik (Chairperson), Stephen Archer, Neil Brass, Michael Curtis, William Ghali, William Hui, Arvind Koshal, Andrew Maitland, Brent Mitchell, Duncan Saunders, and Ms. Colleen Norris. The APPROACH initiative is supported by generous donations from the Weston Foundation, Eli Lilly Canada, Searle Canada, Merck Frosst Canada, Johnson & Johnson and the Guidant Corporation. Financial and technical support received from the Canadian Cardiovascular Outcomes Research Team. Dr. Ghali is supported by a Population Health Investigator Award from the Alberta Heritage Foundation for Medical Research, and by a Government of Canada Research Chair in Health Services Research. Analyses for this paper were supported by an operating grant from Heart and Stroke Foundations of Canada. The authors would also like to thank Drahomir A. Aujesky for providing the translation.

In most administrative databases, researchers do not have access to individual-level socio-economic data. This often makes it necessary to use aggregated census data to fill in missing socio-economic status (SES) information,[1-22] despite the potential for an 'ecological fallacy'.[1,23]

In Canada, there are various levels of census geography and postal geography that researchers need to know about when linking to census data. Postal geography within most provinces begins with forward sortation areas (FSA), defined by the first three digits of the postal code, each with a median of about 20,000 households. FSAs are then further divided into local delivery units, each with a median of about 10 households in urban areas and roughly 500 households in rural areas. These are geographical units created by Canada Post Corporation to facilitate delivery of the mail. Census geography is based on enumeration areas (EA), each of which consists of about 300 households and are the smallest geographic unit for which census data are reported. EAs aggregate to census subdivisions (municipalities) and census division (country-level geographic units). Within the larger census agglomerations and census metropolitan areas, EAs may also be aggregated into census tracts (CT), each with about 8 EAs or roughly 2,400 households. Despite the widespread availability of EA data in Canada,[24,25] numerous studies still use large units like FSAs, CTs or census divisions.[2-7]

Using data on 21,446 cardiac catheterization patients from the Alberta Provincial Project for Outcome Assessment in Coronary Heart Disease (APPROACH), our objectives were 1) to assess the agreement between FSA- and EA-based income quintiles and look for evidence of misclassification between the two methods, and 2) to apply FSA- and EA-based income quintiles to analyses in our research team's ongoing research into survival after cardiac catheterization.

METHODS

Data sources

The study population used was from APPROACH, an inception cohort database that includes all patients in Alberta, Canada undergoing cardiac catheterization. Patients are followed longitudinally, with survival and time from enrollment

catheterization until death ascertained through semi-annual linkage to death records from the Alberta Bureau of Vital Statistics. We analyzed data on 21,812 patients enrolled in APPROACH for calendar years 1995-1998, with complete follow-up of patients through December 31, 1999.[26]

Information on individual income

Statistics Canada Census data from 1996 were used as the source of the median individual income for each FSA and EA.[27] Alberta has a total of 137 FSAs and 4,746 EAs. The Statistics Canada Postal Code Conversion File (PCCF)[28] for May 1999 postal codes contained all 767,381 Canadian postal codes ever used by Canada Post Corporation since 1983 (including many which are now retired). Each postal code in this file is linked to one or more EAs. When there is more than one EA for a postal code, Statistics Canada provides a single link indicator (SLI) to select the most representative EA. After merging the APPROACH data with the PCCF, using only the SLI, we merged the new file with the Census data files containing EA and FSA median individual income. Median individual income was then used to divide the Alberta population into quintiles by both FSA and EA method. Though the same method was used to define quintiles (i.e., percentile cuts of the population), the cut points were different for EA- and FSA-derived quintiles, because the underlying distribution of incomes differed across methods.

Comparison of FSA- and EA-based income levels

We first prepared a scatterplot of individual median income for EA-derived income (y-axis) plotted against FSA-derived income (x-axis). We then used a 5X5 table of EA income quintiles by FSA income quintiles to determine the misclassification of FSA. Sensitivity was calculated using EA-based quintiles as a 'criterion standard'. The kappa statistic was used to measure the degree of nonrandom agreement between FSA and EA.

Application of FSA- and EA-based income levels in survival analyses

We used survival analyses to produce Kaplan Meier plots for the proportion of those surviving over a 4-year period by both EA-based and FSA-based income measures. Univariate proportional hazards models including income quintiles (with the highest income quintile as the baseline category) were also used to derive hazard ratios for both methods for determining income levels.

TABLE I

Characteristics of APPROACH Population and FSA and EA-based Samples

Characteristics	Full Data (N=21,812)	Data Source: FSA-linked (N=21,731)	Data Source: EA-linked (N=21,664)	Data Source: Both EA/FSA linked (N=21,446)
Age (years)				
<65	11,803 (54.1)	11,751 (54.1)	11,726 (54.1)	11,591 (54.0)
65-74	6885 (31.6)	68,661 (31.6)	6834 (31.6)	6781 (31.6)
75+	3124 (14.3)	3114 (14.3)	3104 (14.3)	3074 (14.4)
Male	15,406 (70.6)	15,344 (70.6)	15,303 (70.6)	15,140 (70.6)
Rural	4506 (20.7)	4506 (20.7)	4381 (20.2)	4357 (20.3)

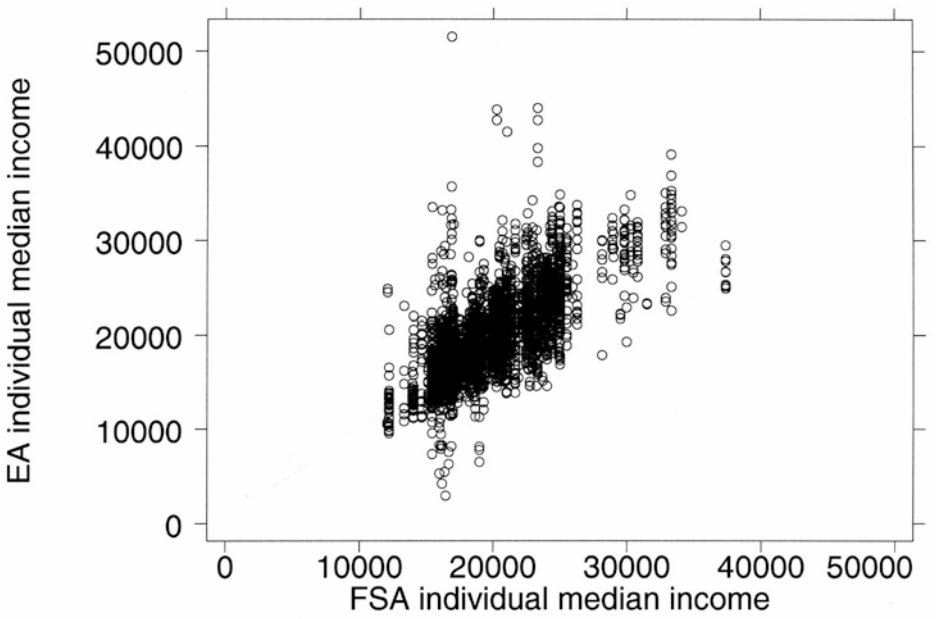

Figure 1. Scatterplot of EA-based individual median income by FSA-based individual median income.

RESULTS

Characteristics of study sample

Our starting point for analyses was an APPROACH 'analysis file' containing 21,812 patients that has been subjected to a data cleaning and management process described elsewhere.[29] Following linkage to EA- and FSA-based income measures, our dataset contained 21,664 patients with complete EA income measures and 21,731 patients with complete FSA income measures. For the analyses that follow, we used only data on 21,446 patients for whom both sources of income levels were available. The characteristics of the data sources defined by these linkages were essentially identical across datasets (Table I).

The ranges of median individual incomes derived from FSA and EA were $12,115-$37,396 and $2,919-$43,963, for FSAs and EAs respectively. For FSA-based quintiles, the medians of the lowest to the highest quintile were $16,190, $17,768, $20,279, $23,271 and $28,984, respectively. For EA-based quintiles, the medians of the lowest to highest quintiles were $14,870, $17,234, $19,433, $22,091 and $26,518 respectively.

Agreement between FSA- and EA-based income levels

Figure 1 is a scatterplot of individual median income for the EA by FSA methods. The data are not tightly clustered along a diagonal line; rather, there is significant variability in the EA-based median individual incomes for any given FSA-based median individual income. For example,

TABLE II

APPROACH Population 1995-1998 by FSA- and EA-based Income Quintiles

FSA Income Quintiles	EA Income Quintiles 1	2	3	4	5	Total
Lowest quintile	**2956**	1828	924	322	144	6174
Second lowest quintile	1069	**1581**	1390	453	234	4727
Middle quintile	334	780	**1182**	1243	462	4001
Fourth quintile	120	380	813	**1557**	1545	4415
Highest quintile	0	96	75	552	**1406**	2129
Total	4479	4665	4384	4127	3791	21,446
% Agreement with EA-based Income quintile (sensitivity)	66.0%	33.9%	27.0%	37.7%	37.1%	

Simple Kappa=0.25 (95% CI 0.24, 0.26)
Weighted Kappa=0.48 (95% CI 0.47, 0.49) - (the higher weighted kappa value accounts for partial agreement).

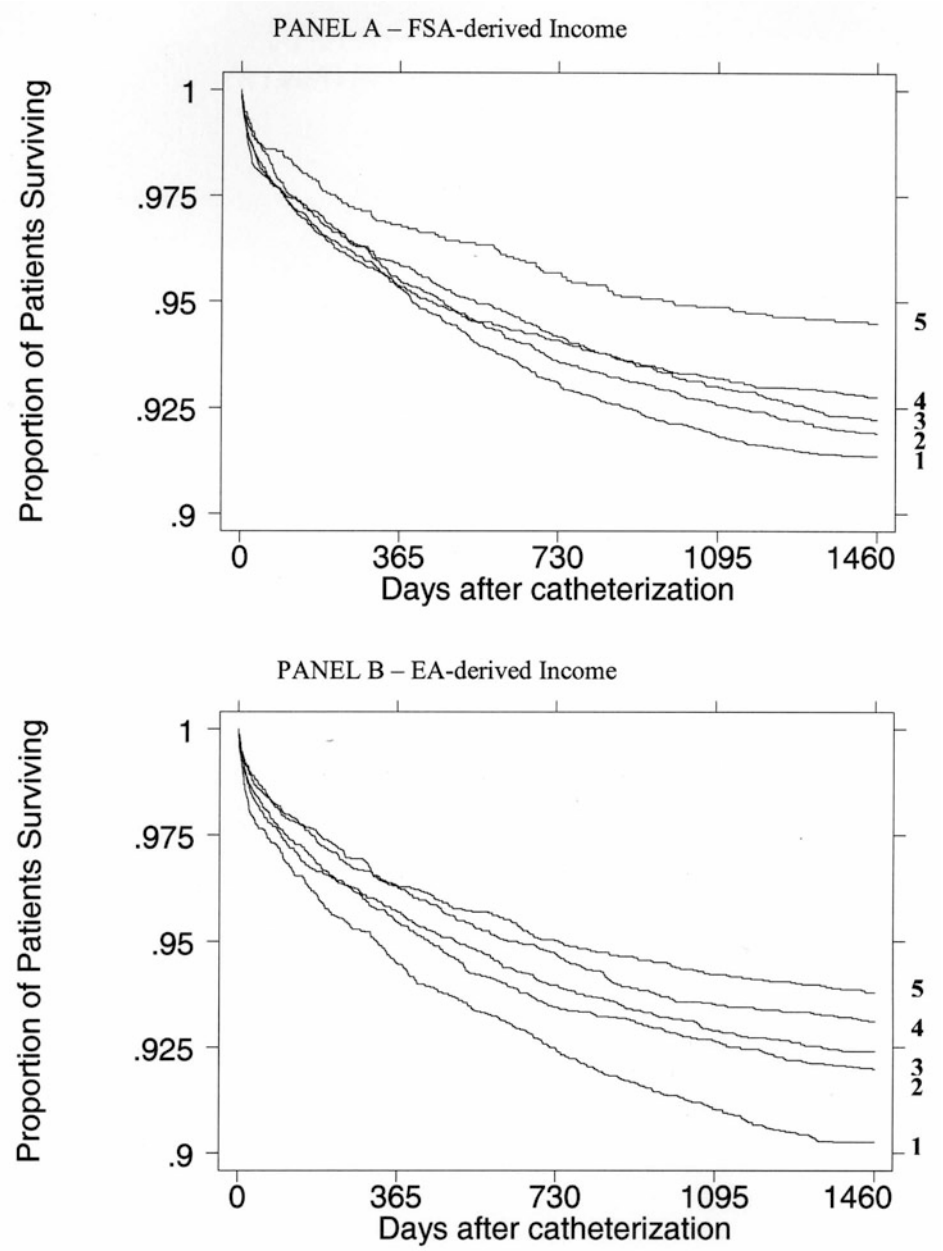

Figure 2. Kaplan-Meier plot of crude proportion surviving up to 4 years by income quintiles derived from FSA (Panel A) and EA (Panel B).

for an FSA income of $20,000, the EA-based median income ranges from approximately $15,000 to about $42,500, a range of over $25,000.

The cross-tabulation of EA by FSA-based income quintiles is shown in Table II. The assignments of income quintile were in agreement for only 40% of cases. Another 43% of assignments were within one level of the diagonal, 12.3% within two levels, 3.6% within 3 levels, and 0.7% within the maximum 4 levels of disagreement. The simple and weighted kappa values (the latter accounts for partial agreement) were 0.25 (0.24-0.26) and 0.48 (0.47-0.49), respectively.

Next, we assumed that EA was less prone to misclassification given the smaller size and therefore used EA as a 'criterion standard' in an analysis of the sensitivity of the FSA-based income quintiles. The results show generally low sensitivity for FSA-derived incomes (Table II).

Application to analyses of 4-year survival

The relationship of income quintiles derived from FSA versus EA to survival after cardiac catheterization differs across methodologies. Figure 2 presents crude Kaplan-Meier plots for survival extending to 4 years by income quintiles derived from FSA (Panel A) and EA (Panel B). Though the difference between the lowest and highest quintiles across methods is not very large, the middle income quintiles show a notable difference across methods, with a clustering of FSA-derived survival curves that does not occur for EA-derived curves.

Corresponding hazards ratios for death following cardiac catheterization from Cox proportional hazards analyses are shown in Figure 3. Notable hazard ratio differences were seen for the second, third and fourth income quintiles (1.48 vs. 1.32, 1.42 vs. 1.24 and 1.31 vs. 1.10, for FSA and EA respectively). Again these findings suggest a clustering of survival rates across the middle FSA quintiles that is not seen for the EA-derived quintiles, where a graded progression of risk is seen across quintiles.

DISCUSSION

When the EA-based quintile is taken as the gold standard, FSA methods misclassify

the income quintile for over 50% of the subjects of our study. We also see a remarkable difference across methods in analyses of survival to 4 years. The pattern of linear increases in survival across increasing EA-derived income quintiles is not seen when using FSA-based methods.

Many authors have argued that individual-level income data should be used whenever possible.[1,6,7,23] Although this is ideal, census-based aggregate measures will continue to be needed in health research since individual level data are often not available. Furthermore, even when available, individual income data can often be incomplete due to non-response that relates to the sensitivity of questions on incomes. Sin et al.[23] compared FSA to individual-based measures in Alberta, but these authors excluded seniors and natives from their analysis because they did not have income data on such individuals. Our analyses have permitted the study of a larger number of patients over 65 (45.9% of cases) and therefore the EA measures used may provide advantages over the approach used by Sin et al.

Optimizing methods of assessing census-derived socio-economic status remains an important research priority. Our work adds to the knowledge base on how to infer SES from Canadian sources. Our findings of probable misclassification of income quintiles by the FSA method complements an Australian study[10] comparing results of analyses based on 4-digit Australian postcode (approximately equivalent to Canadian FSAs) and collector's district (EA equivalent). The investigators found that postcode-based measures misclassified more than half the patients compared to the collector's district-based measures. In addition, the Australian postcode classification underestimated the relationship between SES and health-related measures. The authors conclude that all countries should use collector's district or equivalent small area units (i.e., EA in Canada).

Many have argued that small area statistics can be used to approximate SES.[10-12,22] EA is the smallest unit for which census data are collected and thus is likely to be more homogeneous than other units (i.e., FSA, CT). Statistics Canada updates the PCCF frequently so that even new postal codes may be matched with EAs. Postal codes are constantly added and retired, and FSAs are occasionally added, whereas census data are frozen for five years. When working with EA-based measures, the updated PCCF allows old or discontinued postal codes to match with the EAs. When working with FSA-based measures, newly created FSAs (fortunately a relatively rare occurrence in most parts of Canada) will not match with census income data and will thus have to be dropped from the analysis. This is why 21,731 individuals were available in our FSA analysis, whereas only 21,664 were available for EA analysis.

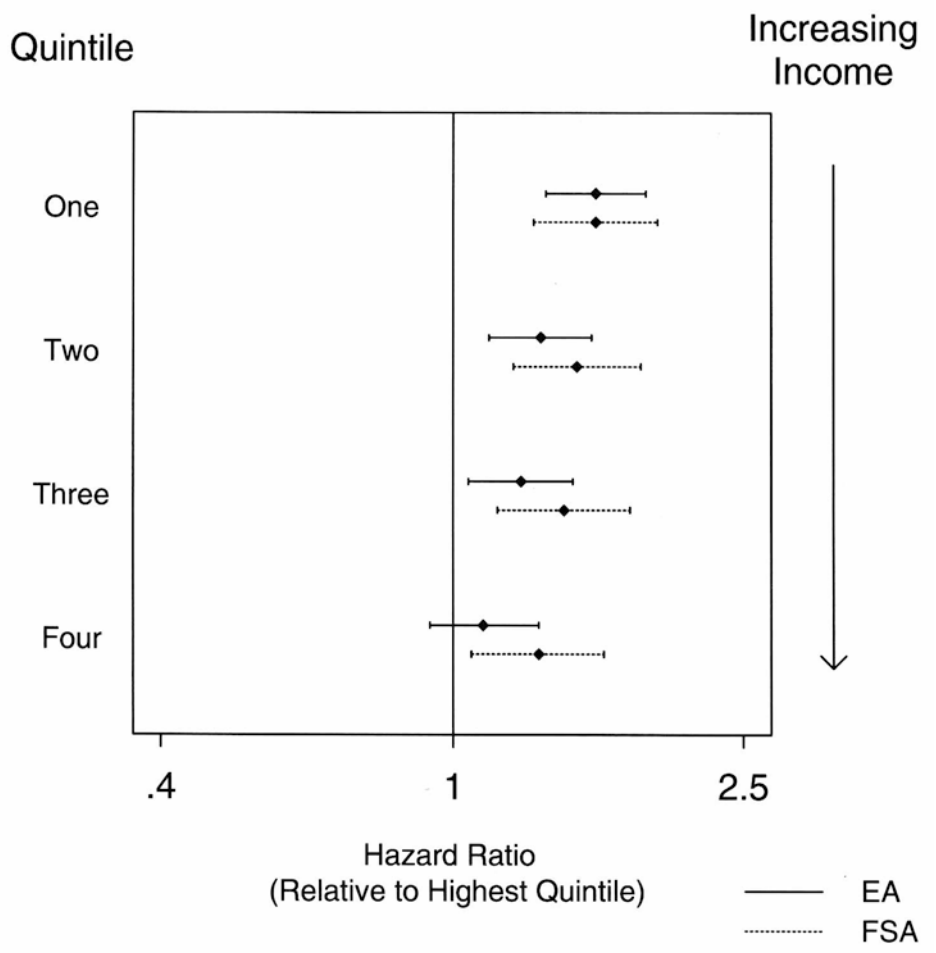

Figure 3. 4-year Crude Hazard Ratios (HR) by EA and FSA-based quintile incomes.

There still are times when FSA-based analyses can be justified, such as in studies making US-Canada comparisons where US data are only available for Zip Code rather than census tract or US block group level; or when the database includes a large population of institutionalized patients, a subgroup normally excluded by EA-based methods since income data would not be available for the institutional EAs.

Several studies in the literature examine the validity of using area-based measures relative to individual level data[1,22,23] and some compared area-based measures to smaller area-based measures (i.e., CT to EA, postcodes to EA).[6,7,10-12] These studies uniformly conclude that aggregate measures should not be used without acknowledging the potential biases that occur when estimates of SES are used regardless of the size of geographical units studied. Mustard et al. concluded that the use of ecologic-level measures of income is valid when individual-level data are not available.[22] Geronimus et al. found that there was very little gain from using smaller area-based measures.[7] In contrast, Soobader et al. argue that measures from smaller geographic units may produce results that are slightly less biased.[6] Krieger found that using CT and census block group (equivalent to EA) derived levels was a valid and useful approach to overcoming the absence of individual level data and also argued

that small area data can be used to construct population-based incidence and prevalence rates stratified by social class. The denominators for incidence or prevalence rates are census derived and therefore can be characterized by the same census-based social class measures as the numerator data.[11,12]

In summary, when EA-based income is taken as the criterion standard, the use of FSA-derived income may misclassify income quintiles for over half of the patients studied. The results from the method comparison of 4-year survival showed a large difference and suggest that the use of EA may reduce some of the potential bias introduced when not using individual level data for SES. We recommend that all Canadian researchers using area-based SES measures use EA rather than FSA to define their measures whenever possible.

REFERENCES

1. Demissie K, Hanley JA, Menzies D, Joseph L, Ernst P. Agreement in measuring socio-economic status: Area-based versus individual measures. *Chron Dis Can* 2000;21(1):1-7.
2. Glazier RH, Badley EM, Gilbert JE, Rothman L. The nature of increased hospital use in poor neighbourhoods: Findings from a Canadian inner city. *Can J Public Health* 2000;91(4):268-73.
3. Jin A, Hertzman C, Peck SHS, Lockitch G. Blood lead levels in children aged 24 to 36 months in Vancouver. *CMAJ* 1995;152(7):1077-86.
4. Spasoff RA, Gilkes DT. Up-to-date denominators: Evaluation of taxation family file for public health planning. *Can J Public Health* 1994;85(6):413-17.
5. Dougherty G, Pless B, Wilkins R. Social class and the occurrence of traffic injuries and deaths in urban children. *Can J Public Health* 1990;81(3):204-9.
6. Soobader M-J, LeClere FB, Hadden W, Maury B. Using aggregate geographic data to proxy individual socioeconomic status: Does size matter? *Am J Public Health* 2001;91(4):632-36.
7. Geronimus AT, Bound J. Use of census-based aggregate variables to proxy for socioeconomic group: Evidence from national samples. *Am J Epidemiol* 1998;148(5):475-86.
8. Hankins CA, Laberge C, Lapointe N, Lai Tung MT, Racine L, O'Shaughnessy M. HIV infection among Quebec women giving birth to live infants. *CMAJ* 1990;143(9):885-93.
9. Gentleman JF, Wilkins W, Nair C, Beaulieu S. An analysis of frequencies of surgical procedures in Canada. *Health Reports* 1991;3(4):291-309.
10. Hyndman JCG, Holman CDJ, Hockey RL, Donovan RJ, Corti B, Rivera J. Misclassification of social disadvantage based on geographical areas: Comparison of postcode and collector's district analyses. *Int J Epidemiol* 1995;24(1):165-76.
11. Krieger N. Overcoming the absence of socioeconomic data in medical records: Validation and application of a census-based methodology. *Am J Public Health* 1992;82(5):703-10.
12. Krieger N. Women and social class: A methodological study comparing individual, household, and census measures as predictors of black/white differences in reproductive history. *J Epidemiol Commun Health* 1991;45:35-42.
13. Locker D, Ford J. Using area-based measures of socioeconomic status in dental health services research. *J Public Health Dentistry* 1996;56(2):69-75.
14. Deonandan R, Campbell K, Ostbye T, Tummon I, Robertson J. A comparison of methods for measuring socio-economic status by occupation or postal area. *Chron Dis Can* 2000;21(3):114-18.
15. Mackillop WJ, Zhang-Salomons J, Groome PA. Associations between community income and cancer incidence in Canada and the United States. *Cancer* 2000;89(4):901-12.
16. Hartford K, Roos LL, Walld R. Regional variation in angiography, coronary artery bypass surgery, and percutaneous transluminal coronary angioplasty in Manitoba, 1987 to 1992: The funnel effect. *Medical Care* 1998;36(7):1022-32.
17. Jolly AM, Orr PH, Hammond G, Young TK. Risk factors for infection in women undergoing testing for chlamydia trachomatis and neisseria gonorrhoeae in Manitoba, Canada. *Sexually Transmitted Dis* 1995;22(5):289-95.
18. Anderson GM, Grumbach K, Luft HS, Roos LL. Use of coronary artery bypass surgery in the United States and Canada: Influence of age and income. *JAMA* 1993;269(13):1661-66.
19. Mackillop WJ, Zhang-Salomons J, Groome PA, Paszat L, Holowaty E. Socioeconomic status and cancer survival in Ontario. *J Clin Oncol* 1997;4(4):1680-89.
20. Boyd C, Zhang-Salomons JY, Groome PA, Mackillop WJ. Associations between community income and cancer survival in Ontario, Canada, and the United States. *J Clin Oncol* 1999;17(7):2244-55.
21. Alter DA, Naylor CD, Austin P, Tu J. Effects of socioeconomic status on access to invasive cardiac procedures and on mortality after acute myocardial infarction. *N Engl J Med* 1999;341(18):1359-67.
22. Mustard CA, Derksen S, Berthelot J-M, Wolfson M. Assessing ecologic proxies for household income: A comparison of household and neighbourhood level income measures in the study of population health status. *Health and Place* 1999;5:157-71.
23. Sin DD, Svenson LW, Paul Man SF. Do area-based markers of poverty accurately measure personal poverty? *Can J Public Health* 2001;92(3):184-87.
24. Statistics Canada. *1996 Census Dictionary – Final Edition*. Ottawa. 1996; Cat 92-351-UIE.
25. Statistics Canada. *Postal Code Conversion File, May 1999 Postal Codes (Reference Guide)*. Ottawa, 1999.
26. Ghali WA, Knudston ML, on behalf of the APPROACH Investigators. Overview of the Alberta Provincial Project for outcome assessment in coronary heart disease. *Can J Cardiol* 2000;16(10):1225-30.
27. Statistics Canada (Geography Division). Postal Code Conversion File, May 1999 version. Ottawa, 1999.
28. Statistics Canada. Census of Canada 1996. Alberta info [computer files].
29. Norris CM, Ghali WA, Knudtson ML, Naylor CD, Saunders LD. Dealing with missing data in observational health care outcome analyses. *J Clin Epidemiol* 2000;53:377-83.

Received: October 12, 2001
Accepted: May 21, 2002

RÉSUMÉ

Contexte : Les méthodes fondées sur le recensement sont souvent utilisées pour l'estimation du statut socio-économique. Nous avons examiné la concordance entre les revenus dérivés des régions de tri d'acheminement (RTA) et ceux dérivés des secteurs de dénombrement (SD) chez tous les patients ayant bénéficié d'un cathétérisme cardiaque en Alberta entre 1995 et 1998.

Méthodes : Nous avons comparé les mesures du revenu dérivées des RTA et des SD pour cerner d'éventuelles erreurs de classement. Ensuite, nous avons appliqué les deux méthodes aux données de 21 446 patients post-cathétérisme cardiaque afin de déterminer leur survie sur quatre ans.

Résultats : Pour toute valeur donnée des revenus dérivés des RTA, la variabilité des revenus dérivés des SD était grande, les revenus ne concordant d'une méthode à l'autre que dans 40 % des cas. Ceci a des conséquences majeures dans l'analyse de l'association entre la survie et les quintiles de revenus : l'estimation de survie du quintile de revenus moyen selon les deux méthodes présente des différences significatives.

Interprétation : Les méthodes dérivées des RTA classifient mal les mesures du revenu et mènent à des résultats incorrects dans l'analyse de survie. Pour cette raison, nous proposons d'utiliser les mesures dérivées des SD si les données individuelles ne sont pas disponibles.