Ecologic Proxies for Household Income

How Well Do They Work for the Analysis of Health and Health Care Utilization?

Murray M. Finkelstein, PhD, MDCM, CCFP

ABSTRACT

Background: Researchers often use census-derived measures of socioeconomic status (SES) when personal information is not available. Theory predicts that the resulting misclassification will blunt associations between outcomes and SES and that control for confounding by SES will be less effective. The purpose of this paper was to examine the magnitude of this problem using data from the National Population Health Survey (NPHS).

Methods: Subjects were 4,037 respondents to the NPHS who were linked to the Ontario Health Insurance Plan. An ecologic measure of income was obtained by linkage of subjects' postal codes to the Census.

Results: The relationships between the ecologic-level measure and health outcomes or health services utilization were attenuated in comparison to the relationships relative to the direct measure of household income. The ecologic measure also produced poorer control for confounding by income in the analysis of other health relationships.

Conclusions: Many interesting public health and health services questions can be addressed only with the use of ecologic level socioeconomic information. While most of the results were qualitatively similar when the direct and ecologic measures were compared, researchers and users of research findings should be aware that attenuated or potentially misleading findings may result from the use of these methods.

Personal or household incomes are important correlates of health status and the utilization of health services in Canada,¹ the United States,² and Europe.³ Information on personal measures of income and socio-economic status (SES) is occasionally available to researchers, particularly when analyzing national surveys such as the National Population Health Survey (NPHS). In many cases, however, personal information is not available, and researchers are obliged to infer personal SES from ecologic measures obtained, for example, from census data.⁴

Canadian researchers make use of administrative data to study questions pertinent to health care and public health policy.^{5,6} There has been worry about the possibility of misclassification of SES in the use of these data. Demissie⁷ found substantial discrepancy between area-based SES measures and SES assessed at the individual level in Montreal. Sin⁸ studied the validity of using postal codes to identify groups of people with low socio-economic status. He reported that the use of Forward Sortation Area (FSA) postal codes as the only marker to identify people with low SES may result in substantial misclassification of personal poverty. Glazier⁹ reported that, although address inaccuracy was found in Ontario's health care registry, serious socio-economic misclassification occurred at a low rate and did not introduce significant bias in the calculation of hospitalization rates by socio-economic group. Southern et al.¹⁰ assessed the agreement between FSA and Enumeration Area (EA) derived income levels in Alberta. They found that the variability in EA-derived income quintiles was large for any given FSA-derived income quintile and recommended that EA-derived measures should be used when individual data are not available.

Measurement theory predicts that a group level measure will contain more measurement error than the direct household measure. Accordingly, the relationship between the group level measure and a health outcome will be attenuated in comparison to the true relationship relative to the direct measure of socio-economic status.¹¹ Mustard and colleagues¹² examined the validity of using ecologic measures of SES as proxies for individual-level measures by using a 5% sample of house-

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

Family Medicine Centre, Mt Sinai Hospital, Toronto, ON; Department of Family Medicine, McMaster University, Hamilton, ON; Department of Family and Community Medicine, University of Toronto
Correspondence: Dr. Murray Finkelstein, Family Medicine Centre, Suite 413, Mt Sinai Hospital, Toronto, ON M5G 1X5, Tel: 416-326-7879, Fax: 416-326-7761, E-mail: murray.finkelstein@utoronto.ca

holds in Manitoba and linking the records of the Manitoba Health Services Insurance Plan and the Census of 1986. They concluded that the hypothesis that risk estimates derived from ecologic income measures will be attenuated relative to estimates obtained from household income was not supported. The purpose of this paper was to make use of a unique data set to re-examine the effect on measures of health status, and health care utilization, of using ecologic rather than personal measures of income.

METHODS

The subjects were the Ontario respondents to the first wave, 1994/95, of the NPHS.¹³ Respondents were asked about their household income, and Statistics Canada derived an index of income adequacy on the basis of household income and household size. In 1994 dollars, the four categories of the index were as follows: lowest income, less than \$10,000; lower middle income, \$10,000 to \$22,499; upper middle income, \$22,500 to \$59,999; and highest income, \$60,000 or more for 1 or 2 people. For households with more than 2 people, the income ranges were scaled accordingly.

The Ontario Ministry of Health linked the NPHS file to the Ontario Health Insurance Plan (OHIP) providers' database, using the respondent's health insurance number as the identifier. The Ministry generated a data file containing a record of each service provided to NPHS respondents by physicians in the year before their interviews. Postal codes from

TABLE I

Distribution of Subjects by Self-reported and Enumeration Area Household Income Categories

Statistics Canada Income Adequacy Category (Based on Self-Reported Income 1994/95)	Number of Subjects	Enumeration Area Categories Based Upon Ranking of EA Median Household Incomes	
		Number of Subjects*	Mean of the "Median Household Incomes"
Lowest (< \$10K)	793	794	\$31,000
Lower Middle (\$10K - 22.49K)	1089	1089	\$43,500
Upper Middle (\$22.5K - 59.9K)	1446	1448	\$58,000
Highest (≥ \$60K)	709	706	\$88,000

* The aim was to create enumeration area categories with the same number of subjects as the Statistics Canada personal response categories. Because of ties in the rank ordering of enumeration areas, the category sizes are not exactly the same.

the OHIP file were used to link subjects to Enumeration Areas from the 1996 Census, and median household incomes in each EA were extracted.

Several measures of self-reported health status, and one of health services utilization, were selected for correlation with direct and proxy measures of household income. Subjects who reported their health as fair or poor were compared with those who reported their health to be good, very good, or excellent. Self-reports of the diagnosis of arthritis, back problems, heart disease, diabetes and the use of diabetes medications, asthma and the use of medication for asthma were also examined. The sum of all Assessment Fees (A or K codes) submitted to OHIP for each subject by General Practitioners in the year prior to the NPHS interview was used as a measure of health services utilization.

The prevalence of dichotomous conditions was analyzed with logistic regression and the cost of health services by multiple linear regression. The variables included in the models are presented in the tables of results. Survey weights were used in all cal-

culations and a robust variance estimator¹⁴ was computed to allow for the complex survey design. The Akaike Information Criterion (AIC)¹⁵ was used to compare non-nested logistic regression models. The smallest value of AIC indicates the best fit to the data. The F-score was used to compare non-nested linear regression models. The larger value of F indicates a better fit to the data.

RESULTS

Study subjects and income groupings

The Ontario component of the 1994/95 NPHS included 4,804 Ontario residents aged 12 years or more, representing some 9.4 million residents of the province. Individual-level household income estimates were available for 4,637 (96.5%) subjects. Linkage of the OHIP-derived postal codes to the 1996 census provided estimates of EA-level median household income for 4,188 (87%) of subjects. Measures of both self-reported and EA income were available for 4,037 (84%) respondents, and these are the subjects of this analysis.

TABLE II

Odds Ratios for Reporting Health Status to be Fair or Poor in Relation to Personal Factors and Self-reported or Enumeration Area Household Incomes

Explanatory Variable	Self-reported Household Income		Median Household Income in Enumeration Area	
	Odds Ratio (Z-Score)†	95% CI	Odds Ratio (Z-Score)	95% CI
Age (Per Year)	1.08 (5.27)	1.05 - 1.12	1.06 (4.01)	1.03 - 1.09
Square of Age	0.9995 (-3.41)	0.9992 - 0.9998	0.9997 (-1.88)	0.9995 - 1.0000
Female	1.10 (0.94)	0.90 - 1.35	1.25 (2.18)	1.02 - 1.52
High School Graduate	0.62 (-4.57)	0.50 - 0.76	0.50 (-6.87)	0.41 - 0.61
Income Category				
Lowest (< \$10K)	1.0 (Baseline)		1.0 (Baseline)	
Lower Middle (\$10K - 22.49K)	0.62 (-3.94)	0.49 - 0.79	0.82 (-1.51)	0.63 - 1.06
Upper Middle (\$22.5K - 59.9K)	0.33 (-8.24)	0.25 - 0.43	0.70 (-2.78)	0.54 - 0.90
Highest (≥ \$60K)	0.19 (-7.92)	0.13 - 0.29	0.45 (-4.53)	0.32 - 0.64
Akaike Information Criterion*	2746		2826	

† Values of the Z-Score greater than 1.96 or less than -1.96 indicate p<0.05.

* The Akaike Information Criterion is a measure of model fit. The lower score indicates better fit of the model to the data.

Figure 1. Distribution of self-reported household income categories in relation to categories of self-reported income. In 1994 dollars, the 4 categories of the index were as follows: 1) lowest income, less than \$10,000 for 1 to 4 people; 2) lower middle income, \$10,000 to \$22,499 for 1 or 2 people; 3) upper middle income, \$22,500 to \$59,999 for 1 or 2 people; and 4) highest income, \$60,000 or more for 1 or 2 people. The numbers of subjects in each category are shown in Table I.

In order to compare income groups of similar size, the EA-derived median incomes were ranked, and subjects were placed into EA-groups of (approximately) the same size as the self-report income categories. The distribution of the subjects by income category is shown in Table I.

Relations between self-reported and enumeration area household incomes

Figure 1 shows the distribution of self-reported incomes within each of 4 categories of Enumeration Area household incomes. Each EA category is clearly not homogeneous with respect to self-reported household income, so that the potential for misclassification exists when using ecologic data for analysis.

Relations between self-reported health status and personal and ecologic measures of income

Table II shows the results of logistic regression models explaining self-reported health status as fair or poor (N=555) versus good or better (N=3,482). These models are adjusted for age and gender and high school graduation status is included as an example of a variable that might be of substantive interest for some analyses, and for which we wish to control for confounding by income. One model includes self-reported income and the other the EA-derived ecologic measure. The Model AICs, reported in the bottom line, indicate that the model containing self-reported income provides a substantially better fit to the data than the model including the eco-

logic measure. This improved fit is reflected in larger z-scores for most of the explanatory variables. For the income categories, the inverse trend in reporting ill health is sharper for the self-report than for the ecologic measure.

Now, suppose that we were not primarily interested in income as an explanatory factor, but that income is included to control for confounding. We see that use of the ecologic measure of income has different effects on the estimates of the “substantive variables” than does the self-reported measure. The odds ratio for educational status is about 20% smaller in the ecologic model. In addition, the odds ratio for gender is significant in the ecologic model but not in the model with self-reported income, suggesting that confounding of

TABLE III

Odds Ratios for Reporting Various Health Problems, and the Use of Medication in Relation to Self-reported or Enumeration Area Household Incomes

Health Condition	Self-reported Household Income		Median Household Income in Enumeration Area	
	Odds Ratio (Z-Score)	95% CI	Odds Ratio (Z-Score)	95% CI
Arthritis	0.78 (-5.32)	0.71 - 0.86	0.85 (-3.53)	0.78 - 0.93
Back Problems	0.85 (-3.86)	(0.78 - 0.92)	0.92 (-1.98)	(0.85 - 1.00)
Diabetes	0.84 (-1.83)	0.70 - 1.01	1.04 (0.47)	0.88 - 1.23
Diabetes Medications	0.84 (-1.36)	0.65 - 1.08	1.08 (0.64)	0.85 - 1.36
Asthma	0.85 (-2.59)	0.76 - 0.96	0.94 (-0.94)	0.84 - 1.06
Asthma Medications	0.83 (-2.69)	0.72 - 0.95	0.88 (0.77 - 1.01)	0.77 - 1.01
Heart Disease	0.80 (-2.94)	(0.70 - 0.93)	0.78 (-3.75)	0.68 - 0.89
Akaike Information Criterion	2746		2826	

Note: The models are adjusted for age and sex. In these analyses, the income categories were coded as integers (lowest = 1 to highest = 4).

TABLE IV

The Cost of General Practitioners' Services in Relation to Personal Factors and Self-reported or Enumeration Area Household Incomes

Explanatory Variable	Self-reported Household Income		Median Household Income in Enumeration Area	
	\$Cost per Unit Increase (Z-Score)	95% CI	\$ Cost per Unit Increase (Z-Score)	95% CI
Constant	42.10 (1.91)	-1.21 to 85.4	-34.6 (1.54)	-9.3 to 78.0
Age (Per Year)	2.49 (-2.86)	0.79 to 4.20	1.64 (1.89)	-0.1 to 3.3
Square of Age	-0.017 (-1.91)	-0.034 to 0.0004	-0.007 (-0.84)	-0.02 to 0.010
Female	33.74 (5.49)	21.7 to 45.8	38.7 (6.31)	26.7 to 50.7
High School Graduate	6.00 (0.89)	-7.1 to 19.1	-5.42 (-0.83)	-18.3 to 7.4
Income Category				
Lowest (< \$10K)	0.0 (Baseline)		0.0 (Baseline)	
Lower Middle (\$10K - 22.49K)	-23.6 (-2.65)	-41.1 to -6.1	-10.0 (-1.12)	-27.5 to 7.5
Upper Middle (\$22.5K - 59.9K)	-52.1 (-5.94)	-69.3 to -34.9	-22.4 (-2.65)	-39.0 to -5.8
Highest (≥ \$60K)	-67.2 (-6.48)	-87.5 to -46.8	-19.3 (-1.94)	-38.9 to 0.3
F(7, 3898)	19.67		12.99	

the effect of gender by income is less well controlled by the ecologic measure of income.

Table III shows the relations between a number of other health conditions reported in the NPHS and the personal and ecologic income measures. For all but heart disease, the inverse trend with income was stronger for the personal measure. For asthma, and the use of asthma medications, the trend was significant for the personal measure but not for the ecologic one.

Relations between the costs of general practitioners' services and personal and ecologic measures of income

Table IV shows the results of linear regression models to explain expenditures on general practitioners' services in the year prior to the NPHS interview. The model F-scores indicate that the model including self-reported income provides a better fit to the data than the model including the ecologic measure. This improved fit is reflected in larger z-scores for most of the explanatory variables and the trend to

decreasing expenditures with higher household income is sharper with direct rather than ecologic measures.

DISCUSSION

The relations of income to health or health services utilization may be of interest in their own right, or researchers might wish to control for income in the analysis of other factors. In the absence of individual-level information on income, use is often made of area-wide measures obtained from census data. Misclassification may thus be introduced. Previous Canadian studies have highlighted a number of considerations in using ecologic level data for the study of income-related effects or for the control of confounding. If postal codes are obtained from medicare databases, researchers must be aware that addresses are frequently not kept current and misclassification might result.⁹ There is also the potential for misclassification if researchers use the 3-character Forward Sortation component rather than the full 6-character postal code.^{8,10}

When the 6-character postal code is available, these are usually mapped to Enumeration Areas. EAs, the geographic area canvassed by one census representative, are the smallest geographic areas for which census data are made available by Statistics Canada. The number of households in an EA varies from 125 in rural areas to 375 in urban areas. Even though the number of households is relatively small, misclassification may be introduced when assigning EA statistics, such as the median income, to individuals. Figure 1, which showed the distribution of self-reported income among NPHS subjects in relation to EA-derived categories, revealed substantial variability in household income at the EA-level. This variability allows the possibility of the Ecologic Fallacy¹⁶ when aggregate level income data is used instead of individual data.

Assuming that self-reported income is the "gold standard" for measurement, it was found that analyses using the EA-derived measure of income have properties predicted for misclassified variables by theory.¹¹ The relationship between the ecologic-level

measure and the health outcome was attenuated in comparison to the relationship relative to the direct measure of household income. In addition, the ecologic measure produced poorer control for confounding.

The results of this research are in contrast to the other Canadian study that compared health and health services utilization in relation to self-report and ecologic measures of income.¹² The Manitoba investigators reported that "the hypothesis that risk estimates derived from ecologic income measures will be attenuated relative to estimates obtained from household income was not supported." Examination of their tables of results finds, however, that for the majority of comparisons the odds ratios were higher for the self-reported measure than for the ecologic measure, even though in most cases the differences were not statistically significant. It is concluded that the attenuation predicted by theory is likely to occur for many measures of health and health care utilization if ecologic measures are used as a proxy for personal measures, particularly if postal codes are obtained from medicare databases.

In summary, many interesting public health and health services questions can only be addressed with the use of ecologic level socio-economic information. Previous investigations have addressed the issues of misclassification resulting from outdated address information and consolidation at too coarse a geographic level. The contribution of this investigation is to show that theoretical predictions of the effect of misclassified income on effect estimates, and the control of confounding, are verified in data from the NPHS. While most of the results were qualitatively similar when the direct and ecologic measures were compared, researchers and users of research findings should be aware that attenuated or potentially misleading findings may result from the use of these methods.

REFERENCES

1. Finkelstein MM. Do factors other than need determine utilization of physicians' services in Ontario? *CMAJ* 2001;165(5):565-70.
2. Lantz PM, House JS, Lepkowski JM, Williams DR, Mero RP, Chen J. Socioeconomic factors, health behaviors, and mortality: Results from a nationally representative prospective study of US adults. *JAMA* 1998;279(21):1703-8.
3. Marmot MG, Smith GD, Stansfeld S, Patel C, North F, Head J, et al. Health inequalities among British civil servants: The Whitehall II study [see comments]. *Lancet* 1991;337(8754):1387-93.
4. 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.
5. Roos NP, Mustard CA. Variation in health and health care use by socioeconomic status in Winnipeg, Canada: Does the system work well? Yes and no. *Milbank Q* 1997;75(1):89-111.
6. Alter DA, Naylor CD, Austin P, Tu JV. 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.
7. 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.
8. Sin DD, Svenson LW, Man SF. Do area-based markers of poverty accurately measure personal poverty? *Can J Public Health* 2001;92(3):184-87.
9. Glazier RH, Creatore MI, Agha MM, Steele LS. Socioeconomic misclassification in Ontario's Health Care Registry. *Can J Public Health* 2003;94(2):140-43.
10. Southern DA, Ghali WA, Faris PD, Norris CM, Galbraith PD, Graham MM, et al. Misclassification of income quintiles derived from area-based measures. A comparison of enumeration area and forward sortation area. *Can J Public Health* 2002;93(6):465-69.
11. Nunnally JC. *Psychometric Theory*. New York: McGraw-Hill, 1978.
12. Mustard CA, Derksen S, Berthelot JM, 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 Place* 1999;5(2):157-71.
13. Statistics Canada. National Population Health Survey 1994-95. Ottawa, Minister of Industry, 1995.
14. White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 1980;48:817-38.
15. Lindsey JK, Jones B. Choosing among generalized linear models applied to medical data. *Stat Med* 1998;17(1):59-68.
16. Greenland S, Robins J. Invited commentary: Ecologic studies – biases, misconceptions, and counterexamples. *Am J Epidemiol* 1994;139(8):747-60.

Received: May 7, 2003
Accepted: September 19, 2003

RÉSUMÉ

Contexte : En l'absence de données d'identification, les chercheurs utilisent souvent des mesures du statut socio-économique (SSE) dérivées du recensement. En théorie, les erreurs de classement qui pourraient en résulter devraient éteindre les associations entre les résultats et le SSE, et il devrait être plus difficile de tenir compte du facteur confusionnel que représente le SSE. Nous avons donc voulu étudier l'ampleur du problème à l'aide des données de l'Enquête nationale sur la santé de la population (ENSP).

Méthode : Nos sujets étaient les 4 037 répondants de l'ENSP reliés au Régime d'assurance-maladie de l'Ontario. Nous avons obtenu une mesure « écologique » [liée au milieu de vie] du revenu des sujets en reliant leurs codes postaux aux données du recensement.

Résultats : Les relations entre la mesure écologique et les résultats sanitaires ou le recours aux services de santé étaient atténuées par rapport aux relations obtenues par la mesure directe du revenu des ménages. En outre, la mesure écologique a moins bien permis de tenir compte du facteur confusionnel que représente le revenu dans l'analyse des autres relations touchant la santé.

Conclusions : Beaucoup de questions intéressantes sur la santé publique et les services de santé ne peuvent être étudiées qu'en employant une information socio-économique à l'échelle écologique. Bien que l'on obtienne des résultats semblables, du point de vue qualitatif, que l'on utilise des mesures directes ou écologiques, il est bon que les chercheurs et les utilisateurs des résultats de recherche sachent que ces méthodes peuvent mener à des constatations atténuées ou potentiellement trompeuses.