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

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.
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.

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. Alberta has a total of 137 FSAs and 4,746 EAs. The Statistics Canada Postal Code Conversion File (PCCF) 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.

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. 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 by FSA-based and EA-based median income.

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

TABLE I
Characteristics of APPROACH Population and FSA and EA-based Samples

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Full Data (N=21,812)</th>
<th>FSA-linked (N=21,731)</th>
<th>EA-linked (N=21,664)</th>
<th>Both EA/FSA linked (N=21,446)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;65</td>
<td>11,803 (54.1)</td>
<td>11,751 (54.1)</td>
<td>11,726 (54.1)</td>
<td>11,591 (54.0)</td>
</tr>
<tr>
<td>65-74</td>
<td>6885 (31.6)</td>
<td>68,661 (31.6)</td>
<td>68,34 (31.6)</td>
<td>67,81 (31.6)</td>
</tr>
<tr>
<td>75+</td>
<td>3124 (14.3)</td>
<td>3114 (14.3)</td>
<td>3104 (14.3)</td>
<td>3074 (14.4)</td>
</tr>
<tr>
<td>Male</td>
<td>15,406 (70.6)</td>
<td>15,344 (70.6)</td>
<td>15,303 (70.6)</td>
<td>15,140 (70.6)</td>
</tr>
<tr>
<td>Rural</td>
<td>4506 (20.7)</td>
<td>4506 (20.7)</td>
<td>4381 (20.2)</td>
<td>4357 (20.3)</td>
</tr>
</tbody>
</table>
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.\textsuperscript{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.\textsuperscript{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\textsuperscript{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.\textsuperscript{10,12,22} EA is the smallest unit for which census data are collected and thus 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.

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.\textsuperscript{1,22,23} Some compared area-based measures to smaller area-based measures (i.e., CT to EA, postcodes to EA).\textsuperscript{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.\textsuperscript{22} Geronimus et al. found that there was very little gain from using smaller area-based measures.\textsuperscript{7} In contrast, Soobader et al. argue that measures from smaller geographic units may produce results that are slightly less biased.\textsuperscript{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

19. Mackillop WJ, Zhang-Salomons J, Groome PA, Paszat L, Holowaty E. Socioeconomic status and