ZIP-Code-based versus Tract-based Income Measures as Long-Term Risk-adjusted Mortality Predictors

Avis J. Thomas¹, Lynn E. Eberly¹, George Davey Smith², and James D. Neaton¹ for the Multiple Risk Factor Intervention Trial (MRFIT) Research Group

¹ Coordinating Centers for Biometric Research, Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN.
² Department of Social Medicine, University of Bristol, Bristol, United Kingdom.

Received for publication November 29, 2005; accepted for publication March 10, 2006.

There is a well-established, strong association between socioeconomic position and mortality. Public health mortality analyses thus routinely consider the confounding effect of socioeconomic position when possible. Particularly in the absence of personally reported data, researchers often use area-based measures to estimate the effects of socioeconomic position. Data are limited regarding the relative merits of measures based on US Census tract versus ZIP code (postal code). ZIP-code measures have more within-unit variation but are also more easily obtained. The current study reports on 293,138 middle-aged men screened in 14 states in 1973–1975 for the Multiple Risk Factor Intervention Trial and having 25-year mortality follow-up. In risk-adjusted proportional hazards models containing either ZIP-code-based or tract-based median household income, all-cause mortality hazard ratios were 1.16 (95% confidence interval: 1.14, 1.17) per $10,000 less ZIP-code-based income and 1.15 (95% confidence interval: 1.13, 1.16) per $10,000 less tract-based income; adding either income variable to a risk-adjusted model improved model fit substantially. Both were significant independent predictors in a combined model; tract-based income was a slightly stronger mortality predictor (hazard ratios 1.05 and 1.11 for ZIP-code-based and tract-based income, respectively). These patterns held across various causes of death, for both Blacks and non-Blacks, and with or without adjustment for ZIP-code-based income diversity or tract-based proportion below poverty.

censuses; geography; income; inequalities; longitudinal studies; neighborhood; social class; socioeconomic factors

Abbreviations: SEP, socioeconomic position; –2log(L), –2log(partial likelihood function).

Low socioeconomic position (SEP) has been clearly established as a risk factor for many causes of death, independent of known biologic risk factors (1, 2). The relative contributions of personal and contextual SEP have not been clearly established (3), but there are strong indications that area-based measures, reflecting community-wide characteristics, are important mortality predictors in their own right and are not merely surrogate estimates of individual SEP (4, 5). Thus, it is often desirable to use area-based SEP estimates in analyzing health outcomes, whether or not personally reported income is available. Controversy exists about the relative merits of measures based on US Census tract versus ZIP code (postal code), with ZIP-code measures usually covering a larger area and having more within-unit variation but also being more easily obtained (6). This investigation, based on 25-year mortality follow-up of 293,138 middle-aged US men screened in 18 cities and 14 states, sought to evaluate the relative...
contributions of ZIP-code-based and tract-based income measures.

MATERIALS AND METHODS

From 1973 to 1975, 361,662 US men aged 35–57 years were screened for the Multiple Risk Factor Intervention Trial (MRFIT) at 22 clinical sites in 18 cities and 14 states. Data were collected on major risk factors for coronary heart disease (age, number of cigarettes smoked per day, blood pressure, serum total cholesterol, use of medication for diabetes, and history of hospitalization for myocardial infarction), race/ethnicity, Social Security number, birth date, and home address (7). Men selected one of six choices for their race/ethnicity: White, Black, Oriental, Spanish American, American Indian, or other. Home addresses were collected and later geocoded and matched to 1980 US Census data by ZIP code and US Census tract (1, 2, 8). The 22 clinical sites were in largely urban settings and used a variety of means to recruit volunteers, including door-to-door canvassing, targeted efforts with civic- and employment-based groups, and recruitment at shopping malls and street fairs.

Mortality was monitored through 1999 (≥25 years) by using Social Security Administration files and the National Death Index. The quality of the National Death Index link was previously assessed for the 12,866 men who were randomized into the Multiple Risk Factor Intervention Trial; a routine National Death Index match would have correctly identified 98.4 percent of 191 known deaths in 1979–1980 and produced 0.03 percent false matches with a matching Social Security number (9). Cause of death was obtained directly from death certificates (coded separately by two or three nosologists) or from the National Death Index Plus service (10).

Baseline summary statistics and analysis of variance models were used to investigate the joint distribution of tract-based and ZIP-code-based income measures. Proportional hazards regressions for various causes of death were performed, stratified by clinical site. The base model included adjustments for major risk factors for coronary heart disease and three racial/ethnic indicators: Black, Hispanic, and Asian American, with Whites as the referent group. ZIP-code-based and tract-based income variables were then added to the base model, separately and in combination. Estimated hazard ratios and 95 percent confidence intervals were compared, as were $-2\log(\text{partial likelihood function})$ ($-2\log(L)$) model fit statistics. Between-model differences in $-2\log(L)$ were assumed to be $\chi^2$-distributed, with degrees of freedom equal to the number of additional variables, and were used to quantify the degree of model improvement obtained from additional variables.

To further assess income-based mortality predictors, the ZIP-code-based Gini coefficient and the tract-based proportion below poverty were considered. The Gini coefficient is a measure of income inequality ranging from 0 to 1, with 0 representing complete equality and 1 complete inequality (6, 11). The federally defined income cutoff for poverty varied by family size, number of children, and age of the householder; for a typical family of four, the threshold was $7,356 in 1979 dollars. All-cause mortality models with these additional income-related adjustments were compared by examining the $-2\log(L)$.

In addition, men were divided into four groups based on higher versus lower tract and ZIP-code income. All-cause mortality regressions examining between-group differences were conducted, using men with high tract and ZIP-code income as the referent group.

Regressions and analysis of variance analyses were conducted with SAS version 8.2 software (SAS Institute, Inc., Cary, North Carolina).

RESULTS

A total of 293,138 (81 percent) of the 361,662 men screened for the Multiple Risk Factor Intervention Trial had complete baseline health data and median household income data and were included in this study. Exclusions encompassed 8,322 men (2 percent) without baseline systolic blood pressure, 15,717 additional men (4 percent) without ZIP-code-based median household income, and another 44,485 men (12 percent) without tract-based median household income (men who did not report a geocodable home address, had an address that was outside the urban areas in which 1980 US Census tracts were defined, or resided in a tract where the Census Bureau suppressed statistics because of insufficient sample size). The men lived in 3,881 ZIP codes and 14,031 US Census tracts. The means of the median ZIP-code-based and tract-based household incomes for the men included in this study were $20,847 (standard deviation, $5,990) and $21,692 (standard deviation, $7,275), respectively, and the correlation coefficient was 0.78. Tract-based percentage below poverty averaged 7.6 percent (standard deviation, 7.5 percent) and had a $-0.68$ correlation coefficient with tract-based income.

Because of the recruitment methods used, there was a wide range in the number of US Census tracts represented within each ZIP code (table 1). A total of 267,891 (91.4 percent) of the men were living in the 1,758 ZIP codes that overlapped five or more US Census tracts represented in the study sample. In an analysis of variance analysis, 66.9 percent of the total variation in tract-based income was attributed to variations between (vs. within) ZIP codes.

ZIP-code-based and tract-based income performed similarly when added to risk-adjusted models (models 1 and 2, table 2). For all-cause mortality, the hazard ratios for $10,000 less ZIP-code-based and tract-based incomes were nearly identical: 1.16 and 1.15 (both $p < 0.0001$), respectively, and $-2\log(L)$, an inverse measure of model fit, was reduced by 428 and 548 (both $p < 0.0001$) compared with the base model with no SEP variables. On a per-standard-deviation basis rather than a per $10,000 basis, hazard ratios for tract-based income and ZIP-code-based income were 1.10 and 1.09, respectively (not shown). A single model using both income measures (model 3, table 2) reduced $-2\log(L)$ by 574 compared with the base model; both income hazard ratios were attenuated, particularly the ZIP-code-based hazard ratio, but both remained significant. The incremental improvements in model fit, versus either
In addition to the analyses shown in table 2, sensitivity analyses were performed to further characterize the relative contributions of tract- and ZIP-code-based income variables to predicting all-cause mortality. Stratifying by the 3,881 ZIP codes, rather than adjusting for ZIP-code-based income, shifted the estimated hazard ratio for $10,000 less tract-based median household income by only 0.01. Adding the tract-based percentage below poverty to model 3 changed the estimated hazard ratios for tract- and ZIP-code-based median incomes only slightly and increased the model’s \(-2\log(L)\) by 25 \(p < 0.0001\) for improvement in model fit; the hazard ratio for the percentage below poverty was 1.42 (95 percent confidence interval: 1.24, 1.63). Adjustment for the ZIP-code-based Gini coefficient had no material effect on the associations of ZIP-code-based and tract-based income with mortality or on the model fit statistic (data not shown).

To allow for possible interaction between ZIP-code-based and tract-based SEP, the men were divided into four groups on the basis of whether ZIP-code-based and tract-based incomes were above or below their respective medians for all study participants. When high ZIP-code income/high tract income was used as the reference group, the risk-adjusted hazard ratio associated with low ZIP code/high tract was 1.05 (95 percent confidence interval: 1.02, 1.07); corresponding hazard ratios for high ZIP code/low tract and low ZIP code/low tract were 1.10 (95 percent confidence interval: 1.07, 1.13) and 1.20 (95 percent confidence interval: 1.17, 1.22). These results were consistent with results obtained when continuous income variables were used (data not shown).

### TABLE 1. Distribution of ZIP codes (postal codes) in the study cohort, by number of US Census tracts represented in that ZIP code (obtained from screening for the Multiple Risk Factor Intervention Trial, 1973–1975)

<table>
<thead>
<tr>
<th>ZIP-code areas in this category</th>
<th>ZIP-code areas overlapping</th>
<th>No.</th>
<th>%</th>
<th>Men represented</th>
</tr>
</thead>
<tbody>
<tr>
<td>only 1 tract represented</td>
<td>1,084</td>
<td>27.9</td>
<td>3,030</td>
<td>1.0</td>
</tr>
<tr>
<td>2–4 tracts represented</td>
<td>1,039</td>
<td>26.8</td>
<td>22,217</td>
<td>7.6</td>
</tr>
<tr>
<td>5–9 tracts represented</td>
<td>882</td>
<td>22.7</td>
<td>65,838</td>
<td>22.5</td>
</tr>
<tr>
<td>≥10 tracts represented</td>
<td>876</td>
<td>22.6</td>
<td>202,053</td>
<td>68.9</td>
</tr>
<tr>
<td>Total</td>
<td>3,881</td>
<td>100.0</td>
<td>293,138</td>
<td>100.0</td>
</tr>
</tbody>
</table>

model with a single SEP variable, were highly significant \(p < 0.0001\). Similar results were found for cause-specific mortality. For deaths due to cardiovascular diseases, coronary heart disease, all cancers, and lung cancer, ZIP-code-based income significantly improved models already containing tract-based income (all \(p \leq 0.006\). Thus, ZIP-code-based and tract-based income estimates were approximately equally predictive of mortality, with tract-based income providing slightly better fit, and with a slight additional model-fitting advantage from including both.


<table>
<thead>
<tr>
<th>Cause of death</th>
<th>No. of deaths</th>
<th>HR* and 95% CI† associated with a $10,000 lower median household income</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Model 1‡: ZIP-code based</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HR 95% CI</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ZIP-code based</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All causes</td>
<td>72,021</td>
<td>1.16 1.14, 1.17</td>
</tr>
<tr>
<td>Cardiovascular diseases</td>
<td>29,778</td>
<td>1.16 1.13, 1.18</td>
</tr>
<tr>
<td>Coronary heart disease</td>
<td>20,578</td>
<td>1.17 1.14, 1.20</td>
</tr>
<tr>
<td>All cancers</td>
<td>25,681</td>
<td>1.10 1.08, 1.13</td>
</tr>
<tr>
<td>Lung cancer</td>
<td>8,061</td>
<td>1.21 1.16, 1.26</td>
</tr>
<tr>
<td>Injury and violence</td>
<td>3,118</td>
<td>1.21 1.13, 1.29</td>
</tr>
</tbody>
</table>

* HR, hazard ratio; CI, confidence interval.
† The \(-2\log(\text{partial likelihood})\) is a measure of model fit. In nested models (i.e., those based on the same participants and with one model's variables a subset of the other model's variables), differences in the \(-2\log(\text{partial likelihood function})\) are assumed to have a \(\chi^2\) distribution with degrees of freedom equal to the degrees of freedom in the additional variables.
‡ Model 0 (not shown): Predictors include age, number of cigarettes smoked per day, serum total cholesterol, systolic blood pressure, diabetes (yes/no), previous myocardial infarction (yes/no), and three racial/ethnic indicators: African American, Hispanic, and Asian American.
§ Model 1: Add median household income, by ZIP code, to model 0.
¶ Model 2: Add median household income, by US Census tract, to model 0.
# Model 3: Add both ZIP-code-based and tract-based median household incomes to model 0.
Analyses were repeated separately for Blacks and non-Blacks and for each state of residence. Estimated hazard ratios were robust across both race/ethnicity and geographic area, although confidence intervals were wider.

DISCUSSION

For this large cohort of men aged 35–57 years when screened in 1973–1975 in 14 US states, with 72,021 deaths over 25 years, the correlation coefficient was 0.78 for US Census-tract- and ZIP-code-based household income, and both performed similarly as strong inverse predictors of all-cause and cause-specific mortality. In proportional hazards mortality models containing them both, the hazard ratio for US Census-tract-based income was further from 1.0, but both made independent statistically significant contributions to the model fit. Further addition of tract-based proportion below poverty also improved the model fit slightly; adjusting for ZIP-code-based income inequalities (represented by the Gini coefficient) had no material effect. Men were divided into four groups on the basis of whether tract-based and ZIP-code-based income were above or below their respective median; the regressions comparing the four groups confirmed that tract-based and ZIP-code-based income were both significant independent predictors of all-cause mortality, with tract-based differences having a greater effect on hazard ratios.

Analyses of tract-based income were conducted with ZIP-code-based income as an adjusting variable or, alternately, with ZIP code itself as a stratifying variable. In either case, tract-based income remained a significant predictor. However, after one income variable was present in the model, adding additional income-related variables (the other median income variable, Gini coefficient, or percentage below poverty) resulted in only small improvements in the model fit (as measured by –2log(L)). In risk-adjusted models lacking SEP information, adding ZIP-code-based income improved the model fit nearly as well as adding tract-based income.

The usefulness of ZIP-code-based medians as mortality predictors was slightly offset by their smaller standard deviation. At the same time, because ZIP-code-based medians are computed from larger populations than tract-based medians, they are likely more stable measures of local income. ZIP-code-based measures were also available for a larger number of study participants; 44,485 screened men, or 12 percent, were excluded solely for lack of tract-based data (primarily because the address could not be geocoded or because US Census-level data had been suppressed for privacy reasons). Finally, an address’s ZIP code can be read directly, without the costly geocoding efforts required for obtaining US Census tracts. ZIP-code-based statistics were thus more readily available for a larger number of screened men than tract-based statistics, and they had a similar effect on model fit as a stand-alone SEP adjustment. This result differs from other studies, which have focused on the superiority of tract-based data (6, 12) or raised significant questions about the accuracy of ZIP-code-based data in assessing an individual’s risk (13). Most of the existing studies on this subject have been limited to just a few locales. The relative value of tract-based and ZIP-code-based data may vary by locale, but the all-cause mortality results in this study were robust across the 14 states considered.

It may be that tract-based and ZIP-code-based measures are significant independent mortality predictors because they capture slightly different effects. Tract-based variables, for instance, could be more relevant predictors of an immediate neighborhood’s “walkability,” a benefit in maintaining cardiovascular health and reducing stress, whereas ZIP-code-based variables might better capture the benefits of high-quality health services within easy driving distance (6).

The address-based income measures correspond to a single snapshot of income rather than continuously gathered data, which is a potential limitation. However, most men aged 35–57 years would have settled into a career path and SEP. Since most people (especially homeowners) are slow to relocate, neighborhood measures tend to reflect long-term SEP rather than short-term job fluctuations and, as such, would have smoothed out some of the vagaries inherent in a single baseline reading.

In conclusion, ZIP-code-based and tract-based median income, which were highly correlated, were both significant independent mortality predictors. Adding either one to a risk-adjusted model substantially improved model fit. ZIP-code-based income, while not quite as strong a predictor as tract-based income, continued to make a statistically significant contribution in a combined model.

ACKNOWLEDGMENTS

The Multiple Risk Factor Intervention Trial was conducted under contract with the National Heart, Lung, and Blood Institute, Bethesda, Maryland. This work was supported by National Heart, Lung, and Blood Institute grants R01-HL-43232 and R01-HL-68140.

Conflict of interest: none declared.

REFERENCES


