LETTERS TO THE EDITOR

RE: "USE OF CENSUS-BASED AGGREGATE VARIABLES TO PROXY FOR SOCIOECONOMIC GROUP: EVIDENCE FROM NATIONAL SAMPLES"

Geronimus and Bound (1) are to be commended for conducting research to improve measuring and monitoring socioeconomic inequalities in health. Their article comparing estimate effects derived from models using individual-level, census tract, and zip code socioeconomic data may detract from this goal, however, given likely limitations of the validity and generalizability of their study findings.

First, complex sample structure and substantial cohort attrition combined with low success rates for geocoding may compromise their results. A number of well-documented problems arise when comparing census data with sample survey data (2–7). Surveys like the Panel Study of Income Dynamics (PSID) used by Geronimus and Bound have a complex sample structure that needs to be taken into account when interpreting the results (8). Any comparative analysis needs to distinguish population differences from sampling variation (9–10).

The PSID is a longitudinal survey that began in 1968 with 4,802 families, 2,930 of which were drawn as a quota sample (not a random sample), and 1,872 of which were drawn from a sample of low-income households. Thus, the “1968 PSID sample is quite unlike the population of the United States” (11). Sample weighting is used to offset biases in the 1968 sample and considerable differential attrition of respondents between 1968 and 1985. The 1985 sample contained only 60 percent of the original set of households sampled in 1968 (1). While sample weighting may be able to correct for biases between sociodemographic groups, it is extremely unlikely that it can also simultaneously correct for inaccuracies at census-tract or other microgeographic level (12). In fact, the weightings used will actually make things worse since if “poor” young people are weighted by, say, 0.5 and “rich” men and women are weighted by, say, 5.0 or 10.0, this may result in the correct number of poor and rich people at national level. However, the rich people with high weightings still live in only one place and can be compared with only one set of census data, even though they have a weighting that means that they represent five or 10 people. Moreover, even if accuracy at census-tract level could be corrected by weighting the results from a relatively small survey like the PSID, it would still be imprecise at small area level (e.g., they would contain a large amount of random error).

Similarly, weighting the PSID cannot correct for biases created by inadequate geocoding. Whereas many other US studies geocoding existing data sets have successfully geocoded over 80 percent, and typically over 90 percent, of addresses to the census-tract or block-group level (13–22), Geronimus and Bound were able to match only 68 percent of respondents in the 1985 sample to 1970 zip codes and only 72 percent to 1980 census-tract data. Assuming an original response rate in 1968 of around 80 percent, the effective sample response rate thus shrinks to only about 35 percent for 1980 Census tracts and 33 percent for 1970 Census tracts. Losing two thirds of the original sample would, in turn, inflate the 95 percent confidence intervals for variables of interest by a significant degree. Yet, despite these sampling problems, the article interprets the larger variability of the microdata compared with the census data as a key indicator of the lack of power of the aggregate census variables. An alternative explanation not considered is that the larger variability of the microdata simply reflects a large sampling uncertainty. The very limited health-related variables on the PSID (e.g., self-reported health status) also pose additional constraints on interpretation, in that while these data may yield accurate aggregated results at national level, they may be expected to be relatively imprecise at the individual and the microgeographic level. The relative imprecision of survey-based estimates (compared with the census) may result in inflated confidence in the individual-level results. Higher correlations within a biased sample may be expected than are seen between a biased sample and a relatively unbiased census.

Second, the article comments on the reliability of census data in estimating health status and outcome without performing a reliability analysis (1). Performed properly, a reliability analysis would provide an estimate of the correlation between the set of health-related census indicators and the “true” scores that would be obtained if the “infinite set” of all possible health-related questions had been asked. It would also provide the average correlation between the set of questions asked in the census and all other possible sets of health-related indicators of equal length (equal number of questions) (23). Yet, rather than following this approach, the article uses chi-square instead of standard reliability analysis techniques such as test-item analysis or Chronbach’s alpha. Given the nonstandard nature of the reliability analysis and the known sample size dependence of the chi-square statistic, it is difficult to be confident in Geronimus and Bound’s interpretation of these results.

Third, disregarding conflicting evidence, it is striking that the article scantily reviews or ignores relevant public health studies conducted in the United States (16, 24–27) and other countries (most notably the United Kingdom (28–33)) comparing effect estimates based on individual- and area-based socioeconomic data across diverse health outcomes. All but the previous (34) and current (1) studies by Geronimus and Bound have consistently reported larger estimates of effect using individual compared with area-based socioeconomic measures, raising the question of why their investigations have found the opposite. Although aggregation bias may be one explanation (1), it is equally plausible that their results are skewed by reliance on a survey sample, attrition of the cohort over time, and an unusually low proportion of successfully geocoded addresses, as described above.

Their finding of comparable economic heterogeneity in census tracts and zip codes (1) is also unusual, given that: 1) tracts on average are over five times smaller than zip codes (4,000 vs. 25,000 residents) and 2) people cluster by socioeconomic characteristics. This is because economic segregation in the housing market results in the people who live next door to you being more like you than the people in the next block,
the people in the next block being more alike than those in the rest of the neighborhood, the people in the neighborhood being more alike those than in the rest of the city, etc. (35–39).

Fourth and finally, the article makes two incorrect assertions about census block-group data. First, far from being "rarely available" (1), census block-group socioeconomic data can easily be obtained from the same census file (STF3A) that contains census-tract data (40). Moreover, given a decision to geocode to the tract level, geocoding to the block-group level imposes little, if any, additional cost (16, 41). In addition, block-group data do not "systematically exclude rural residents" and zip codes are not the only geographic unit with "the potential for complete coverage" (1, p. 476). In fact, since 1990 all geographic areas in the United States have been assigned census-tract codes (or, in rural areas, their equivalent, termed "block numbering areas") and block-group codes (40, 42). Epidemiologists and other public health researchers should therefore be aware that census block-group data provide an invaluable resource for assessing neighborhood socioeconomic characteristics in relatively small regions containing, on average, only 1,000 residents (40–42).

In summary, the finding by Geronimus and Bound that microdata are more accurate and reliable than inferring individual characteristics from aggregated census data is both predictable and already well-demonstrated. What is surprising, however, is their inattention to the anomalous nature of their results and their recommendation to prioritize adding individual-level socioeconomic data to public health databases over geocoding these data sets to enable using area-based socioeconomic measures (1). This recommendation should be tempered by caveats about the validity of their findings and also by the impracticality of their suggestion for monitoring socioeconomic inequalities in health using the many already extant vital statistic and other public health databases lacking socioeconomic data (43). In light of growing evidence of the importance of neighborhood conditions on individuals' health (17, 20–22, 29, 41, 44), it would perhaps be more judicious, both from a scientific and a policy standpoint, to recommend that public health databases include both micro- and macrolevel socioeconomic data.

REFERENCES

To our knowledge, there are only three other studies that have addressed the specific question that we posed—census-based aggregate variables are poor proxies for individual characteristics in health outcome equations. We did not assess the relative importance of contextual versus individual characteristics in relation to health. To our knowledge, there are only three other studies that have even addressed the specific question that we posed (4–6). We offered the following conclusions: 1) census-based aggregate variables are poor proxies for individual characteristics—they have substantially less statistical power than do microlevel measures and are likely to overestimate effects; 2) aggregate socioeconomic variables are highly multicollinear; and 3) somewhat surprisingly, while there is some improvement in precision if a researcher uses aggregate data drawn from a more recent census year compared with one that is 10 years earlier or that uses data aggregated at the census-tract rather than the zip code level, these gains are small.

Krieger and Gordon question our conclusions. They argue that the data we use are unrepresentative, that our statistical methods are flawed, and that our results are at odds with the bulk of research in the area. We disagree.

First, Krieger and Gordon limit their criticisms to only one of the data sets we used, the PSID. Some of their concerns are inapplicable to the NMIHS, including the prominent issue of sample attrition. They also highlight only those PSID subsamples in which we had relatively low rates of matching between the survey sample and the census data (68–72 percent). They fail to mention that we matched 95 percent of the PSID respondents and 89 percent of the NMIHS respondents to 1980 zip codes. Even if one considers only the 1980 zip code analysis, our primary conclusions remain the same.

They question the representativeness of the PSID sample. The many evaluations of the PSID all come to the conclusion that, when weighted, it remains representative of the non-Hispanic US population (7, 8). Of course, these evaluations have not considered the representativeness of the distribution of PSID respondents across zip codes or census tracts, but our own tabulations suggest that respondents are not concentrated in particular zip codes or census tracts or in ones that are unrepresentative of the nation as a whole.

Some issues Krieger and Gordon raise are not pertinent to the validity of estimates based on our PSID samples but, instead, concern their reliability. Thus, for example, the use of sample weights can and typically does reduce the precision of estimates, but does not bias them. Moreover, if one takes into account the fact that weights were used when calculating test statistics, as we did, the decrease in precision will not lead to faulty inference. (Let us add that unweighted tabulations yield results that are similar to the weighted ones we report.) Similarly, geographic clustering does not bias estimates, but typically reduces the efficiency of a sample. We experimented with calculating standard errors using methods that account for clustering, finding that the degree of clustering that exists in our data was sufficiently minor to have little effect on calculated standard errors. Furthermore, the fact that, in the end, we ignored clustering when calculating standard errors would bias the standard errors we estimated on the aggregate variables more than they bias the standard errors on the microvariables (9, 10). Thus, had we reported standard errors that took into account the clustering in our data, it is likely that our finding that census-based proxies lack statistical power relative to their microlevel counterparts would have been strengthened.

Most important, Krieger and Gordon fail to note that samples used in previous research comparing results based on aggregate versus micromeasures have less of a claim to representativeness than do the two data sets we used. Of the three other studies that directly compared the use of census-based aggregate with microlevel socioeconomic variables in health outcome equations, the samples studied were 101 women from Alameda County, California (4), a subsample of participants in the Kaiser Health Insurance plan in one state