DISCUSSION

Maria Elena Gonzalez, Office of Federal Statistical Policy and Standards

"Synthetic Estimates for Local Areas from the Health Interview Survey" is an important paper which compares synthetic and regression estimates with survey estimates from the Health Interview Survey (HIS) and a Baltimore Telephone Survey (BTS) on health. The data presented are extensive and allow us to evaluate the various methodologies used.

First, we will present a procedural comparison of the two surveys which provide the basis for the data analyzed in this paper, that is, the HIS and the BTS. If one were to prepare an error profile, that is, a systematic and comprehensive account of the survey operations that yield survey results, for each of the two surveys one would undoubtedly encounter many possible differences in procedures which might lead to differing data results. However, the objectives of both surveys were to measure health effects in a comparable manner, but the amount of data collected in the interviews differs.

The sample design differs in the type of clusters; also non-telephone household are not included in the BTS. For Baltimore, HIS has has a sample of 400 households and BTS of 2,500. The question wording was the same, but the HIS interview lasted one hour while the BTS lasted 30 minutes. HIS had one interviewer in Baltimore and BTS used 15 interviewers. The response rate for the HIS Baltimore PSU was 95% (nationally 96%) and 76% for the BTS. We do not know how imputation for nonresponse was handled in both surveys. The processing of the data for both surveys, again, was not reported. Estimation procedures were discussed briefly, but differences between the surveys were not analyzed; different seasonal effects were recognized for HIS and BTS. The various procedural differences noted might justify many of the differences found between the BTS and HIS survey data.

The National Center for Health Statistics has contracted with the Survey Research Center in Michigan to develop a telephone interview comparable to the HIS; this project should soon provide more insight into the feasibility of obtaining comparable results in telephone and personal interview surveys on health.

An article on Response Styles in Telephone and Household Interviewing: A Field Experiment by Jordan, Marcus and Reeder (1980) analyzes a personal interview and telephone survey on health behavior and health attitudes in the Los Angeles metropolitan area. In terms of demographic characteristics this study found a significantly higher proportion of questionnaires missing the income information for the telephone survey than for the personal interview data; otherwise there were no differences in demographic characteristics. Table 2 compares the sex-race-age groups used as cells to compute the synthetic estimates for the national HIS, the Baltimore HIS PSU and the BTS before and after adjustments; these estimates show similar distributions; however, no estimates showing income characteristics of the two samples are given.

The main objective of the paper presented by Joe Waksberg is evaluating synthetic and regression estimates of various health characteristics for small areas. The ultimate objective of this paper might be to develop a methodology to obtain estimates of adequate accuracy for health variables in about 200 Health Service Areas (HSA) in the U.S. (or perhaps even for sub-HSA areas). If these small area data were to be collected through a survey it would be extremely costly and, therefore, not a realistic option, except for selected local areas that might have special funds available.

The synthetic estimates were based on national HIS rates for sixteen breakdowns of race, sex and age; the local area population for these same sixteen breakdowns were also used to compute the synthetic estimates. Synthetic estimates based on other alternative breakdowns were not presented. The paper suggests that further analyses might be carried out by considering two categorical variables: degree of urbanization (2 categories) and Census region (4 categories). I agree with introducing variables related to geography; at the same time if these new breakdowns were used to obtain synthetic estimates some breakdown used now might be dropped, since the use of too many cells will not necssarily improve results significantly.

A project in which, I worked about 5 years ago computed synthetic estimates of the unemployment rate based on the 1970 Census of Population; these synthetic estimates were based on divisions; the synthetic estimates based on occupation, race and sex produced a higher correlation (.68) than those based on marital status, race and sex (.57).

Table 3 of the Waksberg paper presents results for the BTS and for synthetic estimates based on national HIS rates for the six component counties of Baltimore and for the SMSA. Remember that in Table 1 we have already seen sharp differences between the results of the telephone survey and HIS: these differences are surely reflected in Table 3. In addition, synthetic estimates tend to reduce the range of variation of the actual estimates; this result has been observed and produces a significant number of large errors particularly for extreme values. In a paper that Gonzalez and Hoza (1978) prepared the relative method error for the 1970 unemployment rate was defined as:

$$\frac{\text{Synthetic} - \text{Census}}{\text{Census}}$$

Of the 2908 counties tabulated, 43% had a relative method error less than ± 0.2; 80% of the counties had an error less than ± 0.5; 95% of the counties had an error less than ± 1.0; and .1% of the counties had an error greater than 2.0.

Table 5 shows relative root mean square errors (RMSE) for synthetic estimates ranging from .210 for number of doctor visits per person per year to 1.01 for work loss days per person per year; these estimates are based on 356 HIS primary sampling units (PSUs). Most of the relative RMSEs are in the range from .3 to .5. One must recognize that synthetic estimates are biased and their biases

are not negligible. Synthetic estimates are frequently used, although not always identified as synthetic estimates. A question to consider is whether introducing more efficient breakdowns in estimating synthetic estimates might improve these relative RMSEs by 10%, 20% or 50%.

Table 4 gives regression estimates based on the 356 PSUs of HIS; the dependent variable is from HIS and various independent variables are used, including synthetic estimates. The R squares presented are quite low (.295 and below). Let me list a few questions that come to mind. Should separate regression equations have been estimated by region? Should outliers have been identified and excluded from the computation of certain steps in the calculation? What other models should have been tried?

In using synthetic or regression estimates outliers might be identified and excluded from the computation of selected steps in the calculation. About ten years ago, I was estimating synthetic estimates of unemployment rates of SMSA's. The estimate obtained for Honolulu was completely unrealistic because the unemployment rate for Black and other races for the Western Region was used to calculate the unemployment rate for Honolulu. I still remember this error. One would probably want to use a local survey to estimate the unemployment rate for Honolulu, even if one were calculating synthetic estimates for other areas.

One of my firm convictions is that in the decade of the 80's the demand for current small area estimates in the fields of health, labor and income (to mention but three areas) will be greatly increased. Because of the great cost of direct data collection to obtain estimates for small areas and because of the major concern of the Federal Government with respondent burden, alternative methodologies which give satisfactory results will have to be developed to meet these needs. For example, exploring the availability of administration records (such as, birth and death registration records) for use in regression equations might provide a useful resource in developing small area estimates. Among the areas for which small area estimates are urgently needed at present are (1) for estimating the undercount of the population for the 1980 Census and (2) for statistical series used for allocation of Federal funds to local areas.

Although the result presented in this paper need to be analyzed in greater depth and further developed, I would like to end by thanking the authors for having tackled a difficult problem and for providing useful pointers for other researchers trying to obtain cost-effective and good quality small area estimates.

References

DiGaetano, Ralph, Ellen MacKensie, Joseph Waksberg and Richard Yaffe (1980), "Synthetic Estimates for Local Areas from the Health Interview Survey," ASA Proceedings of Survey Research Methods Section.

Gonzalez, Maria Elena and Christine Hoza (1978), Small-Area Estimation with Application to Unemployment and Housing Estimates, JASA, pp. 7-15.

Jordan, Lawrence A., Alfred C. Marcus and Leo G. Reeder (1980), Response Styles in Telephone and Household Interviewing: A Field Experiment, Public Opinion Quarterly, pp. 210-222.

Roman, Anthony M. (1980) Memorandum, "Addendum to the Methods Development Survey (MDS) Phase I Report," January 18, 1980, Bureau of the Census.