SYNTHETIC ESTIMATES FOR LOCAL AREAS FROM THE HEALTH INTERVIEW SURVEY *

Ralph DiGaetano and Joseph Waksberg, Westat, Inc.
Ellen MacKenzie, Johns Hopkins U. and Richard Yaffe, Health Care Financing Administration

1. INTRODUCTION

Under the National Health Planning and Resources Development Act of 1974, Health Systems Agencies are required to develop a five-year Health Services Plan (HSP) that covers overall goals and long-range objectives, and an Annual Implementation Plan (AIP) that outlines specific activities for the coming year. To carry out these activities, the HSA's seek relatively recent and reliable data on the health status and needs of the community as well as about their patterns of health services utilization. The Planning Act specifically requires the HSA's to collate and analyze data which are currently available.

The source of data relevant to the planning activities of the HSA's that has generated interest is the NCHS Health Interview Survey (HIS). The Health Interview Survey (HIS) collects data from a continuing nationwide probability sample of the nation's households. Information is available concerning illness, injuries, impairments, disability and the utilization of health services for the civilian, non-institutionalized population of the United States. The sample design and size (approximately 40,000 households per year), however, permit reliable estimates to be calculated only for the U.S. as a whole, for four broad geographic regions and perhaps for certain large standard metropolitan statistical areas (SMSA's). The sample size is not sufficient to allow reliable estimates to be made on health variables for most HSA's or sub-HSA areas. This problem has been recognized for some time and there has been considerable developmental work done on statistical procedures that can be used to develop estimates for small areas using national data sources. One such procedure, synthetic estimation, has been used by NCHS to develop state estimates of disability and utilization of medical services from the HIS data.[1] Other analysts have used multiple regression to generate small-area data.[2] The utility of these statistical methods for health planning at the local (HSA) level remains largely untested. Neither synthetic nor regression estimates applied to local areas are unbiased, and the extent to which they are biased will affect their utility for planning purposes.

The National Center for Health Statistics is engaged in assessing the applicability of these techniques for imputing estimates of HIS variables from national or regional data for small areas and has awarded a contract for such an evaluation to the Health Services Research and Development Center of the Johns Hopkins Medical Institutions. Westat, Inc. is acting as a subcontractor to Johns Hopkins for the Study. This paper contains a description of the methods used to evaluate the estimating techniques and a preliminary analysis of the results available to date. This includes an evaluation of the quality of synthetic and regression estimates through an examination of the HIS data alone. In addition, the paper contains further assessments of these estimates made through comparisons with a random digit dialing telephone survey of about 2,500 households in the Baltimore SMSA. The phone survey was carried out in conjunction with this project.

2. METHODS - COMPUTING SYNTHETIC AND REGRESSION ESTIMATES

The technique of synthetic estimation involves applying national or regional estimates of the characteristic being measured for specific population subgroups to the local area's population composition. The simplest form of synthetic estimation, and the one for which the name is usually reserved, requires computation of a weighted average of the mean values of the characteristics in the subgroups with weights that are proportional to the distribution of the subgroups in the small-area population. A more general approach involves regression analysis. In this approach, the national or regional data are used to estimate a regression equation which relates the independent variables which define the population subgroups to the characteristic of interest. The values of the regression variables for the small area are then used in the equation to obtain estimates of the characteristic for that small area.

In the study discussed in this paper, both techniques described above have been used to derive estimates for 40 key health variables selected from the basic HIS questions and the supplemental questions for 1976-1978. The selection of variables was dictated by data requirements of the HSA and the need to have an adequate range of different types of variables for which the use of synthetic estimation could be evaluated.

Simple synthetic estimates for these 40 dependent variables are derived for a basic set of demographic variables (age, sex and race). Additional independent variables used to obtain regression estimates are of two types: variables that are only available in Census years or for which estimates might be available from other surveys or an inexpensive survey, and variables obtained from the Area Resource File. Variables used in the present analysis include proportion of people in a PSU in households where the head of household completed high school, proportion of persons in a PSU who are heads of households and also are either farm or blue collar workers, proportion of persons in a PSU over 65, proportion of non-white population in a PSU, per capita income in a PSU, number of hospital beds per 100,000 population in a PSU, proportion of people in a PSU who are over 17 years old and married, and number of MD's per 100,000 population in a PSU. Estimates were prepared using all 356 Primary Sampling Units (PSU's) in the national HIS and have been applied to the six counties in the Baltimore SMSA as well as the 20 largest SMSA's.

3. EVALUATION METHODS

Three methods are being used to evaluate the various estimates: (1) Comparison of the results with a telephone survey in the Baltimore HSA (and counties within it) with the various synthetic or regression estimates for the same areas; (2) Comparison of the synthetic and regression estimates for individual PSU's with the direct HIS estimates for the same areas; and (3) Calculation of average mean square errors of the synthetic and regression estimates.

Comparison with Telephone Survey

Since the telephone survey concentrated on a group of items that were also collected in HIS, direct comparisons are possible for synthetic and regression estimates of these items with statistics for the same items from the telephone survey. This is a straightforward method of evaluation. For each item studied, the comparison with the survey estimate serves as a guide to the accuracy of the synthetic or regression estimate. Although such a comparison provides important information on the accuracy of the estimate, it is subject to several limitations. First, it assumes that HIS results are comparable to those from telephone interviewing (more specifically, the particular procedures used in the Baltimore telephone survey). Secondly, the time periods are different. There are sizeable seasonal variations for some of the statistics which complicate the comparisons. Finally, such a comparison is only possible for the Baltimore HSA and for counties within it. The extent to which the Baltimore experience is typical of other areas in the United States is uncertain.

We believe these limitations do not seriously affect the resulting analyses. Other studies have indicated that in most cases telephone surveys produce data quite similar to personal interviews. In regard to seasonal factors, some information on seasonal variation is available from the HIS. A later report will attempt to adjust for the seasonal differences.

It may be helpful to detail some particulars of the telephone survey relative to HIS. The phone survey attempted to simulate HIS to as great a degree as possible (using the same questions, training procedures for interviewers, etc.). The major differences were: (1) Only one respondent was used per family within a household in the phone survey, this respondent providing information on all other family members. HIS encourages every adult in the family to participate in a group session, as it is an "in-person" interview arrangement. (2) Some questions in HIS require the interviewer to show cards to respondents. For the phone survey, cards were mailed to some respondents after initial contact as an experiment. For the 60-65 percent of the respondents who did not use cards, some HIS questions had to be modified. (3) Not all HIS questions for any one year were used in the phone survey. A single interview by phone required approximately 30 minutes, while an HIS interview requires approximately one hour. (4) Non-telephone households are naturally excluded from a phone survey. (5) Clusters of households using the random digit dialing design of the phone survey differ in nature from the clusters of households on a city-block approach used in HIS. (6) Interviewers in the phone survey were closely monitored and there existed a great deal more communication among interviewers working out of a central location than is possible with HIS.

The response rate for the phone survey was 76 percent. There were 2,470 households in this survey consisting of 7,013 people. There were 15 primary interviewers in the study, each logging at least 100 interviewing hours. Information on approximately 1,200 people was obtained from the Baltimore PSU for HIS. HIS uses a single interviewer for the Baltimore PSU.

Comparison with Direct HIS Estimates

In the larger PSU's, the HIS sample size is sufficient to provide the data with fair reliability. For the 20 largest SMSA's in the United States, we have compared direct HIS estimates with those prepared for the same areas using synthetic or regression techniques. For the regression, this is equivalent to examining the distances the observed values are from the regression values.

Average Mean Square Error

The evaluation method described above suffers from three qualifications: (1) It can only be applied to the larger PSU's. The situation for smaller, largely rural, PSU's may be quite different. (2) In making comparisons for a group of areas, there are bound to be variations among the areas in the amount of difference between the direct estimate and the synthetic or regression estimate. A method is needed of summarizing the results so that a conclusion can be reached on whether or not the estimates are satisfactory. (3) The difference between a direct HIS estimate and a synthetic or regression estimate reflects two sources of error: (a) the inaccuracy of the synthetic or regression estimate; and (b) sampling error in the HIS estimate. It is desirable to eliminate the effect of the HIS sampling error in the overall evaluation.

The average mean square error (AMSE) overcomes these three limitations. Using synthetic estimation terminology, the average mean square error is defined as

$$E \frac{1}{M} \sum_{i}^{M} (u_i' - U_i)^2 \qquad (1)$$

where u_i, is the synthetic estimate for area i; U_i is the true value in area i; and M is the number of areas.

The AMSE can be thought of as having characteristics similar to those of sampling variances. That is, the chances will be about two out of three that the synthetic estimate will be equal to the true value plus or minus the square root of the AMSE; the chances are 19 out of 20 that the range within which the synthetic estimate appears will be plus or minus twice the square root of the AMSE, etc.

Of course, in practical situations the value of U_i is not known. Gonzalez and Waksberg[3] have shown that the AMSE of a rate per person can be estimated by

$$\frac{1}{M} \sum_{i} \left[\sum_{j} p_{ij} (u_j - u_{ij}) \right]^2 - \sum_{j} \sum_{i} P_{ij}^2 \sigma_{ij}^2 \qquad (2)$$

where j is an index for the sex-age-etc. groups used in the synthetic estimates; P_{ij} is the population proportion in the i^{th} PSU, in the j^{th} sex-age-etc. category; u_j is the survey estimate of the rate per person in j^{th} demographic group; u_{ij} is the survey estimate of the rate per person in the jth demographic group in the i^{th} PSU; and

σ^2_{ij} is the sampling variance for the item, within the i,jth category.

Calculations of the AMSE have been carried out for the items for which synthetic estimates are prepared.

A similar type of analysis can be made for regression estimates.[4] With regression estimates, the sum of squares of the residuals from the line of regression replaces the first term of equation (3.2). The second term remains the same.

4. AVAILABLE RESULTS

The necessary computations have been completed for 21 of the 40 items in the program, and basic information on the quality of the synthetic and regression estimates for these items are shown in the attached tables. Similar information for the other 19 items will become available at a later time. Even for the 21 items, the discussion and explanation for the statistics that have been produced should be considered preliminary. Further analysis of the data is continuing, and the additional work that is planned, described in the next section, may shed new light on the results.

However, even with the limited analyses done to date some conclusions appear clear, and we believe it is unlikely that they will be revised when the additional information becomes available. The main conclusion is that there is considerable variation in the quality of the estimates among the health-related items studied, and for many of the items neither synthetic nor regression estimates produce very reliable data for areas of the size of typical HSA's. At least this is true with the techniques used for this project. The errors are probably even larger for areas the size of counties, although further evidence is needed on this. The extent to which such data can be used for policy analysis and decisions depends, of course, on the degree of accuracy needed for these uses. The implications of high sampling errors with respect to the applicability of synthetic estimates is a subject which needs to be dealt with separately. Such issues are beyond the scope of this paper.

Before describing the data leading to these conclusions, let us give the specific estimating techniques used, in somewhat more detail than described earlier. The synthetic estimates were prepared by calculating national rates per person separately by race-sex-age, and applying them to the best estimates of the population by race-sex-age in each local area. The race-sex-age classifications consisted of: Race: White vs. non-White; Sex: Male vs. Female; and Age: Under 15, 15-44, 45-64, 65 and over.

The population estimates were the most current estimates prepared by the Census Bureau. At the time this work was done, the Census estimates were for 1977. In addition, for the Baltimore SMSA, other population estimates prepared by the Baltimore Regional Planning Council were also obtained, and formed the basis of alternative synthetic estimates.

For regression estimates, nine independent variables were used. They were: (1) Synthetic estimates for the area (using Census population estimates); (2) Mean per capita income in 1975 (also Census estimates); (3) Percent of blue collar workers; (4) Percent married and 17 years and over; (5) Percent completed high school; (6) Percent 65 years old and over; (7) Percent non-white; (8) Number of MD's per 100,000 persons; and (9) number of hospital beds per 100,000 persons.

As is common in multiple regression, in general, only a few independent variables made an important contribution to the model, and those are the only ones that were eventually used to create estimates.

Table 1 compares synthetic and regression estimates of each of the 21 items for the Baltimore SMSA with both the results of the telephone survey and the direct HIS estimates for Baltimore. Synthetic and regression estimates are fairly close; the two, of course, are not independent since the synthetic estimate variable was usually one of the independent variables making an important contribution to the regression. For many items, synthetic and regression estimates are quite close to the results of the telephone survey. However, there are quite wide differences in a few cases. Differences of 20 to 25 percent are not unusual, and there is a difference of 50 percent for one item (visits to emergency rooms per person per year). These differences are generally far beyond the possible effects of sampling error.

However, a surprising feature of Table 1 is that there are even greater differences between the results of the telephone survey and the direct HIS estimates for Baltimore. They are also beyond any reasonable effects of sampling errors. There seems to be no obvious explanation of these differences. Some part of the differences could be due to the fact that the direct HIS covered the year 1977 while the telephone survey was conducted during the last few months of 1979 and January 1980. It does not seem likely that there are enough changes in health characteristics over this period to account for much of the differences. There is definite seasonal variation for some of the items studied, and this probably explains more of the differences, but it still is far from accounting for most of it. We thought it possible that there might be major differences in the age-sex-race composition of the telephone and direct HIS samples, due to a combination of sampling variation and differential response rates and that this could be a partial explanation. However, as can be seen in Table 2, such differences did not occur.

The HIS conducts interviews on a face-to-face basis, but we doubt that the differences in interviewing techniques contribute importantly to the differences. The question wording in the two interviews was essentially identical. The differences are quite puzzling, but as we will indicate later, we do not believe they vitiate the use of the telephone survey as an evaluation tool of regression and synthetic estimates.

Table 3 shows data similar to Table 1, but for each county in the Baltimore SMSA. (Because of space limitations, Table 3 as presented here consists only of three items. However, the analysis has been done in reference to all 21 items.) For most items, the synthetic and regression estimates are roughly similar to the results of the telephone survey. However, they do not seem to discriminate among counties well. For instance, if one ranked the various counties by size of the estimates, for most items rankings of the telephone survey would not conform very closely to synthetic or regression estimates. There are a few items, however, for which the synthetic and regression estimates come closer to the results of the telephone survey. These are generally items with large differences between the Black and White population. For such items Baltimore city data are quite different from the rest of the SMSA, and these differences persist for all estimators.

Table 4 shows major characteristics of the regression estimator. As indicated earlier, although the regression computations started with nine independent variables, a much smaller number was actually used for most items. A step-wise regression program was initially utilized, with all nine variables. For each item, a smaller number of variables accounting for virtually the entire R^2 were selected and used to prepare the estimates.

The variables used for each item are shown in the second column of Table 4. We were surprised by the variables that show up as important for most items. Synthetic estimates appear as an important variable for only about half the items. We would have expected it to be more prominent; Number of hospital beds per 100,000 population is an important variable for some of the hospital-related statistics, but not all. Demographic characteristics such as percent of blue-collar workers and percent with a high-school education appear more often than we would have expected.

The next to last column shows the contribution each variable makes to the total regression estimate. Where synthetic estimates appear as a variable, it is usually the dominant variable, frequently (although not always) accounting for 70 or 80 percent of the part of the estimates added to the intercepts. This makes it even more puzzling that it does not appear for more items. It is possible that intercorrelations among variables complicate the choice of dominant variables. We have not yet had the opportunity to examine them.

The last column of Table 4 shows the R^2 for each item, and it can be seen they are quite low. The highest R^2 is .30, and there are a few that are about .20. The rest are lower. The low values of R^2 explain the poor ability of regression estimates to simulate the telephone survey in Baltimore. It is possible that a model which includes interaction terms or non-linear relationships may work better. Such models were not examined in this study.

Table 5 shows the reason for the similar poor predictive ability for the synthetic estimates. The root average mean square error has been expressed as a proportion of the estimate. The results are shown in the last column. The relative root mean square error can be thought of as the analogue of the coefficient of variation of a sample survey.

The relative errors are generally in the range of .2 to .5. A few are as high as 1.0. A relative error of .5 implies that when synthetic estimates are prepared for a set of areas, one can expect approximately one-third of the areas to have an error of more than 50 percent of a census value.

Other studies have shown important regional differences for some types of health characteristics. It is possible that using regional parameters, rather than those for the total United States, may improve synthetic or regression estimates, or both. If resources permit, they will be examined in a later phase of the project.

Table 6 contains further insight on the poor predictive power of the synthetic and regression estimates. This table contains both types of estimates, as well as the direct HIS estimates for the largest 20 SMSA's. Somewhat more than 20 areas are shown because several of the largest SMSA's have been split up into subareas. We have selected only a few of the 21 items to keep the table to a reasonable size, but the other items show similar patterns.

It can be seen that the range of variation among areas is much narrower for synthetic and regression estimates than for direct estimates. Furthermore, if one were interested in ranking the areas by size for an item, in order to identify the higher-valued or lower-valued areas, synthetic and regression would generally not simulate the results of sample surveys. As was the case in Baltimore, the differences cannot be attributed to sampling error. These results are consistent with the findings of other studies.[5]

Table 6 contains some other information which appears to be even more surprising than the poor performance of synthetic and regression estimates. Synthetic and regression estimates for the 20 areas appear to have a very small range of variation due to similarities in the independent variables among the large metropolitan areas. However, the direct estimates seem to encompass a much wider range than one would expect. For example, if one looks at the number of visits to a doctor's office per person per year, the direct HIS estimates go from a low of 1.80 for a PSU of New York to a high of 5.39 for a PSU of Philadelphia. The differences cannot be explained by different demographic compositions of the areas or differences in the characteristics used for the regressions. If they were, then synthetic and regression estimates would have better explanatory power. They are also far beyond the limits of sampling. These are the largest self-representing PSU's in HIS and have fairly large sample sizes.

The data seem to imply that, for the items studied, areas are inherently very different. This would explain why predictors based on demographic or economic information, such as the synthetic and regression estimates utilized in this study, do not have much power. However, the large differences among areas appear surprising. The items selected are of a kind that one would think are mostly quite stable. Some of the differences among the areas are no doubt due to inherent geographic variation. However, the dramatic nature of these differences suggest that there may be problems in HIS ability to enforce uniform standards of interviewing. In most of the areas, the HIS interviews were carried out by only a few interviewers and between-interviewer variability may be quite high. This conjecture would help explain the large differences between the direct HIS and the telephone survey in the Baltimore SMSA. HIS, a vehicle designed primarily for obtaining national estimates on health-related data, apparently does not provide direct estimates which are stable enough for small-area estimation needs.

If problems in the HIS are the major reasons for the differences, then one can take a somewhat different attitude towards synthetic and regression estimates. Measured against a standard of the accuracy of data actually achievable in a survey such as the HIS, synthetic and regression estimates may be of acceptable quality for most practical uses. Further analysis in this direction is necessary.

5. PLANNED FUTURE ANALYSES AND OTHER POSSIBLE APPROACHES

The analysis discussed above, done on the 1977 HIS data, will also be done for the years 1976 and 1978. This will allow us to observe the sensitivity of the parameters to sample size. We also plan to repeat the same analyses for all three years combined. In addition, there are several sets of items which are only available for a particular one of the years 1976-1978. For example, health insurance information is available for the 1976 HIS. Some of these items for each of the three years (approximately 20 items altogether) will be examined.

Further research on synthetic estimates would be useful in several areas. First, it would be interesting to examine whether the introduction of an additional cross-classification of the sex-race-age groups produces a significant improvement in the synthetic and regression estimates. Two categorical variables which could be considered are: degree of urbanization (SMSA's over 1,000,000 population, smaller SMSA's, and non-SMSA's) and Census region (Northeast, North Central, South, or West). An urbanization-region cross-classification could be introduced. Again, the data could be evaluated by calculation of the appropriate mean square errors and comparison with results of the telephone survey in Baltimore, Maryland.

Second, it would be informative to calculate synthetic and regression estimates and average mean square errors for specific population subgroups within PSU's. These could be, for example: Sex: female; Age: 17-44, 65+; and Race: Non-white. Statistics for each dependent variable could be calculated for each of these subgroups.

Other areas which warrant investigation include: (1) The use of HIS income information in the regression models; (2) The effect of models, to be examined by excluding PSU's from regressions and comparing the results to those from a complete data base; (3) The validity of the assumption of linearity in the regression models; (4) The preparation of estimates based on varying sizes of PSU's; (5) The utilization of past years' estimates as predictors of the current year's estimates and the examination of other methods for assessing autocorrelative effects; and, (6) The construction of composite estimators consisting of a weighted average of a synthetic (or regression) estimator and a direct estimator.

*Supported in part by Contract No. 223-78-2052 from the National Center for Health Statistics.

FOOTNOTES

[1]National Center for Health Statistics: Synthetic State Estimation of Disability. PHS Publication #1759, Public Health Service, Washington, D. C. 1968.
Namekata, T., Levy, P. S., and O'Rourke, T. W.; Synthetic Estimates of Work Loss Disability for Each State and the District of Columbia. Public Health Report 90:532-538, 1975.
National Center for Health Statistics: Synthetic Estimation of State Health Characteristics based on the Health Interview Survey, by P. S. Levy and D. K. French. Vital and Health Statistics. Series 2, No. 75. DHEW Pub. No. (HRA) 78-1349. Health Resources Administration, Washington, D. C., U. S. Govt. Printing Office, October 1977.

[2]Gonzalez, M.E. and Hoza, C.: Small Area Estimation with Applications to Unemployment and Housing Estimates, Journal of the American Statistical Association, 73, 1978.

[3]Gonzales, Maria and Joseph Waksberg, "Estimation of the Error of Synthetic Estimates", prepared for presentation at the first meeting of the International Association of Survey Statisticians. Vienna, Austria, August 18-25, 1973.

[4]Gonzalez, Maria, and Hoza, C., "Small Area Estimation with Applications to Unemployment and Housing Estimates", Journal of the American Statistical Association, 73: 1978.

[5]Schaible, Wesley; Brock, Dwight; and Schnack, George A., National Center for Health Statistics; "An Empirical Comparison of the Simple Inflation, Synthetic, and Composite Estimators for Small-Area Statistics", American Statistical Association Proceedings of the Social Statistics Section, 1977, Part II, pp. 1017-1021.

Table 1. Comparison of alternative estimates for the Baltimore SMSA

Item and area	Telephone survey		Direct HIS estimate				Synthetic estimate		Regression estimate
			Estimate		σ		Based on Maryland Population figures	Based on Census Population figures	
	Estimate	Approximate σ	SMSA	U.S.	SMSA	U.S.			
Restricted activity days per person per year	18.15	.61	10.34	17.78	1.44	.25	17.97	18.11	20.29
Bed disability days per person per year	7.58	.56	4.31	6.87	.90	.14	6.99	7.06	7.65
Work loss days per person per year	4.11	.40	2.97	2.12	.82	.06	3.10	3.05	3.22
School loss days per person per year	1.39	.19	.65	1.05	.24	.04	.96	1.01	.88
Proportion limited in activity	.162	.005	.119	.135	.013	.001	.134	.135	.138
Proportion unable to carry on major activity	.040	.003	.033	.033*	.008	.001	.036	.037	.036
Proportion with limitation of activity for a duration of 1 year or longer	.134	.005	.106	N.A.	N.A.	N.A.	.115	.116	.118
Number of doctor visit per person per year (annual recall)	3.30	.15	N.A.	N.A.	N.A.	N.A.	3.65	3.67	3.72
Proportion of people with one or more doctor visits in the last year	.746	.006	.766	.752**	.018	.002	.739	.740	.736
Number of dental visits per person per year	1.88	.137	1.38	1.6	.23	.027	1.54	1.53	1.64
Number of short stay hospital episodes per 100 persons per year	12.58	.60	8.58	14.0	5.97	1.05	13.10	13.09	12.09
Number of short stay hospitals days per 100 persons per year	105.37	7.94	83.14	109.20	18.97	2.51	103.20	103.92	118.85
Average length of stay in a hospital	8.38	.74	9.69	7.8	7.03	.61	N.A.	N.A.	5.85
Proportion of persons with one or more hospital episodes in the last year	.108	.004	.074	.104	.007	.001	.105	.104	.093
Visits to doctor's office per person per year	2.878	.150	2.377	3.48*	.301	.045	3.289	3.290	3.570
Visits to Emergency room per person per year	.164	.034	.314	N.A.	N.A.	N.A.	.245	.245	.241
Visits to out patient clinic per person per year	.616	.088	.835	N.A.	N.A.	N.A.	.484	.489	.526
Visits to general practitioners per person per year	1.942	.125	1.640	2.545*	.309	.049	2.480	2.485	2.166
Visits to selected practitioners per person per year	2.81	.172	2.765	3.90*	.348	.051	N.A.	3.709	3.818
Visits for diagnosis or treatment per person per year	3.843	.201	3.200	4.289*	.381	.052	4.110	4.125	4.620
Visits for chronic condition per person per year	1.713	.152	1.885	2.556	.355	.049	2.131	2.143	2.208

*1974 Estimates
** 1975 Estimates

Table 2. Sex-Race-Age distribution for U.S. and Baltimore SMSA based on telephone survey and HIS estimates

Sex-race-age	U.S. estimate from HIS	Baltimore estimate from HIS	Telephone survey: Baltimore estimate before adjustment	Telephone survey: Baltimore estimate after adjustment
White male: <15	10.32	9.34	8.03	8.02
White male: 15-44	19.03	17.99	16.74	17.93
White male: 45-64	8.77	8.27	8.44	7.72
White male: 65+	3.91	3.00	2.68	3.12
Non-white male: <15	2.08	2.92	3.11	3.24
Non-white male: 15-44	2.74	4.10	5.30	5.92
Non-white male: 45-64	.98	2.23	1.96	1.92
Non-white male: 65+	.43	.92	.45	.66
White female: <15	9.85	9.98	8.15	7.47
White female: 15-44	19.69	16.88	18.98	17.73
White female: 45-64	9.52	9.47	8.69	8.31
White female: 65+	5.59	3.85	3.42	4.65
Non-white female: <15	2.05	3.00	3.80	3.34
Non-white female: 15-44	3.30	5.14	6.85	6.65
Non-white female: 45-64	1.16	2.09	2.57	2.35
Non-white female: 65+	.57	.84	.85	.97

Table 3. Comparison of alternative estimates for Baltimore SMSA and component counties

Item and area	Telephone survey: Estimate	Telephone survey: Approximate σ	Synthetic estimate: Based on Maryland population figures	Synthetic estimate: Based on Census population figures	Regression estimate
Restricted activity days per person per year					
Total SMSA	18.15	.608	17.97	18.11	20.29
Anne Arundel	20.40	1.877	16.82	16.86	18.27
Baltimore City	19.77	1.001	19.40	19.74	23.15
Baltimore County	17.57	1.132	17.67	17.55	19.30
Carroll	15.20	1.947	17.08	17.43	15.81
Harford	11.64	1.094	16.42	16.37	18.00
Howard	14.23	1.821	16.23	15.87	18.96
Bed disability days per person per year					
Total SMSA	7.58	0.561	6.99	7.06	7.65
Anne Arundel	6.99	1.290	6.44	6.51	6.15
Baltimore City	9.03	1.022	7.75	7.90	9.87
Baltimore County	7.87	1.145	6.74	6.66	6.78
Carroll	4.59	1.350	6.52	6.65	5.08
Harford	4.85	1.033	6.26	6.37	5.82
Howard	3.93	0.902	6.26	6.11	6.43
Work loss days per person per year					
Total SMSA	4.11	0.404	3.10	3.05	3.22
Anne Arundel	5.56	1.282	3.05	2.89	3.05
Baltimore City	4.94	.735	3.28	3.31	3.30
Baltimore County	3.32	.670	2.98	2.94	3.31
Carroll	2.68	.985	2.85	2.87	3.34
Harford	1.71	.641	2.98	2.70	2.95
Howard	2.76	.834	3.01	2.90	2.94

Table 4. Characteristics of regression estimates

Item	Independent Variables	Coefficients	Estimate of independent variables for Baltimore SMSA	Contribution to total estimate for Baltimore SMSA	R^2
Proportion of people with one or more hospital episodes in last year	Intercept	.1260			.143
	% high school ed.	-.0420	.5911	.35	
	# of hospital beds per 100,000 population	.000046	452.708	.29	
	# of M.D.'s per 100,000 population	-.0001	258.394	.36	
Number of short-stay hospital episodes per 100 persons per year	Intercept	18.8794			.130
	% high school ed.	-9.5157	.5911	.46	
	# of hospital beds per 100,000 population	.0061	452.708	.23	
	# of M.D.'s per 100,000 population	-.0152	258.394	.31	
Number of hospital days per 100 persons per year	Intercept	-278.78			.109
	Synthetic est.	5.3676	103.20	.77	
	% high school ed.	-78.281	.5911	.07	
	% over 65	-967.32	.0939	.13	
	% non-white	-76.631	.2506	.03	
Average length of stay in hospital	Intercept	3.5210			.087
	% blue collar	-9.7882	.0947	.22	
	# of M.D.'s per 100,000 population	.0047	258.394	.29	
	% over 65	21.7070	.0939	.49	
Restricted activity days per person per year	Intercept	-13.840			.048
	Synthetic est.	2.578	17.97	.79	
	% blue collar	24.588	.0947	.05	
	% high school ed.	-8.382	.5911	.08	
	% over 65	-52.290	.0939	.08	
Bed disability days per person per year	Intercept	-9.4183			.098
	Synthetic est.	3.1067	6.99	.83	
	% blue collar	-12.1567	.094	.04	
	% high school ed.	-3.6770	.591	.08	
	% over 65	-14.0559	.093	.05	
Work loss days per person per year	Intercept	-2.8525			.022
	Synthetic est.	1.3834	3.10	.70	
	% blue collar	7.5429	.094	.12	
	Per capita income	.0002	5339.26	.18	
School loss days per person per year	Intercept	.5357			.002
	Synthetic est.	.3718	.96	.65	
	# of hospital beds per 100,000 population	.0002	452.708	.16	
	# of M.D.'s per 100,000 population	-.0004	258.394	.19	
Proportion of people limited in activitiy	Intercept	.1190			.195
	Synthetic est.	.9705	.134	.54	
	% blue collar	-.2354	.094	.09	
	% high school ed.	-.0636	.591	.16	
	Per capita income	-.00000965	5339.26	.21	
Proportion of people unable to carry on major activity	Intercept	.0659			.295
	Synthetic est.	1.0897	.036	.37	
	% blue collar	-.1254	.094	.11	
	% high school ed.	-.0404	.591	.22	
	Per capita income	-.00000616	5339.26	.30	
Proportion of people with limitation of activity one year or longer	Intercept	.1002			.201
	Synthetic est.	1.0362	.115	.44	
	% blue collar	-.2097	.094	.36	
	% high school ed.	-.0534	.591	.02	
	Per capita income	-.00000934	5339.26	.18	
Number of doctor visits per person per year	Intercept	-3.1413			.043
	Synthetic	1.7096	3.65	.87	
	% blue collar	-2.4144	.0947	.03	
	% non-white	.5780	.2506	.02	
	Per capita income	.0001	5339.26	.08	
Proportion of people with at least one doctor visit	Intercept	.5790			.187
	% high school ed.	.1469	.5911	.55	
	% non-white	.0678	.2506	.11	
	Per capita income	.00001	5339.26	.34	
Number of dental visits per person per year	Intercept	-.4352			.163
	% high school ed.	1.0760	.5911	.31	
	Per capita income	.00027	5339.26	.69	
Number of doctor visits in doctor's office per person per year	Intercept	1.1522			
	% married > 17	2.4374	.4477	.35	
	% high school ed.	-.5691	.5911	.11	.037
	% non-white	.5514	.2506	.04	
	Per capita income	-.000285	5339.26	.49	
Number of doctor visits in emergency room per person year	Intercept	.6941			
	% married > 17	-1.3580	.4477	.72	
	% high school ed.	.1967	.5911	.14	.044
	% blue collar	.8574	.0947	.10	
	# of Hospital beds per 100,000 population	-.0000939	452.708	.05	
Number of doctor visits in out clinic per person per year	Intercept	.3584			
	Synthetic est.	.8004	.398	.18	
	% blue collar	-1.6811	.0947	.20	.121
	% over 65	-2.0935	.0939	.26	
	# M.D.'s per 100,000 population	.000794	358.394	.36	
Number of visits to general practitioners per person per year	Intercept	-.9497			
	Synthetic est.	1.4914	2.32	.75	
	% high school ed.	-1.0206	.5911	.13	.059
	% blue collar	4.1333	.0947	.09	
	% non-white	-.5295	.2506	.03	
Number of visits to selected practitioners per person per year	Intercept	2.1513			
	% high school ed.	-.7778	.5911	.23	
	% blue collar	3.5755	.0947	.11	.015
	# of Hospital beds per 100,000 population	.000256	452.708	.07	
	Per capita income	.000312	5339.26	.59	
Number of visits for diagnosis or treatment per person per year	Intercept	3.0038			
	% over 65	-2.1623	.0939	.11	
	% non-white	.6569	.2506	.09	.046
	# M.D.'s per 100,000 population	.00113	258.395	.16	
	Per capita income	.000225	5339.26	.64	
Number of visits for chronic condition per person per year	Intercept	.3146			
	Synthetic est.	.4275	1.879	.42	
	% non-white	.6127	.2506	.08	.039
	# M.D.'s per 100,000 population	.000716	258.395	.10	
	Per capita income	.000146	5339.26	.40	

Table 5. Average Mean Square Error of Synthetic Estimates

Item	Estimate (U.S.)	Average mean square error	Root mean square error	Relative root mean square error
Restricted activity days per person per year	17.8	68.86	8.30	.466
Bed disability days per person per year	6.9	11.55	3.40	.493
Work loss days per person per year	2.12	4.51	2.12	1.00
School loss days per person per year	1.05	.9853	.993	.946
Proportion limited in activity	.135	.0018	.042	.311
Proportion unable to carry on major activity	.033*	.0004	.020	.606
Proportion with limitation of activity for a duration of 1 year or longer	NA	.0015	.039	N.A.
Number of doctor visits per person per year (annual recall)	NA	1.04	1.02	N.A.
Proportion of people with one or more doctor's visits in last year	.752**	.0043	.066	.088
Number of dental visits per person per year	1.6	.7217	.850	.531
Number of short stay hospital episodes per 100 persons per year	14.0	33.41	5.78	.413
Number of short stay hospital days per 100 persons per year	109.2	3045.11	55.18	.505
Proportion of persons with one or more hospital episodes in the last year	.104	.0011	.033	.317
Visits in doctor office, per person per year	3.466*	1.3710	1.171	.338
Visits in emergency room, per person per year	NA	.1413	.376	N.A.
Visits in out-patient clinic, per person per year	NA	.1999	.47	N.A.
Visits to general practitioners, per person per year	2.570*	1.3466	1.160	.451
Visits to selected practitioner, per person per year	3.575*	1.5054	1.227	.343
Visits for diagnosis or treatment, per person per year	4.3301*	2.0018	1.415	.327
Visits for chronic condition, per person, per year	2.581*	.9090	.953	.369

* 1974 Estimate
** 1975 Estimate

Table 6. Comparison of alternative estimates for 20 largest SMSA's for selected items
Item: Bed Disability Days Per Person Per Year

Area and PSU	Direct HIS Estimate	Direct HIS σ	Synthetic estimate	Regression estimate
Chicago				
308 392 (combined)	6.33	.73	6.93	6.95
Los Angeles				
702 762 (combined)	7.63	.88	6.92	6.77
Boston (Suffolk Co. only)				
116	6.73	2.46	7.11	7.40
Philadelphia				
111	6.45	1.03	7.20	7.59
181	7.29	2.26	6.76	6.43
New York				
110	8.44	2.23	7.71	9.51
190	13.23	2.52	7.28	8.39
192	8.46	2.32	7.36	8.88
193	6.10	1.36	7.40	8.27
194	6.79	1.08	6.78	6.30
Detroit				
309	11.17	1.69	6.89	6.85
San Francisco				
703	7.49	1.29	7.03	6.95
Washington, D.C.				
511	7.21	2.72	8.16	10.85
541	4.92	1.55	6.41	5.88
542	5.60	1.58	6.53	6.08
Dallas				
503	8.77	2.02	6.63	6.01
St. Louis				
306	7.60	1.73	7.04	7.05
386	9.59	4.14	6.99	7.16
Pittsburg				
115	5.63	1.14	7.08	6.94
Houston				
509	5.50	1.07	6.50	6.05
Baltimore				
510	4.31	.90	7.06	7.83
Minneapolis				
302	5.71	1.28	6.51	5.09
Newark				
195	10.68	2.46	7.15	7.53
Cleveland				
307	7.12	1.56	7.09	7.22
Atlanta				
508	6.43	1.59	6.69	6.54
Anaheim				
719	4.61	1.06	6.41	5.40
San Diego				
709	6.60	1.62	6.82	6.25

Table 6. Comparison of alternative estimates for 20 largest SMSA's for selected items (Continued)
Item: Proportion of People Limited in Activity

Area and PSU	Direct HIS Estimate	Direct HIS σ	Synthetic estimate	Regression estimate
Chicago				
308 392 (combined)	.100	.006	.133	.124
Los Angeles				
702 762 (combined)	.124	.007	.136	.126
Boston (Suffolk Co. only)				
116	.092	.018	.122	.146
Philadelphia				
111	.129	.011	.144	.140
181	.146	.024	.129	.125
New York				
110	.181	.025	.161	.149
190	.151	.015	.141	.156
192	.161	.023	.142	.165
193	.091	.011	.157	.142
194	.093	.008	.135	.119
Detroit				
309	.131	.010	.132	.123
San Francisco				
703	.153	.014	.140	.121
Washington, D.C.				
511	.143	.028	.149	.141
541	.092	.015	.116	.102
542	.109	.016	.118	.097
Dallas				
503	.081	.010	.121	.111
St. Louis				
306	.115	.014	.138	.134
386	.173	.039	.136	.141
Pittsburg				
115	.112	.012	.148	.144
Houston				
509	.133	.014	.116	.108
Baltimore				
510	.119	.013	.135	.138
Minneapolis				
302	.150	.018	.124	.106
Newark				
195	.148	.018	.141	.130
Cleveland				
307	.096	.011	.140	.133
Atlanta				
508	.131	.017	.121	.113
Anaheim				
719	.092	.011	.122	.109
San Diego				
709	.140	.018	.133	.126

Table 6. Comparison of alternative estimates for 20 largest SMSA's for selected items (Continued)
Item: Proportion of people with at least one doctor visit in past year

Area and PSU	Direct HIS Estimate	Direct HIS σ	Synthetic estimate	Regression estimate
Chicago				
308 392 (combined)	.734	.010	.742	.752
Los Angeles				
702 762 (combined)	.731	.010	.742	.753
Boston (Suffolk Co. only)				
116	.716	.030	.743	.724
Philadelphia				
111	.787	.014	.742	.744
181	.804	.029	.745	.741
New York				
110	.811	.024	.741	.757
190	.764	.017	.742	.731
192	.731	.023	.742	.710
193	.714	.018	.744	.734
194	.755	.013	.745	.760
Detroit				
309	.780	.013	.741	.746
San Francisco				
703	.791	.016	.740	.772
Washington, D.C.				
511	.731	.031	.731	.788
541	.774	.028	.746	.776
542	.819	.026	.742	.787
Dallas				
503	.711	.019	.742	.758
St. Louis				
306	.771	.020	.744	.745
386	.781	.038	.745	.728
Pittsburg				
115	.701	.016	.746	.728
Houston				
509	.743	.017	.739	.749
Baltimore				
510	.766	.018	.740	.738
Minneapolis				
302	.796	.020	.745	.765
Newark				
195	.731	.019	.742	.760
Cleveland				
307	.751	.019	.743	.742
Atlanta				
508	.766	.022	.739	.764
Anaheim				
719	.744	.019	.744	.759
San Diego				
709	.769	.022	.748	.755

Table 6. Comparison of alternative estimates for 20 largest SMSA's for selected items (Continued)
Item: Number of visits to doctors office per person per year

Area and PSU	Direct HIS Estimate	Direct HIS σ	Synthetic estimate	Regression estimate
Chicago				
308 398 (combined)	3.1064	.22	3.3279	3.6054
Los Angeles				
702 762 (combined)	3.7735	.26	3.2628	3.5119
Boston (Suffolk Co. only)				
116	2.3000	.51	3.3765	3.1312
Philadelphia				
111	4.3027	.43	3.3829	3.4700
181	5.3912	1.04	3.3894	3.2547
New York				
110	3.5824	.58	3.4168	3.8855
190	3.2768	.39	3.3170	3.2082
192	3.2148	.55	3.3041	3.0630
193	1.8038	.25	3.4945	3.6836
194	3.7432	.37	3.4537	3.6926
Detroit				
309	4.1211	.39	3.3128	3.4753
San Francisco				
703	4.4083	.47	3.3573	3.7307
Washington, D.C.				
511	3.3262	.78	2.9987	3.8079
541	4.1166	.44	3.3712	3.7008
542	3.7907	.66	3.2719	3.8643
Dallas				
503	3.2234	.46	3.2825	3.5703
St. Louis				
306	3.1481	.44	3.3801	3.4510
386	4.1039	1.09	3.3840	3.3405
Pittsburg				
115	2.4009	.30	3.5291	3.4076
Houston				
509	3.0075	.36	3.2181	3.5950
Baltimore				
510	2.3775	.31	3.2965	3.5556
Minneapolis				
302	3.3466	.46	3.4124	3.4372
Newark				
195	3.8510	.55	3.3640	3.6296
Cleveland				
307	2.7351	.37	3.3958	3.5199
Atlanta				
508	3.0500	.47	3.2142	3.5746
Anaheim				
719	3.6602	.52	3.3939	3.5986
San Diego				
709	4.1989	.64	3.7607	3.3700