THE USE OF MORTALITY DATA IN EVALUATING SYNTHETIC ESTIMATES

Paul S. Levy, National Center for Health Statistics

1. Background and Introduction

For the past 15 years, the National Health Survey through its Health Interview, Health Examination and other surveys has been gathering data on a considerable range of health parameters. The statistics obtained from these surveys are expressed generally as estimates for the United States as a whole or for each of the four broad geographical regions. Estimates for smaller areas such as States cannot be readily obtained because (1) the sample sizes are not adequate for areas this small and (2) the sampling design uses strata which cut across State lines, and this makes it difficult to combine the estimates for strata into estimates for States.

Because of the increasing need for small area estimates of health parameters the National Center for Health Statistics (NCHS) has been exploring methodology for using National Health Survey data to produce estimates for these small areas. This research has resulted in a method called *synthetic estimation* whereby State estimates are produced by using census data on the distribution of a State's population into demographic domains (which we will subsequently refer to as "α-cells") along with the National Health Survey estimates for these domains for the entire United States. An NCHS publication [1] describes this synthetic estimation methodology and uses it to produce synthetic estimates of disability for each State from Health Interview Survey (HIS) estimates for the United States.

Since there is an ever increasing need for small area estimates of health parameters for purposes of local planning and since good local population data essential for synthetic estimation, are now available from the 1970 Decennial Census, an NCHS program to produce small area estimates based on synthetic methodology is seriously being considered. It is first necessary, however, to evaluate the accuracy of the estimates produced by this methodology, and this report presents the results of a study designed to gain insight into the accuracy of synthetic estimates as well as a method which would use ancillary data to obtain an improved synthetic estimate.

2. The NCHS Synthetic Estimate

While the NCHS publication [1] presents the methodology in greater detail, we will outline the methodology below. The synthetic estimate is constructed in two stages with the first stage having the form

$$\tilde{\tilde{X}}_s = \sum_{\alpha=1}^{K} p_{s\alpha} \bar{X}_\alpha \qquad (1)$$

Where

$\tilde{\tilde{X}}_s$ = the first stage synthetic estimate of X-characteristic for the s-th State

$P_{s\alpha}$ = proportion of the population in the s-th State belonging to the α-th demographic cell

$\bar{X}_\alpha$ = the unbiased probability estimate of the X-characteristic for the α-th demographic cell for the U.S. as a whole

and

K = the number of demographic cells.

The final estimate has the form

$$\tilde{X}_s = \tilde{\tilde{X}}_s \left(\frac{\bar{X}'_r}{\sum_{s=1}^{R} \tilde{\tilde{X}}_s P_{rs}} \right) \qquad (2)$$

where $\tilde{X}_s$ = the final synthetic estimate of X-characteristic for State s,

$\bar{X}'_s$ = the unbiased probability estimate of X characteristic for the r-th geographical region

P_{rs} = the proportion of the population of region r belongs to the s-th State

and

R = the number of States in region r.

The basic feature of the first stage synthetic estimator is that for each demographic cell (α-cell) it uses a probability estimate specific to that demographic cell over the entire United States in conjunction with the proportion of the State's population falling into that cell (from census data) to obtain the contribution of that cell to the synthetic estimator. The final synthetic estimate is simply the initial estimate times a post-stratification factor which makes the sum of the synthetic estimates for all States in a geographical region agree with the probability estimate for that region.

Synthetic estimates applied to National Health Survey data have been difficult to evaluate because valid unbiased estimates produced by ordinary probability survey methods are not available for States. Indirect methods such as observing their consistency from one year to another [1] or comparing the synthetic estimates for the largest SMSA's with the HIS unbiased estimators [2] have been used but these have not been altogether convincing.

Since mortality statistics from U.S. Vital Statistics Annual Volumes [3] are available by cause of death for all States by age, sex, and race, a study was planned to compute synthetic estimates for each State for several causes of death using the U.S. mortality data for age, race, and sex α-cells and the corresponding census data on the distribution of each State into these α-cells. If the synthetic estimates should agree well with the true deaths, it would be evidence that the synthetic procedure might produce valid estimates. If not, then some insight might still be gained into the pattern of discrepancy between the synthetic estimate and the true value. This evaluation study is described in the next section.

3. Evaluation Study

3.1 *Methods and Materials*

Mortality data. Deaths from motor vehicle accidents (E810-E835), major cardiovascular renal diseases (330-334, 400-468, 592-594), suicide (E963, E970-E979) and tuberculosis (001-019) for each of forty-nine States (with the District of Columbia and Maryland combined into one "State") were transcribed onto IBM cards from the 1960 U.S. Vital Statistics Annual Volumes [3]. New Jersey was not included because deaths by race were not available for that State. For each State, the deaths were transcribed for the following 40 age-sex-race groups:

Age (under 1 year, 1-4 years, 5-9 years
10-14 years, 15-19 years, 20-24 years, 25-44 years
45-64 years, 65-74 years, 75+)

Race (white, nonwhite)
Sex (male, female).

Population data. The populations in each of the above mentioned age-sex-race groups were transcribed onto IBM cards from the 1960 U.S. Census Volumes for each State in the study.

Data Processing. A program was written in Fortran IV and the analysis was carried out on a remote entry terminal to the UNIVAC 1108 at the National Bureau of Standards in Gaithersburg, Maryland. The basic output was the synthetic estimate, $\tilde{X}_s$ of the number of deaths by specific cause for each State in 1960.

3.2 *Results*

The agreement between the synthetic estimate and the true number of deaths is expressed by the *percentage absolute difference* defined as

$$\left| \frac{X_s - \tilde{X}_s}{\tilde{X}_s} \right| \times 100 \qquad (3)$$

where

X_s = the true number of deaths (by specific cause) for the s-th State

and

$\tilde{X}_s$ = the second stage synthetic estimate of these deaths (obtained from the 40 age-sex-race α-cells).

The results are summarized in Table 1 which gives the frequency distribution of these percentage absolute differences along with the median and mean percentage absolute difference, and the relative variances (V^2 = variance/mean^2) of the percentage absolute differences for each of the four causes of death considered.

The accuracy of the synthetic estimates as measured by the percentage absolute difference varied considerably among the four causes of death examined. The median percentage absolute difference was 5.9% for major cardiovascular-renal deaths, 9.8% for suicides, 15.9% for deaths from motor vehicle accidents, and 24.3% for tuberculosis deaths. Likewise, the variability among States as expressed by the relative variance of the percentage absolute difference was highest for tuberculosis with $V^2 = 1.11$ and lowest for major cardiovascular diseases ($V^2 = 0.57$). The relative variance for deaths from motor vehicle accidents was equal to 1.00 and for suicide was 0.65.

3.3 *Discussion*

From the results presented above and shown in Table 1, it is clear that the accuracy of the synthetic estimates as summarized by the median percentage absolute difference varied considerably among the four causes of death considered. While the agreement between the synthetic estimate and the true value was generally good for major cardiovascular-renal diseases and fairly good for suicide, it was generally poor for motor vehicle accidents and very poor for tuberculosis deaths.

In order to get some insight into the effectiveness of the synthetic estimator against possible competing estimators, we compared it to a regionally adjusted estimator, X'_{as} obtained for each State, S, by multiplying the population of the State by the crude death rate in the geographical region wherein the State lies. The percentage absolute difference between X'_{as} and the true number of deaths X_s was calculated for each State and the median percentage absolute difference over all States was obtained and is shown below next to the comparable figure for the synthetic estimator, $\tilde{X}_s$.

Median Percentage Absolute Difference

Estimator	Motor Vehicle Accidents	Major C.V.R. Diseases	Suicides	Respiratory T.B.
$\tilde{X}_s$	15.9	5.9	9.8	24.3
X'_{as}	14.6	6.3	10.2	32.0

The results above seem to imply that except for respiratory T.B., the synthetic estimate did little or no better than the estimator X'_{as} which is obtained from very crude regional information.

Table 1. Cumulative Percentage Distribution of Percentage Absolute Differences Between the Synthetic State Estimate and the True Number of Deaths ($100 \times | X_s - \tilde{X}_s | / \tilde{X}_s$) for Each of the Four Causes of Deaths Investigated

Percentage Absolute Difference $100 \times \left\| \frac{X_s - \tilde{X}_s}{\tilde{X}_s} \right\|$ %	Frequencies (f) and Cumulative Percentages (cum. %)							
	Motor Vehicle Accidents		Major C.V.R. Diseases		Suicides		Tuberculosis	
	f	cum. %	f	cum. %	f	cum. %	f	cum. %
0- 8.0%	16	32.7	26	53.1	20	40.8%	13	26.5
8.1-16.0%	9	51.0	20	93.9	13	67.3%	6	38.7
16.1-24.0%	9	69.4	3	100.0	6	79.5%	4	46.9
24.1-32.0%	6	81.6			6	91.8%	4	57.1
32.1-40.0%	3	87.8			3	98.0%	9	73.4
40.1% +	6	100.0			1	100.0%	13	100.0
Total	49		49		49		49	
Median % Absolute Difference		15.9		5.9		9.8		24.3
Mean % Absolute Difference		20.2		6.9		13.7		31.6
Relative Variance (V^2)		1.00		0.57		0.65		1.11

4. Improvement by Regression on Ancillary Data

4.1 *Background and motivation*

One of the basic limitations on the synthetic estimator, $\tilde{X}_s$ is that it is adjusted only for the specific set of demographic cells (or a-cells) taken into consideration. If the parameter being estimated is influenced by factors other than those taken in account by the a-cells, then the synthetic estimate will not reflect this influence. Often it is not possible to include in the a-cell array all the variables thought to be of importance in estimating the variable because of the unavailability of data on these variables. For example, even though the risk of a person's dying from a motor vehicle accident may be a function of the amount of time he spends in motor vehicles, there is no way of creating a-cells to reflect this.

A second type of limitation in the synthetic estimator, $\tilde{X}_s$, is that it may not reflect local conditions which are highly related to the parameter being estimated. For example, the probability of a person's dying from a motor vehicle accident is known to be generally higher in States which have lower population densities [4]. Since these types of variables are often available for local areas such as States, it might be possible that a modified synthetic estimate can be constructed which takes into account these variables.

In the following sections, we propose a method of using local variables which might be related to the parameter being estimated in conjunction with the synthetic estimator $\tilde{X}_s$ to produce an improved estimator of the parameter. It is felt that this method will be especially applicable to small area estimates using data from the complex, highly stratified multi-stage nationwide probability surveys such as the Health Interview Survey.

4.2 *Method of estimation*

The method presented below uses the a-cell adjusted synthetic estimate $\tilde{X}_s$ in conjunction with an ancillary variable Z_s to produce an improved estimator. In particular, the following model is assumed:

$$Y_s = a + \beta Z_s + \epsilon_s \qquad (4)$$

where

Z_s = the value of the Z variable for the s-th State

$$Y_s = \frac{X_s - \tilde{X}_s}{\tilde{X}_s} \times 100,$$

$\tilde{X}_s$ = the synthetic estimate of the X-characteristic for the State s,

X_s = the true value of the X-characteristic for State s,

ϵ_s = a term representing random error,

and

a, β = regression parameters to be estimated.

If estimates $\hat{a}$ of a and $\hat{\beta}$ of β were available and substituted into the right hand side of equation (4) with ϵ_s omitted, manipulation of the expression would give us the following estimator $\hat{X}_s$ of X_s:

$$X_s = \tilde{X}_s \left(1 + \frac{\hat{a} + \hat{\beta} Z_s}{100}\right) \qquad (5)$$

Equation (4) merely states that the percentage difference, Y_s, between the synthetic estimate, $\tilde{X}_s$, and the true value X_s is a linear function of some variable Z_s. For example, Z_s might be the population density of State s, $\tilde{X}_s$ the synthetic estimate of deaths from motor vehicle accidents and X_s the true number of deaths for that State. Equation (4) would then state that except for random variation, the percentage difference between the true and synthetic estimates of deaths from motor vehicle accidents for a State is a linear function of its population density.

4.3 *Estimation of* a and β

Since Z_s is assumed to be available for each State, the regression coefficients a and β could be estimated if corresponding values of Y_s were available. The percentage difference, Y_s, however, is a function of the true parameter, X_s, which is not known, as well as the synthetic estimate $\tilde{X}_s$ which can be obtained. If some estimate, X_s' of X_s were available, however, it could be substituted for X_s into the expression for Y_s and estimates of a and β could be obtained from least squares.

One of the problems, however, is that estimates X_s' of X_s are not available for States from the National Health Surveys. One can obtain, however, estimates X_c' of X_c where X_c is the value of characteristic X for the c-th *strata combination.* A *strata combination* is defined here as any unit that can be constructed by combining strata. Since unbiased estimates are available for strata, unbiased estimates X_c' can be obtained for strata combinations. Also, since strata are generally counties or groups of counties, the ancillary variable Z_c can be readily obtained. Likewise, the synthetic estimate, $\tilde{X}_c$ can be obtained for each strata combination. Thus, if we divide the United States into C strata combinations and obtain X_c', Z_c' and $\tilde{X}_c$ in the usual way, we can estimate a and β by least squares from the data pairs (Z_c', Y_c') where

$$Y_c' = \frac{X_c' - \tilde{X}_c}{\tilde{X}_c} \times 100, c = 1, \ldots, C \qquad (6)$$

Once, the estimates $\hat{a}$ and $\hat{\beta}$ are obtained, they can be substituted into equation (5) and estimates $\hat{X}_s$ of X_s can be obtained for each State.

An example of this estimation procedure was constructed from the mortality data on deaths from motor vehicle accidents discussed in Section 3. For the ancillary variable, we let Z represent population density and divided the United States into the 14 State combinations shown in Table 2. From these we obtained the appropriate $\tilde{X}_c$, X_c and Y_c'. The least squares estimates of a and β obtained from these 14 State combinations were $\hat{a} = 12.0626$ and $\hat{\beta} = -.0660$. The correlation between Z and Y as estimated from the 14 sample points was $r = -.6034$. Having obtained $\hat{a}$ and $\hat{\beta}$, the estimated deaths $\hat{X}_s$ were computed for each State and the distribution of percentage absolute differences is shown in Table 3 alongside that for $\tilde{X}_s$. Clearly $\hat{X}_s$ is an improvement over $\tilde{X}_s$ in the sense that the median percentage absolute difference was 10.0 for $\hat{X}_s$ as compared with 15.9 for $\tilde{X}_s$.

5. Some Comments

While the scope of this evaluation study was not large enough to make any final conclusions, about the value and accuracy of synthetic estimation, some extrapolations might prove valuable in planning further studies:

(1) The estimator, $\hat{X}_s$, might be especially suitable to estimate health parameters from National Health Survey Data. Without loss of generality, Z_s might be a vector of ancillary data and the estimator $\hat{X}_s$ would be a multiple regression type estimator. The problem would be to find a set of variables, Z, which might be related to the health characteristic being estimated and different health characteristics would require different sets. There is a wealth of variables available for small areas from the 1970 U.S. Census which might be related to health variables, and this method of estimation could make use of these Census data.

(2) Since there was much variability in the agreement of synthetic estimates with true values not only among States for each cause of death, but also among causes of death with respect to median and mean percentage absolute differences, one might generalize that

Table 2. State Clusters Used to Obtain Regression Coefficients

State Cluster	Population Density (Persons/Mile^2)	Motor Vehicle Deaths		
		Synthetic Estimate $\tilde{X}_c$	True Value X_c	Percentage Difference Y_s $100 \times (X_c - \tilde{X}_c)/\tilde{X}_c$
1. Maine, N. H., Vt.	39.88	267.6	375	40.13
2. Mass., R. I., Conn.	617.98	1154.6	991	-14.17
3. N. Y., Pa.	302.34	3800.8	3857	1.48
4. Ohio, Illinois	204.20	4267.7	3648	-14.52
5. Indiana, Michigan	133.95	2662.2	2884	8.33
6. Wisconsin, Mo., Iowa, Minnesota	55.58	3133.8	3385	8.02
7. N. D., S. D., Neb., Kansas	16.10	1073.4	1220	13.65
8. Del., Md., D. C.	361.90	1071.7	700	-34.68
9. Va., N. C., Fla.	94.13	3378.8	3249	- 3.84
10. W. Va., S. C., Ga., Ky., Tenn., Alabama., Louisiana	73.39	5216.4	5492	5.28
11. Ark., Okl., Texas, Miss.	36.79	3912.2	4140	5.82
12. Alaska, Nevada, Ariz., Montana, Idaho, Wyoming, New Mexico, Utah	4.02	1373.4	1809	31.7
13. Colorado, Oregon, Washington	23.90	1680.7	1556	- 7.42
14. California, Hawaii	100.32	4355.2	4044	- 7.15

$\alpha = 12.0626$

$\beta = -.0660$

Table 3. Distribution of Percentage Absolute Differences for $\hat{X}_s$ and $\tilde{X}_s$ with Respect to Motor Vehicle Accident Deaths 1960

Percentage Absolute Difference	Frequencies (f) and Cumulative Percentages (cp) $\hat{X}_s$		$\tilde{X}_s$	
	f	cp	f	cp
0- 4.0%	10	20.4	7	14.3
4.1- 8.0%	10	40.8	9	32.7
8.1-12.0	9	59.2	6	44.9
12.1-16.0	3	65.3	3	51.0
16.1-20.0	1	67.3	6	63.3
20.1-24.0	5	77.6	3	69.4
24.1-28.0	2	81.6	4	77.6
28.1-32.0	2	85.7	2	81.6
32.1-36.0	0	85.7	2	85.7
36.1-40.0	1	87.8	1	87.8
40.1-44.0	3	93.9	1	89.8
44.1-48.0	1	95.9	0	89.8
48.1+	2	100.0	5	100.0
Total	49		49	
Median % absolute difference	10.0		15.9	
Mean % absolute difference	16.1		20.2	
Relative Variance	1.03		1.00	

great care should be taken in the interpretation of synthetic estimates.

(3) Much more work is needed in the development of methodology to produce estimates of health characteristics for small areas.

REFERENCES

1. U.S. Department of Health, Education, and Welfare: Synthetic State Estimates of Disability – PHS Publication No. 1759. U.S. Government Printing Office (1968).

2. National Center for Health Statistics. Unpublished data (1968).

3. U.S. Department of Health, Education, and Welfare: Vital Statistics of the United States 1960. Volume II – Mortality Part B. U.S. Government Printing Office (1963).

4. Fuchs, V. R. and Leveson, I.: Motor Accident Mortality and Compulsor Inspection of Vehicles, Journal of the American Medical Association, *201*, pages 87-91 (1967).