ASSEMBLY-LINE RESEARCH--WITH QUALITY CONTROL

Herbert I. Sauer, Consultant
909 University Ave., Columbia, Missouri 65201
(and Dept. Family and Community Med., Univ. Missouri)

I. Collecting Data

A substantial volume of data, generally of high quality, has been collected by many agencies, including the Bureau of the Census and the National Center for Health Statistics. These data provide a basis for generating and testing various hypotheses, such as environmental effects upon human health, as well as for mapping.

Collecting data of adequate quality and for a sufficient number of areas for mapping will usually require a greater variety and volume of skills and a larger budget than is available. It is therefore often necessary to utilize data collected by others. As you know, many steps are involved in the work of the Bureau of the Census, from which we obtain the population at risk. Even more steps are involved in the preparation of the detailed death tapes--from clerks in a hospital or undertaker's office, to the certifying physician, the local registrar, the state registrar, state nosologists, and the National Center for Health Statistics.

Data are collected and prepared for many uses--and it is quite possible that some of those conscientiously filling out forms or otherwise participating have done so adequately for most purposes but have possibly not thought about many of the uses of their work. It seems to me to be the responsibility of the investigator, that is, the epidemiologist, the geographer or the statistician presenting the data to recognize that in a sense

he is at the end of an assembly line and is responsible for considering whether his uses of the data are requiring a level of precision higher than is present in the data. Even if costs were not a consideration, it is unreasonable to expect absolute exactness in data collection, making this consideration of precision necessary.

Some of us have purchased new motor vehicles, and then had to return them to the dealer from time to time for correction of specified deficiencies--failures to manufacture the vehicle with sufficient precision for it to function properly. As epidemiologists we have a similar responsibility to be alert to quality control of data at the end of the assembly line.

Our focus in this presentation is upon possibilities for systematic error or bias, which may arise from several sources. Chance fluctuation or standard error, while also extremely important, is not covered in this paper.

II. Low and High Death Rates

The epidemiologist is concerned with identifying epidemics, and in determining their causes. For the chronic diseases it may be more accurate to speak of high endemic rates, as contrasted with other areas or groups with low endemic rates--but this may at times be largely a matter of semantics. An example is our identification of the ten lowest and the ten highest-rate areas, for the cardiovascular diseases, whites age 45-74 (age-sex-adjusted), for 1949-51 (Fig. 1).[1] Areas smaller and more homogeneous than states have been useful--specifically metropolitan areas, the non-metropolitan portions of economic subregions, state economic areas, counties and other standard groupings of counties.[2,3] However, if there are instances of serious systematic error, they are more likely to fall in the "extreme-rate" category, whether very low or very high, than to be rates nearer the median. Therefore an effort is made to continue, from different perspectives, the search for possible bias.

Yet, by using these areas, similar geographic patterns have been identified for different time periods--such as white males age 35-74, for all causes of death, by state economic areas, for the middle 1970's (Fig. 2). Maps have thus enabled us to identify patterns that we

Fig. 1. TEN LOWEST AND TEN HIGHEST DEATH RATE AREAS FOR CARDIOVASCULAR DISEASES, WHITES, AGES 45 - 74, 1949 - 51

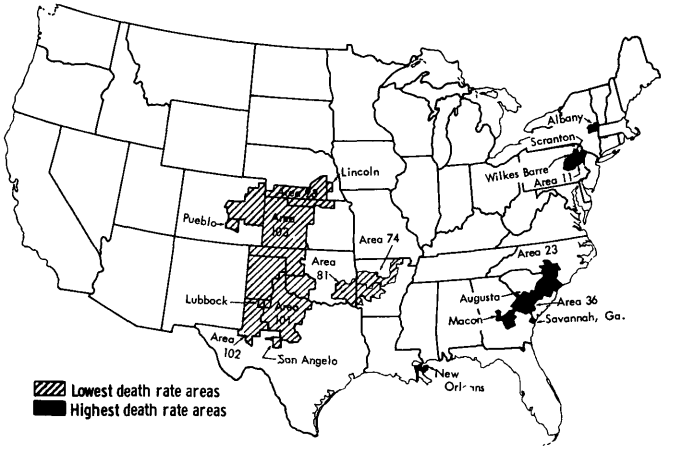
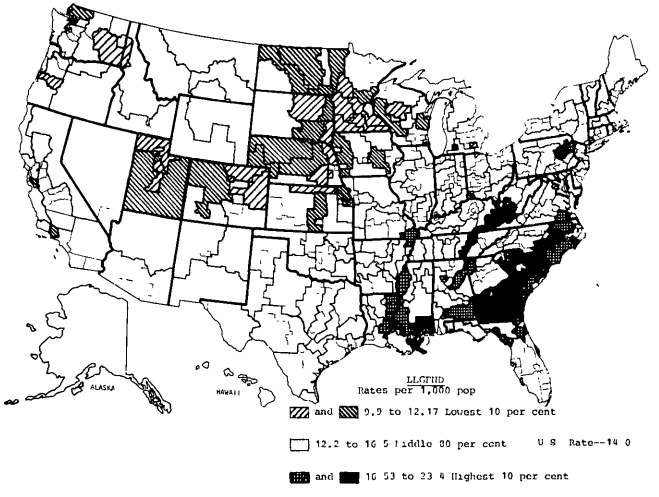


Fig. 2. Death Rates for All Causes, White Males 35-74, 50 Lowest and 50 Highest-Rate State Economic Areas, 1973-1976



otherwise failed to observe--by using broad categories of causes of death, age-race-sex-specific rates, areas smaller than states, and groups of years.

The very substantial degree of consistency in patterns from one time period to another and for different age groups seems to us evidence that the patterns are real--most probably caused by factors in the environment which vary geographically--even though these factors have not yet been adequately identified.[1,4,5] As yet we have been unable to formulate a hypotheses of inaccuracy of data that would explain these geographic differences.

III. Cause of Death

The use of the category "natural causes" or all diseases combined (ICDA 000-796)[6] implies that there has been a reasonably satisfactory distinction between this broad category and the external causes--accidents, suicide and homicide (ICDA E800-E999). These external causes and the broad disease categories, cardiovascular diseases (ICDA 390-448), malignant neoplasms (ICDA 140-209) and "all other diseases" (ICDA 000-138, 210-389, 450-778) are likely to be quite accurate generally.[7] But "Symptoms and ill-defined conditions, ICDA 780-796", with the entry on the death certificate often merely "natural causes", is not really a cause of death, therefore potentially limiting the completeness of other causes. For most areas, however, the number of deaths assigned to this category is so small that, even if ignored, the cause-specific rates will be affected little. But for some areas or groups, these ill-defined deaths constitute ten percent or more of all natural causes deaths. On the basis of available evidence we have concluded that our classification-of-cause error (with present data) can be minimized by including them with the cardiovascular diseases.[8,9]

This and other categories which most need attention for their possible effect on other cause categories include "Other heart diseases, ICDA 420-429" (which is largely symptomatic and ill-defined heart disease) and "Malignant neoplasms of ill-defined and secondary sites, ICDA 195-199". More narrowly defined causes of death are likely to be even more useful for epidemiological study, including mapping--but there is also a greater possibility that there will be slight differences in vocabulary

used in different geographic areas, in certifying cause of death, which may introduce bias.

IV. Age, Race, and Sex

The risk of dying generally increases with age, typified by the experience of black females in Mississippi (Fig. 3). While age-adjusted rates therefore seem necessary, death rates for all ages (age-adjusted) tend to reduce or even obscure differences in risk for middle age.[10] Death rates for ages 35-74, age-adjusted by the direct method by ten-year age groups, tend to produce a maximum contrast in rates consistent with reducing standard error to a minimum.[11] For making comparisons of rates for ages 75 and over, 5-year age groups are more informative. We agree with a distinguished APHA committee, however, that there is no fully adequate substitute for age-specific death rates.[12]

Available evidence indicates that for whites in particular, age is entered accurately on the death certificate, but for nonwhites or blacks, age is frequently not reproducible.[13] The failure of age-specific death rates for blacks in Charleston County, SC, to increase in old age is a sharp contrast with the experience for Mississippi (Fig. 3). This contrast was even more marked for 1949-51 and originally suggested to us the hypothesis of inaccuracy of age (Fig. 4). In a study of deaths in the 1961-63 period, we found that age or birthdate on the death certificate usually did not agree with that on a record prepared while the individual was alive.[14]

Age-specific death rates have generally declined in recent years, as have those for black females in Charleston County under the age of 65, but there has been a clearcut increase in the calculated age-specific rates age 65 and over. Thus Figure 4 suggests to us the hypothesis that accuracy of age has been improving in recent years, and that time trends in age-specific rates may in some instances be misleading. (We are currently working with vital statisticians in South Carolina on this problem.)

A problem concerning race is what ethnic groupings to use. In the western half of the United States especially, "black" (or "Negro") is a better definition of an ethnic group than "nonwhite" or "all other races"

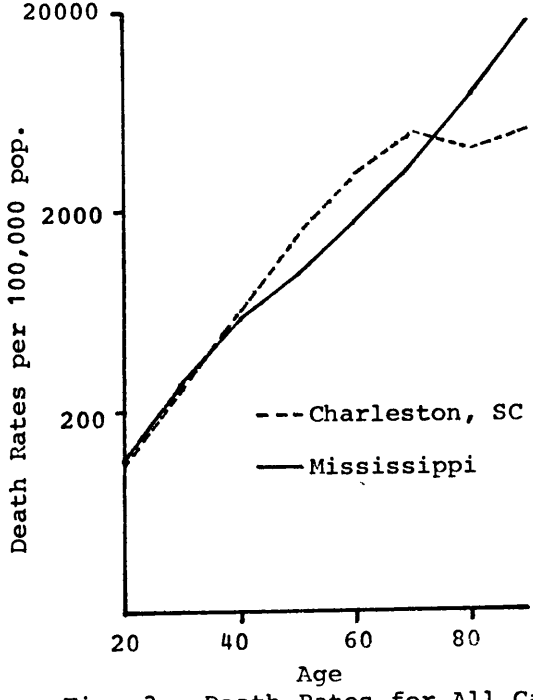


Fig. 3. Death Rates for All Causes, Black Females by Age, Charleston, SC, and Mississippi State, 1968-1972.

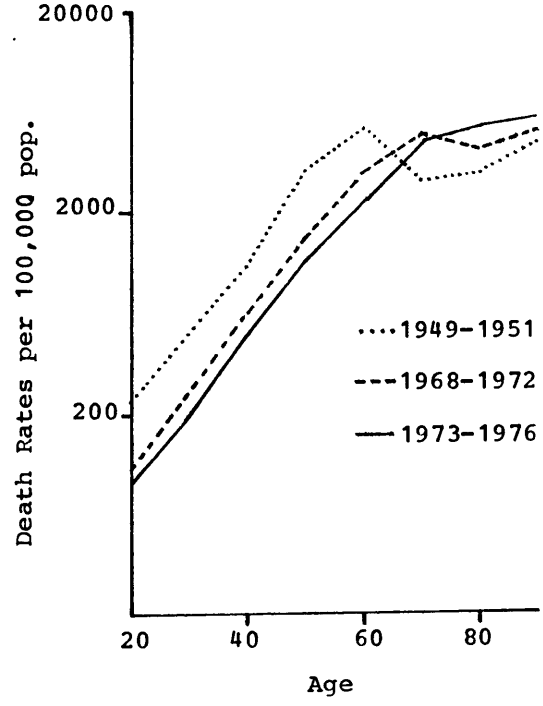


Fig. 4. Death Rates for All Causes, Black Females by Age, Charleston, SC, 1949-1951, 1968-1972, and 1973-1976.

because the latter definition includes ethnic groups that are not at all homogeneous as regards death rates. Among whites, Spanish-Americans tend to have low rates for middle-aged males and in other ways differ from the patterns for other whites. Therefore, in the Southwest and other areas with a substantial proportion of Spanish-Americans, it seems especially important to consider the possible ethnic effect.

V. Geography

Deaths are ordinarily classified by place of usual residence. However, the question is frequently raised as to whether the residents of an area have lived there long enough for their experience to be a measure of the effect of the environment upon human health. While this may be a relevant factor in specific situations, available evidence suggests that migration tends to reduce the amount of contrasts in death rates between states, rather than being the cause of such contrasts.[15] An exception is the low rates for whites of retirement age in Florida areas, which are the result of low rates for those born in the North, now living in Florida.

Residents of mental hospitals and other resident institutions are enumerated by the Bureau of the Census as residents of the county in which the institution is located, whereas for many years decedents were classified as residents of the county from which they were admitted. While the decline in mental hospital populations lessens the effect of resident institutions upon the comparability of death rates, there is still a need for further study of the effect of this factor for some counties.

VI. Summary

The acronym GIGO ("garbage in, garbage out") may be said to imply that data are either perfect or worthless. When we use data prepared by others, it is very easy to assume one of these extremes. Such an assumption is rejected by various sciences in their analytical procedures. Instead, investigators measure or at least estimate the magnitude of error and use feasible methods for reducing error. Obviously, this increases the rigor with which they test hypotheses and interpret data. In the past two decades we have made some progress in applying procedures of this type to mortality

data--particularly as regards comparability of cause of death, age, race, and usual residence of the population at risk. We recognize that much more work is required in order to test more specific hypotheses and to produce more specific maps soundly.

VII. References

1. Sauer, H.I. 1962. AM. J. PUB. HLTH. 52(1):94-105.
2. U.S. Bureau of the Census. 1972. CENSUS OF THE POPULATION: 1970 STATE ECONOMIC AREAS, Final Report PC(2)-10B. U.S. GPO, Washington.
3. U.S. Department of Commerce. 1972. OBERS PROJECTIONS: ECONOMIC ACTIVITY IN THE U.S.: VOLUME 2--BEA ECONOMIC AREAS. U.S. GPO, Washington.
4. Sauer, H.I. 1979. GEOGRAPHIC PATTERNS IN THE RISK OF DYING AND RELATED FACTORS: UNITED STATES 1968-1972. Nat. Center for Hlth. Stat. (NCHS) (in press)
5. _____, R.W. White, and L.E. Reed. 1978. In: TRACE SUBSTANCES IN ENVIRONMENTAL HEALTH-XII, D.D. Hemphill, Ed. (Univ. Mo. Press, Columbia, pp. 62-71)
6. National Center for Health Statistics. 1967. EIGHTH REVISION, INTERNATIONAL CLASSIFICATION OF DISEASES, ADAPTED FOR USE IN THE UNITED STATES, Vol. 1. PHS Pub. No. 1693, U.S. GPO, Washington.
7. Sauer, H.I., and P.E.Enterline. 1959. J. CHRONIC DIS. 10(6):513-524.
8. Spain, D.M., V.A. Bradess and C. Mohr. 1960. JAMA 174:384-388, Sept. 24.
9. Sauer, H.I. 1978. In: GEOCHEMISTRY AND THE ENVIRONMENT, VOLUME III, National Research Council, Washington, pp. 121-125.
10. Erhardt, C.L., and J.E. Berlin, Eds. 1974. MORTALITY AND MORBIDITY IN THE UNITED STATES, pp. 105, 106, 240. Harvard Univ. Press, Cambridge.
11. Sauer, H.I., G.H. Payne, C.R. Council, and J.C. Terrell. 1966. PUB. HLTH REP. 81(5):455-465.
12. Committee on Forms and Methods of Statistical Practice. 1940. AM. PUB. HEALTH ASSOC. YEAR BOOK, 1939-40, Supplement to AM. J. PUB. HLTH. 30:123-131.
13. Hambright, T.Z. 1968. COMPARABILITY OF AGE ON THE DEATH CERTIFICATE AND MATCHING CENSUS RECORD, UNITED STATES, MAY-AUGUST, 1960. PHS Pub. No. 1000-Series 2-No. 29, NCHS.
14. Sauer, H.I. 1966. Presented at the Population Association of America meeting, April 30, 1966.
15. _____, 1967. PROCEEDINGS OF THE SOCIAL STATISTICS SECTION, Am. Stat. Assoc., pp. 399-407.