This research supported by The Occupational Health Studies Group
and NIEHS Training Grant #5-T01-ES00133.

ON THE UTILITY OF PROPORTIONAL MORTALITY ANALYSIS

L.L. Kupper [1], A.J. McMichael[2], and M.J. Symons[1]

Departments of Biostatistics[1] and Epidemiology[2]
Occupational Health Studies Group
School of Public Health
University of North Carolina
Chapel Hill, North Carolina 27514

Institute of Statistics Mimeo Series NO. 988

March 1975

ABSTRACT

The epidemiologic evaluation of longitudinal data typically involves a comparison of two rates. In mortality studies, this comparison utilizes the Standardized Mortality Ratio (SMR), the observed death rate from a specified cause divided by an (age-adjusted) expected death rate. To make such a comparison between absolute death rates necessitates that the size and demographic composition of the "denominator" population-at-risk (PAR) be known. However, the information on the PAR is sometimes lacking, necessitating a comparison of proportional mortality rates. Proportional mortality analysis is generally regarded as suspect because of the inability of relative measures to provide information about absolute rates. However, it is shown in this paper, both theoretically and empirically, that the age-standardized proportional mortality ratio (SPMR) for a specified cause of death provides a close approximation to the SMR for that same cause, when that SMR is expressed relative to the underlying "force of mortality" (i.e., the SMR for all causes of death). It is demonstrated, in situations where longitudinal mortality data accumulates without adequate knowledge of the PAR, that the SPMR can be used to approximate the CSMR ("corrected" SMR) with a degree of accuracy expressible in the form of a statistical confidence interval.

1. GENERAL CONSIDERATIONS

The basic strategy of epidemiology is the comparison of health-related data between two populations. Frequently, the comparison is between a test population (i.e., the population being studied) and some appropriate standard population. The procedure typically utilizes the ratio of observed events to expected events, the latter calculated in "standardized" fashion with respect to such concomitant variables as age, race and sex.

In the analysis of cause-specific mortality data, the comparison can be carried out in either an absolute or a relative context. The former entails the use of the absolute death rates due to the specified cause; the latter involves the relative frequencies of deaths due to the specified cause out of all deaths. Calculation of an absolute death rate within the test population requires knowledge of both the number of deaths due to the specified cause and of the "denominator" population-at-risk (PAR); calculation of a proportional mortality rate necessitates knowing the number of deaths due to the specified cause and the total number of deaths due to all causes, but requires no information on the size of the PAR.

Comparison of the observed absolute death rate with the expected rate demonstrates directly whether the number of deaths due to the specified cause in the test population is more or less than that expected on the basis of the mortality experience of the standard population. Comparison of the observed proportional mortality rate with the expected proportional rate demonstrates only whether the proportion of deaths due to the specified cause is more or less than expected.

Clearly, a cause-specific proportional mortality rate will be influenced by a change in either the number of deaths due to the specified cause or in the number of deaths due to all other causes. That is, whereas the

analysis of absolute cause-specific death rates can be made without reference to death rates from other causes, the analysis of proportional mortality rates necessarily includes reference to deaths from all other causes. This fact underlies the well-recognized limitation of proportional mortality analysis -- namely, that the observation of a relative excess (or deficit) of deaths in the test population does not necessarily indicate an absolute excess (or deficit) in the death rate due to that same cause. (See, for example, McMahon and Pugh, 1970, pp. 59-60; Gilliam, 1954; Moss et al., 1972.) Table 1 below illustrates this point.

TABLE 1: Hypothetical Mortality Data for White Males

	Data Description	Test Population	Standard Population
1.	Stomach Ca. deaths	30	200
2.	All other deaths	970	9,800
3.	Total deaths	1,000	10,000
4.	Population-at-risk	100,000	500,000
5.	Proportion of all deaths due to stomach cancer	0.03	0.02
6.	Death rate for stomach cancer	3/10,000	4/10,000
7.	Death rate for all other causes	97/10,000	196/10,000
8.	Death rate for all causes	100/10,000	200/10,000

Assuming, for simplicity's sake, that the age composition of the two white male populations is identical, then the four rates (items 5-8) can be directly compared. Clearly, the proportion of deaths due to stomach cancer

is higher in the test population than in the standard population (0.03 vs. 0.02), and yet the absolute death rate for stomach cancer is lower in the test population than in the standard population (3/10,000 vs. 4/10,000).

The summary statistic most often used to compare observed and expected death rates, across the full age-range of interest, is the Standardized Mortality Ratio (SMR). The SMR is the ratio of deaths observed in the test population to deaths expected, this latter figure calculated by applying the schedule of age-specific death rates in the standard population to the age-specific PAR's in the test population. Given the assumption of identical age composition for the two populations in Table 1, the following two SMR's can be calculated:

$$\text{Stomach cancer:} \quad \text{SMR} = \frac{3/10{,}000}{4/10{,}000} = 0.75.$$

$$\text{All causes:} \quad \text{SMR} = \frac{100/10{,}000}{200/10{,}000} = 0.50.$$

Now, as mentioned above, the SMR for stomach cancer can be calculated directly, without knowledge of items 7 and 8 in Table 1. However, herein lies a limitation in the meaningfulness of a cause-specific SMR: it necessarily lacks any relativity to the underlying (all-causes) force of mortality within the test population. In other words, simply to say that, compared to the standard population, there is a 25% deficit of stomach cancer deaths in the test population would be to ignore the fact that, since there is an overall mortality deficit of 50% in the test population (i.e., all-causes SMR = 0.50), the death rate from stomach cancer in the test population is higher than would be expected on the basis of the overall death rate. In order to express the actual stomach cancer mortality experience of this population relative to its overall mortality experience, a "corrected" SMR

(or CSMR) can be calculated as:

$$\text{Cause-specific CSMR} = \frac{\text{Cause-specific SMR}}{\text{All-causes SMR}} .$$

For example, the CSMR for stomach cancer (Table 1) is

$$\text{CSMR} = \frac{0.75}{0.50} = 1.50.$$

Note that this procedure produces a mortality ratio that is essentially relative, and therefore one that no longer expresses the absolute deviation of the observed cause-specific death rate from the expected rate. However, the loss of absolute information resulting from this correction procedure is offset by the following considerations:

1. The correction procedure is an addition to, not a substitute for, the initial calculation of a regular SMR. (In practice, this adjustment is often made reflexively by the person scanning a schedule of SMR's that includes the all-causes and various cause-specific SMR's.)
2. In reality, the difference between the regular and corrected SMR's is usually much less than in the above contrived example. The all-causes SMR lies within the range 0.80-1.25 for most test populations. The correction procedure therefore simply makes clearer the meaningfulness of the observed mortality experience from a specific cause.
3. Because of the difficulty in obtaining an "ideal" standard population, the regular cause-specific SMR is often of dubious meaning, since it implies that the value of 1 is the proper baseline figure for comparison purposes. For instance, in the field of industrial

occupational epidemiology, it is common practice to compare the test population to some general community population (often the national population). This is typically a matter of necessity rather than choice. Consequently, because of the "healthy worker" selection process (whereby, to be employable on the production line, a person must be fairly healthy and active), the mortality experience of an industrial population in an industry free of serious mortality hazard usually results in an overall SMR of about 0.80-0.95. Hence, it is really this overall SMR that should be regarded as the "baseline" against which cause-specific SMR's should be evaluated.

In a way exactly analogous to the age-standardization procedure used to obtain an SMR, a standardized proportional mortality ratio (SPMR) can be calculated from the sets of age-specific mortality data for two populations. The SPMR is the ratio of the total observed deaths from a specified cause within the test population to the sum of the age-specific expected deaths (each calculated by multiplying the age-specific proportional mortality rate for the specified cause in the standard population by the age-specific total deaths in the test population). For example, given the assumption of identical age composition for the two populations in Table 1, the stomach cancer SPMR is calculated as 0.03/0.02 = 1.50.

Because of the previously mentioned pitfalls of "numerator" analysis, proportional mortality rates have been used sparingly in mortality studies. For this reason, the SPMR, as a formal summary statistic, has been largely overlooked in the literature. Whereas the analysis of absolute mortality rates typically turns upon the calculation of SMR's, the analysis of proportional mortality rates appears to have not yet assumed any procedural orthodoxy.

Different authors treat "numerator" data in various ways (e.g., Gilliam, 1954; Doll, 1958; Adelstein, 1972; Lloyd et al., 1972), but the potential of the SPMR as an important index of mortality appears to have been largely overlooked.

Note from the above calculations for stomach cancer that the SPMR has exactly the same value as the CSMR. That is, the SPMR, despite the absence of information on the PAR, appears to provide a short-cut to the calculation of the CSMR, whose usefulness has been described above. Of course, the Table 1 data, with its assumption of population identicality with respect to age distribution, is unrealistic. What, therefore, is the general relationship between the SPMR and the CSMR?

In Section 2 below, an analysis of several sets of mortality data, for which all the requisite information was available, indicates <u>empirically</u> that the SPMR is a good approximation to the CSMR. It is also demonstrated that it is possible to place approximate confidence bounds upon the estimated value of the CSMR using the SPMR. It is therefore relevant to identify the circumstances in which the SPMR would be of utility as a central tool in mortality analysis.

The basic circumstance is both obvious and common -- the situation wherein mortality data is available, but the population source cannot be quantified. This situation might arise in several ways. Firstly, death certificates may have accumulated historically over a number of years in a well-defined population setting (e.g., an industrial population), for which insufficient information is available for "re-constructing" the original PAR. (See, for example, Moss et al., 1972.) Secondly, deaths might be occurring in a general community setting for which it is not possible to determine accurately the size and composition of the PAR.

2. STATISTICAL CONSIDERATIONS

It is the purpose of this section to quantify the relationship between the SPMR and CSMR, with specific attention being given to a theoretical examination of why the SPMR empirically seems to be a very useful approximation to the CSMR.

The following notation will be used throughout. Consider a specified cause of death (i, say) and, for standardization purposes, suppose that there are g age groups, with $n_{.j}$ the size of the j-th age group PAR. Let

$d_{i.}$ = observed number of deaths due to cause i in the test population,

$d_{.j}$ = observed number of deaths in the j-th age group of the test population,

$d_{..} = \sum_{j=1}^{g} d_{.j}$ = total observed number of deaths in the test population,

p_{ij} = cause i death rate for the j-th age group of the standard population,

$p_{.j}$ = all-causes death rate for the j-th age group of the standard population.

Then, the usual computational formulae for $SPMR_i$ and $CSMR_i$ can be written as follows:

$$SPMR_i = \frac{d_{i.}}{\sum_{j=1}^{g} a_{ij} d_{.j}},$$

where $a_{ij} = \frac{p_{ij}}{p_{.j}}$ = proportion of all deaths due to cause i in the j-th age group of the standard population; and,

$$CSMR_i = \frac{SMR(\text{cause } i)}{SMR(\text{all causes})} = \frac{d_{i.} / \sum_{j=1}^{g} n_{.j} p_{ij}}{d_{..} / \sum_{j=1}^{g} n_{.j} p_{.j}} .$$

To facilitate the subsequent analysis and to emphasize the similarity in structure between $SPMR_i$ and $CSMR_i$, it is convenient to re-write the above expressions in the following equivalent forms:

$$SPMR_i = \frac{d_{i.}}{d_{..} \sum_{j=1}^{g} a_{ij} \hat{\omega}_j} \qquad (1)$$

and

$$CSMR_i = \frac{d_{i.}}{d_{..} \sum_{j=1}^{g} a_{ij} \omega_j} , \qquad (2)$$

where $\hat{\omega}_j = \frac{d_{.j}}{d_{..}}$ and $\omega_j = n_{.j} p_{.j} / \sum_{j=1}^{g} n_{.j} p_{.j}$ are, respectively, the <u>observed</u> and <u>expected</u> proportions of deaths occurring in the j-th age group of the test population. The notation suggests that $\hat{\omega}_j$ is, in some sense, an approximation to ω_j. Note that $\sum_{j=1}^{g} \hat{\omega}_j = \sum_{j=1}^{g} \omega_j = 1$.

It follows from (1) and (2) for purely mathematical reasons that $SPMR_i = CSMR_i$ either when the $\{a_{ij}\}$ do not vary with j or when $\hat{\omega}_j = \omega_j$ for every j. This latter condition requires that the all-causes SMR's calculated for each age group, the $\{d_{.j}/n_{.j}p_{.j}\}$, must be equal (but not necessarily equal to 1).

Since neither of the above sufficient conditions will hold exactly in actual practice, further study of the utility of the $SPMR_i$ as an approximation to the $CSMR_i$ has led us to an examination of the inequality

$$\left|\frac{CSMR_i}{SPMR_i} - 1\right| < k_i, \tag{3}$$

or, equivalently from (1) and (2),

$$\left|\frac{\sum_{j=1}^{g} a_{ij}\hat{\omega}_j}{\sum_{j=1}^{g} a_{ij}\omega_j} - 1\right| < k_i.$$

It is our goal to specify k_i independently of the $\{n_{.j}\}$, which are unknown quantities when information on the PAR is unavailable.

For completeness, it is worthwhile mentioning that one can find inequalities of the form (3) which are wholly <u>deterministic</u>, e.g.,

$$\left|\frac{CSMR_i}{SPMR_i} - 1\right| \leq \frac{\sum_{j=1}^{g} a_{ij}\ MAX(\hat{\omega}_j, 1-\hat{\omega}_j)}{MIN_j\ a_{ij}}$$

However, such inequalities appear to be much too crude and are generally uninformative in actual practice.

Thus, we turn next to a <u>probabilistic</u> statement of the form

$$Pr\left\{\left|\frac{CSMR_i}{SPMR_i} - 1\right| < k_i\right\} \geq (1-\alpha), \tag{4}$$

or equivalently,

$$Pr\{(1-k_i)SPMR_i < CSMR_i < (1+k_i)SPMR_i\} \geq (1-\alpha). \tag{5}$$

The above statement is exactly in the form of a confidence interval for $CSMR_i$.

In order to proceed further along these lines, we will assume that, conditional on the total number of deaths $d_{..}$, the $\{d_{.j}\}$ are multinomially distributed with parameters $\{\omega_j\}$. If

$$\underset{\sim}{a}_i' = (a_{i1}, a_{i2}, \ldots, a_{ig}),$$

$$\underset{\sim}{\hat{\omega}}' = (\hat{\omega}_1, \hat{\omega}_2, \ldots, \hat{\omega}_g),$$

and

$$\underset{\sim}{\omega}' = (\omega_1, \omega_2, \ldots, \omega_g),$$

then the above assumption implies that

$$E\left(\sum_{j=1}^{g} a_{ij}\hat{\omega}_j\right) = E(\underset{\sim}{a}_i'\underset{\sim}{\hat{\omega}}) = \underset{\sim}{a}_i'\underset{\sim}{\omega}$$

and

$$\mathrm{Var}\left(\sum_{j=1}^{g} a_{ij}\hat{\omega}_j\right) = \underset{\sim}{a}_i' \, V \, \underset{\sim}{a}_i,$$

where $d_{..} V = \mathrm{diag}(\omega_1, \omega_2, \ldots, \omega_g) - \underset{\sim}{\omega}\underset{\sim}{\omega}'$. It can be shown that a good computational formula for $\mathrm{Var}(\underset{\sim}{a}_i'\underset{\sim}{\hat{\omega}})$ is

$$\underset{\sim}{a}_i' \, V \, \underset{\sim}{a}_i = \frac{1}{d_{..}}\left[\sum_{j=1}^{g} a_{ij}^2\omega_j - \left(\sum_{j=1}^{g} a_{ij}\omega_j\right)^2\right];$$

the estimator $\underset{\sim}{a}_i' \, \hat{V} \, \underset{\sim}{a}_i$ of $\underset{\sim}{a}_i' \, V \, \underset{\sim}{a}_i$ is then obtained by using $\hat{\omega}_j$ for ω_j in the above formula.

The multinomial model we are proposing certainly suggests itself in this situation; but, as with any statistical model, it is only expected to provide a reasonable approximation to, and not an exact representation of, the true state of affairs. One saving feature is that we are actually dealing

with a linear combination of random variables (which are constrained to sum to unity), and such a linear combination could be expected to be robust to aberrations in its individual components. The examples we shall presently consider seem to support this contention.

Without making any further assumptions, it is now possible to obtain a statement of the form (4) using Tchebysheff's Theorem (e.g., see Cramér, p. 183). If X is a random variable with mean μ and variance σ^2, this theorem says that

$$\Pr\{|X-\mu| < t\sigma\} \geq 1 - \frac{1}{t^2} \text{ for any } t > 0.$$

Under our model, $\dfrac{CSMR_i}{SPMR_i} = \dfrac{\underset{\sim}{a}_i'\hat{\underset{\sim}{\omega}}}{\underset{\sim}{a}_i'\underset{\sim}{\omega}}$ has mean 1 and variance $\underset{\sim}{a}_i' \, V \, \underset{\sim}{a}_i/(\underset{\sim}{a}_i'\underset{\sim}{\omega})^2$, so that for $t = \alpha^{-\frac{1}{2}}$ we have

$$\Pr\left\{ \left| \frac{CSMR_i}{SPMR_i} - 1 \right| < \alpha^{-\frac{1}{2}} (\underset{\sim}{a}_i' \, V \, \underset{\sim}{a}_i)^{\frac{1}{2}} / \underset{\sim}{a}_i'\underset{\sim}{\omega} \right\} \geq (1-\alpha);$$

this statement is exactly of the form (4) with

$$k_i = \alpha^{-\frac{1}{2}} (\underset{\sim}{a}_i' \, V \, \underset{\sim}{a}_i)^{\frac{1}{2}} / \underset{\sim}{a}_i'\underset{\sim}{\omega} \, .$$

If $\hat{k}_i$ denotes the estimate of k_i obtained by using $\hat{\omega}_j$ for ω_j in the above expression, then, for α=0.01, it follows from (5) that we have approximately

$$\Pr\{(1-\hat{k}_i)SPMR_i < CSMR_i < (1+\hat{k}_i)SPMR_i\} \geq 0.99 \qquad (6)$$

when $\hat{k}_i = 10(\underset{\sim}{a}_i' \, \hat{V} \, \underset{\sim}{a}_i)^{\frac{1}{2}} / \underset{\sim}{a}_i'\hat{\underset{\sim}{\omega}}$.

One reasonable way to evaluate the reliability of the approximate confidence interval (6) is to apply it to several sets of data for which

the requisite information is available and then see how it behaves. Table 2 below summarizes the results of such an application to two sets of data. One data set is taken from a D.H.E.W. Vital Statistics Report on 1950 Tuberculosis Mortality Among U. S. White Miners. The other data source represents preliminary findings concerning the mortality experience of hourly employees and ex-employees at a major tire manufacturing plant in Akron, Ohio over the 9-year period 1/1/64 - 12/31/72. U. S. 1950 mortality data for men with work experience and U. S. national mortality statistics for 1968 provide the respective standard population figures.

From Table 2, it can be seen that in every instance the confidence interval calculated using (6) did indeed enclose the actual CSMR value. While this admittedly does not constitute an unequivocable endorsement of the procedure, it is nevertheless an encouraging finding. We can only hope that other researchers will subject our methodology to similar evaluations using their own data sets.

REFERENCES

1. Adelstein, A.M. Occupational Mortality: Cancer. Ann. Occup. Hyg., Vol. 15, 1972, 53-57.

2. Cramér, Harald. Mathematical Methods of Statistics. Princeton University Press, Princeton, 1963.

3. Doll, R. Cancer of the Lung and Nose in Nickel Workers. Brit. J. Industr. Med., Vol. 15, 1958, 217-223.

4. Gilliam, A.G. A Note on Evidence Relating to the Incidence of Primary Liver Cancer Among the Bantu. J. Nat. Canc. Inst., Vol. 15(1), 1954, 195-199.

5. Lloyd, J.W., Decoufle, P. and Salvin, L.G. Unusual Mortality Experience of Printing Pressmen. Paper presented at meetings of Amer. Ind. Hyg. Assoc., May, 1972, San Francisco.

6. McMahon, B. and Pugh, T.F. Epidemiology - Principles and Methods. Little, Brown & Co., Boston, 1970.

7. Moss, E., Scott, T.S. and Atherley, G.R.C. Mortality of Newspaper Workers from Lung Cancer and Bronchitis, 1952-66. Brit. J. Industr. Med., Vol. 29, 1972, 1-14.

8. U. S. Dept. of Health, Education and Welfare, Vital Statistics - Special Reports, Vol. 53, No. 5: Mortality by Occupation Level and Cause of Death Among Men 20 to 64 Years of Age, United States, 1950. Pages 439-442.

TABLE 2: An Evaluation of the Behavior of the Confidence Interval (6) for Various Data Sets

DATA SOURCE	AGE GROUPS	$n_{.j}$	$1000p_{ij}$	$a_{ij} = \frac{p_{ij}}{p_{.j}}$	$d_{.j}$	$d_{i.}$	$SPMR_i$	$\hat{k}_i$	$(1-\hat{k}_i)SPMR_i, (1+\hat{k}_i)SPMR_i$	$CSMR_i$
TIRE PLANT DATA ICD 200-209 (Leukemia Group)	40-54	16,139	0.1620	0.0200	132	36	1.207	0.02347	1.179, 1.235	1.197
	55-64	15,409	0.4229	0.0178	333					
	65-84	19,799	1.0023	0.0149	1,427					
TIRE PLANT DATA ICD 151 (Stomach Cancer)	40-54	16,139	0.0662	0.0082	132	36	1.646	0.01940	1.614, 1.678	1.652
	55-64	15,409	0.2619	0.0110	333					
	65-84	19,799	0.8032	0.0120	1,427					
TIRE PLANT DATA ICD 160-163 (Resp. Cancer)	40-54	16,139	0.5307	0.0655	132	92	0.877	0.05352	0.830, 0.924	0.865
	55-64	15,409	1.9431	0.0820	333					
	65-84	19,799	3.2353	0.0483	1,427					
WHITE MINER DATA (Tuberculosis)	20-24	74,598	0.1226	0.0628	292	540	1.458	0.04774	1.388, 1.528	1.418
	25-29	85,077	0.1612	0.0830	357					
	30-34	80,845	0.2154	0.0901	341					
	35-44	148,870	0.3396	0.0778	1,095					
	45-54	102,649	0.5682	0.0519	1,784					
	55-59	42,494	0.7523	0.0368	1,554					
	60-64	30,037	0.8237	0.0277	2,051					