CONF-860274--1

A SYSTEM FOR THE ANALYSIS OF COHORT MORTALITY DATA*

Richard McLain

Oak Ridge Associated Universities
Oak Ridge, Tennessee 37830

CONF-860274--1

DE86 008223

Edward L. Frome

Mathematical Sciences Section
Engineering Physics and Mathematics Division
Oak Ridge National Laboratory
Oak Ridge, Tennessee 37831

ABSTRACT

A system is developed for the analysis of cohort mortality data. This Mortality Analysis System (MAS) is designed as a research tool in epidemiologic studies. The system allows a researcher to investigate the effect of one or more factors on the mortality of a study cohort. Variables can be categorized as factors to allow for stratification in the analysis. DATA steps and PROC MATRIX are incorporated in the system to produce the output. Person-years, observed deaths, and expected deaths are calculated and cross-classified by the levels of the factors. The resulting data set can be used to compute the standardized mortality ratios (SMR) for each stratum level. Poisson regression models can then be used for further statistical analysis.

DISCLAIMER

*This work was supported under contract number DE-AC05-76OR00033 between the U.S. Department of Energy, Office of Energy Research, and Oak Ridge Associated Universities and under contract number DE-AC05-84OR21400 with Martin Marietta Energy Systems, Inc. with U.S. Department of Energy.

A SYSTEM FOR THE ANALYSIS OF COHORT MORTALITY DATA

Richard McLain, Oak Ridge Associated Universities
Edward L. Frome, Oak Ridge National Laboratory

SUMMARY

A system is developed for the analysis of cohort mortality data. This Mortality Analysis System (MAS) is designed as a research tool in epidemiologic studies. The system allows a researcher to investigate the effect of one or more factors on the mortality of a study cohort. Variables can be categorized as factors to allow for stratification in the analysis. DATA steps and PROC MATRIX are incorporated in the system to produce the output. Person-years, observed deaths, and expected deaths are calculated and cross-classified by the levels of the factors. The resulting data set can be used to compute the standardized mortality ratios (SMR) for each stratum level. Poisson regression models can then be used for further statistical analysis.

PURPOSE AND DESCRIPTION

Public concern about the health risks of exposure to occupational and environmental hazards has generated much current research in epidemiology. In the study of a cohort there are often many different factors that can influence the force of mortality. Socioeconomic status, attained-age, birth year, calender period, and exposure level are examples of common risk factors. These factors can be potential confounders or effect modifiers. Breslow et al (1983) have described three important methods that can be used in cohort mortality studies. One method that has been widely used requires observed and "expected" deaths grouped according to the levels of the risk factors. The expected deaths are computed using known baseline rates (usually from national vital statistics). The observed deaths are treated as independent Poisson variates and regression methods are used for statistical analysis (see Frome, 1983; Frome and Checkoway, 1985; Whittemore, 1985). The purpose of the MAS is to provide a convenient method for obtaining grouped mortality data.

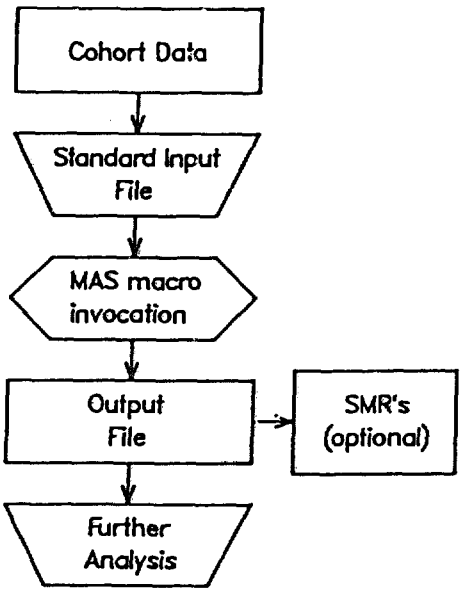
The MAS is designed to help the epidemiologist in achieving a better understanding of complex disease patterns that may occur in cohort mortality studies. Using the system, a researcher can take cohort data and create an output file for a specified cause of death. The variables of interest must be categorized into levels and defined as factors. A macro call will initiate a series of macros that creates the output file. DATA steps accumulate total person-years and observed deaths by levels of the factors. PROC MATRIX is used to calculate expected deaths using the known age and calendar year specific baseline rates. The researcher can request SMR's with the output or use the output data set in subsequent analysis.

Methods

The MAS is generally designed for use with occupational cohorts but is not limited to occupational settings. A file with a standardized format is required as input to the system. This standard input file (SIF) can be

created easily from a typical cohort file containing demographic information, vital status, and risk factors (see, e.g., Frome and Hudson, 1979).

The SIF must contain one record for each study subject and the risk factors must be categorical. Using standard variable names and factor level assignments the data set becomes input for the MAS macro. Options allow the researcher to request specific causes of death or a screen selection of all causes.



The following variables are required by the system:

Date of birth
Date of entry into the study
Date of last termination
Vital status (alive, dead, unknown)
Date last seen
ICD8 code (if dead)
F1 risk factors assigned levels 1,2,...,99
F2
F3
.
.
.
F10

The date last seen should represent the date of death, date lost to follow-up, or date of study closing depending on the vital status for the subject. The MAS will accept up to ten risk factors each with up to 99 levels (subject to operating system constraints).

The macro call allows for the selection of cause of death as an option. The calculations of expected deaths requires a set of rate tables

for the causes of death of interest for some external comparison population. All rates used in the following examples are for U.S. white males from the USDR program by Monson (1974).

The resulting output data set contains factor level indices, person-years, and observed and expected deaths for each cause of death selected. Another macro can be invoked to produce SMR values by the levels of the selected factors. The data set can be cataloged for further analysis using more complex methodology.

Mortality data from an occupational cohort is used here as an example (see Polednak and Frome 1981). Data for over 15,000 white males who worked at a uranium-processing plant during WWII are considered with three risk factors. The factors are socioeconomic status, birth year, and calendar year period. The levels are as follows:

F1 - socioeconomic status (S)	skilled
	unskilled
	professional
F2 - birth year (B)	<1910
	≥1910
F3 - calendar period (P)	1950-1959
	1960-1969
	1970-1974

The output obtained from the MAS for this example is given in Table 1. The overall lung cancer SMR is 119 with 95% confidence limits of (106, 134). Inspection of Table 1 indicates that there is substantial variation in the SMR values with respect to the factors of interest.

Poisson regression analysis is used to evaluate the relative importance of each of the factors (S, B, and P) by fitting a hierarchical sequence of log-linear models. The deviance for each model is recorded in a Poisson ANOVA table (similar to that used in a standard factorial experiment). Consider for example the "main effects" models.

$$E(Y_{ijk}) = {}_{ijk}e^{\phi_i+\beta_j+\delta_k},$$

where Y denotes the observed deaths, μ° the expected deaths, and the indices i, j, k identify the levels of socioeconomic status (S), birth year (B), and period (P). The deviance for this model (S+B+P) is 15.52 with 12 dF (see line 11 in Table 2).

The Akaikie information criterion (A.I.C. = deviance + 2 x number of parameters) is given in the last column of Table 2 (see McCullagh and Nelder, 1983 for further discussion of the deviance and A.I.C.). This is used to identify models that provide a good description of these data. We identified the model with the smallest A.I.C. value that contains all of the main effects as the "best" summary model. The maximum likelihood estimates of parameters for this model (S+B+P+S.B) are given in Table 3. These results indicate that SES is an important risk factor for lung cancer and there is a strong SES by Birth year interaction. Consequently, the use of

a summary SMR adjusted for age and calendar year is clearly not appropriate for this example.

Complete documentation for the portion of the system that uses the SIF to produce the summary output by factor levels is given in an ORAU technical report (see McLain and Frome).

ACKNOWLEDGEMENTS

The authors would like to express appreciation to Janet Kile for assistance in preparation of the manuscript, to Sheryl Byrkit for graphics assistance, and to Donna Hollis for preparing a portion of the macro.

From the Medical and Health Science Division, Center for Epidemiologic Research, Oak Ridge Associated Universities, Oak Ridge, Tennessee 37831-0117. This report is based on work performed under contract number DE-AC05-760R00033 between the Department of Energy, Office of Energy Research, and Oak Ridge Associated Universities, and contract number DE-AC05-840R21400 between the Department of Energy and Martin Marietta Energy Systems.

REFERENCES

Breslow, N. E., Lubin, J. H., Marek, P., and Langholz, B. (1983). Multiplicative Models and Cohort Analysis, Journal of the American Statistical Association, 73, 1-2.

Frome, E. L. and Hudson, D. R. (1980). A General Statistical Data Structure for Epidemiologic Studies of DOE Workers, Proceedings of the 1980 DOE Statistical Symposium, 206-218.

Frome, E. L. (1983). Analysis of Rates Using Poisson Regression Models, Biometrics, 39, 665-674.

Frome, E. L. and Checkoway, H. (1985). Epidemiology Programs for Computers and Calculators: Use of Poisson Regression Models in Estimating Incidence Rates and Ratios, American Journal of Epidemiology, 121, 309-323.

McCullagh, P. and Nelder, J. A. (1983). Generalized Linear Models, London: Chapman and Hall.

McLain, R. W. and Frome, E. L. A System for the Analysis of Cohort Mortality Data: Version I, ORAU Technical Report, Oak Ridge Associated Universities. (In preparation).

Monson, P. R. (1974). Analysis of Relative Survival and Proportional Mortality, Computers and Biomedical Research, 7, 325-332.

Polednak, A. P. and Frome, E. L. (1981). Mortality Among Men Employed Between 1943 and 1947 at a Uranium-Processing Plant, Journal of Occupational Medicine, 23, 169-178.

SAS Institute Inc. (1982). SAS User's Guide: Basics, Cary, NC.

SAS Institute Inc. (1982). SAS User's Guide: Statistics, Cary, NC.

Whittemore, A. S. (1985). Analyzing Cohort Mortality Data. The American Statistician, 39, 437-441.

Table 1. Lung Cancer Data For Workers at Uranium-Processing Plant

Period:		1950s		1960s		1970s	
Byear:		<1910	1910+	<1910	1910+	<1910	1910+
SES:							
Sk	Obs	33	6	75	38	47	35
	Exp	28.91	5.13	64.97	29.41	34.53	27.55
	Pyrs	44697	71011	36585	66796	11152	24621
	SMR	114	117	115	129	136	127
UnSk	Obs	5	1	14	13	7	5
	Exp	8.05	0.61	15.28	3.60	6.90	3.49
	Pyrs	9645	10925	7529	10291	2071	3782
	SMR	62	164	92	361	101	143
Prof	Obs	1	1	3	1	1	1
	Exp	1.79	0.35	3.97	2.15	2.29	2.20
	Pyrs	2750	5785	2258	5582	744	2159
	SMR	56	286	76	47	44	45

NOTE: Expected Deaths based on U.S. White Male rates.

Table 2. Poisson ANOVA For Lung Cancer Data
All Possible Multiplicative Factor Models

Model	df	Deviance	A.I.C.
minimal	17	22.97	25.0
S	15	18.64	24.6
B	16	20.53	24.5
P	15	21.94	27.9
S+B	14	16.05	24.1
S+P	13	17.57	27.6
B+P	14	20.02	28.0
S+B+S.B	12	7.39	19.4
S+P+S.P	9	14.41	32.4
B+P+B.P	12	18.54	30.5
S+B+P	12	15.52	27.5
S+B+P+S.B	10	6.94	22.9
S+B+P+S.P	8	12.29	32.3
S+B+P+B.P	10	14.04	30.0
S+B+P+S.B+S.P	6	3.90	27.9
S+B+P+S.B+B.P	8	5.51	25.5
S+B+P+S.P+B.P	6	10.66	34.7
S+B+P+S.B+S.P+B.P	4	2.71	30.7

Table 3. Parameter Estimates for "Best" Model

For The Lung Cancer Data In Table 1

Referent Group for Internal Comparison is
SES - Skilled Byear <1910 Period - 1950s

Parameter	Estimate(L%)	St. Dev.
SES		
Skilled vs U.S.	10.5	15.2
Unskilled vs. U.S.	-23.0	23.1
Professional vs. U.S.	-56.1	46.5
Period		
1960s vs. 1950s	10.4	17.0
1970s vs. 1950s	11.1	18.3
SES x Byear (Young vs Old)		
Skilled	3.7	14.1
Unskilled	103.4	30.4
Professional	1.3	73.1