

Recent Developments in the Multistage Modeling of Cohort Data for Carcinogenic Risk Assessment

by Sati Mazumdar,* **Carol K. Redmond,***
Joseph P. Costantino,* Rita N. Patwardhan,*
and Susan Y. J. Zhou*

Chairman, Biostatistics
Dept., U. of Pittsburgh
School of Public
Health
412 6243032

The modeling of cohort data based on the Armitage-Doll multistage model of the carcinogenic process has gained popular acceptance as a methodology for quantitative risk assessment for estimating the dose-related relationships between different occupational and environmental carcinogenic exposures and cancer mortality. The multistage model can be used for extrapolation to low doses relevant for setting environmental standards and also provides information regarding whether more than one stage is dose-related, which assists in determining whether different carcinogens affect different stages of the cancer process. This paper summarizes recent developments in the multistage modeling of cohort data and emphasizes practical issues such as handling data arising from large epidemiologic follow-up studies, the time-dependent nature of exposures and statistical issues such as multicollinearity in time-related variables, robustness of parameter estimates, validating of the fitted models, and routine inferential procedures. Problems related to uncertainties of risk estimates are discussed also. Computer programs for fitting multistage models with one or two dose-related stages to cohort data incorporating time-dependent exposure patterns; constructing confidence regions for model parameters; and predicting lifetime risks of dying from cancer adjusting for competing causes of death are detailed. Illustrations are provided using lung cancer mortality in a cohort of nonwhite male coke oven workers exposed to coal tar pitch volatiles.

Introduction

Recent literature on epidemiologic studies of human cancer indicates that the Armitage-Doll multistage model of carcinogenesis can be used to examine the relationships of excess cancer mortality to some carcinogenic exposure and time-related variables such as age at initial exposure; duration of exposure; and time since exposure terminated (1-4). These relationships help to predict benefits of different strategies for reducing carcinogenic exposure in the workplace or environment.

Because multistage models can be formulated to provide information regarding whether more than one stage is dose-related, these models assist in determining whether different carcinogens affect different stages of the cancer process. They also are used in carcinogenic quantitative risk assessment for estimating the dose-response relationships and for extrapolating to low doses relevant for setting environmental standards (5,6). The dose can be constant or time-related, and the response is usually measured by lifetime risk of dying from cancer.

This paper summarizes recent developments in multistage modeling of cohort data detailing methodological as well as

computational advancements. These developments lie mostly in dealing with inferential problems and developing computer programs.

A recently developed computer program package for the direct fitting of multistage models with one or two dose-related stages and time-dependent exposure patterns is described. In addition to tests of significance related to one-at-a-time inference, this package provides computing algorithms for the construction of confidence regions for simultaneous inferences regarding parameters of two-stage, dose-related multistage models and algorithms for computations of lifetime risks adjusting for competing causes of death. Programs in this package can be grouped conveniently according to the need of researchers.

Several statistical concerns that arise when cohort data are analyzed for carcinogenic risk assessment are discussed. New methodological approaches to deal with these concerns are proposed. Specific issues addressed include the appropriateness of statistical inferential procedures, robustness of parameter estimates, model validation, and uncertainties of risk estimates.

Illustrations of the operations of the program are provided, using lung cancer mortality in a cohort of nonwhite, male coke oven workers exposed to coal tar pitch volatiles. Results from the analysis of this data set are used to facilitate discussions of methodological and statistical concerns addressed in the paper.

*Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA 15261.

Address reprint requests to S. Mazumdar, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA 15261.

Armitage-Doll Multistage Model of Carcinogenesis

The simple multistage model of carcinogenesis assumes that a tumor results when k events occur in order in a single cell. The Occurrence rate (at age t) of the i th event given that $i-1$ events have occurred, is assumed to be:

$$\Gamma_i = a_i + b_i D(t) \quad (1)$$

where a_i is the background rate independent of age, b_i is the potency parameter for the i th cellular changes, and $D(t)$ is the dose rate at age t . The cumulative hazard rate (or cumulative incidence rate) by age t is then given by

$$H(t) = \int_0^t \int_0^{u_k} \int_0^{u_{k-1}} \dots \int_0^{u_1} [a_i + b_i D(u_i)] [a_k + b_k D(u_k)] du_1 du_k \quad (2)$$

where t_0 is the time from initiation of tumor to clinical expression.

In accordance with the above formulation, the age-specific cancer death rate $[h(t)]$ at time t can be expressed as follows:

$$h(t) = h_0(t) + h_1(t) \quad (3)$$

where $h_0(t)$ is a baseline age-specific cancer death rate derived either from a standard population or a suitably chosen control population, and $h_1(t)$ is the excess death rate. The expression for the excess death rate depends on the number and the order of the dose-related stage(s) and the nature of the dose rate patterns.

Recent Developments in Multistage Modeling of Cohort Data

The methodology used in the present paper for multistage modeling of cohort data incorporates time-dependent exposure patterns. A brief description of this methodology is given below followed by recent developments in its implementation. Developments in areas of model validation and appropriateness of inferential procedures are also discussed.

Direct Fitting of the Multistage Model

In occupational cohort studies, historical exposure estimates for one or more agents are usually taken as the measure of dose. Simplified expressions for multistage models under the assumption of constant exposure rate are easily amenable to model fitting. Applications of this approach for occupational arsenic and coke oven exposures are given by Brown and Chu (3) and Dong et al. (4).

The assumption of constant exposure rate throughout the work history is not realistic under most occupational settings. The fitting of multistage models to cohort data that reflect realistic exposure experience necessitates the incorporation of time-dependent exposure patterns. A methodology to incorporate exact time-dependent exposure patterns in the calculations of excess death rates has been provided by Crump and Howe (7). A formulation of this approach to epidemiologic data when one stage of the multistage model is dose-related is discussed in a subsequent paper (8).

Our first effort toward enhancing the methodology and com-

Table 1. Lung cancer mortality data of Allegheny County nonwhite, male coke oven workers.

Data	Wue
n	1100
Person-years	17635
Observed lung cancer deaths ¹	50
Expected lung cancer deaths	12.11
Standardized mortality ratio*	412.7
95% Confidence interval	306.3-544.1
99% Confidence interval	271.9-588.2

¹Deaths are coded to the Seventh Revision of the International Classification of Diseases.

*Statistically significant, $p < 0.01$

puter software for the fitting of multistage models to cohort data sets incorporating time-dependent exposure patterns consisted of extending the formulation provided by Crump et al. (8) to the situation when two stages are dose-related and developing a computer program package for model fitting and computing lifetime risks. A brief description of the formulation of the problem with exact expressions for the excess risk function is provided first to facilitate our discussion of the computer package developed.

To illustrate the multistage modeling programs, we used a data set from the Steelworkers Mortality Study, a project conducted in the Department of Biostatistics, University of Pittsburgh (9). The selected cohort consists of 1100 Allegheny County black male coke oven workers who worked between 1951 and 1953 and were followed through 1970. There were 50 lung cancer deaths during this follow-up period (Table 1). Exposure data consist of concentrations of coal tar pitch volatiles (CTPV) in milligrams per cubic meter (10). Exposure profiles were constructed for the workers given the levels of CTPV at jobs over the entire work periods divided into 1-month intervals. These exposure profiles provide the time-dependent exposure rates.

General Formulation of the Two-Stage Dose-Related Multistage Model

In the direct-fitting procedure of the multistage models to cohort data, a discrete version of Eq. (3) is needed. If we divide age into 1-year intervals, x_0 being the age when the worker starts his first job and the exposure rate in the i th interval is assumed to be a constant c_i ($i = 1, 2, \dots$), then for a k stage model with the r th and l th stages dose-related, Eq. (3) takes the form

$$h = \alpha_m + \beta_1 Z_{1m} + \beta_2 Z_{2m} + \beta_3 Z_{3m} \quad (4)$$

where h_m is the age-specific cancer death rate at age x_m , and α_m is the background cancer death rate at age x_m ,

$$Z_{1m} = \sum_{n=1}^{m-t_0} c_n \left[\frac{k!}{(k-r-1)!(r-1)!} \int_{x_{n-1}}^{x_n} (x_n - t_0 - u)^{k-r-1} u^{r-1} du \right] \quad (5)$$

$$Z_{2m} = \sum_{n=1}^{m-t_0} c_n \left[\frac{k!}{(k-l-1)!(l-1)!} \int_{x_{n-1}}^{x_n} (x_n - t_0 - u)^{k-l-1} u^{l-1} du \right] \quad (6)$$

$$Z_{3m} = \frac{k!}{(k-l-1)!(r-1)!(l-1-r)!} \sum_{n=1}^{m-t_0} c_n \int_{x_{n-1}}^{x_n} (x_n - t_0 - u)^{k-l-1} u^{r-1} u^{l-1} du + \sum_{i=1}^{n-1} c_i \int_{x_{i-1}}^{x_i} (x_i - t_0 - u)^{l-1} u^{r-1} du + c_n \int_{x_{n-1}}^{x_n} (u - t_0)^{l-1} u^{r-1} du \quad (7)$$

In Eqs. (4-7), t_0 is the time from initiation of tumor to its clinical expression and β_1, β_2 , and β_3 are potency parameters. In the present formulation of the problem, only β_1, β_2 , and β_3 need to be estimated from data. The constant t_0 is assumed to have a

fixed value which without any loss of generality can be taken to 0.

Denoting by z_{1m} , z_{2m} , and z_{3m} , Eqs. (5-7) for the n th individual and N_n , the total person-years at risk at age x_n , the expected number of cancer deaths at age x_n is given by

$$E_n = N_n \alpha + \beta_1 \sum_{n=1}^N z_{1m} + \beta_2 \sum_{n=1}^N z_{2m} + \beta_3 \sum_{n=1}^N z_{3m} \quad (8)$$

Assuming that the observed number of deaths at age x_n has a Poisson distribution with expected value E_n , the potency parameters β_1 , β_2 , and β_3 can be estimated by grouping observed and background deaths and z values and using a suitable maximum likelihood algorithm. The grouping can be done by the age at risk variable or intervals of z values or any other variable by which deaths and person-years can be grouped suitably.

Using this formulation, we had developed a computer program package for multistage modeling of cohort data when two stages are dose-related. The program package with an application has been reported (11,12).

Our most recent research deals with the methodological and statistical concerns related to the two-stage, dose-related multistage models with time-dependent exposure patterns. This includes adaptation of Armitage-Doll constraints in the general formulation of two-stage dose-related models described earlier and development of additional computer programs for the model fitting and inferential procedures.

Armitage-Doll Formulation of the Two-Stage Dose-Related Model

The formulation of the two-stage, dose-related multistage model described previously provides a more generalized form than given by the Armitage-Doll theory, which requires the following constraints on the model parameters:

$$\beta_3 = \beta_1 \times \beta_2 \text{ and } \beta_1 > 0, \beta_2 > 0 \quad (9)$$

These constraints are usually referred to as the Armitage-Doll constraints.

Detailed formulation of the two-stage, dose-related multistage model for epidemiologic data with the Armitage-Doll constraints, associated inferential procedures, and an application to coke oven workers cohort data set are detailed by Patwardhan (13). The expression given in Eq. (4) is reformulated as

$$h_n = \alpha + \beta_1 z_{1m} + \beta_2 z_{2m} + \beta_1 \times \beta_2 z_{3m}^* \quad (10)$$

where, $z_{3m}^* = z_{3m} (k x_n^{k-1}/a_m)$ and other quantities remain the same. Computer programs have been developed for simultaneous estimation, testing, and construction of confidence regions for model parameters using three distinct statistical procedures: Wald, likelihood ratio, and score.

In our initial program package for multistage modeling, generalized linear interactive modeling (14) was used for the estimation purpose, and the Armitage-Doll constraints could not be imposed on the estimation algorithm. As part of our

most recent work, the computer package has been enhanced by including programs for the constrained estimation methods. In addition, the new package contains several programs for inferential procedures. The package for single-stage, dose-related models is called "Multistage Modeling Programs I" and the package for two-stage, dose-related models is called "Multistage Modeling Programs II." Flowcharts of these two program packages are given in Figures 1 and 2.

Lifetime risks of dying from cancer adjusted for competing causes of death are calculated using the procedure developed by Gail (15). Using the Multistage Modeling Programs I to analyze the coke oven workers data set and specifying the number of stages (k) as 4; the dose-related stage (r) as 1; groupings of deaths and person-years by intervals of z values; and time-dependent exposure patterns as observed, the following information was generated: a) likelihood plot (Fig. 3); b) estimate of the potency parameter, standard error, goodness-of-fit statistics, and Wald, likelihood ratio-based, and score-based confidence intervals and associated test statistics (Table 2); and c) lifetime risks (Table 3).

Using the Multistage Modeling Program II to evaluate the data set and specifying the number of stages (k) as 4; dose-related stages (l, r) as 3 and 1; groupings of deaths and person-years by cross-classified cells of z value vector; and time-dependent exposure patterns as observed, the following information was generated: a) likelihood surface (Fig. 4); b) estimates of potency parameters, standard errors, goodness-of-fit statistics, and Wald, likelihood ratio-based, and score test statistics (Table 4); c) Wald, likelihood ratio-based, and score-based confidence regions (Fig. 5); and d) lifetime risks (not shown).

Coverage and Power Study of Test Procedures

Null hypotheses about the potency parameters of multistage models can be tested using standard inferential procedures such as Wald's procedure, the likelihood ratio-based procedure, and the procedure based on score statistics. The question often arises about the appropriateness and relative merits or differences in the performances of these procedures. To address this question, one must undertake simulation studies that are conducted under conditions corresponding to those of the data set being analyzed. We have developed computer programs to perform such simulations and used them to analyze the coke oven workers data set (15). Results from the simulation study indicated that for a single-stage, dose-related multistage model, confidence intervals obtained by all three procedures have good coverage and power properties. For two-stage dose-related models, the Wald confidence region performs poorly for extreme parameter values, whereas the other confidence regions perform satisfactorily.

Robustness

Assessing robustness of a regression fit is an important step in any modeling endeavor. There are two different aspects of this assessment. The first aspect involves assessing the robustness of the parameter for model-fitting purposes within some fixed distributional assumptions. The second aspect

Table 3. Estimated lifetime lung cancer risk (through age 85+ years) for a hypothetical US black male exposed to coal tar pitch volatiles for 40 years, exposure starting at age 20.

Exposure pattern ^a	Model: four stages	
	First stage dose-related	
	Risk ^b	Ratio ^c
L L L L	0.2104	75
L L H H	0.2566	91
L M H M	0.2864	102
H L L L	0.3488	124
H M L M	0.3800	135
H H L L	0.4013	143
H H H H	0.4251	151

^aL = 1 mg/m³ for 10 years; M = 2 mg/m³ for 10 years; H = 3 mg/m³ for 10 years.

^bU.S. black male age-specific mortality for 1960 used in competing risk calculations.

^cRatio of risk for exposed individual/risk for unexposed individual (lifetime risk for unexposed, 0.0281).

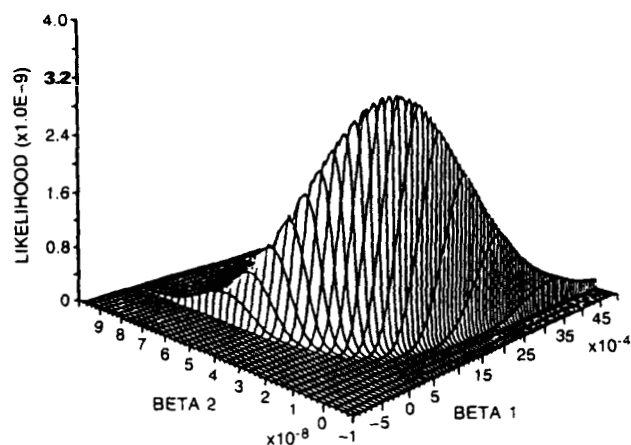


FIGURE 4. Likelihood surface for a first- and third-stage, dose-related, four-stage multistage model fitted to Allegheny County nonwhite male coke oven workers' lung cancer mortality, 1953-1970.

Table 4. Statistics for a first- and third-stage dose-related, four-stage multistage model fitted to Allegheny County nonwhite male coke oven workers' lung cancer mortality data, 1953-1970.

Statistics	
Estimated potency parameters ^a	
β_1	8.209×10^{-9}
	(SE, 10.127×10^{-9})
$\hat{\beta}$	1.454×10^{-9}
	(SE, 1.663×10^{-9})
Goodness-of-fit statistic:	
G ²	8.463
Pearson's X ²	2.990
Test statistics ^c	
Wald	63.936
Likelihood ratio	81.026
Score	180.219

^aEstimation is by grouping Demon-years, deaths, and z-values in 10 intervals.

^bG² and Pearson's X² have chi-square distributions with 17 and 4 degrees of freedom, respectively.

^cEach test statistic has a chi-square distribution with 2 degrees of freedom.

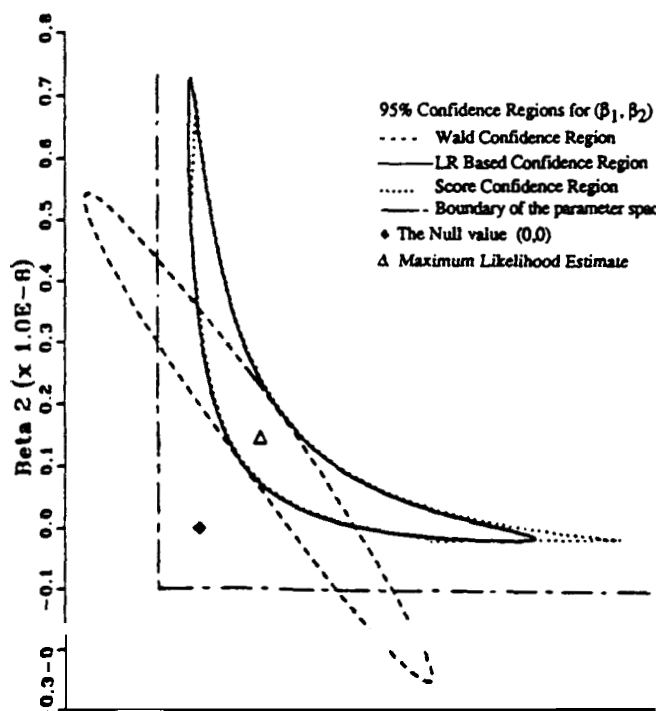


FIGURE 5. Confidence region (95%) for the parameters of a first- and third-stage, dose-related, four-stage multistage model fitted to Allegheny County nonwhite male coke oven workers' lung cancer mortality, 1953-1970.

robust regions of the parameters in the variance function and link function in a generalized linear model (GLM). These robust regions are defined as ranges in the spaces of variance function parameter or link function parameter for which the corresponding estimates of the potency parameter lie within the confidence interval of the potency parameter determined under the basic model assumptions and the model fitting is adequate. Wedderburn's quasiliikelihood approach (17) and several criteria based on commonly used statistics are used to obtain the robust regions. The method applies to the robustness investigation of either the variance function or the link function or both simultaneously.

The GLM method has been applied to the first of four stages dose-related multistage model fitted to the coke oven workers mortality data. Results show that the selected multistage model is sufficiently robust to the change of the relationship between the first two moments of the random component (variance function) but sensitive to the change of the relationship between the systematic component and the mean (link function). Figure 6 displays one such variance-link robust region.

Cross-Validation

The coke oven workers mortality study provides a unique opportunity to cross-validate models fitted to death rates for quantitative carcinogenic risk assessment since the cohort has been followed for about 30 years (the most recent follow-up is through 1982). The extended follow-up allows fitting a model with data through a certain follow-up period and then making

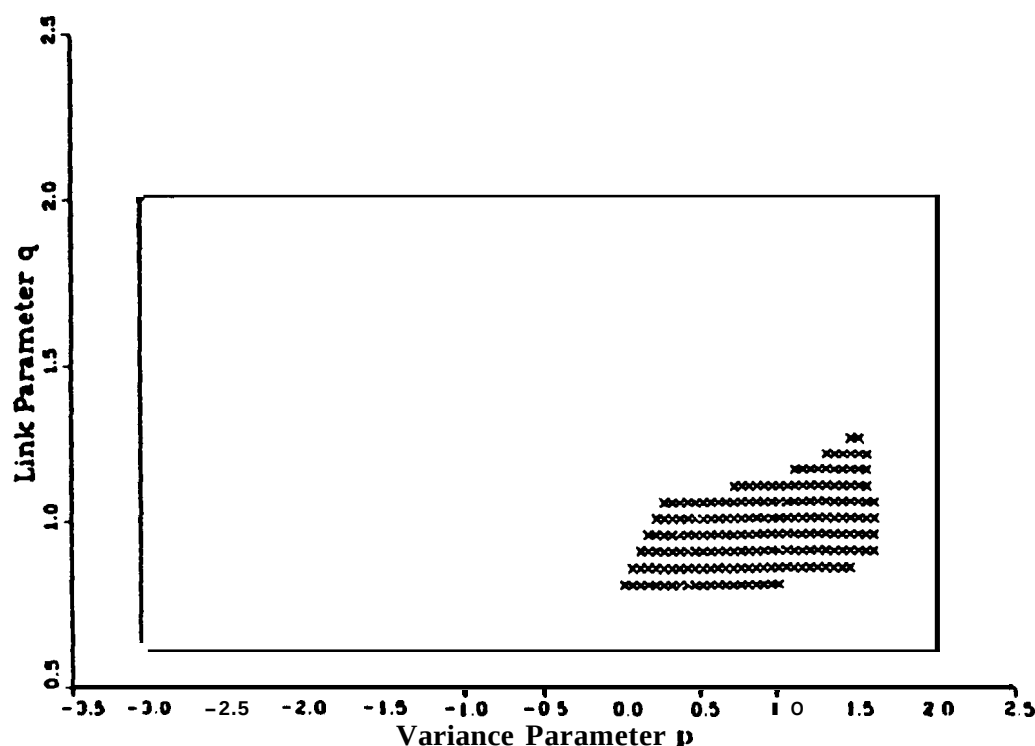


FIGURE 6. Variance-link robust region using four criteria. Criterion I: confidence interval and generalized Pearsonian chi-square; criterion II: confidence interval and deviance; criterion III: test statistic and generalized Pearsonian chi-square; criterion IV: test statistic and deviance.

predictions of excess deaths for subsequent calendar periods. These predictions can then be validated with actual observations from the extended follow-up periods. As commonly used, goodness-of-fit statistics (e.g., deviance and Pearson's chi-square) are found to be not very sensitive in choosing the best fitted model. We believe that the cross-validation method is more meaningful for the modeling of cohort data.

In the context of multistage modeling of cohort data with time-dependent exposure patterns, this cross-validation method requires first the grouping of Z values (vector) by specific calendar year periods and using the estimated potency parameters (s) and calculating the predicted excess death rates beyond the calendar period used in model fitting. The predicted excess death rates can then be compared with the observed excess death rates obtained from the extended follow-up information. A computer program has been developed for grouping deaths and person-years by calendar year periods and was used to cross-validate a multistage model with time-dependent exposure pattern fitted to the lung cancer mortality of a different cohort of coke oven workers (18).

Discussion

In recent years, the use of multistage models has become an integral part of environmental carcinogenic risk assessment. Multistage models provide a conceptual framework that facilitates understanding of the relationships between different

time-dependent variables and tumor incidence, and they also give a reasonable description of the epidemiology of many nonhormonally dependent cancers of epithelial origin.

Concerns about the basis of multistage models have been raised by Moolgavkar and his colleagues in a series of papers (19,20). A deficiency of the Armitage-Doll model is that the target tissue is not allowed to grow and that no allowance is made for the fact that most epithelia constantly shed and replenish cells. Moolgavkar and his colleagues have developed a two-stage model to rectify this defect (18). We have developed some of the necessary software for the modeling of large cohort data sets using two-stage models and will continue efforts to complete the remaining work. Programs that are developed for the multistage modeling purpose will be modified suitably for two-stage modeling.

We have developed methods to assess several statistical issues that arise in multistage modeling including the appropriateness and merits of various procedures for testing the null hypothesis, the robustness of parameter estimates, and the cross-validation of models. However, the methods developed so far fall short of adequately addressing uncertainties in the risk estimates. The confidence limits and the confidence regions of the potency parameters can be used in providing ranges of risk estimates. In addition, we intend to use bootstrap methods to provide alternative estimates of the confidence limits (regions) based on asymptotic theory which then can be used to provide alternate ranges of the risk estimates.

This research has been funded in part by the U.S. Environmental Protection Agency under assistance agreement number CR806815 to the Center for Environmental Epidemiology, Graduate School of Public Health, University of Pittsburgh. This paper does not necessarily reflect the views of the Agency and no official endorsement should be inferred.

We acknowledge the University of Pittsburgh Super Computing Center for allocating the time and facilities to use the Super Computer CRAY System.

REFERENCES

1. Armitage, P., and Doll, R. Stochastic models for carcinogenesis. In: *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability: Biology and Problems of Health*, 4. University of California Press, Berkeley, CA, 1961, pp. 19-38.
2. Day, N. E., and Brown, C. C. Multistage models and primary prevention of cancer. *J. Natl. Cancer Inst.* 64:977-989 (1980).
3. Brown, C. C., and Chu, K. C. A new method for the analysis of cohort studies: implications of the multistage theory of carcinogenesis applied to occupational arsenic exposure. *Environ. Health Perspect.* 50: 293-308 (1983).
4. Dong, M. H., Redmond, C. K., Mazumdar, S., and Costantino, J. P. A multistage approach to the cohort analysis of lifetime lung cancer risk among steelworkers exposed to coke oven emissions. *Am. J. Epidemiol.* 128(4): 860-873 (1988).
5. Brown, C. C., and Chu, K. C. Use of multistage models to infer stage affected by carcinogenic exposure: example of lung cancer and cigarette smoking. *Chron. Dis.* 40: 171S-179S (1987).
6. Gibb, H. J., and Chen, C. W. Multistage model interpretation of additive and multiplicative carcinogenic effects. *Risk Anal.* 6: 167-170 (1986).
7. Crump, K. S., and Howe, R. B. The multistage model with a time dependent dose pattern: applications to carcinogenic risk assessment. *Risk Anal.* 4: 164-176 (1984).
8. Crump, K. S., Allen, B. C., Howe, R. B., and Crockett, P. W. Time related factors in quantitative risk assessment. *Chron. Dis.* 40: 101S-111S (1987).
9. Redmond, C. K., Wiend, H. S., Rockette, H. E., Sass, R., and Weinburg, G. Long-Term Mortality Experience of Steelworkers. NIOSH Research Report (HSM) 997132, U.S. Department of Health and Human Services, Washington, DC, June, 1981.
10. Mazumdar, S., Redmond, C. K., Sollecito, W., and Sussman, N. An epidemiological study of exposure to coal tar pitch volatiles among coke oven workers. *J. Air Pollut. Control Assoc.* 25: 382-389 (1975).
11. Mazumdar, S., Redmond, C. K., Costantino, J. P., Marsh, G. M., Chou, S., and Nalavade, R. Multistage modeling of epidemiologic data: a generalized computer program for quantitative risk assessment. *Arch. Environ. Health* 43(2): 203 (1988).
12. Mazumdar, S., Redmond, C. K., Enterline, P. E., Marsh, G. M., Costantino, J. P., Zhou, S. Y. J., and Parwardhan, R. N. Multistage modeling of lung cancer mortality among arsenic exposed copper-smelter workers. *Risk Anal.*, in press.
13. Parwardhan, R. N. Inferential Procedures for Multistage Models for Carcinogenic Risk Assessment with Applications. Dissertation, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA, 1989.
14. Baker, R. J., and Nelder, J. A. Generalized Linear Interactive Modelling (GLIM) System. Manual for Release 3.77, Numerical Algorithms Group, Oxford, England, 1985.
15. Gail, M. Measuring the benefit of reduced exposure to environmental carcinogens. *Chron. Dis.* 28: 135-147 (1975).
16. Zhou, Y. J. Investigation of Model Robustness Using the Generalized Linear Model Approach with an Application to Multistage Models of Carcinogenesis. Dissertation, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA, 1988.
17. Weederburn, R. W. M. Quasi-likelihood functions, generalized linear models and the Gauss-Newton Method. *Biometrika* 61: 439-447 (1974).
18. Mazumdar, S., Redmond, C. K., Costantino, J. P., Zhou, S. Y. J., and Parwardhan, R. N. Inferential problems in multistage modeling of cohort data with time-dependent exposure patterns. In: *Abstracts of the papers presented at the Joint Statistical Meetings*, New Orleans, LA, August 22-25, 1988.
19. Moolgavkar, S. H., and Knudson, A. G., Jr. Mutation and cancer: a model for human carcinogenesis. *J. Natl. Cancer Inst.* 66: 1037-1052 (1981).
20. Moolgavkar, S. H. Carcinogenesis modelling: from *mutator* biology to epidemiology. *Ann. Rev. Pub. Health* 7: 151-169 (1986).

