American Journal of Epidemiology
Vol. 169, No. 1
DOI: 10.1093/aje/kwn284
Advance Access publication November 6, 2008

Original Contribution

Multistage Modeling of Leukemia in Benzene Workers: A Simple Approach to Fitting the 2-Stage Clonal Expansion Model

David B. Richardson

Initially submitted May 6, 2008; accepted for publication August 14, 2008.

A simple SAS software program (SAS Institute, Inc., Cary, North Carolina) was developed for fitting an exact formulation of the 2-stage clonal expansion model accommodating piecewise constant exposures and left and right censoring of observations. Data on leukemia mortality and occupational exposure to benzene among rubber hydrochloride production workers in Ohio (1940–1996) were analyzed by using this approach. A model in which benzene exposure increased clonal expansion fit these data well; little evidence of an association between benzene exposure and initiation of leukemia was found. The estimated exposure-response association increased in magnitude with age at exposure and decreased with time since exposure. This analysis shows that the 2-stage clonal expansion model can be readily fit to epidemiologic cohort data by using a simple SAS program. The illustrative analyses of leukemia mortality among rubber hydrochloride workers suggest that the effect of benzene on leukemia risk is due to an exposure-induced increase in the proliferation of initiated cells.

benzene; leukemia; models, statistical

Abbreviations: ICD, *International Classification of Diseases*; LRT, likelihood ratio test; TSCE, 2-stage clonal expansion.

Any statistical analysis of cohort data starts with a mathematical model of the underlying process generating the observed data. Epidemiologists tend to draw upon a small set of models, including the logistic model, the exponential rate model, and the proportional hazards model. They seldom attribute any biologic interpretation to the mathematical model at the foundation of their data analysis. Rather, model choice is usually motivated by attributes such as goodness of fit, stability, and ease of implementation (1).

In some situations, however, epidemiologists take a different approach. They posit a mathematical model for the disease process that is informed by a theoretical, or biologic, conception of the disease. Several decades ago, Morrison (2) argued that many diseases may be viewed as arising via a sequence of transitions from a normal, healthy state to a pathologic state. Substantial work has been done to describe how multistage models can be utilized in epidemiologic analyses (3–6). This work dates back a half century to the observation that the rise in mortality rates as a power of age conforms to expectations for the hazard rate from a multistage disease process (7). Shortly thereafter, Armitage and Doll (8) noted that the sequence of transitions between stages of the disease process implies a time-dependent hazard function and that, under a multistage model, the effect of an exposure on disease risk may depend upon age at exposure and time since exposure.

Multistage disease models hold particular appeal in cancer research, where disease is routinely posited to arise because of the sequential transformation of a single cell in a process that involves multiple rate-limiting stages (9). In the 1980s, Moolgavkar and Knudson (5) described a multistage cancer model that included parameters for cell kinetics (the birth and death of clones), building upon work of Armitage and Doll (10); it is referred to as the 2-stage clonal expansion (TSCE) model. Using this model, a data analyst can explore whether the data conform to the patterns implied by exposure-induced changes in disease initiation, promotion, or malignant transformation.

Nonlinear models such as the TSCE model can be more difficult to fit to epidemiologic data than standard linear

Correspondence to Dr. David B. Richardson, Department of Epidemiology, School of Public Health, University of North Carolina, Chapel Hill, NC 27599 (e-mail: david.richardson@unc.edu).

models for disease rates. Nonetheless, multistage disease models offer an interesting complement to purely empirical approaches to epidemiologic data analysis. This paper presents an approach to fitting the TSCE model to epidemiologic cohort data by using the SAS software package (SAS Institute, Inc., Cary, North Carolina). Model fitting is illustrated by using data from a cohort study of mortality among workers exposed to benzene during the production of rubber hydrochloride. The example demonstrates how insights generated by the TSCE model may be obtained that are not readily derived via empirical models for dose-time-response associations.

MATERIALS AND METHODS

Under the TSCE model, the process of cancer induction commences with a population of normal stem cells susceptible to transformation into an intermediate premalignant stage, referred to as initiation, followed by a second transformation resulting in malignant cells (Figure 1). I assume a fixed population of normal stem cells of size X. Initiation occurs at a rate of μ_0; the rate of malignant conversion is described by the parameter μ_1. Initiated cells may increase in number via cell division or diminish in number via cell death, characterized by the parameters α and β, respectively; the net change in the subpopulation of initiated cells may be represented by $\gamma = \alpha - \beta - \mu_1$. The parameter for the rate of cell death β may be defined as a linear combination of the model parameters γ, α, and μ_1. Clonal expansion of the subpopulation of initiated cells increases the number of cells at risk of a second transformation, thereby resulting in a malignant cell. Under a multistage model that allows for clonal expansion, very low clonal expansion rates should lead to negligible divergence from the classical Armitage-Doll multistage model (8); in contrast, a very rapid round of clonal expansion would effectively reduce by 1 the number of steps in the process (11).

Exposure to an agent may affect 1 or more of the parameters of the TSCE model. Under what I will refer to as a linear model, the dose-response relation for a model parameter, θ, is given by $\theta(t) = \theta_0 \times (1 + b_c \times d(t))$, where θ_0 is the value of the model parameter in the absence of exposure, $d(t)$ denotes the concentration of exposure at age t, and b_c is the linear dose-response coefficient. Under what I will refer to as a power model, this relation is given by $\theta(t) = \theta_0 \times \left(1 + b_c \times d(t)^{b_e}\right)$, with b_c and b_e referred to as linear and power coefficients of the dose-response relation.

Fitting the TSCE model

The parameters are estimated by using maximum likelihood methods. The likelihood for an individual is defined in terms of the TSCE model hazard function, h, and the corresponding survival function, S. Expressions for the survival and hazard functions of the TSCE with piecewise constant parameters have been described previously and are discussed further in the Appendix (12). Let us define ts_i as the age at which person i enters the study and tq_i as the age at which the person exits the study (because of death or loss to follow-up). An individual's likelihood is

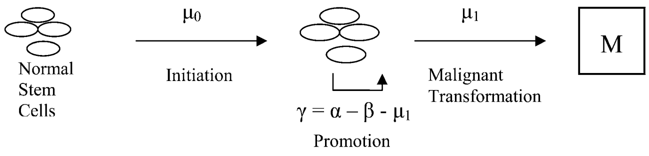

Figure 1. Pictorial depiction of the 2-stage clonal expansion model. Refer to the Materials and Methods section of the text for an explanation of the parameters.

$$L = \begin{cases} \frac{h(tq_i) \times S(tq_i)}{S(ts_i)} & \text{if person is a case;} \\ \frac{S(tq_i)}{S(ts_i)} & \text{otherwise,} \end{cases}$$

thereby accounting for left and right censoring in the calculation of an individual's likelihood. The overall likelihood for a model is the product of the individual likelihoods for the members of the cohort. Allowance for a fixed lag, l, between malignant transformation of a cell and subsequent diagnosis or death due to cancer is implemented by defining $ts_i{}' = ts_i - l$ and $tq_i{}' = tq_i - l$ and then calculating the hazard and survival for the intervals defined by $ts_i{}'$ and $tq_i{}'$.

The likelihood function calculation and its maximization can be performed via the SAS NLMIXED procedure (13). An example of the SAS code used to fit the TSCE model to the study data is provided in the Appendix. Not all parameters of the TSCE model are estimable with typical epidemiologic data; therefore, suitable constraints or identifiable parameter combinations have to be chosen. I have assumed that the number of susceptible stem cells, X, is fixed at $10^7$and imposed the constraint of equality of the spontaneous rates of the first and second mutation ($\mu_0 = \mu_1$). Although alternative parameterizations of the TSCE model can be explored, only a subset of parameter combinations of the TSCE model can be determined from these epidemiologic data.

The baseline rates of transformation and proliferation were estimated first by maximizing the likelihood for the observed outcome by using data for the subcohort of workers who had <1 part per million (ppm)-year of occupational benzene exposure. Next, holding these baseline parameters constant, I fit the TSCE model to the entire cohort to optimize the exposure-response parameters. This approach was preferable to fitting the model to the entire cohort and estimating all parameters simultaneously since it provided better fit of the model to observed data. Model fit was assessed by comparing observed with predicted numbers of leukemia deaths within subgroups of the study population. Predicted numbers of leukemia deaths were generated by using the estimated parameters for the TSCE model to compute the cumulative hazard for each individual in a group; the predicted number of cases for a given group was the summation of the cumulative hazards for all individuals in that group. Lastly, 95% credible intervals for model parameters were derived via Monte Carlo methods; this step was performed with a slight modification of the SAS code shown in the Appendix, invoking the SAS MCMC procedure (13). An example of the modified SAS code is provided as

Table 1. Characteristics of the Cohort of 1,721 Male Rubber Hydrochloride Workers Exposed to Benzene, Ohio, 1940–1996

Characteristic	1st Quartile	2nd Quartile	3rd Quartile	4th Quartile
Age at entry, range in years	17–24	24–29	29–38	38–74
Age at exit, range in years	20–61	61–68	68–74	74–108
Duration of employment, range in years	0.00–0.1	0.1–0.6	0.6–2.9	2.9–36.1
Cumulative exposure, range in ppm-years	0–0	0–1.7	1.7–22.6	22.6–817.7

Abbreviation: ppm, parts per million.

a supplemental Appendix posted on the *Journal*'s website (http://aje.oupjournals.org/).

Example

A cohort mortality study was conducted of workers employed in the manufacture of a natural rubber film (rubber hydrochloride) at 2 locations in Ohio (14, 15). Production activities involved dissolving natural rubber in benzene; then, the benzene was evaporated and recovered while the rubber film was stripped from a conveyor (14). Data for all nonsalaried males employed between January 1, 1940, and December 31, 1965, were included in this analysis. Vital status was ascertained through December 31, 1996. Information was obtained on underlying cause of death for deceased workers, coded according to the revision of the *International Classification of Diseases* (ICD) in effect at the time of death. These analyses focus on death due to leukemia (ICD-6 and ICD-7 code 204, ICD-8 codes 204–207, ICD-9 codes 204–208).

Estimates of benzene exposure rates, in parts per million, by calendar period, plant, department, and job, were developed by Rinsky et al. (14, 15) on the basis of available air sampling data. The US National Institute for Occupational Safety and Health provided a file that contained a plant, department, and job code, and start and finish dates, for each job held by each worker in the study cohort. Benzene exposure histories were estimated for each worker by using this information.

An analytical data file was constructed with 1 observation per worker. The file included the worker's age at start of follow-up, age at date of the last observation, a binary indicator of death due to leukemia, and an array of variables $d(t)$ indexing the benzene exposure during each year of age from birth to the date of the last observation. For example, $d(35)$ represents benzene exposure during the interval 34–35 years of age.

RESULTS

The study cohort included 1,721 male workers and 16 deaths due to leukemia. Table 1 shows the distribution of age at entry, age at exit, duration of employment, and cumulative benzene exposure in the study cohort. As expected, most workers entered follow-up (at the start of employment) in young adulthood and exited follow-up at age 60 years or older. The distribution of cumulative benzene exposure was positively skewed, with 461 workers having a cumulative benzene exposure equal to 0 ppm-year and 814 workers having a cumulative benzene exposure of <1 ppm-year.

Data for the subsample of workers who accrued <1 ppm-year of benzene exposure were used to estimate baseline model parameters for initiation and malignant transformation, μ_0 and μ_1, and net change in the population of initiated cells, γ. Next, data for the entire study cohort were used to fit a model in which benzene affected the rate of transformation to the intermediate stage, the rate of transformation to the malignant state, or the cell kinetics (Table 2). Inclusion of a term describing a linear effect of benzene on the initiation rate, μ_0, or the malignant transformation rate, μ_1, resulted in a significant improvement in model fit (likelihood ratio test (LRT) = 7.86, 1 df and LRT = 6.96, 1 df, respectively). Inclusion of a power term for the effect of benzene on the initiation rate or the malignant transformation rate led to no improvement in model fit (LRT = 0.00, 1 df and LRT = 0.00, 1 df, respectively). The best model fit was obtained via inclusion of a term describing a linear effect of benzene on the cell kinetics (LRT = 16.76, 1 df). Inclusion of a power term for the effect of benzene on the cell kinetics led to a negligible improvement in model fit (LRT = 0.40, 1 df) and therefore was not incorporated.

Table 3 reports the maximum likelihood estimates for a model that allows for a linear effect of benzene exposure on γ. The rate of change in the subpopulation of initiated cells is doubled by an exposure of approximately 15 ppm. The Monte Carlo–based 95% credible interval for the parameter describing the benzene exposure effect on cell kinetics (0.0366, 0.0917) was similar to the Wald-type confidence interval (0.0400, 0.0922).

Table 2. Model Deviances From Fitting of the 2-Stage Clonal Expansion Regression Model With Inclusion of a Term for a Linear Effect of Benzene Exposure on Initiation, Conversion to a Malignant State, or Cell Kinetics in a Study of Leukemia Mortality Among Male Rubber Hydrochloride Workers in Ohio, 1940–1996

Parameter Affected by Benzene	−2LogL	df
No effect	291.9	0
Initiation, μ_0	284.0	1
Promotion, γ[a]	275.1	1
Malignant conversion, μ_1	284.9	1

[a] Applies also to the initiated cells' division rate.

Table 3. Estimates of Parameters for the 2-Stage Clonal Expansion Model

Description	Estimate	Wald 95% CI
Baseline model		
Stem cell population, X	1×10^7	
Initiated cell division rate, α	3	
Initiation and malignant conversion rate, $\mu_0 = \mu_1$	9.914×10^{-8}	−11.66, 31.48
Initiated cells' promotion rate, γ	0.0849	0.0045, 0.1654
Benzene parameter		
Promotion exposure rate coefficient γ_c[a]	0.06611	0.04003, 0.09219

Abbreviation: CI, confidence interval.

[a] Applies also to the initiated cells' conversion rate, α.

Table 4. Numbers of Baseline, Observed, and Predicted Leukemia Deaths by Quartile of Cumulative Benzene Exposure Among Male Rubber Hydrochloride Workers in Ohio, 1940–1996

Quartile (Range in ppm-Years)	Baseline Prediction[a]	Model Prediction[b]	Observed
1st (0–0)	2.70	2.70	1
2nd (0–1.7)	1.93	2.00	3
3rd (1.7–22.6)	2.70	3.08	3
4th (22.6–817.7)	2.96	9.81	9

Abbreviation: ppm, parts per million.

[a] Predicted deaths based on the baseline model shown in Table 3.

[b] Predicted deaths based on the model shown in Table 3 allowing for benzene exposure to influence the promotion (and conversion) rate of initiated cells.

A model that included linear terms for the effect of benzene on cell kinetics and on the initiation rate, μ_0, fit the data no better than a model with just an effect of exposure on cell kinetics (LRT = 0.00, 1 df). A model that included linear terms for the effect of benzene on cell kinetics and on the malignant transformation rate, μ_1, fit the data only slightly better than a model with just an effect of exposure on cell kinetics (LRT = 1.60, 1 df).

No lag was used in these analyses; a model with no lag had better goodness of fit than a model with a 2- or 5-year lag. Sensitivity analyses were conducted to assess the impact on results of assumptions about the size of the stem cell population, X, and the baseline clonal division rate, α. Similar estimates for the effect of benzene on cell kinetics were obtained when replicating the above models under the assumption that the stem cell population was an order of magnitude larger or smaller than the working assumption in these analyses. Results showed greater sensitivity to the assumption about the baseline clonal division rate, α. While the estimate for the effect of benzene on cell kinetics changed very little under the assumption that α was an order of magnitude smaller than the working assumption in these analyses, under the assumption that α was an order of magnitude larger, the estimate of the effect of benzene exposure on cell kinetics increased by a factor of 1.8.

Table 4 reports the numbers of baseline, observed, and predicted leukemia deaths for subcohorts defined by quartiles of lifetime cumulative exposure. The predicted number of leukemia deaths within each group conforms closely to the observed number of deaths. The results based on the TSCE model fitting indicate that approximately 6 of the 16 observed leukemia deaths are excess cases associated with benzene exposure.

Figure 2 illustrates the model prediction for the hazard ratio as a function of attained age for a person exposed to benzene at intensities of 5 ppm and 10 ppm, with exposures commencing at age 20 years and terminating at age 60 years. The figure suggests a relatively prompt decrease in the relative rate of leukemia following termination of exposure. Figure 3 displays the estimated hazard ratio for a person at age 65 years who was exposed to 10 ppm-year of benzene as a function of age at exposure. The figure suggests that the magnitude of the exposure-response association increases with age at exposure, from approximately 1.046 for an exposure at age 20 years to 1.057 for an exposure at age 64 years.

DISCUSSION

This paper illustrates a simple approach to fitting the TSCE model to epidemiologic cohort data. Historically, the TSCE model has been fitted by using specialized FORTRAN computer code (16–19); the approach described in this paper explains implementation of the TSCE model via SAS, a widely used statistical package. The approach implements an exact expression of hazard function for the TSCE model with piecewise constant dosing, as often encountered in occupational studies.

Knoke et al. (20) presented a method for fitting the TSCE model to tabulations of person-time and events by using SAS NLIN; however, that approach implemented a different

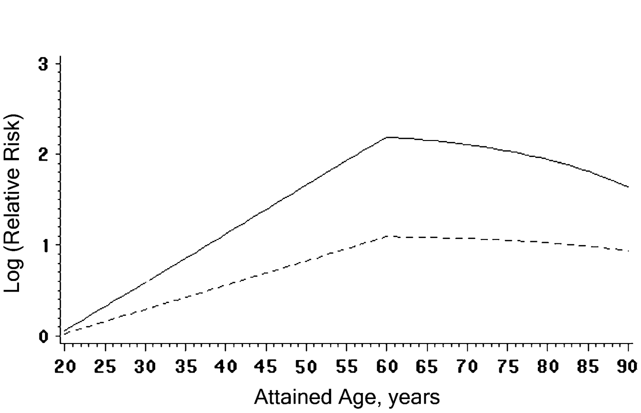

Figure 2. Relative risk of leukemia by attained age. Predicted impact of benzene exposure for rubber hydrochloride workers in Ohio (1940–1996) at intensities of 5 parts per million (dashed line) and 10 parts per million (solid line) based upon fitting of the 2-stage clonal expansion model. Benzene exposure commences at age 20 years and terminates at age 60 years.

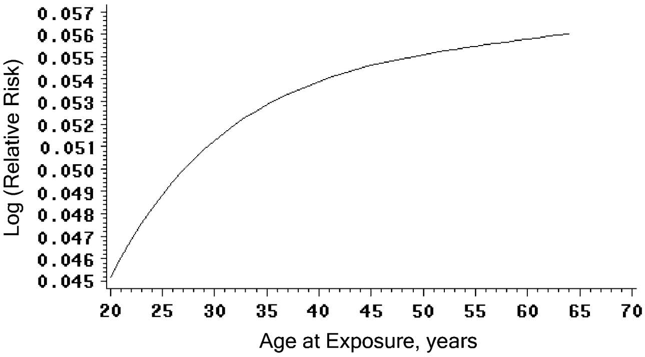

Figure 3. Relative risk of leukemia by age at exposure to benzene. Impact of a single year of exposure at 10 parts per million for a rubber hydrochloride worker aged 65 years, Ohio, 1940–1996.

expression for the TSCE hazard function that subsequent researchers have cautioned against using (20, 21). Lensing and Kodell (22) described how SAS may be used to fit an exact expression for the TSCE model in dosing studies of laboratory animals. However, that approach did not allow for staggered entry into a study and was based on a presentation of data as tabular counts of animals and case failures observed at a number of scheduled sacrifice times.

I illustrated this approach with analyses of the association between occupational benzene exposure and leukemia mortality. Previous analyses, based upon fittings of a Cox proportional hazards model, established that cumulative exposure to benzene was positively associated with leukemia mortality among these rubber hydrochloride workers (15, 23). In the current paper, the magnitude of the benzene exposure–leukemia mortality association derived via the TSCE model (approximately (relative rate at 10 ppm-year = 1.05)) is very similar to an estimate recently derived in analyses fitting a linear relative rate model to these data (23), although, under the fitted TSCE model, the magnitude of this association increases slightly with increasing age at exposure and diminishes promptly following termination of exposure. Prior analyses, utilizing the method of exposure time windows, also suggest that the joint effects of age at exposure and time since exposure are important to consider (23). Those analyses found that the effect of an increment of benzene exposure on leukemia mortality appears promptly, diminishes with time since exposure, and was of greater magnitude for workers exposed at older ages than for those exposed at younger ages.

In contrast to exposure time-window analyses, which impose a piecewise constant model to describe temporal variation in exposure effects, multistage models imply a smooth, time-varying function that jointly describes age at exposure and latency effects. As this paper illustrates, a multistage model can facilitate exploration of exposure-time-response associations in epidemiologic data. Of course, given the relatively small number of leukemia deaths, these results are relatively sensitive to small changes in distribution of events. Furthermore, it is reasonable to suspect that exposure-time-response associations may differ for different types of leukemia; however, given the small numbers of deaths included in the current analysis and the limited information on the death certificate (15), subtype-specific analyses could not be conducted.

In the fitted TSCE model shown in Table 3, benzene exposure influences cell kinetics but not initiation or malignant transformation. Thus, according to the model, leukemia induction over long periods of protracted exposure appears to be dominated by benzene-induced modification of the kinetics of already initiated cells rather than by direct benzene-induced initiation of normal cells. Such an observation is interesting given evidence of association between benzene exposures and proliferative blood disease. Cox (24) posited that benzene metabolites are responsible for the progression of a malignant clone of cells from a few (possibly dormant) transformed cells to a clinically detectable neoplasm. On the other hand, the benzene metabolite, benzene oxide, appear to be mutagenic. The modeling of these data provides modest evidence for benzene effects on malignant transformation rates.

Often, epidemiologists focus on regression modeling as a tool for summarization, pattern detection, and perhaps smoothing of epidemiologic data. Such activities can be difficult in studies of the effects of protracted exposures, when questions arise about the modifying effects of temporal factors such as age at exposure, time since exposure, and exposure rate. Use of multistage models, such as the TSCE, offers a way to incorporate information about exposure rates, and ages at exposure, into exposure-time-response analyses. Clearly, with typical epidemiologic data, a theoretical model of the underlying disease process cannot be accepted or rejected based upon the empirical data alone, and goodness of model fit to a particular data set should not be confused with an indication of the validity of a particular theoretical model for the disease process. It may be wise, therefore, to make very modest claims about the biologic interpretation of these model fittings. However, multistage models do offer a complement to empirical models and an approach to exploring dose-time-response associations within a defined set of constraints based upon some model assumptions about the underlying disease process.

The fundamental idea that a disease process follows from a sequence of rate-limiting steps may be an acceptable proposition in generic terms. The TSCE model introduces the additional complexity of allowing for 1 or more parameters to characterize the kinetics of clonal expansion, a dynamic suggested by current knowledge about the carcinogenic process. Like all modeling exercises, it becomes increasingly likely that, as more parameters are introduced, the fit of a model to a given data set will improve (although the predictive ability of the model may not be good when extrapolating beyond a given data set). However, in this example, the fitted TSCE model involves estimation of a relatively small number of parameters, and Figures 2 and 3 illustrate the potentially useful temporal descriptions of hazard ratios that may be derived from such model fittings.

For simplicity, no additional covariates were incorporated into these models; previous regression analyses of these data found minimal evidence of confounding by measured covariates such as year of birth (15, 23). Model parameters to describe variation in baseline rates with factors such as birth cohort are readily incorporated into the TSCE model; examples are provided in a number of previous studies (25, 26).

The ability to fit the TSCE model with piecewise constant exposures via a relatively simple SAS program should facilitate this approach to cohort analysis.

ACKNOWLEDGMENTS

Author affiliation: Department of Epidemiology, School of Public Health, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina.

This project was supported by grant K01-OH008635 from the National Institute for Occupational Safety and Health of the Centers for Disease Control and Prevention.

The author is grateful to Dr. William Hazelton, Fred Hutchinson Cancer Research Center, Seattle, Washington, for supplying a FORTRAN routine to calculate the exact hazard function for piecewise constant dosing.

Conflict of interest: none declared.

REFERENCES

1. Greenland S. Introduction to regression models. In: Rothman K, Greenland S, eds. *Modern Epidemiology.* 2nd ed. Philadelphia, PA: Lippincott-Raven; 1998.
2. Morrison AS. Sequential pathogenic components of rates. *Am J Epidemiol.* 1979;109(6):709–718.
3. Whittemore AS. The age distribution of human cancer for carcinogenic exposures of varying intensity. *Am J Epidemiol.* 1977;106(5):418–432.
4. Whittemore AS. Quantitative theories of oncogenesis. *Adv Cancer Res.* 1978;27:55–88.
5. Moolgavkar SH, Knudson AG Jr. Mutation and cancer: a model for human carcinogenesis. *J Natl Cancer Inst.* 1981; 66(6):1037–1052.
6. Thomas DC. Models for exposure-time-response relationships with applications to cancer epidemiology. *Annu Rev Public Health.* 1988;9:451–482.
7. Nordling CO. A new theory on the cancer inducing mechanism. *Br J Cancer.* 1953;7(1):68–72.
8. Armitage P, Doll R. The age distribution of cancer and a multistage theory of carcinogenesis. *Br J Cancer.* 1954;8(1):1–12.
9. Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell.* 2000;100(1):57–70.
10. Armitage P, Doll R. A two-stage theory of carcinogenesis in relation to the age distribution of human cancer. *Br J Cancer.* 1957;11(2):161–169.
11. Frank SA. *Dynamics of Cancer: Incidence, Inheritance, and Evolution*. Princeton, NJ: Princeton University Press; 2007.
12. Heidenreich WF, Luebeck EG, Moolgavkar SH. Some properties of the hazard function of the two-mutation clonal expansion model. *Risk Anal.* 1997;17(3):391–399.
13. SAS Institute, Inc. *SAS OnlineDoc 9.2*. Cary, NC: SAS Institute, Inc; 2007.
14. Rinsky RA, Smith AB, Hornung R, et al. Benzene and leukemia: an epidemiologic risk assessment. *N Engl J Med.* 1987;316(17):1044–1050.
15. Rinsky RA, Hornung RW, Silver SR, et al. Benzene exposure and hematopoietic mortality: a long-term epidemiologic risk assessment. *Am J Ind Med.* 2002;42(6):474–480.
16. Moolgavkar SH, Luebeck EG, Krewski D, et al. Radon, cigarette smoke, and lung cancer: a re-analysis of the Colorado Plateau uranium miners' data. *Epidemiology.* 1993;4(3): 204–217.
17. Hazelton WD, Luebeck EG, Heidenreich WF, et al. Analysis of a historical cohort of Chinese tin miners with arsenic, radon, cigarette smoke, and pipe smoke exposures using the biologically based two-stage clonal expansion model. *Radiat Res.* 2001;156(1):78–94.
18. Haylock RG, Muirhead CR. Fitting the two-stage model of carcinogenesis to nested case-control data on the Colorado Plateau uranium miners: dependence on data assumptions. *Radiat Environ Biophys*. 2004;42(4):257–263.
19. Moolgavkar SH, Luebeck EG, Anderson EL. Estimation of unit risk for coke oven emissions. *Risk Anal.* 1998;18(6): 813–825.
20. Knoke JD, Shanks TG, Vaughn JW, et al. Lung cancer mortality is related to age in addition to duration and intensity of cigarette smoking: an analysis of CPS-I data. *Cancer Epidemiol Biomarkers Prev.* 2004;13(6):949–957.
21. Meza R, Hazelton WD, Colditz GA, et al. Analysis of lung cancer incidence in the Nurses' Health and the Health Professionals' Follow-Up Studies using a multistage carcinogenesis model. *Cancer Causes Control*. 2008;19(3):317–328.
22. Lensing SY, Kodell RL. Fitting the two-stage clonal expansion model based on exact hazard to the ED01 data using SAS NLIN. *Risk Anal.* 1995;15(2):233–245.
23. Richardson D. Temporal variation in the association between benzene and leukemia mortality. *Environ Health Perspect*. 2008;116(3):370–374.
24. Cox LA Jr. Biological basis of chemical carcinogenesis: insights from benzene. *Risk Anal.* 1991;11(3):453–464.
25. Hazelton WD, Moolgavkar SH, Curtis SB, et al. Biologically based analysis of lung cancer incidence in a large Canadian occupational cohort with low-dose ionizing radiation exposure, and comparison with Japanese atomic bomb survivors. *J Toxicol Environ Health A*. 2006;69(11):1013–1038.
26. Luebeck EG, Heidenreich WF, Hazelton WD, et al. Biologically based analysis of the data for the Colorado uranium miners cohort: age, dose and dose-rate effects. *Radiat Res.* 1999;152(4):339–351.

APPENDIX

The SAS program below fits the TSCE model, parameterized as shown in Table 3. The model parameter X (denoted by the variable X in the SAS code below) is fixed. The baseline model parameters μ_0 and γ (denoted by the variables *mu0* and *g* in the SAS code below) were estimated from the study data for those workers whose lifetime cumulative benzene exposure was <1 ppm-year; these parameters were subsequently held constant for the analysis reported in Table 3. In this example, the model parameter α (denoted by the variable *alpha* in the SAS code below) was held constant as well; consequently, a single free parameter, *g_c*, describes the linear dose-response relation between benzene exposure and γ.

The source data file, *source*, includes 1 observation per worker. Each record includes the person's age at the start of follow-up (*age_entryy*), age at the date of the last observation (*age_exity*), a binary indicator of death due to leukemia (*leukemia*), and an array of variables *td*(t) indexing the annual benzene exposure rate during the period $t - 1$ to t years of age. The program uses the expressions described by Heidenreich et al. (12) to calculate the survival function for the interval from birth to *age_entryy* and the hazard and survival function for the interval from birth to *age_exity*. These expressions are used to calculate the likelihood for an individual. Briefly, the survival function and its derivative at age t are given by $S(t) = \exp\left\{\sum_{j=1}^{k} \frac{\mu_{0,j}X}{\alpha_j} \ln\left(\frac{q_j - p_j}{f_j\left(t_{j-1}, t_k\right)}\right)\right\}$ and $S'(t) = S(t) \times \sum_{j=1}^{k} \frac{\mu_{0,j}X}{\alpha_j} \frac{\delta}{\delta t_k} \ln\left(f_j\left(t_{j-1}, t_k\right)\right)$, respectively, where k is the number of time periods up to age t, $[t_{\mathrm{j}-1}, t_{\mathrm{j}}]$ denotes the endpoints for the jth interval, and $\mu_{0,\,j}$, α_j, g_j, $\mu_{1,\,j}$ denote the parameter values during the jth period, such that

$$
\begin{aligned}
&\mu_{0,j} = \mu_0(1 + \mu_{0c} d_j^{\mu_{0e}}),\\
&g_j = g(1 + g_c d_j^{g_e}),\\
&\alpha_j = \alpha(1 + g_c d_j^{g_e}),\\
&\mu_{1,j} = \mu_1(1 + \mu_{1c} d_j^{\mu_{1e}}),\\
&p_j = \frac{1}{2}\left(-g_j - \sqrt{g_j^2 + 4\alpha_j \mu_{1,j}}\right),\\
&q_j = \frac{1}{2}\left(-g_j + \sqrt{g_j^2 + 4\alpha_j \mu_{1,j}}\right),\\
&\tilde{y}_k = 0,\\
&\tilde{y}_{j-1} = \frac{\alpha_{j-1}}{\alpha_j} \frac{\left(\tilde{y}_j - p_j\right) q_j e^{q_j\left(t_{j-1} - t_j\right)} + \left(q_j - \tilde{y}_j\right) p_j e^{p_j\left(t_{j-1} - t_j\right)}}{f_j\left(t_{j-1}, t_k\right)},\\
&f_j\left(t_{j-1}, t_k\right) = \left(\tilde{y}_j - p_j\right) e^{q_j\left(t_{j-1} - t_j\right)} + \left(q_j - \tilde{y}_j\right) e^{p_j\left(t_{j-1} - t_j\right)},\\
&\frac{\delta}{\delta t_k} f_k(t_{k-1}, t_k) = \left[e^{q_k(t_{k-1} - t_k)} - e^{p_k(t_{k-1} - t_k)}\right] p_k q_k,\\
&\frac{\delta}{\delta t_k} f_j\left(t_{j-1}, t_k\right)\left[e^{q_j\left(t_{j-1} - t_j\right)} - e^{p_j\left(t_{j-1} - t_j\right)}\right] \frac{\delta}{\delta t_k} \tilde{y}_j,\\
&\frac{\delta}{\delta t_k} \tilde{y}_{k-1} = \frac{\alpha_{k-1}}{\alpha_k} \frac{\left(q_k - p_k\right)^2 e^{(p_k + q_k)(t_{k-1} - t_k)}}{\left(f_k(t_{k-1}, t_k)\right)^2} p_k q_k,\\
&\frac{\delta}{\delta t_k} \tilde{y}_{j-1} = \frac{\alpha_{j-1}}{\alpha_j} \frac{\left(q_j - p_j\right)^2 e^{\left(p_j + q_j\right)\left(t_{j-1} - t_j\right)}}{\left(f_k(t_{k-1}, t_k)\right)^2} \frac{\delta}{\delta t_k} \tilde{y}_j.
\end{aligned}
$$

A useful illustration of the calculation of individual likelihoods is provided by Meza et al. (21); I have attempted, in the SAS code below, to follow the notation used in their article. Illustrative SAS output is provided as supplemental material available with the electronic version of this paper.

```
proc nlmixed data= source;
parms   g_c=0;
bounds  g_c>=0;

x=1E7;             *<= Fixed;
alpha = 3;
mu0= 9.914*1E-8;
g = 0.08491;
mu1=mu0;
mu0_c=0;
    mu1_c=0;

    LAG=0;
    array td EXP_LEVEL1-EXP_LEVEL109;
```

```
DO LOOP = 1 TO 2;
IF Loop = 1 THEN DO; AGE_INDX=(Age_entryy-LAG); w=0; dw=0; sum_ln_Sentry=0; end;
IF Loop = 2 THEN DO; AGE_INDX=(Age_exity-LAG) ; w=0; dw=0; sum_ln_Sexit=0; sum_h=0; end;

DO J= AGE_INDX to 1 by -1;
  agestrt=j-1; agequit=j; dose=td{j};

  mu0_j=    mu0*(1+mu0_c*dose);      * Dose-Response Models;
  g_j=        g*(1+g_c*dose);
  alpha_j=alpha*(1+g_c*dose);
  mu1_j=    mu1*(1+mu1_c*dose);

  root=sqrt((g_j**2)+ (4*alpha_j*mu1_j));
  p_j=(-g_j-root)/2;
  q_j=(-g_j+root)/2;
  dw=dw* alpha_j;
  w= w* alpha_j;
    f = (w-p_j)*exp(q_j*(agestrt-agequit)) + (q_j-w)*exp(p_j*(agestrt-agequit));
    if dw=0 then do; dw=p_j*q_j; end;
    dfdt=dw*(exp(-q_j*(agequit-agestrt))-exp(-p_j*(agequit-agestrt)));
    dw= ((q_j-p_j)**2)*exp(-(p_j+q_j)*(agequit-agestrt))*dw/f/f/alpha_j;
    w = ((w-p_j)*q_j*exp(q_j*(agestrt-agequit))+(q_j-w)*p_j*exp(p_j*(agestrt-agequit)))
/f/alpha_j;
    ln_S= ((mu0_j*x)/alpha_j)*log((q_j-p_j)/f);

  IF LOOP=1 then sum_ln_Sentry=sum_ln_Sentry+ln_S;
  IF LOOP=2 then do;
    h_t= ((mu0_j*X)/alpha_j)*dfdt/f;
    sum_ln_Sexit=sum_ln_Sexit+ln_S;
    sum_h=sum_h+h_t;
  end;
end;    *<= END DO J;
end;    *<= END DO LOOP;

S_entry=exp(sum_ln_Sentry);
S_exit =exp(sum_ln_Sexit);
S_prime_exit=sum_h*S_exit;

ll = (leukemia=0)*log(S_exit/S_entry) + (leukemia=1)*log(S_prime_exit/S_entry);
model leukemia ~ general(LL); run;
```