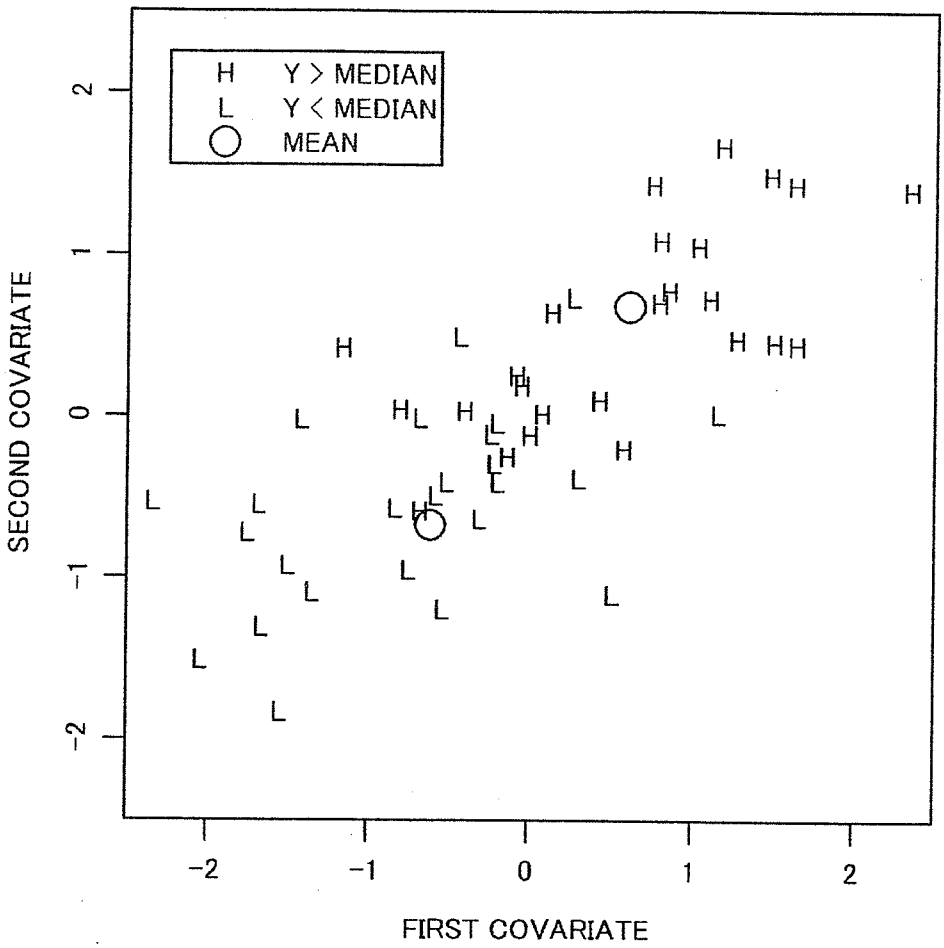

Figure 2. Scatter plot of $z^{(1)}$ and $z^{(2)}$ with correlation $\rho = 0.8$ and $N = 50$ and the sample means $\hat{\mathbf{m}}_H$ and $\hat{\mathbf{m}}_L$. The letters "H" or "L" represent covariates whose response is larger or smaller than the median. The two circles note the means.

In conclusion, the asymptotic approximation is good and not only the improved estimator $\hat{\boldsymbol{\eta}}_N$, but also our simple estimator $\hat{\boldsymbol{\eta}}$ may be expected to perform well even when the covariates are not independent.

5. DISCUSSION

Each element of $\hat{\boldsymbol{\eta}}$ might show the contribution of the corresponding covariate to the response because the difference between two means is close to zero if the covariate is independent of the response. However, the normalized unit vector of $\hat{\boldsymbol{\eta}}$ is interpreted as an estimator of the true EDR vector $\boldsymbol{\eta}_0$. We remark on four points: (1) the median of y_i is an adequate choice of t in general because each slice has the same number of response variables, but it can be replaced by a better value, for example, $t = 0.5$ for binary response; (2) the binary response δ is invariant for a

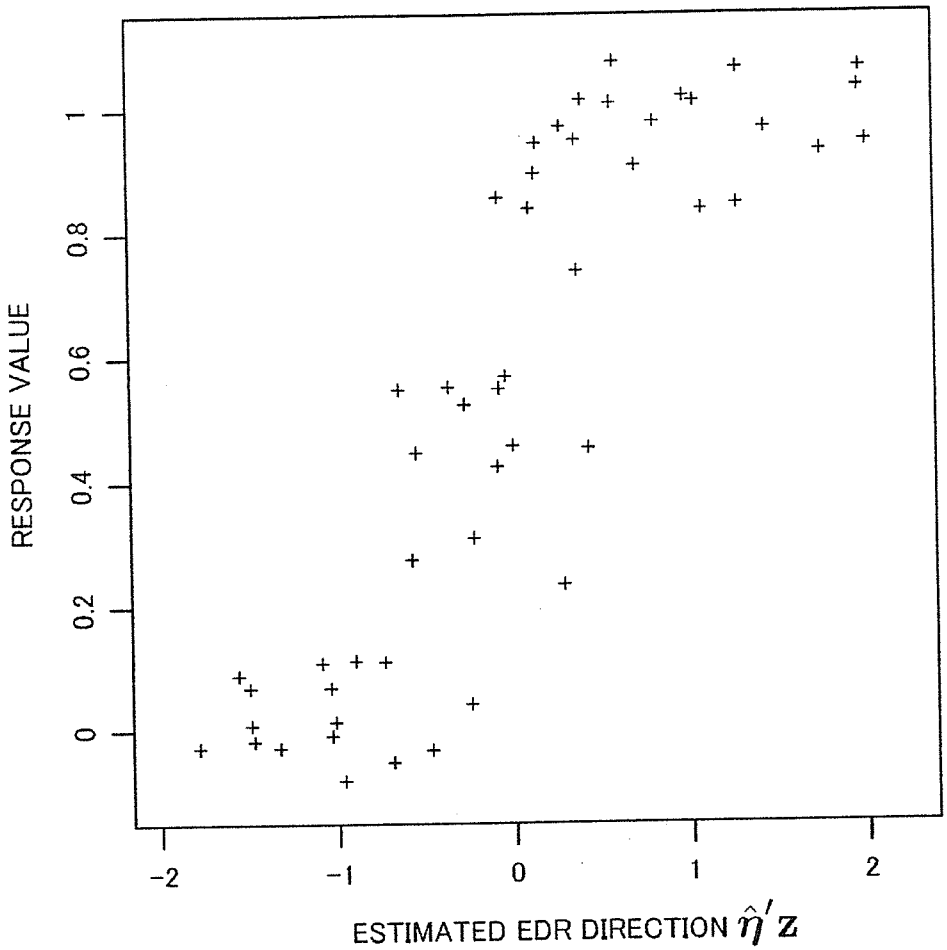

Figure 3. Scatter plot of the estimated EDR direction $\hat{\eta}'\mathbf{z}$ and the response y for $N = 50$ and $\rho = 0.8$.

monotone transformation of the original response variable, for example, the Box–Cox transformation including a logarithmic one (Box and Cox, 1964); (3) for small sample size, $\frac{N^{\frac{1}{2}}}{(N-1)^{\frac{1}{2}}}\hat{\eta}$ might be a better estimator of the correlation coefficient given in Proposition 1, which is corrected by the degrees of freedom on the covariance between δ and $z^{(j)}$ and the variance of δ; and (4) if there are some missing values, we calculate $\hat{\mathbf{m}}_H$ or $\hat{\mathbf{m}}_L$ without them and adjust the means by the number of available samples for each covariate.

Further, for the case when there exists the inverse matrix $\widehat{\Omega}^{-1}$, we have the consistent estimator of the true EDR vector given in Proposition 6, $\hat{\eta}_N = \widehat{\Omega}^{-1}\hat{\eta}$. Note that the matrix $\widehat{\Omega}$ might not be of full rank so $\widehat{\Omega}^{-1}$ is not defined when the number of covariates is large and the number of samples is small; then an alternative to $\widehat{\Omega}^{-1}$ is the generalized inverse matrix (e.g., Searle, 1982). Even when there does not exist the inverse matrix for the entire set of covariates, it might be possible to select a

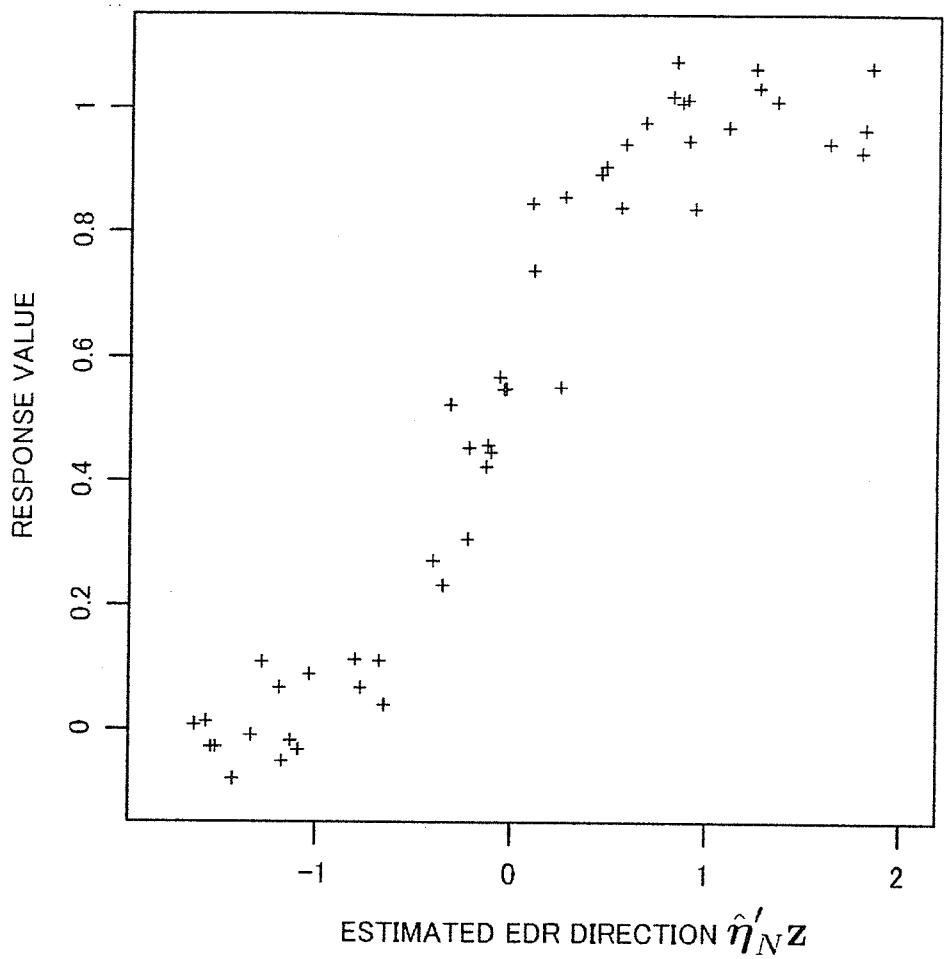

Figure 4. Scatter plot of the estimated EDR direction $\hat{\eta}_N' \mathbf{z}$ and the response y for $N = 50$ and $\rho = 0.8$, where $\hat{\eta}_N = \hat{\Omega}^{-1} \hat{\eta}$.

Table 1. The expectation and approximated standard deviation of $\hat{\eta}^{(j)}$ for $N = 50$ and $\rho = 0.0$ or 0.8 in 100,000 repetitions.

	$\rho = 0.0$			$\rho = 0.8$		
j	E	SD	∞	E	SD	∞
1	0.357	0.122	0.132	0.645	0.072	0.108
2	0.357	0.122	0.132	0.645	0.072	0.108
3	0.357	0.122	0.132	0.071	0.144	0.141
4	0.357	0.122	0.132	0.071	0.144	0.141
5	0.357	0.123	0.132	0.357	0.122	0.132
6	0.000	0.144	0.141	0.000	0.144	0.141

Note: $N^{-0.5} = 0.141$.

Table 2. Evaluation of the estimator by canonical correlation coefficients between $\hat{\eta}'z$ and $\eta_0'z$ and the estimated sample correlation coefficient between the projected value $\hat{\eta}'z$ and the binary response δ for $N = 50$, 100, or 500 and $\rho = 0.0$ or 0.8 in 100,000 repetitions.

	$\rho = 0.0$		$\rho = 0.8$	
N	$Cor(\hat{\eta}'z, \eta_0'z)$	$Cor(\hat{\eta}'z, \delta)$	$Cor(\hat{\eta}'z, \eta_0'z)$	$Cor(\hat{\eta}'z, \delta)$
50	0.936 (0.039)	0.803 (0.034)	0.921 (0.032)	0,769 (0.039)
100	0.967 (0.021)	0.799 (0.023)	0.935 (0.020)	0.762 (0.027)
500	0.993 (0.004)	0.798 (0.010)	0.946 (0.007)	0.758 (0.012)

Note: Numbers in parentheses represent standard deviations.

suitably small set of covariates by the test in Proposition 4, then $\hat{\eta}_N$ with the selected covariates can be computed.

Theoretically, the performance of $\hat{\eta}_N = \Omega^{-1}\hat{\eta}$ with $\mathbf{z} \sim (0_K, \Omega)$ and $\hat{\eta}$ with $\mathbf{z} = \Omega^{-\frac{1}{2}}\mathbf{x} \sim (0_K, I_K)$ is the same. However, the latter estimator is the coefficient of the multivariate standardized covariate. Therefore, each new covariate is different from the original one and it is difficult to interpret the meaning of the coefficient. Thus, we suggest using $\hat{\eta}_N = \widehat{\Omega}^{-1}\hat{\eta}$ without the multivariate transformation $\mathbf{z} = \widehat{\Omega}^{-\frac{1}{2}}\mathbf{x}$.

The monotony of the mean function is not the essential problem, and our theories can be extended directly to a piecewise monotone mean function, for example, a quadratic one.

APPENDIX

Proof of Proposition 1. From the definition of δ, $E(\delta) = 2Pr(y \geq t) - 1 = 0$ and $Var(\delta) = 1$.

Proof of Proposition 2. From Lemma 1, the $\delta_i\mathbf{z}_i$ are asymptotically independent. Then,

$$E(\hat{\eta}\hat{\eta}') = \frac{1}{N^2}\sum_{i=1}^{N} E(\delta_i^2\mathbf{z}_i\mathbf{z}_i') + \frac{2}{N^2}\sum_{i<j} E(\delta_i\mathbf{z}_i)(\delta_j\mathbf{z}_j)'$$

$$= \frac{1}{N}\Omega + \frac{N-1}{N}\eta_*\eta_*'.$$

Proof of Propositions 3 *and* 4. From Lemma 1, the δ_i are independent and the diagonal elements of Ω are one, which leads to the result.

Proof of Proposition 5. From Lemma 1 and Proposition 1, $Cor(\boldsymbol{\eta}'\mathbf{z}, \delta) = \boldsymbol{\eta}'\boldsymbol{\eta}_*(\boldsymbol{\eta}'\Omega\boldsymbol{\eta})^{-\frac{1}{2}}$, where $Cor(\boldsymbol{\eta}'\mathbf{z}, \delta) \geq 0$ without loss of generality. Solving $\frac{\partial}{\partial \boldsymbol{\eta}} \log Cor(\boldsymbol{\eta}'\mathbf{z}, \delta) = 0$, we have $\boldsymbol{\eta} = c\Omega^{-1}\boldsymbol{\eta}_*$, where c is a constant value. Because Ω is a symmetric matrix, $\lambda_1 \leq \frac{\boldsymbol{\eta}_*'\Omega\boldsymbol{\eta}_*}{\boldsymbol{\eta}_*'\boldsymbol{\eta}_*} \leq \lambda_K$ and the result follows.

Proof of Proposition 6. From Lemma 1 and Proposition 1, we can express $Cor_N(\boldsymbol{\eta}'\mathbf{z}, \delta) = \boldsymbol{\eta}'\hat{\boldsymbol{\eta}}(\boldsymbol{\eta}'\widehat{\Omega}\boldsymbol{\eta})^{-\frac{1}{2}}$. Solving $\frac{\partial}{\partial \boldsymbol{\eta}} \log Cor_N(\boldsymbol{\eta}'\mathbf{z}, \delta) = 0$, we have $\hat{\boldsymbol{\eta}}_N = c\widehat{\Omega}^{-1}\hat{\boldsymbol{\eta}}$, where c is a constant.

Proposition 3 implies $Cor(\hat{\boldsymbol{\eta}}_N'\mathbf{z}, \boldsymbol{\eta}_0'\mathbf{z}) = \hat{\boldsymbol{\eta}}'\widehat{\Omega}^{-1}\boldsymbol{\eta}_*/(\hat{\boldsymbol{\eta}}'\widehat{\Omega}^{-1}\Omega\widehat{\Omega}^{-1} \times \hat{\boldsymbol{\eta}} \cdot \boldsymbol{\eta}_*'\Omega^{-1}\boldsymbol{\eta}_*)^{\frac{1}{2}}$ converges to 1.

Proof of Propositions 7 *and* 8. Propositions 3 and 5 lead to the result.

ACKNOWLEDGMENTS

The authors thank Dr. J. B. Cologne for his helpful comments and the referees for their encouragement and valuable advice. This study was supported in part by Grant-in-Aid for Encouragement of Young Scientists from the Ministry of Education Science, Sport Culture of Japan.

REFERENCES

Box, G. P. E., Cox, D. R. (1964). An analysis of transformations (with discussion). *J. R. Statist. Soc. B* 26:211–252.

Cook, R. D., Lee, H. (1999). Dimension reduction in binary response regression. *J. Am. Statist. Assoc.* 94:1187–1200.

Duan, N., Li, K. C. (1991). Slicing regression: A link-free regression method. *Ann. Statist.* 19:505–530.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* 7:1–26.

Härdle, W., Hall, P., Ichimura, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* 21:157–178.

Hooper, J. (1959). Simultaneous equations and canonical correlation theory. *Econometrica* 27:245–256.

Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *J. Am. Statist. Ass.* 86:316–342.

Searle, S. R. (1982). *Matrix Algebra Useful for Statistics.* New York: John Wiley & Sons.

Xia, Y., Tong, H., Li, K. W., Zhu, L. X. (2002). An adaptive estimation of dimension reduction space (with discussion). *J. R. Statist. Soc. B* 64:1–28.

Biometrical Journal 46 (2004) 1, zzz–zzz / DOI 10.1002/bimj.200410000

Incorporation of Inter-individual Heterogeneity into the Multistage Carcinogenesis Model: Approach to the Analysis of Cancer Incidence Data

S. Izumi[*, 1, 2] and **M. Ohtaki**[3]

[1] Department of Computer Science and Intelligent Systems, Faculty of Engineering, Oita University, 700 Dannoharu, Oita, 870-1192, Japan
[2] Department of Statistics, Radiation Effects Research Foundation, 5-2 Hijiyama Park Minami-ku, Hiroshima 732-0815, Japan
[3] Department of Environmetrics and Biometrics, Research Institute for Radiation Biology and Medicine, Hiroshima University, 1-2-3 Kasumi Minami-ku, Hiroshima 734-8553, Japan

Received zzz, revised zzz, accepted zzz

Summary

We investigate a multistage carcinogenesis frailty model to incorporate inter-individual heterogeneity into carcinogenic response. Attention is focused on inference concerning the effects of different sources of population heterogeneity on cancer rates. The authors consider unobserved variability arising from either carcinogen exposure or background characteristics. Gamma and Inverse-Gaussian distributions are selected for frailty models, and the baseline hazard function is the generalized Armitage-Doll model (i.e. non-frailty model) in which exposure effects shift the age scale instead of acting multiplicatively on cancer rates. For illustration, we apply the method to solid cancer data from a cohort of atomic bomb survivors to examine some features of proposed models. The results show that the Gamma frailty model for the heterogeneity of baseline rates provides the best goodness-of-fit of the model and a non-zero frailty variance. Parameter estimates are, for the most part, comparable between the Gamma and Inverse-Gaussian frailty models. In a heterogeneous population the exposure effects on young adulthood cancer rates might be underestimated for the non-frailty model. Meaningful information regarding each source of heterogeneity has been provided by the proposed method. Therefore, the multistage carcinogenesis frailty model approach is useful for analyses of epidemiological cancer data to assess population heterogeneity and heterogeneity-influenced exposure effects.

Key words: Human exposure; Risk assessment; Analysis of follow-up data; Armitage-Doll model; Random effect model; Cancer epidemiology.

1 Introduction

A framework for understanding the time course of carcinogenesis originates with two main stochastic models: the Armitage-Doll multistage model (Armitage and Doll, 1954, 1957) and the Moolgavkar-Venzon-Knudson (MVK) two-stage clonal expansion model (Knudson, 1971; Moolgavkar and Venzon, 1979; Moolgavkar and Knudson, 1981). Both models are based on the theories of Muller (1951) and Nordling (1953). MVK two-stage clonal expansion model involves two phases: initiation, and malignant conversion with progression. Armitage and Doll (1954, 1957) considered that cancer was the end result of the accumulation in a normal cell of a critical number (k, for example) of independent transitions through a series of intermediate states. As shown in Figure 1, first a normal stem cell is initiated as a Poisson process (initiation), the initiated cell goes through $(k-1)$ sequential stages to become a malignant cell (malignant conversion), and finally, after a certain lag time during which progression occurs,

* Corresponding author: e-mail: shizue@csis.oita-u.ac.jp, Phone: +81 97 554 7867, Fax: +81 97 554 7886

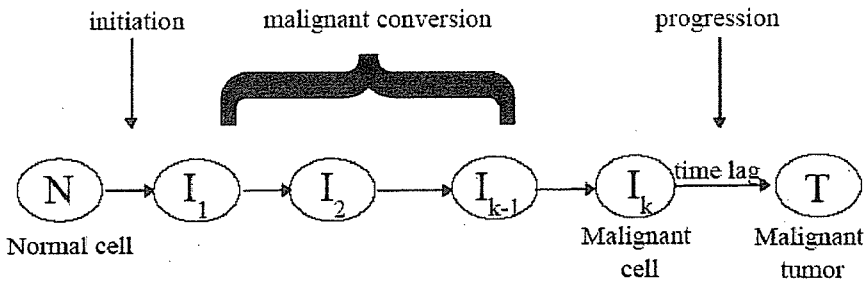

Figure 1 Schematic representation of Armitage-Doll multistage model for carcinogenesis.

the malignant cell becomes part of an observable tumor. For most cancers between 5 and 7 stages are indicated for this model (Doll, 1971). An implicit assumption of virtually all analyses using the Armitage-Doll multistage and MVK two-stage clonal expansion models in the past is homogeneity across the population in the underlying transition rates and their dependence on exposures to carcinogens. Moolgavkar et al. (1999) and Tan (1991) provide excellent mathematical reviews of these models. Ohtaki and his colleagues (Ohtaki, 1981; Ohtaki et al., 1985; Ohtaki and Niwa, 2001) developed a generalized Armitage-Doll model in which the effects of exposure shift the age scale instead of acting multiplicatively on cancer rates, and those effects can be specified by age, dose, and age at exposure. Further, Pierce and Vaeth (2003) illustrated the generalized Armitage-Doll model approach using cancer incidence data among atomic bomb survivors, miners exposed to radon, and cigarette smokers, showing that the model consequences conform well to these data. There are more recent models that do not fall within the Armitage-Doll or MVK frameworks (Little, 1995; Aalen and Tretli, 1999; Luebeck and Moolgavkar, 2002; Nowak et al., 2002; Little and Wright, 2003).

Frailty models provide a useful alternative to standard hazard models when a standard model fails to adequately account for all variability in the observed failure times. Since their introduction by Vaupel et al. (1979), frailty models have been applied to failure-time data in an attempt to account for population heterogeneity resulting from unobserved underlying characteristics of individuals (Andersen et al., 1993; Aalen, 1994; Hougaard, 1995). With recent advances in clinical and epidemiological research, increasing attention is being given to heterogeneity in cancer susceptibility within populations. Although genetic background affects susceptibility to cancer (Caporaso, Debaun, and Rothman, 1995; Feigelson et al., 1996), it is difficult to collect genetic data in population-based epidemiological studies to detect rare gene-environment interactions (Andrieu and Goldstein, 1998). Nevertheless, to assess the true effects of carcinogens, as well as lifestyle factors, the genetic heterogeneity of a population should be incorporated into analyses.

In this paper we examine a frailty model approach for multistage carcinogenic process to test the underlying assumption of population homogeneity when unobserved inter-individual variability arises from either carcinogen exposure or background characteristics. Starting with a review of the generalized Armitage-Doll multistage model (non-frailty model), we propose Gamma and Inverse-Gaussian frailty models for each source of inter-individual variability with the inferential method in Section 2. For illustration, we apply the method to solid cancer data from a cohort of male atomic bomb survivors to examine some features of proposed models in Section 3. Finally, we discuss the implications of our findings.

2 Multistage Carcinogenesis Model

2.1 Generalized Armitage-Doll multistage model

In the Armitage-Doll framework, a stem cell becomes malignant after passing through k stages. Extending the framework to an age-homogeneous Poisson process, the cancer rates at which a malignant cell occurs are asymptotically of the form μt^{k-1} as μ goes to zero, where μ is a function of background

characteristics, and t denotes attained age (years). There is some significant time lag between the emergence of a malignant cell and the development of the first observable clinical malignant tumor, which is referred to hereafter as tumor growth time.

The generalized Armitage-Doll multistage model is proposed based on a series of previous reports (Ohtaki et al., 1985; Pierce and Mendelsohn, 1999; Ohtaki and Niwa, 2001; Pierce and Vaeth, 2003), extending the Armitage-Doll framework to an age-heterogeneous Poisson process. It is not an accelerated failure-time model, and it requires several assumptions summarized by Pierce and Vaeth (2003). When a subject receives a single exposure of dose D at age a (years), the hazard function $h_{AD}(t \mid D,a)$ of the first observable malignant tumor for an individual at attained age t is expressed as:

$$h_{AD}(t \mid D,a) = \mu\left[t+\gamma+\beta D\{1+\omega(a)\}\right]^{k-1},$$

where $\beta(\geq 0)$ is the coefficient of the dose effect, γ is the total excess age effect due to heightened sensitivity in childhood, and $\omega(a)$ is a monotone decreasing function of age at exposure denoting the sensitivity in terms of the occurrence of cellular events. As age increases, $\omega(a)$ goes to zero. In the analysis of an aged cohort (such as the atomic bomb survivors), $\omega(a)$ can be assumed to be zero. Frailty becomes important due to non-zero values of $\omega(a)$ at young ages. The expressions for the hazard functions are approximations to the exact solution. Assuming the same value γ for the tumor growth time, we use the hazard function of cancer risk as follows:

$$h_{AD}(t-\gamma \mid D,a) \cong \mu\left[t+\beta D\{1+\omega(a)\}\right]^{k-1} \cong \mu t^{k-1}\left[1+\frac{(k-1)\beta D\{1+\omega(a)\}}{t}\right].$$

The resulting cancer risk hazard function does not depend on γ; so we rewrite the right hand term as $h_{AD}(t \mid D,a)$, again for simplicity of expression. These approximations are adequate in the left tail of the distribution for waiting-time under a heterogeneous Poisson process (Ohtaki, 1981). Here, exposure effects do not act multiplicatively on the hazard rates; instead, the exposure shifts the age scale by $\beta D\{1+\omega(a)\}$. That is, the exposure accelerates the aging process. The relative risk (RR), the rate ratio between the unexposed and exposed subjects, can then be expressed as

$$RR_{AD}(D \mid t,a) = 1+(k-1)\beta D\{1+\omega(a)\}/t$$

for $t \geq a > 0$. That is, RR increases proportionally to dose, but it decreases toward a value of one with increasing age.

2.2 Multistage frailty model

2.2.1 Heterogeneity of dose response

The latent random variable Z as a measure of frailty is incorporated into the hazard function to account for heterogeneity arising from carcinogen exposure, either due to the measurement error of exposure or the reaction of the body to exposure. The individual hazard can be written as

$$h(t \mid z,D) = h_{AD}(t \mid 0) + z\{h_{AD}(t \mid D,a) - h_{AD}(t \mid 0)\} = \mu t^{k-1}\left[1+\frac{(k-1)\beta D\{1+\omega(a)\}z}{t}\right],$$

which is conditional on the individual frailty under $Z=z$. The subjects with $Z>1$ (i.e., the more frail subjects) face an increased risk, while the subjects with $Z<1$ are less frail than their counterparts. This

000000 Biometrical Journal Heft 1/2004 Art.: 0000/Autor
C:\izumi\kaken\ohtaki2\project\paper\biomj\resubmit4\izumi_2nd_revision_BIMJ.200512141.doc
insgesamt 13 Seiten Diskette 05.11.2006 Bearb.:

frailty is not directly estimated from the data; instead, the frailty distribution is assumed to have unit mean and finite variance σ^2, and the population functions are used for data analysis.

Assuming the Gamma distribution ($\Gamma(\sigma^{-2},\sigma^2)$) for Z, the following density $f.(.)$, hazard $h.(.)$, and survival $S.(.)$ functions of the Gamma (G) frailty model are obtained:

$$\begin{aligned} f_{G1}(t\mid D,a) &= \int_0^{+\infty} h(t\mid z,D)\exp\left\{-\int_0^t h(u\mid z,D)du\right\}g(z)dz \\ &= \mu t^{k-1}\left(1+\frac{(k-1)\beta D\{1+\omega(a)\}}{t}\left[1+\sigma^2\mu t^{k-1}\beta D\{1+\omega(a)\}\right]^{-1}\right) \\ &\quad\times\exp(-\mu k^{-1}t^k)\left[1+\sigma^2\mu t^{k-1}\beta D\{1+\omega(a)\}\right]^{-(1/\sigma)^2}, \end{aligned}$$

$$\begin{aligned} h_{G1}(t\mid D,a) &= f_{G1}(t\mid D,a)/\int_t^{+\infty} f_{G1}(u\mid D,a)du \\ &= \mu t^{k-1}\left(1+\frac{(k-1)\beta D\{1+\omega(a)\}}{t}\left[1+\sigma^2\mu t^{k-1}\beta D\{1+\omega(a)\}\right]^{-1}\right) \\ &= h_{AD}(t\mid 0)+\{h_{AD}(t\mid D,a)-h_{AD}(t\mid 0)\}\left[1+\sigma^2\{H_{AD}(t\mid D,a)-H_{AD}(t\mid 0)\}\right]^{-1}, \end{aligned}$$

$$\begin{aligned} S_{G1}(t\mid D,a) &= \exp\left[-\int_0^t h_{G1}(u\mid D,a)du\right] \\ &= \exp(-\mu k^{-1}t^k)\left[1+\sigma^2\mu t^{k-1}\beta D\{1+\omega(a)\}\right]^{-(1/\sigma)^2} \\ &= S_{AD}(t\mid 0)\left[1+\sigma^2\{H_{AD}(t\mid D,a)-H_{AD}(t\mid 0)\}\right]^{-(1/\sigma)^2}, \end{aligned}$$

where $g(z)$ is the density function of Z, and $H_{AD}(t\mid.)$ and $S_{AD}(t\mid.)$ are the cumulative hazard function and survival function for the generalized Armitage-Doll model, respectively. RR can then be expressed as

$$RR_{G1}(D\mid t,a,\sigma^2) = 1+(RR_{AD}(D\mid t,a)-1)\left[1+\sigma^2\{H_{AD}(t\mid D,a)-H_{AD}(t\mid 0)\}\right]^{-1}.$$

Similarly, assuming the Inverse-Gaussian distribution ($IG(1,\sigma^{-2})$) for Z, the population hazard function of the Inverse-Gaussian (IG) frailty model can be obtained as

$$\begin{aligned} h_{IG1}(t\mid D,a) &= \mu t^{k-1}\left(1+\frac{(k-1)\beta D\{1+\omega(a)\}}{t}\left[1+2\sigma^2\mu t^{k-1}\beta D\{1+\omega(a)\}\right]^{-1}\right) \\ &= h_{AD}(t\mid 0)+\{h_{AD}(t\mid D,a)-h_{AD}(t\mid 0)\}\left[1+2\sigma^2\{H_{AD}(t\mid D,a)-H_{AD}(t\mid 0)\}\right]^{-1}. \end{aligned}$$

RR can then be expressed as

$$RR_{IG1}(D\mid t,a,\sigma^2) = 1+(RR_{AD}(D\mid t,a)-1)\left[1+2\sigma^2\{H_{AD}(t\mid D,a)-H_{AD}(t\mid 0)\}\right]^{-1}.$$

2.2.2 Heterogeneity of baseline rates

Prior to the present work, heterogeneity of baseline rates in the male cancer data was examined by Aalen and Tretli (1999) using a compound Poisson frailty model. In this report the Gamma and Inverse-

Gaussian frailty models are used for the heterogeneity of baseline rates to facilitate comparison of the results with heterogeneity of the dose response.

When we consider heterogeneity from background characteristics due to unobserved environmental or genetic factors, the individual hazard function with a frailty variable $Z = z$ can be expressed as

$$h(t \mid z, D) = \{z h_{AD}(t \mid 0)\} RR_{AD}(D \mid t, a) = z\mu t^{k-1}\left[1 + \frac{(k-1)\beta D\{1+\omega(a)\}}{t}\right].$$

The population hazard function of the Gamma and Inverse-Gaussian frailty models can then be obtained as:

$$h_{G2}(t \mid D, a) = \mu t^{k-1}\left[1 + \frac{(k-1)\beta D\{1+\omega(a)\}}{t}\right]\left(1 + \sigma^2 \mu t^k / k\left[1 + \frac{k\beta D\{1+\omega(a)\}}{t}\right]\right)^{-1}$$
$$= h_{AD}(t \mid D, a)\left[1 + \sigma^2 H_{AD}(t \mid D, a)\right]^{-1}$$

and

$$h_{IG2}(t \mid D, a) = \mu t^{k-1}\left[1 + \frac{(k-1)\beta D\{1+\omega(a)\}}{t}\right]\left(1 + 2\sigma^2 \mu t^k / k\left[1 + \frac{k\beta D\{1+\omega(a)\}}{t}\right]\right)^{-1/2}$$
$$= h_{AD}(t \mid D, a)\left[1 + 2\sigma^2 H_{AD}(t \mid D, a)\right]^{-1/2},$$

respectively. Hougaard (2000) provides general formulas for the Gamma and Inverse-Gaussian frailty models incorporating heterogeneity of baseline rates. RR can also be expressed as

$$RR_{G2}(D \mid t, a, \sigma^2) = RR_{AD}(D \mid t, a)\left(\frac{1 + \sigma^2 H_{AD}(t \mid 0)}{1 + \sigma^2 H_{AD}(t \mid D, a)}\right),$$

and

$$RR_{IG2}(D \mid t, a, \sigma^2) = RR_{AD}(D \mid t, a)\left(\frac{1 + 2\sigma^2 H_{AD}(t \mid 0)}{1 + 2\sigma^2 H_{AD}(t \mid D, a)}\right)^{1/2}.$$

When the frailty variance (σ^2) tends to zero (i.e., $z = 1$ for all subjects), the population and individual hazard functions coincide, and the RR for each frailty model converge to that of the non-frailty model.

2.3 Parameter estimation and hypotheses tests

We construct likelihood functions of the models for individual-based, right-censored failure time data. Given the age at the last follow-up t_i and the age at entry $a_i (> 0)$, the likelihood function can be expressed as

$$L. = \prod_{i \in E} f.(t_i \mid .) / S.(a_i \mid .) \prod_{i \in C} S.(t_i \mid .) / S.(a_i \mid .),$$

where E is the set of indices for failed observations and C is the set of indices for censored observations. Estimation of parameters is performed by maximum likelihood methods using the Spider-II algorithm (Ohtaki and Izumi, 1999). Estimated standard errors are computed from the observed information matrices (Efron and Hinkley, 1978), and are used to compute 95 percent Wald-test-based confidence intervals (CIs) of parameter estimates for this simple illustration. We compute the Akaike Information Criteria (AIC) (Akaike, 1973) as an index of goodness-of-fit. The AIC is minus two times (the number of

parameters minus the logarithm of the maximum likelihood). Regarding the frailty variance (σ^2), the null hypothesis is $H_0^\sigma : \sigma^2 = 0$, and the alternative hypothesis is $H_a^\sigma : \sigma^2 > 0$. Since the null value is on the boundary of the parameter space, we approximate minus two times the logarithm of the likelihood ratio as a 50:50 mixture distribution of χ_0^2 and χ_1^2 (Self and Liang, 1987). All calculations are performed with STATA software (Stata Co., TX, 2003). Macro files used for our analysis might be available upon request.

3 Application to Real Data

3.1 Radiation-related solid cancer incidence data set

We illustrate the above method by analyzing solid cancer data from the cohort study of atomic bomb survivors conducted at the Radiation Effects Research Foundation, Japan between 1958-1996 (Pierce and Preston, 2000). This data set has two desirable features: a large sample size and a strong dose-response relationship to explore whether the population heterogeneity modifies the baseline rates or the dose response. The study cohort includes a baseline sample of 22,638 men who were alive and cancer-free in January 1958, under 40 years of age in either Hiroshima or Nagasaki at the time of the bombings in August 1945, and whose radiation dose was estimated using the DS86 dosimetry system (Roesch, 1987). We used a weighted colon dose (the gamma dose plus 10 times the neutron dose) estimated in the DS86 dosimetry system with adjustment for dose error (Pierce, Stram, and Vaeth, 1990). Mean colon dose was 114 millisieverts (mSv); thus, doses measured in units of 100 mSv were used to examine a linear dose response. Mean age at diagnosis was 62 years (range 25 – 90 years). Cancers in gender-specific organs and the thyroid were excluded from analysis, because the incidence of these cancers differs greatly from most solid cancers in terms of radiation effects. Thus, the analysis here involves solid cancer in nine sites: esophagus, stomach, colon, rectum, liver, gallbladder, pancreas, lung, and urinary bladder. Background rates were adjusted for city and birth year. As seen in the previous reports, effects of radiation exposure decreased with increasing age at exposure, a phenomenon also seen in the present data analysis. In order to simplify the presentation of dose response for this illustration, the term $\beta\{1+\omega(a)\}$ was replaced with β. Detailed results for the age-at-exposure-specific dose response can be obtained from the authors.

First, we compared the results from the non-frailty model with those from frailty models, using solid cancer as cases. Next, we examined whether adjustment for exposure sensitivity on dose response or adjustment for another risk factor on baseline rates would affect the estimates of frailty variance and dose response. For this purpose, we additionally used the data of severe epilation (Sposto, Stram, and Awa, 1987) and smoking status (Pierce, Sharp, and Mabuchi, 2003) based on self-administered questionnaires. Severe epilation (no, yes) was considered to be an indicator of radiation sensitivity for dose response, and it was treated as an effect modifier of the dose response in the present paper. It has other interpretations, such as a surrogate marker for unmeasured genetic factors related to susceptibility to ionizing radiation or an indicator of exposure measurement error. On the other hand, smoking status (never, current, past) was considered to be another risk factor for the baseline rates, and it was treated as a main effect in the baseline rate model. Lack of adjustment for the epilation and smoking variables may result in a large frailty variance. It could be due to the misspecification of the form of the model (Izumi and Ohtaki, 2004) as well as latent biological variability in susceptibility due to genetic or environmental factors (Haiman et al., 2006). Thus, the results for the frailty variance should be interpreted cautiously. Information on epilation and smoking status was available for 22,435 (99%) and 16,024 (71%) subjects, respectively. Analysis was conducted for the subjects with known such status. Dose response for colon cancer among subjects with severe epilation was higher than that among subjects without severe epilation ($p < 0.05$). Among unexposed subjects, lung cancer rates for current or past smokers were higher than those for non-smokers ($p < 0.01$). Thus, colon and lung cancers were additionally used as cases to examine effects of these adjustments in the Gamma frailty models.

Table 1 Estimates of number of stages (k), linear dose response (β) with 95% CI, frailty variance (σ^2) with p values, and AIC.

Solid cancer	k	β	(95% CI)	σ^2	P value for $H_0 : \sigma^2 = 0$	AIC
Non-frailty model	7.0	0.49	(0.32, 0.65)			33192.80
Frailty models						
1. heterogeneity of dose response	7.0	0.62	(0.34, 0.90)	4.2	0.058	33188.33
Gamma frailty	7.0	0.62	(0.34, 0.90)	2.1	0.058	33188.33
Inverse-Gaussian frailty						
2. heterogeneity of baseline rates						
Gamma frailty	8.2	0.54	(0.34, 0.74)	2.2	<0.001	33130.29
Inverse-Gaussian frailty	8.4	0.54	(0.34, 0.74)	3.6	<0.001	33136.85

3.2 Results

3.2.1 Comparison of non-frailty and frailty models

Table 1 shows the estimates of number of stages (k), linear dose response (β) with 95 percent CIs, frailty variance (σ^2) with p values, and AIC. Generally, parameter estimates and p values are consistent between the Gamma and Inverse-Gaussian frailty models. Thus, we compare some aspects among the Gamma frailty models and the non-frailty model.

The estimated number of stages in the non-frailty model is seven, which is the same as that of the frailty models for heterogeneity of dose response and slightly lower than that of the frailty models for heterogeneity of baseline rates. These estimates in the present analysis are similar to those shown by Armitage and Doll (1954, 1957) in their analyses of cancer mortality data. Because the number of stages may differ in different sites of cancer, the estimated values for solid cancer can be interrupted as an average of stages for nine sites.

Frailty models show a somewhat higher linear dose response than the non-frailty model. For example, the attained age, t (years), is increased by the 100 mSv irradiation to $(t+0.49)$ under the non-frailty model, which is consistent with the previously report (Pierce and Vaeth, 2003). Under the frailty models the attained age is increased to $(t+0.62)$ for assuming heterogeneity of dose response and $(t+0.54)$ for assuming heterogeneity of baseline rates. Allowance for frailty also increases the upper 95% confidence limit, while the lower 95% confidence limit varies little.

To illustrate the age-dependent exposure effects, relationships of age and RR following 100 mSv irradiation are shown in Figure 2. RR of the frailty models is derived from the population hazard functions. Generally, RR decreases toward the value of one with increasing age, which is one characteristic of radiation effects found in the studies of these subjects. Estimated RR from the frailty models is about same at younger ages and higher than that at younger ages from the non-frailty model (eg., $RR=1.2$ from the frailty models, $RR=1.15$ from the non-frailty model at age 20 years), which implies that the exposure effects on young adulthood cancer rates might have been underestimated in the non-frailty model. As less frail subjects would tend to remain in the population at advanced ages, it might explain why RR from the frailty models was higher at younger ages and lower at advanced ages than that from the non-frailty model.

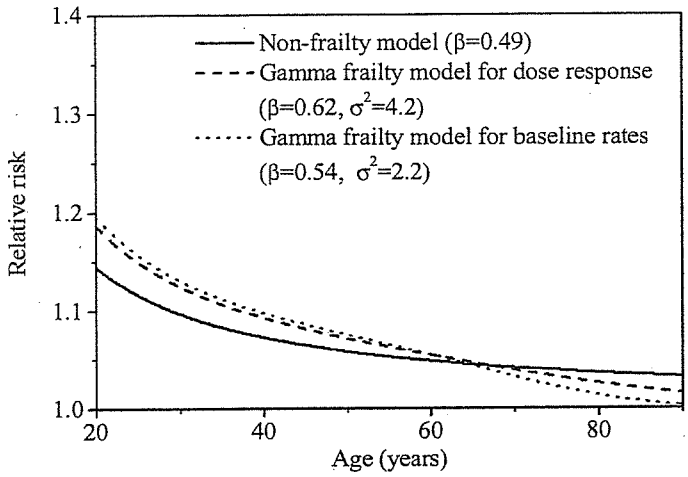

Figure 2 Relationship of age with the relative risk of solid cancer following a 100 mSv irradiation among men.

P values from the test regarding frailty variance suggest that the present data have contained population heterogeneity, though the observed frailty variance is fairly small. For example, estimated frailty variance of the Gamma frailty model is 4.2 ($p = 0.058$) for heterogeneity of dose response and 2.2 ($p < 0.001$) for heterogeneity of baseline rates. Although the observed frailty variance differs between the Gamma and Inverse-Gaussian frailty models, p values are consistent between these models.

Incorporation of population heterogeneity improves the goodness-of-fit of the model. The Gamma frailty model for heterogeneity of baseline rates results in the best goodness of fit among the models considered, with an AIC reduction of 62.51 from the non-frailty model. Thus, observed population heterogeneity is less likely due to heterogeneity of dose response.

3.2.2 Adjustment for population heterogeneity

Tables 2 and 3 show how the adjustment of exposure sensitivity for dose response or the adjustment of smoking status for baseline rates affects the estimates of frailty variance and dose response. Adjustment for severe epilation on dose response reduces the estimated frailty variance by 30% (from 14.6 to 10.3) for colon cancer in the Gamma frailty model for heterogeneity of dose response. Similarly, adjustment for smoking status on baseline rates reduces the estimated frailty variance by 40% (from 14.2 to 8.5) for lung cancer in the Gamma frailty model for heterogeneity of baseline rates. On the other hand, while such adjustment does not improve the goodness-of-fit in the non-frailty model, the estimated frailty variance remains the same or increases after adjustment. A non-zero frailty variance after adjustment may suggest the existence of missing factors explaining residual population heterogeneity. It is also noted that the estimated frailty variance is much larger for colon and lung cancers than solid cancer, which may imply that cancer in different sites has different sources of heterogeneity so that merging several sites into a broad category of cancer may lead the population heterogeneity to null. Thus, the use of specific cancer sites is suggested for identifying population heterogeneity.

In terms of dose response, exposure effects may have been slightly overestimated when the baseline rates were not adjusted for smoking status. Similarly, exposure effects for subjects without severe epilation may have been overestimated, while those for subjects with severe epilation may have been underestimated, similar to that seen in previous report (Sposto et al., 1987). Such different dose response is fur-

ther emphasized by the frailty model, since population heterogeneity regarding radiation sensitivity is taken into account by the model. The best goodness-of-fit is achieved by the Gamma frailty model for heterogeneity of baseline rates before and after adjustment.

4 Discussion

We considered a multistage carcinogenesis frailty model method to account for inter-individual heterogeneity in epidemiological cancer data. Results from the present study suggested that the frailty models, especially the Gamma frailty model for heterogeneity of baseline rates, provided a large improvement of goodness-of-fit and that the choice of frailty distribution to the range of the models considered did not greatly affect the parameter estimates. Also indicated was the exposure effects on young adulthood cancer rates might have been underestimated for the non-frailty model. The proposed frailty models succeeded in assessing population heterogeneity originating from either exposure sensitivity of dose response or an unobserved risk factor of baseline rates.

Table 2 Adjustment of radiation sensitivity for dose response among subjects with known status.

	Estimate of β	(95% CI)	Estimate of σ^2	P value for $H_0 : \sigma^2 = 0$	AIC
1. Lung cancer					
Non-frailty model					
Not adjusted	0.39	(0.03, 0.75)			5927.93
Adjusted for severe epilation: no	0.31	(-0.09, 0.71)			5925.31
yes	0.62	(-0.13, 1.36)			
Gamma frailty model for heterogeneity of dose response					
Not adjusted	0.49	(-0.05, 1.03)	20.6	0.250	5925.48
Adjusted for severe epilation: no	0.31	(-0.24, 0.86)	62.5	0.069	5921.11
yes	2.27	(-2.50, 7.05)			
Gamma frailty model for heterogeneity of baseline rates					
Not adjusted	0.45	(0.03, 0.86)	9.2	0.002	5917.61
Adjusted for severe epilation: no	0.32	(-0.12, 0.75)	9.6	0.001	5914.35
yes	0.91	(-0.26, 2.07)			
2. Colon cancer					
Non-frailty model	0.66	(0.28, 1.05)			5709.82
Not adjusted	0.40	(0.01, 0.78)			5702.12
Adjusted for severe epilation: no	1.44	(0.48, 2.40)			
yes					
Gamma frailty model for heterogeneity of dose response					
Not adjusted	0.82	(0.22, 1.41)	14.6	0.180	5706.98
Adjusted for severe epilation: no	0.44	(-0.01, 0.90)	10.3	0.193	5699.36
yes	1.86	(0.16, 3.55)			
Gamma frailty model for heterogeneity of baseline rates					
Not adjusted	0.77	(0.24, 1.30)	20.0	0.001	5698.92
Adjusted for severe epilation: no	0.47	(-0.02, 0.96)	17.2	0.003	5692.36
yes	2.00	(0.08, 3.92)			

000000 Biometrical Journal Heft 1/2004 Art.: 0000/Autor
C:\izumi\kaken\ohtaki2\project\paper\biomj\resubmit4\izumi_2nd_revision_BIMJ.200512141.doc
insgesamt 13 Seiten Diskette 05.11.2006 Bearb.:

Table 3 Adjustment of smoking status for baseline rates among subjects with known status.

	Estimate of β	(95% CI)	Estimate of σ^2	P value for $H_0 : \sigma^2 = 0$	AIC
1. Lung cancer					
Non-frailty model					
Not adjusted	0.29	(-0.07, 0.66)			4407.49
Adjusted for smoking	0.20	(-0.11, 0.52)			4311.00
Gamma frailty model for heterogeneity of dose response					
Not adjusted	0.42	(-0.22, 1.06)	42.1	0.232	4404.95
Adjusted for smoking	0.33	(-0.24, 0.90)	55.1	0.208	4308.34
Gamma frailty model for heterogeneity of baseline rates					
Not adjusted	0.36	(-0.08, 0.80)	14.2	<0.001	4393.88
Adjusted for smoking	0.25	(-0.11, 0.62)	8.5	0.002	4300.40
2. Colon cancer					
Non-frailty model	0.70	(0.28, 1.11)			4361.76
Not adjusted	0.71	(0.29, 1.13)			4356.07
Adjusted for smoking					
Gamma frailty model for heterogeneity of dose response					
Not adjusted	1.06	(0.21, 1.91)	25.5	0.071	4357.61
Adjusted for smoking	1.07	(0.23, 1.92)	24.9	0.068	4351.86
Gamma frailty model for heterogeneity of baseline rates					
Not adjusted	0.95	(0.26, 1.64)	29.3	<0.001	4346.68
Adjusted for smoking	0.95	(0.26, 1.64)	29.1	<0.001	4341.59

Several motives lead us to postulate the existence of population heterogeneity in the carcinogenesis process. One is unmeasured variability in genetic susceptibility factors (inherited or non-inherited), such as gene polymorphisms related to cancer (Hayashi et al., 2006). Another is unmeasured environmental risk factors such as smoking (Pierce et al., 2003) and diet (Sauvaget et al., 2005). In addition, uncertainties in assessment of environmental exposures, such as random errors in dose estimates (Bennett, Little, and Richardson, 2004), can contribute to population heterogeneity. Another is possible misspecification of the form of the model, such as departure from the assumed model (Izumi and Ohtaki, 2004). The results of analysis with the frailty model may allow for making inference regarding the population heterogeneity, but with censored failure time data among unrelated individuals it is difficult to identify the source of heterogeneity unless further evidence is obtained.

The question whether there is evidence of variation in individual susceptibility to radiation-related cancer risk has been investigated for three decades using the atomic-bomb survivor data. Stewart and Kneale (1984, 2000) suggested that the level of susceptibility to cancer might differ among the survivors exposed to atomic bomb radiation due to damage to the immune system. As an extension of analyses performed by Neriishi et al. (1991) and Stewart and Kneale (2000), this fundamental scientific question was thoroughly examined by Little (2002) using various analyses based on early radiation injury and mortality data among atomic bomb survivors. In all of these analyses—including the present one—variability in individual susceptibility to radiation exposure might be confounded with random dosimetric errors. Although our proposed approaches depend on a fairly strong modeling assumption such as the multistage carcinogenesis model and the problem of the random dosimetric errors remains, the results of our data analysis might provide possible evidence of population heterogeneity (i.e. variation in individual susceptibility to cancer risk).

Regression analyses of epidemiological cancer data include uncertainty associated with the models used to derive the estimates. The process of uncertainty analysis can be used to assess the relative importance of various sources of uncertainty (Matanoski et al., 2001). Although generalized Cox proportional model and Poisson model are flexible for baseline-rate shapes, the proposed method can provide useful guidance to examine the possibility of population heterogeneity in epidemiologic cancer data. Izumi and Ohtaki (2004) showed that predicted cancer rates from the Gamma frailty model for heterogeneity of baseline rates were closer to observed baseline cancer rates in all ages than the Armitage-Doll model, particularly for advanced ages. This implies that frailty models could capture trends of observed baseline cancer rates. An important future extension of the method is to incorporate two sources of heterogeneity (one related to baseline rates, the other related to dose response) together into the present multistage carcinogenesis model. An analysis with the two-component frailty model would reveal not just the relative magnitude of the two sources of heterogeneity, but also whether they are positively or negatively correlated and by how much. The present analysis does not adjust each effect of the population heterogeneity for the other. Additionally, effects of apoptosis on the carcinogenic process and individual variation of tumor growth time need to be considered.

In conclusion, given the performance of the proposed models, we suggest that our approach can be useful for assessing population heterogeneity and heterogeneity-influenced exposure effects in epidemiological cancer data.

Acknowledgements This publication is based on research performed at the Radiation Effects Research Foundation (RERF), Hiroshima and Nagasaki, Japan. RERF is a private nonprofit foundation funded equally by the Japanese Ministry of Health, Labor and Welfare and the United States Department of Energy through the National Academy of Sciences. Supported in part by Grants-in-Aid for Cancer Research from the Ministry of Education, Culture, Sports, Science, and Technology of Japan and Research Program on "Low-Dose Radiation Effects Based on Molecular Biology", from the Japan Atomic Energy Research Institute, by the contract on the Nuclear Safety Research Association. The authors thank Dr. Donald Pierce, Dr. Gerald Sharp, and Dr. John Cologne for providing the example data and the English check. The authors also thank the referees whose comments and suggestions lead to a greatly improved paper.

References

Aalen, O.O. (1994). Effects of frailty in survival analysis. *Statistical Methods in Medical Research* **3**, 227-243.

Aalen, O.O. and Tretli, S. (1999). Analyzing incidence of testis cancer by means of a frailty model. *Cancer Causes & Control* **10**, 285-292.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory.* Akademiai Kiado, 267-281.

Andersen, P.K., Borgan, O., Gill, R.D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes.* Springer, New York.

Andrieu, N. and Goldstein, A.M. (1998). Epidemiologic and genetic approaches in the study of gene-environment interaction: an overview of available methods. *Epidemiologic Review* **20**, 137-147.

Armitage, P. and Doll, R. (1954). The age distribution of cancer and a multi-stage theory of carcinogenesis. *British Journal of Cancer* **8**, 1-12.

Armitage, P. and Doll, R. (1957). A two-stage theory of carcinogenesis in relation to the age distribution of human cancer. *British Journal of Cancer* **11**, 161-169.

Bennett, J., Little, M.P., and Richardson, S. (2004). Flexible dose-response models for Japanese atomic bomb survivor data: Bayesian estimation and prediction of cancer risk. *Radiation and Environmental Biophysics* **43**, 233-245.

Caporaso, N., DeBaun, M.R., and Rothman, N. (1995). Lung cancer and CYP2D6 (the debrisoquine polymorphism): sources of heterogeneity in the proposed association. *Pharmacogenetics* **5**, S129-134.

Doll, R. (1971). The age distribution of cancer. Implication for models of carcinogenesis. *Journal of the Royal Statistical Society* A **134**, 133-155.

Efron, B. and Hinkley, D.V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika* **65**, 457-487.

Feigelson, H.S., Ross, R.K., Yu, M.C., Coetzee, G.A., Reichardt, J.K., and Henderson, B.E. (1996). Genetic susceptibility to cancer from exogenous and endogenous exposures. *Journal of Cellular Biochemistry* **25**, S15-22.

Haiman, C.A., Stram, D.O., Wilkens, L.R., Pike, M.C., Kolonel, L.N., Henderson, B.E., and Marchand, L.L. (2006). Ethenic and racial differences in the smoking-related risk of lung cancer. *The New England Journal of Medicine* **354**, 333-342.

Hayashi, T., Imai, K., Morishita, Y., Hayashi, I., Kusunoki, Y., and Nakachi, K. (2006). Identification of the *NKG2D* haplotypes associated with natural cytotoxic activity of peripheral blood lymphocytes and cancer immunosurveillance. *Cancer Research* **66**, 563-570.

Hougaard, P. (1995). Frailty models for survival data. *Lifetime Data Analysis* **1**, 255-273.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data.* Springer-Verlag, New York.

Izumi, S. and Ohtaki, M. (2004). Aspects of the Armitage–Doll Gamma frailty model for cancer incidence data. *Environmetrics* **15**, 209–218.

Knudson, A.G. Jr. (1971). Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences of the United States of America* **68**, 820-823.

Little, M.P. (1995). Are two mutations sufficient to cause cancer? Some generalizations of the two-mutation model of carcinogenesis of Moolgavkar, Venzon, and Knudson, and of the multistage model of Armitage and Doll. *Biometrics* **51**, 1278-1291.

Little, M.P. (2002). Absence of evidence for differences in the dose-response for cancer and non-cancer endpoints by acute injury status in the Japanese atomic-bomb survivors. *International Journal of Radiation Biology* **78**, 1001-1010.

Little, M.P. and Wright, E.G. (2003). A stochastic carcinogenesis model incorporating genomic instability fitted to colon cancer data. *Mathematical Biosciences* **183**, 111-134.

Luebeck, E.G. and Moolgavkar, S.H. (2002). Multistage carcinogenesis and the incidence of colorectal cancer. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 15095-15100.

Matanoski, G.M., Boice, J.D. Jr, Brown, S.L., Gilbert, E.S., Puskin, J.S., and O'Toole, T. (2001). Radiation exposure and cancer: Case study. *American Journal of Epidemiology* **154**, S91-98.

Moolgavkar, S.H., Cardis, E., Krewski, D., Moller, H., and Woodward, A., eds. (1999). *Quantitative Estimation and Prediction of Human Cancer Risks.* IARC Monograph 131. World Health Organization, Lyon.

Moolgavkar, S.H. and Knudson, A.G. Jr. (1981). Mutation and cancer: a model for human carcinogenesis. *Journal of the National Cancer Institute* **66**, 1037-1052.

Moolgavkar, S.H. and Venzon, D.J. (1979). Two-event models for carcinogenesis: incidence curves for childhood and adult tumors. *Mathematical Biosciences* **47**, 55-77.

Muller, H.J. (1951). Radiation damage to the genetic material. *Sci Prog* **7**, 93.

Neriishi, K., Stram, D.O., Vaeth, M., Minuno, S., and Akiba, S. (1991). The observed relationship between the occurrence of acute radiation effects and leukemia mortality among A-bomb survivors. *Radiation Research* **125**, 206-213.

Nordling, C.O. (1953). A new theory of the cancer inducing mechanism. *British Journal of Cancer* **7**, 68-72.

Nowak, M.A., Komarova, N.L., Sengupta, A., Jallepalli, P.V., Shih, I-M., Vogelstein, B., and Lengauer, C. (2002). The role of chromosomal instability in tumor initiation. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 16226-16231.

Ohtaki, M. (1981). An approximation of the left tail of a distribution for waiting-time in an irreversible point process. *Journal of the Japan Statistical Society* **11**, 111-118.

Ohtaki, M., Fujita, S., Hayakawa, N., Kurihara, M., and Munaka, M. (1985). The age distribution of human adult cancer and an initiation-manifestation model for carcinogenesis. *Japanese Journal of Clinical Oncology* **15**, S325-343.

Ohtaki, M. and Izumi, S. (1999). Globally convergent algorithm without derivatives for maximizing a multivariate function. *Proceedings of Development of Statistical Theories and Their Application for Complex Nonlinear Data.* Hiroshima University, 1-4.

Ohtaki, M. and Niwa, O. (2001). A mathematical model of radiation carcinogenesis with induction of genomic instability and cell death. *Radiation Research* **156**, 672-677.

Pierce, D.A., Mendelsohn, M.L. (1999). A model for radiation-related cancer suggested by atomic bomb survivor data. *Radiation Research* **152**, 642-654.

Pierce, D.A. and Preston, D.L. (2000). Radiation-related cancer risks at low doses among atomic bomb survivors. *Radiation Research* **154**, 178-186.

Pierce, D.A., Sharp, G.B., and Mabuchi, K. (2003). Joint effects of radiation and smoking on lung cancer risk among atomic bomb survivors. *Radiation Research* **159**, 511-520.

Pierce, D.A., Stram, D.O., and Vaeth, M. (1990). Allowing for random errors in radiation dose estimates for the atomic bomb survivor data. *Radiation Research* **123**, 275-284.

Pierce, D.A. and Vaeth, M. (2003). Age-time patterns of cancer to be anticipated from exposure to general mutagens. *Biostatistics* **4**, 231-248.

Roesch, W.C., ed. (1987). *US-Japan Joint Reassessment of Atomic Bomb Radiation Dosimetry in Hiroshima and Nagasaki.* Radiation Effects Research Foundation, Hiroshima.

Sauvaget, C., Lagarde, F., Nagano, J., Soda, M., Koyama, K., and Kodama, K. (2005). Lifestyle factors, radiation and gastric cancer in atomic-bomb survivors (Japan). *Cancer Causes & Control* **16**, 773-780.

Self, S.G. and Liang, K.Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* **82**, 605-610.

Sposto, R., Stram, D.O., and Awa, A.A. (1987). *An investigation of random errors in DS86 dosimetry using data on chromosome aberration and severe epilation.* Radiation Effects Research Foundation (Technical Report 7-90), Hiroshima.

Stewart, A.M. and Kneale, G.W. (1984). Non-cancer effects of exposure to A-bomb radiation. *Journal of Epidemiology and Community Health* **38**, 108-112.

Stewart, A.M. and Kneale, G.W. (2000). A-bomb survivors: factors that may lead to a re-assessment of the radiation hazard. *International Journal of Epidemiology* **29**, 708-714.

Tan, W-Y. (1991). *Stochastic Models of Carcinogenesis.* Marcel Dekker, New York.

Vaupel, J.W., Manton, K.G., and Strallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* **16**, 439-454.

000000 Biometrical Journal Heft 1/2004 Art.: 0000/Autor
C:\izumi\kaken\ohtaki2\project\paper\biomj\resubmit4\izumi_2nd_revision_BIMJ.200512141.doc
insgesamt 13 Seiten Diskette 05.11.2006 Bearb.:

Japanese Journal of Biometrics Vol. 25, No. 2, 61–67 (2004)

Original Article

Genotyping of Single Nucleotide Polymorphisms Based on a Mathematical Model for Two-Dimensional Data

Kenichi Satoh*[1], Keiko Ohtani*[2], Masaru Ushijima*[3], Minoru Isomura*[3], Masaaki Matsuura*[3], Yoshio Miki*[3] and Megu Ohtaki*[1]

*[1]Research Institute for Radiation Biology and Medicine, Hiroshima University, Hiroshima 734-8553, Japan,
*[2]Japan Biological Informatics Consortium, Hiroshima 734-8553, Japan,
*[3]Genome Center, Japanese Foundation for Cancer Research, Tokyo 170-8455, Japan
e-mail: ohtaki@hiroshima-u.ac.jp

Classification methods typically applied to the Invader assay include k-means clustering and the normal mixture model for original two-dimensional data or angle data. Combining the normal mixture model and angle data might result in an inproved method. In fact, such an approach has the advantages that it can be used to evaluate the goodness of classification for each individual and angle data are easily handled. However, the method requires that the data have an origin, which implies that one cluster must be specified before clustering. Therefore, an alternative method using the normal mixture model is desirable. We propose a mathematical model with a latent time variable. Optimization is based mainly on a one-dimensional normal mixture model with two components, which provides stable computational results more quickly than can be obtained using a bivariate normal mixture model.

Key words: SNP typing; Mathematical model; Normal mixture model.

1. Introduction

Single Nucleotide Polymorphisms (SNPs) are one of the most important biological markers for successfully producing tailor-made medicine (Riva and Kohane, 2002). The Invader assay is one method of genotyping SNPs, in which the resultant two-dimensional data must be classified. Several methods have been considered, such as k-means clustering for original two-dimensional data by Renade *et al.* (2001) , Oliver *et al.* (2002) and van den Oord *et al.* (2003), and the normal mixture model for angle data with a fixed origin by Fujisawa *et al.* (2003). Software is available from the pharmaceutical industry (PerkinElmer Inc., SNPscorer) for aseertaining genotype in the FP-TDI assay; a demonstration version is available

Received July 2004. Revised October 2004. Accepted October 2004.