Journal of Carcinogenesis & Mutagenesis

Tan and Zhou J Carcinogene Mutagene 2011, 2:2
DOI: 4172/2157-2518.1000117

Review Article Open Access

A New Stochastic Model of Retinoblastoma Involving both Hereditary and Non-hereditary Cancer Cases

Wai-Yuan Tan[1#*] and Hong Zhou[2#*]

[1]Department of Mathematical Sciences, University of Memphis, Memphis, TN 38152, USA
[2]Department of Mathematics and Statistics, Arkansas State University, State University, AR 72467, USA
[#]Both authors contributed equally to this work

Abstract

Background and purpose: Retinoblastoma is initiated by the mutation or loss or inactivation of the retinoblastoma gene (the Rb gene) in chromosome 13q14. Further, both the germline cells (eggs and sperm) generating the individuals may carry a mutant allele of the Rb gene so that individuals may carry both mutants in the Rb locus at the embryo stage in which case cancer tumor may develop at or before birth. Recent molecular studies have also shown that besides the abrogation of cell dierentiation by the inactivation of the Rb locus, the apoptosis mechanism needs also to be inhibited or abrogated for the generation of retinoblastoma tumor. The purpose of this paper is to develop new stochastic and statistical models for retinoblastoma to incorporate these biological findings.

Results: Based on recent biological studies, in this paper we have developed a discrete-time stochastic multi-stage model and a generalized mixture model for retinoblastoma to account for hereditary cancer cases. We have applied this model to fit and analyze the SEER data of retinoblastoma from NCI/NIH. Our results indicate that a modified MVK (Moolgavkar-Venzon-Knudson) two-stage model with discrete time fits the data much extremely well and better than a three-stage model.

Conclusion: Our studies have shown that retinoblastoma can best be described by a modified MVK two-stage model with discrete time. It appears that this new model would not only provide more insights into retinoblastoma but also would provide useful guidance for its prevention and control and for prediction of future cancer cases.

Keywords: Apoptosis; Generalized mixture model; Multi-stage models of carcinogenesis; Retinoblastoma; Stochastic equations

Introduction

For the human pediatric eye cancer-retinoblastoma, Knudson [1] discovered that the cancer was initiated by the mutation or loss or inactivation of the retinoblastoma gene (Rb gene) in chromosome 13q14. This discovery was further documented by cytogenetic studies by Cavenee et al. [2] (see [3], Chapter 3). For retinoblastoma, Knudson [1] has thus proposed a two-stage model for the development of cancer tumors. According to this model, the person is in the first stage (I_1 stage) if one copy of the Rb gene in an eye stem cell in this person has lost or mutated or inactivated. The person is in the second stage (I_2 stage) if both copies of the Rb gene in an eye stem cell have been mutated or lost or inactivated. According to this model, cancer tumors are derived from primary I_2 cells, where a primary I_2 cell is an I_2 cell generated by an I_1 cells through mutation or deletion or inactivation of the other Rb allele. A specific feature of this model is that both the germ line cells (eggs and sperms) and somatic cells can carry a mutant of the Rb gene leading to inherited cancer cases. The above two stage model for retinoblastoma was taken for granted until the early 1990's when it was discovered that default loss of Rb was not tumorigenesis unless death by apoptosis was also inhibited [4-11]. Based on results of molecular biology, DiCiomme et al. [6] have thus proposed a three-stage biological model for carcinogenesis of retinoblastoma with the first two stages being associated with the Rb gene and with the third stage involving abrogation of apoptosis and /or cell cycle control by some genetic and/or epigenetic changes [12,13]. Let I_3 denote the third stage, where an I_3 cell is an I_2 cell with additional genetic and/or epigenetic changes to abrogate or inhibit apoptosis and/or cell cycle control. Then the 3-stage biological model proposed by DiCommo et al. [6] assumes that retinoblastoma develop by the pathway $N \rightarrow I_1 \rightarrow I_2 \rightarrow I_3 \rightarrow$ Tumor. Despite these biological studies, however, stochastic mathematical models for retinoblastoma based on these biological studies had never been derived, nor had these models been tested against cancer incidence data such as those from SEER from NCI/NIH; for more detail, see Section 4.

Given the above background of retinoblastoma, the objective of this paper is to develop a biologically supported stochastic mathematical model of carcinogenesis for the development of retinoblastoma tumors. Because for retinoblastoma, a large number of babies develop cancer tumors at birth (0) (see data in Table 1), we will also proceed to develop a biologically supported statistical model to incorporate genetic segregation of the mutant allele of the Rb locus in the population.

A biologically supported stochastic model of retinoblastoma incorporating inherited cancer cases

A modified MVK model for retinoblastoma with discrete-time: Consider a discrete-time model with one time unit corresponding to six months or longer (See Section 4 and Remark s 1 and 4). Let N denote normal stem cell and T cancer tumor. Then the MVK model with discrete-time is characterized by the following postulates:

***Corresponding authors:** Wai-Yuan Tan, Department of Mathematical Sciences, The University of Memphis, Memphis, TN 38152, USA, E-mail: waitan@memphis.edu

Hong Zhou, Department of Mathematics and Statistics, Arkansas State University, State University, AR 72467, USA, E-mail: zhou.hong@mathstat.astate.edu

Received April 01, 2011; **Accepted** May 20, 2011; **Published** May 24, 2011

Citation: Tan WY, Zhou H (2011) A New Stochastic Model of Retinoblastoma Involving both Hereditary and Non-hereditary Cancer Cases. J Carcinogene Mutagene 2:117. doi:10.4172/2157-2518.1000117

(a) $N(I_0) \rightarrow I_1 \rightarrow I_2 \rightarrow Tumor$ by genetic changes and/or epigenetic changes.

(b) The I_i cells are subjected to stochastic proliferation (birth) and differentiation (death).

(c) With probability one each I_2 cell generated at time t from an I_1 cell would develop by clonal expansion (i.e., stochastic birth-death process, [14]) into a detectable cancer tumor at time $t + 1$.

(d) All cells develop independently of other cells.

Because this model is not exactly the same as the MVK model proposed by Moolgavkar and Venzon (see [3], Chapter 3; [15]), in what follows we will refer this model as the modified MVK two-stage model. In Section 4 it will be shown that this model fits cancer incidence data (SEER data from NCI/NIH) extremely well and much better than a three-stage model.

The following biological observations also suggest that this discrete-time MVK model is consistent with the observations of the biological model by Diciommo et al. [6]:

(a) Apoptosis is activated only if the number of I_2 stem cells has grown into a cluster of large number of I_2 cells ($10^6 \sim 10^8$ or larger) in which case the probability of some genetic and/or epigenetic changes to induce inhibition or abrogation of apoptosis is greatly enhanced. (The rate of genetic and/or epigenetic changes to induce inhibition of apoptosis is about $10^{-3} \sim 10^{-4}$; see Section (4.3)).

(b) Apoptosis abrogation often occurs in the last stage in the multi-stage model of carcinogenesis [16,17] by that time there are already a large number of I_2 cells.

(c) Cancer tumors are heterogeneous populations of cells with tumor stem cells being a very small minority [18]. These biological findings imply that once an I_2 cell is generated from an I_1 cell, with probability close to one a genetic change or epigenetic change will quickly and almost inevitably develop, so that, as a close approximation , the third mutation can be ignored mathematically. These results are consistent with a discrete-time two-stage model with the assumption that with probability one a primary I_2 at time t would develop into a detectable tumor at time $t + 1$.

Remark 1: To develop stochastic models of carcinogenesis, in the literature [3,19,20] it is conveniently assumed that the last stage cells (i.e. I_k cells in a k-stage model) develop instantaneously into cancer tumors as soon as they are generated. In this case, one may identity I_k cells as cancer tumors so that $T(t)$ is Markov. When k=2, this is the MVK two-stage model (see [3], Chapter 3; [15]). However, as shown by Yang and Chen [14], Yakovlev and Tsodikov [21], Klebanov et al. [22] and Fakir et al. [23] , in many cases T(t) is not Markov. Nevertheless, if one assumes a discrete time model with one time unit to correspond to six months or one year, then because the growth of I_k cells is very rapid, with probability close to one an I_k cell generated at time t will develop into a detectable tumor by time t + 1 [16,17] in these cases, one may practically assume T(t) as Markov.

A stochastic model of retinoblastoma involving inherited cancer cases

As first documented by Knudson[1], both the germ line cells (eggs and sperms) and somatic cells may carry the mutant Rb allele r. If both germ line cells (egg and sperm) generating the individual carry the mutant Rb allele r, then this individual is at the I_2 stage at the embryo stage (fertilized egg stage) so that for this individual cancer tumors develop at and /or before birth; see Remark 2. As shown in Table 1, this is clearly the case since the retinoblastoma incidence from the SEER data of NCI/NIH gives the highest cancer rate at birth; see Remark 3 in Section 3 for SEER data.

To account for inherited cancer cases in the stochastic model of retinoblastoma, let p be the frequency of the r gene in the population so that $q = 1 - p$ is the frequency of the R allele (Normal allele of the Rb locus) in the population. Assume that the population is very large and that mating (marriage) between people is random with respect to the retinoblastoma locus. For each individual, let the embryo stage denote the time of the fertilized egg in his/her mother's womb from which this individual is developed. Then, by the Hardy-Weinberg law [24,25] the frequency of individuals with genotypes RR, Rr and rr at the embryo stage in the population are given by q^2, $2pq$ and p^2 respectively.

To develop stochastic models to incorporate inherited cancer cases, observe that during pregnancy the proliferation rates of all stem cells are very high; hence, as a close approximation one may practically assume that with probability one individuals with genotype rr at the embryo stage would develop detectable cancer tumors at or before birth. Similarly for individuals with genotype Rr at the embryo stage, the R allele in some stem cells may be mutated or lost or inactivated during pregnancy in which case with positive probability these individuals would carry stem cells with genotype rr at or before birth to develop cancer tumor at birth. Because spontaneous mutation of genes in normal individuals is very low ($10^{-6} \sim 10^{-8}$) during pregnancy [17,19,20], one would expect that normal people at the embryo stage would remain to be normal people at birth. Hence, practically one may assume that individuals with genotype RR at the embryo stage would remain to have genotype RR at birth.

Remark 2: Without exception, every human being develops from the embryo in his/her mother's womb, when stem cells of different organs divide and differentiate to develop different organs respectively (See Weinberg [26], Chapter 10). To develop stochastic models of retinoblastoma with genetic component, we thus let time at the embryo stage to be the starting time for carcinogenesis.

The probability distributions for developing detectable cancer tumors

For retinoblastoma, the development of cancer tumors of individuals depend on his/her genotype at the embryo stage in his/her mother's womb; see Figure 1.

Embryo stage: Given that the individual has genotype rr at the embryo stage (an I_2 person), then with probability one this individual would develop cancer tumors before or at birth. On the other hand, if the individual has genotype Rr at the embryo stage, then with probability α, the R allele in some stem cells of this individual may

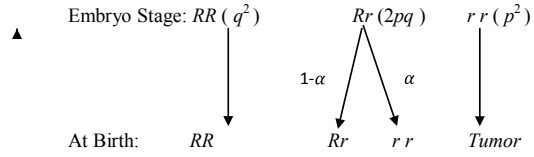

Figure 1: Genotypes of Individuals with Respect to Rb Locus and Their Frequencies in the Population at the Embryo Stage.

be mutated or lost to generate I_2 stem cells in this individual during pregnancy to develop cancer at birth. Thus for each individual with genotype *Rr* at the embryo stage, with probability α this individual would develop detectable cancer at birth, and with probability $1-\alpha$ this individual would remain an I_1 stage individual at birth in which case one more stage is required to develop retinoblastoma after birth ($I_1 \to I_2 \to$ Tumor). If the individual has genotype *RR* (normal people) at the embryo stage, then one may practically assume that with probability one this individual remains a normal people at birth so that after birth, retinoblastoma are developed by a two-stage model $N \to I_1 \to I_2 \to Tumor$. This is described in Figure 1.

Let $Q_{Rr}(j)$ ($Q_{RR}(j)$) denote the probability that an individual with genotype *Rr* (*RR*) at birth develops cancer during the *j*–th time period $(t_j-1, t_j]$ $(j > 1)$ after birth. For individuals with genotype *RR* at the embryo stage, the probability that this individual would develop cancer during the *j*–th time period after birth is $Q_{RR}(j)$. For individuals with genotype *Rr* at the embryo stage, the probability that these individuals would develop cancer during the *j*–th period after birth is $(1-\alpha)Q_{Rr}(j)$.

To derive $Q_{RR}(j)$ and $Q_{Rr}(j)$, let β_1 be the transition rate per cell division of $I_1 \to I_2$ and β_0 the transition rate per cell division of $N \to I_1$. Let $I_1^{(1)}(t) = I_1(t|RR\ att_0)$, $\left(I_1^{(2)}(t) = I_1(t|Rr\ att_0)\right)$ denote the number of I_1 cells at time *t* in individuals who have genotype *RR* (*Rr*) at birth (0). Let $R_i(t) = \sum_u^t t_o u_i(u)\beta_1, i = 1,2,$ where $u_i(s) = EI_1^{(i)}$ (s) is the expected number of $I_1^{(i)}$ (s). Then, by using similar methods given in Tan [16], Tan et al. [17], Tan and Chen [27], Tan et al. [28-30], it can be shown that

$$Q_{RR}(j) = E\{e^{-\xi_1(t_{j-1})} - e^{-\xi_1(t_j)}\} \approx e^{-R_1(t_{j-1})} - e^{-R_1(t_j)} \quad (1)$$

$$Q_{Rr}(j) = E\{e^{-\xi_2(t_{j-1})} - e^{-\xi_2(t_j)}\} \approx e^{-R_2(t_{j-1})} - e^{-R_2(t_j)} \quad (2)$$

where $\xi_i(t) = \sum_u^t t_o u_i(u)\beta_1$.

From equations (1)-(2), observe that approximately the probability of developing cancer during the j-th time period after birth in each individual are functions of $R_i(t_j)$'s which in turn are functions of the expected number of I_1 stem cells in the individual during the j–th time period. Let N_0 be the number of normal stem cells at birth and denote by b_1 the birth rate per cell division (proliferation rate) of I_1 cells and d_1 the death rate per cell division (differentiation rate) of I_1 cells. Using stochastic difference equations as described in Tan et al. [17], Tan and Chen [27], Tan et al. [28,29] and Tan et al. [31], it can easily be shown that the expected number of I_1 cells at time t from normal individuals at birth and from individuals who are I_1 people at birth are given respectively by:

$$u_1(t) = \lambda_1\beta_1^{-1}\left\{\left(1+\gamma_1\right)^t - 1\right\}, and\ u_2(t) = \lambda_2\beta_1^{-1}\left(1+\gamma_1\right)^t,$$

where $\gamma_1 = b_1 - d_1$, $\lambda_1(t) = N_0\beta_0\beta_1/\gamma_1$, and $\lambda_2(t) = u_2(t)\beta_1$. In the above equation, observe that γ_1 is the proliferation rate per cell division of I_1 cells so that the number of I_1 cells will increase if $\gamma_1 > 0$ and decrease if $\gamma_1 < 0$. Using the basic result $\sum_{I=0}^n a^i = (a^{n+1}-1)/(a-1)$ and noting $R_i(t) = \sum_{u=t_0}^t u_i(u)\beta_1, (i = 1,2)$ we obtain

$$R_1(t) = \gamma_1^{-1}\lambda_1\left\{\left(1+\gamma_1\right)^{t-t_0} - 1 - \gamma_1(t-t_0)\right\}, \text{ and}$$

$$R_2(t) = \gamma_1^{-1}\lambda_2\left\{\left(1+\gamma_1\right)^{t-t_0} - 1\right\}.$$

A statistical model and the probability distribution of the number of detectable tumors

The data available for modeling carcinogenesis are usually cancer incidence over different time periods. For example, the SEER data (see Remark 3) of NCI/NIH for retinoblastoma are given by $\{(y_0, n_0), (y_j, n_j), j = 1, \ldots, n\}$, where y_0 is the number of cancer cases at birth and n_0 the total number of birth and where for $j \geq 1$, y_j is the number of cancer cases during the *j*-th age group of a one year period and n_j is the number of normal people who are at risk for cancer and from whom y_j of them have developed cancer during the *j*-th age group. Given in Table 1 are the SEER data for retinoblastoma cases during the period 1973-2007. From this data set, notice that there are a large number of cancer cases at birth implying a large number of inherited cancer cases. In this section, based on the models in Sections (2)-(3) we will develop a statistical model for these data sets.

Note:

(1) The observed Incidence rates per 10^6 individuals from SEER are zero after 10 years old. Hence we fit only data up to 10 years old.

(2) The data in Table 1 are yearly data with 0 denoting birth and j the jth years old, j=1,… 10.

Remark 3: The SEER data are data compiled by the Surveillance and End Results (SEER) Program of the National Cancer Institute/ NIH, a premier source for cancer statistics in the United States. This program collects information on incidence, prevalence and survival from specific geographic areas representing 28 percent of the US population and compile reports on all of these plus cancer mortality for the entire country; for more retail about SEER, the readers are referred to the web of Google Search.

The probability distribution of observed cancer incidence incorporating inherited cancer cases: To incorporate inherited cancer cases, among the n_j people at risk for retinoblastoma, let n_{1j} be the number of individuals who have genotype *RR* at the embryo stage, n_{2j} the number of individuals who have genotype *Rr* at the embryo stage, and $n_{3j} = n_j - n_{1j} - n_{2j}$ the number of individuals who have genotype *rr* at the embryo stage. Then, from results in Section (2.2) and by the Hardy-Weinberg law, the probability that a random individual from the population has genotype RR, Rr and rr at the Rb locus is q^2, 2pq and p^2 respectively. Hence, from basic probability theory [32], the conditional probability distribution of (n_{2j}, n_{3j}) given n_j is multinomial with parameters $\{n_j; 2pq, p^2\}$. That is, $(n_{2j}, n_{3j})|n_j \sim ML\{n_j; 2pq, p^2\}$. It follows [32] that $n_{3j}|n_j \sim Binomial\{n_j, p^2\}$.

The probability distribution of y_0: Because y_0 is the number of cancer cases at birth, y_0 derives either from individuals who have genotype *rr* at the embryo stage or from individuals who have genotype *Rr* at the embryo stage. Because with probability one an individual with genotype *rr* at the embryo stage would develop cancer at or before birth, all n_{30} individuals with genotype *rr* at the embryo stage would develop retinoblastoma at birth; on the other hand, with probability α $(0 < \alpha < 1)$, each individual with genotype *Rr* at the embryo stage would develop cancer at birth. Hence, from basic probability theory [32], $y_0 = n_{30} + z_0$, where $z_0|n_{20} \sim Binomial\{n_{20}, \alpha\}$. Furthermore, as shown in Section (3.1), $(n_{20}, n_{30})|n_0 \sim Multinomial\{n_0; 2pq, p^2\}$. Hence, we have, because n_0 is very large and *p* is very small:

$$y_0|n_0 \sim Binomial\{n_0, p^2 + 2pq\alpha\} \sim Poisson\{\chi_0\} \quad (3)$$

where $\chi_0 = n_0\left(p^2 + 2pq\alpha\right)$.

Years Old	Population at Risk	Observed Incidence Rate per 10^6 Individuals	Observed Cancer Cases	Predicted Cancer Cases 2-Stage Model	Predicted Cancer Cases 3-Stage Model
0	11,687,938	2.584	302	300	294
1	11,437,360	1.478	169	164	76
2	11,347,783	1.084	123	113	55
3	11,344,946	0.582	66	67	38
4	11,388,462	0.220	25	32	29
5	11,403,795	0.114	13	16	28
6	11,365,139	0.079	9	9	30
7	11,430,904	0.044	5	4	37
8	11,215,416	0.027	3	3	48
9	11,650,697	0.026	3	2	58
10	11,773,403	0.025	3	1	68

Table 1: Retinoblastoma Incidence SEER Data (1973-2007) from NCI/NIH.

The probability distribution of y_j ($j \geq 1$) : To derive the probability distribution of y_j ($j \geq 1$) in the j–th age group after birth, observe that y_j can only be developed from individuals who have genotype *Rr* or *RR* at the embryo stage.

Among the y_j cancer cases, let y_{1j} be the number of cancer cases generated by people with genotype *RR* at the embryo stage, and y_{2j} the number of cancer cases generated by people with genotype *Rr* at the embryo stage. Then $y_j = y_{1j} + y_{2j}$; y_{1j} are developed from n_{1j} people only and y_{2j} are developed from n_{2j} people only.

Because cancer develops in each individual independently of other individuals and because individuals with the same genotype are expected to yield the same phenotype , to derive the conditional probability distribution of (y_{1j} , $y_{2j} = y_j - y_{1j}$) given (n_{ij} , i = 1, 2, n_j), one may practically assume that each individual develops cancer independently of other people and that all individuals with the same genotype at the embryo stage develop cancer by the same mechanism. Then, because the probabilities $\{P_1(j) = Q_{RR}(j), P_2(j) =(1 - \alpha)Q_{Rr}(j)\}$ are very small, the conditional probability distribution of $\{ y_{1j} , y_{2j} = y_j - y_{1j} \}$ given $\{n_{1j} , n_{2j} , n_j \}$ is

$$P\left\{y_{1j}, y_{2j} | n_{ij}, i=1,2,n_j\right\} = h\{y_{1j}; n_{1j}P_1(j)\}h\{y_{2j}; n_{2j}P_2(j)\} \quad (4)$$

where h (y; λ) is the density at y of the Poisson distribution with mean λ.

Put $Q_T(j) = n_{1j} P_1(j) + n_{2j} P_2(j)$. From Equation (4), the conditional distribution of y_j given $\{n_{ij} , i = 1, 2, n_j \}$ is Poisson with mean $Q_T(j)$. It follows that the probability distribution of y_j given n_j is

$$P\left(y_j | n_j\right) = \sum_{n_{1j}=0}^{n_j} \sum_{n_{2j}=0}^{n_j - n_{1j}} g\left(n_{1j}, n_{2j}; n_j, q^2, 2pq\right) h\left\{y_j; Q_T(j)\right\}, \quad (5)$$

The probability distribution P (y_j $|n_j$) given by equation (5) is a mixture of Poisson distributions with mixing probability distribution given by the multinomial distribution of $\{n_{1j}, n_{2j}\}$ given n_j. This mixing probability distribution represents individuals with different genotypes at the embryo stage in the population.

Let Θ be the set of all unknown parameters (i.e. the parameters (p, α) and the birth rates, the death rates and the mutation rates of N and I_1 cells). Based on data ($y_j, j = 0,1,\ldots,k$), the likelihood function of Θ is $L\left\{\Theta | y_j, j = 0,1,\ldots,k\right\} = h(y_0; \chi_0)\prod_{j=1}^{k} P(y_j | n_j)$. From Section (2.3), this likelihood function depends on the unknown parameters via the estimable parameters (p, α) and the estimable parametric functions $\{\gamma_1$, λ_i, i=1,2$\}$.

The fitting of the model and applications

To illustrate the application of the models in Sections 2-3, in this section we apply the model to some of the NCI/NIH SEER data of retinoblastoma to derive some important information about retinoblastoma. Because the biological findings in references [4]~[13] suggest a three stage model, we will also fit a three stage model which assumes that retinoblastoma develop by the pathway $N \rightarrow I_1 \rightarrow I_2 \rightarrow I_3 \rightarrow$ Tumor. For fitting this model to the SEER data, we give in the Appendix the probability distributions of cancer incidence data and some basic formula for implementing the fitting of SEER. Because this model is a special case of the general k-stage model described and analyzed by Tan [16] and Tan et al. [17], for more detail and mathematical theories, we refer the readers to Tan [16] and Tan et al. [17].

Methods for fitting data

Because it is well documented that the Bayesian inference procedures are the most efficient procedures [33,34], we will use the Bayesian approach via the data augmentation and Gibbs sampling procedures to estimate the unknown parameters and to derive predicted cancer cases. For the model of this paper, the basic approach are summarized in five steps:

(a) Expand the model and data (n_j , y_j , j = 0, 1, . . . , m) to include the un-observable variables $\{n_{20}, n_{30}, n_{1j}$, n_{2j} , y_{1j} , y_{2j} , j = 1, . . . , $m\}$ and derive the joint probability distribution of all random variables. This probability distribution is given in Equation (6) below.

(b) Derive the conditional probability distribution of $\{ y_{1j}$, y_{2j} , j = 1, . . . , m $\}$ given $\{n_{ij}$, i = 1, 2, n_j , y_j , j = 1, . . . , $m\}$. This probability distribution is given in Equation (4).

(c) Derive the conditional probability distribution of $\{ n_{1j}$, n_{2j} , j = 1, . . . , $m\}$ given $\{n_j$, y_j , j = 1, . . . , $m)$. Combining this probability distribution with the conditional probability distribution given in equation (5), one may apply the Weighted bootstrap method to generate observed values of $\{ n_{1j}$, n_{2j} , j = 1, . . . , m $\}$ given $\{ n_j$, y_j , j = 1, . . . , m $\}$. This is illustrated in detail in Tan et al. [17], Tan [24] (Chapter 3) and in Tan et al. [28].

(d) Derive the general conditional posterior distribution of all parameters given $\{y_j$, y_{1j} , n_{ij} , i =1, 2, j = 1, . . . , $m\}$. Notice that this posterior distribution is proportional to the product of the prior distribution $P(\Theta)$ and the joint distribution of all the variables given by Equation (6). (The prior distribution of the parameters is given in the next Section (4.2).)

(e) Apply the Gibbs sampling procedure to estimate the unknown parameters and the state variables. The details of these steps and the Gibbs sampling procedure and its applications to

cancer modeling have been illustrated in great detail in [17,24,28].

Put $Y=\left(y_{ij}\ \tilde{\imath}\tilde{=}\quad j=\ \ldots\ k\right)\qquad N=\left(n_{ij},i=1,2;j=1,\ldots,k\right)$, $y=(y_j,j=1,\ldots,k)$, and $n=(y_j,j=1,\ldots,k)$. For the SEER data, the joint density P{**Y**, **y**, **N**| **n**, Θ} of {**Y**, **y**, **N**} given { **n**, Θ } is:

$$P\{Y,y,N\,|\,n,\Theta\}=h[y_0;n_0\left(p^2+2pq\left(1-\alpha\right)\right)]$$

$$\prod_{j=1}^{k}\{g(n_{1j},n_{2j};n_j,q^2,2pq)\}\prod_{i=1}^{2}h\{y_{ij};P_i\left(j\right)\}\qquad(6)$$

Notice that the above distribution is a product of multinomial distributions and Poisson distributions. The above joint density will be used as the kernel for estimating the unknown parameters and for predicting the state variables.

Fitting of the model by cancer incidence data

To fit the SEER retinoblastoma cancer data, we let one time unit correspond to 6 months after birth and let the time at birth be 0; see Remark 4. For the prior distributions of Θ, because biological information have suggested some lower bounds and upper bounds for the mutation rates and for the proliferation rates, we assume $P(\Theta)$ α c, (c > 0) where *c* is a positive constant if these parameters satisfy some biologically specified constraints, and equal to zero for otherwise. These biological constraints are:

(1) $0 < \alpha < 1$ as α denotes a probability measure, and because the estimated frequency of most mutant genes in human population are from $10^{-3} \sim 10^{-5}$ [25], we let $0 < p < 10^{-2}$;

(2) Because $\gamma_1 = b - d$ is the proliferation rate per cell division of cells with genotype Rr at the Rb locus and since Rb is a tumor suppressor gene, we let $-0.01 < \gamma_1 < 1$; see also estimates from Tan [16], Tan et al. [17], Tan and Chen [27], Tan et al.([28-30]) and Luebeck and Moolgavkar [35].

(3) Because the estimates of the transition rate from I_j to I_{j+1} per cell division in almost all multistage models in human beings are $10^{-4} \sim 10^{-8}$ [16,17,27-30,35], we let $10^{-8} < \beta_i < 10^{-3}$, i = 0, 1 ;

(4) Because the estimate of the total number of normal stem cells in most tissues [36] is around 10^8, we let $10^6 < N_0$, $u_2(t_0) < 10^9$; hence we let $10^{-1} < \gamma_1 < 10^4$ and $10 < \gamma_2 < 10^6$.

We will refer the above prior as a partially informative prior which may be considered as an extension of the traditional non- informative prior given in Box and Tiao [37].

Remark 4: In our model and the fitting of data we have assumed 6 month for one time unit since for most of human cancers, the last stage cells grow very fast and 6 months is a long time interval for the last stage cell to develop into a detectable tumor [3,16,17,27-30]. Luebeck and Moolgavkar [35] had used one year as the time unit for fitting human colon cancer. In our computation, we have tried 3 months, 6 months and one year as time unit and found that the 6 month period is the best time for fitting retinoblastoma.

Using this prior distribution and applying the method in Section (4.1) to the models in Sections (2)-(3) and the SEER data in Table 1, we have estimated the unknown parameters and the predicted cancer cases. Given in Table 2 are the estimated parameter values and its standard errors. Given in Figure 2 are the plots of predicted cancer cases from the modified MVK two-stage model of Section (2) and from a three-stage model respectively. (For a multi-stage model of carcinogenesis and its fitting to the data [16,17]. For comparison purposes, we have provided numbers of predicted cancer cases from the modified MVK two-stage model and from a three-stage model together with the observed cancer cases over time from SEER. We have also examined SEER data from 1973-2005 and SEER data from 1973-2006 and found that the observed cancer incidence per 10^6 are identical to those from 1973-2007. This indicates that the epidemic of retinoblastoma cancer has reached a steady state condition in US population.

Note: $\gamma_1 = b_1 - d_1, \lambda_1 = N_0\beta_0\beta_1/\gamma_1,\ \lambda_2 = u_2(t_0)\beta_1$

From results in Table 2, notice that the standard errors for the estimates of λ_1 and λ_2 are quite large. These results imply wider confidence intervals and higher uncertainty in the estimation of these two parametric functions. These results are expected because λ_1 is a function of 4 parameters each of which are subjected to uncertainty in estimation and because there is considerable uncertainty in the estimation of $u_2(t_0)$.

Fitting results

From results in Tables 1 and 2 and from Figure 2, we have made the following observations:

(a) As shown by results in Table 1 and Figure 2, it appeared that the modified MVK two-stage model fitted the SEER data extremely well; on the other hand, the three-stage model can not fit the SEER data; see Table 1 and Figure 2. The AIC (Akaike Information Criterium) and the BIC (Bayesian Information Criterium) for the modified MVK two-stage model and for the three-stage model are given by (AIC=1588.7, BIC=1586.3) and (AIC=2423.8, BIC=2431.3) respectively. For the modified MVK two-stage model, the Chi-square test statistic is $\chi^2=\sum_{j=0}^{10}\frac{(y_j-\hat{y}_j)^2}{\hat{y}_j}=7.909$ with degrees of freedom 11, giving a *p-value* $=\mathrm{P}\left\{\chi_{11}^2\geq 7.909\right\}=0.7214$.

(b) From Table 2, it is observed that the estimate of γ_1 is quite small (the estimate is of order $10^{-3} \sim 10^{-4}$) indicating that the phenotype of *Rr* is quite close to that of *RR*; this is consistent with the biological hypothesis that the Rb gene is a tumor suppressor gene.

(c) From Table 2, the estimates $\hat{\lambda}_j(j=1,2)$ of λ_j are {5.2168*10², 9.9277*10⁴}. Notice that the estimate of $\lambda_1=\frac{N_0\beta_0\beta_1}{\gamma_1}$, is of order $\{10^2 \sim 10^3\}$ and the estimate of $\lambda_2=u_2(t_0)\beta_1$ is order of $\{10^4 \sim 10^5\}$, respectively. If we follow Potten et al. [36] to

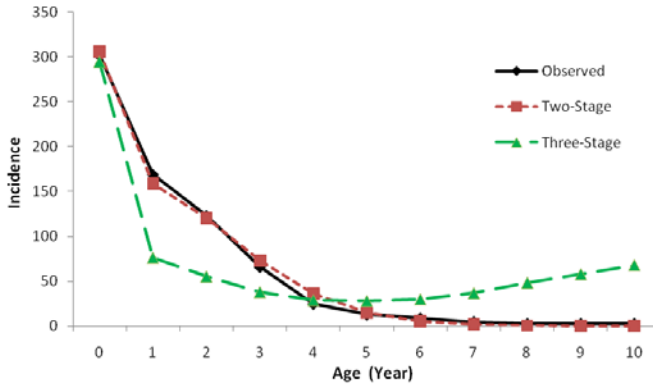

Figure 2: The Observed and Predicted Cancer Cases.

Parameters	p	α	λ_1	λ_2	γ_1
Estimates	5.6114E-03	8.3676E-01	5.2168E+02	9.9277E+04	2.4951E-04
St.D	2.9257E-04	7.3018E-02	1.3067E+02	5.6250E+03	1.2788E-04
95%CI-Lower	5.0380E-03	6.9364E-01	2.6556E+02	8.8252E+04	0
95%CI-Upper	6.1848E-03	9.7987E-01	7.7780E+02	1.1030E+05	5.0016E-04

Table 2: Estimates of Parameters under a MVK Two-Stage Model of Carcinogenesis.

assume $N_0=N(t_0) = u_2(t_0) \sim 10^8$, then β_0 is of order $\{10^{-7} \sim 10^{-6}\}$ and β_1 is of order $\{10^{-4} \sim 10^{-3}\}$.

(d) From Table 2, the estimate of p from the SEER data is of order 10^{-3} indicating that in the US population, the frequency of the recessive allele r of the Rb gene is approximately of order 10^{-3}. Table 2 also showed that the estimate of α was 0.83, indicating that most individuals with genotype Rr would develop cancer at birth. This is consistent with the observation that there are high observed cancer incidences at birth for retinoblastoma in the SEER data.

Discussion and Conclusions

In this paper, by assuming discrete time we have proposed a modified MVK two- stage model to model cancer progression and tumor development of retinoblastoma. We found that the biological findings (a)-(d) given in Section (2.1) imply that this model is consistent with the observation and results of molecular biology studies by DiCiommo et al. [6], Laurie et al. [7], Gallie et al. [12] and Corson and Gallie [13]. To account for inherited cancer cases in the stochastic model of retinoblastoma, we have also developed a generalized mixture model for retinoblastoma in human beings. In this mixture model, the mixing probability is a multinomial distribution to account for the distribution of the three genotypes (rr, Rr, RR) of the Rb locus in chromosome 13q14 among individuals in the population. This mixture model allows us to estimate for the first time the frequency p of the mutant allele of the Rb gene in the US population.

To illustrate the usefulness and applications of our models and methods, we have applied our models and methods to the retinoblastoma SEER data of NCI/NIH. Our analysis clearly showed that the proposed modified MVK two-stage model fitted the data almost perfectly (see Table 2 and Figure 2); on the other hand, the stochastic three-stage model fitted the data poorly (see Table 2 and Figure 2) even though in the three stage model there are four more parameters (see Appendix). Notice, however, our modified MVK two-stage model is more general than the classical MVK two-stage model as described in Tan [3] in that we postulate that cancer tumors develop from primary I_2 cells by clonal expansion [14]. (The stochastic multi-stage models in the literature assume that cancer tumors develop from last stage cells immediately as soon as they are generated, ignoring completely cancer progression [23].

Applying our models and methods to the SEER data of retinoblastoma, we have derived for the first time some useful information on the epidemic of retinoblastoma in the population. Specifically, we mention:

(1) For the first time, we have estimated the frequency of the mutant allele of the Rb gene in the US population ($\hat{p} \sim 5.81 \times 10^{-3}$).

(2) The estimate of the proliferation rate (γ_1) of I_1 cells (i.e. cells with genotype Rr) is $\hat{\gamma}_1 = 2.4954 \times 10^{-4} \sim 0$. This is consistent with the biological hypotheses that the Rb gene is a tumor suppressor gene, and unlike the p53 gene in chromosome 17p [38], there is little or no haploid-insufficiency for the Rb gene in cells with genotype Rr.

Using models and methods of this paper, one can easily predict future cancer cases for retinoblastoma in the population. Thus, by comparing results from different populations, our models and methods can be used to assess cancer prevention and control procedures. This will be our future research topics; we will not go any further here.

Acknowledgements

The research of this paper by W.Y. Tan is supported by a grant from NCI/NIH, grant number R15 CA113347-01. The research by H. Zhou is partially supported by Arkansas State University faculty research fund from July 1, 2010 to June 30, 2011.

References

1. Knudson AG (1971) Mutation and cancer: Statistical study of retinoblastoma. Proc Natl Acad Sci USA 68: 820-823.
2. Cavenee WK, Hansen MF, Nordenskjold M, Kock E, Maumenee I, et al. (1985) Genetic origin of mutations predisposing to retinoblastoma. Science. 228: 501-503.
3. Tan WY (1991) Stochastic Models of Carcinogenesis. Marcel Dekker, New York.
4. Clarke AR, Maandag ER, van Roon M, van der Lugt NM, van der Valk M, et al. (1992) Requirement for a functional Rb-1 gene in murine development. Nature 359: 328-330.
5. Jacks T, Fazeli A, Schmir EM, Bronson RT, Goodell MA, et al. (1992) Effects of an Rb mutation in the mouse. Nature 359: 295-300.
6. Di Ciommo D, Gallie BL, Bremner R (2000) Retinoblastoma: the disease, gene and protein provide critical leads to understand cancer. Semin Cancer Biol 10: 255-269.
7. Laurie NA, Donovan SL, Shih CS, Zhang J, Mills N, et al. (2006) Inactivation of the p53 pathway in retinoblastoma. Nature 444: 61-66.
8. Robanus-Maandag E, Dekker M, van der Valk M, Carrozza ML, Jeanny JC, et al. (1998) p107 is a suppressor of retinoblastoma development in pRb-deficient mice. Genes Dev 12: 1599-1609.
9. Chen D, Livne-bar I, Vanderluit JL, Slack RS, Agochiya M, et al. (2004) Cell-specific effects of RB or RB/p107 loss on retinal development implicate an intrincically death-resistant cell-of-origin in retinoblastoma. Cancer Cell 5: 539-551.
10. MacPherson D, Sage J, Kim T, Ho D, McLaughlin ME, et al. (2004) Cell type-specific effects of RB deletion on the murine retina. Genes Dev18: 1681-1694.
11. Zhang J, Schweers B, Dyer MA (2004) The first knockout mouse model of retinoblastoma. Cell Cycle 3: 952-959.
12. Gallie BI, Campbell C, Devlin H, Duckett A, Squire JA (1999) Developmental basis of retinal-specific induction of cancer by RB mutation. Cancer Res 59: 1731-1735.
13. Corson TW, Gallie BL (2007) One hit, two hit, three hit, more? Genomic changes in the development of retinoblastoma. Genes, Chromosomes Cancer 46: 617-634.
14. Yang GL, Chen CW (1991) A stochastic two-stage carcinogenesis model: a new approach to computing the probability of observing tumor in animal bioassay. Math. Biosci 104: 247-258.

15. Moolgavkar SH, Venzon DJ (1979) Two event model for carcinogenesis: Incidence curves for childhood and adult tumors. Math. Biosciences 47: 55-77.

16. Tan WY (2010) Stochastic multi-stage models of carcinogenesis as Hidden Markov models: A new approach. Int J Systems and Synthetic Biology 1: 313-337.

17. Tan WY, Chen CW, Zhang LJ (2008) Cancer Biology, Cancer models and Stochastic Mathematical Analysis of Carcinogenesis. In: " Handbook of Cancer Models and Applications." eds. Tan,W.Y. and Hanin, L. Chapter 3, World Scientific, River Edge, NJ.

18. Al-Hajj M, Becker MW, Wicha M, Weissman I, Clarke MF (2004) Therapeutic implications of cancer stem cells. Current Opinion in Genetics and Development 14: 43-47.

19. Little MP, (2008) Cancer models, ionization and genomic instability: A review. In: "Handbook of Cancer Models with Applications." (eds. Tan W.Y. and Hanin L.) World Scientific, River Edge, NJ. Chapter 5.

20. Zheng Q (2008) Stochastic multistage cancer models: A fresh look at an old approach. In: " Handbook of Cancer Models and Applications." (eds. Tan, W.Y. and Hanin, L.) World Scientific, River Edge, NJ. Chapter 2.

21. Yakovlev AY, Tsodikov AD (1996) Stochastic Models of Tumor Latency and Their Biostatistical Applications. World Scientific, Singapore and River Edge, New Jersey.

22. Klebanov LB, Rachev ST, Yakovlev AY (1993) A stochastic model of radiation carcinogenesis: Latent time distributions and their properties. Math. Biosciences 113: 51-75.

23. Fakir H, Tan WY, Hlatky L, Hahnfeldt P, Sachs RK (2009) Stochastic population dynamic effects for lung cancer progression. Radiation Research 172: 383-393.

24. Tan WY (2002) Stochastic Models With Applications to Genetics, Cancers, AIDS and Other Biomedical Systems. World Scientific, River Edge, New Jersey.

25. Crow JF, Kimura M (1970) An Introduction to Population Genetics Theory. Harper and Row, New York.

26. Weinberg RA The Biology of Cancer. GS Garland Science, Taylor and Francis Group, New York, USA.

27. Tan WY, Chen CW (2005) Cancer stochastic models. In:" Encyclopedia of Statistical Sciences, Revised edition". John Wiley and Sons, New York.

28. Tan WY, Zhang LJ, Chen CW (2004) Stochastic modeling of carcinogenesis: State space models and estimation of parameters. Discrete and Continuous Dynamical Systems. Series B 4: 297-322.

29. Tan WY, Chen CW, Zhang LJ (2008) Cancer risk Assessment by State Space Models. In: " Handbook of Cancer Models and Applications. " eds. Tan, W.Y. and Hanin, L. Chapter 12, World Scientific, River Edge, NJ.

30. Tan WY, Zhang LJ, Chen W, Zhu JM (2008) A stochastic model of human colon cancer involving multiple pathways. In: "Handbook of Cancer Models with Applications." eds. Tan W.Y. and Hanin L. Chapter 11, World Scientific, River Edge, NJ.

31. Tan WY, Ke WM, Webb G (2009) A stochastic and state space model for tumor growth and applications. Math. Comp. Methods in Medicine 10: 1-21.

32. Hogg RV, Tanis EA (2010) Probability and Statistical Inference. Eight edition. Prentice Hall, NJ.

33. Carlin BP, Louis TA (2008) Bayesian methods for data analysis. Third edition, Chapman and CRC, Boca Raton, FL.

34. Gelman A, Carlin J, Stern H, Rubin D (2004) Bayesian methods for data analysis. Second edition, Chapman and CRC, Boca Raton FL.

35. Luebeck EG, Moolgavkar SH (2002) Multistage carcinogenesis and the incidence of colorectal cancer. Proc. Natl Acad Sci USA 99: 15095-15100.

36. Potten CS, Booth C, Hargreaves D (2002) The small intestine as a model for evaluating adult tissue stem cell drug targets. Cell Prolif 36: 115-129.

37. Box GEP, Tiao GC (1973) Bayesian Inference in Statistical Analysis. Addison-Wesley, Reading, MA.

38. Lynch CJ, Milner J (2006) Loss of one p53 allele results in four-fold reduction in p53 mRNA and protein: A basis for p53 haplo-insufficiency. Oncogene 25: 3463-3470.