Hiroshima J. Med. Sci.
Vol. 58, No. 1, 25~35, March, 2009
HIJM 58–4

A Cure Weibull Gamma-Frailty Survival Model and Its Application to Exploring the Prognosis Factors of Neuroblastoma

Mohammad DOKHI[1)], Megu OHTAKI[2)] and Eiso HIYAMA[3)]

1) Graduate School of Biomedical Science, Hiroshima University, 1-2-3 Kasumi, Minami-ku, Hiroshima 734-8551, Japan
2) Research Institute for Radiation Biology and Medicine, Hiroshima University, 1-2-3 Kasumi, Minami-ku, Hiroshima 734-8551, Japan
3) Natural Science Centre for Basic Research and Development, Hiroshima University, 1-2-3 Kasumi, Minami-ku, Hiroshima 734-8551, Japan

ABSTRACT

The log rank test and the Cox regression, or modifications thereof, emphasize the effect of covariates on survival rate parameter. In some cases, cured individuals, i.e., individuals who may not experience the event of interest may exist in the population of interest. In this situation, we may wish to examine the effect of covariates on both survival rate and cured fraction parameters. Motivated by the Japanese neuroblastoma dataset, we consider a cure model that accounted for the effect of covariates on both of the abovementioned parameters. To deal with heterogeneity that is not explained by covariates, as well as individual random heterogeneity, we perform a frailty variable. Moreover, some nested models are fitted to deal with the principle of parsimony. The effect of covariates was then evaluated by the best nested model. From a statistical point of view, we found that the model of analysis is flexible and adequate to describe the abovementioned dataset. From a medical point of view, we confirmed AGE and STAGE to be the most dominant prognosis factor of neuroblastoma. We also conclude that *NMYC* and *FERRITIN* are the other most important prognosis factors. The analysis designated that some of the prognosis factors of neuroblastoma probably just affected the median life of patients and some others are the fatal prognosis factor indicated by their effect which significance on both of survival rate and cured fraction parameters. The present model of analysis is also potentially extendable to facilitate other aspects of inferences.

Key words: *Neuroblastoma, Weibull-distribution, Gamma-distribution, Maximum-likelihood, Information-matrix*

The cure model is a method of modeling time-to-death when cured individuals, i.e., individuals who may not experience the event of interest, exist in the population of interest. In this situation, we need to fit the model with an explicit parameter of the cured fraction to be estimated. Ignoring the cured component in the model may lead to a distorted result. A common approach is to formulate the population of interest as a mixture of cured and susceptible individuals. The cure model was introduced by Boag[4)], who proposed a method by which to estimate the proportion of patients who were cured of breast cancer. This initial research was followed by other papers[3,18)]. Numerous subsequent analyses have been conducted using the concept of the cured component. More general concepts and examples of the application of the cure model in various aspects of human life can be found in[16)]. In recent decades, the development of modern cancer treatments has meant that the proportion of cured cancer patients has increased dramatically, and the cure model will likely become more attractive, especially in medical statistics.

The standard cure model assumes that individuals in the population of interest are homo-

Corresponding author: Megu OHTAKI,
Research Institute for Radiation Biology and Medicine, Hiroshima University, 1-2-3 Kasumi, Minami-ku, Hiroshima 734-8551, Japan
E-mail: ohtaki@hiroshima-u.ac.jp

geneous with respect to risk. However, in reality, individuals are inherently dissimilar. The heterogeneity may be modeled in part by covariates, but there will always be an unexplained residual[1]. Such a residual may be induced by unobserved covariates, as well as an immeasurable risk factor. Failing to account for such heterogeneity may lead to an imprecise conclusion[6]. The model that accounts for such heterogeneity is known as the frailty model, introduced at the first time by Vaupel et al[26]. The basic concept of frailty is that individuals have different frailties, and individuals who are frailer will die earlier than the others. Further discussion on the extended frailty model can be found in[10,15], for example.

Kuk and Chen[13] proposed a method by which to integrate the mixture component of a population into a nonparametric proportional hazard model without a frailty component. Maller and Zou[16] and Sposto[21] modeled the parameters of the cure model as a function of covariates, but their model did not account for the frailty effect either. Price and Manatunga[20] discussed the modeling of Weibull survival cure models using some frailty distributions, but they applied the models without covariates. In order to investigate the Japanese neuroblastoma dataset and its constraints, we consider a model that accounts for the effect of covariates on both the cured fraction and the survival rate. Moreover, in order to deal with unobserved heterogeneity, as well as individual random heterogeneity, we apply a frailty variable. The general set-up of this analysis assumes that the covariates of interest affect the survival function through the links function of the model parameters, whereas the effect of the other covariates composite individual random heterogeneity explained by the frailty variable. Such a model can be considered as an extension of the frailty model to allow the effect of covariates on both cured fraction and survival rate parameters.

The remainder of the present paper is organized as follows. Section 2 reviews the motivating dataset and its constraints. In Section 3, we discuss the proposed model, a method by which to deal with heterogeneity, modeling and fitting nested models, as well as a practical interpretation. We then evaluate the model by applying it to an actual dataset in Section 4. In the final section, we present the results of this evaluation and some concluding remarks.

MOTIVATING DATA

The database used in this study was approved by the "Research Group for Evaluation of the Japanese Neuroblastoma Mass-Screening Project". This group is represented by graduate schools of medicine and related research centers of several universities in Japan, including the Natural Science Center for Basic Research and Development of Hiroshima University as the correspondence address. The consortium gathered data for patients with neuroblastoma from the cancer registries of the Japanese Society of Pediatric Surgery and the Japanese Society of Pediatric Oncology, two major Japanese organizations that deal with neuroblastoma patients. Both of these organizations enrolled neuroblastoma cases directly from hospitals at which the patients had been treated. Moreover, the consortium referenced the database of the Japanese Infantile Neuroblastoma Cooperative Study Group to confirm the status of clinical pathologic data of infants with neuroblastoma. The new single database was then assembled with a careful check of duplicate data.

For the registration years of 1985 to 2000, they obtained data for 3,720 children diagnosed as neuroblastoma patients. The dataset contains the core variables of the patients, including a unique sequential ID, the year of registration, the follow-up time, and the outcome of diagnosis. The dataset also contains some of the most important clinical and histological prognosis factors, as well as the treatment information for each patient. In study of neuroblastoma it is known that age at diagnosis (referred to as *AGE*) and stage of diseases (referred to as *STAGE*) are the most important prognosis factors[24]. Here, STAGE was classified according to the criteria of the Japanese Society of Pediatric Surgery. Table 1a shows the distribution of the remaining observations after validating core variables, *AGE* and *STAGE*.

The major constraint of the dataset is a lot of missing data on the different covariates and the different patients. Consequently, the number of observable samples decreases dramatically when we increase the number of covariates in the considered model. This contradiction is a problem because, theoretically, we need a greater number of samples when the model is more complicated. For convenience, we refer to this problem as the contradiction problem. Table 1b shows a code sheet for the variables considered in this paper and the number of remaining observations after validation on *AGE*, *STAGE* and an additional covariate (*AC*).

MODEL AND ITS IMPLEMENTATION

1. Cured Fraction Indication and Standard Cure Model

Due to the heterogeneity of individuals, the survival rate of a disease typically decreases with time. Patients with a high risk die earlier than others, and after some time, the low-risk individuals become dominant in the society. If the following time is sufficient, the remaining individuals will not die of disease and may be consid-

ered *cured individuals*. In the context of survival time analyses, the cured individuals are subject to right censoring. In our dataset, there are three kinds of right censoring, i.e. alive-free-tumor, alive-remaining-tumor, and death for other reason. The first group consists of individuals whom we assume are cured. If the follow-up time is long enough (more than five years from diagnosis) then we also assume the second group to be cured individuals. As we know, the Kaplan-Meier survival curve does not step down when the patient is censored. Consequently, even though we cannot identify which individuals will be cured in our dataset, their presence is signaled by a leveling-off of the Kaplan-Meier survival curve at any positive value.

If cured individuals appear to exist in the dataset, then it is reasonable to assume that the population of interest consists of two sub populations, a population that is susceptible to the disease and a cured population. Suppose that an individual is either cured with probability 1-ϕ or has a proper survival function $S(t)$ with probability ϕ. If T denotes the random time to the event of interest, then the cure survival function of T can be specified as

$$S_c(t) = 1-\phi+\phi S(t), \quad t \geq 0, 0 < \phi \leq 1. \tag{1}$$

In parametric terminology, we assume any theoretical function of $F(t)$ for modeling the probability distribution of random variable T and then calculate $S(t)$=1-$F(t)$. In application, usually we have initial information about the shape of the hazard function, and it is preferable to directly model the survival function by the hazard function rather than by the distribution of T. If F is assumed to be continuous with respect to the density function $f(t)=dF(t)/dt$, then the hazard function is defined as $h(t)=f(t)/[1-F(t)]$. Based on the hazard function, the survival function can be expressed as

$$S(t) = e^{-\int_0^t h(u)du} = e^{-H(t)}, \quad t \geq 0, \tag{2}$$

where $h(\cdot)$ and $H(\cdot)$ are the baseline and cumulative hazard function of $S(t)$, respectively. The cure survival function given in (1) deals with the simplest case, i.e., we assume that individuals in the population of interest are homogeneous with respect to risk. In the following two sections, we discuss how to deal with the heterogeneity of individuals in a population of interest.

2. Cure Weibull Gamma-Frailty Model

As mentioned in Section 1, we assume that unobserved heterogeneity composite individual random heterogeneity is taken into account by using the frailty term. The frailty variable Z is assumed to be a random variable varying across the population with $E(Z)$=1 and variance var $(Z)=v$. This variable has as a multiplicative effect on the hazard function. The survival function (2) conditional on frailty $Z=z$ can then be expressed as

$$S_{F|Z}(t \mid z) = e^{-\int_0^t zh(u)du} = e^{-zH(t)}, \quad t \geq 0, z \geq 0. \tag{3}$$

The unconditional survival function is characterized by the Laplace transform of the Z distribution, as follows:

$$S_F(t) = E_Z[S(t \mid Z)] = E_Z[e^{-ZH(t)}] = L_Z(H(t)), \; t \geq 0, \tag{4}$$

i.e., the Laplace transform Lz of the cumulative hazard function $H(t)$. By substituting (4) into (1), we obtain the cure frailty model, as follows:

$$S_{CF}(t) = 1-\phi+\phi L_Z(H(t)), \quad t \geq 0, 0 < \phi \leq 1. \tag{5}$$

If we choose the distribution of the frailty variable Z that has an explicit form of the Laplace transform, then we can use the ordinary maximum likelihood method for parameter estimation.

In the present paper, we consider the Weibull distribution as a baseline for the hazard function of T and gamma as the distribution of Z. The model is then referred to as the *Cured Weibull Gamma-Frailty* (CWGF) model. The Weibull formula is easy to write down and often provides a good fit in survival time distributions, which are observed in practice[16]. Furthermore, the Laplace transform of the gamma distribution is relatively simple and computable. Then, it follows from (5) that the CWGF model can be expressed as

$$S_{CWGF}(t) = 1-\phi+\phi\left(1+\frac{(\lambda t)^{\alpha}}{\theta}\right)^{-\theta}, \tag{6}$$
$$t \geq 0, \lambda \geq 0, \theta \geq 0, \alpha \geq 0, 0 < \phi \leq 1,$$

where $v=1/\theta$ is the variance of the gamma distribution, and component $(\lambda t)^{\alpha}$ is the cumulative hazard function of the Weibull distribution. The primary interests are ϕ and λ parameters. Here, ϕ characterizes the susceptible proportion in the population of interest, which is consequently related to both fatal and non-fatal cases of the patients. On the other hand, λ is the hazard rate parameter related to the survival length of fatal cases, so that λ is concerned only with fatal cases.

3. The use of Covariates

If our dataset has no additional information on covariates related to T, then model (6) is an alternative solution to deal with the heterogeneity of individuals. In many cases, the dataset includes information of covariates. If the covariate is categorical data, the common practice is to distinguish between its categories (levels). The other important task is to control the effects of a particular covariate using the other covariates.

The most commonly used method to deal with covariates in survival analyses is Cox regression. Cox regression accounts for the effect of covariates by multiplying the baseline hazard $h(t)$ by a

positive function of covariates. Consequently, in a non-cure model, the hazard of any individual i is independent of t. Due to the cured fraction ϕ on (6), λ is not constant with respect to t. Therefore, it is rather unnatural to apply the proportional hazard in the cure model. Moreover, the Cox model also calls attention to the effect of covariates on survival rate only, rather than their effect on the cured fraction, which in some cases may be the primary goal of analysis. As an alternative method by which to deal with these restrictions, we use a CWGF model, which assumes that both the survival rate (represented by λ) and the susceptible probability (represented by ϕ) of any individual i are related to covariates through the following link functions:

$$\lambda_i = e^{\boldsymbol{\beta}'\mathbf{x}_i} \quad \text{and} \quad \phi_i = \frac{e^{\boldsymbol{\delta}'\mathbf{x}_i}}{1+e^{\boldsymbol{\delta}'\mathbf{x}_i}}, \tag{7}$$

where β and δ are vectors of parameters that are to be estimated and $\mathbf{x}_i$ is the covariate vector of individual i.

4. Parameter Estimation

Let N be the number of observations, and t_i is the value of the independent random variable T that represents the observed lifetime, possibly censored, of individual i, i=1,2,3,...,N. If c_i is a censoring indicator defined as

$$c_i = \begin{cases} 1, & \text{if } i \text{ is a death-event case} \\ 0, & \text{if } i \text{ is a censored case,} \end{cases}$$

then based on (6) and (7), the likelihood function is defined as

$$L(\boldsymbol{\eta}\,|\,t_i) = \prod_{i=1}^{N} \left(f_{CWGF}(\boldsymbol{\eta}\,|\,t_i)\right)^{c_i} \left(S_{CWGF}(\boldsymbol{\eta}\,|\,t_i)\right)^{1-c_i}. \tag{8}$$

Here, $f_{CWGF}(\boldsymbol{\eta}\,|\,t_i) = -\frac{\partial}{\partial t_i} S_{CWGF}(\boldsymbol{\eta}\,|\,t_i)$ is the probability density function of individual i, and $\boldsymbol{\eta}$ is the vector (β, δ, α, θ). For convenience in the computation, we re-parameterize γ =ln(α) and Ψ=ln(θ). The original parameter α and θ can be obtained by one-to-one inverse transformation $\alpha=e^{\gamma}$ and $\theta=e^{\Psi}$. By this transformation, $\boldsymbol{\eta}$ (β, δ, γ ,Ψ) and we allow the impact on all of the parameters in $\boldsymbol{\eta}$ to vary unrestrictedly in$(-\infty,+\infty)$. This condition is more reasonable for the maximization of the likelihood function (8). In order to simplify the calculation, we maximized the logarithm of (8) rather than the original function.

The most common method by which to find the maximum likelihood estimate (MLE) of parameters from any function as (8) or its logarithm is the Newton-Raphson procedure. However, the implementation of this procedure requires the inverse of the observed information matrix to be calculated upon each iteration. Due to the cured component, Sposto[22] noted the difficulty caused by the possibility of a singular information matrix. Moreover, Price and Manatunga[20] noted the difficulty of likelihood function optimization due to the possibility of multiple or boundary maxima. To deal with such difficulties, we use the spider optimization proposed by Ohtaki and Izumi[19]. This algorithm is suitable for maximizing the multivariate function without its derivative and without the inverse of the information matrix of each iteration. Using this algorithm, we then plug the MLE into the second derivative matrix of the logarithm (8) and take its inverse to determine the asymptotic standard error of the parameter estimates.

5. Interpretation

As mentioned in the previous section, the motivation for applying categorical covariates is group comparison among its categories. In the proposed CWGF model, practically speaking, it is important to distinguish differences among the levels of covariates both on λ and ϕ. To simplify the discussion, we consider only one covariate with two levels (two-category case), but the method should work for the general case. Referring to (7), we assume a dummy vector $\mathbf{x}_i$ associated with individual i to be two-dimensional, as follows:

$$\mathbf{x}_i = \begin{cases} (1\ 0), & (i \in \text{category 1}) \\ (1\ 1), & (i \in \text{category 2}) \end{cases} \tag{9}$$

The extra component 1 in (9) is needed to allow for the intercept term. In logistic regression terminology, this coding method is called *reference cell coding*[11]. In general, this coding method specifies that all of the elements of the dummy vector are equal to zero for the reference category (except for the element of the intercept), while setting a single element equal to 1 for each dummy vector of the other category. Corresponding to vector (9), we define $\beta = (\beta_0\ \beta_1)'$ and $\delta = (\delta_0\ \delta_1)'$, such that the survival rates of category 1 and category 2 are $\lambda_1 = e^{\beta_0}$ and $\lambda_2 = e^{\beta_0+\beta_1}$, respectively. The relative risk (RR) between the two groups then becomes

$$RR = \frac{\lambda_2}{\lambda_1} \equiv e^{\beta_1}. \tag{10}$$

Similarly, based on the odds of group 1 and group 2, i.e., $\phi_1/(1-\phi_1)$ and $\phi_2/(1-\phi_2)$, we can define the odds ratio (OR) as

$$OR = \frac{\phi_2/(1-\phi_2)}{\phi_1/(1-\phi_1)} \equiv e^{\delta_1}. \tag{11}$$

Let $\hat{\beta}_1$ be the MLE of β_1, and let $\hat{\delta}_1$ be the MLE of δ_1, the estimates of RR and OR can then be expressed as

$$\widehat{RR} = e^{\hat{\beta}_1} \quad \text{and} \quad \widehat{OR} = e^{\hat{\delta}_1} \tag{12}$$

and a $100\times(1-\breve{\alpha})$ estimate of the confidence intervals (CI) of RR and OR can be given by expression $e^{\hat{\beta}_1 \pm z_{1-\breve{\alpha}/2} \times \widehat{SE}(\hat{\beta}_1)}$ and $e^{\hat{\delta}_1 \pm z_{1-\breve{\alpha}/2} \times \widehat{SE}(\hat{\delta}_1)}$, respectively. Here, z is the value of the random variable $Z\sim N(0.1)$ corresponding to the significance level $\breve{\alpha}$.

APPLICATION TO A REAL DATASET

1. Analysis of Age at Diagnosis and Stages of Disease as the Most Dominant Covariates

In our neuroblastoma dataset, all of the covariates are categorical data, except for the age at diagnosis (referred to as *AGE*). *AGE* is the most important factor in assessing the prognosis of neuroblastoma. The younger the patient, the more favorable the prognosis and the better the chance of survival. In order to guarantee that the follow-up time is sufficient, we use all observations where the longest follow-up time is 19.8 years. *AGE* is also potentially confounded by most of the other covariates. Moreover, its effect on the survival function is typically non-linear. In order to address this difficulty, we categorized *AGE* into two levels, with 1 year as the cut-off point. This cut-off point is the most common in the study of neuroblastoma, although the use of 1.5 years has also been suggested. The other important prognosis factor is disease stage. Here, we use the Japan staging system (referred to hereinafter as *STAGE*). Originally, this covariate consists of six stages/categories, namely, stages I, II, III, IVA, IVB, and IVS. We excluded stages I and IVS because the number of observations or events were insufficient to perform analyses (see Table 1a). We then consider only four stages, namely, stages II, III, IVA, and IVB. Before performing *AGE* and *STAGE* in subsequent analyses, we check their extraordinary effect separately and jointly.

Figure 1 shows the Kaplan-Meier survival curves of *AGE* and *STAGE* by their categories. From this figure, both *AGE* (left panel) and *STAGE* (right panel) are extremely different in the cured proportion of their categories, as indicated by differences in their leveling-off values. Individuals in *AGE* category 2 (> 1 year) failed far more quickly than those in category 1 (≤ 1 year) and level off at a value lower than 0.40 after 9

Table 1a. Number of observations by *AGE*, *STAGE* and outcome status

STAGE	*AGE* ≤ 1 year		*AGE* > 1 year	
	Censored	Death-event	Censored	Death-event
I	842	12	227	4
II	692	6	86	14
III	368	2	78	56
IVA	94	50	116	278
IVB	64	10	23	44
IVS	194	14	26	8

Table 1b. Code sheet and the number of remaining observations after validation on *AGE*, *STAGE****) and an additional covariate(AC)

Additional Covariate	Abbreviation	Category code*)	N	NE**)
NMYC Amplification	*NMYC*	Non amplified=0, Amplified=1	1234	221
Vanillye Mandelic Acid	*VMA*	Normal=0, High=1	1891	437
Serum Ferritin	*FERRITIN*	Normal=0, High=1	1068	262
Homovanilic Acid	*HVA*	Normal=0, High=1	1857	428
Lactate dehydrogenase	*LDH*	Normal=0, High=1	1421	364
DNA Hyperploidy	*DNA PLOIDY*	Normal=0, High=1	409	91
Neuro Specific Enalose	*NSE*	Normal=0, High=1	1375	347
Phatologic classification	*PATHOL. CLASS*	Ganglio-NB=0, NB=1	1803	411
Size of tumor	*TUMOR SIZE*	≤10 cm=0, >10 cm=1	1131	203
Liver metastasis	*LIVER META*	H_0=0, H_1+H_2+H_3=1	1890	429
Bone metastasis	*BONE META*	B_0=0, B_1=1	1919	450
Bone-marrow metastasis	*BM META*	BM_0=0, BM_1=1	1868	430
Orib metastasis	*ORIB META*	E_0=0, E_1=1	1896	436
Lymph node metastasis	*LN META*	LN_0=0, LN_1=1, LN_2=2, LN_3=3	1806	393
Primary tumor infiltration	*PTI*	PTI_0=0, PTI_1=1, PTI_2=2, PTI_3=3	1848	423
Sex	*SEX*	Female=0, Male=1	1927	452

*) Code=0 is the reference category
**) Number of death-events
***) Not including *STAGE-I* and *IVS*

years because there were no events after that. While the survival time of individuals in *AGE* category 1 declined slowly and leveled off at a value greater than 0.90 after 9.3 years. From the right-hand panel of Fig. 1, visual investigation of *STAGE* reveals a similar trend. Even though *STAGE- IVA* and *IVB* initially decline following a similar pattern, after 2 years, *STAGE-IVA* declines faster and levels off at a value significantly lower than that for *STAGE-IVB*.

Figure 1 shows the results of univariate analyses. Fitting a series of univariate models rarely provides an adequate analysis because of the possibility of a confounding effect among covariates. Next, we evaluate the controlled effect of *AGE* and *STAGE* on each other by fitting them univariately and bivariately. The analyses reveal that both *AGE* and *STAGE* significantly affect the cured fraction ϕ, for both the univariate and bivariate analyses. Table 2a presents the results of analyses on ϕ. Table 2a shows that *AGE* and *STAGE* are the vital prognosis factors of neuroblastoma, where significantly high values of odds ratio are estimated both in univariate or bivariate analyses. The effects of *AGE* and *STAGE* also potentially confounded each other, as indicated by their $\widehat{OR}$ in bivariate analyses, which are significantly lower compared to those in univariate analyses. Without control by *STAGE*, the death risk in *AGE* category 2 is 5.57 times higher compared to *AGE* category 1, but it is reduced to 3.72 after being adjusted for *STAGE* in bivariate analyses. Those odds ratios indicate that the proportion of fatal cases in younger patients is relatively low compare to that in older patients. The other distinction is the $\widehat{OR}$ of *STAGE-IVA*, which is reduced from 7.09 to 5.39 after control by *AGE*. The estimates of logistic function coefficients $\hat{\delta}$ presented in Table 2a have a high level of significance ($p<0.001$). Since $\widehat{OR}$ is a function of $\hat{\delta}$, such significances also impact the significances of $\widehat{OR}$ represented by its confidence interval.

Table 2b shows the effects of *AGE* and *STAGE* on survival rate parameter λ. As we can see from this table, the estimate of hazard ratio/relative risk ($\widehat{RR}$) of group *AGE* >1 to group *AGE* $\leq$ 1 is lower than 1 on both univariate and bivariate analysis. It is indicated that even though the number of fatal cases in younger patients is extremely less, their survival lengths are shorter compared to the fatal cases in older patients. This suggests that the progression of diseases for fatal cases is quicker in younger patients compared to older patients. This table also shows that the effects of *STAGE* on λ are insignificant ($p > 0.05$), which is probably due to the small number of events on its reference category (*STAGE*-II). This restriction impact on estimates of its standard error became large. Furthermore, the $\widehat{RR}$ of *STAGE-IVA* as the most fatal group is lower than 1. It may be attributed to the data distribution: around 73% (394 of 538) cases of this stage are patients with *AGE* > 1 years, while reversely for the reference group (*STAGE*-II) around 88% (698 of 788) cases are patients with *AGE* $\leq$ 1 year (see Table 1a for details of data distribution).

2. *Nested Models and Their Selection Criteria*

Due to the contradiction problem mentioned in Section 2, we decided to evaluate the other covariates one by one, each controlled by *AGE* and *STAGE*. This means that we use *AGE* and *STAGE*

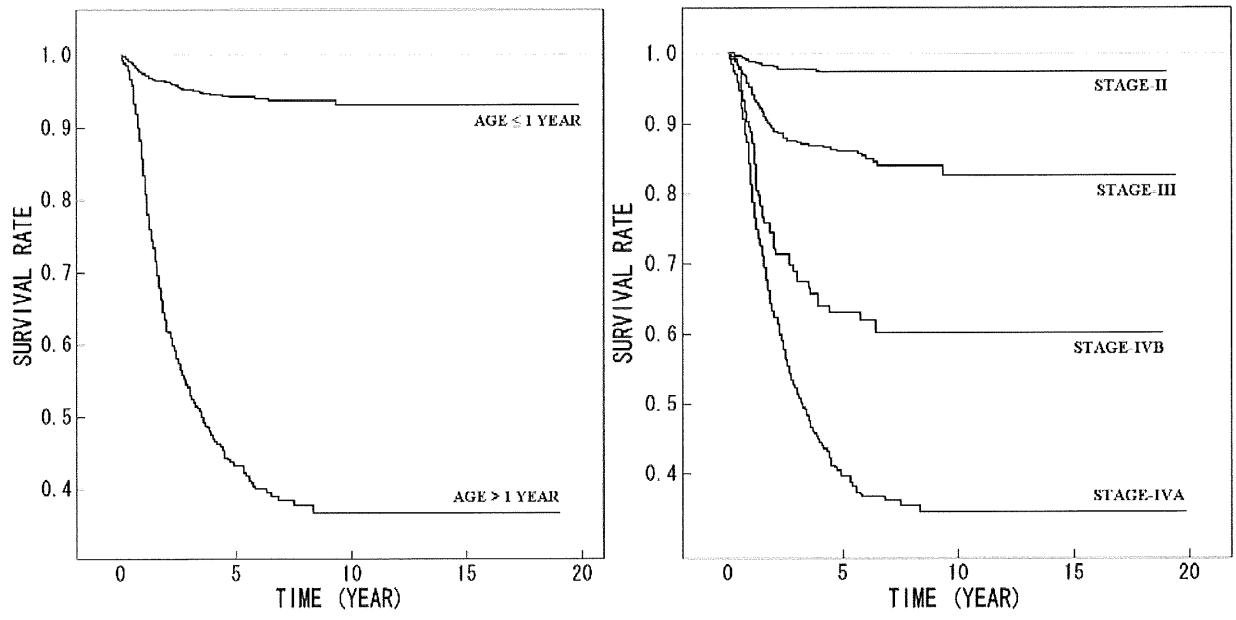

Fig. 1. Kaplan-Meier survival curves for *AGE* (left panel) and for *STAGE* (right panel) by their categories.

as the basis covariates when assessing the effect of the other covariate of interest. For convenience, we hereinafter refer to the covariate of interest (other than *AGE* and *STAGE*) as the *additional covariate* (*AC*). Using this nomenclature, the explicit forms of (7) for individual i, i=1, 2,...,N become

$$\lambda_i = e^{\beta_0+\beta_1\times AGE_i+\beta_2\times STAGE_i+\beta_3\times AC_i}$$
$$\text{and } \phi_i = \frac{e^{\delta_0+\delta_1\times AGE_i+\delta_2\times STAGE_i+\delta_3\times AC_i}}{1+e^{\delta_0+\delta_1\times AGE_i+\delta_2\times STAGE_i+\delta_3\times AC_i}}. \quad (13)$$

For example, when we consider the effect of additional covariate *NMYC*, we replace AC_i in (13) by $NMYC_i$.

The CWGF model (6) with link functions (13) is the most complicated model when performing *AGE* and *STAGE* as basis covariates with a single additional covariate. In order to deal with the principle of parsimony, we need to investigate the simplest possible model, while adequately describing our dataset, and, according to the purpose of the analyses, the model must be suitable for inference. To deal with this theoretical framework, we fit a number of nested models, as listed in Table 3. Functions $f(\cdot)$ and $g(\cdot)$ in Table 3 refer to the link functions defined in (13). Table 3 shows that model M_0 is the simplest. However, according to the purposes of our analyses, this model is not suitable for inference. We need to fit this model as the reference for evaluating the other possible models.

To select the most adequate model among all nested models, as shown in Table 3, we use the Bayesian information criterion (*BIC*) proposed by Kass and Wasserman[12]. For an exposed model parameterized by vector $\boldsymbol{\eta}_j$ with dimension p_j, they defined

$$BIC(M_j) = -2\{\ell_j(\hat{\boldsymbol{\eta}}_j) - \ell_0(\hat{\boldsymbol{\eta}}_0)\} + (p_j - p_0)\times\ln(N), \quad (14)$$

where $\ell_j(\hat{\boldsymbol{\eta}}_j)$ and $\ell_0(\hat{\boldsymbol{\eta}}_0)$ are the maximized log likelihoods under model M_J and a reference model M_0. Here, N is the observable sample size under link function (13). Without the penalty term

Table 2a. Effects of *AGE* and *STAGE* on cured fraction parameter

Covariate	Category	Univariate Analysis				Bivariate Analysis			
		$\hat{\delta}$	p	$\widehat{OR}$	95%CI	$\hat{\delta}$	p	$\widehat{OR}$	95%CI
AGE	≤1 year*)	0.000	-	1.00	-	0.000	-	1.00	-
	> 1 year	1.718	<0.001	5.57	(4.64,6.70)	1.315	<0.001	3.72	(3.05,4.55)
STAGE	II*)	0.000	-	1.00	-	0.000	-	1.00	-
	III	0.849	<0.001	2.34	(1.85,2.96)	0.736	<0.001	2.09	(1.61,2.70)
	IVA	1.958	<0.001	7.09	(5.60,8.96)	1.684	<0.001	5.39	(4.05,7.17)
	IVB	0.817	<0.001	2.26	(1.94,2.64)	0.695	<0.001	2.00	(1.67,2.40)

Note : *) = Reference category

Table 2b. Effects of *AGE* and *STAGE* on survival rate parameter

Covariate	Category	Univariate Analysis				Bivariate Analysis			
		$\hat{\beta}$	p	$\widehat{RR}$	95%CI	$\hat{\beta}$	p	$\widehat{RR}$	95%CI
AGE	≤1 year*)	0.000	-	1.00	-	0.000	-	1.00	-
	> 1 year	-0.1771	0.0124	0.84	(0.73,0.96)	-0.1859	0.0076	0.83	(0.72,0.95)
STAGE	II*)	0.000	-	1.00	-	0.000	-	1.00	-
	III	0.0818	0.1678	1.09	(0.97,1.22)	0.0715	0.2283	1.07	(0.96,1.21)
	IVA	-0.0790	0.0856	0.92	(0.84,1.01)	-0.0872	0.0693	0.92	(0.83,1.01)
	IVB	0.0374	0.3311	1.03	(0.96,1.12)	0.0291	0.4546	1.03	(0.95,1.11)

Note : *) = Reference category

$(p_j - p_0) \times \ln(N)$, the above *BIC* criterion reduces to the standard likelihood ratio test statistic. By criterion (14), the best model is the model with the smallest *BIC*.

3. *Exploring Prognosis Factors of Neuroblastoma*

Since, model M_9 has the largest degrees of freedom (*dof*), theoretically, this model is a candidate for the best fit to empirical data. However, this model may be too complicated for inference. On the other hand, even if model M_0 fits our dataset well, it is unsuitable for inference because it has no covariate. A compromise is to search for a model between models M_0 and M_9 by evaluating their *BIC* values. To illustrate this procedure, we consider Pathologic Classification (*PC*) and NMYC as examples of additional covariates. Their *BIC* values are presented in Fig. 2 (left panel). From the left panel of Fig. 2, model M_3 of *PC* appears comparable to model M_6. Due to a smaller number of parameters, in this case, we may choose model M_3, rather than model M_6, for inference. In the case of *NMYC*, the lowest *BIC* is model M_3, which is relatively low compared to the most complicated model M_9. To clarify the goodness of fit visually, the right-hand panel shows a comparison of models M_3 and M_0 and M_9 of *NMYC*. Due to the frailty effect, model M_0 provides a good fit with respect to survival rate (up to the last event at 9.3 years), but provides a poor fit with respect to cured fraction (after 9.3 years). Model M_9 provides a poor fit with respect to both survival rate and cured fraction. This is probably attributable

Table 3. Some nested models of additional covariates (AC) using *AGE* and *STAGE* as the basis covariates

Model	Function of λ	Function of ϕ
M_0	$\lambda = g(-)$	$\phi = f(-)$
M_1	$\lambda = g\,(AC)$	$\phi = f\,(AC)$
M_2	$\lambda = g\,(AC)$	$\phi = f\,(AGE,AC)$
M_3	$\lambda = g\,(AC)$	$\phi = f\,(AGE,STAGE,AC)$
M_4	$\lambda = g\,(AGE,AC)$	$\phi = f\,(AC)$
M_5	$\lambda = g\,(AGE,AC)$	$\phi = f\,(AGE,AC)$
M_6	$\lambda = g\,(AGE,AC)$	$\phi = f\,(AGE,STAGE,AC)$
M_7	$\lambda = g\,(AGE,STAGE,AC)$	$\phi = f\,(AC)$
M_8	$\lambda = g\,(AGE,STAGE,AC)$	$\phi = f\,(AGE,AC)$
M_9	$\lambda = g\,(AGE,STAGE,AC)$	$\phi = f\,(AGE,STAGE,AC)$

Note : $g(\cdot)$ and $f(\cdot)$ refer to the link functions of (13)

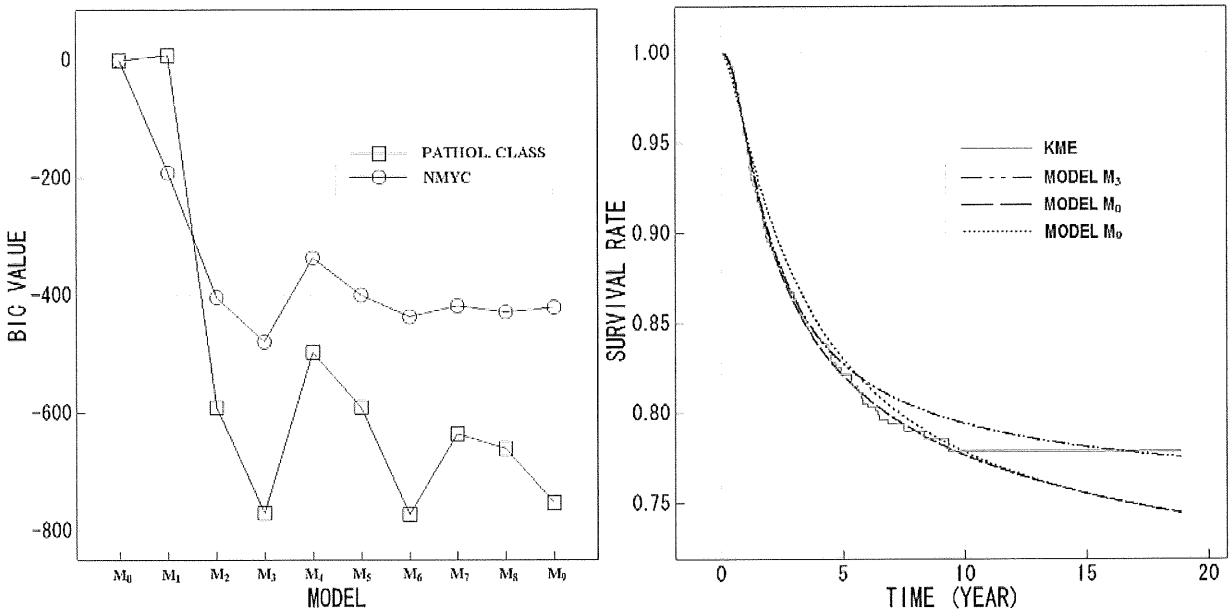

Fig. 2. *BIC* values of the nested model from *NMYC* and *PATHOL. CLASS* (left panel) and fitted models M_0, M_3, and M_9 of *NMYC* (right panel).

to the use of *STAGE* in the link function of λ. As mentioned in the previous subsection, *STAGE* was not significantly affecting the λ parameter. The use of *STAGE* in the link function of λ has an effect on a noise for the survival function. Then, the compromise model is Model M_3, which appears to adequately fit both survival rate and cured fraction. When the number of events is sufficiently high (up to 5 years after diagnosis), model M_3 fits the empirical data represented by the Kaplan-Meier survival curve very well. When the number of events decreases (after 5 years), the curve starts to flat and approaches the leveling-off value of the Kaplan-Meier survival curve. The comprehensive results of parameter estimates of each additional covariate based on the best nested model are shown in Table 4. With the level of $\breve{\alpha}$ =0.05, we found that a number of additional covariates statistically significantly affect both the cured fraction and the survival rate. This is not a contradiction to the result on table 2b because their inferences are based on their own best nested model. In the case of *NMYC*, for example, its inference is based on model M_3 which only used *AGE* as a basis covariate on λ. Some additional covariates have a significant effect on either the cured fraction or the survival rate, whereas some others have not. In general, we found that the best nested model for all of the additional covariates in Table 4 is model M_3 or model M_4. As we can see from Table 3, neither of these model use *STAGE* as a basis covariate on their λ. This is consistent with the result in Table 2b which showed that the effect of *STAGE* on λ was insignificant.

4. Software

For the purpose of computation, we developed a computer program written in FORTRAN. This program begins by fitting the model with the fewest parameters, model M_0, and then moves, one by one, to the more complicated model. For the purpose of visualization, we also developed a program to calculate the value of the Kaplan-Meier estimate as well as the fitted of all nested models in Table 3 and their *BIC* values.

DISCUSSION

We have investigated a cure survival model using the Weibull hazard function. To deal with the individual random heterogeneity that we believe to exist in our dataset, we applied a frailty variable having a gamma distribution. In order to evaluate the effect of covariates on both the cured fraction ϕ and the survival rate λ, we used the logistic link function for ϕ and the log linear link function for λ. To deal with the principle of parsimony, we then fitted some nested models based on different combinations of covariates used in ϕ and λ.

We have extended the model of Price and Manatunga[20] by entering the covariates. The proposed model differs from the model of Kuk and Chen[13] in that their model is a non-parametric model without a frailty variable. The proposed

Table 4. Effects of additional covariates (*AC*) controlled by *AGE* and *STAGE* (estimated by the best nested model defined in Table 3).

Additional Covariate	Category *)	Effect on survival rate				Effect on cured fraction			
		$\hat{\beta}$	*p*	$\widehat{RR}$	95%*CI*	$\hat{\delta}$	*p*	$\widehat{OR}$	95%*CI*
NMYC	Amplified	0.192	0.00	1.21	(1.11,1.33)	0.892	0.00	2.44	(1.65,3.60)
VMA	High	0.120	0.00	1.13	(1.04,1.22)	0.316	0.00	1.37	(1.14,1.65)
FERRITIN	High	0.009	0.90	1.01	(0.88,1.16)	0.714	0.00	2.04	(1.49,2.79)
HVA	High	0.018	0.68	1.02	(0.94,1.11)	0.248	0.01	1.28	(1.06,1.55)
LDH	High	0.207	0.02	1.23	(1.03,1.47)	0.164	0.30	1.18	(0.86,1.61)
DNA PLOIDY	>2 ploidy	0.027	0.83	1.03	(0.80,1.31)	0.277	0.28	1.32	(0.80,2.18)
NSE	High	0.068	0.48	1.07	(0.89,1.29)	0.170	0.26	1.19	(0.88,1.59)
PTL CLASS	NB	0.125	0.02	1.13	(1.02,1.25)	0.171	0.13	1.19	(0.95,1.48)
TUMOR SIZE	> 10cm	0.063	0.31	1.06	(0.94,1.20)	0.281	0.03	1.32	(1.03,1.71)
LIVER META	H1+H2+H3	0.099	0.01	1.10	(1.03,1.19)	0.156	0.07	1.17	(0.99,1.38)
BONE META	B1	-0.083	0.04	0.92	(0.85,1.00)	0.382	0.01	1.46	(1.11,1.94)
BM META	BM1	-0.014	0.74	0.99	(0.90,1.07)	0.242	0.07	1.27	(0.99,1.65)
ORIB META	E1	0.010	0.72	1.01	(0.95,1.07)	0.224	0.06	1.25	(0.99,1.58)
LN META	LN1	0.078	0.37	1.08	(0.91,1.28)	0.070	0.66	1.07	(0.78,1.47)
	LN2	0.084	0.22	1.09	(0.95,1.24)	0.229	0.12	1.26	(0.94,1.68)
	LN3	0.124	0.02	1.13	(1.02,1.25)	-0.159	0.19	0.85	(0.67,1.08)
PTI	PTI1	0.078	0.70	1.08	(0.73,1.61)	0.114	0.74	1.12	(0.58,2.17)
	PTI2	0.079	0.71	1.08	(0.71,1.64)	0.244	0.49	1.28	(0.64,2.54)
	PTI3	0.237	0.28	1.27	(0.83,1.94)	0.196	0.59	1.22	(0.60,2.47)
SEX	Male	0.047	0.36	1.05	(0.95,1.16)	0.105	0.31	1.11	(0.91,1.36)

*) Compare to the reference category defined on table 1b

model also differs from Maller and Zou[16] and Sposto[21] with respect to the frailty term. They did not achieve the frailty variable in their models. Furthermore, the present analysis differs from previous analyses in other aspects. Most previous studies on the survival cure model emphasized the effect of covariates on only the cured fraction component. The present study was an attempt to evaluate the effect of covariates on both the cured fraction and the survival rate. In addition, the concept of fitting the nested models based on different combinations of covariates used in ϕ and λ is a fresh concept in the study of the survival cure model with covariates. Moreover, the novel idea of performing a serial analysis of additional covariates controlled by the basis covariates was inspired by the contradiction problem of the dataset considered herein.

From a statistical point of view, we conclude that the best nested model provided an adequate fit to the dataset considered herein. We found that the use of the Weibull function as a baseline hazard function affects model flexibility. As any $\hat{\delta}_j \to +\infty$, $\hat{\phi} \to 1$ and the CWGF model reduces to a frailty model (without the cured fraction). When $\hat{\psi} \to +\infty$, the estimate of gamma variance $\hat{\nu} \to 0$, and the CWGF model reduces to a standard cure model (without frailty). Furthermore, if any $\hat{\delta}_j \to +\infty$ and $\hat{\psi} \to +\infty$, then the CWGF model reduces to an ordinary Weibull model. The model may also be modified in a straightforward manner, for example, by the use of some other candidate for the link function of ϕ or λ. Other theoretical functions that are suitable for modeling the frailty variable are also available. In order to explore the effect of a covariate in greater detail, the link functions of ϕ and λ may also be extended to allow interaction among covariates.

From a medical point of view, we confirmed *AGE* and *STAGE* to be the most dominant prognosis factors of neuroblastoma. This fact is indicated by their huge effect on the cured fraction with very high reliability. Due to the enormity of their effect on ϕ, we suggest using *AGE* and *STAGE* as the basis covariates in assessing the effect of other neuroblastoma prognosis factors. With $\widehat{OR}$=2.24 for *NMYC* and 2.04 for *FERRITIN*, these two covariates are the other most important prognosis factors, after *AGE* and *STAGE*. *NMYC* is also significant with respect to its hazard ratio with $\widehat{RR}$ =1.21. The comprehensive results in Table 4 also show the irregular effect of the metastasis factors. Clinically speaking, the effect of metastasis should significantly affect the survival function, but we found that most of the metastasis covariates in Table 4 were not significant with respect to either ϕ or λ. The logical explanation of this irregular result is that their effects might be absorbed by *STAGE*. Since *STAGE* is not a generic covariate, it is specified by the combination of other covariates, with metastasis as the basis factor.

Finally, in the present study, we investigated the prognosis factor. In the future, we intend to investigate the effect of treatment of neuroblastoma. Due to the complexity of the effect of treatment, we may need to verify the interaction component in the model. It may also be necessary. to verify the model by other link functions of ϕ and λ Moreover, we will verify further nested models in order to find a more adequate model for the purpose of inference.

ACKNOWLEDGMENTS

This study was partially supported by Grants-in-Aid for Scientific Research (B) of the JSPS (Grant Number 18300095) and by the 3rd-term Comparative Ten-year Strategy for Cancer Control. We thank the Committees on Tumor Registrations in the Japanese Society of Pediatric Surgeons and Japanese Society of Pediatric Oncology for the use of their neuroblastoma database.

(Received October 21, 2008)
(Accepted December 19, 2008)

REFERENCES

1. **Aalen, O.O.** 1988. Heterogeneity in survival analysis. Statistics in Medicine **7**:1121-1137.
2. **Angelis, R.D., Capocaccia, R., Hakulinen, T., Soderman, B. and Verdecchia, A.** 1999. Mixture models for cancer survival analyses: application to population-based data with covariates. Statistics in Medicine **18**:441-454.
3. **Berkson, J. and Gage, R.P.** 1952. Survival curves for cancer patients following treatment. Journal of the American Statistics Association **47**:501-515.
4. **Boag,** 1949. Maximum likelihood estimate of proportion of patients cured by cancer theraphy. Journal of the Royal Statistical Society, Series B **11**:15-44.
5. **Evans, A.E. and D'Angio, G.J.** 2005. Age at diagnosis and prognosis in children with neuroblastoma. Journal of Clinical Oncology. **23**:6443-6444.
6. **Farewell, V.T.** 1982. The use of mixture model for the analysis of survival data. Biometrics **38**:1041-1046.
7. **Farewell, V.T.** 1986. Mixture model in survival analysis: are they worth the risk? The Canadian Journal of Statistics **14**:257-262.
8. **Hiyama, E., Lehara, T., Sugimoto, T., Fukuzawa, M., Hayashi, Y., Sasaki, F., Sugiyama, M., Kondo, S., Yaneda, A., Yamaoka, H., Tajiri, T., Akazawa, K. and Ohtaki, M.** 2004. Screening at 6 months of age reduced mortality of neuroblastoma: A retrospective population-based cohort study including more than 13 million Japanese screened infants. Natural Science Center for Basic Research and Development, Hiroshima University, Hiroshima, Japan.
9. **Hiyama, E., Lehara, T., Sugimoto, T., Fukuzawa, M., Hayashi, Y., Sasaki, F., Sugiyama, M., Kondo, S., Yaneda, A., Yamaoka, H., Tajiri, T.,**

Akazawa, K. and Ohtaki, M. 2008. Effectiveness of screening for neuroblastoma at 6 months of age: a retrospective population-based cohort study. Lancet **371**:1173-1180.

10. **Hougaard, P., Myglegaard, P. and Borch-Johnsen, K.** 1994. Heterogeneity models of disease susceptibility, with application to diabetic nephropathy. Biometrics **50**:1178-1188.
11. **Hosmer, D.W. and Lemeshow, S.** 1989. Applied logistic regression. John Wiley & Sons.
12. **Kass, R.E. and Wasserman, L.** 1995. A reference Bayesian test for nested hypothesis with large samples. Journal of the American Statistics Association **90**:928-934.
13. **Kuk, A.C. and Chen, C.H.** 1992. A mixture model combining logistic regression with proportional hazard regression. Biometrika **79**:531-541.
14. **Lambert, P.C. and Thompson, J.R.** 2007. Estimating and modeling the cure fraction in population-based cancer survival analysis. Biostatistics **8**:576-594.
15. **Longini, I.M. and Halloran, M.E.** 1996. A frailty mixture model for estimating vaccine efficacy. Applied Statistics **45**:165-173.
16. **Maller, R. and Zhou, X.** 1996. Survival analysis with long-term survivors. John Wiley & Sons.
17. **McGilchrist, C.A. and Aibett, C.W.** 1991. Regression with frailty in survival analysis. Biometrics **47**:461-466.
18. **Mould, R.F. and Boag, J.W.** 1975. A test of several parametric statistical models for estimating success rate in the treatment of carcinoma cervix uteri. British Journal of Cancer **32**:529-550.
19. **Ohtaki, M. and Izumi, S.** 1999. Globally convergent algorithm without derivatives for maximizing a multivariate function. In Proceedings of symposium on "Exploratory methods and analyses for non linear structures of data with random variation" at Hiroshima.
20. **Price, D.L. and Manatunga, A.K.** 2001. Modeling survival data with a cured fraction using frailty models. Statistics in Medicine **20**:722-729.
21. **Sposto, R.** 2002. Cure model analyses in cancer: an application to data from the children's cancer group. Statistics in Medicine **21**:293-312.
22. **Sposto, R.** 1992. A comparison of test of the difference in population of patients who are cured. Biometrics **48**:87-99.
23. **Spix, C., Aarellid, T., Steler, C., Magnani, C., Kaatsch, P. and Michaelis, J.** 2001. Survival of children with neuroblastoma: time trends and regional differences in Europe, 1978-1992. European journal of cancer **37**:1515-1527.
24. **Swank, R.L., Fetterman, G.H., Sieber, W.K. and Kiesewetter, W.B.** 2002. Prognostic factor in neuroblastoma. Annals of Surgery **173**:428-435.
25. **Sy, J.P. and Taylor, J.M.G.** 2000. Estimation in Cox proportional hazard cure model. Biometrics **56**:227-236.
26. **Vaupel, J.W., Manton, K. and Stallard, E.** 1979. The impact of heterogeneity in individual frailty on the dynamics of mortality. Demography **16**:439-454.