Journal of Data Science **5**(2007), 41-51

Application of EM Algorithm to Mixture Cure Model for Grouped Relative Survival Data

Binbing Yu[1] and Ram C. Tiwari[2]
[1]*Information Management Services, Inc. and* [2]*National Cancer Institute*

Abstract: The interest in estimating the probability of cure has been increasing in cancer survival analysis as the cure of some cancer sites is becoming a reality. Mixture cure models have been used to model the failure time data with the existence of long-term survivors. The mixture cure model assumes that a fraction of the survivors are cured from the disease of interest. The failure time distribution for the uncured individuals (latency) can be modeled by either parametric models or a semi-parametric proportional hazards model. In the model, the probability of cure and the latency distribution are both related to the prognostic factors and patients' characteristics. The maximum likelihood estimates (MLEs) of these parameters can be obtained using the Newton-Raphson algorithm. The EM algorithm has been proposed as a simple alternative by Larson and Dinse (1985) and Taylor (1995). in various setting for the cause-specific survival analysis. This approach is extended here to the grouped relative survival data. The methods are applied to analyze the colorectal cancer relative survival data from the Surveillance, Epidemiology, and End Results (SEER) program.

Key words: Mixture cure model, relative survival.

1. Introduction

In the study of cancer incidence and mortality of the population, mixture cure models Boag (1949) have been used for failure time data with long term survivors. These models assume that a fraction of the patients are cured from the disease of interest. The cured individuals will never experience the event. However, the uncured patients are at risk of eventual failure from the disease, and the event would be observed with certainty if the complete follow-up were possible. The uncured patients experience excessive mortality rates in excess of the general population. Information on the patients' causes of death may not always be suitable for correcting survival rates as it may be inaccurate or unavailable. The relative survival rate Ederer, Axtell, and Cutler (1961). , defined as the ratio of the observed survival rate for the group of patients under consideration to the survival rate expected for a group taken from the general population matched to

the patients on age, race, sex and year of diagnosis, estimates the impact of the cancer on the mortality.

Population-based survival data are often presented in the form of a life table. The period of observations is grouped into a series of time intervals and the survival probabilities are estimated for each interval. Suppose that the survival data are categorized into I strata by demographical groups such as sex, age and the prognostic factors, e.g. historical stage, histologic type. The event times are grouped into J intervals $(t_{j-1}, t_j], j = 1, ..., J$, for each stratum, where $t_1 = 0$ and $t_1, ..., t_J$ are distinct event times and $t_J = \tau$ is the end of the follow-up. The observed data in stratum i and interval j consist of n_{ij}, the number of individuals who are alive at the beginning of the interval; d_{ij}, the number of individuals who die during the interval; l_{ij}, the number of individuals lost to follow-up during the interval. The number of people at risk during the interval adjusted for uniform loss is $n_{ij}^{'} = n_{ij} - \frac{1}{2}l_{ij}$ and the adjusted number of people surviving the interval is $s_{ij} = n_{ij}^{'} - d_{ij}$. The expected survival rate, E_{ij}, is the probability of surviving interval I_j for the comparable general population.

When the information on cause of death is not available or unreliable, relative survival is used as the measure of excess mortality (net survival) due to cancer of interest. Let $S(t) = P(T > t|x, \theta)$ be the net survival function, where θ is the parameter vector and x is the covariate vector. The interval specific net survival function is

$$r_{ij}(\theta; x_i) = \frac{S(t_j; x_i)}{S(t_{j-1}; x_i)}. \tag{1.1}$$

Let $p_{ij}(\theta; x_i, E_{ij}) = P(T > t_j|T \geq t_{j-1}; \theta, x_i, E_{ij})$ be the probability that an individual in stratum i survives until time t_j from all causes given that she is alive at t_{j-1}. It is generally thought that an additive hazards model is biologically more plausible and most appropriate for the population-based cancer survival analysis. The additive hazards model implies that $p_{ij}(\theta; x_i, E_{ij}) = r_{ij}(\theta; x_i)E_{ij}$. The loglikelihood function for the grouped relative survival data $(\mathbf{x}, s, d, E) = \{(x_i, s_{ij}, d_{ij}, E_{ij}), i = 1, ..., I, j = 1, ..., J\}$ is

$$\ell(\theta|\mathbf{x}, s, d, E) = \sum_{i=1}^{I}\sum_{j=1}^{J}\Big\{s_{ij}\log p_{ij}(\theta; x_i, E_{ij}) + d_{ij}\log(1 - p_{ij}(\theta; x_i, E_{ij}))\Big\}. \tag{1.2}$$

and the maximum likelihood estimate (MLE) of θ is obtained by maximizing the loglikelihood.

When there is no cure, i.e., $c = 0$, the survival function $S(t; x)$ is usually specified by parametric or semi-parametric models. Prentice and Gloeckler (1978) discussed the semi-parametric proportional hazards (PH) regression model for the

grouped cause-specific survival data. Hakulinen and Tenkanen (1987) extended the PH regression model to the relative survival data and provided a simpler estimation method. When cure is a possibility, the survival function is usually modeled by a mixture model Boag (1949), Farewell (1982):

$$S(t;x) = c(x) + (1 - c(x))G(t;x), \tag{1.3}$$

where $c(x)$ is the probability of cure and $G(t;x)$ denotes the survival (latency) distribution for the uncured individuals (susceptibles). Define a binary cure status z where $z = 0$ indicates that an individual will experience the event eventually, and $z = 1$ indicates that the individual will never experience the event, i.e., will be cured. The cure probability is typically modeled as Farewell (1982):

$$c(x) = P(z = 1|x) = \frac{\exp(\beta_c x)}{1 + \exp(\beta_c x)}. \tag{1.4}$$

Let the latency distribution be $G(t) = \exp\{-\Lambda(t)\}$, where $\Lambda(t)$ is the cumulative hazard function. The Cox PH model assumes that

$$\Lambda(t) = \Lambda_0(t) \exp(\beta_\lambda x) \tag{1.5}$$

where $\Lambda_0(t)$ is the unspecified baseline cumulative hazard function and for the Weibull model, $\Lambda(t) = [\lambda_x t]^\delta$, where δ and λ_x are the shape and scale parameters, respectively and λ_x is usually modeled as

$$\lambda_x = \exp(\beta_\lambda x). \tag{1.6}$$

Note that when $\delta = 1$, $G(t)$ reduces to an exponential model. Other parametric models, e.g., loglogistic model and lognormal models, have also been used to model $G(t)$ (Yu *et al.*, 2004). Here, we will focus on the Cox model and the Weibull model.

The parametric mixture cure models have been used to analyze population-based survival data by Gamel *et al.* (2000) and De Angelis *et al.* (1999). Sy and Taylor (2000) provided a semi-parametric mixture cure model to the continuous survival data. Recently, Yu *et al.* (1999) developed a software for analyzing the grouped relative survival data using parametric cure models. The software is available for public use at http://www.srab.cancer.gov/cansurv. In this paper, we provide a simple alternative, the EM algorithm, to estimate the mixture cure model for grouped relative survival data. The rest of the paper is organized as follows. In Sections 2 and 3, we describe the EM algorithm for mixture cure model (1.3). In Section 4, we illustrate the methods using the colorectal cancer survival data from SEER 9 registries.

2. EM Algorithm for the Mixture Cure Model for Grouped Relative Survival Data

When there is no cure, i.e., $c = 0$, the survival functions for the Weibull and Cox regression models are given by

$$S(t_j; x_i) = \begin{cases} \exp[-\{\exp(\beta_\lambda x_i)t_j\}^\delta] & \text{Weibull model,} \\ \exp\{-\Lambda_0(t_j)\exp(\beta_\lambda x_i)\} & \text{Cox model.} \end{cases}$$

The interval specific survival probabilities are

$$r_{ij}(\theta; x_i) = \begin{cases} \exp\{-\exp(\beta_\lambda x_i \delta)(t_j^\delta - t_{j-1}^\delta)\} & \text{Weibull model,} \\ \exp\{-\exp(\alpha_j + \beta_\lambda x_i)\} & \text{Cox model,} \end{cases} \tag{2.1}$$

where $\alpha_j = \log(\Lambda_0(t_j) - \Lambda_0(t_{j-1}))$ for the Cox model. Notice that $\alpha_j, j = 1, ..., J$ are the coefficients of the interval indicators $I_j, j = 1, ..., J$.

Because the number of people surviving through the interval I_{ij} follows a binomial distribution $Bin(n_{ij}^{'}, p_{ij}, (\theta; x_i, E_{ij}))$, equation (2.1) implies generalized linear models (GLM) with link functions

$$\beta_\lambda x_i = \frac{1}{\delta}\Big[\log\Big\{-\log\Big(\frac{\mu_{ij}}{n_{ij}^{'}E_{ij}}\Big)\Big\} - \log(t_j^\delta - t_{j-1}^\delta)\Big] \tag{2.2}$$

for Weibull model, and

$$\alpha_j + \beta_\lambda x_i = \Big[\log\Big\{-\log\Big(\frac{\mu_{ij}}{n_{ij}^{'}E_{ij}}\Big)\Big\} - \log(t_j - t_{j-1})\Big] \tag{2.3}$$

for Cox model, where $\mu_{ij} = n_{ij}^{'} p_{ij}(\theta; x_i, E_{ij})$ is the mean response. Other commonly used forms of latency distribution such as the exponential and log-logistic models can also be put in the framework of GLM with special link functions (Weller *et al.*, 1999). The parameters to be estimated are $\beta_G = (\beta_\lambda, \delta)$ for the Weibull model and $\beta_G = (\beta_\lambda, \alpha_1, ..., \alpha_J)$ for the Cox model. The MLEs of the parameters can be found by using the standard statistical packages such as GLIM and SAS (Weller *et al.*, 1999; Hakulinen and Tenkanen (1987).

When the mixture cure model (1.3) is applied, the parameters of interest are the cure parameter β_c and the latency parameter β_G. Gamel *et al.* (2000); De Angelis *et al.* (1999); and Yu *et al.* (1999) used the Newton-Raphson method to find the MLEs of the parameters. Here, we propose the EM algorithm, a simple alternative which is easy to implement using the standard statistical packages.

Let $S_0(t, x)$ denote the expected cumulative survival probability. We can write $S_0(t_j; x_i) = \prod_{k=1}^{j} E_{ij}$. Let $c_i = c(x_i), S_i(t_j) = S(t_j; x_i), G_i(t_j) = G(t_j; x_i)$

and $S_{0i}(t_j) = S_0(t_j; x_i)$ for $i = 1, ..., I$. After plugging (1.1) into the loglikelihood (1.2), the loglikelihood function can be expressed as

$$\begin{aligned}
&\ell(\theta|\mathbf{x}, s, d, E) \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J}\Big\{m_{ij}\log\left[S_i(t_j)S_{0i}(t_j)\right] + d_{ij}\log\left[S_i(t_{j-1})S_{0i}(t_{j-1}) - S_i(t_j)S_{0i}(t_j)\right]\Big\} \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J}\Big\{m_{ij}\log\left[c_iS_{0i}(t_j) + (1-c_i)G_i(t_j)S_{0i}(t_j)\right] \\
&\quad + d_{ij}\log[(1-c_i)(G_i(t_{j-1})S_{0i}(t_{j-1}) - G_i(t_j)S_{0i}(t_j)) \\
&\quad + c_i(S_{0i}(t_{j-1}) - S_{0i}(t_j))]\Big\},
\end{aligned}$$

where

$$m_{ij} = \begin{cases} s_{ij} - (s_{i,j+1} + d_{i,j+1}) = \frac{1}{2}(l_{ij} + l_{i,j+1}) & \text{if} \quad j < J \\ s_{iJ} & \text{if} \quad j = J. \end{cases}$$

Note that m_{ij} can be interpreted as the number of individuals censored right at time t_j. If we have observed the number of cured and censored patients at interval j in stratum i, the individuals can be classified into one of the following four groups : not cured and die; not cured and censored; cured but die from causes other than cancer; cured and censored (Table 1).

Table 1: Classification of individuals by cure and censoring status

Number	Status	Probability
d^*_{ij}	Not cured, die	$(1-c_i)[S_{0i}(t_{j-1})G_i(t_{j-1})$ $- G_i(t_j)S_{0i}(t_j))]$
m^*_{ij}	Not cured, censored	$(1-c_i)S_{0i}(t_j)G_i(t_j)$
$d_{ij} - d^*_{ij}$	Cured, die from other causes	$c_i[S_{0i}(t_{j-1}) - S_{0i}(t_j)]$
$m_{ij} - m^*_{ij}$	Cured, censored for other cause	$c_iS_{0i}(t_j)$

The loglikelihood based on the complete data $(d, m, d^*, m^*) = \{(d_{ij}, m_{ij}, d^*_{ij}, m^*_{ij}), i = 1, ..., I, j = 1, ..., J\}$ is

$$\begin{aligned}
&\ell(\theta|\mathbf{x}, d, m, d^*, m^*, E) \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J}\Big[(d_{ij} - d^*_{ij})\log\{c_i(S_{0i}(t_{j-1}) - S_{0i}(t_j))\} \\
&\quad + (m_{ij} - m^*_{ij})\log\{c_iS_{0i}(t_j)\} + d^*_{ij}\log\{(1-c_i)(S_{0i}(t_{j-1})G_i(t_{j-1}) \\
&\quad - G_i(t_j)S_{0i}(t_j))\} + m^*_{ij}\log\{(1-c_i)S_{0i}(t_j)G_i(t_j)\}\Big]. \qquad (2.4)
\end{aligned}$$

The loglikelihood can be decomposed into two parts:

$$\begin{aligned}
\ell(\beta_c) &= \sum_{i=1}^{I}\sum_{j=1}^{J}\Big\{(d_{ij}-d^*_{ij}+m_{ij}-m^*_{ij})\log c_i + (d^*_{ij}+m^*_{ij})\log(1-c_i)\Big\},\\
\ell(\beta_G) &= \sum_{i=1}^{I}\sum_{j=1}^{J}\Big[m^*_{ij}\log\{G_i(t_j)S_{0i}(t_j)\} + d^*_{ij}\log\{G_i(t_{j-1})S_{0i}(t_{j-1})\\
&\quad - G_i(t_j)S_{0i}(t_j)\} + (m_{ij}-m^*_{ij})\log S_{i0}(t_j)\\
&\quad + (d_{ij}-d^*_{ij})\log\{S_{0i}(t_{j-1})-S_{0i}(t_j)\}\Big].
\end{aligned}$$

Let

$$s^*_{ij} = \begin{cases} m^*_{iJ} & \text{if} \quad j = J\\ m^*_{ij} + s^*_{i,j+1} + d^*_{i,j+1} & \text{if} \quad j < J. \end{cases}$$

Here, s^*_{ij} are the uncured individuals who survive the interval $I_j = (t_{j-1}, t_j]$. The loglikelihood $\ell(\beta_G)$ for uncured individuals can be expressed as

$$\ell(\beta_G) = \sum_{i=1}^{I}\sum_{j=1}^{J}\Big\{s^*_{ij}\log p^*_{ij}(\theta; x_i, E_{ij}) + d^*_{ij}\log(1-p^*_{ij}(\theta; x_i, E_{ij}))\Big\} + \Delta, \quad (2.5)$$

where

$$p^*_{ij}(\theta; x_i, E_{ij}) = \frac{G_i(t_j)S_{0i}(t_j)}{G_i(t_{j-1})S_{0i}(t_{j-1})} = \frac{G(t_j; x_i)}{G(t_{j-1}; x_i)}E_{ij}$$

and

$$\Delta = (m_{ij}-m^*_{ij})\log S_{i0}(t_j) + (d_{ij}-d^*_{ij})\log\{S_{0i}(t_{j-1})-S_{0i}(t_j)\}$$

is a constant given the complete data.

In the E-step, compute the conditional expectation of the missing data (m^*_{ij}, d^*_{ij}) given the observed data (m_{ij}, d_{ij}) and current estimates $\hat\theta$, i.e.,

$$E(m^*_{ij}|d_{ij}, m_{ij}, \hat\theta) = m_{ij}w_{ij} \quad \text{and} \quad E(d^*_{ij}|d_{ij}, m_{ij}, \hat\theta) = d_{ij}\eta_{ij}, \qquad (2.6)$$

where

$$w_{ij} = \frac{(1-c_i)G_i(t_j)}{c_i + (1-c_i)G_i(t_j)} \quad \text{and} \quad \eta_{ij} = 1 - \frac{c_i(S_{0i}(t_{j-1}) - S_{0i}(t_j))}{S_i(t_{j-1})S_{0i}(t_{j-1}) - S_i(t_j)S_{0i}(t_j)}.$$

In the M-step, the MLE of β_c from $\ell(\beta_c)$ is obtained by fitting a logistic regression; the MLE of the parameters β_G can be obtained by GLIM or PROC GENMOD. For details see Weller *et al.* (1999) and Hakulinen and Tenkanen (1987).

The covariance matrix of $\hat{\theta}$ is given by the inverse of the information matrix, and this can be estimated by the covariance matrix of the score vectors. The score vectors are $U_{ij} = \frac{\partial \ell_{ij}(\theta)}{\partial \theta}$, where

$$\frac{\partial \ell_{ij}(\theta)}{\partial \theta_k} = \sum_{i=1}^{I}\sum_{j=1}^{J}\Big\{m_{ij}\frac{\frac{\partial S_i(t_j)}{\partial \theta_k}}{S_i(t_j)} + d_{ij}\frac{\frac{\partial S_i(t_{j-1})}{\partial \theta_k} - E_{ij}\frac{\partial S_i(t_j)}{\partial \theta_k}}{S_i(t_{j-1}) - E_{ij}S_i(t_j)}\Big\},$$

and

$$\frac{\partial S_i(t)}{\partial \theta_k} = \begin{cases} x_{ik}c_i(1-c_i)(1-G_i(t)) & \text{if } \theta_k \in \beta_c, \\ x_{ik}\delta(1-c_i)G_i(t)\log G_i(t) & \text{if } \theta_k \in \beta_G, \\ (1-c_i)G_i(t)\log G_i(t)\{\log(-\log G_i(t))\}/\delta & \text{if } \theta_k = \delta. \end{cases}$$

3. Application

Colorectal cancer is the third most common cancer and the second common cause of cancer death in the US, with about 145,290 new cases and 56,290 deaths expected in 2005. When men and women are considered separately, colorectal cancer is the third most common cause of cancer death in each sex. Over the past decade, colorectal cancer incidence and mortality rates have modestly decreased or remained level. Until age 50, men and women have similar incidence and mortality rates; after age 50, men are more vulnerable. There are striking differences with respect to racial and ethnic groups in both incidence and mortality. As cancer treatments progress, the cancer patients may live long enough and die from other causes.

Several different types of treatments are often combined to treat colorectal cancer. Surgery to remove the tumor is the cornerstone of treatment for tumors found to be potentially curable. Additional chemotherapy and, in cases of rectal cancer, chemotherapy and radiation therapy, have been proven to improve a patient's chance for cure and longer life. It is of interest to estimate the cure fractions by race, sex and historical stages using the population-based cancer survival data.

The Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute is an authoritative source of information on the cancer incidence and survival in the United States. Case ascertainment for the SEER-9 registries began on January 1, 1973, in the states of Connecticut, Iowa, New Mexico, Utah, Hawaii, the metropolitan areas of Detroit, San Francisco-Oakland and Atlanta and the 13-county Seattle-Puget Sound area. The SEER Registries routinely collect data on patient demographics, primary tumor site, morphology, stage at diagnosis, first course of treatment, and follow-up for vital status. Here

we used the method described in Section 2 to analyze the colorectal cancer relative survival data from SEER-9 registries.

The colorectal cancer incidence and mortality rates are more than 35% higher in men than in women (American Cancer Society, 2005). Hence, the data were analyzed by male and female, separately. The parameter estimates and standard errors of the Weibull cure models are presented in Table 2. For both men and women, race and historic stages are significantly related to both the cure fraction and the short-term survival rate (latency parameter λ_x), except that the difference of short-term survival between the white females and the black females is not significant. The cure fraction estimates are listed in Table 3. The localized colorectal cancer has very high cure fractions 74.2-79.3%, which are much higher than those for the regional (40.4-50.4%) and the distant colorectal cancer (4.6-6.8%). This indicates that early detection of colorectal cancer can substantially improve the cure rates, which reinforces the benefit of colorectal cancer screening. Also notice that the cure rates are higher for the whites than those for the blacks.

Table 2: Parameter estimates of the Weibull cure models for the colorectal cancer relative survival data

	Male		Female	
Parameters	Estimate	Std-Error	Estimate	Std-Error
Parameter in cure $c(x)$				
Intercept	1.288	0.015	1.345	0.014
Race (Black vs. White)	-0.212	0.045	-0.287	0.036
Localized				
Regional	-1.463	0.017	-1.328	0.015
Distant	-4.109	0.030	-3.958	0.025
Parameters in scale λ_x				
Intercept	-1.734	0.016	-1.691	0.017
Race (Black vs. White)	0.106	0.013	0.009	0.013
Localized				
Regional	0.369	0.016	0.509	0.016
Distant	1.687	0.016	1.774	0.017
Shape Parameter δ				
δ	0.980	0.002	0.940	0.002

Table 3: Estimates of cure fraction by sex, race and historic stage

Sex	Race	Localized	Regional	Distant
Female	White	79.3	50.4	6.8
	Black	74.2	43.3	5.2
Male	White	78.4	45.6	5.6
	Black	74.6	40.4	4.6

4. Discussion

This paper provides an EM algorithm to fit the mixture cure model to the grouped relative survival data. It can fit both a parametric or a semi-parametric mixture cure model. This algorithm utilizes the standard statistical software to achieve the M-step and is easier to implement than the Newton-Raphson. The EM algorithm is usually stabler than the Newton-Raphson method (Yu *et al.*, 1999) and the convergence of the EM algorithm is generally fast for the grouped survival data. A SAS macro is available to implement the EM algorithm.

A word of caution is needed about the existence of cure fraction and stability of cure fraction estimates. The mixture cure model generally requires a sufficiently long follow-up and large samples to identify the parameters in cure fraction and latent survival distribution for uncured individuals (Farewell, 1986). The cure fraction estimates may be sensitive to the specification of latency distributions when the follow-up time is not sufficient (Yu *et al.*, 2004). We need to be cautious in interpreting the cure fraction estimate. Hence, we would only suggest to use the mixture cure models in situations where it is clear that a cured group exists and where there is sufficient follow-up beyond the time when most of the events occur.

Acknowledgments

The authors wish to thank Drs. Eric J. Feuer, K. Cronin and L. Clegg for their insightful suggestions and comments.

References

American Cancer Society (2005). *C*olorectal Cancer Facts and Figures Special Edition 2005. American Cancer Society.

Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society* **11**, 15-44.

De Angelis, R., Capocaccia, R., Hakulinen, T., Soderman, B. and Verdecchia, A. A. (1999). Mixture models for cancer survival analysis: application to population-based data with covariates *Statistics in Medicine* **18**, 441-454.

Ederer, F., Axtell, L. M. and Cutler, S. J. (1961). *The Relative Survival Rate: A Statistical Methodology Monograph No. 6,*, Bethesda: National Cancer Institute.

Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, **38**, 257-262.

Farewell, V. T. (1986). Mixture models in survival analysis: Are they worth the risk? *Canadian Journal of Statistics*, **14**, 257-262.

Gamel, J. W., Weller, E. A. , Wesley, M. N. and Feuer, E. J. (2000). Parametric cure models of relative and cause-specific survival for grouped survival times. *Computer Methods and Programs in Biomedicine*, **61**, 99-110.

Hakulinen, T. and Tenkanen, L. (1987). Regression analysis of relative survival rates. *Applied Statistics*, **36** 309-317.

Larson, M. G. and Dinse G. (1985). A mixture model for the regression analysis of competing risks data. *Applied Statistics* **34**, 201-211.

Prentice, R. L. and Gloeckler, L. A. (1978). Regression analysis of grouped survival data with application to breast cancer data. *Biometrics* **34**, 57-67.

Surveillance, Epidemiology and End Results Program (1999). *The Portable Survival System/Mainframe Survival System*, National Cancer Institute.

Sy, J. P. and Taylor, J. M. G. (2000). Estimation in a Cox proportional hazards cure model. *Biometrics*, **56**, 227-36.

Taylor J. M. G. (1995). Semiparametric estimation in failure time mixture models. *Biometrics* **51**, 899-907.

Weller, E. A., Feuer, E. J., Frey, C. M. and Wesley, M. N. (1999). Parametric relative survival regression using generalized linear models with application to Hodgkin's lymphoma. *Applied Statistics*, **48**, 79-89.

Yu, B. and Tiwari, R. C., Cronin, K. A. and Feuer, E. J. (2004). Cure fraction estimation from the mixture cure models for grouped survival data. *Statistics in Medicine* **23**, 1733-1747.

Yu, B., Tiwari, R. C., Cronin, K. A., McDonald, C. and Feuer, E. J. (2008). CANSURV: a Windows program for population-based cancer survival analysis. *Computer Programs and Methods in Biomedicine* **80**, 195-203.

Received July 19, 2005; accepted January 17, 2006.

Binbing Yu
Information Management Services, Inc.,
12501 Prosperity Dr. Suite 200
Silver Spring, MD 20910, USA.
yub@imsweb.com

Ram C. Tiwari
Statistical Applications and Research Branch
National Cancer Institute
6116 Executive Boulevard
Bethesda, MD 20892, USA
tiwarir@mail.nih.gov