CANSURV: A Windows program for population-based cancer survival analysis

Binbing Yu, Ram C. Tiwari, Kathleen A. Cronin, Chris McDonald, Eric J. Feuer

*Corresponding author.

Summary: Patient survival is one of the most important measures of cancer patient care (the diagnosis and treatment of cancer). The optimal method for monitoring the progress of patient care across the full spectrum of provider settings is through the population-based study of cancer patient survival, which is only possible using data collected by population-based cancer registries. The probability of cure, "statistical cure", is defined for a cohort of cancer patients as the percent of patients whose annual death rate equals the death rate of general cancer-free population. Mixture cure models have been widely used to model failure data. CANSURV (CANCer SURVival) is a Windows software fitting both the standard survival models and the cure models to population-based cancer survival data. CANSURV can analyze both cause-specific survival data and, especially, relative survival data, which is the standard measure of net survival in population-based cancer studies. It can also fit parametric (cure) survival models to the individual data. The program is available at http://srab.cancer.gov/cansurv. The colorectal cancer survival data from the Surveillance, Epidemiology and End Results (SEER) program [Surveillance, Epidemiology and End Results Program, The Portable Survival System/Mainframe Survival System, National Cancer Institute, Bethesda, 1999.] of the National Cancer Institute, NIH is used to demonstrate the use of CANSURV program. Published by Elsevier Ireland Ltd.
way that the results are representative of the population. For example, Feuer et al. [2] and Weller et al. [3] use survival models to describe the impact of breakthrough clinical trials on population survival, while Mariotto et al. [4] uses models of this type to extrapolate survival from areas covered by the SEER program to the rest of the U.S.

1.1. The population-based survival data

The cancer registries record all new cancer cases and the cancer patients are followed-up until death or censored. The patients who are alive at the end of the study or are lost to follow-up during the study are censored. When the outcome event is death due to the cancer of interest, the patients who die due to a competing event other than the cancer of interest are also considered to be censored. The survival experience of a heterogeneous population may be associated with a vector of covariates \( x \), such as race, age and year of diagnosis, etc.

When the survival time \( T \) and censoring status \( \delta \) are recorded for each individual with covariates \( x \), the survival data are listed individually as \( \{ (T_i, \delta_i, x_i), i = 1, \ldots, n \} \). Usually \( \delta_i = 1 \) when the time \( t_i \) is a death time. The data from clinical survival study usually takes this format.

The survival data based on large populations are often presented as a life table, where the survival times are grouped into a series of time intervals and the population is stratified by demographic variables. Let \( 0 = t_0 < t_1 < \cdots < t_j \) be the grouped survival times and let the \( j \)th interval be \( (t_{j-1}, t_j) \). For the \( j \)th stratum with covariates \( x_i \), let \( n_j \) be the number of people alive at the beginning of the \( j \)th interval and let \( d_j(l_j) \) be the number of people dying (lost to follow-up) during the \( j \)th interval, respectively.

1.2. Cause-specific and relative survivals

A quantity of interest is the proportion of cancer patients who would have survived \( t \) years or more following diagnosis in the hypothetical situation where cancer was the only possible cause of death. This theoretical quantity is known as the net survival (excess mortality) rate \( S(t) \).

One method of estimating \( S(t) \) is to consider only deaths directly caused by the cancer of interest as outcome events and to consider the survival times of patients who die of other causes to be censored, giving rise to what is called the cause-specific survival rate. The cause-specific survival \( S_c(t) \) is a measure of net survival in the competing risks setting. It requires reliably coded information on cause of death and usually assumes that the cancer mortality is independent of the mortality due to other competing risks. Even when the cause of death information is available to the cancer registry via death certificates, it is often difficult to ascertain whether or not the cancer of interest is actually the primary cause of death. Various definitions can be used for counting a person as having died of their cancer. A "narrow" definition uses only the specific cancer of interest while a "wider" definition may use all cancers, common metastatic sites and possible other non-cancer causes of death associated with more immediate causes of death from a particular cancer. When using the wider definition, often the analysis is restricted to those who only have the single cancer of interest, both prior to or after the diagnosis of that cancer. While this restriction helps assure that death from another cancer is not associated with the diagnosis of another primary cancer, it limits the generalizability of the analysis. These and associated issues are discussed in Boer et al. [5].

An alternative method of estimating the net survival \( S(t) \) is to use the relative survival \( S_R(t) \), defined as the observed survival rate due to all causes of death in the patient group, \( S_A(t) \), divided by the expected survival rate of a comparable group from the general population, \( S_E(t) \), who are assumed to be free of the cancer of interest; that is \( S_R(t) = S_A(t)/S_E(t) \). A major advantage of this measure is that information on cause of death is not required, thereby circumventing problems with the inaccuracy or non-availability of death certificates. The expected survival rate is usually estimated from the nationwide population life tables stratified by age, sex, calendar time and where applicable, by race.

2. The survival models

Many statistical models, including parametric models and semi-parametric models, have been used to model \( S(t) \). The standard survival models assume that all the patients will die from the cancer of interest. As medical treatments progress, a proportion of the cancer patients for some cancer sites survive a long time and some of them become eventually cured. The mixture cure models and several alternatives have been used to model the survival data when the cure is a possibility. The mixture cure model [7, 8] assumes that

\[
S(t) = c + (1 - c)G(t),
\]

where \( c \) is the cure fraction and \( G(t) \) is the (latency) survival function for the uncured patients. The latency distribution \( G(t) \) can take the form of parametric or semi-parametric distributions. Among the
parametric models, lognormal (LN), loglogistic (LL), Weibull (WB) and Gompertz (GP) distributions are widely used to model the survival time. After reparametrization [10], these survival functions can be expressed as

\[
G(t; \mu, \sigma) = \begin{cases} 
1 - \phi \left( \frac{\log t - \mu}{\sigma} \right), & \text{lognormal;} \\
1 + \exp \left( - \frac{\log t - \mu}{\sigma} \right)^{-1}, & \text{logistic;} \\
\exp \left( - \exp \left( - \frac{\log t - \mu}{\sigma} \right) \right), & \text{Weibull;} \\
\exp \left( \sigma(1 - e^{\mu}) / \mu \right), & \text{Gompertz.}
\end{cases}
\]

The parameters \((c, \mu, \sigma)\) may depend on the covariates as \(c = \{1 + \exp(- \frac{\rho_i x_i^{(3)})}{\sigma}\}^{-1}\), \(\mu(x) = \beta_c^x x_i^{(3)}\) and \(\sigma(x) = \exp(\beta^x x_i^{(3)})\). When \(c(x) = 0\), this model reduces to the standard parametric survival models. In (1), covariates \(x_i\) can appear in \(x_i^{(1)}\), \(x_i^{(2)}\) and \(x_i^{(3)}\) simultaneously.

The latency survival function \(G(t)\) can also be a semi-parametric proportional hazards model,

\[
G(t|x) = G_0(t)^{\exp(\hat{\beta} x)},
\]

where the baseline survival function \(G_0(t)\) is modeled by piecewise exponential distribution.

A generalization of model (1) with a power function \(\delta\) has been used in the estimation of complete-ness index of cancer prevalence [11]:

\[
S(t) = [c + (1 - c) G(t)]^{\delta},
\]

where the power function \(G(t|x) = \exp(\hat{\beta} x)\).

Although the cure model (3) reduces to the simple mixture model (1) when \(c(x) = 1\), in practice we do not allow the same covariates \(x_i\) to be used in \(c\) and \((\mu, \sigma)\) simultaneously because this may cause singularity of the estimated information matrix.

The parameters of interest are \(\hat{\beta} = (\hat{\beta}_c, \hat{\beta}_d, \hat{\beta}_e)\) for the parametric cure models and \(\hat{\beta} = (\hat{\beta}_c, \hat{\beta}_T)\) for the semi-parametric proportional hazards model. If one is interested in interpreting which parameters depend on particular covariates, then it is preferable to model them as a function of \((c, \mu, \sigma)\) directly. If fit, rather than interpretability, is of interest, the power parameter \(\delta\) has the advantage that no decision has to be reached with respect to which parameters of the cure model need to vary as a function of the covariates. Except in unusual circumstances, a covariate would not be used in a model in all three parameters \((c, \mu, \sigma)\) simultaneously. Yu et al. [9] explored the cure fraction estimation from the mixture cure models for grouped survival data and found that the cure fraction estimate may be sensitive to the assumptions of latency distribution and follow-up time. In these cases, the interpretation of cure fraction estimate needs caution. However, the mixture cure models can still provide a better estimate of long-term survival than the standard survival models.

CANSURV is a Windows application to analyze population-based survival data. The software is designed for the relative survival data, but it can be used for cause-specific survival data as well. It can handle both individual and grouped survival data. Except that the semi-parametric Cox model is not available for individual survival data, CANSURV can fit standard survival models (no cure), e.g., parametric and semi-parametric Cox models, two types of parametric cure models (1) and (3) to both individual and grouped survival data.

In Section 3, we discuss the computational methods to obtain the maximum likelihood estimates of the parameter \(\hat{\beta}\). The interface and capabilities of the CANSURV software are described in Section 4. In Section 5, the colorectal cancer survival data from SEER program are used as an application to demonstrate the usage of the software.

### 3. Computational methods

In this section, we first describe the computational methods for the relative survival analysis and then show that the computation for cause-specific survival can be regarded as a special case. If the relative survival \(S(t)\) is used as the measure of net survival \(S(t)\) due to cancer, then the deaths due to all causes are recorded and the corresponding survival function is \(S(t) = S(t)S(t)\), where \(S(t)\) is the expected survival function estimated from general population.

When the survival data are grouped and presented as a life table, the observed data are \([x_i, t_1, t_2, n_i, d_i, l_i, f_1, \ldots, f_j, \ldots, J]\) where \(d_i\) is the number of deaths due to all causes during the \(j\)th interval in the \(i\)th stratum, \(f_1, \ldots, f_j, \ldots, J\) and \(E(t_i|x_i) = S(t_i|x_i)/S(t_i)\) is the interval expected survival probability. By actuarial assumption that the censoring occurs uniformly throughout the interval \((t_i-1, t_i)\), the effective number of people at risk is \(\hat{n}_i = n_i - \frac{1}{2}d_i\). The number of people dying during the interval \(d_i\) follows a binomial distribution, i.e., \(d_i = B(n_i - \hat{n}_i, 1 - p_0(\hat{\beta}))\) with parameters \(p_0(\hat{\beta}) = P(T > t_i | T > t_i-1)\),

\[
\hat{\beta} = \begin{cases} 
\frac{S_0(t_i|x_i)}{S(t_i-1|x_i)} = \frac{S(t_i|x_i)}{S(t_i-1|x_i)S(t_i)}
\end{cases}
\]
The loglikelihood function for the grouped survival data is

\[
L(\theta) = \prod_{i=1}^{n} \left( \sum_{j=1}^{d_i} \log \left(1 - p_j(\theta)\right) \right) + (n_0 - d_i) \log p_{d_i}(\theta).
\]

The individual data are presented as \((x_i, t_i, \delta_i, \lambda_i, i = 1, \ldots, n)\), where \(\lambda_i\) is the expected hazard rate. Let \(f(t|x)\) be the density function of the expected survival. The expected hazard rate for the \(i\)th individual

\[
\lambda_i = \lambda(t|x_i) = \frac{f(t|x_i)}{S(t|x_i)}
\]

is known and treated as a constant.

Because \(S(t) = S(t|S(t))\), the density function for the observed survival is

\[
f_a(t) = -\frac{dS(t)}{dt} = -S(t)\frac{dS(t)}{dt} - S(t)\frac{dS(t)}{dt}.
\]

The loglikelihood function for the individual data is

\[
L(\theta) = -\sum_{i=1}^{n} \left( \lambda_i\log f_a(t_i|x_i) + (1 - \lambda_i)\log S(t_i|x_i) \right)
= -\sum_{i=1}^{n} \left( \lambda_i\log \left(S(t_i|x_i)\lambda_i - \frac{dS(t_i|x_i)}{dt}\right) + (1 - \lambda_i)\log S(t_i|x_i) \right) + \Delta,
\]

where \(\Delta = \sum_{i=1}^{n} \log S(t_i|x_i)\) is a constant independent of \(\theta\).

The maximum likelihood estimates (MLEs) of the parameters \(\theta\) can be obtained by the Newton–Raphson method. The estimate of \(\theta\) is updated by iterations

\[
\theta^{(k+1)} = \theta^{(k)} - \gamma L'(\theta^{(k)}) [L'(\theta^{(k)})]^{-1}, \quad k = 0, 1, \ldots,
\]

where \(\gamma = 1\) and \(L'\) and \(L''\) are the first and second partial derivatives of \(L\) with respect to \(\theta\). Convergence is achieved when either the distance \(\|\theta^{(k+1)} - \theta^{(k)}\|\) or the maximum absolute value of each element of \(L'(\theta^{(k)})\) falls below a specified threshold \(\epsilon\) with default value \(10^{-4}\). Upon convergence, the covariance matrix of the parameters is given by the inverse of the observed information matrix \(-L''(\theta^{(k)})\). On occasions, the procedure described above may fail to converge to a stable solution. Such instability may be overcome by reducing \(\gamma\) until we have an increased loglikelihood.

The initial estimates \(\theta^{(0)} = (\mu_0^{(0)}, \mu_1^{(0)}, \mu_2^{(0)}, \mu_3^{(0)}, \rho_0^{(0)}, \rho_1^{(0)}, \rho_2^{(0)}, \rho_3^{(0)})\) for grouped survival data have been discussed by Gamel et al. [10]. For individual survival data, the initial estimate for \(\rho_0^{(0)}\) is obtained by fitting the logistic regression \(P(t = 1) = [1 + \exp(-\rho_0 x(t))]^{-1}\) and similarly, the initial estimate for \(\rho_0^{(0)}\) is obtained by fitting regression \(\log(t_i) = \rho_0 x(t_i) + e_i\) for uncensored patients. Sometimes, the obtained final estimates \(\hat\theta\) are not the global MLEs. To ensure that the global maximum is obtained, different initial parameter values are tried by oscillating one component of the initial parameters \(\theta^{(0)}\), for example, using \((\mu_0^{(0)}, (1 \pm u)\mu_1^{(0)}, \mu_2^{(0)}, \mu_3^{(0)})\) where the default value of \(u\) is 0.5.

When the measure of net survival is the cause-specific survival function, i.e., \(S(t) = S(t|S(t))\), then for the grouped survival data, \(d_i\) are the number of deaths due to the cancer of interest; and for the individual survival data, \(\delta_i = 1\) only if the cause of death is the cancer of interest. The loglikelihood functions are the same except that \(S(t) = S(t|S(t))\) and \(S(t)\) is replaced by \(S(t|S(t))\) in the computation.

4. Description of CANSURV

The CANSURV (http://srab.cancer.gov/cansurv) software has three tabs, “Input File”, “Parameter Specifications” and “Output Specifications” to allow the user to specify the model and output options. This section gives a review of these options. Because the software is designed for the population-based cancer survival data, e.g., the SEER program data [1], some features are specifically applicable to the data extracted from SEER-STAT software (see http://seer.cancer.gov/seerstat).

4.1. Input file

This tab (Fig. 1) allows the users to specify the input data. If the input data is extracted from SEER-STAT, the CANSURV identifies whether it is grouped or individual survival data and whether the net survival is the relative survival or the cause-specific survival. The appropriate analysis method is automatically chosen depending on the survival data type.

4.2. Model specifications

The “Model Specifications” tab (Fig. 2) allows the user to specify the survival models and to control the computation.
4.2.1. The survival models and latency distribution G(t)

The options for the survival models are:

- **Standard survival models (no cure, i.e., c = 0):** The standard survival model can be either a parametric model with latency distribution G(t) or a semi-parametric Cox proportional hazards (PH) model.
- **Parametric cure models:** The cure models can be either the regular mixture cure model (1) or the mixture cure model with power function (3).

4.2.2. Computation specifications

CANSURV also has specifications to control the Newton–Raphson algorithm. The default number of iterations is 50 and the default convergence criteria are $L'(t) < 10^{-6}$ or $|L(t)^{(k+1)} - L(t)^{(k)}| < 10^{-6}$. The user can also specify whether to try different initial estimates to ensure a global maximum.

4.2.3. How to use the variables?

The matrix below shows how to determine how to use the variables in the analysis. An example for the mixture cure model is shown below:

<table>
<thead>
<tr>
<th>Variable name</th>
<th>Categorical</th>
<th>Stratum</th>
<th>$\mu$</th>
<th>$\sigma$</th>
<th>$c$</th>
<th>$\delta$</th>
<th>Reference value $x_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Race recode</td>
<td>X</td>
<td>X</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
</tr>
<tr>
<td>SEER historical stage</td>
<td>X</td>
<td>$\emptyset$</td>
<td>X</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
<td></td>
</tr>
<tr>
<td>Year of diagnosis</td>
<td>$\emptyset$</td>
<td>X</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
<td></td>
</tr>
</tbody>
</table>

Here, X means the option is checked and $\emptyset$ means the option is not applicable. The first column lists all the variables in the data. The 'X' in the second column indicates the variables are used as
categorical variables. The ‘X’ in the third column indicates the variables are used as stratum variables. Only a categorical variable can be used as a stratum. The columns corresponding to \( \mu_i, \sigma_i, \xi_i \) indicate whether to include the variables as covariates in the analysis. A stratum variable cannot be used as a covariate in \( \mu_i, \sigma_i, \xi_i \).

4.2.4. Variable transformation

Three variable transformations are available: logarithm \( \log(x - x_0) \), square of deviation \( (x - x_0)^2 \) and truncation \( (x - a)^+ \), where \( x_0 \) is the reference value and \( a \) is specified by the users. The truncation \( (x - a)^+ = x - a \) if \( x > a \) and 0 otherwise. This transformation is usually used when \( a \) is a change-point. Usually, the change-point \( a \) is pre-specified or estimated from the JOINPOINT software (see the website: http://srab.cancer.gov/joinpoint/index.html).

4.3. Output specifications

The output (Fig. 3) includes the main report on the parameter estimates and the graphics for model diagnosis. The graphic options are available only for the grouped survival data.

4.3.1. The main report

The main report consists of the model specifications, parameter estimates, their standard errors and the Wald chi-square tests.

4.3.2. Plot of observed and estimated survivals: \( S^{\text{obs}}(t(x_i)), S^{\text{est}}(t(x_i)) \) versus \( t \)

For each cohort defined by a vector of covariates \( x_i \), the observed survival curves \( S^{\text{obs}}(t(x_i)) \) is the actuarial estimate of the survival data. The estimated survival curve is \( S^{\text{est}}(t(x_i)) = \left[ \hat{\ell} + (1 - \hat{\ell}) G(t(\hat{\mu_i}, \hat{\sigma_i})) \right]^X \).

4.3.3. Plot of deviance residuals

The deviance residual for stratum with covariate \( x_i \) and interval \( j \) is defined as [12]

\[
\hat{r}_{ij} = \text{Sign}(s_{ij} - \hat{s}_{ij}) \times \sqrt{2 \left[ s_{ij} \log \left( \frac{s_{ij}}{\hat{s}_{ij}} \right) + \left( \hat{n}_j - s_{ij} \right) \log \left( \frac{\hat{n}_j - s_{ij}}{\hat{n}_j - \hat{s}_{ij}} \right) \right]^{1/2}},
\]

where \( s_{ij} \) is the observed number of survivors and \( \hat{n}_j \) is the adjusted number of people at risk for the \( j \)th interval and cohort \( x_i \). The predicted number of survivors is \( \hat{s}_{ij} = \hat{n}_j \hat{P}_{ij} \), where \( \hat{P}_{ij} = \frac{\hat{S}(t_j | x_i)}{\hat{S}(t_{j-1} | x_i)} \) for the relative survival and \( \hat{P}_{ij} = \frac{\hat{R}(t_j | x_i)}{\hat{R}(t_{j-1} | x_i)} \) for the cause-specific survival. The deviance residual plot is the plot of \( \hat{r}_{ij} \) versus \( t_j \). This plot can also be used to examine the goodness of the model fit. Usually, the residuals from a good model fit should be within ±3 from the 0 line.

4.3.4. Plot of \( k \)-year survival

This option is available only for the grouped survival data and at least one variable \( x \) should be in the
data. The k-year survival plot shows the plots of the observed and the estimated k-year survival probabilities, \((S^\text{obs}(t_k)), S^\text{est}(t_k))\), with respect to one of the covariate \(x\). For example, when \(x\) is the diagnosis year and \(k = 5\), this plot gives the observed and estimated 5-year survival rates by diagnosis year.

4.3.5. Profile loglikelihood plot: \(\log L(c)\) versus \(c\)

This option is available when the model is cure model and no covariate is used in the cure parameter \(c\). The profile loglikelihood plot of \(\log L(c)\) versus \(c\) is a visual tool to examine the curvature of the loglikelihood function. The profile loglikelihood \(\log L(c)\) is defined as

\[
\log L(c) = \max \log L(c, \beta_{\text{PAR}}, \beta_{\text{SCE}}, \beta_{\text{AR}})
\]

for fixed value of \(c\), where \((\beta_{\text{PAR}}, \beta_{\text{SCE}}, \beta_{\text{AR}})\) are the MLE’s of the parameters \((\beta_{\text{PAR}}, \beta_{\text{SCE}}, \beta_{\text{AR}})\) given fixed value of \(c\). If the curve \(\log L(c)\) has single mode, the cure fraction estimate \(\hat{c}\) is unique and stable. If the curve is flat, a range of \(c\) values can produce similar loglikelihood values and the cure fraction estimate \(\hat{c}\) has a large variance.

5. Applications

To demonstrate the use of the CANSURV program, we use the grouped colorectal cancer survival data from SEER-9 registries. The patients were followed-up until December 31, 2000, in order to ascertain the date of death. The survival time is the time between diagnosis and death; survival times are grouped into annual intervals. The maximum follow-up time is 20 years, so the interval takes a value from \(\{1, \ldots, 20\}\). When a patient is still alive at the end of the study or at the time the patient is lost to follow-up, then his or her survival time is censored.

5.1. Colorectal cancer survival by cancer stage

In order to look at survival by stage over a long period of time without changes in how stage is categorized, SEER uses a staging scheme, “SEER historic stage”, which is different from clinical practice. SEER historic stage categorizes stage at diagnosis into localized, regional, distant and unstaged based on the extent of the cancer at the time of diagnosis. First, the data is stratified into localized, regional and distant stages. We fit the standard survival models and cure models to the data, interpret the results and give cautionary notes. The standard survival models include the semi-parametric Cox proportional hazards model and Weibull model, and the cure model is the Weibull mixture cure model.

The parameter estimates from the standard survival models are listed in Table 1. All the covariates are significant \((p < 0.001)\) because of the large sample. The relative risk (hazard ratio) of dying of cancer is \(\exp(-\hat{\beta}_\text{PAR})\) for the Cox model and \(\exp(-\hat{\beta}_\text{SCE}/\exp(\hat{\beta}_\text{PAR}))\) for the standard Weibull model, respectively. For example, the risk of cancer death for the regional cancer relative to the risk for the localized cancer is \(\exp(1.281) = 3.6\) from the Cox model and \(\exp (-1.281) = 3.6\) from the Weibull model. The median survival time for the Weibull model is \(\exp(x)(\log 2)^{-1}\), which is 73.8, 10.4 and 0.6 for localized, regional and distant cancer, respectively.

Table 2 shows the parameter estimates from the Weibull mixture cure model. As the historic stage is an important prognostic factor, it is included in both \(c\) and \(\mu\). The second and third columns are the estimates and standard errors from the CANSURV output. In the last column, the cure rates and median survival times for uncured patients for localized, regional and distant cancer patients are calculated from the parameter estimates. Fig. 4 shows the estimated and observed relative survival curves from the Cox model, Weibull model and Weibull mixture cure model.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Cox model</th>
<th>Weibull model</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\hat{\beta})</td>
<td>S.E. ((\hat{\beta}))</td>
<td>S.E. ((\hat{\beta}))</td>
</tr>
<tr>
<td>(\hat{\beta}_{\text{PAR}})</td>
<td>4.889</td>
<td>0.026</td>
</tr>
<tr>
<td>SEER historic stage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Localized</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regional (\hat{\beta}_{\text{SCE}})</td>
<td>-1.281</td>
<td>0.016</td>
</tr>
<tr>
<td>Distant (\hat{\beta}_{\text{SCE}})</td>
<td>-3.046</td>
<td>0.015</td>
</tr>
<tr>
<td>(\hat{\beta}_{\text{PAR}})</td>
<td>0.471</td>
<td>0.004</td>
</tr>
</tbody>
</table>
observed and estimated survival curves from the Cox model, standard Weibull model and Weibull mixture cure models. The Cox model is semi-parametric and fits the observed survival curves well. The standard Weibull model assumes no cure, hence the estimated survival decreases until zero. The survival curves from the Weibull cure model level off after a certain number of years. The fit from the Weibull cure model is much better than the standard Weibull model.

5.2. Localized colorectal cancer survival by race and diagnosis year

The localized colorectal survival data are further stratified by race (white/black) and diagnosis year (1975–1995). The Weibull mixture cure model is fitted. Both race and diagnosis year are put in the parameters $c$ and $\mu$, the estimates and standard errors of the parameters are shown in Table 3. The black patients has significantly lower cure rate than the white patients ($p=0.013$) and the black patients who are not cured survive shorter than the white patients ($p=0.039$). The diagnosis year does not affect the cure rate ($p=0.269$), but significantly increases the survival time for the uncured patients ($p=0.000$). The estimated and observed $k$-year survival probabilities ($k=1, 5, 10$) by diagnosis year are plotted for both races (Fig. 5). The survival probabilities increase with the diagnosis year, but the variation for the black patients is larger.

---

**Table 2** The estimates and standard errors of the parameters of the Weibull mixture cure model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$\hat{\beta}$</th>
<th>S.E. ($\hat{\beta}$)</th>
<th>Cure rate (%)</th>
<th>Median survival time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cure ($c$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept $\beta_0$</td>
<td>1.215</td>
<td>0.034</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEER historic stage</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Localized</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regional $\beta_1$</td>
<td>$-1.350$</td>
<td>0.035</td>
<td>68.0</td>
<td></td>
</tr>
<tr>
<td>Distant $\beta_2$</td>
<td>$-4.002$</td>
<td>0.040</td>
<td>50.3</td>
<td></td>
</tr>
<tr>
<td>$\mu$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept $\beta_{\mu}$</td>
<td>1.775</td>
<td>0.040</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEER historic stage</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Localized</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regional $\beta_{\mu 1}$</td>
<td>$-0.482$</td>
<td>0.041</td>
<td>4.0</td>
<td></td>
</tr>
<tr>
<td>Distant $\beta_{\mu 2}$</td>
<td>$-1.779$</td>
<td>0.041</td>
<td>2.5</td>
<td></td>
</tr>
<tr>
<td>$\sigma$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept $\beta_{\sigma}$</td>
<td>0.037</td>
<td>0.006</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Fig. 5 Observed and estimated $k$-year survival probabilities by diagnosis year: (a) white and (b) black.
Table 3 The estimates and standard errors of the parameters of the Weibull mixture cure model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$\hat{\beta}$</th>
<th>S.E. ($\hat{\beta}$)</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cure $c$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>0.797</td>
<td>0.080</td>
<td>0.000</td>
</tr>
<tr>
<td>Race (black)</td>
<td>−0.348</td>
<td>0.140</td>
<td>0.013</td>
</tr>
<tr>
<td>Year of diagnosis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(75—95)</td>
<td>0.011</td>
<td>0.010</td>
<td>0.269</td>
</tr>
<tr>
<td>$\mu$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>1.829</td>
<td>0.112</td>
<td>0.000</td>
</tr>
<tr>
<td>Race (black)</td>
<td>−0.348</td>
<td>0.169</td>
<td>0.039</td>
</tr>
<tr>
<td>Year of diagnosis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(75—95)</td>
<td>0.054</td>
<td>0.012</td>
<td>0.000</td>
</tr>
<tr>
<td>$\delta$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>0.228</td>
<td>0.028</td>
<td>0.000</td>
</tr>
</tbody>
</table>

6. Discussions

We described the chief features of the CANSURV software, which is designed to analyze the population-based cancer survival data. The CANSURV software is quite versatile as the survival data that can be either grouped or individual; the net survival can be either cause-specific or relative survival. The standard survival models include parametric models and semi-parametric Cox model. The cure models include the mixture cure model and the proportional hazards model with cure component. Although the cause-specific survival data can be analyzed by standard packages, such as SAS or Splus, no software is available for the analysis of relative survival data. The CANSURV is primarily developed for analyzing the population-based grouped survival data from SEER program. However, it can be used for analyzing the individual survival data as well.

Also, CANSURV has other potential applications. The estimated survival models from CANSURV can be used in estimating cancer prevalence [11]. Thus, the CANSURV software provides a unified and comprehensive tool for analyzing any survival data using the standard survival models or the mixture survival models.

Acknowledgement

The authors would like to thank Dr. John Gamel whose work on methods for the estimation of cure model parameters and associated software provided a basis for the development of CANSURV.

References