Special Proceedings: ISBN #: 978-81-950383-5-0
26th Annual Conference, 26-28 February 2024; pp 111-126

Issues in Estimating Disease Specific Incidence Rates in Long-Term Follow-up in Childhood Cancer Survivors

Deo Kumar Srivastava[1], Kirsten Ness[2], Melissa Hudson[2,3], Sarmistha Das[4,5,6] and Shesh N. Rai[4,5,6]

[1]*Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, Tennessee, 38105, USA*

[2]*Department of Epidemiology & Cancer Control, St. Jude Children's Research Hospital, Memphis, Tennessee, 38105, USA*

[3]*Department of Oncology, St. Jude Children's Research Hospital, Memphis, Tennessee, 38105, USA*

[4]*Department of Biostatistics, Health Informatics and Data Science, University of Cincinnati, Cincinnati, Ohio, 45267, USA*

[5]*Cancer Data Science Center, University of Cincinnati College of Medicine, Cincinnati, Ohio, 45267, USA*

[6]*Biostatistics and Informatics Shared Resource, University of Cincinnati Cancer Center, Cincinnati, Ohio, 45267, USA*

Received: 18 June 2024; Revised: 10 July 2024; Accepted: 12 July 2024

Abstract

With significant advances in the treatment and supportive care, the overall 5-year survival rate for pediatric cancers in high-income countries, such as USA, exceeds 85%, see Ehrhardt, *et al.* (2023), SEER (2022) and it is expected that the number of survivors will exceed 580,000 by year 2040. However, this comes at a high cost of treatment and cancer related long-term sequalae. To better characterize and to develop interventions/screening guidelines to mitigate the long-term effects of these adverse events researchers in North America and Europe established large-cohort retrospective studies with prospective follow-up assessments, see Robison, *et al.* (2009), Winther, *et al.* (2015), Park, *et al.* (2012). However, it is logistically impossible to follow the survivors continuously and this information is usually collected through cross-sectional surveys at various times from cancer diagnosis, which leads to interval censored data since the exact time of the onset of the adverse event of interest is unknown. However, if this risk could be characterized in a continuous manner, then appropriate screening guidelines or interventions could be implemented. Our primary focus is on estimating the incidence rates (cumulative incidence) of a particular outcome of interest *e.g.* cardiovascular events using interval censored data. In this exposition we utilize SJLIFE cohort and propose the use of multi-state survival framework for modeling incidence rates and risk factors associated with it. We also highlight the use of multi-state models for analyzing more complicated relationships and identify some challenges associated with the analysis of such data.

Corresponding Author: Deo Kumar Srivastava
Email: kumar.srivastava@stjude.org

Key words: Cross-sectional survey data; Prospective follow-up; Multi-state model, Interval-censored data.

1. Introduction

The 5-year survival rate for childhood cancer survivors now exceeds 85% but they are at an increased risk of developing long-term chronic health conditions as a result of their cancer or its treatment. As the childhood cancer survivor population increases, understanding the long-term impact of cancer on health during adulthood is important to guide the development of interventions to improve the quality of life and duration of survival. The Cancer Control and Survivorship Program in the Comprehensive Cancer Center at St. Jude Children's Research Hospital is a multidisciplinary research program that strives to improve the quality of life of individuals surviving childhood cancer by translating the research findings into effective strategies to reduce treatment related complications. The St. Jude Lifetime Cohort Study (SJLIFE) is a cancer epidemiologic cohort established to facilitate longitudinal clinical evaluation of health outcomes in childhood cancer survivors across the lifespan. A detailed description of the study along with schema for longitudinally follow-up can be obtained by visiting the St. Jude Cloud portal (https://www.stjude.cloud/) and going through the Cancer Survivorship tab and clicking on St. Jude LIFE study (SJLIFE), Howell, *et al.* (2021). SJLIFE was activated in 2007 with initial eligibility for participation including 10+ year survivor of pediatric cancer, treated, or followed at St. Jude from 1962 to 2012 who were at least 18 years of age. Eligibility criteria were modified in 2015 to include 5-year survivors. Eligible survivors are periodically invited to return to St. Jude for comprehensive clinical evaluations that involves completion of questionnaire with patient-reported outcomes, collection of biological specimens, and systematic evaluation of organ function including metabolic, cognitive and neuromuscular status.

Howell, *et al.* (2021) reported an update of the cohort progress. Among 8192 eligible survivors, 6560 have agreed to participate and 5,223 have completed baseline on-campus evaluations. The median [range] age at evaluation was 32 [7.0 – 71.9]. Participants are invited to return for follow-up visits in 3–5 years intervals. Study findings from these evaluations have enabled characterization of multimorbidity experiences by survivors many years after treatment for childhood cancer. This is highlighted in the study by Bhakta, *et al.* (2017) that used St. Jude modified National Cancer Institute's Common Terminology Criteria for Adverse Events (CTCAE) 4.03, Hudson, *et al.* (2017), and graded 168 chronic conditions within 13 organ systems. The CTCAE grades correspond to grade 1 (mild), grade 2 (moderate), grade 3 (severe/disabling), grade 4 (life-threatening) and grade 5 (death). Often, the focus is on modeling grade 3 or higher chronic conditions. The details of all the chronic conditions within each organ system can be obtained from Bhakta, *et al.* (2017), Supplementary Table $S1$. They grouped the chronic conditions into 13 organ systems as shown in Table 1.

In addition to studying different outcomes the identification of the appropriate study cohort at risk for developing the outcomes of interest is equally important. There are multiple factors that need to be considered in the selection of eligible subjects (survivors) in the study cohort. Because this cohort was originally constructed retrospectively, cohort entry and exit

Table 1: Number of chronic conditions within each grouped category for each organ system

Organ System	Grouped Condition Category
Cardiovascular	Myocardial Infarction (1), Arrhythmias (6), Cardiovascular Dysfunctions (4), Structural Heart Defects (4), Vascular Diseases (4), Essential Hypertension/Dyslipidemia (3)
Respiratory	Asthama (1), Obstructive Respiratory Disorders (2), Functional Pulmonary Deficits (3), Respiratory Parenchymal Diseases (7)
Gastrointestinal	Esophageal Disorders (3), Disorders of the GI Tract (11), Inflammatory Disorders (8), Hepatic Disorders (6), Disorder of the Gallbladder (1)
Reproductive	Disorders of the Female Reproductive System (8), Disorders of the Male Reproductive System (5), Condition affecting the Pituitary (1)
Endocrine	Growth Hormone Deficiencies (2), Overweight/Underweight (2), Thyroid Disorders (4), Parathyroid Disorders (2), Abnormal Glucose Metabolism (1), Conditions affecting the Pituitary (3)
Renal	Kidney Injuries (2), Obstructive Urinary Disorders (3), Hematuria (1)
Musculoskeletal	Amputation (1), Osteoporosis (1), Joint Diseases (3), Peripheral Musculoskeletal Disorders (6), Spine Disorders (4)
Neurology	Strokes (3), Central Nervous System Disorders (9), Mixed Nervous System Disorders (4), Peripheral Nervous System Disorders (6), Seizure (1), Severe Headache (1)
Immunology and Infections	Immunologic Disorders (2), Frequent/Recurrent Infections (8), Chronic Infections (7)
Hematology	Hematologic Disorders (7)
Auditory	Hearing Loss (1)
Second Neoplasms	Secondary and Recurrent Malignancy (1)
Ocular	Ocular Disorders (4)

Note: The numbers in the brackets indicate the number of chronic conditions within each grouped category within an organ system.

are heterogenous, *i.e.* their follow-up times are not equally spaced, and subjects may enter the cohort or leave the cohort at any follow-up times, including baseline (T_0), longitudinal follow-up (T_L) and date of death (T_D) *etc.* These are outlined in Table 2.

A thorough understanding of these chronic conditions, their prevalence and associated risk factors can provide valuable information which could be used to improve future treatment plans. For our discussions, we will focus on Cardiovascular Dysfunction (CD) within the Cardiovascular System (22 individual chronic conditions), which has four individual chronic conditions (cardiomyopathy (CAD), Right ventricular systolic dysfunction (RVSD), Cor Pulmonale (CP) and Pulmonary Hypertension (PH). The discussion can be easily generalized to all chronic conditions within cardiovascular system or across other organ systems as well. CAD refers to problems with heart muscles that make it harder for the heart to pump blood and, if untreated, can lead to heart failure or cardiac arrest. Similarly, RVSD if untreated could lead to heart failure or myocardial infarction, CP is an alteration in the structure and function of the right ventricle of the heart caused by a primary disorder

Table 2: Issues of heterogeneity in identifying the study cohort

Time points	Challenges
T_0	1. Should $T_0 = 5$ years since the primary diagnosis (eligibility criterion). What is the rationale for 5 years, why not 2, 3, 4, 6, 7, 8, 9, 10 years? 2. Should T_0 depend on multiple factors, including patient's age, time since treatment, disease type *etc.*?
T_L	1. If we need to choose only on long-term follow-up time to collect information, should T_L be largest or shortest or somewhere in between, when patients have more than one follow-up visits.
T_D	1. Eligibility restricted to 5-year survivors potentially introduces survival bias 2. Not every eligible patient in the the SJLIFE cohort visits clinics. They may be lost to follow-up (true censoring) or died but this information may not be accurately recoded. Should T_D and patient characteristics, including comorbidity at the time of death be obtained from other sources (such as death registry).

of the respiratory system, and PH is a condition that affects blood vessels in the lungs and makes heart work harder than normal to pump blood into lungs.

Among childhood cancer survivors, cardiovascular events (CEs) are among the top nonmalignant causes of death (Armstrong, *et al.* (2009)). This is due to the damage to cardiomyocytes caused by chemotherapy and chest radiation therapy received during the cancer treatment (Hammoud, *et al.* 2024). Even though certain chemotherapy exposures such as anthracycline are well known for associations with cardiotoxicity (Ehrhardt, *et al.* 2023), they continue to be used to treat cancer because of their curative benefits. Improved characterization of the cumulative incidence of CEs may facilitate opportunities for intervention to improve/preserve cardiac health. This motivates us to estimate the cumulative incidence (CI) of the CEs in childhood cancer survivors who completed their baseline evaluation because such information could be used to help researchers identify the best time to intervene. More information regarding the causes, treatment, and prevention of cardiotoxicity can be found in a comprehensive review by Koutsoukis, *et al.* (2018).

For our discussion we will focus on CAD (a CD), whose exact timing of development is unknown since SJLIFE participants are not followed continuously in real time. However, health related information is ascertained when participants came to campus for their baseline and follow-up visits, either on their scheduled visits or in an ad hoc manner to participate in ancillary studies. Although the exact time of CAD is unknown, the current status is available and from there we know that CAD symptoms that motivated medical attention occurred sometimes in the interval between the two follow-up visits. Such dataset can be characterized as case I interval-censored data, see Sun (2006), Rai (2008).

Remark: It may be noted that the approaches presented in this article would be applicable to the phase IV clinical trials in the context of drug development where the focus would be on monitoring for long term toxicities/side effects of an approved drug, see Zhang, *et al.* (2016).

We propose the use of multi-state models for estimating the cumulative incidence

rate for the event of interest. Multi-state models are extensions of the survival models, which are usually analyzed under the Cox proportional hazard model assumption. However, multi-state models have the advantage of providing more insight into disease process and progression as each transition, from one state to another, can be modeled and the covariates could be incorporated *e.g.* using Cox proportional hazards model. It further allows for simultaneous modeling of competing causes of death or morbidities; see Eulenburg, *et al.* (2015). The simplest extension of the survival model is the continuous time progressive three-state model *e.g.* see van den Hout (2017, Chapter 3). It is assumed that the survivors are followed longitudinally and the status of the survivors alive at the pre-specified observation time is available and that follow-up time for survivors on study can vary. In a progressive illness/death model as illustrated in Figure 1. For example, in the context of the study discussed above all survivors who survived for at least 5 years from their date of diagnosis will be assumed to be in State 1 (all are alive with no CAD) and then after a median follow-up of about 25 years were enrolled in SJLIFE and systematically evaluated for CAD. At that point each survivor could take one of the three paths: 1) remain alive with no CAD (State 1); 2) develop CAD but remain alive (moving from State 1 to State 2); 3) progress to death due to CAD or otherwise (moving from State 1 to State 3). For patients that reached State 2, they could also have two options: 1) remain alive with CAD (State 2); or 2) progress towards death or cardiac failure (moving from State 2 to State 3).

$\lambda_1(u)$ represents the transition intensity from State 1 to State 2; $\lambda_2(t)$ is the transition intensity from State 1 to State 3; $\lambda_3(t|u)$ is the transition intensity from State 2 to State 3. In general, the focus could be on estimating $\lambda_1(u)$, the transition intensity rate for patients

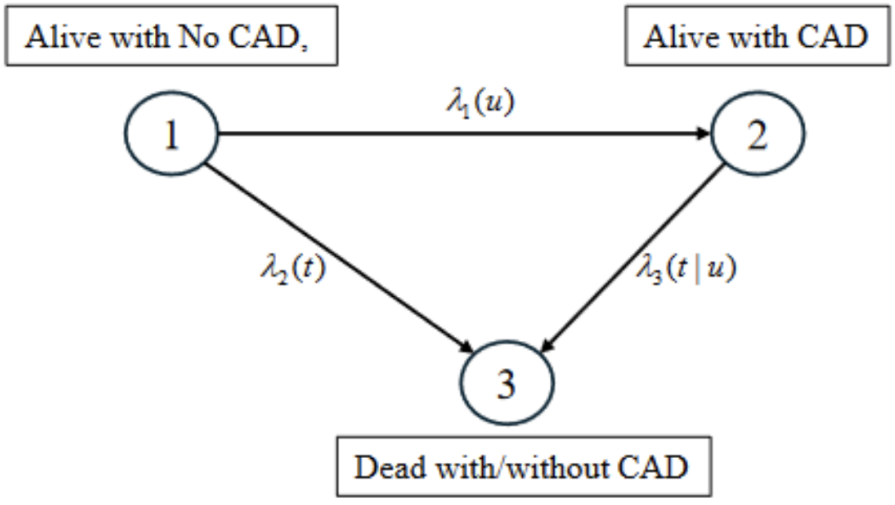

Figure 1: Three-states illness/death model of patients developing CAD

progressing from the initial state (State 1) of being normal to developing CAD (State 2) but one could be interested in estimating the incidence rates $\lambda_2(t)$ or $\lambda_3(t|u)$. This can be done using a parametric or a semi-parametric framework. In section 2 and 3 we will outline the parametric and semi-parametric approaches, respectively. In section 4, we outline the complexities involved when the interest could be in estimating the incidence rates of two or more events of interest.

2. Parametric modeling

Let $\{X(t)\}$ represent the stochastic process which identifies the state occupied a survivor at time t if we start with n survivors in the cohort then at time $t = 0$ all survivors will be in state 1 and has not experienced CAD (the event of interest). Let T be a random variable denoting the observation time (death, CAD or no event observed in the study duration) and U denotes the time to CAD. Thus, $X(t) = 1$, $X(t) = 2$ and $X(t) = 3$ represent the current status of the survivor; alive without CAD, alive with CAD or dead with or without CAD, respectively. We assume that the occurrence of CAD is irreversible. With the intensity functions shown in Figure 1 one can easily obtain the pseudo-survival functions, see Rai, *et al.* (2013) and van den Hout (2017), as follows:

$$Q_i(t) = exp\{-\int_t^0 \lambda_i(\nu)d\nu\}, \text{for } i = 1, 2 \tag{1}$$

and

$$Q_3(t|u) = exp\{-\int_t^u \lambda_3(\nu|u)d\nu\} \tag{2}$$

It may be noted that probability of surviving without experiencing CAD or death beyond time t can be represented as,

$$Q(t) = exp\{-\int_0^t (\lambda_1(\nu) + \lambda_2(\nu))d\nu\} = Q_1(t)Q_2(t), \tag{3}$$

and the survival function can be obtained as,

$$\begin{aligned} S(t) &= P(X(t) = 1) + P(X(t) = 2) \\ &= Q(t) + \int_0^t \lambda_1(u)Q(u)Q_3(t|u)du \end{aligned} \tag{4}$$

2.1. Construction of the likelihood

The likelihood for the three-state model can be constructed in the following manner. Let θ denote the vector of parameters including transition intensities. Let t be the realization of the *r.v.* T and let Δ_i denote the contribution to the likelihood for the i^{th} survivor for $i = 1, 2, \ldots, n$. Then the likelihood function $L(\theta) = \prod_{i=1}^n \Delta_i$ Within this framework the survivor will be in one of the four distinct types of observations and their contribution to the likelihood will be as follows:

(i) Death without CAD, $T = t$, $X(t^-) = 1$, and $L_1(t) = \lambda_2(t)Q(t)$,

(ii) Alive without CAD, $T > t$, $X(t) = 1$, and $L_2(t) = Q(t)$,

(iii) Death with CAD, $T = t$, $X(t^-) = 2$, and $L_3(t) = \int_0^t \lambda_1(u)Q(u)\lambda_3(t|u)Q_3(t|u)du$,

(iv) Alive with CAD, $T > t$, $X(t) = 2$, and $L_4(t) = \int_0^t \lambda_1(u)Q(u)Q_3(t|u)du$

The likelihood function depends on in addition to the observation time and status, but is suppressed for convenience.

Parametric modeling is appealing as one can easily obtain the estimates and perform the inference using likelihood approaches. Among the class of parametric distributions, *e.g.* see Srivastava, *et al.* (2018), ven den Hout (2017), commonly used distributions are exponential, piecewise exponential and Weibull but other distributions such as log-normal, Gamma, log-logistic or Gompertz distributions could also be used. The data observed for each survivor i, $i = 1, 2, \ldots, n$, at a particular time, consists of triplet $(t_i, \delta_i, \gamma_i)$ where t_i is the observation time and,

$$\delta_i = \begin{cases} 1, & \text{if } i^{th} \text{ survivor died at time } t_i \\ 0, & \text{if } i^{th} \text{ survivor was alive without CAD at time } t_i \end{cases}$$

and

$$\gamma_i = \begin{cases} 1, & \text{if } i^{th} \text{ survivor had CAD at time } t_i \\ 0, & \text{if } i^{th} \text{ survivor did not have CAD at time } t_i \end{cases}$$

Then, the log-likelihood function can be written as,

$$l(\theta) = \sum_{i=1}^{n} [a_i log L_1(t_i) + b_i log L_2(t_i) + c_i log L_3(t_i) + d_i log L_4(t_i)], \tag{5}$$

where, $a_i = \delta_i(1 - \gamma_i)$, $b_i = (1 - \delta_i)(1 - \gamma_i)$, $c_i = \delta_i\gamma_i$, and $d_i = (1 - \delta_i)\gamma_i$.

The contributions that a survivor makes to the likelihood depends on the status and the underlying distribution. It may be worth noting that often the interest would be in estimating the cumulative incidence of the event of interest (CAD) at particular time points, *e.g.* 5-year of 10-year. Exponential distribution is the simplest model because it assumes the intensity function to be constant across time. However, this is not a plausible assumption when the follow-up time is long and there is the possibility of the intensity function changing over time. To provide more flexibility in modeling such data Rai, *et al.* (2013) proposed to use piecewise exponential distribution in estimating $\lambda_1(u)$. However, this poses the problem of knowing exactly when the incidence rate changes, how many change points are needed and that the incidence rates are constant within each piece. Srivastava, *et al.* (2018) used Weibull distribution to circumvent these limitations. Pradhan and Kundu (2014) also used Weibull distribution as the underlying lifetime distribution for the interval-censored data but suggested using the EM algorithm approach.

For exponential distribution the intensity rates are constant, *i.e.* $\lambda_i(t) = \lambda_i$ for $i = 1, 2$ and $\lambda_3(t|u) = \lambda_3$. Then, the contributions to the likelihood will be as follows:

$$L_1(t) = \lambda_2 \exp(-(\lambda_1 + \lambda_2)t) \tag{6}$$

$$L_2(t) = \exp(-(\lambda_1 + \lambda_2)t) \tag{7}$$

$$L_3(t) = \frac{\lambda_1\lambda_3}{\lambda_1 + \lambda_2 - \lambda_3}(\exp(-\lambda_3 t) - \exp(-(\lambda_1 + \lambda_2)t)) \tag{8}$$

$$L_4(t) = \frac{\lambda_1}{\lambda_1 + \lambda_2 - \lambda_3}(\exp(-\lambda_3 t) - \exp(-(\lambda_1 + \lambda_2)t)) \tag{9}$$

Now if one assumes that the incidence rate for CAD may change say at t_c years (say,

$t_c = 5$) then one could use piecewise exponential distribution that would imply that $\lambda_1 = \lambda_{11}$ if $t < t_c$ and λ_{12} if $t \geq t_c$. The piecewise exponential assumption could be extend to other pieces of multi-state model if there is evidence of shift in incidence rates over time. The contributions to the likelihood will be as follows:

$$
\begin{aligned}
&L_1(t) = \lambda_2 \exp(-(\lambda_{11} - \lambda_{12})t_c - \lambda_{12}t) \exp(-\lambda_2 t) \\
&L_2(t) = \exp(-(\lambda_{11} - \lambda_{12})t_c - \lambda_{12}t) \exp(-\lambda_2 t) \\
&L_3(t) \\
&= \frac{\lambda_{11}\lambda_3}{\lambda_{11} + \lambda_2 - \lambda_3}(\exp(-\lambda_3 t) - \exp(-(\lambda_{11} + \lambda_2 + \lambda_3)t)) \text{ if } t < t_c \\
&= \frac{\lambda_{11}\lambda_3}{\lambda_{11} + \lambda_2 - \lambda_3} \exp(-\lambda_3 t)(1 - \exp(-(\lambda_{11} + \lambda_2 + \lambda_3)t_c) + \frac{\lambda_{12}\lambda_3}{\lambda_{12} + \lambda_2 - \lambda_3} \\
&\exp(-(\lambda_{11} - \lambda_{12})t_c - \lambda_3 t)(\exp(-(\lambda_{12+\lambda_2-\lambda_3})t_c)) - \exp(-(\lambda_{12} + \lambda_2 - \lambda_3)t) \text{ if } t \geq t_c \\
&L_4(t) = L_3(t)/\lambda_3
\end{aligned}
$$

For Weibull distribution the intensity function can be defines by $\lambda_i(t) = \eta_i \omega_i t^{\omega_i - 1}$, for $i = 1, 2$ and $\lambda_3(t|u) = \lambda_3^{SM}(t|u) = \eta_3\omega_3(t)^{\omega_3 - 1}$ under the assumption of a semi-Markov process or $\lambda_3(t|u) = \lambda_3^{M}(t|u) = \eta_3\omega_3(t-u)^{\omega_3 - 1}$ under the assumption of a Markov process. This leads to $Q_3^{SM}(t|u) = \exp(-\eta_3(t^{\omega_3} - u^{\omega_3}))$ under semi-Markov assumption and $Q_3^{M}(t|u) = \exp(-\eta_3(t-u)^{\omega_3})$ under Markov assumption, see Kalbfleisch and Lawless (1985) and Hazerlak, *et al.* (2003). The contribution to the likelihood is provided below:

$$
\begin{aligned}
&L_1(t) = \eta_2\omega_2 t^{\omega_2 - 1} \exp(-\eta_1 t^{\omega_1} - \eta_2 t^{\omega_2}) \\
&L_2(t) = \exp(-\eta_1 t^{\omega_1} - \eta_2 t^{\omega_2}) \\
&L_3(t) = \int_0^t Q(u)\lambda_1(u) Q_3^{SM}(t|u)\lambda_3^{SM}(t|u)du \quad \text{under semi-Markov assumption} \\
&= \int_0^t Q(u)\lambda_1(u) Q_3^{M}(t|u)\lambda_3^{M}(t|u)du \quad \text{under Markov assumption} \\
&L_4(t) = \int_0^t Q(u)\lambda_1(u) Q_3^{SM}(t|u)du \quad \text{under semi-Markov assumption} \\
&= \int_0^t Q(u)\lambda_1(u) Q_3^{M}(t|u)du \quad \text{under Markov assumption}
\end{aligned}
$$

Now, using the above contributions to the likelihood one can perform the likelihood estimation and obtain confidence intervals.

2.2. Incorporating covariates

As we have noted before, one of the long-term consequences of cancer patients treated with cardiotoxic therapy (treated with anthracycline and/or chest radiation) is that they are at a very high risk of developing CAD. Let us denote this group as AR (At Risk group), and let NR represent the groups of survivors who were not treated with cardiotoxic therapy.

Within this context, it would be important not only to know the onset time of these CADs but it would be important to know, from a clinician's perspective, the risk associated with therapeutic exposure and other risk factors such as age at diagnosis or sex *etc.* Thus, incorporation of covariates is an important issue. van den Hout (2017) proposes to use proportional hazard type to regression framework represented as

$$h(t|\boldsymbol{X}) = h_0(t)\exp(\boldsymbol{\beta'X}) \tag{10}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ is a vector of parameters and $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)$ is the vector of fixed covariates without an intercept term and $h_0(t)$ denotes the baseline hazard which could be modeled using any of the above mentioned parametric distributions. Alternatively, one can assess the impact of the covariates on the incidence rates by modeling the scale parameter as function of the parameters. For exponential distribution it would be equivalent to modeling $\lambda = \exp(\boldsymbol{\beta'X})$ and for Weibull distribution it would amount to modeling $\eta = \exp(\boldsymbol{\beta'X})$.

3. Semi-parametric modeling

Both Rai, *et al.* (2013) and Srivastava, *et al.* (2018) used a parametric approach with the assumption that the transition intensity rates from different risk groups are independent. Subjects in different groups are indeed independent, but the rates can be modeled assuming a dependence structure. In other words, the parametric approach may not be suitable if the underlying parametric assumptions do not hold. To overcome this limitation, we propose a semi-parametric approach with EM algorithm to model CI within the illness-death framework. Rai and Matthews (1997) introduced discrete scale models for estimating the transition intensity rate in a survival-sacrifice experiment using EM algorithm. Later, Rai, *et al.* (2000) extended a similar methodology to a mixed-scale model with an EM algorithm approach as well. EM algorithm is a powerful procedure to use when no closed-form solution can be obtained from the likelihood function. Besides the traditional EM algorithm, Rai and Matthews (1993) introduced a modified approach that could potentially save time (fewer iterations). A detailed application demonstration of EM algorithm can be found in Gunaratnam and Rai (2019). Additionally, when dealing with high dimensional parameters, the regular simple case likelihood function will not work. Rather, the use of profile likelihood should be considered, see Murphy and Van Der Vaart (2000).

Given that the data on CAD is collected intermittently the occurrence of the events falls in the category of interval-censored data as discussed above. To better understand the problem at hand, let us define Year 1 as 5 years after treatment completion and Year 2 as 6 years after treatment completion, a patient can choose to participate in the study either at Year 1 or Year 2, or even later. The challenge comes when we build the likelihood functions that will be discussed in the next section.

The following table reflects the data characteristics in a complete data setting. Table 3 corresponds to survivors who have come for SJLIFE evaluations for the first time and assume the maximum follow-up time is 5 years after cohort entry.

It is assumed that we are interested in looking at intensity rates up to five years. Let λ_i, $i = 1, 2, \ldots, 5$ be the intensity rates at Year i (Year 1, Year 2, *etc.*). Let n_{+i} $i = 1, 2, \ldots, 5$

Table 3: Data characteristics in a complete data setting

Year	n_{+1}	n_{+2}	n_{+3}	n_{+4}	n_{+5}	Rate
1	r_{11}	r_{12}	r_{13}	r_{14}	r_{15}	λ_1
2		r_{22}	r_{23}	r_{24}	r_{25}	λ_2
3			r_{33}	r_{34}	r_{35}	λ_3
4				r_{44}	r_{45}	λ_4
5					r_{55}	λ_5
	r_{+1}	r_{+2}	r_{+3}	r_{+4}	r_{+5}	

be the number of subjects that came to the clinic in year i. For example, n_{+1} represents the patients who came to the clinic with 1 year of follow-up, and n_{+2} represents the patients who came to the clinic with 2 years of follow-up. Within any year i let there be r_{+i}, $i = 1, 2, \ldots, 5$ events (survivors with CAD, abnormal). For example, when an abnormal survivor visited in Year 2, the survivor may have become abnormal either in Year 1 or in Year 2, but that information is unknown. Therefore, we define $r_{+2} = r_{12} + r_{22}$ as the total number of events in Year 2 in which r_{11} represents the events that occurred during Year 1 and r_{22} represents the events that occurred during Year 2.

In summary, estimating transition intensity rates in a three-states illness-death model, such as the SJLIFE study, is not a simple task. The existing approach such as constructing the likelihood function is very complicated since the function is not in a closed form. Furthermore, sometimes the exponential model and the Weibull model might not be appropriate as the underlying assumptions might not always hold. This motivates us to construct a non-parametric model. In combination with EM algorithm, the transition intensity rate can be easily obtained.

Likelihood based approach

Including all the characteristics of the data in the likelihood is somewhat challenging in our situation. However, here we present an approach that would be appropriate for the data that we have in hand. For simplicity purposes, we will only show time points up to three years ($M = 3$).

We define the table above as complete data since we specifically know which survivors are abnormal. For incomplete data, we define (n_{+k}, r_{+k}), $k = 1, 2, \ldots, M$ and $r_{+k} = \sum_{j=1}^{k} r_{jk}$, $k = 1, 2, \ldots, M$ represents the total number of survivors with abnormality with k years of follow-up and n_{+k} represents the survivors who come to the clinic with k years of follow-up. In our case, we can write it as $r_{+1} = r_{11}$, $r_{+2} = r_{12} + r_{22}$, $r_{+3} = r_{13} + r_{23} + r_{33}$. In the incomplete data setting, we only know r_{+k} but not the actual number of abnormal cases within each year of follow-up. Since we have two independent groups, AR (At Risk) and NR (Not at Risk), the likelihood function can be defined for each group independently, which have a similar form. Let

$$
\begin{aligned}
z_{+j} &= n_{+j} - r_{+j}, j = 1, 2, 3 \\
p_{+1} &= \lambda_1 \\
p_{+2} &= \lambda_1 + (1 - \lambda_1)\lambda_2 \\
p_{+3} &= \lambda_1 + (1 - \lambda_1)\lambda_2 + (1 - \lambda_1)(1 - \lambda_2)\lambda_3
\end{aligned}
$$

Thus, in general,

$$p_{+j} = \lambda_1 + \sum_{k=1}^{j-1} \Big[\prod_{i=1}^{k} (1-\lambda_i) \Big] \lambda_{k+1} \tag{11}$$

z_{+j} represents the number of subjects at risk, p_{+j} is the prevalence of events occurring in the j^{th} year. For incomplete data, the likelihood functions corresponding to the first three time points will be:

$$L_1^{IC} = p_{+1}^{r_{+1}}(1-p_{+1})^{z_{+1}} = \lambda_1^{r_{+1}}(1-\lambda_1)^{n_{+1}-r_{+1}}$$
$$L_2^{IC} = p_{+2}^{r_{+2}}(1-p_{+2})^{z_{+2}} = [\lambda_1 + (1-\lambda_1)\lambda_2]^{r_{+2}}[(1-\lambda_1)(1-\lambda_2)]^{n_{+2}-r_{+2}}$$
$$L_3^{IC} = p_{+3}^{r_{+3}}(1-p_{+3})^{z_{+3}} = [\lambda_1 + (1-\lambda_1)\lambda_2 + (1-\lambda_1)(1-\lambda_2)\lambda_3]^{r_{+3}}[(1-\lambda_1)(1-\lambda_2)(1-\lambda_3)]^{n_{+3}-r_{+3}}$$

The generalized form of the likelihood function at each time point can be written as:

$$L_j^{IC} = \Big[\lambda_1 + \sum_{k=1}^{j-1}\prod_{i=1}^{k}(1-\lambda_i)\lambda_{k+1}\Big]^{r_{+j}} \Big[\prod_{k=1}^{j}(1-\lambda_j)\Big]^{z_{+j}}, j = 1, 2, \ldots, M \tag{12}$$

For complete data, assuming the number of events to follow a multinomial distribution, the likelihood functions of the first three time points can be presented as:

$$L_1^C = \lambda_1^{r_{11}}(1-\lambda_1)^{z_{+1}}$$
$$L_2^C = \lambda_1^{r_{12}}[(1-\lambda_1)\lambda_2]^{r_{22}}[(1-\lambda_1)(1-\lambda_2)]^{z_{+2}}$$
$$L_3^C = \lambda_1^{r_{13}}[(1-\lambda_1)\lambda_2]^{r_{23}}[(1-\lambda_1)(1-\lambda_2)\lambda_3]^{r_{33}}[(1-\lambda_1)(1-\lambda_2)(1-\lambda_3)]^{z_{+3}}$$

The generalized form of the likelihood function at each time point can be written as:

$$L_j^C = \lambda_1^{r_{1j}}[(1-\lambda_1)\lambda_2]^{r_{2j}} \ldots [\prod_{k=1}^{j-1}(1-\lambda_k)\lambda_j]^{r_{jj}}[\prod_{k=1}^{j}(1-\lambda_k)]^{z_{+j}} \tag{13}$$

Now assuming the intensity rates to be λ_i and λ_i^*, $i = 1, 2, \ldots, M$, for the AR and NR groups respectively, one can obtain the complete likelihood functions, for details see Qian, *et al.* (2023). Qian, *et al.* (2023) used logit link, see Agresti (2013), Rai and Matthews (1997), to establish a relationship between and to provide for a parsimonious modeling of the data. Specifically, they assumed,

$$\lambda_k^* = \frac{e^\beta \lambda_k}{1 + (e^\beta - 1)\lambda_k}, \text{ and } (1-\lambda_k^*) = \frac{1-\lambda_k}{1+(e^\beta-1)\lambda_k} \tag{14}$$

It may be noted that within this framework other covariates of interest could be modeled by replacing $\boldsymbol{\beta}$ with $\boldsymbol{\beta'X}$ in the above equation. Now it is easy to see that r_{+j} is known in the incomplete data and follows a binomial distribution,

$$r_{+j} \sim B(n_{+j}, 1 - \prod_{k=1}^{j}(1-\lambda_k)), \;\; j = 1, 2, \ldots, M \tag{15}$$

and the conditional distribution of r_{kj} given r_{+j} will also follow a binomial distribution given

by,

$$r_{kj}|r_{+j} \sim B(r_{+j}, \frac{\prod_{i=1}^{k-1}(1-\lambda_i)\lambda_k}{1-\prod_{i=1}^{j}(1-\lambda_i)}), \;\; k = 1, 2, \ldots, j; j = 1, 2, \ldots, M \tag{16}$$

Qian, *et al.* (2023) use the fact the sufficient statistics corresponding to the complete log-likelihood $l^C(\lambda, \beta)$ are:

$$\sum_{j=k}^{M}(r_{kj} + r^*_{kj}), \;\; \text{and} \;\; \sum_{j=k}^{M} r^*_{kj} \tag{17}$$

and propose to estimate the parameters using EM algorithm. The basic steps are outlined below:

E-step: Start with the initial estimates of $\lambda = \lambda^{(0)}$ and $\beta = \beta^{(0)}$ then one can obtain the values of $r^{(1)}_{kj}$ for $k = 1, 2, \ldots, j$ and for all $j = 1, 2, \ldots, M$.

M-step: Then, using the $r^{(1)}_{kj}$'s and the initial value $\lambda^{(0)}$ and one can obtain the estimate of β, $\beta^{(1)}$, using profile likelihood. Then, using $r^{(1)}_{kj}$'s and $\beta^{(1)}$ obtain updated estimate $\lambda^{(1)}$ using the complete likelihood. This iterative process continues until the distance between ($\lambda_1^{(q)}$, $\lambda_2^{(q)}$, $\lambda_3^{(q)}$, $\beta^{(q)}$) and ($\lambda_1^{(q-1)}$, $\lambda_2^{(q-1)}$, $\lambda_3^{(q-1)}$, $\beta^{(q-1)}$) at the q^{th} iteration is smaller than a pre-specified constant C_0. Qian, *et al.* (2023) also performed simulation studies to show that the performance of the EM approach is reasonable.

4. Extension of multi-state model for competing events

So far, our focus has been on estimating the CI for CAD. However, there are multiple events of interest such as RVSD and PH making the modeling becomes even more complicated. In this section we outline some of the issues in modeling such data and discuss some analytical approaches. For simplicity let us first assume that all three types of CEs are mutually exclusive and if the interest is in estimating CI for all three types of events, then we can proceed to model it according to the illness-death model proposed below.

In the above setting we assume that the three CDs of interest are mutually exclusive, and each survivor can have only one event during the follow-up time. In such situations, one can extend the parametric models proposed in Section 2 using either exponential or Weibull distribution. Although, in principle, the approach could be easily implemented but the likelihood representation may be somewhat complicated, and the estimations process could be computationally more involved. The pseudo survival functions can be obtained as,

$$Q_{1a}(t) = \exp\{-\int_0^t \lambda_{1a}(\nu)d\nu\}, Q_{1b}(t) = \exp\{-\int_0^t \lambda_{1b}(\nu)d\nu\}, \;\; \text{and}$$
$$Q_{1c}(t) = \exp\{-\int_0^t \lambda_{1c}(\nu)d\nu\}, Q_2(t) = \exp\{-\int_0^t \lambda_2(\nu)d\nu\}$$

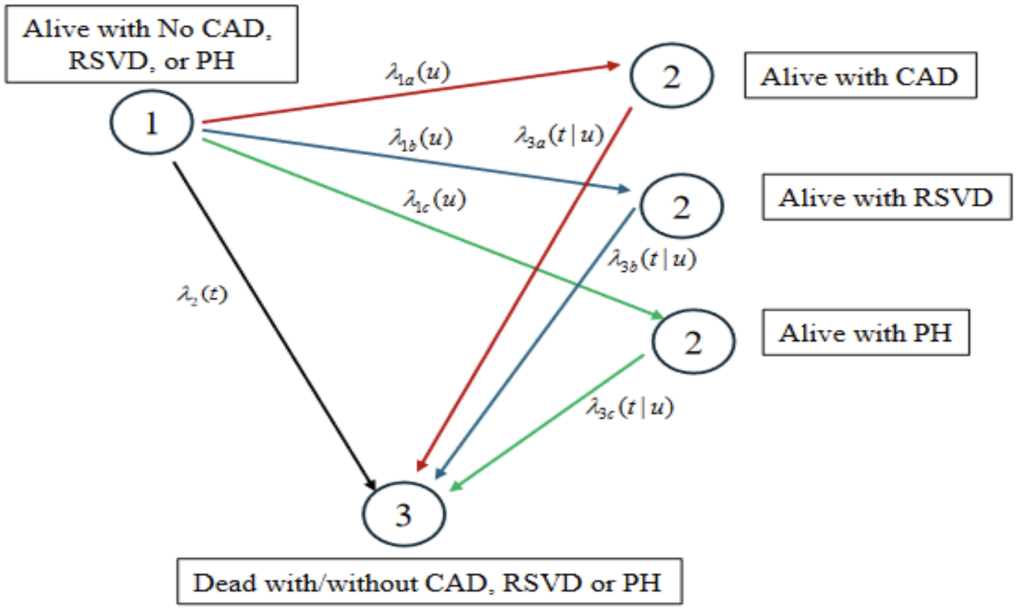

Figure 2: Three-states illness/death model of patients developing cardiovascular dysfunction

and

$$Q_{3a}(t|u) = \exp\{-\int_u^t \lambda_{3a}(\nu|u)d\nu\}, Q_{3b}(t|u) = \exp\{-\int_u^t \lambda_{3b}(\nu|u)d\nu\},$$
$$Q_{3c}(t|u) = \exp\{-\int_u^t \lambda_{3c}(\nu|u)d\nu\}$$

It may be noted that probability of surviving without experiencing and any CD or death beyond time t can be represented as,

$$Q(t) = \exp\{-\int_0^t (\lambda_{1z}(\nu) + \lambda_{1b}(\nu) + \lambda_{1c}(\nu) + \lambda_2(\nu))d\nu\} = Q_{1a}(t)Q_{1b}(t)Q_{1c}(t)Q_2(t) \quad (18)$$

and the survival function can be obtained as,

$$\begin{aligned} S(t) &= P(X(t) = 1) + P(X(t) = 2) \\ &= Q(t) + \int_0^t \lambda_{1a}(u)Q(u)Q_{3a}(t|u)du + \int_0^t \lambda_{1b}(u)Q(u)Q_{3b}(t|u)du + \\ &\int_0^t \lambda_{1c}(u)Q(u)Q_{3c}(t|u)du \end{aligned} \quad (19)$$

Then, using the above pseudo survival functions one can obtain the contributions to the likelihood made by each survivor depending on their outcome which can be described as follows:

(i) Death without any CD, $T = t$, $X(t^-) = 1$, and $L_1(t) = \lambda_2(t)Q(t)$,

(ii) Alive without any CD, $T > t$, $X(t) = 1$, and $L_2(t) = Q(t)$,

(iii) Death with CAD, $T = t$, $X(t^-) = 2$, and $L_{3a} = \int_0^t \lambda_{1a}(u)Q(u)\lambda_{3a}(t|u)Q_{3a}(t|u)du$

(iv) Alive with CAD, $T > t$, $X(t) = 2$, and $L_{4a} = \int_0^t \lambda_{1a}(u)Q(u)\lambda_{3a}(t|u)du$

(v) Death with RSVD, $T = t$, $X(t^-) = 2$, and $L_{3b} = \int_0^t \lambda_{1b}(u)Q(u)\lambda_{3b}(t|u)Q_{3b}(t|u)du$

(vi) Alive with RSVD, $T = t$, $X(t) = 2$, and $L_{4b}(t) = \int_0^t \lambda_{1b}(u)Q(u)Q_{3b}(t|u)du$

(vii) Death with PH, $T > t$, $X(t^-) = 2$, and $L_{3c}(t) = \int_0^t \lambda_{1c}(u)Q(u)\lambda_{3c}(t|u)Q_{3c}(t|u)du$

(viii) Alive with PH, $T = t$, $X(t) = 2$, and $L_{4c}(t) = \int_0^t \lambda_{1c}(u)Q(u)Q_{3c}(t|u)du$

Based on the above quantities one can write down the likelihood function and obtain the estimates of the parameters of interest using the theory of maximum likelihood or EM algorithms as discussed in Sections 2 and 3. Extension of semi-parametric approach for simultaneous modeling of the three types of CDs requires more theoretical development and is proposed as future work.

5. Conclusions and discussions

In this manuscript, we have provided an overview of the parametric and semi-parametric approaches that could be adopted for modeling CI of one or more competing events of interest with death being an absorbing state.

When the survivors are followed longitudinally then, under the assumption of continuous time Markov process, one can easily adopt the likelihood approach to model the transition probabilities as discussed in van den Hout (2017, Chapter 4). The development of semi-parametric approach needs to be developed and is left as future work.

The SJLIFE cohort study is a unique study to evaluate the association of childhood cancer treatment with the long-term adverse effect. The discussed approach can be extended to any interval-censored data or any multi-state models and could be extremely useful in the prediction of adverse outcomes.

There are multiple challenges to drawing statistical inference from such studies. Robustness of results may depend on the selection of the cohort and time of data collection as discussed in Table 2. Estimation can be based on parametric, non-parametric or semi-parametric models. The number of parameters not only depends on estimation procedure but also number of stages as included in Figures 1 and 2. Incorporating covariate effects on different parameters in Figures 1 and 2 makes inference much more cumbersome.

Funding

S.N. Rai was partly supported by the University of Cincinnati Cancer Center, College of Medicine. D.K. Srivastava, K. Ness and M. Hudson were in part supported by the St. Jude Children's Research Hospital Cancer Center Support Grant No. 5P30CA021765-33, the St. Jude Lifetime Cohort Study Grant No. U01 CA195547, and the American Lebanese Syrian Associated Charities (ALSAC).

Acknowledgements

We thank Dr. Vinod Gupta, Chair Editor, Statistics and Applications President, Society of Statistics, Computer and Applications (SSCA); ssca.org.in/ and President, Governing Body, Institute of Applied Statistics and Development Studies.

References

Agresti, A. (2013) *Categorical Data Analysis*. Wiley-Interscience, A John Wiley & Sons, INC., New York.

Armstrong, G. T., Liu, Q., Yasui, Y., Neglia, J. P., Leisenring, W. L., *et al.* (2009). Late mortality among 5-year survivors of childhood cancer: a summary from the childhood cancer survivor study. *Journal of Clinical Oncology*, **27**, 2328-2338.

Bhakta, N., Liu, Q., Ness, K. K., Bassiri, M., Eissa, H., *et al.* (2017). The cumulative burden of surviving childhood cancer: An initial report from the St. Jude Lifetime Cohort Study. *LANCET*, **390**, 2569-2582.

Ehrhardt, M. J., Krull, K. R., Bhakta, N., Liu Q., Yasui Y., *et al.* (2023). Improving quality and quantity of life for childhood cancer survivors globally in the twenty-first century. *Nature Reviews Clinical Oncology*, **20**, 678-698.

Eulenburg, C., Manher, S., Woeiber, L., and Wegscheider, K. (2015). A systematic model specification procedure for an illness-death model without recovery. *PLoS One*, **10**, e0123489.

Gunaratnam, B. and Rai, S. N. (2019). Comparing the variability using Louis' method and resampling methods. *Journal of Biometrics and Biostatistics*, **10**.

Hammoud, R. A., Liu, Q., Dixon, S. B., Onerup, A., Mulrooney, D. A., *et al.* (2024). The burden of cardiovascular disease and risk for subsequent major adverse cardiovascular events in survivors of childhood cancer: a prospective, longitudinal analysis from the St. Jude Lifetime Cohort Study. *The Lancet Oncology*, **25**, 811-822.

Harezlak, J., Gao, S., and Hui, S. L. (2003). An illness-death stochastic model in the analysis of longitudinal dementia data. *Statistics in Medicine*, **22**, 1465-1475.

Howell, C.R., Bjornard, K. L., Ness, K. K., Alberts, N., Armstrong, G. T., *et al.* (2021). Cohort Profile: The St. Jude Lifetime Cohort Study (SJLIFE) for pediatric cancer survivors. *International Journal of Epidemiology*, **50**, 39-49.

Hudson, M. M., Ehrhardt, M. J., Bhakta, N., Baassiri, M., Eissa, H., *et al.* (2017). Approach for Classification and Severity Grading of Long-term and Late-Onset Health Events among Childhood Cancer Survivors in the St. Jude Lifetime Cohort. *Cancer Epidemiology, Biomarkers & Prevention*, May; **26**, 666-674. doi: 10.1158/1055-9965.EPI-16-0812. EPI-16-0812. Epub 2016 Dec 29. PMID: 28035022; PMCID: PMC5413397.

Kalbfleisch, J. D. and Lawless, J. F. (1985). The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association, Theory and Methods*, **80**, 863-871.

Koutsoukis, A., Ntalianis, A., Repasos, E., Kastritis, E., Dimopoulos, M. A., *et al.* (2018). Cardio-oncology: A Focus on Cardiotoxicity. *European Cardiology Review*, **13**, 64-69.

Murphy, S. A. and Van der Vaart, A. W. (2000). On Profile Likelihood. *Journal of the American Statistical Association*, **95**, 449–465.

Park, J.R., Bagatell, R., London, W.B., Maris, J.M., Cohn, S.L., *et al.* (2012). Children's Oncology Group's 2013 blueprint for research: neuroblastoma. *Pediatric Blood Cancer*, **60**, 985-993.

Pradhan, B. and Kundu, D. (2014). Analysis of Interval-Censored Data with Weibull Lifetime Distribution. *Sankhya B*, **76**, 120–139.

Qian, C., Srivastava, D. K., Pan, J., Hudson, M. M., and Rai, S. N. (2023). Estimating transition intensity rate on interval-censored data using semi-parametric with EM algorithm approach. *Communications in Statistics - Theory and Methods* **53**, 1-17.

Rai, S. N. and Matthews, D. E. (1993). Improving the EM algorithm. *Biometrics*, **49**, 587-591.

Rai, S. N. and Matthews, D. E. (1997). Discrete scale models for survival-sacrifice experiments. *Applied Statistics*, **48**, 93-109.

Rai, S. N., Matthews, D. E., and Krewski, D. R. (2000). Mixed-scale models for survival/sacrifice experiments. *The Canadian Journal of Statistics*, **28**, 65-80.

Rai, S. N. (2008). Analysis of occult tumor studies. In: Tan WY and Hanin L (Ed.), *Handbook of Cancer Models with applications.* World Scientific Press.

Rai, S. N., Pan, J., Sun, J., Hudson, M. M., and Srivastava, D. K. (2013). Estimating incidence rate on current status data with application to a Phase IV cancer trial. *Communications in Statistics: Theory and Methods*, **42**, 2417-2433.

Robison, L. L., Armstrong, G. T., Boice, J. D., Chow E. J., Davies, S. M., *et al.* (2009). The childhood cancer survivor study: a National Cancer Institute – supported resource for outcome and intervention research. *Journal of Clinical Oncology*, **27**, 2308-2318.

Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER *Stat Database: Incidence – SEER Research Data, 8 Registries, Nov 2021 Sub (1975-2019) – Linked To County Attributes – Time Dependent (1990-2019) Income/Rurality 1975-2020 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, released April 2022, based on the November 2021 submission.

Srivastava, D. K., Zhu, L., Hudson, M. M., Pan, J., and Rai, S. N. (2018). Robust estimation and inference on current status data with applications to Phase IV cancer trial. *Journal of Modern Applied Statistical Methods*, **17**, Article no. 18, 1-20.

Sun, J. (2006). *The Statistical Analysis of Interval-Censored Failure Time Data.* Springer: New York.

van den Hout A. (2017). *Multi-State Survival Models For Interval Censored Data.* CRC Press, Taylor and Francis Group, New York.

Winther, J. F., Kenborg, L., Byrne, J., Hjorth, L, Kaatsch, P., *et al.* (2015). Childhood cancer survivor cohorts in Europe. *Acta Oncologica*, **54**, 655-668.

Zhang, X., Zhang, Y., Ye, X., Guo, X., Zhang, T., *et al.*. (2016). Overview of phase IV clinical trials for postmarket drug safety surveillance: a status report from the ClinicalTrials.gov registry. *British Medical Journal (open)*, **6**, e010643.