Imputing Partial Status and Estimating Incidence Rate in an Illness-death Model with Application to a Phase IV Cancer Trial

Deo Srivastava
St. Jude Children's Research Hospital
Jianmin Pan
University of Louisville
Chen Qian
University of Louisville
Melissa Hudson
St. Jude Children's Research Hospital
Shesh Rai (shesh.rai@louisville.edu)
University of Louisville

Research Article

Keywords: Phase IV clinical trial, Imputation, Cross-sectional survey data, Interval censored data, K-M Method, Missing Value

Posted Date: December 2nd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-108160/v1

Imputing Partial Status and Estimating Incidence Rate in an Illness-death Model with Application to a Phase IV Cancer Trial

Deo K Srivastava[a], Jianmin Pan[b], Chen Qian[b], Melissa M Hudson[c], Shesh N Rai[b¶]

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

[b] *Biostatistics and Bioinformatics Facility, James Graham Brown Cancer Center, University of Louisville, Louisville, Kentucky, 40202, USA*

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

¶ Corresponding Author:

Dr. Shesh N. Rai, Biostatistics and Bioinformatics Facility, James Graham Brown Cancer Center, University of Louisville, Louisville, Kentucky, 40202, USA

[†] *E-mail: shesh.rai@louisville.edu*

Abstract

Background

Phase IV clinical trials are designed to monitor the long-term toxic effects of drugs in cancer survivors. Evaluations to study the long-term effects of the cancer treatment are often made with cross-sectional surveys. This leads to interval censored data since the exact time of the onset of toxicity is not known. In addition to finding prognostic factors for log-term survival outcome, estimating and comparing the cumulative incidence rates for adverse outcomes of interest for interval censored data is also desired. However, the analysis of such data is further complicated by many issues, such as incomplete data, competing risks and selection bias. For example, one such study was designed by Hudson et al. to study the effect of anthracyclines exposure, received as part of treatment for childhood cancer, to cardiotoxicity. Rai et al. had utilized a parametric approach for assessing the effect of anthracycline on the cumulative incidence of cardiotoxicity but excluded the patients with missing information on the parameters used for assessing cardiotoxicity.

Methods

In this paper our focus is on imputing the missing data and then using the current status regression methods, previously described in Rai et al. for estimating and comparing cumulative incidence rates in an illness-death/failure model.

Results

We undertook a comprehensive simulation study to evaluate the performance of our imputation approach and applied it to a Phase IV clinical trial to evaluate the effect of anthracycline exposure on long-term cardiotoxicity in childhood cancer survivors, which had missing cardiotoxicity information.

Conclusions

Our simulations suggest that the results obtained by imputing the missing values using regression methods are significantly more efficient than those obtained without imputation. The proposed approach is easy to implement, and we demonstrate its usefulness by applying it to the data reported in Rai et al. and compare the results reported there to our approach that utilizes imputation.

Keywords: Phase IV clinical trial; Imputation; Cross-sectional survey data; Interval censored data; K-M Method; Missing Value.

Background

As a result of more modern therapies and better supportive care the 5-year survival rate for childhood cancer has improved significantly and currently exceeds 80%[3-6]. In 2011 there were 390,000 survivors of childhood cancer living in the US and it is expected that by 2020 there will be more than 500,000 childhood cancer survivors[7]. This improvement in survival rate comes at a price as these survivors are at an elevated risk of experiencing long-term morbidity and early mortality as a result of their cancer and its treatment. The purpose of Phase IV clinical trials is to monitor long-term sequala and develop interventions to mitigate their effect in long-term. A chemotherapy agent Anthracycline has served as the backbone for many pediatric malignancies because of its therapeutic effects but it is also well known to be cardiotoxic[8-9]. One such study was undertaken by Hudson et al.[1] to evaluate the effect of anthracycline exposure on cardiotoxicity using non-invasive modalities.

Cardiotoxicity is the occurrence of heart electrophysiology dysfunction or/and muscle damage. The heart becomes weaker and is not as efficient in pumping, and therefore, circulating blood. There are many measures of electrophysiology dysfunction or/and muscle damage, including shortening fraction, afterload, QTc interval, and ejection fraction[1, 9, 10]. It is not economic and feasible to evaluate patients very frequently to estimate the onset time of cardiotoxicity, and hence estimate the incidence rates. Usually patients are followed longitudinally in the clinics, but not all follow a routine pattern. Therefore, it is convenient to design cross-sectional surveys for estimating the effect of long-term side effect of treatments and its predictors. We only know the current status of the patient with onset prior to current status but not the actual onset times of these events. These types of incomplete data are referred to as interval censored data since the actual onset time of the events are unknown[2, 8-11] and our interest is in estimating the onset rate or the cumulative incidence rate.

Nonparametric procedures for analyzing interval censored failure time data have been extensively studied and discussed in the literature.[11-15] Another issue in the cross-section survey study is that results need to be generalized to the specific population. There can be competing toxic effects from the same drug. Sun[11] provides an extensive survey of non-parametric methods of estimation using EM algorithm in studies involving interval censored data. In this paper, we have the same interest, as Rai et al.[2], in estimating the cumulative incidence rates in a parametric setting but focus on improving the accuracy by imputing the missing observations using multivariable regression method.

In practice, most investigators exclude observations with missing values and incomplete cases. While using only complete cases has its simplicity, one may lose the important information in the incomplete cases and ignore the possible systematic differences between the complete and incomplete cases. Hence, the resulting inference may not be applicable to the population of all cases, especially with a smaller number of complete cases. It is well known that imputation is a widely used method for handling missing data. Little and Rubin[16] and Buuren[17] provide an excellent overview of the methods for conducting analyses with missing data. For further information on multiple imputations see Rubin[18-19], and Rubin, Stern, and Vehovar[20] discuss imputation of missing

discrete data. King et al.[21] review many of the practical costs and benefits of multiple imputations. For routine imputation of missing data, Schafer[22] presents a method based on multivariate normal distribution. Liu[23] uses the t distribution, and Van Buuren, Boshuizen, and Knook[24] use interlocking regressions. Furthermore, Troxel, Ma, and Heitjan[25] present a method to study the sensitivity of inferences to missing-data assumptions.

This paper is organized as follows. In the following subsection, we provide the details of the motivating example to introduce the problem. In Methods section, we give a brief description of the procedure introduced in Rai, et al.[2] and construct corresponding likelihood function. The data from the motivation example is analyzed by imputing the missing values and compared with the results obtained without imputation in the Results section. An extensive simulation experiment to study the performance of the imputation approach are summarized as well. The Discussion section is devoted to miscellaneous remarks.

Motivation Example

A study was undertaken by Hudson et al.[1] to evaluate the effect of anthracycline on cardiotoxicity using 12-lead ECG and echocardiography, non-invasive technique. For the study, the cancer survivors were recruited from St. Jude Children's Research Hospital After Completion of Therapy Clinic. The survivors were classified into two groups; the first group of survivors consisted of survivors that received cardiotoxic therapy (anthracycline and/or thoracic radiation) and the other group did not receive cardiotoxic therapy (no anthracycline nor thoracic radiation). The details of the study can be found in Hudson et al.[1]. The study was approved by the institutional review boards at St Jude Children's Research Hospital and Stanford University (Stanford, CA). All study participants or their parents provided informed consent. Survivors with cardiotoxic therapy were designated as At-Risk (AR) and those without cardiotoxic therapy as Not At-Risk (NR).

Onetime clinical assessment was made by the primary oncologists to identify the survivors with signs of heart failure using New York Heart Association classification. Along with the clinical evaluation non-invasive testing based on 12-lead ECG and echocardiography within 24 hours of the clinical assessment.

When the cardiac measures, discussed above, are out of the normal range, patients are declared to have cardiotoxicity. Following Hudson et al.[1], we consider two outcome measures, fractional shortening (*FS*) and afterload (*AF*). The main measure is defined as *FS= (LVEdD-LVEsD)/LVEdD*, where *LVEdD* is the left ventricular end-diastolic diameter, *LVEsD* is the left ventricular end-systolic diameter. The other measure *AF* can be described as the pressure that the chamber of the heart has to generate in order to eject blood out of the chamber. The study was planned to enroll almost equal number of patients from each group to detect a medium effect size[26] increase in mean *AF* (or decrease in mean *FS*) at α =0.05 and β =0.20, without adjusting for multiple outcomes or multiple comparisons.

A short summary about the cohort is summarized in Table 1. Further description is available in Hudson et al.[1] On closer examination of the data it was seen that there were missing data in outcome measures *AF* and *FS*. A total of 40 survivors had missing *AF* values and 6 of those were also missing *FS* values. To keep the discussion simple and straight forward these 6 observations

were deleted from our analysis and we focused our attention on imputing 34 missing *AF* observations. Although, the imputation approach discussed could easily be applied to data missing in several variables in a recursive manner. The scatter plot of *AF* and *FS* displayed in Figure 1, does not show any clear missing pattern; the missing values of AF are in the entire range of values of FS. Also note that the missing proportion of *AF* values in the NR (7/54 = 13%) and AR (27/218 = 12%) were almost similar.

Table 1. Characteristics of 278 Patients Enrolled Onto the Noninvasive Cardiac Study

	At-Risk Group (n=223)		**Not At-Risk Group (n=55)**		**Total (n=278)**	
Demographics	**N**	**%**	**N**	**%**	**N**	**%**
Sex						
Male	108	51.6	31	56.4	139	50
Female	115	48.4	24	43.6	139	50
Treatment group						
Anthracycline	157	70.4	0	0	157	56.5
Anthracycline + Radiation	60	26.9	0	0	60	21.6
Radiation	6	2.7	0	0	6	2.1
None	0	0	55	100	55	19.8
Race/ethnicity						
White	183	82.1	44	80.0	227	81.7
Black	30	13.5	11	20.0	41	14.7
Other	10	4.4	0	0	10	3.6
Diagnosis						
Leukemia	67	30.0	10	18.2	77	27.7
Sarcomas	60	26.9	14	25.4	74	26.6
Lymphoma	54	34.2	2	3.6	56	20.1
Embryonal tumors	42	18.8	29	52.7	71	25.6
Age at Cancer Diagnosis, Years						

N	223	55	278
Mean	7.37	5.77	7.05
Median	5.46	3.11	4.68
Range	0.01-23.56	0.29-20.06	0.01-23.56
Afterload			
N	191	47	238
Mean	57.50	45.73	55.18
Median	55.43	42.18	51.88
Range	15.38-147.32	25.66-95.02	15.38-147.32
Fractional Shortening			
N	218	54	272
Mean	0.33	0.36	0.34
Median	0.33	0.36	0.34
Range	0.20-0.57	0.24-0.49	0.20-0.57

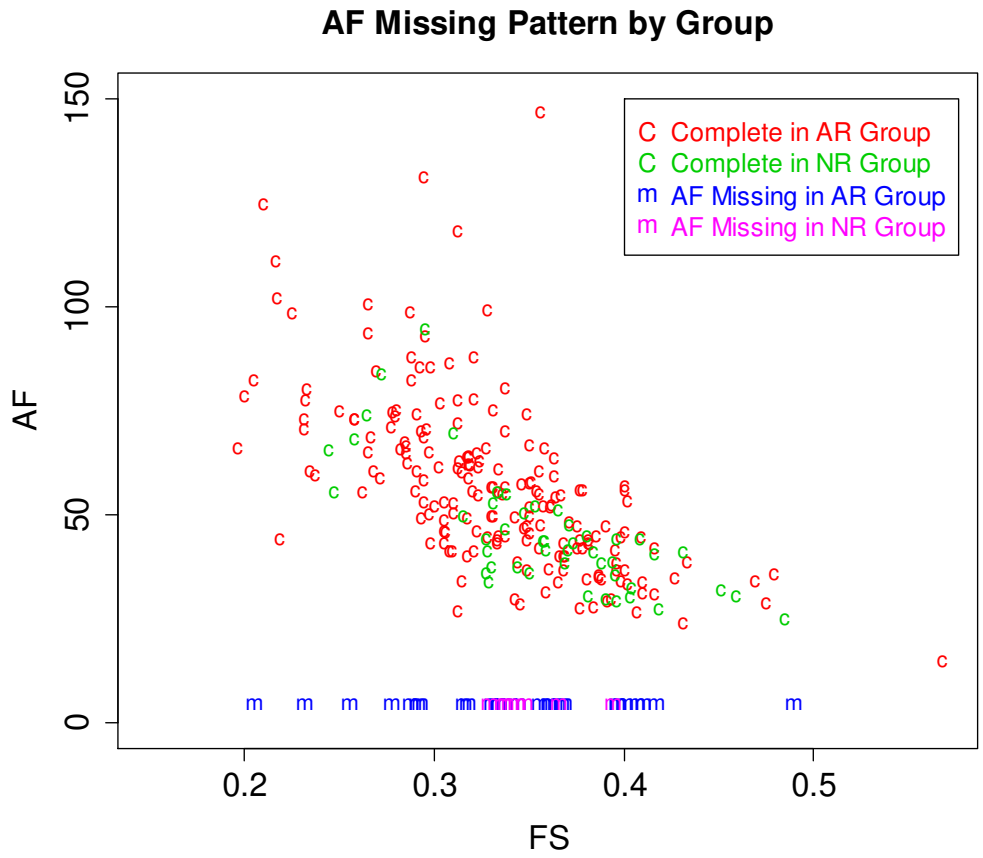

Figure 1: Scatter plot of *AF* and *FS* within AR and NR group of patients. Complete data (labeled c) displays strong correlation between AF and FS. Missing values of AF are in almost entire range of FS values (labeled m).

Using actual measures of these dependent variables *FS* and *AF*, the threshold values were used to classify patients as abnormal if (*FS*< 0.28) or (*AF*> 74 g/cm^2). Threshold values for *FS* and *AF* were determined based on published normative data[27-28]; these are well accepted norms. Let *AFS* and *AAF* denote the indicators of these abnormalities. The 278 patients participated in the study; 223 were designated AR and 55 were designated NR based on treatment. Data on each individual also included demographics, date of cancer diagnosis, time since treatment completion, disease related variables (such as type, histology, and stage of cancer), treatment related variables (such as chemotherapy drugs, doses and irradiation). In the AR group, noninvasive assessment identified subclinical dysfunction with *FS* in 37 (13.6%) of 272 and *AF* in 33 (13.9%) of 238; prolonged QTc interval in 11 (4.0%) of 273. These are the estimates of prevalence of cardiac abnormalities. Among others, one main objective of the study is to estimate cumulative incidence rates of *AFS* and *AAF*.

In this study echocardiography was performed as a research measure and not in response to clinical symptoms. Individuals with previously established cardiac disease were excluded from participation. As formally assessed by New York Heart Association classification, none of the study participants reported clinical symptoms of cardiac dysfunction at enrollment. The imaging quality

in echocardiography is dependent on obtaining a clear acoustic window from which ventricular volumes are estimated based on geometric assumptions. Operator experience and variations in thoracic structures can contribute to difficulties in obtaining technically satisfactory data in a given study. These factors randomly contributed to missing data among study participants. *AF* is not a standardly used assessment in clinical practice, thus, despite training of ultrasonographers for this study, this factor may have contributed to a higher prevalence of missing *AF* measurements compared to *FS*. In other words, *AF* is either under detected or over detected, but operators might miss it, and therefore, causing missing values. Thus, we feel that there is no selection bias related to those with and without abnormal *FS* and *AF* identified as part of the study. The missing values of *AF* are displayed in Figure 1, also do not show any pattern.

A crude approach to estimating the incidence rates and obtaining confidence intervals is to apply the Kaplan-Meier estimator with the assumption of the evaluation time as the onset time. Then, incidence rate of each type of toxicity is estimated. Some of these toxicity measures could be missing. In this paper, we impute the missing measurements using regression method first and then use a parametric approach to estimate incidence rates of specific toxicity. For the cardiotoxicity data the 34 missing *AF* values were imputed using a multivariable regression with *FS* and other covariates, such as age, diagnosis, risk status, BMI as predictors. The estimates of cumulative incidence rates were derived based on the data after imputation and then compared with those derived in Rai, et al.[2] without imputation.

Methods

Cardio-Measures Abnormality Model

The descriptive statistics of all participants can be found in Table 1. In Hudson et al.,[1] the subjects who died or had cardiac failure during the treatment or during the follow-up after completion of therapy were excluded because the number of deaths at the time of the analysis were too few. However, this information is available from the medical record abstraction and with longer follow-up the number of deaths would increase. Hence, we present the general theory here for a cross-sectional data with indicators of cardiac abnormality and death/cardiac failure, and time since the treatment to the survey or the death/cardiac failure, as depicted in Figure 2. We also assume cardiac abnormality is the precursor for cardiac failure.

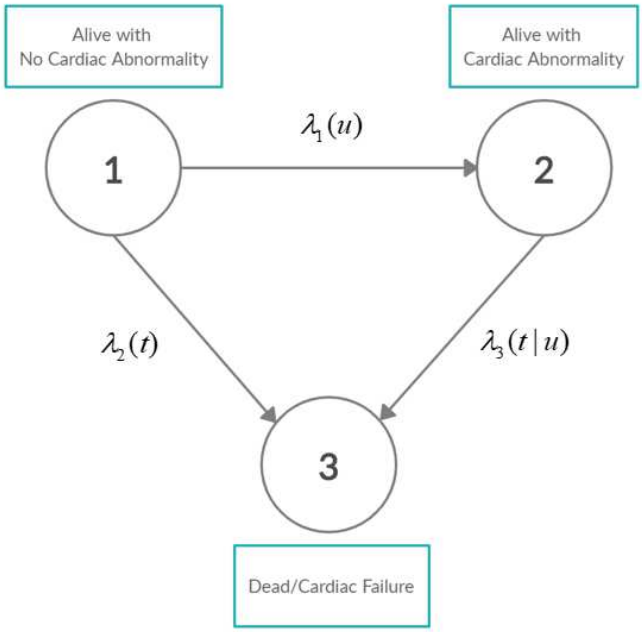

Figure 2: An abnormal cardiac measure-death/cardiac failure model involving three states. State 1 corresponds to patients who are alive with normal value. Patients who are alive with abnormal value are in state 2. State 3 is an absorbing state and corresponds to death or cardiac failure.

Let stochastic process $\{X(t)\}$ identify the state occupied by a patient at time t. For simplicity, we suppose that n patients in state 1 at time $t = 0$ are those who are identified with different disease groups and are planned for treatment, where we have assumed that no patient has cardiac abnormality at time $t = 0$. Let the random variable T denote the observation time (survey, death/cardiac failure) from the study evaluation and U the time of *AFS* or *AAF* from the study evaluation. Thus, at any time t, $X(t) = 1, X(t) = 2$ and $X(t) = 3$ indicate the patient alive with normal cardiac measure, alive with abnormal cardiac measure and died or had cardiac failure with or without cardiac abnormality, respectively. We also assume that the development of abnormality is an irreversible event without the treatment for cardiotoxicity, and therefore, transitions from state 2 to state 1 do not occur, as illustrated in Figure 2. According to practice in this study, the patients are chosen for survey independent of their health status, which ensures that the survey results can be regarded as independent of the times of the events of interest. The intensities $\lambda_1(u)$, $\lambda_2(t)$ and $\lambda_3(t|u)$, shown in Figure 2, are corresponding transitions rates, where t is the observation time and u is the time of *AAF* or *AFS*.

The survival function and the cardiac abnormality prevalence function are derived in Rai et al.[2] as follows

$$S(t) = Q(t) + \int_0^t \lambda_1(u)Q(u)Q_3(t|u)du \quad \text{and} \quad \pi(t) = \frac{\int_0^t \lambda_1(u)Q(u)Q_3(t|u)du}{S(t)},$$

where $Q(t) = Q_1(t)Q_2(t)$ is the probability that the time to the first event— alive with abnormal value or death with normal value — exceeds t, and

$$Q_i(t) = exp\{-\int_0^t \lambda_i(v)dv\}$$

for $i = 1$ and 2, and

$$Q_3(t|u) = exp\{-\int_u^t \lambda_3(v|u)dv\}.$$

are pseudo-survival functions corresponding to the intensities $\lambda_1(u)$, $\lambda_2(t)$ and $\lambda_3(t|u)$. We are interested in estimating $\Lambda_1(t) = \int_0^t \lambda_1(u)du$, the cumulative incidence function (CIF), but focus on the comparison of CIFs between AR and NR groups based on the original data and imputation data.

Table 2 identifies the various types of observations which occur in this illness-death/failure model and the corresponding contribution to the likelihood, denoted as $L_1(t)$ to $L_4(t)$, which are functionals of intensities and pseudo-survival functions. Rai et al.[2] derive the explicit form of $L_1(t)$ to $L_4(t)$ for both constant and piecewise exponential model and the likelihood functions, which are summarized in appendix.

Table 2: Likelihood Contributions for Anthracycline Cardiac Toxicity Study

Observation Type	Outcome	Likelihood Contribution
Death with No Cardiac Abnormality	$T = t, X(t^-) = 1$	$L_1(t) = \lambda_2(t)Q(t)$
Alive with No Cardiac Abnormality	$T > t, X(t) = 1$	$L_2(t) = Q(t)$
Death/Cardiac Failure with Cardiac Abnormality	$T = t, X(t^-) = 2$	$L_3(t) = \int_0^t \lambda_1(u)Q(u)\lambda_3(t \mid u)Q_3(t \mid u)du$
Alive with Cardiac Abnormality	$T > t, X(t) = 2$	$L_4(t) = \int_0^t \lambda_1(u)Q(u)Q_3(t \mid u)du$

Imputation Model

Let the cardiac measure, such as *AF*, be denoted by *Y*. Assume that $Y = (Y_1, Y_2)^T$ be a $n \times 1$ response vector with Y_1 $(n_1 \times 1)$ observed and Y_2 $(n_2 \times 1)$ missed, and $X = (X_1, X_2)^T$ be corresponding $n \times p$ matrix comprised of covariates including other cardiac measures (other response variables).

There are several methods for imputation which can be broadly classified as single imputation or multiple imputation (MI). In MI approach several copies of the complete data set are created and then the appropriate statistical method is applied to each data set and the results from these analyses are then combined to provide the final results. Usually, MI approaches are preferred over single

imputation as they incorporate variability due to imputation[17, 29-30]. There are many MI approaches discussed in literature, but two most commonly used approaches based on joint multivariate modeling or fully conditional specification perform quite well in the regression setting, as seen in Huque et al[31]. It may be noted that PROC MI can perform imputations for data that have monotone or arbitrary missing patterns. PROC MI with FCS option, a standard feature in SAS version 9.4[32], utilizes the conditional distribution and can incorporate both continuous and categorical variables appropriately, see Liu and De[33]. In our setting we had missing values only in *AF* and we wanted to take advantage of the relationship between *AF* and other covariates of interest that included categorical variables. Therefore, we preferred to perform the imputations using PROC MI in SAS with FCS option. The method can be briefly described as follows:

A multivariable regression model $y = x\beta + \varepsilon$ is fitted based on the complete data Y_1 and X_1, and the least squared estimator $\hat{\beta}$ of β $(p \times 1)$ and associated variance-covariance matrix is obtained. Then, missing values in Y_2 are imputed using the posterior predictive distributions, see PROC MI in SAS[32] for details. It is natural to use the imputation data $(Y_1, \hat{Y}_2)^T$ instead of only Y_1 and is anticipated that the imputed information in $\hat{Y}_2$ will improve the related results in statistical analysis.

To each complete data set likelihood ratio test was applied to compare the two risk groups (AR and NR). In the regression framework one could use PROC MIANALYZE in SAS to combine the results from multiple imputations to conduct inference that incorporates inherent variability introduced due to imputations[17,29-30]. However, in our setting p-values associated with each imputation are obtained based on the likelihood ratio test. Then, the overall conclusion can be based on some type of summary measure of all the p-values such as mean or median. We prefer to report the results based on median as that would be much more robust than mean.

Results

Application: Cancer Survivor Study

In this section we obtained the imputed data for the cardiotoxicity example and applied the theory for the exponential model described in appendix to evaluate the effect of anthracyclines on cardiotoxicity. Furthermore, the results obtained using the imputation approach are then compared with those obtained without imputation, reported in Rai et al.[2], under the assumption of no deaths/cardiac failures. The simplest model is the Parametric-1, which is one parameter Exponential model. Since there are very few events before 5 years and after 10 years, we also fit two piecewise Exponential models; Paremeter-2, based on two incidence rates, one up-to year five and the second for year 5 and above, and Parameter-3, based on three incidence rates one up-to year 5, second between years 5 and 10 and the last one for year 10 and above, (see Figure 3).

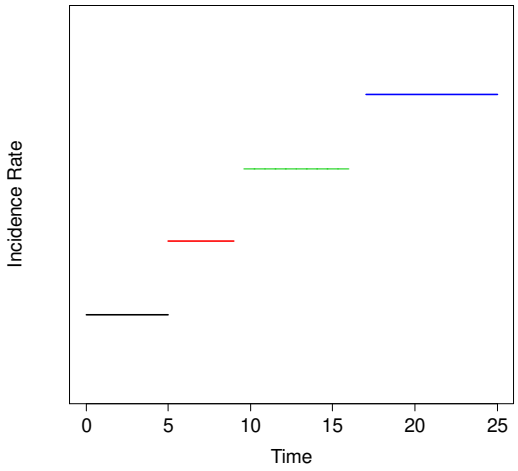

Figure 3: Piecewise Incidence Rate

In the cardiotoxicity example, there are 278 subjects, Leukemia (n=77), Sarcoma (n=74), Lymphoma (n=56) and Embryonal (n=71), and 34 measurements were missing for *AF* and 6 were missing in both measurements AF and FS, but no covariate information was missing. We exclude the 6 with both missing. Hence, we have 272 subjects and employ the multivariable regression method to estimate the 34 missing measurements in *AF* based on the values of *FS* and corresponding covariates, like age at diagnosis, race, gender, BMI, QTC, diagnosis group and risk group (AR/NR) [27-28] . Based on 4 diagnosis groups, we define three dummy variables as follows:

$$\text{Diag1} = \begin{cases} 1 & \text{Leukemia} \\ 0 & \text{Otherwise,} \end{cases} \quad \text{Diag2} = \begin{cases} 1 & \text{Sarcoma} \\ 0 & \text{Otherwise,} \end{cases} \quad \text{Diag3} = \begin{cases} 1 & \text{Lymphoma} \\ 0 & \text{Otherwise.} \end{cases}$$

Before conducting the regression analysis, the Shapiro-Wilk test of normality was applied to original *AF* and *FS* measurements and a few commonly used transformations for making the underlying distributions of *AF* and *FS* more normal. Log(*AF*) was normally distributed (p=0.701) but original AF (p<0.001) was not. On the other hand, Log(FS), Logit(FS) and FS were not normally distributed with p values <0.001, 0.007 and 0.026, respectively. This suggests fitting the regression model using logarithm transformation of *AF* and original *FS*. The significant predictors with coefficients, their p-values and $\boldsymbol{R^2}$ in the model are presented in Table 3a. That is,

$$\log(\text{AF}) = 5.534 - 3.891\text{FS} + 0.008\text{Age} + 0.081\text{Risk} + 0.098\text{Diag2} - 0.010\text{BMI} + \varepsilon \qquad (4.1)$$

where Risk=1 for patient in AR group and 0 in NR group. Based on the regression model, the values of *FS* and the covariates, the 34 missing *AF* values are imputed. Thus, after imputing the missing values the total sample size is 272.

Table 3a: Coefficients and P-values in Regression Model			
Variable	**Estimator**	**SE**	**p-value**
Intercept	5.534	0.106	<0.001

FS	-3.891	0.274	<0.001
Age at Diagnosis	0.008	0.003	0.005
Risk Group	0.081	0.038	0.033
Diag2	0.098	0.035	0.006
BMI	-0.010	0.003	<0.001
R^2	0.586		

Remark: It may be noted that the R^2 for the above model is 0.586 which represents a reasonable fit but may not be the best model fit. However, in general, it would not be unreasonable to expect that the imputation process would be more efficient if a better model with higher value of R^2 can be obtained.

Because *AF* is the only variable with missing values in our data (n=272), we consider the data including age at diagnosis, BMI, diagnosis group, risk group, *FS*. Based on each imputation data, we calculate the Cumulative Incidence Function (CIF) for constant exponential model, two or three piecewise exponential model using the method. Then, the methods described in Appendix were applied to the imputed data sets for each group, AR and NR, and both groups combined. Then, based on the likelihood ratio test, the corresponding p-value for group effect for the variable *AF* is 0.014 (median of $m = 20$ imputations). On the other hand, the p-value without imputation is 0.020^2. The results with imputation seem to be little bit more sensitive for the group effect in *AF* compared to that without imputation. Note that not using the illness/death model as proposed and, instead, using logistic regression, without imputation, the group effect was only marginally significant (p=0.065)[1-2]. Thus, the approach base on illness/death model led to a better understanding of this data and motivated the current development. A summary of the results is given in Table 3b.The cumulative incidence function (CIF) was derived for exponential and piecewise exponential models for the imputed data using the above regressions model and SAS procedure (PROC MI, with m=5, 10 and 100 imputations) and were compared to those based on the original data (without imputation).

In Table 3b, the group effects are reported for both data without imputation and with imputation for m=5, 20 and 100. For imputed data, we reported the p-values of group effect as mean, minimum, maximum, and median. A comparison of cumulative incidence rates can be found in Figure 4.

Table 3b. AF p-values for Group Effect using Likelihood Ratio Test

Procedure	Without Imputation	With Imputation (MI)			
		mean	median	min	max
Parameter-1					
m=5	0.0199	0.0125	0.0121	0.0098	0.0187
m=20		0.0180	0.0138	0.0098	0.0561
m=100		0.0160	0.0123	0.0051	0.0696

Parameter-2					
m=5	0.0118	0.0082	0.0077	0.0063	0.0126
m=20		0.0120	0.0091	0.0063	0.0397
m=100		0.0105	0.0080	0.0031	0.0487
Parameter-3					
m=5	0.0782	0.0564	0.0519	0.0469	0.0805
m=20		0.0743	0.0621	0.0469	0.1983
m=100		0.0656	0.0565	0.0206	0.2164

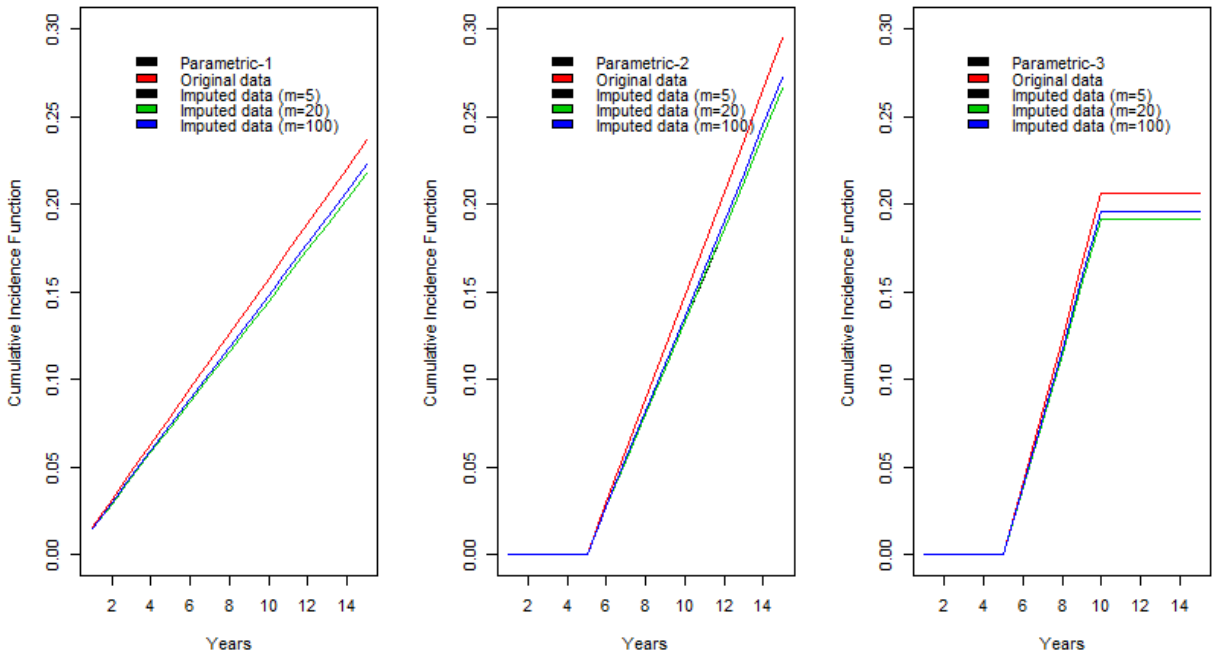

Figure 4: Cumulative Incidence Comparison for AF Based on Original Data and Imputed Data (Using the mean of intensity estimates) Corresponding to m =5, 20 and 100.

Note: The lines for m=5 and 20 are almost overlapped.

Simulation Study

To assess the performance of the imputation approach we conducted simulation studies as described below.

The primary focus is on assessing the performance of imputing *AF* in Anthracycline Cardiac Toxicity data for comparing the cumulative incidences of cardiac toxicity in the illness-death model as discussed above. The detail steps are described as follows.

Step 1: From Table 3a it is clear that log(*AF*) values are associated with the risk group (AR and NR), diagnosis group (Sarcomas vs. others), Age, FS, and BMI and using the equation (4.1) we first filled all the missing values of *AF* with the mean predicted values and obtained a complete copy of the data set.

Step 2: Then, for simulation studies we first created four subgroups:

Group 1: The patients which are in AR group diagnosed with Sarcomas (sample size n_1)

Group 2: The patients which are in AR group diagnosed with other cancers (sample size n_2)

Group 3: The patients which are in NR group diagnosed with Sarcomas (sample size n_3)

Group 4: The patients which are in NR group diagnosed with other cancers (sample size n_4)

Now to keep the covariance structure consistent with the observed data the sample mean and variance-covariance matrix were obtained for the variables LAF=log(*AF*), FS, Age, BMI and LTime=log(Time), where Time is the length of follow-up from diagnosis to the time of survey and the individual can be in one of the three states as shown in Figure 2.

Step 3: (Generate multi-normal data by group): Generate a sequence of random vectors, $(\text{LAF}, \text{FS}, \text{Age}, \text{BMI}, \text{LTime})_j, j = 1, \dots, n_i$, from multi-normal distribution for Group i $(i = 1, 2, 3, 4)$. That is, we assumed that,

$$(LAF, FS, Age, BMI, LTime)_j \sim i.i.d.\ \ MVN(v_i, \Sigma_i) \text{ for } j = 1, 2, \dots, n_i, i = 1, 2, \dots, 4$$

where $n_1 = n_2 = n/3$ and $n_3 = n_4 = n/6$ for $n = 180,\ 240,$ or 300 (the sample sizes for three simulation studies) and v_i and $\sum_i$ are the sample mean vector and covariance matrix. In this simulation study we used unequal samples size to reflect higher proportion of At-risk survivors in our cohort.

Note: We noticed the means of FS and LAF were significantly low (high) for Group 1 which would have led to highly significant p-values for comparing the two groups (AR vs. NR). Therefore, we adjusted the mean values for these two variables in the simulation studies as follows:

Mean of FS and LAF in Anthracycline Cardiac Toxicity data	
	Group

		1	2	3	4
Original	FS	0.307	0.342	0.359	0.358
	LAF	4.209	3.932	3.865	3.786
Adjusted	FS	**0.330**	0.342	0.359	0.358
	LAF	**4.000**	3.932	3.865	3.786

Step 4: (Incomplete Data in AF): From the sample size generated in Step 2, we randomly deleted R% (R=20 or 30) of AF values, and got incomplete data with sample sizes (100-R)%n.

Step 5: (Imputed data): Using SAS procedure PROC MI with FCS option we imputed AF values and obtained a complete copy of the data set.

Step 6 (Calculate p-value for group Effect): For the one parameter exponential distribution, the p-values for group effect (comparing AR with NR) were obtained using likelihood ratio test for complete (originally generated), incomplete and imputed data set.

Step 7: The imputation process (Step 5) was repeated 20 times to obtain 20 copies of complete data sets, which resulted in 20 p-values. A description of the p-values in terms of mean, median, minimum, and maximum is summarized in Table 4.

Step 8: Steps 2 – 7 were repeated 10 times to assess the performance of the imputation approach on 10 independently generated data sets. The results of the simulation study are summarized in Table 4.

From Table 4, it is seen that, in general, the median of the p-values is much closer to the p-value obtained from the complete data compared to those obtained from the incomplete data. However, there are some extreme situations where the results from the complete data and those obtained from imputations and incomplete data are not in agreement and this could be due to chance that more observations were deleted from a particular group and the regression is not able to completely exploit the underlying correlation structure. For example, for the 4th simulation, when the sample size is 180 and 30% observations are imputed the p-value for the complete data set is 0.007 but those corresponding to incomplete data and imputed data are 0.059 and 0.469, respectively. This clearly suggests that the manner in which data are generated and the observations are randomly deleted might have changed the underlying structure particularly those who are in the AR and NR

360 groups. However, for lager sample size (n=300) we see that in all simulations the median p-value
361 based on imputation is much closer to the p-value obtained from the complete data. Thus, it is clear
362 that with imputation approach we are able to exploit the underlying correlation structure and obtain
363 nearly unbiased conclusions.

364

Table 4. p-values for Group Effects using Likelihood Ratio Test

			Complete Data	Incomplete Data	Imputed Data (p-values)				
n	R%	Simulation #	p-value	p-value	Median	Min	Max	Mean	SD
180	20	1	<.001	<.001	<.001	<.001	0.011	0.001	0.003
		2	0.285	0.322	0.256	0.127	0.628	0.300	0.143
		3	0.004	0.019	0.006	0.002	0.014	0.006	0.003
		4	0.007	0.057	0.025	0.003	0.202	0.045	0.050
		5	0.003	0.003	0.004	<.001	0.025	0.007	0.008
		6	<.001	0.001	0.005	<.001	0.073	0.011	0.016
		7	0.013	0.031	0.020	0.002	0.088	0.027	0.027
		8	0.761	0.937	0.770	0.539	0.995	0.801	0.135
		9	0.007	0.001	<.001	<.001	0.100	0.008	0.024
		10	0.011	0.099	0.026	0.004	0.191	0.045	0.053
180	30	1	<.001	0.130	0.004	<.001	0.091	0.015	0.027
		2	0.285	0.002	0.035	0.005	0.130	0.046	0.037
		3	0.004	0.003	<.001	<.001	0.003	<.001	0.001
		4	0.007	0.059	0.469	0.115	0.889	0.493	0.230
		5	0.003	0.033	0.314	0.022	0.935	0.330	0.253
		6	<.001	0.040	0.001	<.001	0.060	0.007	0.015
		7	0.013	0.262	0.008	<.001	0.090	0.020	0.026
		8	0.761	0.003	0.035	0.003	0.150	0.050	0.044
		9	0.007	0.005	<.001	<.001	0.032	0.005	0.009
		10	0.011	0.250	0.140	0.011	0.372	0.140	0.100
240	20	1	<.001	<.001	<.001	<.001	0.003	<.001	0.001
		2	0.087	0.075	0.087	0.011	0.248	0.094	0.072
		3	0.020	0.041	0.040	0.003	0.117	0.043	0.026
		4	0.011	0.005	0.007	<.001	0.086	0.017	0.022
		5	0.002	0.016	0.009	0.001	0.74	0.025	0.060
		6	<.001	<.001	<.001	<.001	0.001	<.001	<.001
		7	0.001	0.001	0.001	<.001	0.008	0.001	0.002
		8	0.221	0.390	0.347	0.109	0.933	0.388	0.197
		9	0.001	<.001	<.001	<.001	0.003	0.001	0.001
		10	<.001	0.001	<.001	<.001	0.033	0.002	0.007
		1	<.001	<.001	<.001	<.001	<.001	<.001	<.001

240	30	2	0.087	0.327	0.105	0.102	0.499	0.154	0.141
		3	0.020	0.060	0.023	0.004	0.135	0.037	0.038
		4	0.011	0.028	0.011	0.001	0.140	0.027	0.033
		5	0.002	0.014	0.008	<.001	0.042	0.011	0.013
		6	<.001	<.001	<.001	<.001	0.004	0.001	0.001
		7	0.001	0.002	0.001	<.001	0.081	0.011	0.022
		8	0.221	0.121	0.041	0.004	0.201	0.051	0.047
		9	0.001	0.019	0.012	0.001	0.079	0.022	0.026
		10	<.001	0.005	0.015	<.001	0.159	0.041	0.054
300	20	1	<.001	<.001	<.001	<.001	0.001	<.001	<.001
		2	0.268	0.107	0.155	0.023	0.809	0.212	0.184
		3	0.026	0.011	0.014	0.001	0.077	0.022	0.022
		4	<.001	<.001	<.001	<.001	0.003	<.001	0.001
		5	<.001	0.001	0.001	<.001	0.005	0.001	0.002
		6	<.001	0.028	0.007	0.001	0.056	0.014	0.015
		7	0.001	0.003	0.002	<.001	0.016	0.004	0.005
		8	0.107	0.193	0.118	0.031	0.370	0.148	0.098
		9	<.001	0.001	0.002	<.001	0.017	0.004	0.005
		10	0.001	0.013	0.007	0.001	0.336	0.047	0.090
300	30	1	<.001	<.001	<.001	<.001	<.001	<.001	<.001
		2	0.268	0.759	0.383	0.078	0.889	0.389	0.231
		3	0.026	0.024	0.019	0.003	0.103	0.032	0.026
		4	<.001	0.003	<.001	<.001	0.007	0.001	0.002
		5	<.001	0.009	0.001	<.001	0.056	0.006	0.013
		6	<.001	0.015	0.003	<.001	0.042	0.009	0.014
		7	0.001	0.015	0.003	<.01	0.022	0.002	0.006
		8	0.107	0.169	0.110	0.007	0.620	0.142	0.142
		9	<.001	<.001	<.001	<.001	0.003	<.001	0.001
		10	0.001	0.002	0.003	<.001	0.114	0.015	0.027

Discussion

In this paper, we have employed a well-established methodology of illness-death/Failure model[2] and imputed the missing observations for a phase IV clinical trial study as an example. Although, we assumed a very simple parametric model, it is straightforward to expand to other parametric or semi-parametric models. From a clinician's point of view the simple approaches such as log rank test and KM survival curves[34] are most commonly used and understood. However, in the setting of interval censored data the approach proposed here is simple to use and can be implemented easily to estimate fixed-time cumulative incidence function with or without imputation[2].

When studying the long-term effects of treatment, there can be multiple unwanted events identified at the time of observation. Some of these events can be competing and others are not correlated. This leads to multivariate time-to-event data. One simple approach is to study the incidence of first event and then incidence of specific event. In our example, cardiotoxicity measures included abnormal *AF* and *FS* but there are some other measures to evaluate cardiotoxicity. For some reason, not all patients had both measures and the models based on bi-variate time-to-event outcomes would include only those patients who have data on both outcomes and this would reduce the sample size and potentially ignore important information. Based on this consideration, the multivariable regression method was used to impute the missing observations and to apply the parametric method to the imputed data and compare the results with those obtained without imputation.

As stated before, the problem of evaluating possible toxic effects of cancer therapies in a Phase IV trial setting is an important problem. Among the many issues in such studies, missing data is a key aspect that can influence the inferences. It is also important to understand the nature of impact of missing data on the analysis and the interpretation of the study data. Also note that imputation increases the sample size, and thus increases statistical power to detect the same effect size. But if the model assumptions are not correct, the inference may not be valid. Thus, it is recommended to report the p-values with and without imputation. However, with higher absolute correlations between two outcome measures (the primary outcome measure, AF, with higher missing and the secondary outcome measure, FS, with little or no missing), produced efficient results, a rigorous simulation study with different amount of correlations between two outcome measures, amount of missing and model uncertainty is underway to consider this aspect and will be reported elsewhere; this is along the lines our work for a randomized clinical study[33].

The study involved all the patients visiting the clinic in a pre-specified time frame (such as 1 year of accrual) and represents a somewhat unbiased survey of patients. Since the outcome measure may depend on disease type, an almost equal allocation was used to enroll patients. It has been reported in Hudson et al. [1] that the prevalence depends on disease type; hence it will be another research direction to adjust the sampling allocation and variability due to sampling and modeling for generalizing the results for the entire patient population[35].

Another limitation of this study is that this is a cross-sectional survey to estimate the long-term effect of cardiotoxicity of the primary treatment of cancer. Dodge[36] defined cross-sectional survey "A method of data collection whereby a battery of questions is asked of participation at one single point or in a relatively small interval of time. Inferences about a population must be anchored to the time period in which the sample was taken. Data from cross-sectional surveys are typically unable to be used to prove the existence of cause-and-effect relationships." Even though this is based on enrolling consecutive eligible patients in a very homogeneous environment (St. Jude Children's Hospital treats patients without charge to patients), effect of this limitation is minimized but cannot be reduced to zero. Generalization of the results to a general population should be done with caution.

Conclusions

Based on simulation output, it is suggested that the results obtained by imputing the missing values using regression methods are significantly more efficient than those obtained without imputation. It is recommended to carefully impute missing values in certain dataset in order to retain most of the information from the incomplete cases.

List of Abbreviations

AR: At-Risk

NR: Not At-Risk

FS: Fractional shortening

AF: Afterload

LVEdD: Left ventricular end-diastolic diameter

LVEsD: Left ventricular end-systolic diameter

BMI: Body mass index

CIF: Cumulative incidence function

Declarations

Ethics approval and consent to participate

The study was approved by the institutional review boards at St Jude Children's Research Hospital and Stanford University (Stanford, CA).

All study participants or their parents provided informed consent.

All methods were carried out in accordance with relevant guidelines and regulations.

Consent to publish

Not applicable

Availability of data and materials

The datasets generated and/or analyzed during the current study are not publicly available due to privacy restrictions.

Competing interests

The authors declare that they have no competing interests.

Funding

D.K. Srivastava was in part supported by the Grant CA21765 from the National Institutes of Health (NIH) and by the American Lebanese Syrian Associated Charities (ALSAC).

C. Qian was supported by the University of Louisville Fellowship.

S. N. Rai was partly supported with Wendell Cherry Chair in Clinical Trial Research Fund, multiple National Institutes of Health (NIH) grants (5P20GM113226, PI: McClain; 1P42ES023716, PI: Srivastava; 5P30GM127607-02, PI: Jones; 1P20GM125504-01, PI: Lamont; 2U54HL120163, PI: Bhatnagar/Robertson; 1P20GM135004, PI: Yan; 1R35ES0238373-01, PI: Cave; 1R01ES029846, PI: Bhatnagar; 1R01ES027778-01A1, PI: States;), and Kentucky Council on Postsecondary Education grant (PON2 415 1900002934, PI: Chesney).

Author's contributions

Conception: All Authors

Data analysis: JP

Original Draft: All Authors

Review & Editing: All Authors

Approval of the Final Version: All Authors

Acknowledgements

We thank Dr. Donald Miller for his support for this research work.

References

1. Hudson MM, Rai SN, Nunez C, Merchant TE, Marina NM, Zalamea N, Cox C, Phipps S, Pompeu R and Rosenthal D. Noninvasive evaluation of late anthracycline cardiac toxicity in childhood cancer survivors. *Journal of Clinical Oncology* 2007; **25**: 3635-3643.

2. Rai SN, Pan J, Sun J, Hudson MM and Srivastava Dk. Estimating incidence rate on current status data with application to a Phase IV cancer trial. *Communications in Statistics: Theory and Methods* 2013; **42**:2417-2433.

3. Armstrong GT, Chen Y, Yasui Y, et al. Reduction in late mortality among 5-year survivors of childhoodcancer. N Engl J Med 2016;**374**(9):833–42.

4. Howlader N, Noone AM, Krapcho M, et al. SEER Cancer Statistics Review (CSR), 1975-2012. Bethesda,MD: National Cancer Institute (internet). 2015. http://seer.cancer.gov/csr/1975_2012/ (accessed November 18, 2020).

5. Phillips SM, Padgett LS, Leisenring WM, et al. Survivors of childhood cancer in the United States: prevalence and burden of morbidity. Cancer Epidemiol Biomarkers Prev 2015;**24**(4):653–63.

6. Fidler MM, Reulen RC, Winter DL, et al. Long term cause specific mortality among 34 489 five year survivors of childhood cancer in Great Britain: population based cohort study. BMJ 2016;354: i4351.

7. Robison LL, Hudson MM. Survivors of childhood and adolescent cancer: life-long risks and responsibilities. Nat Rev Cancer 2014;**14**(1):61–70.

8. Krischer JP, Epstein S, Cuthbertson DD, et al. Clinical cardiotoxicity following anthracycline treatment for childhood cancer: The Pediatric Oncology Group experience. *Journal of Clinical Oncology* 1997; **15**: 1544-1552.

9. Sorensen K, Levitt GA, Bull C, et al. Late anthracycline cardiotoxicity after childhood cancer: A perspective longitudinal study. *Cancer* 2003; **97**: 1991-1998.

10. Pein F, Sakiroglu O, Dahan M, et al. Cardiac abnormalities 15 years and more after adriamycin therapy in 229 childhood survivors of a solid tumor at the Institut Gustave Roussy. British *Journal of Cancer* 2004; **91**: 37-44.

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

12. Ayer M, Brunk HD, Ewing, GM, Reid WT, and Silverman E. An empirical distribution function for sampling with incomplete information. *Ann. Math. Statist*.1955; **26**: 641-647.

13. Van Eeden C. Maximum likelihood estimation of ordered probabilities. *Indagationes Mathematicae* 1956; **18**: 444-455.

14. Peto R. Experimental survival curves for interval-censored data. *Appl. Statist*. 1973; **22**: 86-91.

15. Turnbull BW. The empirical distribution function with arbitrarily grouped, censored and truncated data. *J. R. Statist*. Soc. B 1976; **38**: 290-295.

16. Little RJ, Rubin DB. *Statistical Analysis with Missing Data*, 2nd edition, New York: John Wiley, 2002.

17. Burren Sv. *Flexible Imputation of Missing Data*, 2nd edition, Chapman & Hall/CRC, Florida, 2012.

18. Rubin DB. *Multiple Imputation for Nonresponse in Surveys.* J. Wiley & Sons, New York, 1987.

19. Rubin DB. Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association,* **91**: 473-489

20. Rubin DB, Stern H, and Vehovar V. (1995), Handing "don't know" survey responses: The case of the Slovenian plebiscite. *Journal of the American Statistical Association* 1995; **90**: 822-828.

21. King G, Honaker J, Joseph A, and Scheve K. Analyzing incomplete political science data: an alternative algorithm for multiple imputation. *The American Political Science Review* 2001; **95**: 49–69.

22. Schafer JL. *Analysis of Incomplete Multivariate Data.* Chapman & Hall, London, 1997.

23. Liu C. Missing data imputation using the multivariate t distribution. *Journal of Multivariate Analysis* 1995; **53**: 139-158.

24. Buuren Sv, Boshuizen HC and Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine* 1999; **18**: 681-694.

25. Troxel A, Guoguang M and Heitjan DF. An index of local sensitivity to nonignorability. *Statistica Sinica* 2004; **14**: 1221-1237.

26. Cohen, J. *Statistical power analysis for the behavioral sciences*, 2nd edition. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.

27. Henry WL, Gardin JM and Ware JH. Echocardiographic measurements in normal subjects from infancy to old age. *Circulation* 1980; 62:1054-1061.

28. Colan SD, Parness IA, Spevak PJ, et al. Developmental modulation of myocardial mechanics: Age- and growth-related alterations in afterload and contractility. *Journal of the American College of Cardiology*, 1992; 19:619-629.

29. Molenberghs G and Fitzmoaurice G. Incomplete data: Introduction and overview. In *Handbooks of Modern Statistical Methods; Longitudinal Data Analysis*, (Chapter 17, pages 395-408), Chapman & Hall/CRC, Florida, 2009.

30. Kenward MG and Carpenter JR. Multiple Imputation. In In *Handbooks of Modern Statistical Methods; Longitudinal Data Analysis*, (Chapter 21, pages 477-500), Chapman & Hall/CRC, Florida, 2009.

31. Huque, M.H., Carlin, J.B., Simpson, J.A. *et al.* A comparison of multiple imputation methods for missing data in longitudinal studies. *BMC Med Res Methodol* **18,** 168 (2018).

32. SAS/STAT®14.1 User's Guide: The MI Procedure, SAS Institute Inc., Cary, NC, USA, 2015.

33. Liu Y, and De A. Multiple imputations by fully conditional specification for dealing with missing data in a large epidemiologic study. *International Journal of Statistics in Medical Research*, 4(3): 287-295, 2015.

34. Kaplan EL and Meier P. Nonparametric estimation from incomplete observation. *Journal of the American Statistical Association* 1958; **53**: 457-481.

35. Kovacevic MS and Rai SN. Log-Linear Modeling of Change Using Longitudinal Survey Data. *Communications in Statistics—Theory and Methods* 2002; **31**: 1815-1835.

36. Dodge Y. *The Oxford Dictionary of Statistical Terms.* The International Statistical Institute, Oxford Press, 2003

Appendix

A1: Details of the Likelihood Approach

Due to the actual time of onset of abnormality, U, is not known, the observed quantity for each patient includes the observation time, T, and two indicators of status, δ and γ, at the time of survey, where δ is an indicator of patient alive with no cardiac failure or dead/cardiac failure, and γ is an indicator of patient with a normal or abnormal value. Let t_i be the observation time (death, cardiac failure or survey) for the $i^{t\square}$ subject. That is,

$$\delta_i = \begin{cases} 1, & \text{if unit } i \text{ dead or cardiac failure at time } t_i \\ 0, & \text{if unit } i \text{ alive and no cardiac failure at time } t_i, \end{cases}$$

and

$$\gamma_i = \begin{cases} 1, & \text{if unit } i \text{ with abnormal value at time } t_i \\ 0, & \text{if unit } i \text{ with normal value at time } t_i. \end{cases}$$

The simplified forms of intensities $\lambda_i(t) = \lambda_i$ for $i = 1$ or 2 and $\lambda_3(t|u) = \lambda_3$ lead to $Q_i(t) = e^{-\lambda_i t}$ for $i = 1$ or $2, Q_3(t|u) = e^{-\lambda_3(t-u)}$ and $Q(t) = e^{-(\lambda_1+\lambda_2)t}$. We derive the corresponding likelihood contributions from $L_1(t)$ to $L_4(t)$ for the four observation types in Table 2 and then the log-likelihood function as follows

$$\begin{aligned} l(\lambda_1, \lambda_2, \lambda_3) = & \sum_{i=1}^{n} a_i [log\, \lambda_2 - (\lambda_1 + \lambda_2) t_i] - \sum_{i=1}^{n} b_i (\lambda_1 + \lambda_2) t_i] \\ & + \sum_{i=1}^{n} c_i [log\, \lambda_1 + log\, \lambda_3 - log(\lambda_1 + \lambda_2 - \lambda_3) + log(e^{-\lambda_3 t_i} - e^{-(\lambda_1+\lambda_2)t_i}] \\ & + \sum_{i=1}^{n} d_i [log\, \lambda_1 - log(\lambda_1 + \lambda_2 - \lambda_3) + log(e^{-\lambda_3 t_i} - e^{-(\lambda_1+\lambda_2)t_i}], \end{aligned}$$

where $a_i = \delta_i(1-\gamma_i), \mathrm{b}_i = (1-\delta_i)(1-\gamma_i), c_i = \delta_i\gamma_i$ and $\mathrm{d}_i = (1-\delta_i)\gamma_i$ are the indicators corresponding to observation type 1 to type 4 in Table 2. Then the maximum likelihood estimators $\hat{\lambda}_1, \hat{\lambda}_2$ and $\hat{\lambda}_3$ of λ_1, λ_2 and λ_3 are derived from the following equations[2].

$$\frac{D_3 + D_4}{\lambda_1} - \frac{D_3 + D_4}{\lambda_1 + \lambda_2 - \lambda_3} - T_2 + \sum_{i=1}^{n} \frac{(c_i + d_i) t_i}{e^{(\lambda_1+\lambda_2-\lambda_3)t_i} - 1} = 0$$

$$\lambda_2 = \frac{D_1}{D_3 + D_4} \lambda_1$$

$$\lambda_3 = \frac{D_3}{T_1 \lambda_1 - D_3 - D_4} \lambda_1$$

where

$$D_1 = \sum_{i=1}^{n} a_i\,, D_2 = \sum_{i=1}^{n} b_i, D_3 = \sum_{i=1}^{n} c_i\,, D_4 = \sum_{i=1}^{n} d_i$$

$$T_1 = \sum_{i=1}^{n} (a_i + b_i + c_i + d_i)\, t_i \text{ and } T_2 = \sum_{i=1}^{n} (a_i + b_i)\, t_i.$$

It is further extended to the model to allow the intensity λ_1 with piecewise constant[2]. Assume two intervals: less than t_c years and above t_c (includingt_c) years (say, $t_c = 5$) and let these two rates be λ_{11} and λ_{12}. Then the log-likelihood function is derived as

$$l(\lambda_{11}, \lambda_{12}) = \sum_{i=1}^{n} [b_i \, log \, L_2(t_i) + d_i \, log \, L_4(t_i)]$$

$$= \sum_{i \in S_1} [-b_i t_i \lambda_{11} + d_i \, log(1 - e^{-t_i \lambda_{11}})] + \sum_{i \in S_2} \{-b_i [t_c \lambda_{11} + (t_i - t_c)\lambda_{12}] + d_i \, log(1 - e^{-t_c \lambda_{11} - (t_i - t_c)\lambda_{12}})\},$$

for a special case with no deaths/cardiac failures, that is, $a_i = c_i = 0$, $\lambda_2 = \lambda_3 = 0$, where $S_1 = \{i: \mathrm{t}_i < t_c\}$ and $S_2 = \{i: \mathrm{t}_i \geq t_c\}$. Hence, the estimates of λ_{11} and λ_{12} can be derived easily from following score equations

$$\sum_{i \in S_1} b_i t_i - \sum_{i \in S_1} \frac{d_i t_i}{e^{\lambda_{11} t_i} - 1} + \sum_{i \in S_2} b_i t_c - \sum_{i \in S_1} \frac{d_i t_c}{e^{\lambda_{11} t_c + (t_i - t_c)\lambda_{12}} - 1} = 0$$

$$\sum_{i \in S_2} b_i (t_i - t_c) - \sum_{i \in S_2} \frac{d_i (t_i - t_c)}{e^{\lambda_{11} t_c + (t_i - t_c)\lambda_{12}} - 1} = 0.$$

For a general model with piecewise constants in parameterλ_1, the log-likelihood function is

$$l(\lambda_{11}, \lambda_{12}, \lambda_2, \lambda_3) = \sum_{i=1}^{n} [a_i \, log \, L_1(t_i) + b_i \, log \, L_2(t_i) + c_i \, log \, L_3(t_i) + d_i \, log \, L_4(t_i)]$$

,

where $L_1(t) = \lambda_2 Q(t)$, $L_2(t) = Q(t)$, $L_4(t) = L_3(t)/\lambda_3$ and if $t < t_c$,

$$L_3(t) = \frac{\lambda_{11}\lambda_3}{\lambda_{11} + \lambda_2 - \lambda_3} e^{-\lambda_3 t} (1 - e^{-(\lambda_{11} + \lambda_2 - \lambda_3)t})$$

if $t \geq t_c$,

$$L_3(t) = \frac{\lambda_{11}\lambda_3}{\lambda_{11} + \lambda_2 - \lambda_3} e^{-\lambda_3 t_c} (1 - e^{-(\lambda_{11} + \lambda_2 - \lambda_3)t_c})$$

$$+ \frac{\lambda_{12}\lambda_3}{\lambda_{12} + \lambda_2 - \lambda_3} e^{-(\lambda_{11} - \lambda_{12})t_c - \lambda_3 t} (e^{-(\lambda_{12} + \lambda_2 - \lambda_3)t_c} - e^{-(\lambda_{12} + \lambda_2 - \lambda_3)t}),$$

in which $Q_1(t) = e^{-\lambda_{11} t}$ if $t < t_c$, and $Q_1(t) = e^{-(\lambda_{11} - \lambda_{12})t_c - \lambda_{12} t}$ if $t \geq t_c$; $Q_2(t) = e^{-\lambda_2 t}$, $Q_3(t|u) = e^{-\lambda_3 (t-u)}$ and $Q(t) = Q_1(t) Q_2(t)$. Based on the log-likelihood equation, the maximum likelihood estimates of $\lambda_{11}, \lambda_{12}, \lambda_2$ and λ_3 can be computed from the score equations using numerical method. It is similar to derive the likelihood function for exponential model if λ_1 has three or more pieces.

Figures

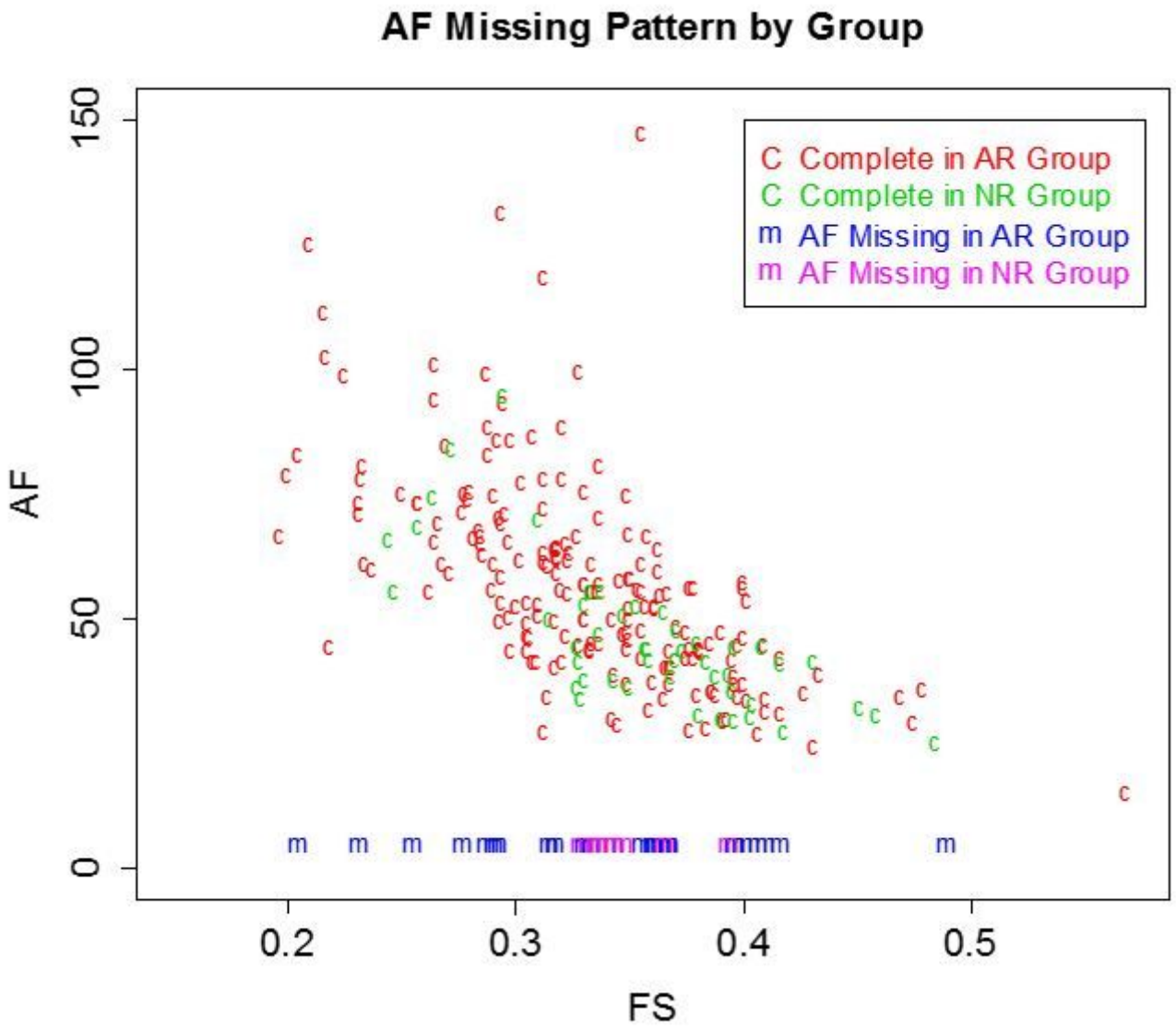

Figure 1

Scatter plot of AF and FS within AR and NR group of patients. Complete data (labeled c) displays strong correlation between AF and FS. Missing values of AF are in almost entire range of FS values (labeled m).

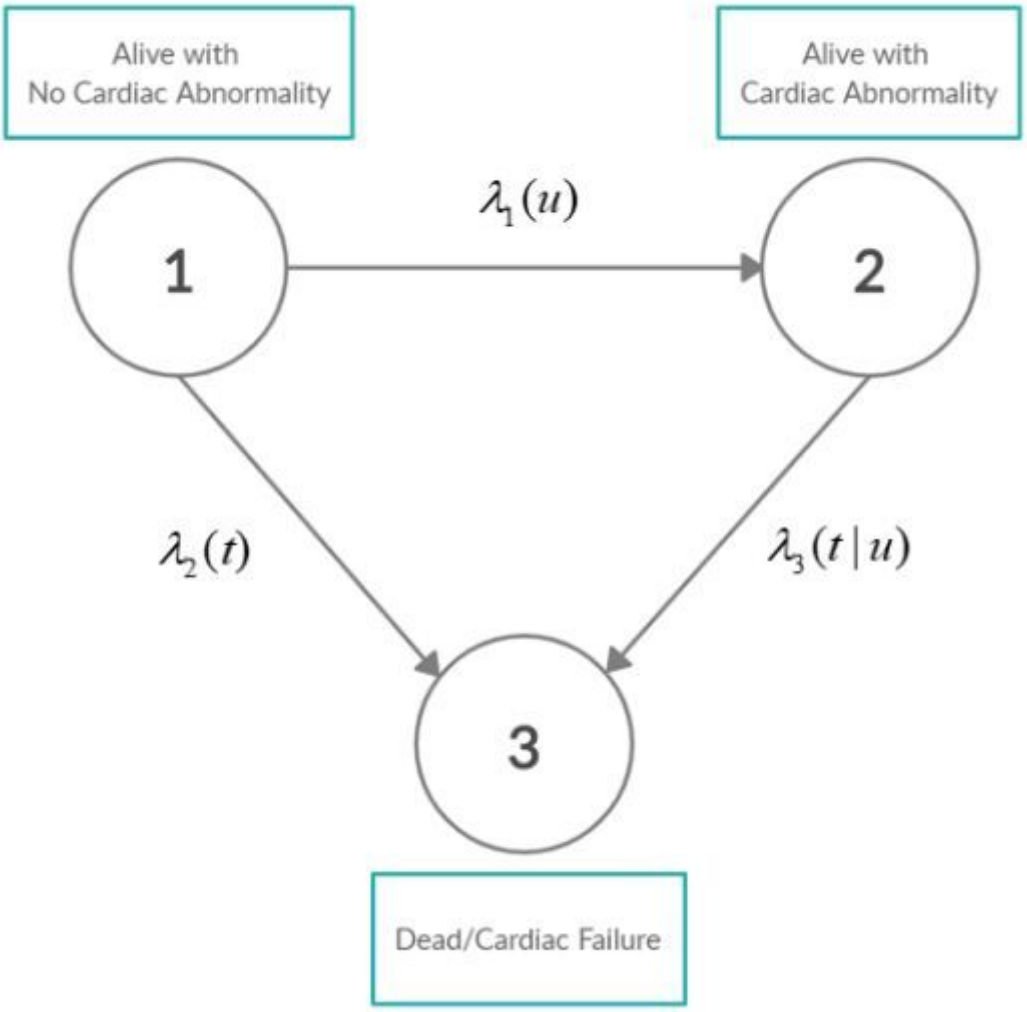

Figure 2

An abnormal cardiac measure-death/cardiac failure model involving three states. State 1 corresponds to patients who are alive with normal value. Patients who are alive with abnormal value are in state 2. State 3 is an absorbing state and corresponds to death or cardiac failure.

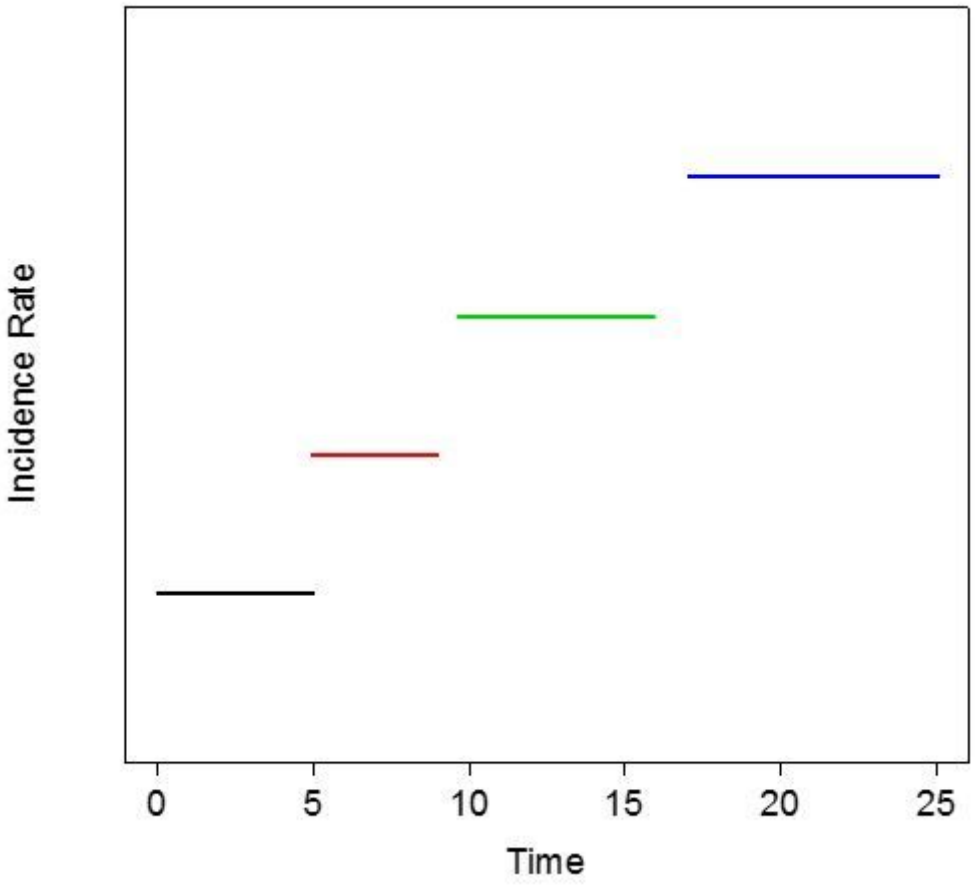

Figure 3

Piecewise Incidence Rate

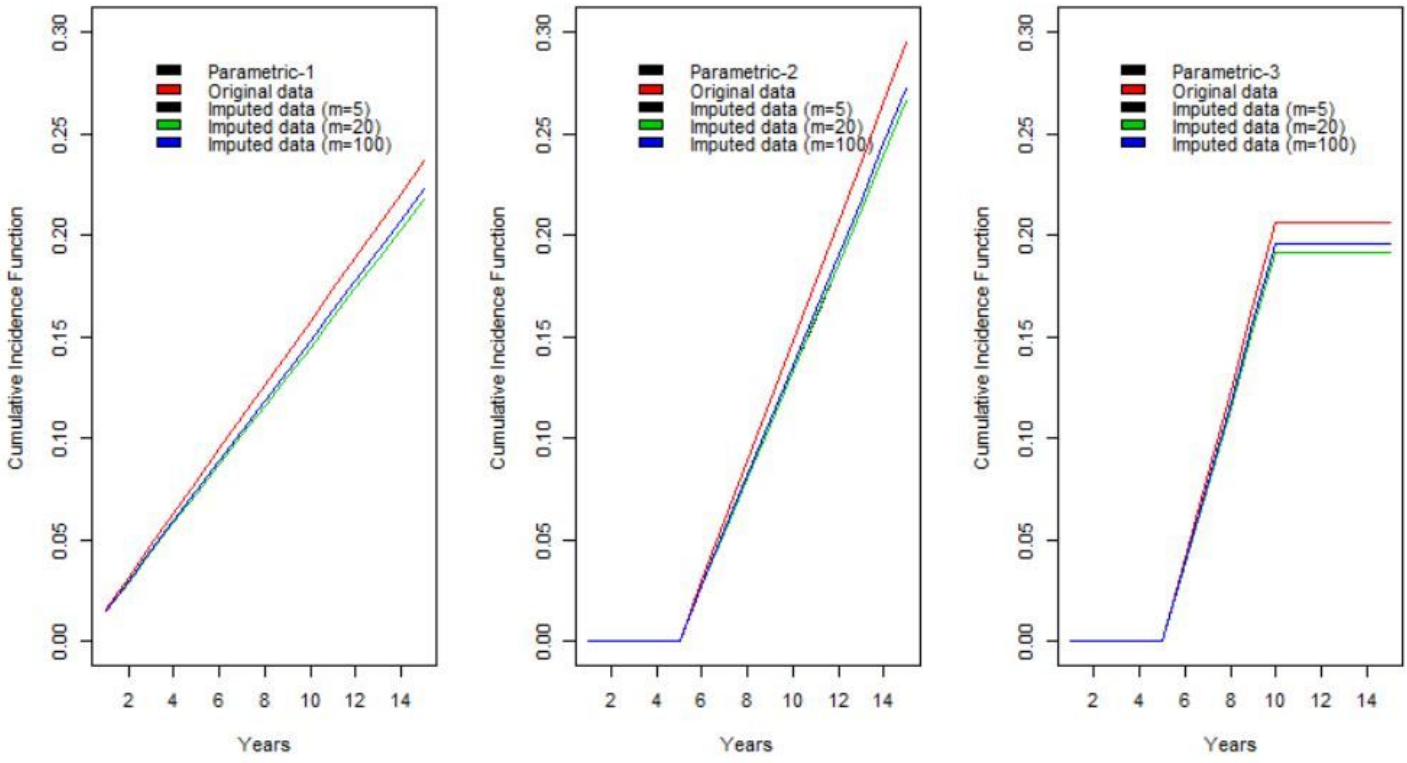

Figure 4

Cumulative Incidence Comparison for AF Based on Original Data and Imputed Data (Using the mean of intensity estimates) Corresponding to m =5, 20 and 100. Note: The lines for m=5 and 20 are almost overlapped.