A COMPARISON OF KAPLAN-MEIER AND CUMULATIVE INCIDENCE ESTIMATE IN THE PRESENCE OR ABSENCE OF COMPETING RISKS IN BREAST CANCER DATA

by

Bintu N. Sherif

B.S., University of Pittsburgh, 2004

Submitted to the Graduate Faculty of

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2007

UNIVERSITY OF PITTSBURGH

Graduate School of Public Health

This thesis was presented

by

Bintu N. Sherif

It was defended on

December 14, 2007

and approved by

Vincent C. Arena, PhD, Associate Professor, Department of Biostatistics, Graduate School Public Health , University of Pittsburgh

Christine E. Ley, PhD,MPH,MSW, Associate Director, Behavioral and Community Health Sciences, Graduate School Public Health, University of Pittsburgh

John W. Wilson, PhD, Assistant Professor, Department of Biostatistics, Graduate School Public Health , University of Pittsburgh

Jong-Hyeon Jeong, PhD, Associate Professor , Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh
Thesis Advisor

A COMPARISON OF KAPLAN-MEIER AND CUMULATIVE INCIDENCE ESTIMATE IN THE PRESENCE OR ABSENCE OF COMPETING RISKS IN BREAST CANCER DATA

Bintu N. Sherif, M.S.

University of Pittsburgh, 2007

Statistical techniques such as Kaplan-Meier estimate is commonly used and interpreted as the probability of failure in time-to-event data. When used on biomedical survival data, patients who fail from unrelated or other causes (competing events) are often treated as censored observations.

This paper reviews and compares two methods of estimating cumulative probability of cause-specific events in the present of other competing events: 1 minus Kaplan-Meier and cumulative incidence estimators. A subset of a breast cancer data with three competing events: recurrence, second primary cancers, and death, was used to illustrate the different estimates given by 1 minus Kaplan-Meier and cumulative incidence function. Recurrence of breast cancer was the event of interest and second primary cancers and deaths were competing risks.

The results indicate that the cumulative incidences gives an appropriate estimates and 1 minus Kaplan-Meier overestimates the cumulative probability of cause-specific failure in the presence of competing events. In absence of competing events, the 1 minus Kaplan-Meier approach yields identical estimates as the cumulative incidence function.

The public health relevance of this paper is to help researchers understand that competing events affect the cumulative probability of cause-specific events. Researchers should use methods such as the cumulative incidence function to correctly estimate and compare the cause-specific cumulative probabilities.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would first and foremost like to thank my thesis advisor Dr. Jong-Hyeon Jeong, for all the help and guidance he has provided during the work on this thesis. He has not only assisted me with his excellent advice, but he also elevated my interest in survival analysis, introduced me to a new statistical software package and improved my comprehension of theoretical statistics.

Second, I would like to thank my thesis committee members, Dr. Vincent Arena, Dr. Christine Ley and Dr. John Wilson. Thanks to Dr. Arena and Dr. Wilson for contributing greatly to my learning of biostatistics in my graduate study, and willingness to consult with me about problems encountered. Thanks to Dr. Ley for helping me understand different aspects of community health science theories.

I am also grateful to everyone at the department for adding some pleasure to the work. I would especially like to thank Phyllis Fisher who always listened and provided encouraging words.

Finally, I would like to thank my parents, Seku and Hawa, my sisters, Maryam, and Fatima, brother, Malik, as well as my friends Genevieve and Solomon for their unconditional love and support. I am blessed and proud to call them my family.

1.0 INTRODUCTION

Statistical techniques such as Kaplan-Meier product limit estimate (Kaplan and Meier 1958), which take into account censored data, are primarily used in the medical and biological sciences for estimating the probability of failure in time-to-event data “survival data”. The term “survival data” is widely used to describe data involving time to the occurrence of an event. Events may be death, the appearance of a cancerous tumor, the development of some disease, recurrence of a disease, cessation of smoking, conception, and so forth. We have also seen survival analysis widely been used in the social sciences, where interest is on analyzing time to events such as job changes, marriage, birth of children and so forth. The engineering sciences have also contributed to the development of survival analysis which is called "reliability analysis" or "failure time analysis" in this field, since the main focus is in modeling the time it takes for machines or electronic components to break down. The developments from these diverse fields have for the most part been consolidated into the field of "survival analysis" (Allison, 1984). In the past decades, applications of the statistical methods for survival data analysis have been extended beyond biomedical and reliability research to other fields, for example, felons’ time to parole (criminology), length of newspaper or magazine subscription (marketing), workmen’s compensation claims (insurance), health insurance practice, business and economics. The study of survival data has previously focused on predicting the probability of response, survival, or mean lifetime, and comparing the survival distributions of experimental animals or of human patients. In recent years, the identification of risk and/or prognostic factors related to response, survival, and the development of a disease has become equally important (Lee (1992) Ch. 1).

The analysis of survival data can be complicated by issues of censoring. In biomedical data, censoring arises when an individual's life length is only partially known in a certain period of time. Types of censoring includes right censoring- where the event occurs after the follow-up time, left censoring- where the event time occurred before the observation time, or interval censoring, where observation is not continual, but occurs at discrete times. Only the times between which the event occurred is known. Censored observations are contributed not only by losses to follow-up but also by deaths from other causes and sometimes by other events if they preclude development of the endpoint under consideration (Pepe, 1991). For example, in a study of the disease-free survival in lymph node-negative breast carcinomas (Kuru et al study, 2003) patients with pathologically proven breast carcinoma and with negative axillary lymph nodes, who had been operated on for primary breast cancer, were followed-up for 60 months. The primary event of interest was death due to breast carcinoma. Patients who died from causes other than breast carcinoma were treated as censored observations. Many other studies tend to use the same type of approach; including Martinez (2007) in which the primary event of interest was AIDS related deaths (if the primary cause of death was an AIDS-defining condition) and death by other causes were censored.

Ideally, the survival period is determined by following a group of patients until each of them has been reviewed for a set period of time or until an event has occurred. Emerging evidence now suggests that in the presence of competing risks, which will be further discussed, the cumulative incidence function, a method which takes into account competing risks occurrence, is the appropriate method use to estimate the probability of occurrence of the event of interest in the presence of other events. However, researchers often use the Kaplan Meier approach (1-KM) to evaluate the survival probability of occurrence of a cause-specific endpoint, even if the appropriate data contain competing-risk events (Gooley, Leisenring et al. 1999). In the clinical oncology and epidemiology literature it is still quite common to see this probability incorrectly estimated as the 1 - KM estimator (Gaynor et al., 1993). This could result in an over-estimation of the cumulative probability of cause-specific failure.

There can be different types of failure in a time-to-event analysis under competing risks. For illustration purposes I will make the same assumption as Gooley et al (1999), that is, the existence of two failure types; events of interest and all other events. This paper evaluates the advantages and statistical appropriateness of using the cumulative incidence estimate over the Kaplan Meier estimates (1-KM) method in biomedical survival analysis under right censoring. The introduction and background are presented in Section 1. Section 2 reviews the hazard function estimate, commonly used the Kaplan Meier approach and the cumulative incidence estimate, as well as the definition of competing risks. Section 3 contains the description of a breast cancer dataset, used for comparison and illustrates the difference between cumulative incidence estimate and the 1 minus Kaplan Meier estimate. Section 4 contains the Statistical methods. Numerical results of comparing the two types of estimates are provided in Section 5. Section 6 is a discussion of the results, limitations, suggestions for possible future application of this method, and suggested modifications of this method to fit different types of competing risks.

2.0 ESTIMATORS OF CUMULATIVE PROPOTION UNDER COMPETING RISKS

2.1 THE HAZARD FUNCTION ESTIMATE

A central quantity in survival analysis is the hazard function, (also known as the failure rate, hazard rate, or force of mortality)

It is defined by:

$$\lambda(t) = \lim_{\delta t \to 0} \frac{P[t \le T \le t + \delta t \,|\, T \ge t\rangle}{\delta t} = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)}$$

Where *f(t), F*(*t*) and *S*(*t*) are the probability density function (p.d.f), the cumulative distribution function and the survivor function of *T* , respectively.

Thus, *λ (t)* δ*t* can be seen as the conditional probability that the event of interest occurs in the interval [*t*, *t* + δ *t*), given that it has not occurred before time *t*. It is clear that the hazard function is finite and nonnegative. This function is particularly useful in determining the appropriate failure distributions when competing risks are present.

A quantity related to the hazard function is the probability of an individual surviving beyond time *t*, the *survival function*. The survival function, *S*(*t*) , is the exponential of the negative of the cumulative hazard function, i.e.

$$S(t) = \exp(-H(t)) = \exp\left(-\int_0^t \lambda(u)du\right)$$

where $H(t)$ is the cumulative hazard function.

2.2 KAPLAN MEIER ESTIMATE

The standard nonparametric estimator of the survival function is the Kaplan-Meier estimate. The Kaplan–Meier, or product limit estimator, first derived by Kaplan and Meier (1958), estimates the survival probability beyond time t in right-censored data. It is very often useful to summarize the survival experience, in particular groups of patients in terms of the empirical survival function S(t):

k distinct event times $t_1 < t_j < \ldots < t_k$

at each event time t_j there are n_j individuals at risk

d_j is the number of subjects who have the event at the time

$$\hat{S}(t) = \prod_{j:t_j \leq t} [1 - \frac{d_j}{n_j}]$$

When there are no censored observations, the Kaplan Meier estimator is simple and intuitive, as the proportion of failures times > t. When there is censoring, the Kaplan Meier provides an estimate of *S(t)* that takes censoring into account. The Kaplan-Meier estimator is a step function with jumps at the observed event times. The pattern of these jumps depends not only on the number of events observed at each event time t_j, but also on the pattern of the censored observations prior to t_i. Figure 1 shows an example of a Kaplan-Meier survival curve for breast cancer patients assigned to tamoxifen treatment. Details of this study will be described in section 3.

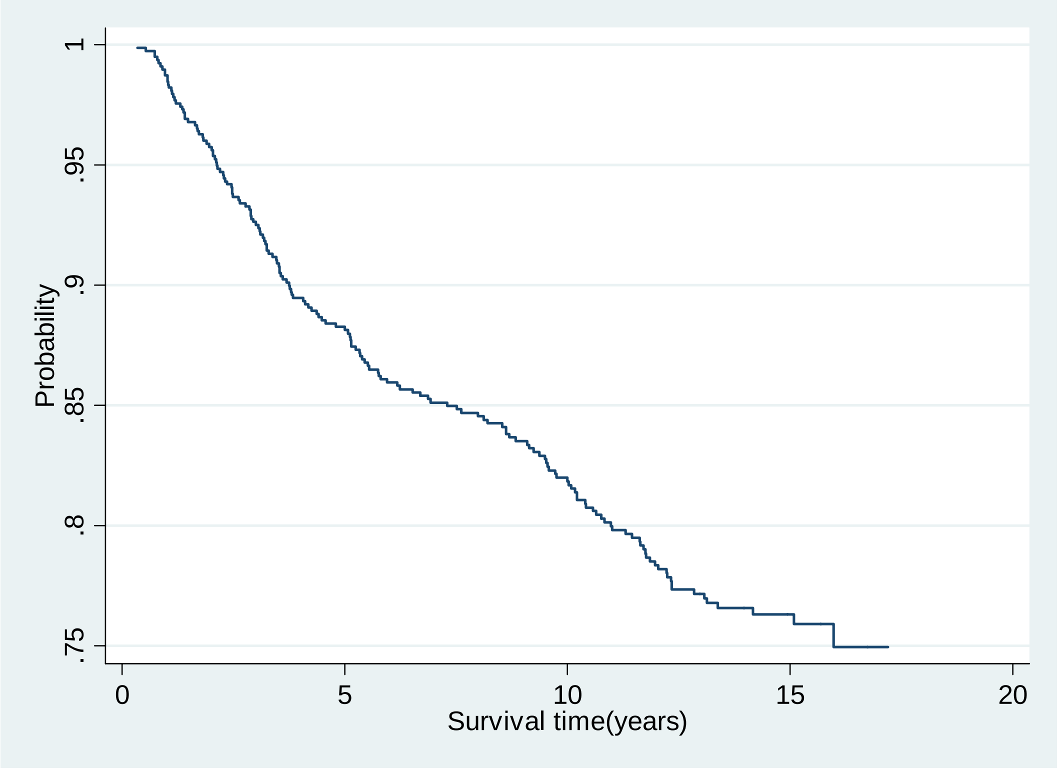

Figure 1. An example of a Kaplan Meier curve

2.3 COMPETING RISK

Gooley et al (1999) defined a competing risk as an event whose occurrence either precludes the occurrence of another event under examination or fundamentally alters the probability of occurrence of this other event. It is assumed that there is a potential failure time associated with each of the *p* risks to which the event is exposed. Thus, T_j represents the time-to-event failure from cause j ($j = 1, . . . ,p$) and the smallest T_j dictates the time to overall event failure. In classical competing risks only U = min $\{T_1, . . . , T_p\}$ and C, where U = T_c, are observable; U and C are respectively the time of failure and the cause, or type of failure. It is assumed for the present that ties cannot occur, i.e. pr $(T_j = T_k) = 0$ for all $j \neq k$. Otherwise C is not so simply defined, Crowder (1994). In other words, a unit is exposed to several risks at the same time, but it is assumed that the eventual failure of the unit is due to only one of these risks, which is called a “cause of failure” (McKeague et al, 1994).

When competing risks are present it is assumed that the subjects contribute iid observations to the data; the component fails when the first of all the competing failure mechanisms reaches a failure state; each of the *k* failure modes has a known life distribution model. (Pepe, 1991; Crowder, 1994). One can assume that each failure mechanism leading to a particular type of failure proceeds independently of each other, including the risk of the event of interest, at least until a failure occurs. However, this is often not likely to be true, particularly when there is causal-effect between events. To assume independence one must be sure that a failure of one type of event has no effect on the likelihood of any other events (Crowder, 1994).

Competing risks data can arise in many different situations and in relation to many different research questions. A common example found in the literature is allogeneic bone marrow transplant among patients with leukemia. Following bone marrow transplant, a patient may die from chronic graft-versus-host disease (CGVHD), recurrence of leukemia, pneumonia, other transplant related toxicities, or other causes during a study period. If the failure of interest is CGVHD, the other causes of death would be competing events. Each event can cause failure to occur; regardless of whether it is the failure of interest. Competing risks can be seen in women who start using an intrauterine device (IUD). They are subject to several risks, including accidental pregnancy, expulsion of the device, removal for medical reasons (including pelvic inflammatory disease) and removal for personal reasons. (Progress, 2002). Competing events can also be seen in patients on a waiting list for renal transplant, because not all patients who are registered will eventually be transplanted. Competing events arise when registered patients may developed contraindications from transplantation and will be removed from the waiting list, or patients may die while awaiting a donor kidney (Smits, 1998). Time to any first event of these can constitute competing risks data.

Kalbfleisch and Prentice (1980 p.164) identified three distinct problems that arise in analyzing data with competing risks:

1. The estimation of the relationship between covariates and the rate of occurrence of cause-specific failure.
2. The study of the interrelation between failure types under a specific set of conditions.
3. The estimation of failure rates for certain types of failure given the removal of some or all other failure types; this is regarded as a classic competing risk analysis.

Kalbfleisch and Prentice (1980) suggested the use of cumulative incidence estimates as a solution to the first problem and Tsai (1982) proposed 1 minus a Kaplan-Meier estimator as a way for estimating each event of interest in the presence of other competing events. The problem with that is 1 minus Kaplan-Meier overestimates the cumulative probability of a cause-specific failure (Gooley et al. (1999); Gaynor et al. (1993)) suggest the use of cumulative incidence estimator to estimate the cumulative probability when competing risks are present.

2.4 CUMULATIVE INCIDENCE ESTIMATE

The cumulative incidence function, also referred to as the cause-specific failure probability (Gaynor et al., 1993), can be interpreted as the cumulative probability that a failure of type *k* occurs on or before time *t* (Bryant and Dignam, 2004). The cumulative incidence function helps to determine patterns of failure and to assess the extent to which each component contributes to overall failure. For competing risks data one often wishes to estimate the cumulative incidence probability of failure of a specific cause, k, at time t, that is:

$$P_k(t) = P(T_j \leq t, \varepsilon_j = k) = \int_0^t \lambda_k(s) S(s) du \, ,$$

where ε_j indicates the cause of type of failure, *S(s)* is the overall survival probability, and $\lambda_k(s)$ is the cause-specific hazards for cause *k*(Scheike, 2003). The cumulative incidence estimator can be expressed in terms of the Kaplan Meier estimates,

$$CI(t) = \sum_{t_i \prec t} \frac{d_i}{n_i} K(t_i)$$

t_i is the distinct ordered observed times.
n_i is the number of patients who at risk beyond t_i.
d_i is the number of event of interest at t_i.
$K(t_i)$ is the Kaplan - Meier estimate of the probability of free of all events at time t_i.

3.0 BREAST CANCER DATA

3.1 BREAST CANCER BACKGROUND

According to the American Cancer Society, other than skin cancer, breast cancer is the most frequent malignant disease of women in the US with 180,000 new cases being diagnosed each year. It is the second leading cause of cancer death and the leading cause of cancer death for women 40-55 years of age. As in other cancers, the majority of women diagnosed with breast cancer are elderly, making aging one of the single greatest risk factors for the development of new breast cancer, with the estimated risk of new breast cancer at 1 in 14 for women aged 60 to 79 compared with 1 in 24 women aged 40 to 59 and 1 in 228 women aged 39 and younger (Holmes et al., 2003). While some breast cancer patients die from this disease, most of them die from other causes because breast cancer patients are usually old with many other health conditions (Holmes et al., 2003).

3.2 BREAST CANCER DATASET

The data used to illustrate the difference between Kaplan Meier estimates and the cumulative incidence estimate consists of 788 eligible participants with follow-up from the National Surgical Adjuvant Breast and Bowel Project (NSABP) databases. The participants were enrolled in the B-20 study between Oct 17, 1988 and March 5, 1993 and they were randomly assigned to 3 regimes including tamoxifen only treatment, which was used here for illustration. The dataset included time to event, measured in years, and event indicator categorized as: event free, recurrence, death (not event related), and second primary cancer. Median follow-up years were 13.09, and among those 788 patients, 478 were event free at the last follow-up. 170 patients had breast-cancer recurrence, 100 patients developed other cancers (secondary primary cancer) and 40 patients die before they developed those events mentioned earlier. In many clinical and epidemiological studies, subjects can potentially experience recurrent events. Here recurrence of breast cancer, death prior due to other diseases, and second primary cancers are competing events.

4.0 METHODS

4.1 STATISTICAL METHODS

The data were labeled and recoded using the Stata version 9.2 statistical software package (StataCorp., 2005). To convert time into years, the original time variable was divided by 120. The data were declared to be a survival- time data with recurrence as failure event, and the sts and generate command was used to create the 1 minus Kaplan-Meier estimates. The stcompet command in Stata was used to generate the cumulative incidence estimates for each competing events, recurrence, second primary cancers and death. The focus will only be on the cumulative incidence estimates for recurrence. Overlaid two-way graphs were plotted using the 1 minus Kaplan-Meier and cumulative incidence estimates for recurrence versus time. The Kaplan-Meier estimates and cumulative incidence estimates were compared, using the compare command, to see how many observations were different.

When there are no competing events the Kaplan-Meier estimates and the cumulative incidence estimates are identical. To illustrate this, a new time variable was generated, and event times were replaced with the maximum time 17.2 years for other events: event free, second primary cancers and death. The data were again declared to be a survival-time data with recurrence as failure event. The 1 minus Kaplan-Meier estimates and cumulative incidence estimates were generated, plotted and compared.

5.0 RESULTS

5.1 COMPARISION IN THE PRESENCE OF COMPETING EVENTS

The breast cancer recurrence was used as the event of interest (failure event) to illustrate the differences in cumulative probability of cause-specific failure estimates given by the two estimators (1 minus Kaplan Meier and cumulative incidence). The graphs are identical for the first five observations (Table 1) because there are no competing events before t= 0.7833 years. Figure 2 and Table 2 show a clear difference between the two estimates. The 1 minus Kaplan Meier estimator over-estimates the probability of recurrence among the breast cancer patients. The difference is very noticeable after 5 years and increases with more competing events i.e. death, second primary cancers. The 1 minus Kaplan Meier gives a larger estimate than the Cumulative Incidence estimator.

Table 1: 1-KM and CI estimates for the first 5 breast cancer data observations

Probability of breast cancer recurrence		
time	1-KM	CI
.342	.0013	.0013
.533	.0025	.0025
.725	.0051	.0051
.725	. 0051	.0051
.783	.0063	.0063

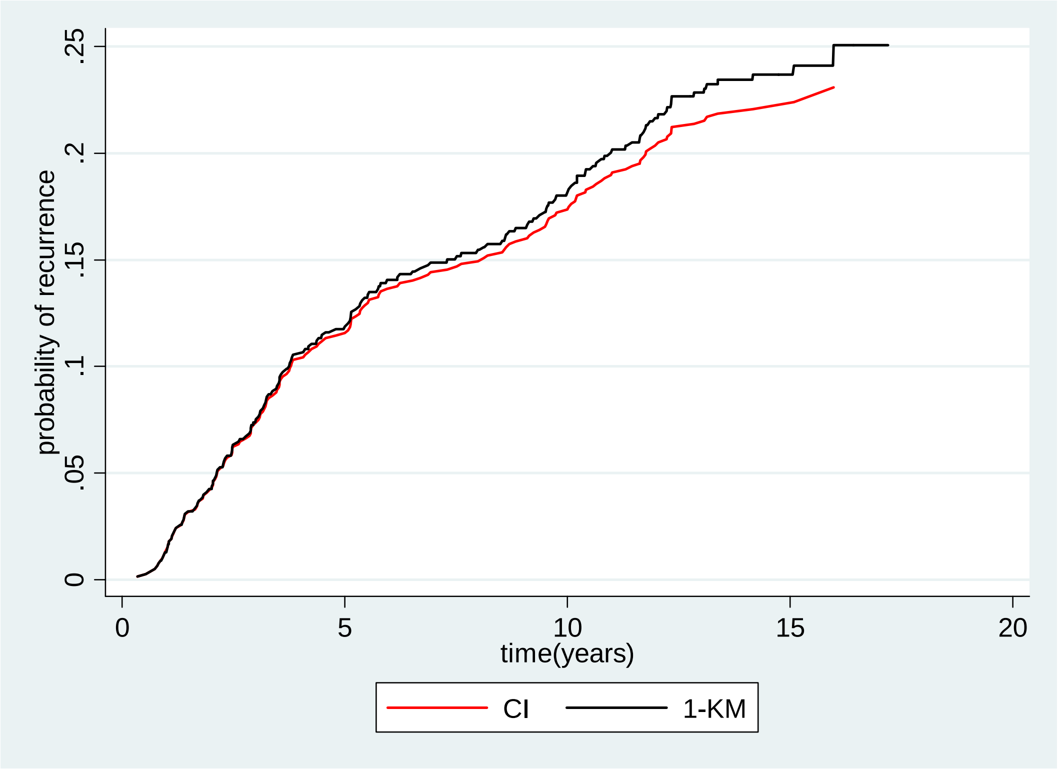

Figure 2. 1- KM estimate and the CI estimate of recurrence for the breast cancer dataset

Figure 2 and shows the relationship between follow-up time in years and the probability of recurrence of breast cancer, which was estimated by the cumulative incidence function and Kaplan-Meier methods.

Table 2: Selected observations of CI and 1 - KM for the breast cancer data

Probability of breast cancer recurrence		
Time t (years)	1-KM	CI
1.1	.0191	.0190
5.0	.1187	.1157
10.23	.1893	.1802
15.08	.2409	.2239

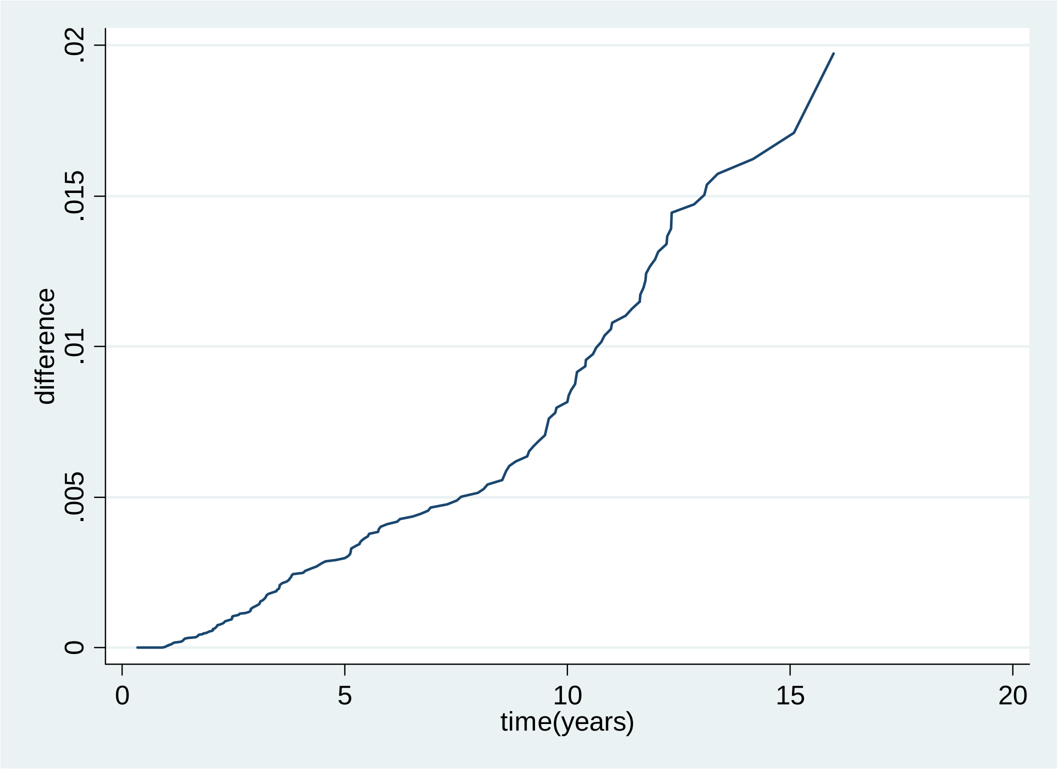

Figure 3. The difference between 1-KM and CI

Figure 3 shows the difference between the two estimators, and the compare command in Stata generated (Table 3), which shows the Kaplan Meier estimates greater than the cumulative incidence for 165 out of 170 observations. The difference between the 1 minus Kaplan-Meier and cumulative incidence probabilities of breast cancer recurrence is positive over time. The 1 minus Kaplan Meier is a non-interpretable and biased estimate for the probability of recurrence, is due to the censoring of observations that are failures from a competing events (Gooley et al., 1999)

Table 3: 1-KM and CI comparison.

		difference		
	count	minimum	average	maximum
KM1=CI1	5			
KM1>CI1	165	1.62e-06	.0045642	.0197285
jointly defined	170	0	.0044299	0197285
CI1missing only	618			
total	788			

With two competing events i.e. second primary cancer and death, the graphs and tables shows an evident over-estimate of the 1 minus Kaplan-Meier estimator, however researchers continue to use 1 minus Kaplan-Meier estimates to interpret the cumulative probability of cause-specific failure. “We feel the primary reason for this misuse is a fundamental misunderstanding among clinical researchers of the assumptions required for interpretable Kaplan-Meier estimates, coupled with a lack of thorough comprehension of how CI is computed” Gooley et al,1999.

5.2 COMPARISON IN THE ABSENCE OF COMPETING EVENTS

The Kaplan Meier estimates are identical to the cumulative incidence estimates if no failures from competing events are encountered (Gooley et al., 1999). According to Gaynor et al., there are two hypothetical assumptions that must be met for the two estimators to be identical: (1) eliminate competing events, failures due to other causes and (2) the events of interest must remain the same given assumption (1). In Figure 4,the breast cancer dataset was used to validate both assumptions.

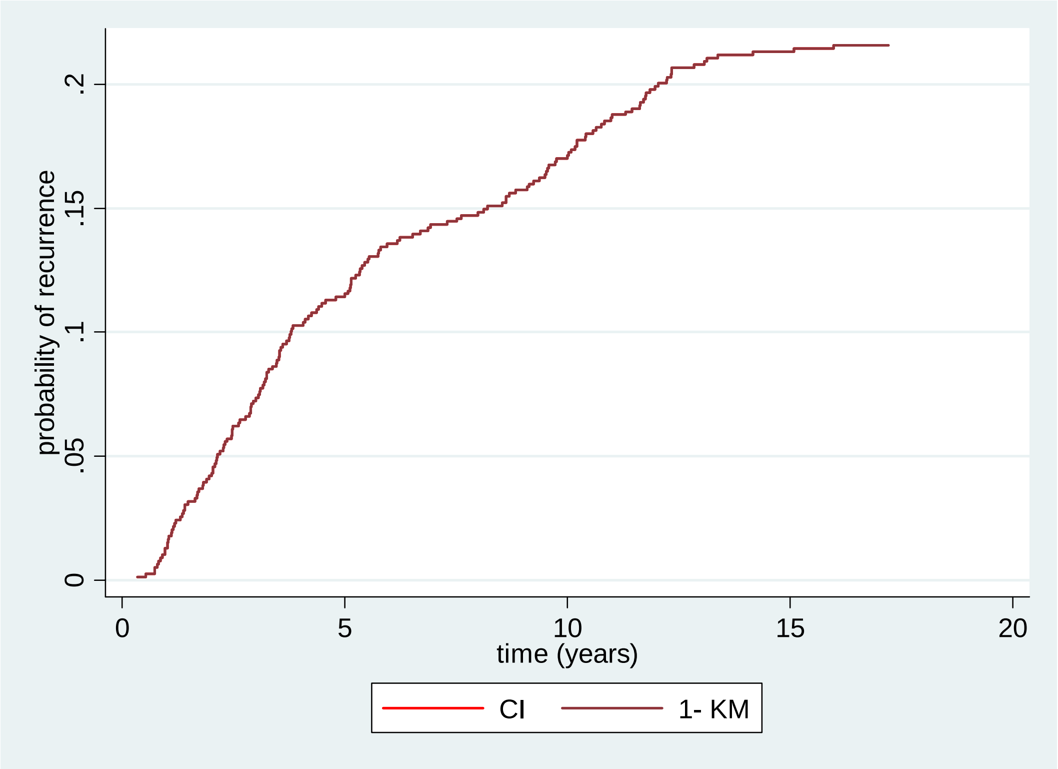

Figure 4. 1-KM and CI estimates in the absence of competing events

Figure 4 shows one curve, which indicates that the cumulative incidence function is equivalent to the 1 minus Kaplan-Meier. Table 4 below also verifies the equality of these two estimators.

Table 4:1-KM and CI comparison in the absence of competing events

		difference		
	count	minimum	average	maximum
KM2=CI2	170			
jointly defined	170	0	0	0
CI2missing only	618			
total	788			

6.0 DISCUSSION

In analyzing competing risks data, it is important to realize the possible contributions of sound statistical methodology to the adequate exploration of the data. The use of 1 minus Kaplan-Meier to estimate cause-specific cumulative probability is based on the incorrect assumption that the probability of failing prior to time *t* from cause *k* is equal to 0. This incorrect assumption can lead to an inflated estimated of the proportion of patients who are at risk of failure at time t, causing 1 minus Kaplan-Meier to overestimate the cause-specific failure probability. The breast cancer dataset demonstrated the bias in using the 1 minus Kaplan-Meier approach.

Discrepancy between the 1 minus Kaplan Meier and the cumulative incidence methods is instantly recognizable, so the cumulative incidence should always be used if an estimate of the cumulative probability of cause-specific events is desired. One should avoid censoring competing events at the event times for convenience to use the 1 minus Kaplan-Meier approach. The censored events are informative, because it changes the probability of the event of interest occurring.

APPENDIX A

PROGRAM CODE

```
                                                          ___  ____  ____  ____  ____tm
                                                         /__    /   ____/   /   ____/
                                                        ___/   /   /___/   /   /___/
                                                          Statistics/Data Analysis

                                                            User: BINTU N. SHERIF
                                                         Project: Thesis
 1 . tab event
   > -------------
          log:  C:\data\bintu.log
     log type:  text
    opened on:  25 Nov 2007, 20:29:53

 2 . fdause "C:\Documents and Settings\Owner\Desktop\THESIS\bintu.xpt"

 3 . generate time= toff/120

 4 . generate event=0 if dstf==1
   (310 missing values generated)

 5 . replace event=1 if dstf==2
   (133 real changes made)

 6 . replace event=1 if dstf==6
   (37 real changes made)

 7 . replace event=2 if dstf==3
   (40 real changes made)

 8 . replace event=3 if dstf==5
   (100 real changes made)

 9 . label define evt 0 "event-free"

10 . label define evt 1 "recurrence", add

11 . label define evt 2 "death", add

12 . label define evt 3 "SecPrimCan", add

13 . label values event evt

14 . tab event

         event |      Freq.     Percent        Cum.
   ------------+-----------------------------------
    event-free |        478       60.66       60.66
    recurrence |        170       21.57       82.23
         death |         40        5.08       87.31
    SecPrimCan |        100       12.69      100.00
   ------------+-----------------------------------
         Total |        788      100.00

15 . stset time, failure(event==1)

        failure event:  event == 1
   obs. time interval:  (0, time]
    exit on or before:  failure

   ------------------------------------------------------------------------------
         788  total obs.
           0  exclusions
   ------------------------------------------------------------------------------
         788  obs. remaining, representing
         170  failures in single record/single failure data
    8817.242  total analysis time at risk, at risk from t =         0
                                earliest observed entry t =         0
                                     last observed exit t =      17.2
```

```
16  . sts gen surv = s

17  . gen km=1- surv

18  . stcompet CumInc = ci SError = se, compet1(2) compet2(3)

19  . gen CumInc1 = CumInc if event==1
    (618 missing values generated)

20  . gen CumInc2 = CumInc if event==2
    (748 missing values generated)

21  . gen CumInc3 = CumInc if event==3
    (688 missing values generated)

22  . twoway (line surv time, sort connect(stairstep)), ytitle(Probability) xtitle(Survival time (ye
    > ars))

23  . label var km "1- KM"

24  . label var CumInc1 "CI"

25  . twoway (line km time, sort connect(stairstep)) (line CumInc1 time, sort lcolor(red) connect(st
    > airstep)),
    > ytitle(Probability of recurrence) xtitle(time (years))

26  . gen time2=17.2

27  . replace time2=time if event==1
    (170 real changes made)

28  . gen evt2=0

29  . replace evt2=1 if event==1
    (170 real changes made)

30  . stset time2, failure(evt2==1) scale(1)

         failure event:  evt2 == 1
    obs. time interval:  (0, time2]
     exit on or before:  failure

    ------------------------------------------------------------------------------
          788  total obs.
            0  exclusions
    ------------------------------------------------------------------------------
          788  obs. remaining, representing
          170  failures in single record/single failure data
     11588.01  total analysis time at risk, at risk from t =         0
                                 earliest observed entry t =         0
                                      last observed exit t =      17.2

31  . sts gen survb = s

32  . gen km2=1- survb
```

```
33  . stcompet CI = ci SE = se, compet1(2) compet2(3)

34  . gen CumInc1a = CI if evt2==1
    (618 missing values generated)

35  . label var CumInc1a "CI"

36  . label var km2 "1- KM"

37  . twoway (line km2 time2, sort connect(stairstep)) (line CumInc1a time2, sort lcolor(red) connec
    > t(stairstep
    > )), ytitle(Probability of recurrence) xtitle(time (years))

38  . gen diff= km- CumInc1
    (618 missing values generated)

39  . twoway (line diff time, sort), ytitle(Difference) xtitle( time (years))

40  . compare km CumInc1

                                               ---------- difference ----------
                                    count       minimum      average     maximum
    ------------------------------------------------------------------------
    km=CumInc1                          5
    km>CumInc1                        165      1.62e-06     .0045642    .0197285
                               ----------
    jointly defined                   170             0     .0044299    .0197285
    CumInc1 missing only              618
                               ----------
    total                             788

41  . compare km2 CumInc1a

                                               ---------- difference ----------
                                    count       minimum      average     maximum
    ------------------------------------------------------------------------
    km2=CumInc1a                      170
                               ----------
    jointly defined                   170             0            0           0
    CumInc1a missing only             618
                               ----------
    total                             788

42  . edit
    - preserve
    - sort time

43  . list time  km CumInc1 in 1/50

         +---------------------------------+
         |     time          km    CumInc1 |
         |---------------------------------|
      1. | .3416667     .001269    .001269 |
      2. | .5333334    .0025381   .0025381 |
      3. |     .725    .0050761   .0050761 |
      4. |     .725    .0050761   .0050761 |
      5. | .7833334    .0063452   .0063452 |
         |---------------------------------|
      6. | .7833334    .0063452          . |
      7. | .8166667    .0076158   .0076142 |
      8. |     .825    .0076158          . |
      9. | .8666667    .0088881   .0088832 |
     10. |     .875    .0088881          . |
         |---------------------------------|
     11. |       .9     .010162          . |
     12. |       .9     .010162   .0101523 |
     13. | .9083334     .010162          . |
     14. | .9583333    .0127165   .0126904 |
      15. | .9583333   .0127165   .0126904 |
          |--------------------------------|
      16. | .9583333   .0127165          . |
      17. | .9666666   .0127165          . |
      18. | .9916667   .0127165          . |
      19. | 1.016667   .0152808          . |
      20. | 1.016667   .0152808   .0152284 |
          |--------------------------------|
      21. | 1.016667   .0152808   .0152284 |
      22. | 1.033333   .0165647   .0164975 |
      23. | 1.041667   .0165647          . |
      24. |     1.05   .0178502          . |
      25. |     1.05   .0178502   .0177665 |
          |--------------------------------|
      26. |      1.1   .0191375   .0190355 |
      27. |    1.125   .0204247   .0203046 |
      28. |     1.15   .0217119   .0215736 |
      29. |     1.15   .0217119          . |
      30. | 1.183333   .0230008   .0228426 |
          |--------------------------------|
      31. |      1.2   .0242897   .0241117 |
      32. | 1.308333   .0255786   .0253807 |
      33. | 1.316667   .0255786          . |
      34. | 1.333333   .0255786          . |
      35. |     1.35    .026871   .0266497 |
          |--------------------------------|
      36. |    1.375   .0281633   .0279188 |
      37. | 1.408333    .030748   .0304569 |
      38. | 1.408333    .030748   .0304569 |
      39. |    1.475   .0320403   .0317259 |
      40. | 1.583333   .0320403          . |
          |--------------------------------|
      41. | 1.641667   .0333344   .0329949 |
      42. |    1.675   .0346285    .034264 |
      43. |      1.7   .0359225    .035533 |
      44. |    1.725   .0372166    .036802 |
      45. | 1.733333   .0372166          . |
          |--------------------------------|
      46. | 1.816667   .0385124   .0380711 |
      47. |    1.825   .0398082   .0393401 |
      48. |      1.9    .041104   .0406091 |
      49. |     1.95   .0423998   .0418782 |
      50. | 1.991667   .0423998          . |
          +--------------------------------+

44  . list time2  km2 CumInc1a in 1/50

          +--------------------------------+
          |    time2        km2   CumInc1a |
          |--------------------------------|
       1. | .3416667    .001269    .001269 |
       2. | .5333334   .0025381   .0025381 |
       3. |     .725   .0050761   .0050761 |
       4. |     .725   .0050761   .0050761 |
       5. | .7833334   .0063452   .0063452 |
          |--------------------------------|
       6. |     17.2    .215736          . |
       7. | .8166667   .0076142   .0076142 |
       8. |     17.2    .215736          . |
       9. | .8666667   .0088832   .0088832 |
      10. |     17.2    .215736          . |
          |--------------------------------|
      11. |     17.2    .215736          . |
      12. |       .9   .0101523   .0101523 |
      13. |     17.2    .215736          . |
      14. | .9583333   .0126904   .0126904 |
      15. | .9583333   .0126904   .0126904 |
          |--------------------------------|
    16. |     17.2    .215736          . |
    17. |     17.2    .215736          . |
    18. |     17.2    .215736          . |
    19. |     17.2    .215736          . |
    20. | 1.016667   .0152284   .0152284 |
        |--------------------------------|
    21. | 1.016667   .0152284   .0152284 |
    22. | 1.033333   .0164975   .0164975 |
    23. |     17.2    .215736          . |
    24. |     17.2    .215736          . |
    25. |     1.05   .0177665   .0177665 |
        |--------------------------------|
    26. |      1.1   .0190355   .0190355 |
    27. |    1.125   .0203046   .0203046 |
    28. |     1.15   .0215736   .0215736 |
    29. |     17.2    .215736          . |
    30. | 1.183333   .0228426   .0228426 |
        |--------------------------------|
    31. |      1.2   .0241117   .0241117 |
    32. | 1.308333   .0253807   .0253807 |
    33. |     17.2    .215736          . |
    34. |     17.2    .215736          . |
    35. |     1.35   .0266497   .0266497 |
        |--------------------------------|
    36. |    1.375   .0279188   .0279188 |
    37. | 1.408333   .0304569   .0304569 |
    38. | 1.408333   .0304569   .0304569 |
    39. |    1.475   .0317259   .0317259 |
    40. |     17.2    .215736          . |
        |--------------------------------|
    41. | 1.641667   .0329949   .0329949 |
    42. |    1.675    .034264    .034264 |
    43. |      1.7    .035533    .035533 |
    44. |    1.725    .036802    .036802 |
    45. |     17.2    .215736          . |
        |--------------------------------|
    46. | 1.816667   .0380711   .0380711 |
    47. |    1.825   .0393401   .0393401 |
    48. |      1.9   .0406091   .0406091 |
    49. |     1.95   .0418782   .0418782 |
    50. |     17.2    .215736          . |
        +--------------------------------+

45 . log close
          log:  C:\data\bintu.log
     log type:  text
    closed on:  25 Nov 2007, 20:43:44
   -----------------------------------------------------------------------------------------------
   > -----------
```

BIBLIOGRAPHY

Allison, P. D. (1984). Event History Analysis: Regression for Longitudinal Event Data. Beverly Hills, Sage Publications.

Bender, A. P., J. Punyko, et al. (1992). "A standard person-years approach to estimating lifetime cancer risk. The Section of Chronic Disease and Environmental Epidemiology Minnesota Department of Health." Cancer Causes Control **3**(1): 69-75.

Bryant, J. and J. J. Dignam (2004). "Semiparametric models for cumulative incidence functions." Biometrics **60**(1): 182-90.

Chen, B. E., J. L. Kramer, et al. (2007). "Competing Risks Analysis of Correlated Failure Time Data." Biometrics.

Cheng, S. C., J. P. Fine, et al. (1998). "Prediction of cumulative incidence function under the proportional hazards model." Biometrics **54**(1): 219-28.

Cronin, K. A. and E. J. Feuer (2000). "Cumulative cause-specific mortality for cancer patients in the presence of other causes: a crude analogue of relative survival." Stat Med **19**(13): 1729-40.

Crowder, M. (1996). "On assessing independence of competing risks when failure times are discrete." Lifetime Data Anal **2**(2): 195-209.

Crowder, M. (1997). "A test for independence of competing risks with discrete failure times." Lifetime Data Anal **3**(3): 215-23.

Fisher, B., J. H. Jeong, et al. (2004). "Treatment of lymph-node-negative, oestrogen-receptor-positive breast cancer: long-term findings from National Surgical Adjuvant Breast and Bowel Project randomised clinical trials." Lancet **364**(9437): 858-68.

Gaynor, J. J., E. J. Feuer, et al. (1993). "On the Use of Cause-Specific Failure and Conditional Failure Probabilities: Examples From Clinical Oncology Data." Journal of the American Statistical Association, **88**(No. 422): 400-409.

Gooley, T. A., W. Leisenring, et al. (1999). "Estimation of failure probabilities in the presence of competing risks: new representations of old estimators." Stat Med **18**(6): 695-706.

Holmes, C. E. and H. B. Muss (2003). "Diagnosis and treatment of breast cancer in the elderly." CA Cancer J Clin **53**(4): 227-44.

Jeong, J. H. and J. P. Fine (2007). "Parametric regression on cumulative incidence function." Biostatistics **8**(2): 184-96.

Kalbfleisch, J. D. and R. L. Prentice (1980). The Statistical Analysis of Failure Time Data. New York, Wiley.

Kaplan, E. L., Meier, P. (1958). "Nonparametric Estimation from Incomplete Observations " Journal of the American Statistical Association, **53**(282): 457-481.

Kuru, B., M. Camlibel, et al. (2003). "Prognostic factors affecting survival and disease-free survival in lymph node-negative breast carcinomas." J Surg Oncol **83**(3): 167-72.

Lee, E. T. (1992a). Statistical Methods for Survival Data Analysis. Oklahoma City, John Wiley & Sons, Inc.

Lin, D. Y. (1997). "Non-parametric inference for cumulative incidence functions in competing risks studies." Stat Med **16**(8): 901-10.

Martinez, E., A. Milinkovic, et al. (2007). "Incidence and causes of death in HIV-infected persons receiving highly active antiretroviral therapy compared with estimates for the general population of similar age and from the same geographical area." HIV Med **8**(4): 251-8.

Maurice, J. (2002). A plethora of IUD: but how safe, how effective? Progress in Reproductive Health Research. **60:** 3.

McKeague, E., et al. (1994). "Some Tests for Comparing Cumulative Incidence Functions and Cause-Specific Hazard Rates." Journal of the American Statistical Association **89**(427): 994-999.

Pepe, M. S. (1991). "Inference for Events With Dependent Risks in Multiple Endpoint Studies." Journal of the American Statistical Association **86**(415): 770-778.

Prentice, R. L., J. D. Kalbfleisch, et al. (1978). "The analysis of failure times in the presence of competing risks." Biometrics **34**(4): 541-54.

Schairer, C., P. J. Mink, et al. (2004). "Probabilities of death from breast cancer and other causes among female breast cancer patients." J Natl Cancer Inst **96**(17): 1311-21.

Scheike, T. H. and M. Zhang (2003). Predicting Cumulative Incidence Probability by Direct Binomial Regression.

Smits, J. M., H. C. van Houwelingen, et al. (1998). "Analysis of the renal transplant waiting list: application of a parametric competing risk method." Transplantation **66**(9): 1146-53.

Sobolev, B. G., A. R. Levy, et al. (2006). "The risk of death associated with delayed coronary artery bypass surgery." BMC Health Serv Res **6**: 85.

Tsai, W.-Y. **(1982),** "Bivariate Survival Time and Censoring," unpublished Ph.D. dissertation, University of Wisconsin, Dept. of Biostatistics.

Tsiatis, A. (1975). "A nonidentifiability aspect of the problem of competing risks." Proc Natl Acad Sci U S A **72**(1): 20-2.

Wei, L. J. and D. V. Glidden (1997). "An overview of statistical methods for multiple failure time data in clinical trials." Stat Med **16**(8): 833-9; discussion 841-51.

Wu, M. C., S. Hunsberger, et al. (1994). "Testing for differences in changes in the presence of censoring: parametric and non-parametric methods." Stat Med **13**(5-7): 635-46.