Competing risk analysis of PFS considering patients who switch therapy prior to progression

Lixia Jiao* Liji Shen* Zhenming Shun*

Abstract

Standard survival analysis methods, such as Kaplan Meier curves, log-rank test and Cox proportional hazard model, are widely accepted tools to compare the cause-specific hazards when there is only one event of interest and the time to event and time to censoring are independent. However, competing risks are often encountered in clinical research, where multiple failure types exist and one type of event either precludes the occurrence of another event or fundamentally alters the probability of occurrence of the other event. In the analysis of competing risks data, the standard analysis methods may lead to biased results by treating the competing event as censored at the time this event occurs. This way, it is assumed that the patients failing from a competing risk are no more or less likely to fail from the cause of interest than the patients still at risk beyond this time. The newly developed methods, such as Gray's test and Pepe and Mori's method, take into account of the competing risks and provide different clues regarding the effect of a covariate. Gray proposed a class of generalized linear rank statistics for testing equality of cumulative incidence functions. Pepe and Mori proposed a different class of test statistics, not based on ranks, for comparing cumulative incidence functions and conditional probability functions. In standard progression free survival (PFS) analysis, patients who change their cancer therapy prior to progression will be labeled as censored at the time of stopping randomized treatment. As changing cancer therapy alters the probability of progression, it should be considered as a competing risk event and the newly developed methods apply.

Key Words: Survival analysis; Kaplan Meier; Log-rank test; Competing risks; Cumulative incidence function.

1. Introduction

It is challenging in the estimation of the probability of failure for time-to-event endpoints in randomized clinical trials where the competing risk events are present. Gooley et al. (1999) define the concept of competing risks as the event whose occurrence either precludes the occurrence of another event under investigation, or fundamentally alters the probability of occurrence of this other event. For example, cause-specific death due to prostate cancer is the event of interest in a randomized clinical trial, whereas deaths due to other causes (i.e. car accident) are competing risk events as 'death due to other causes' precludes occurrence of 'death due to prostate cancer'. Patients who change their cancer therapy prior to progression will be labeled as censored at the time of stopping randomized treatment in the standard progression free survival analysis. Should changing cancer therapy prior to progression be considered as a competing risk? Changing cancer therapy alters the probability of progression. When there is only one type of events, i.e. progression, the censoring is assumed non-informative and the time to event and time to censoring are independent. However, in the competing risk context such as changing cancer therapy, censoring due to competing risks event is no longer non-informative and therefore time to event and time to censoring are no longer independent. Therefore, this necessitates a carefully assessment in regard to whether the standard approach of treating competing risk event as censored observations introduces serious bias and whether the newer approaches recently developed in dealing with competing risk can properly address this question.

*sanofi-aventis, 55 Corporate Dr., Bridgewater, NJ 08807

There are two estimators for probability of failure: the complement of a Kaplan-Meier estimate (1-KM), and the cumulative incidence estimate (CI). (1-KM) is frequently misused in the competing risks framework. To estimate (1-KM), the failures from a competing event are treated as censored at the time this event occurs. This way, we assume that the patients failing from a competing risk are no more or less likely to fail from the cause of interest than the patients still at risk beyond this time point. A hypothetical probability assuming the probability of failure from the event of interest would not change if the competing risk event is removed. However, the cumulative incidence function is a function of the hazards of all the competing events and not solely of the hazard of the event to which it refers. Furthermore, the sum of all cumulative incidences has the nice feature that it equals $1 - S(t)$, the complement of the overall Kaplan-Meier estimate of survival considering failures of any kind.

Standard methods (i.e. the logrank test and the Cox's proportional hazards model) compare the cause-specific hazards as if other types of events did not exist. It is a good way to find the biological mechanisms underlying the specific event. In the presence of competing risks, comparing cumulative incidence functions is more direct to the comparison of probability of failure, which accounts for all types of events and does not assume independence between the time to different types of events. It is worth noting that the cumulative incidence function for event of interest can be low because risk of a competing risk is high, hence there arises a need to compare cumulative incidence function for the competing risk too.

In this paper we will discuss three competitive approaches in adjusting competing risk events: (1) Gray's test that is a modified log-rank score using the modified number at risk; (2) Pepe and Mori's method that compares the cumulative incidence functions of the two treatment groups using a score function; (3) A conditional probability approach that incorporates both the event of interest and the competing risk event. These methods are applied to a real oncology trial example. In addition, a comprehensive simulation study is conducted to evaluate the statistical performance of these three newer approaches and to compare to the standard logrank test in terms of power and type I error rate.

2. DESCRIPTION OF METHODS

This work is motivated by a randomized oncology trial that compares an investigational cancer treatment to the standard cancer therapy on progression free survival in patients with prostate cancer. A total of 439 patients, 219 in the investigation treatment group and another 219 in the standard treatment group, were randomized in the trial. The primary event of interest is a composite endpoint of either progression or death, whichever comes earlier. Since patients may change to other anti-cancer therapy prior to progression due to either insufficient treatment effects or unfavorable side effects to the study medication, progression or death occurred after the switch of cancer therapy cannot be used the same way as data for patients who have not experienced any switch. In this trial, as shown in Table 1, 185 patients either developed progression or died before study cut-off in the investigational treatment group, without interference of other anti-cancer therapy; whereas, 174 patients experienced progression/death in the standard treatment group. There are 22 patients in the investigational treatment group and 27 patients in the standard treatment group who switched to another anti-cancer therapy prior to progression or death before the study cut-off date. In addition, 12 patients in the investigational treatment group and 18 in the standard treatment group did not progress or die or switch to other anti-cancer therapy before study cut-off date and are deemed as truly non-informative censored observations. Figure 1 depicts the traditional progression free survival censoring scheme that is com-

monly deployed in the oncology arena, where changing cancer therapy prior to progression or death is treated as regular censored observations as if these events are independent of the primary event of interest, i.e. progression/death. However, as changing cancer therapy alters the probability of progression and the underlying assumption of independence is no longer valid. Therefore, newer approaches that properly deal with the competing risks are evaluated below.

2.1 1 - KM versus CIF

1-KM is a function of the hazard of failure due to event of interest and does not depend on the hazard of failure due to competing risk event. Whereas, CIF is a function of the hazard of failure due to event of interest and competing risk event. Therefore, cumulative incidence function is more appropriate in estimating the probability of failure due to event of interest when competing risk events are present.

For the purpose of illustration, we assume that there are two types of events, namely the even of interest and the competing risk event, represented by $k = 2$. Each patient will experience one and only one of the following three outcomes: failure from event of interest, failure from completing risk event, or survival from event of interest and competing risk event, at the ordered time points such that $t_1 \leq t_2 \leq ... \leq t_n$.

The following notations are defined:
n : Number of patients who are initially at risk of failure;
e_j : Number of patients who failed from event of interest at time t_j;
r_j : Number of patients who failed from competing risk event at time t_j;
c_j : Number of patients who are censored at time t_j;
n_j : Number of patients who are at risk of failure beyond time t_j;
$S(t_j)$: Probability of free of any kind of event at time t_j;
$S_e(t_j)$: Probability of free of event of interest at time t_j.

Then the estimate of $1 - KM_e(t)$ for event of interest is expressed as

$$1 - \hat{KM}_e(t) = \sum_{all j, t_j < t} \frac{e_j}{n_j} \hat{S}_e(t_{j-1})$$

The estimate of CIF ($F_e(t)$) for event of interest is expressed as

$$\hat{F}_e(t) = \sum_{all j, t_j < t} \frac{e_j}{n_j} \hat{S}(t_{j-1})$$

It is easily shown that

$$\hat{F}_e(t) \leq 1 - \hat{KM}_e(t)$$

.

2.2 Gray's test

Gray's k-sample test (1988) is a modified log-rank test, which compares the weighted averages of the hazard function $\gamma_i(t)$ for event of interest. $\gamma_i(t)$ is the hazard of the subdistribution for the event of interest in treatment group i. Again, we assume $k = 2$ which means there are two types of event: the event of interest and the competing risk event. And $t_1 \leq t_2 \leq ... \leq t_n$ are ordered time points of event.

We further define
e_{1j} : Number of patients who failed from event of interest for Group 1 at time t_j;
n_{1j} : Number of patients who are at risk of failure for Group 1 beyond time t_j;

The modified log-rank score for Gray's test can then be expressed as

$$Z_1 = \sum_{i=1}^{n}(e_{1j} - R_{1j}\frac{e_j}{R_j})$$

Where R_j is the modified number at risk and can be expressed as

$$R_{1j} = n_{1j}\frac{1 - \hat{F}_1(t_{j-1})}{\hat{S}_1(t_{j-1})}$$

$$R_j = R_{1j} + R_{2j}$$

In comparison, the standard log-rank is formulated as

$$Z = \sum_{i=1}^{n}(e_{1j} - n_{1j}\frac{e_j}{R_j})$$

2.3 Pepe and Mori's method

Pepe and Mori's method compares the CIFs of the two groups using a score function s that is the weighted area between the two CIFs. The score function s is defined as

$$s = \sqrt{\frac{N_1 N_2}{N_1 + N_2}}\sum_{j=1}^{n}\{W(t_j)[\hat{F}_1(t_j) - \hat{F}_2(t_j)](t_{j+1} - t_j)\}$$

Where $\hat{F}_i(t)$ denotes the cumulative incidence function of group i at time t and N_i denotes the total number of patients in group i. The weight function $W(t_j)$ is defined as

$$W(t_j) = \frac{(N_1 + N_2)\hat{C}_1(t_{j-1})\hat{C}_2(t_{j-1})}{N_1\hat{C}_1(t_{j-1}) + N_2\hat{C}_2(t_{j-1})}$$

Where $1 - \hat{C}_i(t)$ is the left continuous Kaplan-Meier estimator of the censoring distribution function where events are censored or competing risk events in the ith group at time t.

The score function s follows a normal distribution:

$$s \sim N(0, \sigma)$$

2.4 Conditional probability approach

Pepe also proposed a conditional probability approach for comparing general functions, which provides a way to incorporate 2 types of events, i.e. event of interest and competing risk events. In general, it can be expressed as

$$CP_i(t) = Prob(T \leq t, C = i \mid \texttt{no other type of event by time } t)$$

Again, assume that there are two type of events: Type 1 event which is the event of interest and Type 2 event which is the competing risk event. Therefore, the probability of failure from the event of interest given there is no competing risk event by time t can be expressed as

$$CP(t) = \frac{P(T \leq t, C = C_1)}{1 - P(T \leq t, C = C_2)} = \frac{F_1(t)}{1 - F_2(t)}$$

Where $CP(t)$ can be estimated using estimators for $F_1(t)$ and $F_2(t)$.

2.5 Application

In order to illustrate the three newer approaches as described above within the framework of competing risks, we apply these three methods along with the standard logrank approach to the oncology example introduced earlier in the section where changing cancer therapy is considered as a competing risk to the primary event of interest, i.e. progression or death event, whichever occurs earlier. Figure 2 and Figure 3 provide visual representation of the cumulative incidence functions by treatment group for the event of interest (progression or death) and the competing risk event (changing cancer therapy prior to progression or death), respectively. Figure 4 and Figure 5 depict the conditional probability functions by treatment group for the event of interest and the competing risk event, respectively. The results containing p-values for treatment comparisons between the investigational cancer treatment and the standard cancer treatment after applying the four different methods are summarized in Table 2. The first row shows that the p-value for the primary endpoint of interest (progression/death) is 0.199 by using the standard logrank test; whereas, they are 0.235, 0.500 and 0.207, respectively, by using Gray's test, Pepe and Mori's method, and the conditional probability approach. This indicates that there is no statistically significant difference between the two treatment groups even with the application of the three newer approaches that adjust for completing risk events. When using these three newer approaches, it is important to compare cumulative incidence functions or conditional probability functions for competing risk as well. For this particular example, there is no statistically significant difference between the two treatment groups for competing risk enpoint with p-values equal to 0.424, 0.651, and 0.726, respectively for Gray's test, Pepe and Mori's method, and the conditional probability approach. These results are fairly consistent with each other as the incidence of competing risk events is relatively low in this example with a rate of approximately 10% and are comparable between the two treatment groups.

3. SIMULATION STUDY

To compare the performance of the various methods (i.e. standard logrank test, Gray's test, Pepe & Mori's method, and conditional probability approach) in different competing risk settings, a simulation study is conducted to compare the power and type I error rate of the various tests for detecting treatment differences in terms of the primary event of interest. We consider a trial comparing a new therapy (treatment A) to a standard therapy (treatment B) in the presence of two types of competing causes of failure where type 1 failure is the primary event of interest and type 2 failure is the competing risk event. The simulation data are generated using a bivariate exponential model for the latent failure times. A sample size of $n = 200$ per arm is used in these simulations and 500 realizations for each parameter configuration are repeated to estimate the type I error rate and power of the different approaches. Table 3 and 4 present power based on empirical rejection probability at the two-sided 5% significance level under various different settings that correspond to different combinations of the relative treatment effects between treatment A and treatment B on the failure type 1 in the presence of failure type 2.

In Tables 3, the simulation results in terms of power are shown for the following six scenarios:

(1) For treatment A, there is a large benefit on type 1 failure , and no benefit on type 2 failure. And the risk of type 2 failure is high;

(2) For treatment A, there is a large benefit on type 1 failure , and no benefit on type 2 failure. And the risk of type 2 failure is moderate;

(3) For treatment A, there is a large benefit on type 1 failure , and no benefit on type 2 failure. And the risk of type 2 failure is low;

(4) For treatment A, there is a moderate benefit on type 1 failure , and no benefit on type 2 failure. And the risk of type 2 failure is high;

(5) For treatment A, there is a moderate benefit on type 1 failure , and no benefit on type 2 failure. And the risk of type 2 failure is moderate;

(6) For treatment A, there is a moderate benefit on type 1 failure , and no benefit on type 2 failure. And the risk of type 2 failure is low.

For example, in the first setting of the table, the hazard rates of type 1 failures are 1 and 1.6 for treatment A and treatment B, respectively, which indicates a large treatment benefit for treatment A as compared to treatment B. The hazard rates of type 2 failures are 1 and 1 for treatment A and B, respectively, which indicates there is no treatment benefit on type 2 failures for treatment A as compared to treatment B, while the amount of failure caused by type 2 events are relatively high. The power based on the empirical rejection probability are 0.948, 0.778, 0.198 and 0.904 for the standard logrank test, Gray's test, Pepe & Mori's method and conditional probability method, respectively. The standard logrank test, which compares the cause-specific hazards and treats the competing risk events as censored observations, demonstrates the highest power in detecting the treatment difference on type 1 failures, the primary events of interest. On the other hand, the Pepe & Mori's method that compares the cumulative incidence functions through a weighted score between the two cumulative incidence curves, provides the lowest power among the four different methods. Gray's test that compares the cumulative incidence functions through a modified log-rank score, and the conditional probability approach that compares the conditional probability functions, both have reasonably good power property with that of the conditional probability approach more close to that of the standard logrank approach. These results somehow indicate that the standard logrank test may be useful in discovering the biological mechanisms that underly the specific event by ignoring the existence of other competing types of events. While the three newer methods developed taking account of all types of events through direct comparisons of the probability of failure or conditional probability of failure are more of the comprehensive and adjusted approaches; the only concern however lies with Pepe & Mori's method which seems to be extremely overly-adjusted and leads to a substantial loss of power. In the second and third settings of the table, as the risk of type 2 failures is decreasing while the other assumptions are held unchanged compared to the first setting, the power performance is seemingly increasing for the standard log rank test, Gray's test and the conditional probability method, where the conditional probability method is almost as powerful as the standard logrank test. Meanwhile, the Pepe & Mori's method remains to be a poor power performer. In settings 4, 5 and 6 of the table, where there is only a moderate treatment benefit on type 1 failure and no benefit on type 2 failure for treatment A and the risk of competing events varies from high grade to low grade, similar conclusions can be reached in terms of the power performance of these four various approaches.

In Table 4, the simulation results in terms of power are shown for the following three scenarios:

(7) For treatment A, there is no benefit on type 1 failure, and a moderate benefit on type 2 failure;

(8) For treatment A, there is no benefit on type 1 failure, and a large benefit on type 2 failure;

(9) For treatment A, there is no benefit on type 1 failure, and a very large benefit on

type 2 failure.

In the three settings demonstrated in Table 4, there is no treatment benefit on type 1 failures for treatment A, however, there are various degree of treatment benefit simulated on type 2 failure data. It can be seen that Gray's test is most sensitive in terms of power when there is a larger benefit on competing risk events with power dramatically increasing as the treatment difference on the type 2 failures is drastically increased between the two treatments. As a comparison, the standard logrank rest is almost insensitive to the difference in competing risk events regardless of the magnitude of such difference. Whereas, Pepe & Mori's method and the conditional probability method provide a relatively moderate adjustment on power when there is a larger benefit on the competing risk events.

In Table 5, the simulation results in terms of type I error rate are shown for the following three scenarios:

(10) There is no benefit on either type 1 or type 2 failures. And the risk of type 2 failure is high.

(11) There is no benefit on either type 1 or type 2 failures. And the risk of type 2 failure is moderate.

(12) There is no benefit on either type 1 or type 2 failures. And the risk of type 2 failure is low.

In the three settings presented in Table 5, there is no treatment benefit on either type of failures for treatment A except that there are various degree of risk embedded on type 2 failure data. It can be seen that the standard logrank test causes slight inflation on Type I error rate under all three scenarios. Whereas, the rest three methods that take into consideration of the existence of the competing risk events control Type I error rate reasonably well. In conclusion, when putting both power and type I error rate into the perspective, Gray's test and the conditional probability approach have relatively more robust profile and therefore are recommended approaches in the competing risk framework based on this simulation work. Since it is also important to understand the underlying biological mechanism of the specific event, it is always desirable to present the standard logrank analysis as a complementary perspective.

4. Discussion

In this paper, three competing newer approaches are discussed and compared to the standard logrank test through simulations in the analysis of time-to-event endpoint when there exists a confounding factor of the competing risk events. In particular, Gray's k-sample test (1988) compares the weighted averages of the $\gamma_i(t)$ for the event of interest, where $\gamma_i(t)$ is the hazard of the sub-distribution for event of interest in treatment group i. This can be interpreted as the probability of observing an event of interest in the next time interval while knowing that either the event of interest did not happen until then or that a competing risk event was observed. Instead, Pepe & Mori (1993) proposed a different class of test statistics, which were not based on ranks, for the comparison of cumulative incidence functions and conditional probability functions. Of these, one uses a score function that is a weighted area between the two cumulative incidence functions for treatment comparison; another uses the concept of conditional probability that is simply the proportion of patients who have experienced the event of interest among those surviving the competing risk events, which has a rather straightforward interpretation.

Freidlin & Korn (2005) undertook a large number of simulations from correlated bi-

variate exponential data and concluded that the cause-specific log-rank test is robust in the sense of preserving the nominal level of the test and has good power properties for testing the treatment differences. Their work however only recommended the use of the cumulative incidence based approach as one of the components required for a comprehensive comparison of treatment groups in the presence of multiple types of failure. Pintilie (2007) also indicated that the standard logrank test and the cumulative incidence based approaches give different information and thus it is necessary to do both. He viewed the the cause-specific hazard modeling as providing information on how covariate influences the outcome of interest in a laboratory setting where competing risks do not exist. On the other hand, the modeling of the hazard of the sub-distribution gives the real effect seen in the dataset under study. Gooley et al (1999) thought the logrank test is appropriate for inference in treatment effect evaluation on the hazard of failure from the cause of interest, since it is a function solely of the hazard of failure from the cause of interest and failures from the competing risk can therefore be censored. However, if the interest lies on the comparison of the actual probability of failure between two groups, it is appropriate to use tests based on cumulative incidence functions, that is those that depend on the hazards of each type of failure. He stressed the importance of understanding the relationship of treatment to the various causes of failure in order to draw appropriate conclusions.

It has been demonstrated in our simulation work that when there is a large treatment benefit on the event of interest, the standard logrank test, Gray's test and the conditional probability approach all provide reasonably good power properties when the risk of competing events varies from low risk to high risk however are balanced between the two treatment groups. Among these three, the power performance of the standard logrank test is the best and that of the conditional probability approach is slightly lessened compared to the standard logrank test. As a contrast, the power performance of Pepe & Mori's method is extremely unsatisfactory in all these simulation settings being studied. At the mean time, when there is no treatment benefit existing for both the event of interest and the competing risk event, the standard logrank test slightly inflates the type I error rate while the other three methods provide slightly better control on the type I error rate. It is worth noting that when there is no treatment benefit for the event of interest however there remains a substantial difference in terms of the risk for competing events, Gray's test would pick up the difference caused by the competing risk and produces a very high power performance. In comparison, the standard logrank test shows indifference to the impact of the competing risk benefit. As a conclusion, when looking at the comprehensive statistical properties of the various approaches, Gray's test and the conditional probability approach show relatively more robust profile in the competing risk framework based on simulations. However, the importance of the standard logrank analysis should not be undermined, as it provides some important clues to assist in the proper understanding of the underlying biological mechanism of the specific event. Therefore, it is necessary to provide both the standard analysis and the newer analysis approaches comparing either the cumulative incidence functions or the conditional probability functions. In all occasions, Pepe & Mori's method should be used with extreme caution considering it's unrealistically low power performances.

REFERENCES

Feidlin B, Korn EL (2005). "Testing treatment effects in the presence of competing risks." *Statistics in Medicine*, **Vol. 24**: 1703-1712.

Gaynor JJ, Feuer EJ, Tan CC, Wu DH, Little CR, Straus DJ, Clarkson BD, Brennan MF (1993). "On the use of cause-specific failure and conditional probabilities: examples from clinical oncology data." *Journal of the American Statistical Association.*, **Vol. 88, No. 422**: 400-409.

Gooley TA, Leisenring W, Crowley J, Storer BE (1999). "Estimation of failure probabilities in the presence of

competing risks: new representation of old estimators." *Statistics in Medicine*, **Vol. 18**: 695-706.

Gray RJ (1988). "A class of k-sample tests for comparing the cumulative incidence of a competing risk." *Annals of Statistics*, **Vol. 16**: 1141-1154.

Kim HT (2007). "Cumulative Incidence in competing risks data and competing risks regression analysis." *Clinical Cancer Research* , **Vol. 13, No. 2**: 559-565.

Pintilie M (2007). "Competing Risks, A practical Perspective." *Wiley*.

Pepe MS, Mori M (1993). "Kaplan-Meier, marginal or conditional probability curves in summarizing competing risks failure time data?" *Statistics in Medicine*, **Vol. 12**: 737-751.

Pintilie M (2002). "Dealing with competing risks: testing covariates and calculating sample size." *Statistics in Medicine*, **Vol. 21**: 3317-3324.

Table 1: Progression free survival data presentation for the oncology trial example.

	Progression Death	Switching cancer treatment	Censored observations
Investigational therapy	185	22	12
Standard therapy	174	27	18

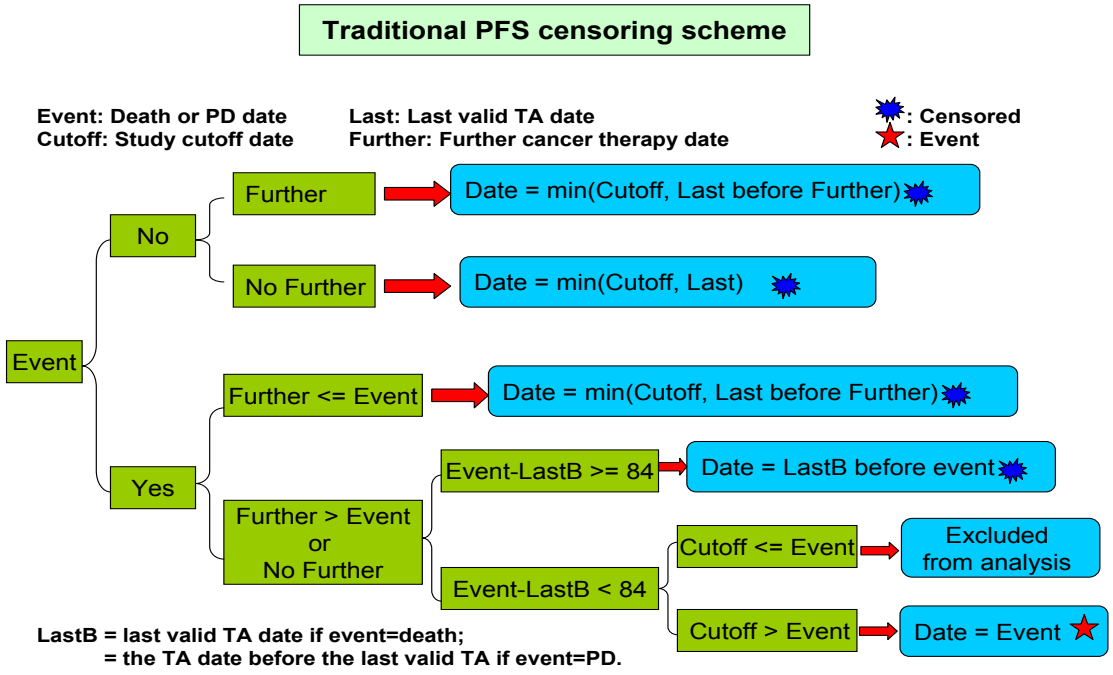

Figure 1: Traditional progression free survival censoring scheme

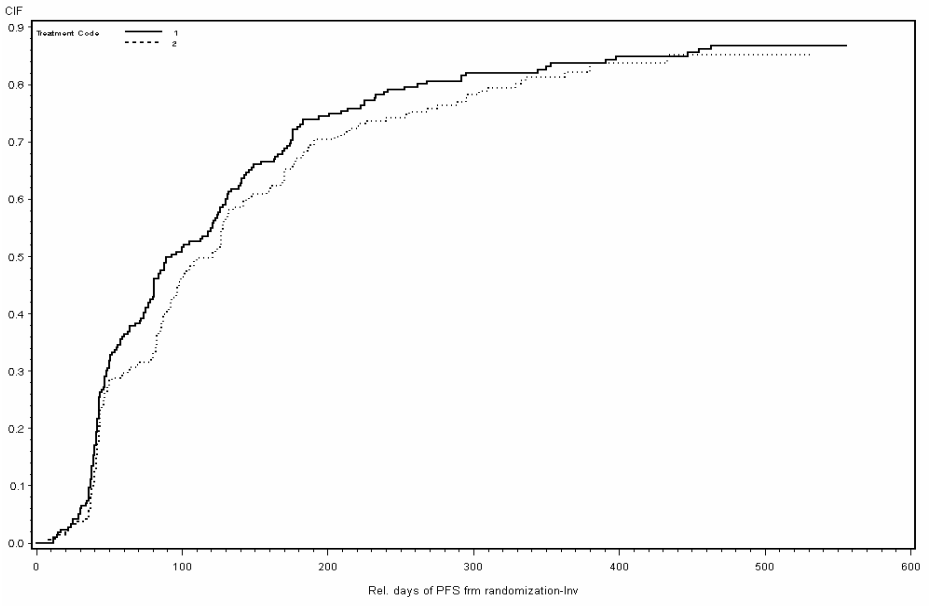

Figure 2: Cumulative incidence function for event type I by treatment group

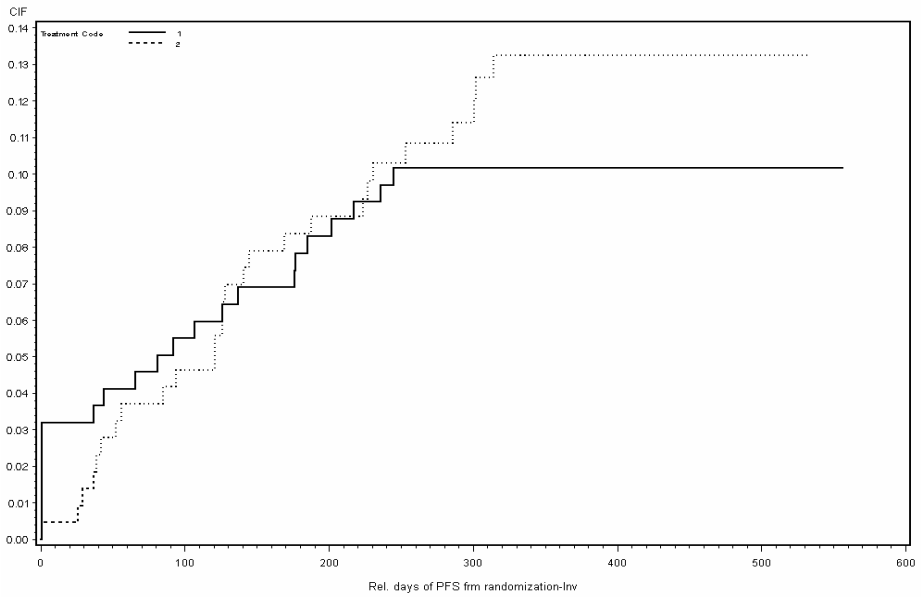

Figure 3: Cumulative incidence function for event type II by treatment group

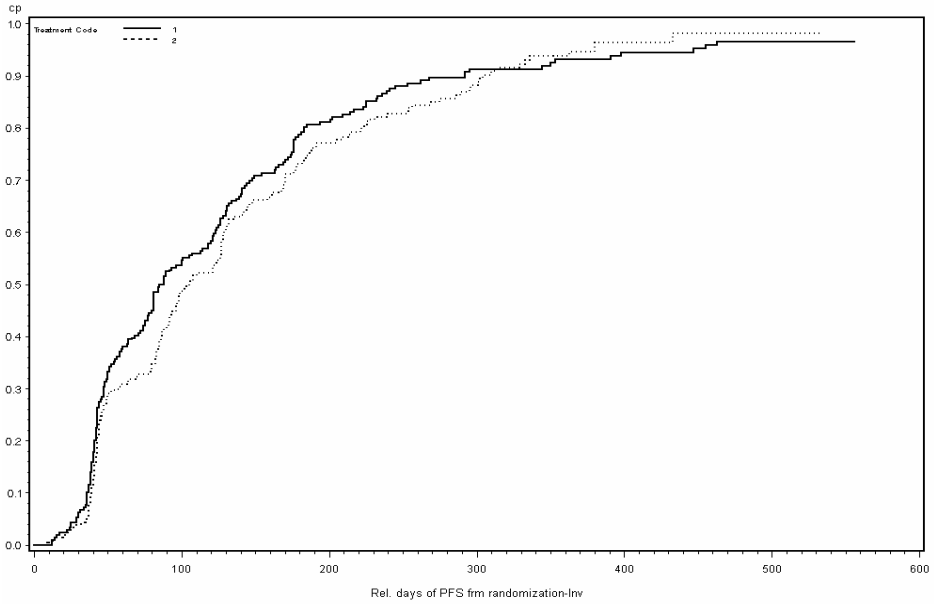

Figure 4: Conditional probability for event type II by treatment group

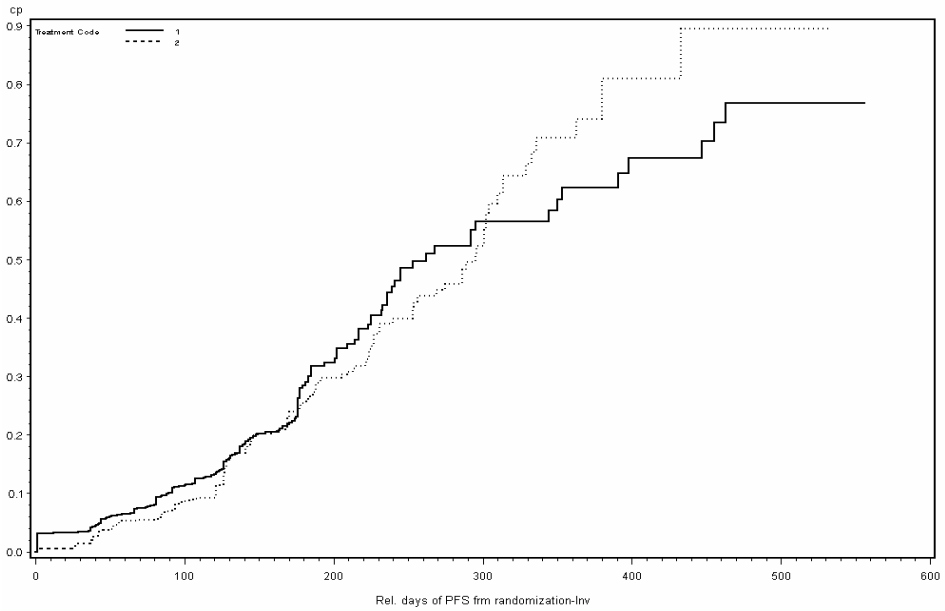

Figure 5: Conditional probability for event type II by treatment group

Table 2: Oncology example - results of the standard logrank test and the three newer approaches

	Logrank test	Gray's test	Pepe & Mori's method	Conditional probability
Progression or death	0.199	0.235	0.500	0.207
Switching cancer treatment		0.424	0.651	0.726

The first row contains p-values of applying the four different methods for comparing the investigational cancer treatment to the standard cancer treatment for the primary event of interest - progression or death, whichever is earlier.

The second row contains p-values of applying the four different methods for for treatment comparison for the competing risk event - switching cancer treatment prior to progression/death.

Table 3: Simulation results in terms of power when there is moderate to large benefit on Type 1 failure, and no benefit on Type 2 failure

	Logrank test	Gray's test	Pepe and Mori's	Conditional probability
(1) $\lambda_1^A = 1, \lambda_1^B = 1.6, \lambda_2^A = 1, \lambda_2^B = 1$	0.948	0.778	0.198	0.904
(2) $\lambda_1^A = 1, \lambda_1^B = 1.6, \lambda_2^A = 0.5, \lambda_2^B = 0.5$	0.980	0.848	0.154	0.960
(3) $\lambda_1^A = 1, \lambda_1^B = 1.6, \lambda_2^A = 0.3, \lambda_2^B = 0.3$	0.988	0.886	0.156	0.982
(4) $\lambda_1^A = 1, \lambda_1^B = 1.2, \lambda_2^A = 1, \lambda_2^B = 1$	0.278	0.200	0.010	0.250
(5) $\lambda_1^A = 1, \lambda_1^B = 1.2, \lambda_2^A = 0.5, \lambda_2^B = 0.5$	0.324	0.184	0.002	0.304
(6) $\lambda_1^A = 1, \lambda_1^B = 1.2, \lambda_2^A = 0.3, \lambda_2^B = 0.3$	0.342	0.222	0.002	0.332

λ_1^A and λ_1^B are hazard rates of type 1 failure for Treatment A and B, respectively. λ_2^A and λ_2^B are hazard rates of type 2 failure for Treatment A and B, respectively.

Table 4: Simulation results in terms of power when there is no benefit on Type 1 failure, and moderate to very large benefit on Type 2 failure

	Logrank test	Gray's test	Pepe and Mori's	Conditional probability
(7) $\lambda_1^A = 1, \lambda_1^B = 1, \lambda_2^A = 1, \lambda_2^B = 1.2$	0.066	0.118	0.010	0.058
(8) $\lambda_1^A = 1, \lambda_1^B = 1, \lambda_2^A = 1, \lambda_2^B = 1.6$	0.066	0.494	0.046	0.144
(9) $\lambda_1^A = 1, \lambda_1^B = 1, \lambda_2^A = 0.5, \lambda_2^B = 1.6$	0.054	0.990	0.396	0.468

λ_1^A and λ_1^B are hazard rates of type 1 failure for Treatment A and B, respectively. λ_2^A and λ_2^B are hazard rates of type 2 failure for Treatment A and B, respectively.

Table 5: Simulation results in terms of Type I error rate

	Logrank test	Gray's test	Pepe and Mori's	Conditional probability
(10) $\lambda_1^A = 1, \lambda_1^B = 1, \lambda_2^A = 1, \lambda_2^B = 1$	0.066	0.054	0.000	0.056
(11) $\lambda_1^A = 1, \lambda_1^B = 1, \lambda_2^A = 0.5, \lambda_2^B = 0.5$	0.054	0.054	0.000	0.046
(12) $\lambda_1^A = 1, \lambda_1^B = 1, \lambda_2^A = 0.3, \lambda_2^B = 0.3$	0.070	0.060	0.000	0.056

λ_1^A and λ_1^B are hazard rates of type 1 failure for Treatment A and B, respectively. λ_2^A and λ_2^B are hazard rates of type 2 failure for Treatment A and B, respectively.