응용통계연구 제 **20**권 *2*호, *2007*년, *pp. 357–371*

경쟁위험 하에서의 누적발생함수 추정량 성능 비교

김동욱[1] 안치경[2]

요 약

경쟁위험(competing risk) 하에서의 누적발생함수(cumulative incidence function)는 일반적으로 비모수적 방법으로 추정된다. 그러나 관심 있는 원인에 의한 사건이 다른 원인에 의한 사건보다 상대적으로 적게 발생하는 경우에 비모수적 방법으로 추정된 누적발생함수는 이산성으로 인해 다소 정확하지 않게 된다. 이와 같은 경우에 Bryant와 Diagnam(2004)는 관심 있는 원인에 대한 원인특정적 위험함수(cause-specific hazard function)를 모수적으로 모형화하고 다른 원인에 의한 사건은 비모수적으로 추정하는 준모수적 방법을 제안했다. 본 연구에서는 준모수적 누적발생함수 추정량을 재표현하고 와이블분포모형과 대수 정규분포모형으로 확장하였다. 또한 대수 정규분포 원인특정적 위험모형일 경우 누적발생함수에 대한 비모수적 추정량, 와이블분포 준모수적 추정량과 대수 정규분포 준모수적 추정량의 효율성을 비교하며 준모수적 추정량의 성능과 모형 오설정이 미치는 영향을 살펴보았다.

주요용어: 누적발생함수, 경쟁위험, 준모수적 추정량, 상대효율, 와이블분포, 대수 정규분포.

1. 서론

생존시간에 대한 연구에서 연구자가 사건 전체에 대해서 관심이 있는 것이 아니라 특정 원인에 의한 사건에만 관심이 있는 경우가 있다. 여기서 관심원인 이외의 다른 원인에 의한 사건을 경쟁위험(competing risk)이라 한다. 경쟁위험과 관련된 초창기 연구에서는 경쟁위험에 의한 사건 발생을 중도절단으로 처리한 KM 추정량(Kaplan-Meier estimator)이 사용되었다. 그리고 1-KM 추정값을 특정 원인에 의한 누적 발생확률로 해석하였으나 이후 잘못된 해석이라는 것이 몇몇 연구를 통해서 밝혀졌다(Gooley 등, 1999; Gaynor 등, 1993). 특히, Gooley 등(1999)은 1-KM 추정량이 누적 발생확률을 과대 추정한다는 사실을 도약함수(jump function)를 사용하여 수식으로 표현하였다.

Kalbfleisch와 Prentice(1980)는 경쟁위험 하에서 특정원인에 의한 사건발생 형태를 설명할 수 있는 누적 발생함수(cumulative incidence function: CIF)를 정의하였다. 누적발생함수는 특정 원인에 기인한 사건 발생의 누적확률로 해석할 수 있고 추정 가능한 확률이다.

1) (110-745) 서울특별시 종로구 명륜동 3가 53, 성균관대학교 통계학과, 교수
E-mail: dkim@skku.edu
2) (110-745) 서울특별시 종로구 명륜동 3가 53, 성균관대학교 통계학과, 박사과정
E-mail: acg11@skku.edu

이러한 이유로 경쟁위험 하에서의 생존분포에 대한 연구에서 누적발생함수가 많이 사용된다(Gray, 1988; Korn과 Dorey, 1992; Gaynor 등, 1993; Pepe와 Mori, 1993).

누적발생함수는 비모수적 방법으로 주로 추정된다. 그러나 관심 있는 원인에 의한 사건이 다른 원인에 의한 사건보다 상대적으로 덜 발생한다면 비모수적 방법은 그 추정량 분포의 이산성(discreteness)으로 인해 관심 있는 원인에 대한 누적발생함수를 정확하게 추정하지 못한다. 이 경우 관심 있는 원인특정적 위험함수는 모수적으로 모형화하는 반면에 다른 원인에 의한 원인특정적 위험함수들은 비모수적으로 처리할 수 있다. 이렇게 얻어진 누적발생함수 추정량은 계산 가능하며, 비모수적 CIF 추정량보다 효율적이다. 특히 소표본일 경우 비모수적 CIF 추정량은 이산성이 강해 사건이 발생하는 시점에서 큰 도약(jump)이 발생하는 계단함수(step function)가 된다. 그러나 준모수적 CIF 추정량은 자료범위내의 임의의 시점에서 CIF 추정값을 계산할 수 있으며 더 정확하게 추정할 수 있다.

본 연구에서는 2장에서 원인특정적 위험함수와 누적발생함수 그리고 그 추정량에 대해 간단히 살펴본다. 3장에서는 누적발생함수의 준모수적 모형에 대하여 설명하고 준모수적 모형에 따른 누적발생함수의 형태를 알아본다. 그리고 Bryant와 Diagnam(2004)에서 표현된 일반적인 형태의 준모수적 CIF 추정량을 재표현한다. 4장에서 원인특정적 위험함수가 대수 정규분포를 따를 때 준모수적 CIF의 추정량의 성능을 연구한다. 즉, 비모수적 CIF에 대한 준모수적 CIF 추정량의 상대효율을 모의실험으로 구하여 준모수적 CIF 추정량의 효율성과 모형 오설정(model misspecification)에 대한 영향을 연구한다. 5장은 호지킨병 자료를 이용한 실증분석이며 6장은 결론이다.

2. 누적발생함수의 비모수적 추정

2.1. 원인특정적 위험함수

경쟁위험 하에서 연구대상자는 여러 위험에 동시에 노출된다. 여기서 특정 원인의 위험에만 연구자가 관심이 있는 경우 다중 위험에 노출된 대상 중에서 특정 원인에 의한 위험만이 관측 가능할 것이다. Prentice 등(1978)은 특정 원인에 의한 위험을 정의하기 위해 원인특정적 위험함수(cause-specific hazard function) $\lambda_l(t)$를 다음과 같이 정의하였다.

$$\lambda_l(t) = \lim_{\Delta t \to 0} \frac{Pr[t \le T < t + \Delta t, L = l \mid T \ge t]}{\Delta t}, \quad l = 1, \ldots, C. \tag{2.1}$$

여기서 T는 생존 시간이고 L은 사건 발생의 원인이며 C는 원인의 수를 나타낸다. 원인특정적 위험함수 $\lambda_l(t)$는 t시점까지 사건이 발생하지 않았을 때, l번째 원인으로 바로 다음 순간에 사건이 발생할 조건부 확률이다. 전체 위험함수(overall hazard function) $\lambda(t)$와 $\lambda_l(t)$는 $\lambda(t) = \sum_{l=1}^{C} \lambda_l(t)$ 의 관계를 가진다.

r개의 중복이 없는 생존시간, $0 < t_{(1)} < t_{(2)} < \cdots < t_{(r)}$, 에서 $t_{(0)} = 0$이라 할 때,

$$\lambda_l(t) = \begin{cases} \lambda_{lj}\,, \; t = t_{(j)}\text{인 경우}, \; l = 1, \ldots, C, \; j = 1, \ldots, r, \\ 0, \quad \text{그 외} \end{cases} \tag{2.2}$$

이다. λ_{lj}는 $t_{(j-1)}$시점까지 사건이 발생하지 않았을 때, $t_{(j)}$시점에서 l번째 원인으로 사건이 발생할 조건부 확률이다. λ_{lj}의 ML 추정량은 $\hat{\lambda}_{lj} = d_{lj}/n_j$이며 d_{lj}는 t_j시점에서 l번째 원인으로 사건이 발생한 사람의 수이며 n_j는 t_j시점 직전까지 위험에 노출된 사람의 수를 나타낸다.

2.2. 누적발생함수

원인 l의 조(crude) 누적발생함수 $I_l(t)$는 다음과 같다.

$$I_l(t) = \Pr(T \le t,\ L = l), \quad l = 1, \ldots, C.$$

조 누적발생함수 $I_l(t)$는 C개의 원인이 존재하는 경쟁위험 상황에서 t시점 이전에 특정한 원인 l로 사건이 발생할 누적확률이다. 이 CIF는 추정가능하고, 경쟁위험 하에서 해석이 가능하다. 또한 각 생존시간 T_l들 사이에 독립성 가정이 없더라도 해석이 가능하다. 따라서 CIF는 원인특정적 누적 위험함수(cause-specific cumulative hazard function) $\Lambda_l(t)$와 같은 다른 추정 가능한 함수보다 활용도가 높다.

$$\Lambda_l(t) = \int_0^t \lambda_l(u)du\,, \qquad S_{cs.l}(t) = \exp[-\Lambda_l(t)]. \tag{2.3}$$

식 (2.3)의 원인특정적 누적위험함수를 이용하여 전체 생존함수(overall survival function)가 유도된다.

$$\Lambda(t) = \int_0^t \lambda(u)du = \sum_{l=1}^{C} \Lambda_l(t)\,, \qquad S(t) = \exp\left[-\sum_{l=1}^{C} \Lambda_{l(t)}\right] = \prod_{l=1}^{C} S_{cs.l}(t). \tag{2.4}$$

각각의 원인에 의한 위험이 독립이라면 $S_{cs.l}(t)$는 l번째 원인에 의한 생존함수로 생각할 수 있다. 원인의 개수 C가 2이상이고 각각의 원인에 대한 독립성 가정이 없다면, $S_{cs.l}(t)$는 l번째 원인에 의한 생존함수로 해석할 수가 없다(Tsiatis, 1975). 여기서 말하는 l번째 원인에 의한 생존함수는 순 생존확률을 의미하며 추정할 수 없는 확률이다.

원인특정적 실패 누적발생함수(cause-specific failure cumulative incidence function) $I_{cs.l}(t)$ 는 다음과 같이 정의된다.

$$I_{cs.l}(t) \;=\; 1 - S_{cs.l}(t) = \int_0^t S_{cs.l}(u)\lambda_l(u)du = \int_0^t f_{cs.l}(u)du,$$

여기서 $f_{cs.l}(u) = \lambda_l(u)S_{cs.l}(u)$이다.

각각의 원인에 의한 위험이 독립이라면 $S_{cs.l}(t)$는 l번째 원인에 의한 생존함수로 생각할 수 있으므로 $I_l(t)$는 생존시간 T_l의 분포함수로 볼 수 있다. 앞에서 살펴본 $S_{cs.l}(t)$의 경우와 마찬가지로 각 원인에 대한 독립성 가정이 없다면 $I_{cs.l}(t)$는 직접적으로 해석이 불가능한 함수이다.

2.3. 누적발생함수의 추정

원인 l의 조 누적발생함수 $I_l(t)$는 다음과 같이 정의될 수 있다.

$$I_l(t) = \int_0^t f_l(u)du = \int_0^t \lambda_l(u)S(u)du. \tag{2.5}$$

여기서 $S(t) = \exp\{-\int_0^t (\sum_{l=1}^{C} \lambda_l(u))du\}$이고 원인특정적 실패확률(cause-specific failure probability) $f_l(t)$는 $f_l(t) = \lambda_l(t)S(t)$이다.

식 (2.5)를 이산형으로 표현하면 다음과 같다.

$$\begin{aligned} f_l(t_{(j)}) &= \lambda_{lj} \cdot S(t_{(j-1)}), \\ I_l(t) &= \sum_{\{j|t_{(j)} \le t\}} \lambda_{lj} \cdot S(t_{(j-1)}). \end{aligned}$$

따라서 $I_l(t)$의 ML 추정량 $\hat{I}_l(t)$는 다음과 같이 구해진다(Marubini와 Valsecchi, 1995; Gooley 등, 1999).

$$\hat{I}_l(t) = \sum_{\{j\ |t_{(j)} \le t\}} \frac{d_{lj}}{n_j} \hat{S}(t_{(j-1)}). \tag{2.6}$$

여기서 전체 생존함수의 ML 추정량인 $\hat{S}(t_{(j-1)})$은 $\hat{S}(t_{(j-1)}) = \prod_{i=1}^{j-1}(1 - d_i/n_i)$이며 이는 $t_{(j)}$시점 직전에서의 KM 추정량이다. $\hat{I}_l(t)$를 crude CIF 추정량이라 한다.

3. 누적발생함수의 준모수적 모형

3.1. 준모수적 누적발생함수의 형태

누적발생함수의 준모수적(semi-parametric) 모형은 다음과 같이 가정한다(Bryant와 Dignam, 2004). 관심 있는 원인, 즉 1의 원인에 대한 원인특정적 위험함수에 대하여 모수적 모형을 $\lambda_1(t) = \lambda_1(t;\theta)$로 설정한다. 모수 θ는 자료로부터 추정되며, 다른 원인에 의한 사건은 비모수적 방법으로 다룬다. 또한 1의 원인이 아닌 모든 다른 실패의 원인은 하나의 범주로 묶어서, 총 원인의 개수 C는 $C = 2$로 생각한다.

모든 원인에 대한 위험함수인 $\lambda(\,\cdot\,;\theta)$는 $\lambda(\,\cdot\,;\theta) = \lambda_1(\,\cdot\,;\theta) + \lambda_2(\,\cdot\,)$이 되며 모든 원인에 의한 생존함수 $S(t;\theta)$는 다음과 같이 된다.

$$S(\,t\,;\,\theta\,) = S_{cs.1}(t\,;\,\theta) \cdot S_{cs.2}(t).$$

따라서 원인 1에 대한 준모수적 누적발생함수는 식 (2.5)으로 부터 다음과 같다.

$$\begin{aligned} I_{sp.1}(t;\theta) &= \int_0^t \lambda_1(u;\theta)S(u;\theta)du \\ &= \int_0^t S_{cs.2}(u)\ f_{cs.1}(u;\theta)\,du \end{aligned} \tag{3.1}$$

여기서 $f_{cs.1}(u;\theta) = \lambda_1(u;\theta)S_{cs.1}(u;\theta)$이다. $f_{cs.1}(u;\theta)$는 각 원인 간에 독립성 가정이 충족된다면 관심 있는 1의 원인에 대한 생존시간 T_1의 확률밀도함수로 볼 수 있으나 독립성 가정 없이는 해석이 불가능한 함수다.

식 (3.1)을 통해 준모수적 누적발생함수 추정량 $\hat{I}_{sp.1}(t)$를 구할 수 있다.

$$\hat{I}_{sp.1}(t) = \int_0^t \hat{S}_{cs.2}(u)\, f_{cs.1}(u;\hat{\theta})\, du \tag{3.2}$$

여기서 $\hat{\theta}$는 θ의 ML 추정량이고 $\hat{S}_{cs.2}(\cdot)$은 원인 1에 의한 사건을 중도절단으로 계산한 원인 2에 의한 생존 시간의 KM 추정량이다.

3.2. 준모수적 CIF 추정량의 재표현

식 (3.2)의 $\hat{S}_{cs.2}(u)$는 조각별 연속(piecewise continuous)이기 때문에 임의의 $t_{(h)}$에서의 준모수적 CIF 추정량은 다음과 같이 표현할 수 있다.

$$\begin{aligned}
\hat{I}_{sp.1}(t_{(h)}) &= \int_0^{t_{(h)}} \hat{S}_{cs.2}(u)\cdot f_{cs.1}(u;\hat{\theta})du \\
&= \sum_{j=0}^{h-1} \hat{S}_{cs.2}(t_{(j)}) \int_{t_{(j)}}^{t_{(j+1)}} f_{cs.1}(u;\hat{\theta})du \\
&= \sum_{j=0}^{h-1} \hat{S}_{cs.2}(t_{(j)})\Big\{I_{cs.1}(t_{(j+1)};\hat{\theta}) - I_{cs.1}(t_{(j)};\hat{\theta})\Big\}.
\end{aligned} \tag{3.3}$$

여기서 $\hat{S}_{cs.2}(t_{(j)})$는 2의 원인만을 발생사건으로 보고 나머지는 중도절단으로 보았을 때의 KM 추정량이고, $I_{cs.1}(t;\hat{\theta})$는 $I_{cs.1}(t;\hat{\theta}) = \int_0^t f_{cs.1}(u;\hat{\theta})du$이다.

위의 준모수적 CIF 추정량은 $I_{cs.1}(t)$를 알고 있을 때 편리하다. 원인특정적 위험함수가 $\lambda_1(t;\theta) = \theta$일 때 임의의 $t_{(h)}$에서의 준모수적 CIF 추정량은

$$\hat{I}_{sp.1}(t_{(h)}) = \sum_{j=0}^{h-1} \hat{S}_{cs.2}(t_{(j)}) \cdot \left(e^{-\hat{\theta}t_{(j)}} - e^{-\hat{\theta}t_{(j+1)}}\right)$$

이 된다. 또한 관심 있는 원인이 와이블분포를 따를 때 임의의 $t_{(h)}$에서의 준모수적 CIF 추정량은

$$\hat{I}_{sp.1}(t_{(h)}) = \sum_{j=0}^{h-1} \hat{S}_{cs.2}(t_{(j)}) \cdot \left(\exp[-(\hat{\lambda}t_{(j)})^{\hat{p}}] - \exp[-(\hat{\lambda}t_{(j+1)})^{\hat{p}}]\right)$$

이 된다. 준모수적 CIF 추정량의 장점 중 하나는 사건이 발생하지 않은 임의의 시점에서도 CIF의 추정량을 계산할 수 있다는 것이다. 식 (3.3)을 확장하여 관측되지 않은 시점에서의 CIF 추정량을 다음과 같이 구할 수 있다. $t_{(h)} < t < t_{(h+1)}$ 일 때,

$$\begin{aligned}
\hat{I}_{sp.1}(t) &= \int_0^t \hat{S}_{cs.2}(u)\cdot f_{cs.1}(u;\hat{\theta})du \\
&= \hat{I}_{sp.1}(t_{(h)}) + \hat{S}_{cs.2}(t_{(h)})\Big\{I_{cs.1}(t;\hat{\theta}) - I_{cs.1}(t_{(h)};\hat{\theta})\Big\}.
\end{aligned} \tag{3.4}$$

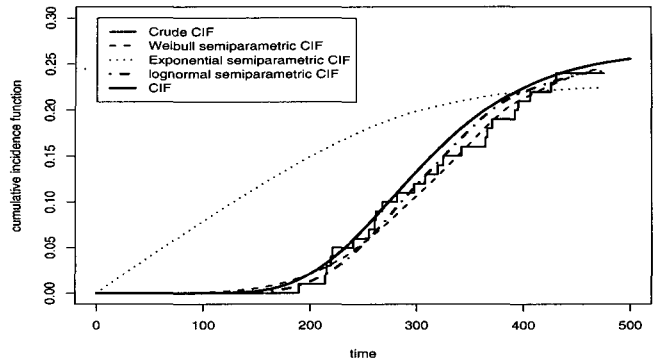


그림 4.1: 4개의 CIF 추정량과 실제 CIF의 분포 (표본수 100)

4. crude CIF 추정량에 대한 준모수적 CIF 추정량의 상대효율

준모수적 CIF 추정량은 crude CIF 추정량과 비교해 볼 때 CIF 추정을 위해 원인특정적 위험함수에 대한 가정, 즉 관심 있는 1의 원인에 대한 생존시간 T_1의 분포에 대한 가정이 더 필요하다. 생존시간 T_1과 T_2가 이변량 대수 정규분포를 따를 때, T_1의 분포를 와이블로 가정할 경우 모형 오설정에 따른 준모수적 CIF 추정량의 성능을 알아본다. 또한 생존시간의 분포에 대한 가정을 하지 않은 crude CIF 추정량에 대한 와이블 준모수적 CIF 추정량의 상대효율을 구하려고 한다. 그리고 T_1의 분포를 대수 정규분포로 바르게 가정한 경우 crude CIF 추정량에 대한 대수 정규분포 준모수적 CIF 추정량의 상대효율도 계산하여 준모수적 CIF 추정량의 성능을 비교한다.

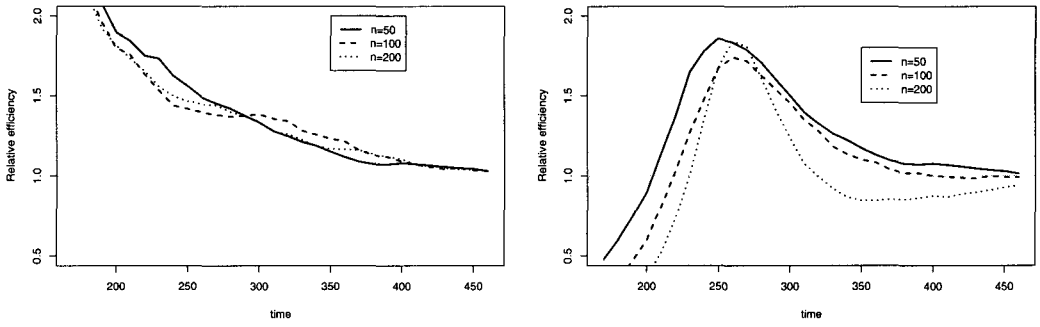
4.1. 이변량 대수 정규난수자료 발생

모의실험을 위해 이변량 대수 정규분포를 따르는 n개의 중도절단이 없는 생존시간 난수를 다음과 같이 발생시킨다. 평균벡터는 $(5.9, 5.7)'$이며 대각원소가 0.1이고 비대각원소가 0.05인 2×2 분산-공분산 행렬을 가지는 정규난수 벡터 $(Y_1, Y_2)'$를 발생시킨다. 잠재적인 생존시간 T_1과 T_2를 각각 $T_1 = \exp(Y_1)$, $T_2 = \exp(Y_2)$로 한다. 그 후 생존시간 T는 T_1과 T_2중 최소값으로 하고 발생 원인은 $T_1 < T_2$이면 $L = 1$, $T_1 > T_2$면 $L = 2$로 한다.

위의 과정으로 표본수가 100인 난수자료를 발생시켜서 실제 CIF 추정량의 분포와 함께 crude CIF 추정량, 1의 원인을 각각 지수분포, 와이블분포, 대수 정규분포로 가정한 준모수적 추정량의 분포를 그림 4.1에 나타냈다. 지수분포를 가정한 준모수적 CIF 추정값은 이 자료에서 실제 CIF와 많은 차이가 나는 것을 알 수 있다.

4.2. 중도절단이 없는 자료에서의 상대효율 비교

위의 모의실험 자료에서 지수분포를 가정한 준모수적 CIF 추정량은 실제 CIF에 잘 적합 되지 않으므로, crude CIF에 대한 와이블분포와 대수 정규분포 준모수적 CIF 추정량의



(a) 대수정규분포 가정 (b) 와이블분포 가정

그림 4.2: $\hat{I}_1(t)$에 대한 $\hat{I}_{sp.1}(t;\hat{\theta})$의 상대효율

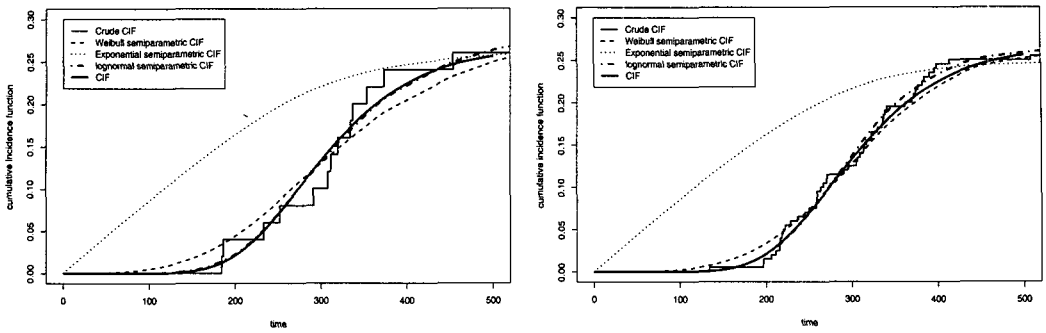
상대효율을 각각 구하였다. 즉, 이변량 대수 정규난수를 50, 100, 200개 발생 시켜서 특정 시점(t)에서의 crude CIF 추정값 $\hat{I}_1(t)$와 실제 CIF 값 $I_1(t)$ 그리고 준모수적 CIF 추정값 $\hat{I}_{sp.1}(t;\hat{\theta})$을 와이블분포로 가정한 경우와 대수 정규분포로 가정한 경우에 대해 각각 계산하였다.

그 후 해당 시점에서의 각 추정량의 제곱오차 $(\hat{I}_1(t) - I_1(t))^2$과 $(\hat{I}_{sp.1}(t;\hat{\theta}) - I_1(t))^2$을 2000번 반복하여 각 추정량의 제곱오차의 합을 계산한 후, crude CIF 추정량에 대해 와이블분포와 대수 정규분포 준모수적 CIF 추정량의 상대효율을 계산하였다.

그림 4.2(a)는 표본수별로 시간에 따른 $\hat{I}_1(t;\hat{\theta})$에 대한 대수 정규 분포가정 준모수적 CIF 추정량 $\hat{I}_{sp.1}(t;\hat{\theta})$의 상대효율을 나타낸다. 위의 결과를 보면 T_1의 분포를 대수 정규분포로 바르게 가정한 대수 정규분포 준모수적 CIF 추정량은 crude CIF 추정량보다 효율성이 상당히 높은 것을 볼 수 있다. 특히 시간의 초반부에서 효율성이 더 높은데, 이는 Bryant와 Diagnam(2004)에서 관심 있는 원인의 생존분포가 지수분포일 경우 crude CIF 추정량에 대한 지수분포 준모수적 CIF 추정량의 상대효율이 시간의 초반부에서 높은 것과 일치한다.

그림 4.2(b)는 관심 있는 원인에 대한 생존시간 분포를 와이블분포로 잘못 가정한 와이블분포가정 준모수적 CIF 추정량의 효율성을 나타낸다. 표본수가 200인 경우 crude CIF 추정량의 분포는 이산성이 약해지므로 crude CIF 추정량과 비교해서 와이블분포 준모수적 CIF 추정량은 시간의 후반부에서는 효율성이 좋지 않다. 그러나 표본수가 50인 경우와 같이 표본수가 적고 관심 있는 원인에 의한 사건이 경쟁위험에 비해 적게 발생하는 경우 T_1의 분포를 잘못 가정하더라도 전반적으로 crude CIF 추정량보다 효율성이 좋게 나오는 것을 알 수 있다.

그림 4.3은 표본수가 각각 50, 200일 때 4개의 CIF 추정량과 실제의 CIF를 비교하였다. 표본수가 50일 때는 crude CIF 추정값의 이산성이 큰 것을 확인할 수 있다. 반면에 표본수가 200정도 되면 crude CIF 추정량은 덜 이산적인 분포를 한다. 초반부의 crude CIF 추정량이 와이블분포 준모수적 CIF 추정량보다 효율성이 높게 나온 것은 초반부에서 crude CIF 추정값은 원인 1에 의한 사건이 발생하지 않아 추정값이 0인데 반해 와이블분포 준모수적



(a) 표본수 50 (b) 표본수 200

그림 4.3: 4개의 CIF 추정량과 실제 CIF의 비교

CIF 추정값은 과대추정되는 것에 기인한다고 판단된다.

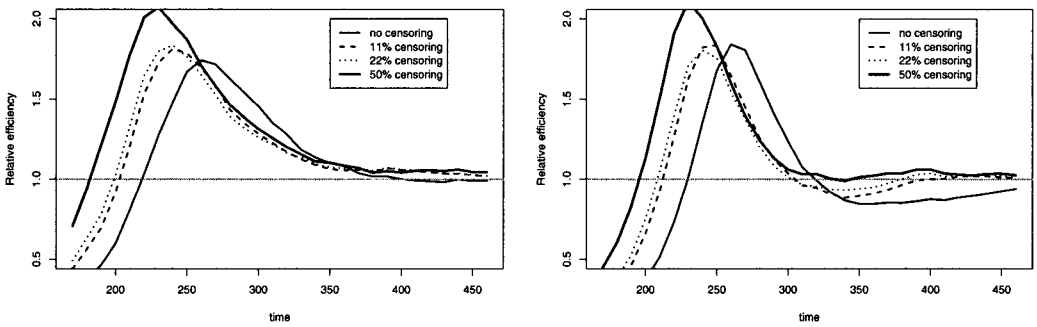
4.3. 중도절단이 있는 자료에서의 상대효율 비교

앞 절에서는 중도절단이 없는 완전 자료에서 상대효율을 연구하였다. 이번에는 중도절단이 있는 경우로 확장시켰다. 앞에서 발생시킨 이변량 대수 정규난수와 함께 잠재적인 중도절단 시간으로써 균일분포 난수 $C \sim U(a, 600)$을 발생시켰다. 실제 생존시간 T와 잠재적인 중도절단 시간 C에 대해 관측 생존시간 $T^* = \min(T, C)$로 하였고 C가 T보다 작으면 중도절단(censoring)으로 처리하였다. 또한 a값을 바꾸면서 중도절단의 비율을 조사하였다. 10만 번 이상 반복하여 표본수 100의 난수를 발생시킨 결과 평균적으로 $a = 300$일 때 중도절단 비율이 11%, $a = 220$일 때 중도절단 비율이 22%, $a = 0$일 때 중도절단 비율이 50%로 나타났다.

표본수와 중도절단의 비율에 따라 관심 있는 원인에 의한 사건의 평균 발생수는 표 4.1과 같이 나타났다. 같은 표본수라도 중도절단의 비율이 높을수록 관심 있는 원인에 의한 사건이 적어져서 그 분포가 더 이산적으로 되며, crude CIF는 덜 정확해진다.

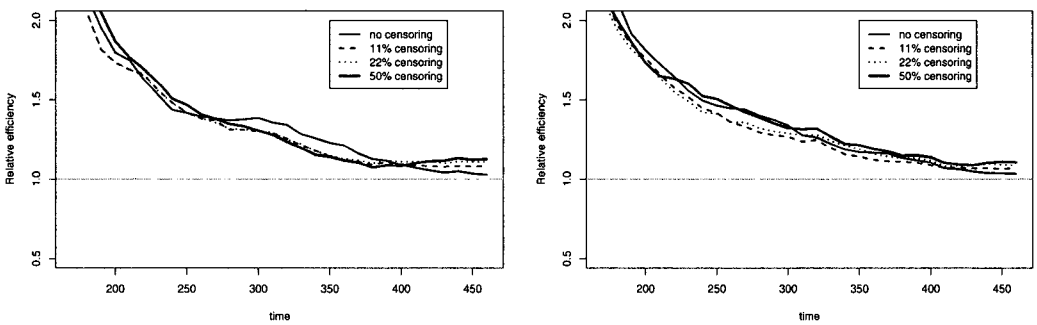
표 4.1: 표본수와 중도절단 비율에 따른 관심 있는 원인에 의한 사건의 평균 발생수

표본수	중도 절단의 비율 (a의 값)			
	0% (없음)	11% (300)	22% (220)	50% (0)
50	13.11	11.46	9.78	6.31
100	26.22	22.93	19.56	12.63
200	52.45	45.87	39.12	25.26



(a) 표본수 100 (b) 표본수 200

그림 4.4: 중도절단 비율에 따른 와이블 $\hat{I}_{sp.1}(t;\hat{\theta})$의 상대효율



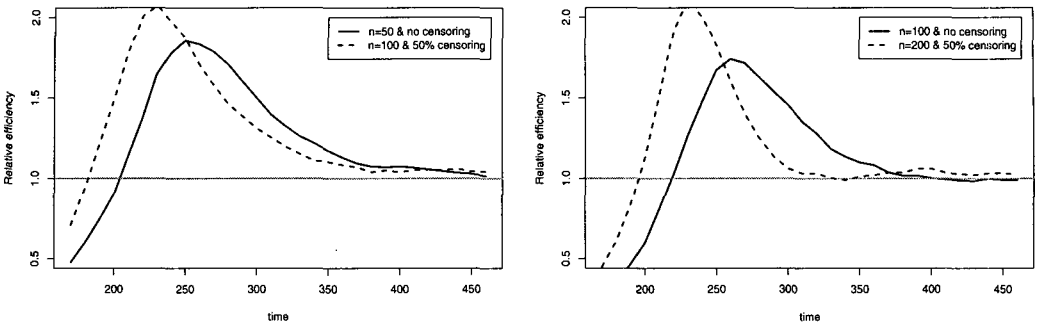
(a) 표본수 100 (b) 표본수 200

그림 4.5: 중도절단 비율에 따른 대수 정규 $\hat{I}_{sp.1}(t;\hat{\theta})$의 상대효율

그림 4.4(a)와 (b)는 표본수 $n = 100$과 200인 경우 중도절단비율에 따른 $\hat{I}_1(t)$에 대한 와이블분포 가정 $\hat{I}_{sp.1}(t;\hat{\theta})$의 상대효율을 나타낸다. 분포를 잘못 가정한 와이블분포 준모수적 CIF 추정량의 효율성이 시간의 전반부에서 중도절단 비율이 높을수록 효율성이 많이 높아진 것을 알 수 있다. 이는 중도절단 비율이 높을수록 준모수적 CIF 추정량이 시간의 전반부에서 모형 오설정에 대한 영향이 작아진다고 생각할 수 있다. 그러나 분포를 바르게 가정한 대수 정규분포 준모수적 CIF 추정량은 그림 4.5에서와 같이 중도절단 비율에 상관없이 효율성이 높다.

다음으로 관심 있는 원인에 의한 사건의 평균 발생수가 비슷한 경우 중도절단 비율이 높을수록 시간의 전반부에서 준모수적 CIF 추정량의 모형 오설정에 대한 영향이 작아지는지 살펴보겠다.

표 4.1을 보면 중도절단이 없고 표본수 50일 경우와 중도절단 비율이 50%이고 표본수



(a) n=50 & no censoring vs. n=100 & 50 % censoring(b) n=100 & no censoring vs. n=200 & 50 % censoring

그림 4.6: 평균 발생수가 비슷한 경우 와이블 $\hat{I}_{sp.1}(t;\hat{\theta})$의 상대효율

가 100인 경우가 관심 있는 원인에 대한 평균 사건 발생수가 비슷하다. 그리고 중도절단이 없고 표본수 100인 경우와 중도절단 비율이 50%이고 표본수가 200인 경우가 관심 있는 원인의 평균 사건 발생수가 비슷하다. 그림 4.6은 이들의 상대효율을 비교한 것이다.

그림 4.6에서 보듯이 관심 있는 원인에 의한 평균 사건의 발생수가 비슷한 경우 시간의 초반부에서는 중도절단 비율이 높은 경우가 와이블분포 준모수적 CIF 추정량의 효율성이 더 높았다. 나머지 시간대에서도 대부분 효율성이 높으며 같은 표본수인 경우 중도절단 비율이 높아질수록 효율성이 낮은 구간이 줄어들었다. 표 4.2 와 표 4.3은 모의실험 결과중 일부분을 나타낸다.

5. 실증분석

본 연구에 사용된 실증분석 자료는 호지킨병(Hodgkin's disease)이 있는 Princess Margaret 병원의 환자에 대한 자료이다(출처 : www.uhnresearch.ca/hypoxia/People_Pintilie.htm). 총환자수는 865명이며, 모든 환자는 초기 질병(Ⅰ기 또는 Ⅱ기)을 가지고 있고 방사선요법(radiation; RT)이나 화학요법(chemotherapy; CMT)을 받았다. 결과로 기록된 것에는 호지킨병 이후에 진단된 악성종양 또는 사망이다. 생존시간의 측정단위는 연수(years)이고, 진단을 받은 날부터의 시간을 나타낸다.

본 실증분석에서 crude CIF와 준모수적 CIF 추정량을 사용하여 악성종양에 걸릴 확률을 추정하려한다. 악성종양에 걸리지 않고 사망한 경우를 경쟁위험으로 처리하였다. 그림 5.1(a)는 호지킨병 자료에서 악성종양에 대한 원인특정적 누적 위험함수를 나타낸다. 그림 5.1(a)를 살펴보면 시간에 따라 원인특정적 누적위험함수가 체증적으로 증가하므로 지수분포 준모수적 CIF 추정량은 적당하지 않다. 그림 5.1(b)는 $\log t$에 따른 $\log(-\log(\hat{S}(t)))$를 나타낸다. 그림 5.1(b)를 살펴보면 전반부를 제외하고는 선형이 되므로 준모수적 CIF 추정량은 와이블분포를 사용하였다.

그림 5.2에서 보듯이 와이블분포 준모수적 CIF 추정량의 분포는 거의 모든 시간대에서

표 4.2: 표본수와 중도절단의 비율에 따른 와이블 $\hat{I}_{sp.1}(t;\hat{\theta})$의 상대효율

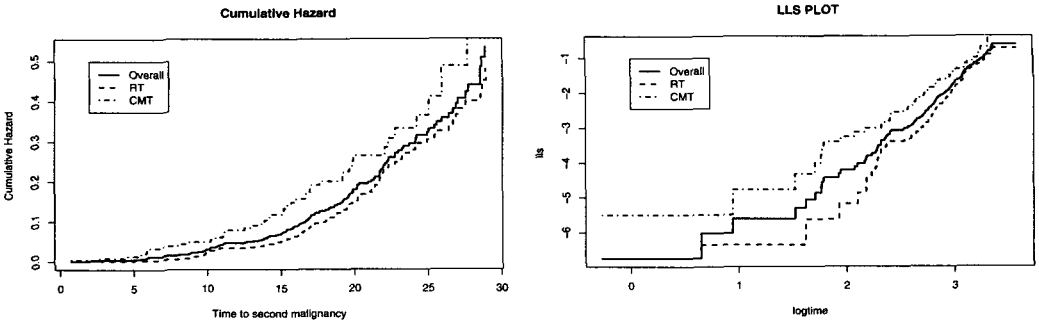
시간	0% 중도절단			11% 중도절단			22% 중도절단		50% 중도절단	
	n=50	n=100	n=200	n=50	n=100	n=200	n=100	n=200	n=100	n=200
170	0.48	0.28	0.15	0.66	0.44	0.25	0.49	0.28	0.71	0.45
190	0.74	0.46	0.27	1.06	0.70	0.46	0.78	0.53	1.22	0.83
210	1.14	0.81	0.52	1.57	1.21	0.93	1.33	1.05	1.78	1.51
230	1.65	1.27	1.01	1.88	1.72	1.63	1.80	1.71	2.07	2.11
250	1.86	1.67	1.68	1.86	1.79	1.84	1.76	1.76	1.87	1.83
270	1.79	1.72	1.81	1.67	1.59	1.45	1.52	1.37	1.58	1.40
290	1.60	1.54	1.41	1.46	1.35	1.12	1.32	1.09	1.39	1.13
310	1.40	1.35	1.08	1.32	1.23	0.96	1.22	0.97	1.26	1.03
330	1.26	1.18	0.92	1.23	1.12	0.91	1.13	0.94	1.15	1.00
350	1.17	1.10	0.85	1.16	1.06	0.90	1.08	0.94	1.10	1.01
370	1.10	1.04	0.85	1.10	1.06	0.94	1.07	0.98	1.07	1.04
390	1.07	1.02	0.86	1.06	1.07	0.99	1.07	1.03	1.05	1.06
410	1.07	0.99	0.87	1.05	1.05	1.00	1.06	1.01	1.05	1.04
430	1.05	0.98	0.90	1.03	1.03	1.02	1.03	1.03	1.05	1.02
450	1.03	0.99	0.93	1.00	1.02	1.01	1.02	1.02	1.04	1.03

표 4.3: 표본수와 중도절단의 비율에 따른 대수 정규 $\hat{I}_{sp.1}(t;\hat{\theta})$의 상대효율

시간	0% 중도절단			11% 중도절단			22% 중도절단		50% 중도절단	
	n=50	n=100	n=200	n=50	n=100	n=200	n=100	n=200	n=100	n=200
170	2.36	2.30	2.20	2.14	2.23	2.18	2.23	2.13	2.28	2.17
190	2.08	1.96	1.92	2.01	1.82	1.85	1.83	1.82	2.05	1.87
210	1.84	1.75	1.73	1.83	1.69	1.67	1.69	1.64	1.77	1.65
230	1.73	1.53	1.56	1.62	1.56	1.52	1.55	1.49	1.61	1.60
250	1.56	1.42	1.46	1.48	1.42	1.42	1.42	1.41	1.47	1.51
270	1.45	1.38	1.44	1.40	1.37	1.33	1.36	1.36	1.38	1.42
290	1.38	1.38	1.37	1.32	1.31	1.27	1.31	1.30	1.33	1.35
310	1.28	1.36	1.27	1.26	1.29	1.24	1.29	1.28	1.28	1.31
330	1.21	1.28	1.22	1.21	1.21	1.20	1.22	1.24	1.20	1.27
350	1.15	1.23	1.17	1.14	1.14	1.14	1.15	1.19	1.14	1.22
370	1.09	1.16	1.16	1.10	1.11	1.11	1.12	1.14	1.10	1.17
390	1.07	1.12	1.12	1.06	1.11	1.10	1.12	1.13	1.09	1.15
410	1.07	1.07	1.07	1.07	1.09	1.07	1.11	1.08	1.10	1.10
430	1.06	1.04	1.05	1.07	1.08	1.07	1.10	1.08	1.12	1.09
450	1.04	1.04	1.04	1.06	1.08	1.06	1.11	1.09	1.12	1.11

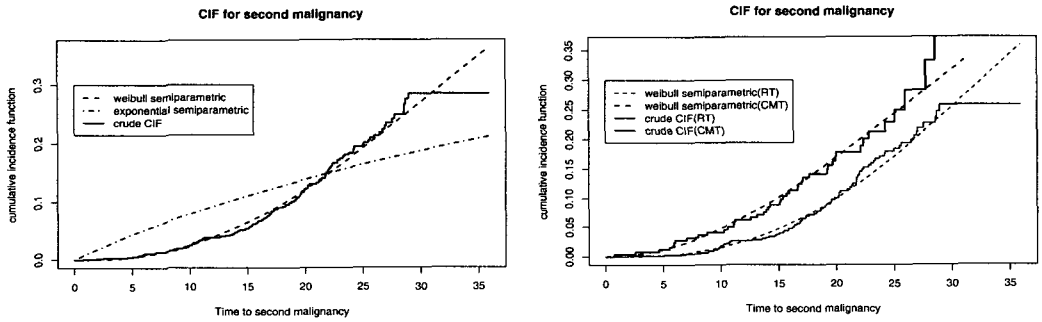
crude CIF 추정량의 분포와 유사하다. 따라서 와이블분포 준모수적 CIF 추정량은 주어진 자료에 적합이 잘되는 것을 알 수 있다. 그러나 지수분포 준모수적 CIF 추정량의 분포는 두 추정량의 분포와 많이 상이하다. 악성종양 원인특정적 위험함수가 시간에 대해서 일정하지 않고 시간이 지남에 따라 증가한다는 것을 알 수 있다.

방사선요법을 받은 환자가 진단 후 15년 이내에 악성종양에 걸릴 확률은 crude CIF로 추정했을 때 0.0375, 와이블분포 준모수적 CIF 추정량으로 추정하였을 때 0.0468이다. 한편 화학요법을 받은 환자의 진단 후 15년 이내에 악성종양에 걸릴 확률은 crude CIF로 추정했을 때 0.0888, 와이블분포 준모수적 CIF 추정량으로 추정하였을 때 0.1013이다. 방사선요법



(a) 원인특정적 누적 위험함수 (b) $\log t$에 따른 $\log(-\log(\hat{S}(t)))$

그림 5.1: 호지킨병 자료의 악성종양에 대한 분포진단 그래프



(a) 모든 환자 (b) 치료방법별

그림 5.2: 호지킨병 자료의 악성종양에 대한 비모수 및 준모수적 CIF 그래프

이 화학요법보다 악성종양의 발생을 억제하는 효과가 높은 것으로 판단된다. 지수분포 준모수적 CIF 추정량은 각각 0.0995, 0.1359로써 위의 두 추정값과 차이가 많이 난다. 앞에서 살펴보았듯이 지수분포 준모수적 CIF 추정량을 사용하기에는 적당하지 않다. 또한 crude CIF 추정량인 경우 이산성의 정도가 영향을 미친다.

그림 5.3은 총 865명의 환자 중 무작위로 100명을 추출한 자료에서 치료방법별 악성종양에 대한 CIF를 추정한 것이다. 그림 5.2(b)와 비교하였을 때 crude CIF 추정량이 매우 이산적인 것을 볼 수 있다. 이 경우 방사선요법을 받은 환자가 진단 후 15년 이내에 악성종양에 걸릴 확률은 crude CIF로 추정했을 때 0.0412, 와이블분포 CIF 추정량으로 추정하였을 때 0.0415이다. 한편 화학요법을 받은 환자의 진단 후 15년 이내에 악성종양에 걸릴 확률은 각각 0.1435, 0.1109이다. crude CIF 추정량은 이산성이 심해 화학요법을 받은 환자가 15년 이내에 악성종양에 걸릴 확률을 과대추정하였다.

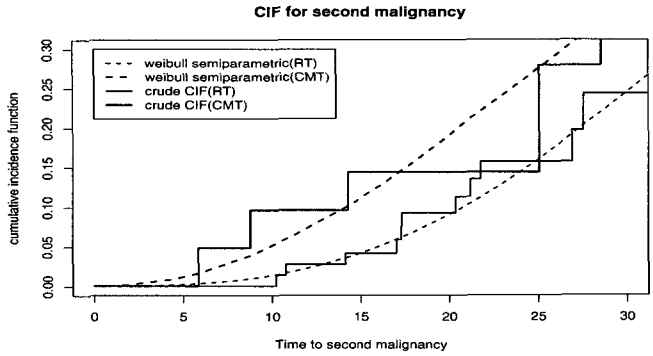


그림 5.3: 호지킨병 자료의 악성종양에 대한 비모수 및 준모수적 CIF 그래프 (표본수 100)

표 5.1: 15년 이내 악성종양에 걸릴 확률에 대한 각 방법별 CIF 추정값

집단	전체자료			100개의 무작위 표본	
	crude	Weibull (semi)	지수 (semi)	crude	Weibull (semi)
방사선	0.0375	0.0468	0.0995	0.0412	0.0415
화학요법	0.0888	0.1013	0.1359	0.1435	0.1109

6. 결 론

본 연구에서는 경쟁위험 하에서 비모수적 그리고 준모수적 누적발생함수 추정량에 대한 성능비교를 하였다. 경쟁위험 하에서는 관심 있는 원인에 대한 순수한 생존확률을 구할 수가 없고 원인특정적 위험함수를 통해 구할 수 있는 원인특정적 생존함수는 그 의미를 해석할 수가 없다. 따라서 경쟁위험 하에서는 생존함수의 형태가 아닌 CIF로써 관심 있는 원인에 대한 위험을 나타낸다.

CIF는 경쟁위험 하에서 추정이 가능하며 직접적으로 해석이 가능한 확률이기 때문에 경쟁위험 하에서 유용한 함수이다. 일반적으로 CIF를 추정할 때 비모수적 추정량인 crude CIF 추정량을 사용하며 이 추정량은 관심 있는 원인에 의한 사건이 적게 발생할 경우 매우 이산적인 분포를 한다. 그러나 관심 있는 원인에 대한 원인특정적 위험함수를 모수적으로 가정하여 CIF를 추정한 준모수적 CIF 추정량은 연속형 함수의 형태로 추정되며 crude CIF 추정량보다 정확하게 추정된다.

모의실험 결과, 관심 있는 원인에 대한 생존시간 분포가 대수 정규분포일 때 대수 정규분포 준모수적 CIF 추정량은 표본수, 자료의 중도절단 비율, 관심 있는 사건의 발생수에 상관없이 crude CIF 추정량보다 더 효율적으로 나타났다. 특히 시간의 초반부에서 상대효율이 더 좋게 나왔다. 와이블분포 준모수적 CIF 추정량은 crude CIF 추정량과 비교해서 표본

수가 적을수록, 관심 있는 원인에 대한 사건이 적게 발생할수록 효율성이 좋게 나왔다. 또한 시간의 초반부에 중도절단 비율이 높을수록 효율성이 좋게 나왔다.

결론적으로 표본수가 충분하고 관심 있는 원인에 대한 사건의 발생이 많은 자료에서는 CIF를 추정할 때 비모수적으로 crude CIF 추정량을 사용하는 것이 효율적이고 안정적이라 생각된다. 관심 있는 원인에 대해 시간에 따른 위험의 형태를 알고 있거나 자료를 통해 파악이 가능한 경우에는 원인특정적 위험함수를 모수적 가정을 하여 준모수적 CIF 추정량을 사용하는 것이 효율적이다. 원인특정적 위험함수에 대한 정보가 없더라도 표본수가 적거나 관심 있는 원인에 대한 사건의 발생이 적거나 또는 중도절단이 많은 경우, crude CIF 추정량은 이산성으로 인해 부정확하게 추정되는데 반해 준모수적으로 누적발생함수를 추정하면 crude CIF 추정량보다 효율성 높게 추정할 수 있다.

참고문헌

Bryant, J. and Diagnam, J. J. (2004). Semiparametric models for cumulative incidence functions, *Biometrics,* **60**, 182–190.

Gaynor, J. J., Feuer, E. J., Tan, C. C., Wu, D. H., Little, C. R., Straus, D. J., Clarkson, B. D. and Brennan, M. F. (1993). On the use of cause-specific failure and conditional failure probabilities: examples from clinical oncology data, *Journal of the American Statistical Association,* **88**, 400–409.

Gooley, T. A., Leisenring, W., Crowley, J. and Storer B. E. (1999). Estimation of failure probabilities in the presence of competing risks; new representations of old estimators, *Statistics in Medicine,* **18**, 695–706.

Gray, R. J. (1988). A class of K-sample tests for comparing the cumulative incidence of a competing risk, *The Annals of Statistics,* **16**, 1141–1154.

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

Korn, E. L. and Dorey, F. J. (1992). Applications of crude incidence curves, *Statistics in Medicine,* **11**, 813–829.

Marubini, E. and Valsecchi, M. G. (1995). *Analysing Survival Data from Clinical Trials and Observational Studies*, John Wiley & Sons, England.

Pepe, M. S. and Mori, M. (1993). Kaplan-Meier, marginal, or conditional probability curves in summarizing competing risks failure time data, *Statistics in Medicine,* **12**, 737–751.

Prentice, R. L., Kalbfleisch, J. D., Peterson, A. V., Jr, Flournoy, N., Farewell, V. T. and Breslow, N. E. (1978). The analysis of failure times in the presence of competing risks, *Biometrics,* **34**, 541–554.

Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks, *Proceedings of the National Academy of Science,* **72**, 20–22.

[2007년 3월 접수, 2007년 4월 채택]

Performance Comparison of Cumulative Incidence Estimators in the Presence of Competing Risks

Donguk Kim[1] Chikyung Ahn[2]

ABSTRACT

For the time-to-failure data with competing risks, cumulative incidence functions (CIFs) are commonly estimated using nonparametric methods. If the cases of events due to the cause of primary interest are infrequent relative to other cause of failure, nonparametric methods may result in rather imprecise estimates for CIF. In such cases, Bryant *et al.* (2004) suggested to model the cause-specific hazard of primary interest parametrically, while accounting for the other modes of failure using nonparametric estimator.

We represented the semiparametric cumulative incidence estimator and extended to the model of Weibull and log-normal distribution. We also conducted simulations to access the performance of the semiparametric cumulative incidence estimators and to investigate the impact of model misspecification in log-normal cause-specific hazard model.

Keywords: CIF, competing risks, semiparametric estimator, relative efficiency, Weibull distribution, log-normal distribution.

1) Professor, Department of Statistics, Sungkyunkwan University, Seoul 110-745, Korea
E-mail : dkim@skku.edu
2) Graduate Student, Department of Statistics, Sungkyunkwan University, Seoul 110-745, Korea
E-mail: acg11@skku.edu