응용통계연구 제 17권 2호, 2004년, pp. 253-267

모의실험을 통한 두 처리군간 치료율 비교방법 연구 *

박미라 [1] 이재원 [2] 진서훈 [3]

요 약

임상시험중에는 소아암연구에서와 같이 환자 중의 상당수에서 사망 또는 재발이 오랜 기간 일어나지 않고 완치된 것으로 보이는 경우가 있다. 이 경우 연구자는 생존함수의 전반적인 비교보다는 치료율의 비교에 더 관심이 있을 것이다. 본고에서는 치료율의 비교를 위한 여러 모수적, 비모수적 방법들을 소개하고, 생존분포, 치료율, 중도절단율 등을 다양하게 설정한 모의실험을 통하여 각 방법들의 검정력과 유의수준을 비교하였다.

주요용어: 치료율 비교, 모의실험, 혼합모형, 비례위험모형, 카플란-마이어 추정치

1. 서론

임상시험중에는 환자 중 상당수가 일정기간이 지나도록 사망 또는 재발 등의 반응을 보이지 않고 완치된 것으로 판단되는 경우가 있다. 이러한 현상은 치료효과가 높은 소아암(pediatric cancer)분야 등의 연구에서 종종 찾아 볼 수 있다. 이 때 환자들의 생존곡선은 일정 시점 이후에 평평하게 지속되는 형태를 보일 것이다. 이 경우에 두 종류의 처리를 비교하는 데 있어서 연구자는 단순히 생존함수의 전반적인 비교를 하는 것보다는 환자들의 치료율을 비교하는 데 더 관심이 있을 것이다. 예컨대 그림 1.1은 급성 골수성백혈병 환아들의 end-of-induction부터의 Disease-free 생존곡선이다. 환자들은 골수이식수술을 받았는지 화학치료만을 받았는지에 따라 두 그룹으로 나뉘어졌으며 두 그룹 모두 일정시점 이후에는 생존곡선이 길게 수평으로 유지되고 있어 환자들 중 상당비율이 치료되었음을 시사하고 있다(cf. Lee and Sather, 1995). 이러한 치료율의 비교를 위해 몇 가지 통계적 방법들이 사용되고 있다. Farewell(1982)은 모수적 방법으로 치료율을 추정하기 위한 혼합모형(mixture model)을 제시하였으며, Yamaguchi(1992), Peng et al.(1998)은 모수적방법의 제한점을 줄이기 위하여 분포에 대한 가정을 보다 완화하는 방법을 제안하였다. 그 밖에 모수적 방법에 대한 논의로 Jones et al.,(1981), Goldman(1984), Arbutiski(1985), Ghitany et al.(1994)등이 있다. Kuk and Chen(1992)은 이를 준모수적(semi-parametric) 방법으로 일반

* 본 연구는 2002년도 범석학술연구비 지원으로 수행되었음

1) (301-832) 대전시 중구 용두동 143-5, 을지의과대학교 의예과, 조교수
E-mail: mira@eulji.ac.kr
2) (301-701) 서울시 성북구 안암동 5가 1, 고려대학교 통계학과, 교수
E-mail: jael@korea.ac.kr
3) (150-757) 서울 영등포구 여의도동 15-22, 국민은행 카드마케팅팀, 과장
E-mail: bobbiej@passmail.to

화한 방법을 제안하였으며, 최근 Peng and Dear (2000)는 추정방법을 달리한 비례위험혼합모형을 제안하였다. 준모수적 방법에 관한 연구는 아직 많이 이루어지지 않고 있다. 치료율의 비교를 위한 비모수적 접근방법으로 Gray and Tsiatis(1989)가 두 처리군간의 비례생존분포함수를 정의하고 비모수적 방법에 근거하여 검정력을 최대화하는 선형최적검정(linear optimal test)을 제안하였다. Laska and Meisner(1992)는 치료율의 일반화 최대우도 누적한계추정치(generalized maximum likelihood product limit point estimator)를 구하고, 이에 대한 근사적인 우도비 검정을 개발하였으며, 최근에도 이러한 연구가 이어지고 있다(cf. Peng and Dear,2000; Sy and Taylor,2000). 한편, 생존함수의 카플란-마이어 추정치(Kaplan-Meier estimate)를 이용한 비모수적 검정을 할 수 있으며 이를 proportion 검정, 또는 PL(product limit)검정이라고 한다.

본고에서는 치료율 비교를 위한 여러 방법 중 자주 사용되거나 인용되는 대표적인 방법들을 선택하여 소개하고 생존분포, 치료율, 중도절단율 등을 다양하게 설정한 모의실험(simulation)을 통하여 각 방법들을 비교하였다. 또한 생존분포의 비교에 흔히 쓰이는 로그-순위 검정(log-rank test)과 Gehan의 검정을 모의실험에 포함시켜 이들 방법과의 비교를 행하였다.

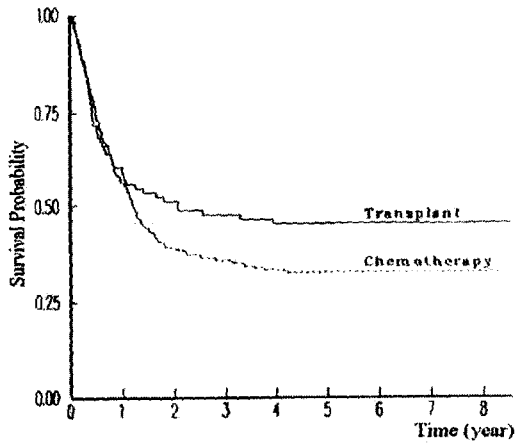

그림 1.1: 급성골수성백혈병 환아들의 Disease-free survival

2. 치료율 비교를 위한 통계적 방법들

2.1. Farewell(1982)의 방법

모수적 모형으로 가장 많이 쓰이는 것은 Farewell의 방법일 것이다. Farewell(1982)은 치료율에는 로지스틱(logistic)모형을, 분포함수에는 와이블(Weibull)모형을 적용하는 혼합모형(mixture model)을 제안하였다. 환자가 오랜 기간동안 사망하지 않고 생존할 때(즉, 완치

되는 경우) $Y=0$, 사망할 때 $Y=1$로 정의한다. 환자의 생존에 영향을 주는 공변량 x가 있다고 하면, Y의 분포는 다음과 같은 로지스틱모형으로 나타낼 수 있다 (Cox,1970).

$$P(Y=1|x)=\exp(\beta_0+\beta_1 x)/\{1+\exp(\beta_0+\beta_1 x)\} \tag{2.1}$$

$Y=1$인 환자들의 생존시간은 다음과 같은 와이블분포로 모형화할 수 있다.

$$f(t|Y=1,x)=\delta\lambda(\lambda t)^{\delta-1}\exp\{-(\lambda t)^{\delta}\} \tag{2.2}$$

여기서 δ와 $\lambda=\exp(-\gamma_0-\gamma_1 x)$는 분포의 형태를 결정하는 모수이며, γ_0와 γ_1은 미지의 회귀계수이다. 이러한 두 모형을 혼합한 모형에서의 모수추정은 최대우도함수를 통하여 이루어진다. 시간 t에서 환자가 사망할 확률은

$$P(Y=1|x)f(t|Y=1,x)$$

이 된다. 사망이 관측되지 않고 시간 t까지 추적된 환자는 완치되는 환자이거나 또는 t시간 이후의 어느 시점에서 사망하는 환자이므로 이 때의 확률은

$$\{1-P(Y=1|x)\}+P(Y=1|x)\int_t^{\infty} f(t^*|Y=1,x)dt^*$$

와 같이 표현할 수 있다.

두 처리의 효과를 비교하기 위해 처리군 1일 때 $x=0$, 처리군 2일 때 $x=1$과 같이 더미변수(dummy variable)로 놓으면 두 처리군의 치료율을 비교할 수 있게 된다. j번째 처리군$(j=1,2)$의 치료율을 $\theta_j=P(Y=1|x=j-1)$ 라고 하고, 완치되지 않은 환자의 생존함수를 $H_j(t)$라고 하자. j번째 처리군에서의 환자수를 n_j, j번째 처리군의 i번째 환자에서 사망 또는 중도절단이 관측된 시간을 t_{ji} 라하고 환자가 사망했을 때 $c_{ji}=1$, 그렇지 않을 때 $c_{ji}=0$으로 놓으면 우도함수는 다음과 같이 된다.

$$L(\beta_0,\beta_1;t)=\prod_{j=1}^{2}\prod_{i=1}^{n_j}[(1-\theta_j)f_j(t_{ji})]^{c_{ji}}[\theta_j+(1-\theta_j)H_j(t_{ji})]^{1-c_{ji}}$$

여기서 $f_j(t)=\partial(1-H_j(t))/\partial t$이다. 두 치료군간의 치료율이 동일하다는 가설은

$$H_0:\theta_1=\theta_2 \quad \text{또는} \quad H_0:\beta_1=0$$

으로 표현할 수 있다. 이 모형에서는 두 처리군의 치료율이 같더라도 생존분포는 같지 않을 수 있다. 이러한 우도함수에 근거하여 $H_0:\beta_1=0$에 대한 우도비(likelihood ratio) 검정, Wald 검정 및 Rao의 Score 검정 등을 할 수 있게 된다.

2.2. Kuk and Chen(1992)의 방법

Kuk and Chen(1992)은 Farewell(1982)의 방법을 준모수적(semi parametric)방법으로 일반화한 방법을 제안하였다. $Y=1$인 환자들의 위험함수 $h(t|Y=1,x)$를 다음과 같은 식으로 표현하자.

$$h(t|Y=1,x)=\exp(\eta x)h_0(t|Y=1) \tag{2.3}$$

여기서 $h_0(t|Y=1)$은 조건부 기저위험함수(conditional baseline hazard function)이다. Farewell(1982)의 방법은 여기서 $\eta = -\delta\gamma_1$이고, 조건부 기저위험함수가 모수 δ와 $\lambda = e^{-\gamma_0}$를 갖는 와이블 함수인 경우이다. Kuk and Chen(1992)은 조건부 위험함수를 와이블족으로 제한하지 않고 임의의 위험함수로 확장하고 식 (2.1)과 (2.3)을 결합한 준모수적 혼합모형을 사용하였다.

계수 β_0, β_1과 η를 추정하기 위해서는 위험함수의 비례성을 이용하여 다음과 같은 주변우도함수(marginal likelihood function)를 사용한다.

$$L(\beta_0, \beta_1, \eta) = \sum_{y_c \in \Omega} L(\beta_0, \beta_1, \eta; y_C)$$

여기서 $y_c \in \Omega$는 $Y_c = (Y_i, i \in C)$의 실현치이고,

$$L(\beta_0, \beta_1, \eta; y_C) = \prod_{i \in D} P_i \prod_{i \in C} [P_i^{Y_i}(1-P_i)^{Y_i}] \prod_{i=1}^{k} \left[\psi_{(i)} / \{ \sum_{j=i}^{k} (\psi_{(j)} + \sum_{l \in C(j)} Y_l \psi_l) \} \right]$$

이다. 여기서 $\psi_{(i)} = \exp(\eta x_{(i)})$, $\psi_l = \exp(\eta x_l)$이며, $x_{(j)}$는 사망이 관측된 시간을 오름차순으로 배열했을 때 (즉, $t_{(1)} < \cdots < t_{(k)}$), j번째 사망시간 $t_{(j)}$에 해당되는 x의 값이다. 또한 $D = \{i : c_i = 1\}$는 사망이 관측된 경우, $C = \{i : c_i = 0\}$는 사망이 관측되지 않은 경우의 부호들의 집합을 가리키고, $C(j)$는 구간 $[t_{(j)}, t_{(j+1)})$에서 중도절단된 부호들의 집합이다. 주변우도함수를 최대화시킴으로써 회귀계수의 추정치를 얻을 수 있다. 두 처리군의 비교를 하기 위해서는 x를 처리군을 나타내는 더미변수로 놓고 주변우도함수에 근거하여 $H_0 : \beta_1 = 0$에 대한 검정을 하게 된다.

최근 Peng and Dear (2000)는 이른바 비례위험혼합모형을 제안하였다. 여기서는 치유되지 않은 환자들에 대한 생존시간에 대한 공변량 효과로 비례위험모형을 사용하였으며 추정방법으로 Cox의 비례위험모형에 대한 주변우도함수와 EM 알고리즘을 이용하였다. 이 모형은 Kuk and Chen의 방법과 유사하나, 추정방법에서 기저생존함수를 EM 알고리즘에서 제외시키지 않고 추정하여 사용한다는 점에서 차이가 있다.

2.3. Proportion 검정

생존함수의 카플란-마이어 추정치(Kaplan-Meier estimate)를 이용하여 치료율의 비교를 위한 간단한 비모수적 검정을 수행할 수 있다. 한 그룹의 환자들의 생존시간을 순서대로 나열한 것을 $t_{(1)} < t_{(2)} < \cdots < t_{(m)}$이라 하고, d_k를 t_k에서 사망한 환자수, r_k를 t_k직전의 위험환자수라고 하자. 시간 t에서의 생존율에 대한 카플란-마이어 추정치는

$$\hat{S}(t) = \prod_{k:t_{(k)}<t} (1 - d_k/r_k)$$

이며, 이의 분산의 추정치는 다음과 같다 (Kalbfleish and Prentice,1980).

$$\hat{\sigma}^2(t) = \hat{S}(t)^2 \sum_{k:t_{(k)}<t} \frac{d_k}{r_k(r_k - d_k)}$$

두 그룹의 생존분포를 비교하기 위한 통계량으로

$$\chi_P^2 = \frac{(\hat{S}_1(u) - \hat{S}_2(v))^2}{\hat{\sigma}_1(u)^2 + \hat{\sigma}_2(v)^2}$$

을 생각할 수 있다. 여기서 $\hat{S}_j(t)$와 $\hat{\sigma}_j(t)^2$는 각각 처리군 j의 시점 t에서의 카플란-마이어 추정치와 그의 분산추정치이다. 이 때 χ_P^2는 귀무가설

$$H_0 : S_1(u) = S_2(v)$$

하에서 자유도 1인 χ^2-분포를 따르게 된다. 연구자는 각 그룹에서 카플란-마이어 추정치의 곡선이 완만해지는 적당한 시점 u, v를 정하여 생존시간을 비교할 수 있다.

Sposto et al. (1992)는 이 방법을 PL(product limit) 검정이라고 칭하였으며, 모의실험을 통하여 이를 Farewell(1982)의 방법과 비교한 결과 여러 상황에서 이 검정법이 모수적 방법에 못지 않은 검정력을 가지고 있음을 보인 바 있다.

2.4. Gray and Tsiatis(1989)의 방법

Gray and Tsiatis(1989)는 생존분포의 후반부의 차이에 대한 검정력에 중점을 둔 대립가설을 고려하여, 이러한 대립가설에 관한 검정력을 최대화하는 최적선형순위검정법(optimal linear rank test)을 제안하였다. j번째 처리를 받고 완치된 환자의 모비율을 θ_j라고 하고, 완치되지 않은 환자의 조건부생존함수를 $H_j(t)$라고 했을 때 각 처리군에서 모집단의 생존함수를 다음과 같이 표현할 수 있다.

$$S_j(t) = \theta_j + (1 - \theta_j)H_j(t), \quad j = 1, 2.$$

주 관심이 치료율의 비교에 있으므로 치료되지 않은 환자들의 조건부 생존분포는 동일하다고 가정할 수 있을 것이다. 즉, $H_1(t) = H_2(t) = H(t)$ 라고 가정한다. 이제 두 처리군의 치료율간에는 다음과 같은 관계식이 성립하게 된다.

$$1 - S_2(t) = \frac{1 - \theta_2}{1 - \theta_1}(1 - S_1(t)). \tag{2.4}$$

이 때 귀무가설은 생존분포의 동일성

$$H_0 : S_1(t) = S_2(t)$$

으로 치료율자체의 동일성에 관한 것은 아니다. 식 (2.4)로부터 귀무가설하에서는 두 처리군의 치료율이 같다는 것을 의미하며, 두 처리군에서의 생존분포는 대립가설하에서도 비례성을 가진다는 것을 알 수 있다.

두 생존분포의 비교를 위해서 선형순위통계량(linear rank statistic)이 자주 사용된다. Tarone and Ware(1977)은 다음과 같은 형태의 점수에 근거한 2-표본 비모수 검정법을 개발하였다.

$$\int_0^T w(t)\frac{r_1(t)r_2(t)}{r_1(t) + r_2(t)}\left[\frac{dN_1(t)}{r_1(t)} - \frac{dN_2(t)}{r_2(t)}\right]. \tag{2.5}$$

여기서 $N_j(t)$는 처리군 j에서 t시점까지 사망이 관측된 환자수이고, $r_j(t)$는 처리군 j에서 t 직전에서의 위험그룹의 환자수이며, $w(t)$는 관심있는 대립가설에 대한 검정력에 따라 선택되는 값이고, T는 이 시점이후까지 생존한 환자는 치료된 것으로 간주하는 시점이다. Gray and Tsiatis(1989)는 대립가설하에서 검정력을 최대화하는 가중함수는 카플란-마이어 추정치(Kaplan-Meier estimate)의 역함수, $w(t) = KM^{-1}(t-)$임을 보였다. 이를 근거로 한 Gray and Tsiatis(1989)의 검정통계량은 다음과 같이 정리된다.

$$Z_{-1} = \frac{\sum_{t_i \le T} [KM(t_i-)]^{-1}[\Delta N_1(t_i) - p(t_i)]}{\{\sum_{t_i \le T} [KM(t_i-)]^{-2} p(t_i)[1-p(t_i)]\}^{1/2}}$$

여기서 $p(t) = r_1(t)/[r_1(t) + r_2(t)]$로서 t에서 1명이 사망했을 때 그 시점에서 처리군 1의 기대 사망환자수(expected number of death)이다. 이 통계량 Z_{-1}는 귀무가설하에서 근사적으로 표준정규분포를 따르므로 이를 이용하여 검정하게 된다.

통상적인 로그-순위검정법이나 proportion 검정도 선형순위통계량으로서, 식(2.5)에서 가중치가 $w(t) = 1$인 경우에는 로그-순위검정이 되며, $w(t) = [r_1(t) + r_2(t)]/[r_1(t)r_2(t)]$가 되면 proportion 검정이 된다. Gray and Tsiatis(1989)의 방법은 로그-순위검정법보다 후반기의 차이에 더 많은 비중을 두게 되므로, 비례성의 가정이 충족되지 않더라도 로그-순위검정법보다는 치료율의 검정에 보다 적합할 것이다.

2.5. Laska and Meisner(1992)의 방법

Laska and Meisner(1992)의 방법도 비모수적 방법으로서 치료율에 대한 일반화 최대우도 누적한계추정치(generalized maximum likelihood product limit point estimator)를 구하고, 이에 대한 근사적인 우도비 검정을 개발하였다. 여기서도 역시 다음과 같은 모형을 고려한다.

$$S_j(t) = \theta_j + (1-\theta_j)H_j(t), \quad j = 1, 2.$$

이때에는 Gray and Tsiatis(1989)와 달리 치료되지 않은 환자들의 조건부 생존분포가 같다($H_1(t) = H_2(t)$)는 가정은 필요하지 않으며, 귀무가설은 특정한 점 t^*에서의 생존확률의 동일성

$$H_0 : S_1(t^*) = S_2(t^*)$$

이다. 여기서 t^*는 중도절단시간들이 모두 이보다 작게 되도록 충분히 큰 어떤 시점으로 정해진다.

우도비 통계량을 구하기 위해서 치료율이 동일($\theta_1 = \theta_2$)하다는 조건하에서의 일반화 최대우도추정치를 구한다. j번째 처리군에서 생존확률의 증분을 $p_{ij} = S_j(t_{(i-1)}) - S_j(t_{(i)})$라고 하자. 이의 합은 j번째 처리군에서 치료되지 않은 확률 $1-\theta_j$가 된다. λ_{ij}를 조건부 확률

로서 $\lambda_{ij} = p_{ij}/(1 - \sum_{j=1}^{i-1} p_{ij})$라고 하자. 우도함수는

$$L = \prod_{j=1}^{2} L_j = \prod_{j=1}^{2} \prod_{i=1}^{n_j} (\lambda_{ij})^{c_{ij}} (1 - \lambda_{ij})^{n_j + m_j - i}$$

가 된다. 무제한 모형(unrestricted model)과 귀무가설하에서의 λ_{ij}의 최대우도추정치는 각각

$$\begin{aligned} \hat{\lambda}_{ij} &= c_{ij}/(n_j + m_j - i + c_{ij}) \\ \hat{\lambda}^0_{ij} &= c_{ij}/(n_j + m_j - i + 1 - t^*) \end{aligned}$$

로 구해진다. 여기서 m_j와 n_j는 각각 그룹 j에서 치료된 사람수와 치료되지 않은 사람수를 가리킨다. 무제한모형의 우도를 L_Ω, 귀무가설이 맞을 때의 우도를 L_ω라고 했을 때 $-2\log\frac{L_\omega}{L_\Omega}$가 자유도 1인 χ^2-분포를 따른다는 사실을 이용하여 검정한다.

3. 모의실험 및 결과

3.1. 모의실험의 구조

실제 임상자료를 분석하는 통계학자들에게는 주어진 상황에서 가장 적당한 검정법을 선택하는 것이 가장 중요한 문제일 것이다. 여기서는 치료율 비교를 위한 방법들 중 Farewell 방법, Gray-Tsiatis 방법, proportion 검정과 Laska-Meisner방법 등에 대한 모의실험을 수행하여 다양한 상황하에서 이들의 검정력과 유의수준을 비교하였다. 또한 생존분포의 비교를 위해 흔히 사용되는 로그-순위 검정과 Gehan의 검정이 함께 사용되었다. Kuk and Chen 방법의 결과는 관련된 우도함수의 Monte Carlo 근사에 의존하게 되는 불편함이 있어 모의실험에서는 제외하였다.

모의실험의 구조는 다음과 같다. 먼저 중도절단율(censoring rate;C_k)로서 0%, 25%, 50%의 세 수준이 고려되었으며, 두 그룹의 치료율(cure rate;θ_1, θ_2)은 귀무가설이 참인 경우의 세 수준 (0.1, 0.1) (0.3, 0.3) (0.5, 0.5)과 거짓인 경우의 세 수준 (0.1, 0.3) (0.2, 0.4) (0.4, 0.6)을 고려하였다. 또한 조건부 생존분포로서 표 3.1과 같은 6가지의 분포경우를 고려하였다. (i)과 (iv)의 경우에는 두 그룹의 조건부 생존함수가 같으므로 치료율이 같다는 귀무가설은 결국 두 처리군의 생존함수가 같다는 것을 의미한다. 또한 처리1은 처리2보다 치료율이 낮은 경우이므로 (ii)와 (v)의 경우에는 교차하는 생존곡선을 갖게 된다. (iii)과 (vi)은 조건부 생존함수가 같지 않으나 교차하지 않는 경우이다. 분포(i)-(iii)은 와이블분포(여기서는 지수분포)를 가정하는 경우이다. 비례위험모형을 만족하지 않는 경우를 만들기 위해 지수분포의 혼합형태를 사용하였으며 이는 분포 (iv)-(vi)과 같이 주어졌다. 각 처리군당 50명씩 총 100명의 환자가 할당된 것을 가정하였으며, 각 경우에서 1000회의 모의실험을 하였다. 중도절단은 $U(0, T_c)$에서 발생하도록 하였는데, 여기서 T_c는 기준그룹에서 완치되지 않은 환자들이 중도절단될 확률이 C_k가 되도록 하는 값이다.

표 3.1: 모의실험에서 가정된 조건부 생존분포

	처리1	처리2	기준그룹
(i)	exp(1)	exp(1)	처리1
(ii)	exp(1)	exp(2)	처리1
(iii)	exp(2)	exp(1)	처리2
(iv)	exp(0.5)+exp(1)	exp(0.5)+exp(1)	처리1
(v)	exp(0.5)+exp(1)	exp(1)+exp(2)	처리1
(vi)	exp(1)+exp(2)	exp(0.5)+exp(1)	처리2

3.2. 모의실험의 결과 및 해석

표 3.2부터 표 3.7까지는 유의수준 5%에서 양측검정을 1000회 시험하였을 때의 기각 횟수로서 조건부 생존분포에 따른 검정력과 유의수준을 보여준다. 여기서 L-M은 Laska-Meisner의 방법을, G-T는 Gray-Tsiatis의 방법을 가리킨다. 표3.2는 두 처리군의 조건부 생존분포가 모두 exp(1)인 경우의 모의실험 결과이다. 이 경우는 두 그룹에서 조건부생존분포가 모두 지수분포를 따르며 동일한 조건부 생존분포를 갖는 경우이다. Farewell방법의 조건이 만족되는 경우이나 치료율이 낮을 때에는 유의수준이 정상수준보다 작게 되는 경향이 있다. 또한 기대와는 달리 이 모의실험에서는 비모수적인 방법인 Laska-Meisner방법과 Proportions 검정의 검정력이 Farewell 방법보다 오히려 약간 더 높게 나타났다. 이는 모수적 방법의 가정이 맞는 경우에도 비모수적 방법인 Laska-Meisner방법이나 Proportions 검정이 경쟁력 있음을 시사하는 것으로 생각된다. Gray-Tsiatis방법의 경우 중도절단이 없고 치료율이 낮은 경우에는 정상수준보다 유의수준이 조금 높게 나타난다. 이는 Gray-Tsiatis방법이 카플란-마이어의 역함수를 가중치로 사용하는데 이러한 경우처럼 카플란-마이어 추정치가 작게 되는 상황에서는 검정이 불안정한 경향이 있기 때문이다. 그러나 정상수준에서 크게 벗어나지는 않고 있으며 중도절단 비율이 증가하면 유의수준이 정상수준에 가까이 간다는 것을 알 수 있다. 로그-순위검정이나 Gehan의 검정의 경우 대체로 정상적인 유의수준을 만족하나 검정력은 다른 방법에 비해 약간 작다. 모든 방법에서 중도절단율이 증가할수록 검정력은 작아지며, Gray-Tsiatis 방법의 경우는 그 감소율이 다른 검정에 비해 크다. 또한 치료율이 높을 때에는 검정력간의 차이가 작아짐을 알 수 있다.

표 3.3은 처리1의 조건부 생존분포함수가 exp(1)이고 처리2의 분포 함수는 exp(2)인 경우이다. 두 조건부 생존분포가 다르므로 이 경우에는 치료율이 같더라도 생존분포는 달라지게 되는 경우이며, 또한 교차하는 생존곡선을 갖는다. 대부분의 경우 proportion 검정의 검정력이 가장 높고, Laska-Meisner 방법과 Farewell방법의 검정력도 높다. 여기서도 Farewell방법은 치료율이 낮을 때 유의수준이 정상수준보다 작게 나타난다. Gray-Tsiatis방법, 로그-순위검정과 Gehan의 검정은 생존분포의 동일성을 비교하는 것으로 이 경우에는 적합한 검정법이 아니다. 이들의 유의수준은 5%를 훨씬 상회하는 값들임을 알 수 있다. 그

러나 Gray-Tsiatis방법의 경우 치료율이 높고 중도절단율이 낮은 경우 유의수준이 정상적인 수준에 가까이 가며, 검정력도 크게 떨어지지 않는다. 반면 로그-순위 검정이나 Gehan 검정의 검정력은 매우 낮다. 표 3.4도 역시 조건부 생존분포가 다른 경우로서 Gray-Tsiatis방법, 로그-순위검정, Gehan의 검정이 적합한 검정법이 아니다. 그러나 이 경우는 처리1의 조건부생존분포가 exp(2), 처리2의 조건부생존분포가 exp(1)인 경우로서, 처리1의 치료율이 처리2보다 작으므로 생존분포함수가 교차하지 않아 두 처리군의 생존분포함수가 표3의 경우보다는 비례위험모형에 가깝게 된다. 그 결과 로그-순위검정과 Gehan의 검정의 유의수준은 여전히 매우 크지만 검정력은 상당히 높아졌음을 알 수 있다. Gray-Tsiatis방법은 치료율이 낮은 경우를 제외하고는 유의수준과 검정력 면에서 모두 좋은 결과를 보여준다.

표 3.5는 두 처리군에서의 조건부생존분포가 exp(0.5)와 exp(1)을 따르는 것들이 50%씩 섞여진 분포이다. 여기서는 와이블 가정이 깨어졌을 때 모수적 방법에 의한 검정이 적합한가를 알아볼 수 있다. Farewell 방법의 경우 와이블가정이 위반되었음에도 유의수준과 검정력에 있어서 모두 다른 검정법들과 큰 차이가 나지 않음을 보이고 있어, 와이블가정에 대해 로버스트한 검정법임을 알 수 있다. Gray-Tsiatis방법의 경우 중도절단이 없는 경우에는 가장 큰 검정력을 보이나 중도절단비율이 커 갈수록 점차 검정력이 떨어지는 양상을 보인다. 표 3.6과 표 3.7 의 경우도 마찬가지로 Farewell검정이 유의수준 및 검정력에 있어서 대체로 좋은 결과를 보이고 있다. Proportion 검정의 경우에는 중도절단이 없고 치료율도 낮은 경우에서 유의수준이 커져 편의가 발생하였다.

표 3.2: 분포 (i)의 경우 모의실험 결과 :1000회중 기각횟수
(처리군 1: exp(1), 처리군 2: exp(1)일 때)

Censoring	Test	Cure rates (θ_1, θ_2)					
		(.1, .1)	(.3, .3)	(.5, .5)	(.1, .3)	(.2, .4)	(.4, .6)
None	Farewell	30	53	71	708	586	512
	L - M	63	50	70	736	584	505
	Proportion	67	60	70	753	600	505
	G - T	81	61	71	760	619	512
	Log-rank	64	64	53	661	554	486
	Gehan	57	45	57	431	453	449
25%	Farewell	24	45	40	549	444	368
	L - M	48	51	42	587	468	378
	Proportion	59	57	44	604	485	391
	G - T	57	53	39	583	452	370
	Log-rank	47	62	41	484	424	378
	Gehan	49	46	47	320	337	346
50%	Farewell	29	49	40	394	326	286
	L - M	51	57	46	415	346	301
	Proportion	61	62	51	432	353	304
	G - T	49	54	42	397	333	277
	Log-rank	53	52	50	352	310	282
	Gehan	54	47	44	237	256	263

표 3.3: 분포 (ii)의 경우 모의실험 결과 :1000회중 기각횟수
(처리군 1: exp(1), 처리군 2: exp(2)일 때)

Censoring	Test	Cure rates (θ_1, θ_2)					
		(.1, .1)	(.3, .3)	(.5, .5)	(.1, .3)	(.2, .4)	(.4, .6)
None	Farewell	30	53	71	708	587	512
	L - M	63	52	71	722	574	378
	Proportion	69	61	71	741	591	498
	G - T	75	59	71	720	592	510
	Log-rank	413	141	61	202	242	332
	Gehan	644	300	112	56	75	190
25%	Farewell	24	45	40	549	444	368
	L - M	48	51	42	587	468	378
	Proportion	59	57	44	604	485	391
	G - T	104	62	41	451	379	334
	Log-rank	336	118	61	136	168	253
	Gehan	536	254	88	53	64	145
50%	Farewell	29	49	40	395	325	285
	L - M	51	57	46	415	346	301
	Proportion	61	62	51	432	353	304
	G - T	159	61	46	225	216	233
	Log-rank	300	99	61	92	126	180
	Gehan	430	189	82	58	59	107

표 3.4: 분포 (iii)의 경우 모의실험 결과 :1000회중 기각횟수
(처리군 1: exp(2), 처리군 2: exp(1)일 때)

Censoring	Test	Cure rates (θ_1, θ_2)					
		(.1, .1)	(.3, .3)	(.5, .5)	(.1, .3)	(.2, .4)	(.4, .6)
None	Farewell	30	53	71	681	581	497
	L - M	71	53	71	724	613	504
	Proportion	69	61	71	743	629	504
	G - T	75	59	71	750	633	497
	Log-rank	413	141	61	943	811	653
	Gehan	644	300	112	962	909	702
25%	Farewell	24	45	40	524	479	363
	L - M	55	51	42	557	491	373
	Proportion	59	57	44	570	506	388
	G - T	104	62	41	714	572	410
	Log-rank	336	118	61	857	684	527
	Gehan	536	254	88	897	819	574
50%	Farewell	29	49	40	360	323	259
	L - M	54	57	46	388	334	278
	Proportion	61	62	51	406	343	281
	G - T	159	61	46	610	472	317
	Log-rank	300	99	61	726	562	383
	Gehan	430	189	82	779	675	470

표 3.5: 분포 (iv)의 경우 모의실험 결과 :1000회중 기각횟수
(처리군 1: $0.5\exp(1) + 0.5\exp(2)$, 처리군 2: $0.5\exp(1) + 0.5\exp(2)$일 때)

Censoring	Test	Cure rates (θ_1, θ_2)					
		(.1, .1)	(.3, .3)	(.5, .5)	(.1, .3)	(.2, .4)	(.4, .6)
None	Farewell	26	49	70	703	586	503
	L - M	53	51	72	661	535	474
	Proportion	53	56	72	668	547	474
	G - T	80	57	73	760	614	510
	Log-rank	58	61	52	647	544	485
	Gehan	42	40	58	419	437	441
25%	Farewell	26	62	53	519	461	374
	L - M	54	63	54	562	470	384
	Proportion	61	67	54	579	487	396
	G - T	52	62	51	565	464	378
	Log-rank	46	53	48	480	435	363
	Gehan	44	45	43	303	333	345
50%	Farewell	32	41	35	367	334	284
	L - M	50	43	42	399	349	296
	Proportion	64	44	43	415	360	303
	G - T	52	41	36	367	329	283
	Log-rank	49	43	38	339	310	287
	Gehan	52	48	41	236	262	271

표 3.6: 분포 (v)의 경우 모의실험 결과 :1000회중 기각횟수
(처리군 1: $0.5\exp(0.5) + 0.5\exp(1)$, 처리군 2: $0.5\exp(1) + 0.5\exp(2)$일 때)

Censoring	Test	Cure rates (θ_1, θ_2)					
		(.1, .1)	(.3, .3)	(.5, .5)	(.1, .3)	(.2, .4)	(.4, .6)
None	Farewell	29	51	70	706	593	507
	L - M	101	66	77	546	437	422
	Proportion	103	74	77	565	457	422
	G - T	80	58	73	713	583	500
	Log-rank	375	128	61	225	250	330
	Gehan	612	282	105	46	86	205
25%	Farewell	37	56	46	535	453	369
	L - M	56	56	53	568	466	374
	Proportion	64	63	56	586	483	390
	G - T	101	63	49	433	376	329
	Log-rank	311	121	63	152	180	257
	Gehan	522	237	93	44	64	145
50%	Farewell	27	46	43	363	328	273
	L - M	50	53	50	400	348	284
	Proportion	57	58	53	414	356	290
	G - T	138	64	43	222	224	232
	Log-rank	276	101	48	92	123	185
	Gehan	405	209	74	41	59	114

표 3.7: 분포 (vi)의 경우 모의실험 결과 :1000회중 기각횟수
(처리군 1: $0.5\exp(1)+0.5\exp(2)$, 처리군 2: $0.5\exp(0.5)+0.5\exp(1)$일 때)

Censoring	Test	Cure rates (θ_1,θ_2)					
		(.1, .1)	(.3, .3)	(.5, .5)	(.1, .3)	(.2, .4)	(.4, .6)
None	Farewell	29	51	70	680	586	489
	L - M	72	51	72	810	670	560
	Proportion	103	74	77	824	686	560
	G - T	80	58	73	757	638	500
	Log-rank	375	128	61	937	804	655
	Gehan	612	282	105	955	907	692
25%	Farewell	37	56	46	514	459	360
	L - M	58	59	52	552	482	375
	Proportion	64	63	56	569	498	388
	G - T	101	63	49	697	575	406
	Log-rank	311	121	63	847	683	519
	Gehan	522	237	93	890	812	561
58%	Farewell	27	46	43	355	347	274
	L - M	52	53	50	396	359	293
	Proportion	57	58	53	409	366	303
	G - T	138	64	43	623	481	334
	Log-rank	276	101	48	724	567	396
	Gehan	405	209	74	784	692	462

4. 토론

생존함수를 비교하는데 있어서 흔히 로그-순위검정이나 Gehan의 검정등을 사용한다. 그러나 이러한 방법은 연구의 주목적이 처리의 완치효과를 비교하는 데에 있을 때에는 그다지 유용하지 않게 된다. 또한 치료율의 비교를 위해 고안된 방법이라 하더라도 기저분포나 치료율의 크기, 중도절단비율 등에 따라 검정력의 차이가 있을 수 있다. 본 고에서는 치료율의 비교를 위한 여러 가지 통계적 방법들을 소개하고 두 처리군의 생존분포, 치료율, 중도절단율 등을 다양하게 설정한 모의실험(simulation)을 통하여 여러 상황하에서 각 방법들의 검정력을 비교하였다.

모의실험결과를 유의수준 면에서 살펴보면, Farewell 방법은 치료율이 낮은 경우에 유의수준이 약간 작은 경향이 있으나, 와이블분포를 따르지 않을 때에도 대체로 정상수준을 보여 와이블분포의 가정에 대해 로버스트함을 보였다. Gray-Tsiatis방법은 두 처리군의 조건부 생존분포가 같을 때 중도절단이 없을 때 약간 큰 유의수준을 보였고 중도절단율이 증가할수록 정상수준에 가까워졌으며, 두 처리군의 조건부 생존분포가 같지 않을 때에는 중도절단율이 높으면서 치료율은 낮은 경우를 제외하고는 대체로 정상수준을 보였다. 로그-순위 검정과 Gehan의 검정은 두 조건부생존분포가 서로 다를 때에는 매우 큰 값을 보였다.

검정력 면에서 보면, 두 조건부생존분포가 동일할 때 Laska-Meisner방법, proportion검정, Gray-Tsiatis방법이 Farewell방법보다 검정력이 컸으며, 로그-순위검정, Gehan의 검정

은 이보다 작았다. 또한 Gray-Tsiatis는 중도절단율이 높을 때 검정력이 작아지는 경향을 보였다. 한편, 조건부생존분포가 상이할 때에도 Laska-Meisner방법과 proportion 검정은 Farewell 방법보다 검정력이 크게 나타났고, 중도절단이 없는 경우에는 Gray-Tsiatis방법도 비슷한 수준으로 나타났다. 반면, 로그-순위검정과 Gehan 검정의 검정력은 매우 낮았다. 조건부 생존분포가 와이블족이 아닐 때 Farewell 방법의 검정력은 다른 검정과 비슷한 수준으로 로버스트한 성질을 보였고, 이 경우 중도절단이 없을 때에는 Gray-Tsiatis방법이 가장 검정력이 큰 것으로 나타났다.

전반적으로 Laska-Meisner방법과 proportion 검정이 대부분의 상황에 우수한 것으로 나타났다. Farewell방법은 조건부 생존분포가 와이블족이 아닌 경우에도 대체로 높은 검정력을 보였으며, Gray-Tsiatis방법은 조건부 생존분포가 동일하지 않은 경우에도 생존곡선이 교차하지 않고 치료율이 낮지 않을 때에는 좋은 결과를 보여주었다. 예상대로 로그-순위검정이나 Gehan의 검정 등은 치료율의 비교에 적절치 않음을 확인하였다. 다음의 표 4.1은 각 상황별로 모의실험결과를 정리한 것이다.

표 4.1: 각 상황별 가정 및 적절한 검정방법

상황	조건부생존분포가 와이블족인가?	두 그룹의 조건부생존분포가 동일한가?	두 그룹의 생존곡선이 교차하는 형태인가?	가정에 부합되지 않는 검정법	사용가능한 방법 (검정력크기순)
(i)	o	o	×	없음	proportion, L-M, Farewell, G-T(중도절단율과 치료율이 동시에 낮지 않을때)
(ii)	o	×	o	G-T, Log-rank, Gehan	proportion, L-M, Farewell,
(iii)	o	×	×	G-T, Log-rank, Gehan	G-T(치료율이 낮지 않을때) proportion, L-M, Farewell,
(iv)	×	o	×	Farewell	Proportion L-M, G-T(중도절단율과 치료율이 동시에 낮지 않을때) Farewell
(v)	×	×	o	Farewell G-T, Log-rank, Gehan	Proportion(중도절단율과 치료율이 동시에 낮지 않을때) L-M(중도절단율과 치료율이 동시에 낮지 않을때) Farewell,
(v)	×	×	×	Farewell G-T, Log-rank, Gehan	G-T(치료율이 낮지 않을때) Proportion(중도절단율과 치료율이 동시에 낮지 않을때) L-M Farewell,

치료율을 비교하기 위한 임상시험에서 로그-순위검정이나 Gehan의 검정을 사용하는 것은 위험하다는 것을 알 수 있다. 따라서 치료율의 비교를 위해 고안된 검정법들을 적절히 선택하는 것이 필요하다. 위의 모의실험결과는 연구자로 하여금 각자의 상황에 적합한 분석방법을 선택하는 지표가 될 수 있을 것이다.

참고문헌

Arbutiski, T. (1985). A family of multiplicative survival models incorporating a long-term survivorship parameter C as a function of covariates, *Communications in statistics A -Theory and Methods*, **14**, 1627-1642.

Cox, D. R. (1970). *The Analyis Binary Data*, Methuen, London.

Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors, *Biometrics*, **38**, 1041-1046.

Goldman, A. (1984). Survivorship analysis when cure is a possibility: a Monte Carlo study, *Statistics in Medicine*, **3**, 153-163.

Gray, R. J. and Tsiatis, A. A. (1989). A linear rank test for use when the main interest is in difference in cure rates, *Biometrics*, **45**, 899-904.

Guitany, M. E., Maller, R. A., and Zhou, S. (1994). Exponential mixture models with long-term survivors and covariates. *Journal of Multivariate Analysis*, **49**, 218-241.

Jones, D. R., Powels, R. L., Machin, D., and Sylvester, R. J. (1981). On estimaing the proportion of cured patients in clinical studies, *Biometrie-Praximetrie*, **21**, 1-11.

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

Kuk, Y. C. and Chen, C. H. (1992). A mixture model combining logistic regression with proportional hazards regression, *Biometrika*, **79**, 531-541.

Laska, E. M. and Meisner, M. J. (1992). Nonparametric estimation and testing in a cure model, *Biometrics*, **48**, 1223-1234.

Lee, J. W. and Sather, H. L. (1995). Group sequential methods for comparison of cure rates in clinical trials. *Biometrics*, **51**, 756-763.

Peng, Y., Dear K. B. G., and Denham, J. W. (1998). A generalized F mixture model for cure rate estimation. *Statistics in Medicine*, **17**, 813-830.

Peng, Y. and Dear K. B. G. (2000). A nonparametric mixture model for cure rate estimation, *Biometrics*, **56**, 237-243.

Sposto, R., and Sather, H. N. and Baker, S. A. (1992). A comparison of tests of the difference in the proportion of patients who are cured, *Biometrics*, **48**, 87-99.

Sy, J. P. and Taylor, J. M. G. (2000). Estimation in a Cox proprtional hazards cure model, *Biometrics*, **56**, 227-236.

Tarone, R. E. and Ware, J. (1977). On distribution free tests for equality of survival distributions, *Biometrika*, **64**, 156-160.

Yamaguchi, K. (1992). Accelerated failure-time regression models with a regression model of surviving fraction: an application to the analysis of permanent employment in Japan, *Journal of the American Statistical Association*, **87**, 284-292.

[2003년 6월 접수, 2003년 10월 채택]

Simulation Study for Statistical Methods in Comparing Cure Rates between Two Groups *

Mira Park [1] Jae Won Lee [2] Seohoon Jin [3]

ABSTRACT

In some clinical trials, one may see that a significant fraction of patients are cured and their original disease does not recur even after termination of treatment and prolonged follow-up. This situation occurs frequently in pediatric cancer trials where there are excellent therapeutic results. In such cases, interest concentrated on the difference of cure rates rather than other types of differences in failure distributions. Various authors have investigated the parametric and nonparametric methods for testing the difference of cure rates.

In this study, we compare by simulation the power and size of a parametric test and five nonparametric tests in a various range of the alternatives, censoring rates and cure rates. Our objectives are to determine if any test was preferable on the basis of size and power in various situation, and to investigate the effect of the model misspecification.

Keywords: Cure rate; Simulation; Mixture model; Proportional hazard model; Kaplan-Meier estimate.

* This research was supported by Bum-Suk Academic Scholarship Foundation.

1) Assistant Professor, Dept. of Premedicine, Eulji University, 143-5 yongdu-dong chung-gu, Daejon, 301-832, Korea.
E-mail: mira@eulji.ac.kr

2) Professor, Department of Statistics, Korea University, 5-1 anam-dong seongbuk-gu, 136-701, Korea.
E-mail: jael@korea.ac.kr

3) Manager, Credit Card Marketing Team, Kookmin Bank, 15-22 yoido-dong yongdeungpo-gu, 150-757, Korea.
E-mail: bobbiej@passmail.to