응용통계연구 제 19권 3호, 2006년, pp. 395–406

기간분석에 따른 수정된 누적한계 추정량*

김진흠[1] 안윤옥[2]

요 약

임상시험 연구나 역학 연구에서 환자들의 예후는 흔히 생존율 추정을 통해 수량화 되곤 한다. 하지만 코호트분석이나 완전분석에 의한 생존율 추정량들은 수년 전에 진단된 환자에 크게 의존하기 때문에 실재 생존율보다 더 낮게 추정하곤 한다. 본 연구에서는 최근의 생존정보를 잘 반영하는 생존율 추정을 위해 기간분석 방법을 통한 누적한계 추정량을 제안하였고, 그 방법을 1993년 1월-1997년 12월 사이에 조사된 서울시 암등록 자료(Ahn 등, 2002)에 적용하여 결과를 고찰하였다.

주요용어: 기간분석, 누적한계 추정량, 암등록 자료, 좌절단

1. 서론

임상시험 연구나 역학 연구에서 환자들의 예후는 흔히 생존율의 추정을 통해 수량화 되곤 한다. 일반적으로 관심 있는 몇 년 후 생존율은 질병에 따라 변하는데, 만성 질환인 경우에는 몇 십 년씩이 되기도 한다. 예를 들어 암 환자들의 생존율을 예측하고자 할 때 연구자들은 보통 5년, 10년 등과 같은 장기간 생존율에 관심을 갖는다. 전통적으로 흔히 사용되는 생존율 추정법 즉, 코호트분석 또는 actuarial 추정법(Cutler와 Ederer, 1958)과 누적한계 추정법(Kaplan과 Meier, 1958)를 포함하는 완전분석을 통해 장기간 생존율을 추정했을 때, 그 값이 실제 생존율보다 낮게 나오는 경향이 있다. 다시 말해 코호트분석이나 완전분석에 의한 생존율 추정법은 최근의 생존정보에 민감하지 못한 단점을 갖고 있는데, 이는 전통적인 추정법이 최근에 연구에 포함된 환자보다 오래 전에 연구에 들어온 환자에 의존하기 때문이고, 특히 암 예후에 진보가 있을 때 이에 대한 초기 탐지가 전통적인 추정법으로는 불가능하기 때문이다. 이와 같은 경향은 이전보다 우수한 새로운 치료법이 개발되어 생존율이 점차 증가하는 예에서 두드러지게 나타난다.

만성 질환을 앓고 있는 환자들의 생존시간은 대체로 길고, 시간이 지남에 따라 새로운 치료법이 지속적으로 개발되어 이전에 연구에 들어온 환자의 생존율과 최근에 진단된 환자의 생존율이 달라 전통적인 추정법을 통해 추정된 생존율은 대체적으로 실제보다 작게 추정될 수밖에 없다. 따라서 최근 개발된 치료법으로 처치를 받았을 때 실제로는 그 보다 높

* 이 논문은 2005년도 정부재원(교육인적자원부 학술연구조성사업비)으로 한국학술진흥재단의 지원을 받아 수행된 연구임(KRF-2005-041-C00094)

1) (445-743) 경기도 화성시 봉담읍 와우리 산 2-2호, 수원대학교 자연과학대학 통계정보학과, 부교수
E-mail: jinhkim@suwon.ac.kr
2) (110-799) 서울시 종로구 연건동 28번지, 서울대학교 의과대학 예방의학교실, 교수
E-mail: yoahn@plaza.snu.ac.kr

은 생존율을 기대할 수 있음에도 불구하고 전통적인 추정법에 의해 실제보다 낮은 생존율 정보를 제공함으로서 최근 진단된 환자로 하여금 치료에 대한 용기를 잃게 할 수도 있다.

본 연구에서는 장기간 생존자료의 생존율 추정에서 최근의 자료 정보 즉, 최근에 개발된 치료법으로 처치를 받은 환자들의 정보를 잘 반영할 수 있는 추정법을 제안하고자 한다. 이를 위해 2절에서는 전통적인 추정법을 먼저 간략히 소개한 후, 기간분석에 따른 누적한계 추정량을 제안하고자 한다. 3절에서는 제안한 추정법을 1993.1.1-1997.12.31 기간 동안 등록되어 2001.12.31까지 추적 조사된 79,707명의 서울시 암등록 자료(Ahn 등, 2002)에 적용하여 그 결과를 고찰하고자 한다.

2. 생존율 추정법의 비교

표 2.1은 2000년부터 2005년 사이에 진단되어 2005년까지 조사된 환자들에 대해 3년 생존율 유도하고자 할 때 생존율 계산에 포함되는 추적기간을 분석 방법에 따라 나타낸 가상의 예이다. 예를 들어 2000년에 진단된 환자가 만일 2000년에 사망하거나 센서링되었다면 이 환자는 1년 생존율 계산에 기여하므로 '1'로 표시하였고, 만일 2001년에 사망하거나 센서링되었다면 이 환자는 추적기간에 따라 1년 이하, 1년 이상 모두 가능하기 때문에 1년 또는 2년 생존율 계산에 기여하므로 '1/2'로 표현하였다. 이와 동일한 방법으로 진단연도와 이벤트(사망 또는 센서링) 발생연도에 따라 생존율 계산에 기여하는 연수를 표 2.1과 같이 표시할 수 있다.

코호트분석에서 3년 생존율을 계산하고자 할 때 표 2.1에서 살펴볼 수 있듯이 2000-2005 기간에 포함된 환자들이 모두 분석대상이 되지는 않는다. 관심 있는 생존기간보다 크거나 같은 환자들만이 코호트에 포함되기 때문에 2003-2005 기간에 진단된 환자들은 코호트에서 제외된다. 다시 말해 코호트분석에 의한 3년 생존율은 2000-2002 기간에 진단된 환자들 중에서 3년이 지난 이후에도 생존해 있는 환자들의 비율로 추정된다. 한편 생존율 추정에서 코호트분석보다 널리 쓰이는 것은 actuarial 추정법 또는 누적한계 추정법과 같은 완전분석 방법이다. 완전분석 방법은 코호트분석 방법과 달리 조사 기간 2000-2005 내에 있는 모든 환자를 분석대상에 포함하여 생존율을 추정한다. 따라서 완전분석 방법은 최근 즉, 2003-2005 기간에 진단된 환자들의 초기 생존경험을 포함하기 때문에 코호트분석 방법보다는 최근의 생존정보를 더 잘 반영할 뿐만 아니라 생존율을 더 정확히 추정할 수 있는 장점을 지니고 있다.

2.1. 기간분석에 의한 actuarial 추정량

장기간 생존자료의 생존율 추정에서 전통적인 코호트분석이나 완전분석 방법이 새로운 치료법의 진보된 예후를 잘 반영하지 못하기 때문에, Brenner와 Gefeller(1996, 1997)는 정의된 최근 기간의 자료에만 actuarial 추정법을 적용한 소위 기간분석 방법을 제안하였다. 기간분석은 최근에 변화된 생존정보를 잘 반영할 수 있는 추정량을 얻기 위해 완전분석 방법처럼 조사기간 내 전체자료를 분석대상으로 하지 않고 최근 일정기간(소위 P:달력 날짜 기준)의 자료만을 대상으로 한다. 따라서 완전분석 방법과 다르게 정의된 기간 P 이전에

표 2.1: 생존율 추정법에 따른 진단연도와 3년 생존율 계산에 포함되는 추적기간의 비교

추정법	진단연도	이벤트(사망 또는 센서링) 발생연도					
		2000	2001	2002	2003	2004	2005
코호트	2000	1	1/2	2/3	3		
	2001		1	1/2	2/3	3	
	2002			1	1/2	2/3	3
완전	2000	1	1/2	2/3	3		
	2001		1	1/2	2/3	3	
	2002			1	1/2	2/3	3
	2003				1	1/2	2/3
	2004					1	1/2
	2005						1
기간	2000				3		
	2001				2/3	3	
	2002				1/2	2/3	3
	2003				1	1/2	2/3
	2004					1	1/2
	2005						1

사망했거나 센서링된 자료는 분석에 더 이상 포함되지 않는다. 이렇게 최근 일정기간을 정함으로 해서 오랜 시간 이전에 진단된 환자들의 생존정보가 생존율에 미치는 영향이 감소하고, 최근에 진단된 환자들의 정보가 추정량에 미치는 영향이 상대적으로 증가하기 때문에 최근에 진단된 환자의 생존율 추정에서 완전분석 방법보다 더 타당한 생존율을 제공해 줄 수 있다. 표 2.1에서 보여주듯이 만일 2003-2005 기간의 생존경험만을 갖고 3년 생존율을 추정할 때 2000-2002 기간에 사망했거나 센서링 된 환자들은 분석에서 제외되기 때문에, 시간이 지남에 따라 새로운 치료법이 개발되어 점진적으로 생존율이 커지는 예에서는 다른 분석방법보다 기간분석을 통해 생존율을 추정하는 것이 타당한 측면이 있는 반면에,정의된 기간에 포함되지 않는 자료를 제외함으로 인해 자료의 손실을 낳는 부정적인 측면 또한 갖고 있다. 결국 이와 같은 유효 자료수의 감소는 추정량의 분산을 크게 하는 결과로 이어질 수도 있다. 이런 단점에도 불구하고 암등록 자료와 같이 대용량 임상자료의 경우에는 특정 기간의 정의로 인한 표준오차의 증가가 크지 않기 때문에 요즘 장기간 생존자료의 생존율 추정법으로 기간분석 방법이 널리 사용되고 있으며, 몇 가지 임상 사례들을 Brenner 등(2001), Brenner(2002), Brenner와 Hakulinen(2002), Brenner 등(2002) 등에서 찾아 볼 수 있다.

상술한 기간분석에 의한 actuarial 추정법을 좀 더 구체적으로 설명하면 다음과 같이 요약할 수 있다(Brenner와 Gefeller, 1996). 생명표의 간격은 일반성을 잃지 않고 1년으로 할

수 있다. 구간 $I_x = (x-1, x], x = 1, \ldots, l,$ 이라 놓을 때, x는 진단된 달력 연도로부터 추적기간 연수, l은 최대 추적기간 연수를 뜻한다. p_x를 구간 I_x의 시작점에서 생존해 있는 개체가 1년 후 생존할 확률이라 놓으면, 추적기간 k년 이후 생존율은 아래와 같이 표현된다.

$$S_k = \prod_{x=1}^{k} p_x.$$

R_x를 구간 I_x의 시작점에서 위험에 여전히 노출되어 있는 인구 집단의 크기, d_x를 구간 I_x에서 사망한 사망자 수라 놓으면, 상술한 분석 방법에 관계없이 p_x는 $(1 - d_x/R_x)$로 추정될 수 있고 따라서 생존율 추정이 가능하다. 여기서 주목할 점은 구간 내에서 센서링된 개체에 대해서는 절반의 정보만이 R_x에 포함된다. 코호트분석에서는 선정한 코호트, 완전분석에서는 전체자료에 각각 이와 같은 추정법을 한번만 적용하여 생존율을 추정한다. 반면에 기간분석에서는 두 번의 과정을 거쳐 생존율을 추정한다. 첫 번째 과정에서는 완전분석 방법에서처럼 전 연구기간 자료(f), 두 번째 과정에서는 관심기간 P의 시작연도까지 만의 자료(r)를 가지고 각 구간에서 $R_x^{(f)}(R_x^{(r)})$와 $d_x^{(f)}(d_x^{(r)})$를 구한 후, 그 차이 $R_x = R_x^{(f)} - R_x^{(r)}$와 $d_x = d_x^{(f)} - d_x^{(r)}$를 써서 생존율을 추정한다. 표 2.1의 예로 기간분석을 다시 설명하면, 전체 연구기간이 2000-2005이고 관심기간이 2003-2005이므로 2000-2005 기간의 자료(f)와 2000-2002 기간의 자료(r)에 상술한 방법을 적용하여 생존율을 추정한다.

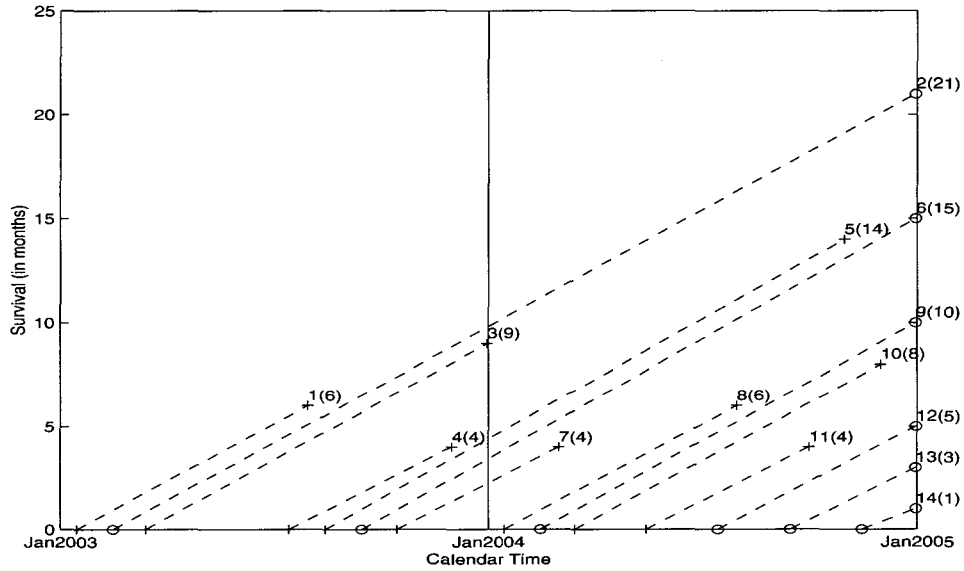

그림 2.1: 2003-2004 기간에 진단된 환자들의 진단시점과 생존시간에 따른 Lexis 다이어그램 (숫자는 환자번호(생존시간 혹은 센서링 시간), '+'는 사망, 'o'는 센서링을 표시)

2.2. 기간분석에 의한 누적한계 추정량

일반적으로 actuarial 추정법에서는 이미 그룹화 되어 관측된 생존시간 자료의 생존율을 추정하거나 또는 관측한 생존시간 자료를 일정한 간격(예, 몇 개월 단위, 몇 년 단위 등)에 따라 그룹화 한 후 생존함수를 조밀(fine)하게 추정하기 보다는 설정한 구간의 끝점에서만 생존율을 추정하는 데 초점이 맞추어 있다. 반면에 누적한계 추정법에서는 관측한 생존시간 각각을 개별 그룹으로 형성하여 다루기 때문에 actuarial 추정법 보다는 조밀한 생존함수 추정이 가능케 된다. 한편 그룹화 하는 간격을 작게 하면 할수록 actuarial 추정량은 누적한계 추정량으로 수렴하기 때문에 이 때 두 추정법은 서로 동등(equivalent)하다고 말할 수 있다(Breslow와 Crowley, 1974; Thompson, 1977). 본 연구에서는 최근에 변화된 생존정보를 잘 반영할 수 생존함수 추정을 위해 전체 자료를 다 분석에 포함하는 완전분석 방법 대신에 Brenner와 Gefeller(1996, 1997)가 제안한 기간분석 방법을 따르고자 하며, 조밀한 생존율 추정을 위해 actuarial 추정법 대신에 누적한계 추정법을 이용한 생존함수 추정량을 제안하고자 한다.

기간분석에 포함되는 환자들은 최근부터 이전 일정 기간 P 내에 관찰되는 환자들이기 때문에 기간분석의 생존자료는 마치 좌절단(left-truncated) 되고 우센서링(right-censored) 된 자료라고 할 수도 있다. 다시 말해 관심기간 P의 시작시점(달력 기준 날짜) 이전에 이벤트가 발생한 환자들은 연구대상에서 제외되고 그 이전에 진단되어 시작시점까지 살아 있는 환자나 관심기간 P 내에 진단된 환자만이 연구대상에 포함되기 때문이다. 기간분석에 포함되는 자료와 좌절단되고 우센서링된 자료가 이와 같이 유사한 측면이 있지만, 관심기간 이전에 진단되어 관심기간의 시작시점까지 이벤트가 발생하지 않은 환자의 경우에 전자는 관심기간 이전 시점까지의 생존경험을 함께 포함하는 반면, 후자는 관심기간 이후 시점부터의 생존경험만을 포함하는 면에서는 서로 다르다고 할 수 있다.

따라서 생존율 추정량을 정의함에 있어 중요한 점은 형태적으로는 누적한계 추정량을 따르지만, 각 생존시간에서 위험집합(risk set)과 사망집합(failure set)을 적절히 정의하는 데 있다. 여기서 또 한 가지 유의할 것은 생존시간의 정의인데, 관심기간 P의 시작시점과 이벤트 발생시점 간의 차이가 아니라 진단시점과 이벤트 발생시점 간의 차이라는 점이다. 서로 다른 생존시간을 $t_j(j = 1, \ldots)$, t_j에서 위험집합에 속하는 환자수를 r_{tj}, 사망집합에 속하는 환자수를 d_{tj} 라고 하면, 진단된 시점부터 t 시간 이후 생존율 S_t는 아래와 같이 추정된다.

$$\hat{S}_t = \prod_{t_j \le t} \left(1 - \frac{d_{tj}}{r_{tj}}\right).$$

이 때 P의 양 끝점은 달력 기준 날짜이고, 생존시간은 달력 기준과 무관하기 때문에 r_{tj}와 d_{tj}를 명시적으로 정의할 때 주의가 요망된다. 이를 위해 r_{tji}는 i번째 환자가 관심기간 P 동안 생존시간 t_j에서 위험에 노출되어 있으면 1, 그렇지 않으면 0으로 정의하고, d_{tji}는 i번째 환자가 관심기간 P동안 생존시간 t_j로 이벤트가 발생하면 1, 그렇지 않으면 0으로 정의하자. 그러면 r_{tj}와 d_{tj}는 아래와 같이 표현된다.

$$r_{tj} = \sum_i r_{tji}, \quad d_{tj} = \sum_i d_{tji}.$$

표 2.2: 그림 2.1 자료의 P=2004년에 대한 환자별 위험집합과 사망집합

	r_{tji}(위험집합)				d_{tji}(사망집합)			
i(환자번호)	t_1	t_2	t_3	t_4	t_1	t_2	t_3	t_4
1	-	-	-	-	-	-	-	-
2	0	0	0	1	0	0	0	0
3	-	-	-	-	-	-	-	-
4	-	-	-	-	-	-	-	-
5	0	1	1	1	0	0	0	1
6	1	1	1	1	0	0	0	0
7	1	0	0	0	1	0	0	0
8	1	1	0	0	0	1	0	0
9	1	1	1	0	0	0	0	0
10	1	1	1	0	0	0	1	0
11	1	0	0	0	1	0	0	0
12	1	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0
합계	7	5	4	3	2	1	1	1

그림 2.1의 예를 갖고 위험집합과 사망집합을 간략히 설명하면, 14개의 자료 중에서 관측된 생존시간(단위: 월)은 $t_1 = 4$(환자 7, 11), $t_2 = 6$(환자 8), $t_3 = 8$(환자 10), $t_4 = 14$(환자 5)이므로 각 t_j에서 $r_{tji}, d_{tji}(i = 1, \ldots, 14; j = 1, \ldots, 4)$는 표 2.2와 같이 주어진다. 2004년 이전에 진단되어 $P = 2004$의 시작시점에 생존해 있는 환자(환자 2, 5, 6, 7)만 먼저 살펴보면, 환자 2의 경우는 센서링 시간이 21개월이지만 P에 진입한 시점은 진단 후 9-10개월 후이기 때문에 $t_1 - t_3$ 에서는 위험집합에서 제외되고 $t_3 = 8$ 이후부터 위험집합에 포함되며, 환자 5의 경우는 생존시간이 14개월이지만 P에 진입한 시점은 진단 후 4-5개월 후이기 때문에 $t_1 = 4$ 이후부터 위험집합에 포함되고 $t_4 = 14$에서 사망집합에 포함된다. 환자 6의 경우는 센서링 시간이 15개월이지만 P에 진입한 시점은 진단 후 3-4개월 후이기 때문에 $t_1 = 4$ 부터 위험집합에 포함되고, 환자 7의 경우는 생존시간이 4개월이고 P에 진입한 시점은 진단 후 1-2개월 후이기 때문에 $t_1 = 4$에서 위험집합과 사망집합에 동시에 포함된다. 관심기간 $P = 2004$ 기간 동안에 진단된 환자(환자 8-14)의 경우에는 우센서링에만 영향을 받기 때문에 위험집합과 사망집합에 기여하는 정도를 쉽게 알 수 있다. 결국 $t_j(j = 1, \ldots, 4)$에서 생존율은 각각 $\hat{S}_4$=0.714, $\hat{S}_6$=0.571, $\hat{S}_8$=0.429, $\hat{S}_{14}$=0.286 으로 추정된다. 한편 완전분석 방법에 따른 누적한계 추정량으로 생존율을 추정하면 $\hat{S}_4$=0.75, $\hat{S}_6$=0.563, $\hat{S}_8$=0.469, $\hat{S}_9$=0.375, $\hat{S}_{14}$=0.25와 같이 얻어져 두 분석방법 간에 차이가 있음을 알 수 있다.

3. 적용 예

본 절에서는 1993.1.1-1997.12.31 기간 동안에 등록되여 2001.12.31까지 추적 조사된 79,707명의 서울시 암등록 자료(Ahn 등, 2002)에 제안한 추정법을 적용하여 그 결과를 고찰하고자 한다. 본 연구에서는 성별과 연령의 조합에 따라 8개의 층으로 층화하여 완전분석과 기간분석 방법에 따른 actuarial 추정량과 누적한계 추정량을 서로 비교하였다. 기간분석을 위해 관심기간은 P=[1997.1.1, 2001.12.31]로 설정하였는데 시작연도는 1년 생존율을 추정할 수 있으며 또한 최근 생존정보를 잘 담아 낼 수 있도록 정하였다.

표 3.1에서 알 수 있듯이 성별 분포는 남자(54%), 여자(46%)이고, 연령 분포는 20세 미만(2.2%), 20세-40세(11.8%), 40세-60세(41.6%), 60세 이상(44.4%)으로 나타났다. 설정한 관심기간으로 인해 좌절단되는 자료의 비율은 층에 따라 15%-43%에 이르렀으며, 성별로는 남성이 여성보다, 연령별로는 고령층으로 옮아 갈수록 절단되는 비율이 커짐을 알 수 있었다. 센서링 비율에 있어서는 성별, 연령에 관계없이 완전분석 방법의 경우보다 기간분석의 경우에 더 높게 나타났다.

표 3.1: 성별과 연령에 조합에 따른 자료수(n), 센서링 비율(%, f_c: 완전분석 방법, f_p: 기간분석 방법), 기간분석에서 제외되는 대상자 비율(%, h)

표본	남자				여자			
통계량	<20	20-40	40-60	60<	<20	20-40	40-60	60<
n	1,004	3,258	17,969	20,519	772	6,162	15,089	14,578
f_c	53	51	39	28	61	72	63	36
f_p	73	73	62	50	82	85	79	59
h	27	31	37	43	25	15	20	39

먼저 완전분석과 기간분석 방법에 따른 actuarial 추정량을 1년 단위로 1년-8년 후 생존율을 각각 살펴보았는데, 그중 1년, 3년, 7년 후 생존율(표준오차 포함)만을 표 3.2에 제시하였다. 표준오차는 Greenwood 공식(Greenwood, 1926)을 통해 구했으며, 표 3.2에서 c_{lt}, p_{lt}는 각각 완전분석과 기간분석에 따른 actuarial 추정량을 나타낸다. 성별과 연령 조합에 따른 8개의 모든 층에서 완전분석 방법을 따랐을 때보다 기간분석 방법을 따랐을 때 더 큰 값의 생존율 추정값을 얻을 수 있었다. 한편, 예상했던 것처럼 완전분석 방법의 경우보다 기간분석 방법의 경우에 표준오차 또한 크게 추정됐지만 완전분석 방법에 의한 actuarial 추정량의 값이 기간분석 방법에 의한 생존율의 95% 신뢰구간 내에 포함되지 않기 때문에 두 방법은 유의수준 5%에서 통계적으로 차이가 있음을 알 수 있었다.

아울러 본 연구에서 제안한 기간분석에 의한 누적한계 추정량과 완전분석에 의한 누적한계 추정량을 비교하였는데, 그 결과를 성별과 연령 조합에 따라 그림 3.1에 나타냈다. 그림 3.1에서 포인트 없는 점선과 실선은 남자의 경우에 해당하며 각각 완전분석과 기간분석에 따른 누적한계 추정값을 나타내며, 포인트 있는 점선과 점선은 여자의 경우에 해당하며

표 3.2: 성별과 연령 조합에 따른 진단 1년, 3년, 7년 후 생존율에 대한 actuarial 추정량(표준오차)(c_{lt}: 완전분석 방법, p_{lt}: 기간분석 방법)과 누적한계 추정량(표준오차)(c_{pl}: 완전분석 방법, p_{pl}: 기간분석 방법)의 비교

	방	남자				여자			
년	법	<20	20-40	40-60	60<	<20	20-40	40-60	60<
1	c_{lt}	.7440	.6949	.6071	.5201	.7668	.8616	.8142	.5808
		(.0138)	(.0081)	(.0036)	(.0035)	(.0152)	(.0044)	(.0032)	(.0041)
	p_{lt}	.8028	.7415	.6610	.5802	.8128	.8966	.8531	.6375
		(.0236)	(.0147)	(.0067)	(.0064)	(.0264)	(.0070)	(.0051)	(.0074)
	c_{pl}	.7430	.6943	.6069	.5195	.7668	.8614	.8138	.5805
		(.0138)	(.0081)	(.0036)	(.0035)	(.0152)	(.0044)	(.0032)	(.0041)
	p_{pl}	.7865	.7276	.6544	.5702	.8042	.8909	.8453	.6229
		(.0245)	(.0150)	(.0067)	(.0063)	(.0271)	(.0073)	(.0054)	(.0075)
3	c_{lt}	.5837	.5479	.4395	.3353	.6451	.7597	.6872	.4171
		(.0156)	(.0087)	(.0037)	(.0033)	(.0172)	(.0054)	(.0038)	(.0041)
	p_{lt}	.6292	.5901	.4868	.3803	.7123	.7934	.7273	.4647
		(.0243)	(.0145)	(.0061)	(.0054)	(.0273)	(.0081)	(.0056)	(.0067)
	c_{pl}	.5837	.5479	.4394	.3350	.6451	.7593	.6871	.4169
		(.0156)	(.0087)	(.0037)	(.0033)	(.0172)	(.0055)	(.0038)	(.0041)
	p_{pl}	.6251	.5827	.4817	.3750	.7117	.7886	.7200	.4549
		(.0248)	(.0146)	(.0061)	(.0053)	(.0278)	(.0082)	(.0058)	(.0067)
7	c_{lt}	.5219	.4960	.3732	.2632	.6016	.7128	.6180	.3474
		(.0164)	(.0091)	(.0038)	(.0033)	(.0182)	(.0061)	(.0042)	(.0042)
	p_{lt}	.5629	.5341	.4128	.2975	.6680	.7446	.6543	.3861
		(.0238)	(.0140)	(.0057)	(.0048)	(.0271)	(.0083)	(.0057)	(.0062)
	c_{pl}	.5225	.4967	.3742	.2638	.6022	.7131	.6189	.3482
		(.0165)	(.0090)	(.0038)	(.0034)	(.0183)	(.0061)	(.0042)	(.0042)
	p_{pl}	.5589	.5286	.4102	.2954	.6635	.7408	.6488	.3803
		(.0241)	(.0141)	(.0057)	(.0048)	(.0278)	(.0085)	(.0058)	(.0062)

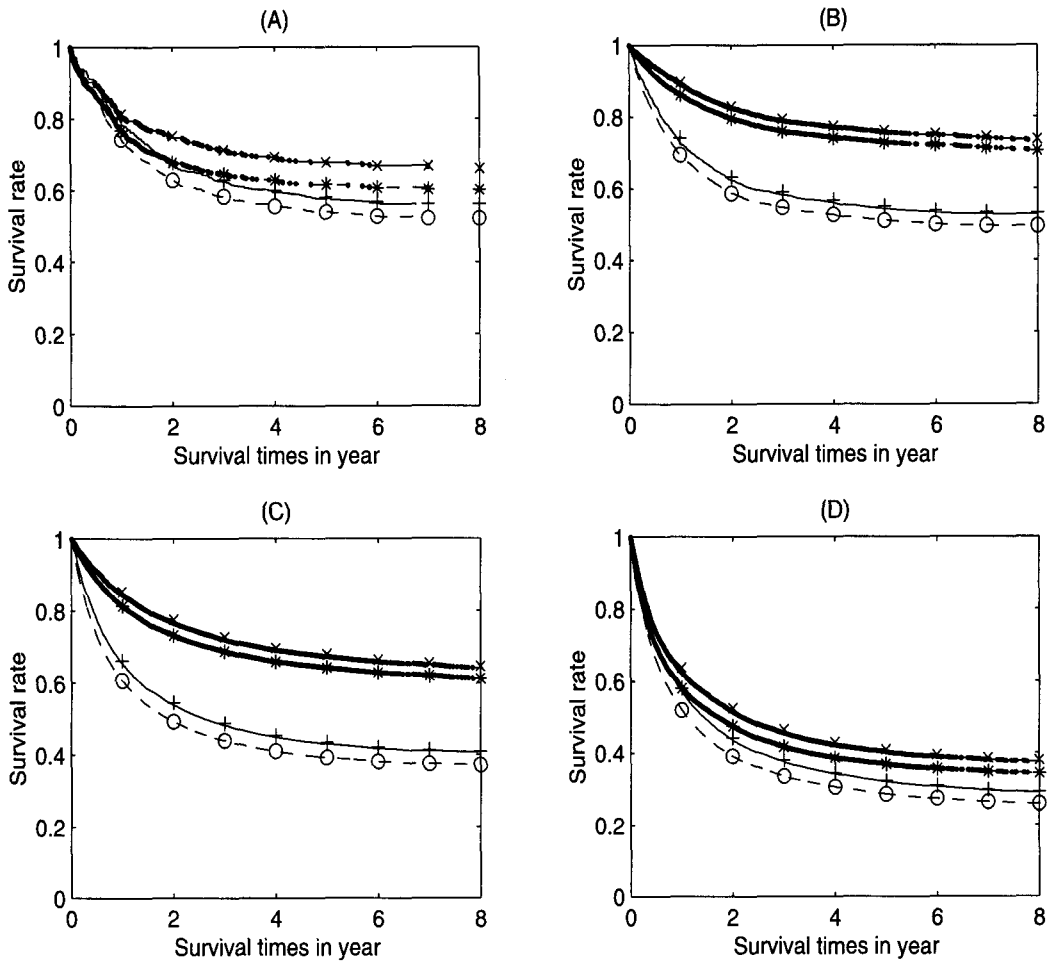

그림 3.1: 성별과 연령(A: 20세 이하, B: 20-40세, C: 40-60세, D: 60세 이상) 조합에 따른 생존함수 추정량의 비교 (기간분석에 의한 누적한계 추정량-남자:포인트 없는 실선, 여자: 포인트 있는 실선; 완전분석에 의한 누적한계 추정량-남자:포인트 없는 점선, 여자: 포인트 있는 점선; 기간분석에 의한 actuarial 추정량-남자:'+',여자:'×'; 완전분석에 의한 actuarial 추정량-남자:'o',여자:'∗')

각각 완전분석과 기간분석에 따른 누적한계 추정값을 나타낸다. 연령층에 무관하게 남자의 생존율이 여자보다 낮게 나왔으며 그 차이는 중간 연령층(20-60세)에서 두드러졌다. 성별에 무관하게 연령이 증가할수록 생존율이 낮았으며 인접 연령 층간 차이는 고령층으로 옮아 갈수록 커짐을 알 수 있었다. 마지막으로 제안한 추정량과 완전분석에 의한 누적한계 추정량과의 비교에서는 고려한 모든 층에서 제안한 추정량의 값이 크게 나와 기간분석에 의한 누적한계 추정량이 완전분석에 의한 누적한계 추정량보다 최근의 생존정보를 더 잘 반영하고 있음을 알 수 있었다.

한편 1년 단위 생존율을 중심으로 actuarial 추정량과 누적한계 추정량을 비교하기 위해 표 3.2에는 완전분석과 기간분석에 따른 누적한계 추정량(c_{pl}, p_{pl})을 actuarial 추정량과 함께 제시하였고, 그림 3.1에는 1년 단위 actuarial 추정량을 누적한계 추정량의 결과와 함께 나타냈다. 그림 3.1에서 'o'과 '+'은 남자의 경우 각각 완전분석과 기간분석에 따른 actuarial 추정량에 해당하고, '*'과 '×'은 여자의 경우 각각 완전분석과 기간분석에 따른 actuarial 추정량에 해당한다. 표 3.2와 그림 3.1를 통해 동일한 분석방법 내에서는 년 단위 생존율에 있어 두 추정량(c_{lt} vs. c_{pl}; p_{lt} vs. p_{pl}) 간에 통계적으로 유의한 차이가 없음을 알 수 있었으나, 그 차이는 완전분석에서 보다는 기간분석에서 약간 크게 나왔다. 이와 같은 결과는 1년 단위로 그룹화된 자료에 해당하는 것이고 그룹 간격에 따라 결과는 변화하기 때문에 결과에 대한 해석은 주어진 그룹 간격으로 제한하는 것이 바람직하다고 생각한다.

4. 고찰 및 향후 과제

본 연구에서는 장기간 생존자료의 생존율 추정에서 최근 생존정보를 잘 반영할 수 있는 추정법을 제안하고, 그 추정법을 암등록 자료에 적용하여 그 결과를 살펴보았다. 의료기술 및 의학의 발달로 매년 한층 우수한 치료법이 지속적으로 개발되고 있기 때문에 암 환자들의 생존율은 증가 추세에 있다고 말할 수 있다. 따라서 완전분석 방법보다는 기간분석 방법이 그와 같은 변화를 잘 탐지할 수 있다고 여겨진다. 분석한 자료의 결과에 따르면 두 방법 간 차이가 통계적으로 유의할 뿐만 아니라 기간분석에 의한 누적한계 추정량이 완전분석에 의한 누적한계추정량보다 더 크게 나와 전자가 증가하는 실제 생존율을 더 잘 반영하고 있음을 알 수 있었다. 다른 한편으로는 관심기간 이전에 이벤트가 발생한 환자들이 분석대상에서 제외됨으로서 유효 표본수가 줄어들고 그로 인해 제안한 추정량의 표준오차는 누적한계 추정량의 표준오차보다 커질 수밖에 없는 단점을 갖고 있지만 대용량 자료에서는 그 차이가 크지 않기 때문에 심각한 문제로 받아드려지고 있지는 않다.

향후에는 수정된 누적한계 추정량에 대한 점근적 성질을 고찰하고, 생존율이 시간의 변화에 따라 증가하는 가상의 모형하에서 두 생존함수 추정법을 비교하고자 한다.

감사의 글

귀중한 자료를 제공해 주시고 토론해 주신 성균관대학교 의과대학 사회의학교실 신명희 교수님에게 깊이 감사드립니다.

참고문헌

Ahn Y. O., Shin M. H., Kim J. P. (2002). Korea, Seoul, In *Cancer Incidence in Five Continents, Vol. VIII*, IARC Scientific Publications No. 155 (Parkin, D.M., Whelan S.L., Ferlay, J., Teppo, L., and Thomas, D.B. eds), IARC, Lyon, 276-277.

Brenner, H. and Gefeller, O. (1996). An alternative approach to monitoring cancer patient survival, *Cancer*, **78**, 2004-2010.

Brenner, H. and Gefeller, O. (1997). Deriving more up-to-date estimates of long-term patient survival, *Journal of Clinical Epidemiology*, **50**, 211-216.

Brenner, H., Gefeller, O., Stegmaier, C., and Ziegler, H. (2001). More up-to-date monitoring of long-term survival rates by cancer registries: an empirical example, *Methods of Information in Medicine*, **40**, 248-252.

Brenner, H. (2002). Long-term survival rates of cancer patients achieved by the end of the 20th century: a period analysis, *Lancet*, **360**, 1131-1135.

Brenner, H. and Hakulinen, T. (2002). Up-to-date long-term survival curves of patients with cancer by period analysis, *Journal of Clinical Oncology*, **20**, 826-832.

Brenner, H., Soderman, B., and Hakulinen, T. (2002). Use of period analysis for providing more up-to-date estimates of long-term survival rates: empirical evaluation among 370,000 cancer patients in Finland, *International Journal of Epidemiology*, **31**, 456-462.

Breslow, N. E. and Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship, *Annals of Statistics*, **2**, 437-453.

Cutler, S. J. and Ederer, F. (1958). Maximum utilization of the life table method in analyzing survival, *Journal of Chronic Diseases*, **8**, 699-712.

Greenwood, M. (1926). The natural duration of cancer, In *Reports on Public Health and Medical Subjects*, **33**, Her Majesty's Stationery Office, London, 1-26.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association*, **58**,457-481.

Thompson, W. A., Jr. (1977). On the treatment of grouped observations in life studies, *Biometrics*, **33**, 463-470.

[2006년 3월 접수, 2006년 6월 채택]

Modified Product-Limit Estimator via Period Analysis*

Jinheum Kim[1] Yoonok Ahn[2]

ABSTRACT

Long-term survival rates are the most commonly used outcome measures for patients with cancer. However, traditional long-term survival statistics, which are derived by cohort analysis or complete analysis, essentially reflect the survival expectations of patients diagnosed many years ago. They are often outdated at the time they become available. In this article, we propose a modified product-limit method to obtain up-to-date estimates of long-term survival rates via a period analysis. The proposed method is illustrated with cancer registry data collected from January 1993 to December 1997.

Keywords: Cancer registry data, Left truncation, Modified product-limit estimator, Period analysis

* This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD)(KRF-2005-041-C00094)

1) Associate Professor, Department of Applied Statistics, University of Suwon, Gyeonggi-Do, 445-743, Korea
E-mail: jinhkim@suwon.ac.kr

2) Professor, Department of Preventive Medicine, Seoul National University College of Medicine, Seoul, 110-799, Korea
E-mail: yoahn@plaza.snu.ac.kr