제2상 시험의 설계 및 통계분석법

한림대 통계학과

정신호, 박병주*, 손민주, 김지연

= Abstract =

Design and Analysis of Phase II Clinical Trials

Sin-Ho Jung, Byung-Joo Park*, Min-Joo Sohn, Ji-Yun Kim

Department of Statistics, Hallym University, Chuncheon, Korea

Department of Preventive Medicine, Seoul National University, Seoul, Korea.

A clinical trial may be defined as a carefully designed, prospective medical study which attempts to answer a precisely defined set of questions with respect to the effects of a particular treatment or treatments. Major types of clinical trials include phase I, II, III, and IV studies, although a more strict definition may limit to phase III studies only. A phase II study is to characterize, but not to compare, the therapeutic efficacy, toxicity, or quality of life of one or more therapeutic strategies for a given disease or condition, and then decide whether the regimen(s) merit further study in subsequent phase III trials. In designing of a phase II study, we decide the number of patients for a multi-stage design through an optimal strategy. In completion of a phase II study, one of major statistical analyses is to obtain the exact confidence interval of response (or toxicity) rate. This article reviews the existing statistical methodologies used in designing and analyzing phase II clinical trials or anti-cancer therapeutics.

KEY WORDS : Confidence interval · Multiple stage design · Number of patients · Toxicity · Response.

서울대학교 의과대학 예방의학교실*

교신저자: 정신호, 한림대학교 통계학과,

주소: 강원도 춘천시 옥천동 1(우: 200-702)

TEL: 0361-240-1443, FAX: 0361-56-3424

I. 서 론

어떤 암치료제(방사선 치료를 포함하여)에 대한 2상시험은 그 치료제의 항암 효과에 관한 첫 평가단계이다. 1상 시험에서는 그 치료법의 최대 허용 용량(maximum tolerable dose)을 정하는데, 이것은 대부분의 암치료제가 최대효과를 얻기 위하여 최대 용량으로 투여되어야 하기 때문이다. 그러나 1상 시험에서는 각 단계의 용량에서 단지 3내지 6명의 환자만을 투여하며, 이 환자들은 각기 다른 다양한 종류의 암을 가지고 있고 다양한 예후 상태를 가지고 있으므로 1상 시험으로서는 항암 효과를 적절히 평가할 수 없다. 2상시험의 주 관심사는 암부위의 크기가 50% 이상 줄어든 환자들의 비율(이하 반응률이라 부른다)을 파악하는 것이다. 이러한 2상시험들은 본격적인 임상시험 단계인 3상시험과는 달리 환자들의 예후 요인(prognostic factor)들에 의해 통제되지 않으며 대조치료군이 없이 신약 치료군 단독으로 시행된다. 더우기 2상 시험은 특정 암에 대한 새로 개발된 치료제의 역할이나 치료효과를 결정짓기 위한 것이 아니고, 이 신약이 더 이상의 연구가치가 있는지 검토하는 것이다. 즉, 이 신약을 완치를 기대할 수 있는 초기의 암환자들에 투여하여 치료효과를 평가해보고, 기존의 표준치료제(대조군의 치료제)와 비교하여 암환자들의 수명을 연장하는 효과가 있는가를 평가하는 3상 시험의 실시등을 진행할 것인지를 판단하기 위한 근거를 확보하고자 하는 것이 주된 목적이다.

본 논문에서는 2상시험과 관련하여 시험 설계와 결과 분석을 위하여 현재까지 발표되어 사용되었거나 현재 사용되고 있는 통계 방법들을 소개하고자 한다. 여기서는 이러한 통계 방법들의 개념만을 묘사하므로, 구체적인 내용은 해당 문헌을 참고해야 할 것이고 이들은 주로 복잡한 계산 과정을 요구하므로 실제 2상 시험에 사용하기 위해서는 효율적인 프로그램과 컴퓨터가 필요할 것이다.

II. 제2상시험의 설계

2상시험의 설계시엔 신약의 암치료제로서의 기능을 평가하기 위하여 그것의 효능을 어떻게 측정할 것인가를 정해야 한다. 2상시험에서 암 치료제의 기능에 관한 가장 일반적인 측정치는 반응여부이다. 설계 단계에서는 시험 결과 얻어질 반응률의 추정치가 어느 정도의 통계적 효율을 가질 수 있도록 시험에 필요한 환자수를 정한다. 환자수 결정에는 서로 상충하는 몇가지 기본원칙이 적용된다. 첫째로, 정확한 추정치를 얻기 위하여는 많은 수의 환자를 모집해야 한다. 둘째로, 그러나 시험 치료제의 효과가 낮거나 지나치게 독성이 강한 경우에 투여될 환자숫자를 최소화해야 하고, 셋째로, 신약의 효능에 대한 적절한 평가가 이루어지고 나면 더 이상의 환자를 투입하지 않아야 한다는 것이다. 2상시험의 설계단계에서는 시험에 필요한 환자수를 결정하고 동시에 시험이 진행되는 동안 신약이 효과가 있나 없나를 결정할 때 사용할 규칙을 미리 정해야 하는데 이러한 설계 과정에는 적절한 통계방법이 사용되어야 한다.

예로써 전이된 대장암 환자들에 대한 9-AC의 치료효과를 보는 2상 시험(phase II study for 2-AC in previously untreated patients with metastatic colorectal carcinoma)을 고려해 보자. 이러한 대상환자들에 대해 현재 사용되는 치료제들의 효능을 고려해 볼 때 신약의 반응률은 20%는 되어야 재고의 가치가 있어 보여서, 2상시험을 통하여 9-AC의 반응률이 20%를 넘는지를 알아보고자 한다. 2상시험에 필요한 환자수 계산을 위해 제안된 여러 가지 시험설계 방법들을 이 예에 적용시켜 보자.

1. 단일단계 설계(Single-Stage Design)

2상시험의 가장 단순하면서도 원시적인 형태의 설계로 단일단계 시험이 있다. 단일단계 2상시험에서는

시험에 배당된 경비와 모집 가능한 환자수 등을 고려하여 신약의 반응률에 대한 신뢰구간이 적절한 길이를 갖도록 필요한 환자수 N 을 정한다. 일반적으로 반응률 p 의 95% 신뢰구간의 길이 즉,

$$2 \times 1.96 \sqrt{\frac{p(1-p)}{N}} \text{ 가 } 30\% \text{에서 } 40\% \text{ 사이가 되도록}$$

환자수는 25명에서 40명 사이에서 결정된다. 환자수 N 이 결정되면 가능한 빠른 시일내에 이만큼의 환자를 모집하고 신약에 대한 각 환자의 반응 여부를 평가해서 반응률의 추정치, $\hat{p} = (\text{반응을 보인 환자수}) / (\text{총 환자수}, N)$,을 계산하고 이항분포를 사용하여 참 반응률에 대한 (정규분포를 사용한 근사가 아닌) 정확한 신뢰구간(예를 들어 Rosner의 Table 7 등을 참조)을 구한다¹⁾. 아울러 이 치료제가 충분한 효과(반응률)를 가지고 있는지도 검정한다.

9-AC시험의 예에서 $p=0.2$ 이고 반응률의 95% 신뢰구간의 길이가 30%~40% 정도가 되도록 하기 위해 $0.3 \leq 2 \times 1.96 \sqrt{0.2 \times 0.8 / N} \leq 0.4$ 를 만족하는 N 을 구해보면 16명 내지 28명이 된다. 이 시험에 24명의 환자를 모집했고 그 중 7명이 반응을 보였다고 하자. 이때 반응률의 추정치는 $7/24=0.29$ 이고 참 반응률에 대한 정확한 95% 신뢰구간은 (0.12, 0.52)가 된다. (Rosner, 1990의 Table 7 참조) 그리고 참반응률이 20%를 넘는가에 대한 검정의 p-value는 실제 반응률이 20%일 때 24명중 7명 이상이 반응을 보일 확률

$$= \sum_{k=7}^{24} \frac{24!}{k!(24-k)!} 0.2^k (1-0.2)^{24-k} = 0.1889 \text{ 로서}$$

시험결과 9-AC의 반응률은 유의하게 높아보이지는 않는다고 할 수 있다. 여기서 $k! = k \times (k-1) \times \dots \times 2 \times 1$ 이다.

단일단계 2상시험은 그 설계와 결과분석에 필요한 통계이론이 단순하다는 장점이 있으나 시행상에 그리고 윤리상에 이 장점을 능가하는 문제점들을 안고 있다. 일반적으로 임상시험에서는 필요한 수만큼의 환자수를 모두 확보한 다음 신약을 동시에 투여하는 것은 아니다. 환자들은 시간 간격을 두고 한 명씩 모집되며

필요한 수의 환자들이 모두 모집되기 전에 일찌기 모집되어 치료받은 환자들에서는 신약에 대한 반응 여부가 판가름나게 될 것이다. 이러한 상황에서, 일찌기 치료받은 환자들로부터 실통한 반응이 관찰되지 않으면 의사들은 더 이상의 환자를 이 시험에 투입하는 걸 꺼려하게 될 것이고 그 결과로 환자 모집 속도는 떨어져서 연구의 진행은 지지부진하게 될 것이다. 단일단계 2상시험의 윤리적인 문제점으로는 시험의 중반까지 충분한 반응을 보이지 않는 시험에 환자들을 끝까지 계속 모집하는 것이 합리화 될 수 있는가 하는 것이다. 단일단계 시험의 이러한 문제점들을 피하기 위하여 오늘날 대부분의 2상시험은 다단계로 설계되고 미국의 N.I.H. 나 F.D.A.와 같은 임상시험 감독 기관에서는 다단계 설계를 의무화하고 있다.

2. Gehan's Two-Stage Design (Gehan, 1961)

2상시험의 다단계 설계를 중 가장 일찌기 만들어졌고 비교적 널리 사용되었던 것으로서 Gehan의 2단계 설계가 있다²⁾. 여기서는 1단계에 n_1 명 그리고 2단계에 n_2 명을 투여한다. 1단계에서 n_1 명 환자들 중 한명도 반응을 보이지 않으면 시험 신약은 효과가 없다고 결정을 내리고 시험은 여기서 중단된다. 신약 채택 여부의 경계가 되는 반응률 p 에 대해 n_1 명 중 한명도 반응을 보이지 않을 확률 $(1-p)^{n_1}$ 이 매우 작도록 (주로 이 확률이 0.05정도 되도록) n_1 값이 정해진다. 1단계에서 1명 이상이 반응을 보이면 2단계로 넘어가 시험은 계속되는데, 2단계에서 필요한 환자수 n_2 는 전체환자수 $N = n_1 + n_2$ 를 사용하여 반응률을 추정했을 때 추정치의 표준오차가 적당히 작도록 (10% 정도) 결정된다.

9-AC의 시험에서 $p=0.2$ 이고 $(1-0.2)^4=0.044 \approx 0.05$ 이므로 1단계에서 $n_1=4$ 명이 필요할 것이다. 만약 1단계 시험결과 4명중 4명이 반응을 보여 (이 경우 1단계에 의한 반응률의 추정치는 $\hat{p}_1=4/4=1$) 2단계로 진행하는 경우를 고려해 보자. 이때 2단계에 필요한 환자수 n_2 는 다음과 같이 얻어진

다. 2단계 시험이 종결되었을 때 반응을 추정치의 표준오차 공식은 $\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1+n_2}}$ 로 주어진다. 여기서 $\hat{p}$ 대신 $\hat{p}_1=0.29$ 와 $n_1=14$ 를 사용하여 이것이 10%가 되도록 하는 n_2 는 $\sqrt{0.29 \times 0.71 / (14+n_2)} = 0.1$ 를 풀어서 $n_2=7$ 로 얻어진다. 2단계에서 7명을 더 모집하여 총 $N=21$ 명 중 7명이 반응을 보였다면 (이때 2단계에서 반응자수는 3명), 9-AC의 반응률의 추정치는 $7/21=0.33$ 이고 참 반응률이 20%를 넘는가 하는 검정에 대한 p-value는 $p\text{-value} =$

$$\sum_{k=7}^{21} \frac{21!}{k!(21-k)!} 0.2^k (1-0.2)^{21-k} = 0.1085$$

로서 별로 유의하지 않다는 결론을 내리게 된다. 위에서 1단계 후 표준오차 계산에서 환자수 n_1 이 충분히 크다면 $\hat{p}_1$ 이 참반응률 p 에 충분히 가까울 것이라는 가정하에서 $\hat{p}$ 대신 $\hat{p}_1$ 을 사용한 것인데 $n_1=14$ 명으로는 이 가정의 타당성을 믿기 힘들기 때문에, 어느정도 여유있는 n_2 값을 얻기 위하여 Gehan은 $\hat{p}$ 대신 $\hat{p}$ 의 75% 근사 신뢰구간의 상한선 $\hat{p}_1 + z_{0.875} \times$

$\sqrt{\hat{p}_1(1-\hat{p}_1)n_1}$ 을 사용하기를 권한다 ($z_{0.875}=1.15$). 이 경우에는 $\hat{p}$ 대신

$0.29 + 1.15 \sqrt{0.29 \times 0.71 / 14} = 0.43$ 을 이용하여 표준오차 $\sqrt{0.43 \times 0.57 / (14+n_2)} = 0.1$ 을 풀면 2단계에서는 $n_2=11$ 명의 환자를 필요로 하게 된다. 이 숫자는 위에서 $\hat{p}_1$ 을 사용하여 얻어진 $n_1=7$ 명 보다 약간 크다.

이와같이 Gehan의 2단계설계에서는 제 2단계 시험에 몇 명의 환자가 필요한지가 2단계시험 결과에 의하여 결정된다. Gehan의 2단계 설계는 반응률의 추정을 위하여 만들어졌지만 시험이 2차까지 진행되었을 때는 실제 반응률에 대한 검정을 실시하기도 한다. 검정은 이 시험이 마치 단일단계 시험으로 시행된 것처럼 간주하여 전체환자수 N 과 관찰된 반응자수에 의거하여 2.1절에서와 같이 이항분포를 이용한 p-value를 계산한다.

그런데 Gehan의 2단계 설계는 전체환자수 N 을 최소로 유지하려는 시도가 없어서 최적의 설계 방법이라

고 볼 수는 없다. 2상시험의 설계로서 가장 널리 사용되는 것으로 다음에 소개되는 Simon의 최적설계가 있다.

3. Simon's Optimal Two-Stage Design

Simon은 일반적으로 많이 사용되는 2단계의 2상 시험에 소요되는 환자수를 줄이는 최적 설계를 제안하였다³⁾. 여기서는 시험이 초기에 (1단계에서) 중단되기 위한 조건으로 시험 치료제의 반응률이 아주 낮은 경우만 고려한다. 먼저 명백히 낮은 반응률 중 최대값 p_0 와 명백히 높은 반응률 중 최소값 p_a 를 정하고 각각의 반응률에 대한 1종오류 확률 α (주로 0.05)와 검정력 $1-\beta$ (주로 0.8~0.95)를 설정한다. 시험신약의 참반응률 p 에 대해 귀무가설 $H_0: p \leq p_0$ 와 대립가설 $H_a: p \geq p_a$ 에 대한 검정을 위한 2단계 2상 시험의 설계단계에서는 각 단계에서 필요한 환자수 n_1, n_2 와 채택치 a_1, a_2 를 결정하는데, 이 설계를 이용한 2상시험은 다음과 같이 실시된다.

제 1단계 n_1 명의 환자를 모집하여 그 중 반응자수가

i) a_1 보다 작거나 같으면 H_0 를 채택하고 시험 중단

ii) a_1 보다 크면 제 2단계로 진행

제 2단계 a_2 명의 환자를 추가 모집하여 총 $N=n_1+n_2$ 명 중 반응자수가

i) a_2 명보다 작거나 같으면 H_0 를 채택

ii) a_2 명보다 크면 H_a 를 채택

정해진 두 반응률 p_0 와 p_a 에 대해 1종 오류확률이 α 를 넘지않고 검정력이 최소한 $1-\beta$ 를 보장하는 (n_1, n_2, a_1, a_2) 조합들을 후보로 하여 Simon은 다음과 같은 두가지의 다른 방법으로 최선의 조합을 고르는 설계를 제시한다.

최적설계(optimal design) : 각 후보 조합들에 대해 실제로 시험 신약의 반응률이 낮을 때 ($p=p_0$ 일 때) 앞에서 묘사된 2단계 시험에 필요한 환자수의 기댓치 $E(N | p_0)$ 를 계산할 수 있다. Simon의 최적

표 1. Simon의 최적설계와 최소최대설계. 각 (p_0, p_a) 값에 대해 세 개의 열들은 각각 $(\alpha, 1-\beta) = (0.10, 0.90)$, $(0.05, 0.80)$, $(0.05, 0.90)$ 인 경우에 해당되며, $PET(p_0)$ 는 실제 반응률이 p_0 일 때 시험이 조기에 종결될 확률이다.

1(a) $p_a - p_0 = 0.20$ 일 때

Optimal Design						Minimax Design			
p_0	p_a	H_0 채택		$E(N p_0)$	$PET(p_0)$	H_0 채택		$E(N p_0)$	$PET(p_0)$
		$\leq a_1/n_1$	$\leq a_2/N$			$\leq a_1/n_1$	$\leq a_2/N$		
0.05	0.25	0/9	2/24	14/5	0.63	0/13	2/20	16.4	0.51
		0/9	2/17	12/0	0.63	0/12	2/16	13.8	0.54
		0/9	3/30	16/8	0.63	0/15	3/25	20.4	0.46
0.10	0.30	1/12	5/35	19.8	0.65	1/16	4/25	20.4	0.51
		1/10	5/29	15.0	0.74	1/15	5/25	19.5	0.55
		2/18	6/35	22.5	0.71	2/22	6/33	6.33	0.62
0.20	0.40	3/17	10/37	26/0	0.55	3/19	10/36	28.3	0.46
		3/13	12/43	20.6	0.75	4/18	10/33	22.3	0.50
		4/19	15/54	30.4	0.67	5/24	13/45	31.2	0.66
0.30	0.50	7/22	17/46	29.9	0.67	7/28	15/39	35.0	0.36
		5/15	18/46	23.6	0.72	6/19	16/39	25.7	0.48
		8/24	24/63	34.7	0.73	7/24	21/53	36.6	0.56
0.40	0.60	7/18	22/46	30.2	0.56	11/28	20/41	33.8	0.55
		7/16	23/46	24.5	0.72	17/34	20/39	34.4	0.91
		11/25	32/66	36.0	0.73	12/29	27/54	38.1	0.64
0.50	0.70	11/21	26/45	29.0	0.67	11/23	23/39	31.0	0.50
		8/15	26/43	23.5	0.70	12/23	23/37	27.7	0.66
		13/24	36/61	36.0	0.73	14/27	32/53	36.1	0.65
0.60	0.80	6/11	26/38	25.4	0.47	18/27	24/35	28/5	0.82
		7/11	30/43	20.5	0.70	8/13	25/35	20.8	0.65
		12/19	37/53	29.5	0.69	15/26	32/45	35.9	0.48
0.70	0.90	6/9	22/28	17.8	0.54	11/16	20/25	20.1	0.55
		4/6	22/27	14.8	0.58	19/23	21/26	23.2	0.95
		11/15	29/36	21.2	0.70	13/18	26/32	22.7	0.67

1(b) $p_a - p_0 = 0.15$ 일 때

Optimal Design						Minimax Design			
p_0	p_a	H_0 채택		$E(N p_0)$	$PET(p_0)$	H_0 채택		$E(N p_0)$	$PET(p_0)$
		$\leq a_1/n_1$	$\leq a_2/N$			$\leq a_1/n_1$	$\leq a_2/N$		
0.05	0.20	0/12	3/37	23.5	0.54	0/18	3/32	26.4	0.40
		0/10	3/29	17.6	0.60	0/13	3/27	19.8	0.51
		1/21	4/41	26.7	0.72	1/29	1/29	32.9	0.57
0.10	0.25	2/21	7/50	31.2	0.65	2/27	6/40	33.7	0.48
		2/18	7/43	24.7	0.73	2/22	7/40	28.8	0.62
		2/21	10/66	36.8	0.65	3/31	9/55	40.0	0.62
0.20	0.35	5/27	16/63	43.6	0.54	6/33	15/58	45.5	0.50
		5/22	19/72	35.4	0.73	6/31	15/53	40.4	0.57
		8/37	22/83	51.4	0.69	8/42	21/77	58.4	0.53
0.30	0.45	9/30	29/82	51.4	0.59	16/50	25/69	56.0	0.68
		9/27	30/81	41.7	0.73	16/46	25/65	49.6	0.81
		13/40	40/110	60.8	0.70	27/77	33/88	78.5	0.86
0.40	0.55	16/38	40/88	54.5	0.67	18/45	34/73	57.2	0.56
		11/26	40/84	44.9	0.67	28/59	34/70	60.1	0.90
		19/45	49/104	64.0	0.68	24/62	45/94	78.9	0.47
0.50	0.65	18/35	47/84	53.0	0.63	19/40	41/72	58.0	0.44
		15/28	48/83	43.7	0.71	39/66	40/68	66.1	0.95
		22/42	60/105	62.3	0.68	28/57	54/93	75.0	0.50
0.60	0.75	21/34	47/71	47.1	0.65	25/43	43/64	54.4	0.46
		17/27	46/67	39.4	0.69	18/30	43/62	43.8	0.57
		21/34	64/95	55.6	0.65	48/72	54/84	73.2	0.90
0.70	0.85	14/20	45/59	36.2	0.58	15/22	40/52	36.8	0.51
		14/19	46/59	30.3	0.72	16/26	39/49	34.4	0.56
		18/25	61/79	43.3	0.66	33/44	53/68	48.5	0.81
0.80	0.95	5/7	27/31	20.8	0.42	5/7	27/31	20.8	0.42
		7/9	26/29	17.7	0.56	7/9	26/29	17.7	0.56
		16/19	37/42	24.4	0.76	31/35	35/40	35.3	0.94

표 2. Chang등 (1978)의 최적설계. 여기서 $E(N) = \{E(N | p_0) + E(N | p_a)\}/2$ 이다.

총 환자수																
		각단계 환자수					하한선			상한선			환자수 기대값			
p_0	p_a	Size α	Power $1-\beta$	N	n_1	n_2	n_3	a_1	a_2	a_3	r_1	r_2	r_3	$E(N p_0)$	$E(N p_a)$	$E(N)$
05	20	05	70	40	20	20		2	6		4	7		21.2	24.1	22.7
			80	40	20	20		1	5		4	6		25.0	26.9	25.9
			90	40	20	20		0	4		5	5		32.8	32.4	32.6
			70	40	15	15	10	1	4		3	5		17.0	18.5	17.7
			80	40	15	15	10	1	2	4	4	5	5	18.4	23.5	20.9
10	30	05	90	40	15	15	10	0	1	4	4	5	5	26.5	26.3	26.4
			70	30	10	10	10	1	4	4	4	5	5	12.5	15.0	13.8
			80	30	10	10	10	0	3	6	4	6	7	17.5	18.9	18.2
			90	30	10	10	10	—	—	—	1	—	—	—	—	—
			70	35	20	15		4	8		6	9		20.5	22.7	21.6
20	40	05	80	35	20	15		3	7		5	8		21.4	22.0	21.7
			90	35	20	15		—	—	—	—	—	—	—	—	—
			70	35	15	10	10	3	4	6	5	5	7	16.0	18.1	17.1
			80	35	15	10	10	2	4	7	5	6	8	17.1	19.6	18.3
			90	35	15	10	10	1	3	6	6	7	7	21.5	24.7	23.1
30	50	05	70	50	25	25		8	16		10	17		25.7	28.8	27.3
			80	50	25	25		7	16		9	17		26.6	28.0	27.3
			90	50	25	25		5	15		10	16		34.2	34.9	34.5
			70	50	20	15	15	6	10	16	8	12	17	21.0	22.8	21.9
			80	50	20	15	15	5	10	16	8	12	17	22.8	25.0	23.9
30	50	05	90	50	20	15	15	3	8	15	9	12	16	31.7	31.1	31.4
			70	50	25	25		10	21		12	22		26.3	28.3	27.3
			80	50	25	25		9	21		13	22		29.3	34.6	32.0
			90	50	25	25		—	—		—	—		—	—	—
			70	50	20	15	15	8	14	21	11	16	22	21.7	25.7	23.7
30	50	05	80	50	20	15	15	7	12	21	11	16	22	24.7	28.7	26.7
			90	50	20	15	15	—	—		—	—		—	—	—

설계는 이 기대치를 최소로 하는 조합 (n_1, n_2, a_1, a_2) 를 선택하는 것이다.

최소최대설계(minimax design) : Simon은 2단계까지 필요한 전체 환자수 $N = n_1 + n_2$ 를 최대환자수라 부르고 후보 조합들 중 최대환자수를 최소로 하는 조합을 선택하는 것을 최소최대설계라 부른다.

표 1 에는 $(\alpha, 1 - \beta) = (0.10, 0.90), (0.05, 0.80)$ 혹은 $(0.05, 0.90)$ 이고, $p_a - p_0 = 0.20$ 혹은 0.15 인 경우에 대해 Simon(1989)의 논문에서 발췌된 최적설계와 최소최대설계들을 예시하였다. 9-AC의 시험을 Simon의 방법으로 설계해보자. 9-AC의 효과여부를 반응률 20%를 기준으로 하나, 실제 반응률이 $p_0 = 10\%$ 이하면 $1 - \alpha = 0.95$ ($\alpha = 0.05$)의 높은 확률로 9-AC를 버리고 싶고, 실제 반응률이 $p_a = 30\%$ 이상이면 $1 - \beta = 0.8$ 의 높은 확률로 9-AC를 채택하고 싶다. 이 경우 표 1(a)로 부터 최적설계는 $n_1 = 10, N = n_1 + n_2 = 29$ (즉, $n_2 = 19$), $a_1 = 1, a_2 = 5$ 로 주어지며, 9-AC의 실제반응률이 $p_0 = 0.1$ 이라면 소요될 환자수의 기대값은 $E(N | p_0) = 15$ 명이고 이 시험이 1단계에서 종결될 확률은 $PET(p_0) = 0.74$ 이다. 아울러 최소최대설계는 $n_1 = 15, N = 25$ (즉, $n_2 = 10$), $a_1 = 1, a_2 = 5$ 이고, 9-AC의 실제반응률이 $p_0 = 0.1$ 이라면 이 시험에 소요될 환자수의 기대값은 19.5명이고 이 시험이 1단계로 종결될 확률은 0.55이다.

일반적으로 최적설계와 최소최대설계에 필요한 환자수는 일치하지 않는다. 때때로 최적설계보다는 최소최대설계가 선호될 수도 있다. 두 설계에 의한 소요 환자수의 기대값의 차이는 작고 모집 가능한 환자수가 제한된 경우에 특히 그러하다. 구체적인 예로써, $p_0 = 0.1, p_a = 0.3, \alpha = 0.1, \beta = 0.9$ 일 때를 고려해보자. 이 경우 최적설계는 $(n_1, n_2, a_1, a_2) = (12, 23, 1, 5)$ 이고 소요 환자수의 기대값은 $E(N | p_0) = 19.8$ 이며 최대환자수는 $N = 25$ 인 반면, 최소최대설계는 $(16, 9, 1, 14)$ 이며 소요 환자수의 기대값은 $E(N | p_0) = 20.4$ 이고 최대환자수는 $N = 25$ 이다. 만약 예상되는 모집 환자수가 1년에 10명 정도라면 시험이 2

단계까지 진행될 경우에 최적설계를 사용한 2단계 시험은 최소최대설계를 사용한 2단계 시험보다 1년이 더 소요될 것이다. 실제로 Simon의 2단계 시험 설계를 사용할 때는 두 설계 방법에 의한 $E(N | p_0)$ 와 N 을 비교하여 보고 모집 가능한 환자수 등을 고려하여 상황에 맞는 설계 방법을 선택하여야 할 것이다.

4. Chang, Therneau, Wieand & Cha Optimal Design

앞에서 논의된 Gehan의 2단계 설계나 Simon의 최적 설계법은 1단계에서 시험 신약에 대한 충분한 효과를 입증할 수 있을 정도로 아주 높은 반응률을 관찰한 경우에도 2단계로 넘어가서 시험을 계속 진행해야 한다. 현재 시험되어야 할 신약이 많이 기다리고 있다면 이 문제점은 제한된 수의 환자들을 많은 임상 시험에 경제적으로 사용하지 못하게 되는 결과를 초래할 것이다. 우리는 중간단계에서 신약의 반응률이 매우 낮을 때 뿐만 아니라 매우 높다고 판정될 경우에도 2상시험을 조기 종결하고 3상시험으로 진행하는 것을 고려해 볼 수 있을 것이다. 이러한 경우 다단계 2상 시험은 다음과 같이 진행된다. 시험에 필요한 (혹은 모집가능한) 환자수 N , 검정 단계 횟수 K , 그리고 각 단계에서 투입될 환자수 $n_1, \dots, n_K$ 가 정해지면 각 단계에서 $H_0: p \leq p_0$ 와 $H_a: p \geq p_a$ 를 검정하기 위한 H_0 의 채택치 $(a_1, \dots, a_K)$ 와 기각치 $(r_1, \dots, r_K)$ 가 적절한 방법으로 정해진다. 여기서 p_0 는 명백히 낮다고 생각되는 반응률의 최대치이고 p_a 는 명백히 높다고 생각되는 반응률의 최소치로서 시험 설계 당시에 설정된다. 마지막 검정 단계인 K 번째 단계에서는 H_0 를 채택하든지 기각하든지 둘 중 한 선택을 해야 하므로 $r_K = a_K + 1$ 이다. 그러면 $k(k = 1, \dots, K)$ 번째 단계에서 현재까지 반응을 보인 총 환자수 S_k 에 대해 $S_k \leq a_k$ 이면 H_0 를 채택하고 시험을 중단하고, $S_k \geq r_k$ 이면 H_a 를 채택하고 시험을 중단하며, $a_k < S_k < r_k$ 이면 결정을 보류하고 $(k + 1)$ 번째 단계로 진행한다.

시험 설계시에는 귀무가설 하에서의 치료율 p_0 와 대립가설 하에서의 치료율 $p_a(p_0 < p_a)$ 를 정하고 동시에 최대환자수 N , 사용할 단계수 K (주로 2~3단계), 각 단계에 모집할 환자수 $n_1, \dots, n_K (N = n_1 + \dots + n_K)$, 허용된 1종오류 확률 α , 최소한의 검정력 $1 - \beta$ 등을 상황에 맞게 정한다. 그러면 Chang 등의 최적설계는 다음과 같은 방법으로 $(a_1, \dots, a_K)$ 와 $(r_1, \dots, r_K)$ 를 정한다 ($r_K = a_K + 1$). 주어진 1종오류확률과 검정력에 의한 제한 조건을 만족하는 모든 $\{(n_1, \dots, n_K), (a_1, \dots, a_K), (r_1, \dots, r_K)\}$ 들의 조합에 대해서 H_0 하에서의 필요한 환자수의 기대치 $E(N | p_0)$ 와 H_a 하에서의 필요한 환자수의 기대치 $E(N | p_a)$ 를 계산할 수 있다. 제한조건을 만족하는 후보 조합들이 존재하기 위해 N 은 적절히 큰 수이어야 한다. Chang등의 최적설계는 $E(N | p_0) + E(N | p_a)$ 를 최소화하는 채택치와 기각치의 조합을 선택하는 것이다. 이 최소화 과정에서 모든 가능한 조합들에 대해 환자수의 기대치를 계산하고 이들을 비교하는 것은 엄청난 계산 시간을 요구하는데, Chang 등은 우도비 검정법을 적절히 이용하여 최적설계와 거리가 먼 후보 조합들을 95% 이상 사전에 제외시킴으로써 최적 조합을 찾기위한 검색과정을 단순화 하고 있다.

표 2에는 Chang등의 논문에서 발췌한 최적조건들이 예시되었다⁴⁾. 9-AC의 시험을 위해 $N=35$ 명 정도의 환자를 모집할 수 있다고 하자. 그리고 앞에서와 같이 $p_0=0.1$, $p_a=0.3$ 으로 두고 $\alpha=0.05$, $\beta=0.8$ 로 설정하자. 이 시험을 $K=2$ 단계로 실시하고자 하는 경우 최적설계는 $n_1=20$, $n_2=15$, $a_1=3$, $a_2=7$, $r_1=5$, $r_2=8$ 로 주어진다. 이 시험에 소요될 환자수의 기대값은, 9-AC의 실제 반응률이 $p_0=0.1$ 일 때는 21.4명이고 $p_a=0.3$ 일 때는 22.0명이다. 이 시험을 $k=3$ 단계로 실시하는 경우에는 $n_1=n_2=n_3=10$, $a_1=0$, $a_2=3$, $a_3=6$, $r_1=4$, $r_2=6$, $r_3=7$ 의 설계나 $n_1=15$, $n_2=n_3=10$, $a_1=3$, $a_2=4$, $a_3=6$, $r_1=5$, $r_2=5$, $r_3=7$ 의 설계 중 하나를 선택할 수 있다. 그런데 최소소요환자수에 있어서는 전자가 5명 작고(30명 : 35명), 소요환자수의 기대값은 실제 반응률이 0.

1일 때는 후자가 $E(N | p_0)$ 이 17.5명 : 17.1명), 그리고 실제반응률이 0.3일 때는 전자가 $E(N | p_0)$ 이 18.9명 : 19.6명) 약간 작다.

지금까지 소개된 설계 방법외에도 Bayesian이론을 이용한 Herson의 방법, 통계적 결정론을 이용한 Sylvester & Staquet의 2단계 설계, 단일 의료기관에서 수행되는 2상시험에 사용하기 적절한 Lee, Staquet, Simon, Catane & Muggia의 제한된 환자 모집 설계 (limited accrual design)방법 등이 있다^{5~7)}.

III. 치료효능으로서 반응률의 신뢰구간

2상시험이 끝나면 우리는 연구대상 신약에 대한 치료효능으로서 반응률을 추정하고 그것의 신뢰구간을 구하게 된다. 반응률의 추정치는 $\hat{p} = (\text{반응을 보인 환자수}) / (\text{전체환자수})$ 로 얻어진다. 반응률의 신뢰구간을 구하는 공식으로 누구나 쉽게

$\hat{p} \pm Z_{1-\alpha/2} \sqrt{\hat{p}(1-\hat{p})/N}$ (여기서 $Z_{1-\alpha/2}$ 는 표준정규 분포의 $100\alpha/2$ 백분위수이며 $\alpha=0.05$ 일때 $Z_{1-\alpha/2}=1.96$)을 떠올리게 될 것이다. 그런데 이 공식은 이항분포를 하는 반응률을 N 이 충분히 클때 근사적으로 정규분포 즉 따르는 것으로 간주하는 것인데 이항분포를 근사한 것인데 일반적으로 2상시험에 선정되는 환자수는 이 근사가 적절할 만큼 충분히 크지 않다. 더 우기 다단계로 시행되는 2상시험은 극단적인(너무 작거나 너무 큰) 반응자수를 관찰할 때에 조기 중단되므로 단일단계 시행시 사용되는 신뢰구간은 바이어스(bias)를 갖게된다. 따라서 2상시험결과로 관찰된 반응률의 구간추정에는 시험이 다단계로 시행되었음을 고려하여 정규근사가 아닌 이항분포 그 자체를 사용한 정확한 신뢰구간을 구하는 방법을 써야 할 것이다. 이러한 일반적인 2상시험의 제한점을 고려하여 반응률의 신뢰구간 계산에 가장 널리 사용되고 있는 Jennison & Turnbull의 방법을 소개한다⁸⁾.

다단계 2상시험의 설계는 최대 단계수 K , 각 단계에 모집할 환자수 $(n_1, \dots, n_K)$, 각 단계의 채택치 $(a_1, \dots, a_K)$ 와 기각치 $(r_1, \dots, r_K)$ 로 정의된다. 채

택치만 사용하는 Simon의 방법은 $k=1, \dots, K-1$ 에 대해 $r_k = n_k + 1$ 그리고 $r_K = a_K + 1$ 을 사용한다고 볼수 있다. 이러한 다단계 2상시험이 끝나면 우리는 단계수 M 과 그때까지 관찰된 총 반응자수 S 를 관찰한다. Jennison과 Turnbull은 2상시험 결과 관찰될 모든 (M, S) 의 조합을 실제 반응률이 낮을 때 관찰될 가능성이 큰 조합부터 실제 반응률이 높을 때 관찰될 가능성이 큰 조합까지 다음과 같이 일렬로 나열한다.

$(M=1, S=0), (M=1, S=1), \dots, (M=1, S=a_1),$

$(M=2, S=a_1+1), (M=2, S=a_1+2), \dots, (M=2, S=a_2),$

$\vdots$

$(M=K, S=a_{K-1}+1), (M=K, S=a_{K-1}+2), \dots, (M=K, S=r_K),$

$(M=K, S=r_K), (M=K, S=r_K+1), \dots, (M=K, S=r_{K-1}-1+n_K),$

$(M=K-1, S=r_{K-1}), (M=K-1, S=r_{K-1}+1), \dots, (M=1, S=r_{K-2}-1+n_{K-1}),$

$\vdots$

$(M=1, S=r_1), (M=1, S=r_1+1), \dots, (M=1, S=n_1)$

임의의 반응률 $p(0 < p < 1)$ 에 대해 $(M=m, S=s)$ 의 확률을 계산할 수 있는데 임상시험 결과 관찰된 (M, S) 의 값을 (m^*, s^*) 라 하면 반응률의 $100(1-\alpha)\%$ 신뢰구간 (p_L, p_U) 는

$$\Pr\{(M, S) \leq (m^*, s^*) \mid p = p_U\} = \alpha/2$$

$$\Pr\{(M, S) \leq (m^*, s^*) \mid p = p_L\} = \alpha/2$$

의 관계에 의해 얻어진다. 여기서 (M, S) 의 실현치들에 대한 크기의 비교는 위에 나열된 순서에 의해 결정된다.

시험결과 H_0 가 채택되는 경우들에 대해 반응률의 90% 혹은 95%의 신뢰구간을 계산하여 시험약정서에 첨부하는 것이 좋다. $H_0: p \leq 0.1$ (즉 $p_0=0.1$)과 $H_a: p \geq 0.3$ (즉 $p_a=0.3$)의 가설을 2단계 ($K=2$)를 사용하여 1단계에서 $n_1=15$ 명, 2단계에서

$n_2=15$ 명의 환자를 모집하기로 하고 각 단계 채택치가 $a_1=1, a_2=5$ 인 2상시험에서 H_0 가 채택되는 경우들에 대해 실제반응률 p 에 대한 90% 신뢰구간은 다음과 같이 주어진다.

시험중단단계	총반응자수	Jennison-Turnbull 1의 90% 신뢰구간
M	S	
1	0	(0.000, 0.201)
1	1	(0.003, 0.311)
2	2	(0.024, 0.312)
2	3	(0.036, 0.312)
2	4	(0.052, 0.316)
2	5	(0.071, 0.353)

2상시험의 결과 시험 신약의 반응률에 대한 신뢰구간을 구하는 방법으로는 Jennison & Turnbull 방법 외에도 Duffy & Santner의 방법이 있다⁹⁾.

IV. 안전성(독성률) 평가를 위한 2상시험

지금까지는 치료제의 효능을 평가하는 경우를 기술하였다. 그런데 우리는 종종 2상시험을 통하여 치료제의 안전성(독성률)을 평가하기도 한다. 이때 사용되는 기본적인 통계이론은 반응률을 평가하는 경우와 동일하나, 효능은 높을수록 좋은 반면 독성률은 낮을수록 좋으므로 앞에서 언급된 이론을 적용하는 데 있어서 약간의 수정이 필요하다.

명백히 높다고 생각되는 독성률 중 최소값을 q_0 , 명백히 낮다고 생각되는 독성률 중 최대값을 q_a 라고 하자. 그러면 시험치료제의 참독성률 q 에 대해 귀무가설은 $H_0: q \leq q_0$, 대립가설은 $H_a: q \leq q_a$ 로 주어진다. 이 두 가설에서 부등호의 방향은 반응률 평가의 경우와 반대이므로 앞에서 언급된 방법들을 직접 사용할 수는 없다. 여기서 $p=1-q$ (즉, 독성을 안보일 환자들의 비율), $p_0=1-q_0, p_a=1-q_a$ 라 두면 귀무가설과 대립가설은 각각 $H_0: p \leq p_0, H_a: p \geq p_a$ 로 주어지므로 두 가설의 부등호방향은 반응률 평가시와

같아졌고, 따라서 앞에서 논의된 다단계설계방법을 사용할 수 있다.

예를 들어, $q_0=0.4$, $q_a=0.2$ 일 때 $\alpha=0.05$, $1-\beta=0.8$ 인 Simon의 최적 2단계시험을 설계해 보자. 이 때 $p_0=1-0.4=0.6$, $p_a=1-0.2=0.8$ 이므로 표 1(a)로부터 ($n_1=11$, $n_2=N-n_1=43-11=32$)이고 ($a_1=7$, $a_2=30$)이므로 이 시험은

제 1단계 $n_1=11$ 명의 환자들 중

i) 7명이하만 독성을 안보이면 H_0 를 채택하고 시험중단

ii) 8명이상이 독성을 안보이면 2단계로 진행

제 2단계 $n_2=32$ 명을 더 모집해서 총 $N=43$ 명의 환자들 중

i) 30명이하만 독성을 안보이면 H_0 를 채택

ii) 31명이상이 독성을 안보이면 H_a 를 채택

으로 진행된다. 이 과정을 독성을 보인 환자수에 대해 기술하면

제 1단계 $n_1=11$ 명의 환자들 중

i) 4명 또는 그 이상이 독성을 보이면 H_0 를 채택하고 시험중단

ii) 3명이하가 독성을 보이면 2단계로 진행

제 2단계 $n_2=32$ 명을 더 모집해서 총 $N=43$ 명의 환자들 중

i) 13명 또는 그 이상이 독성을 보이면 H_0 를 채택

ii) 12명이하가 독성을 보이면 H_a 를 채택하게 된다.

독성률의 신뢰구간 추정은 3장에 소개된 반응률의 신뢰구간 추정방법을 수정없이 그대로 이용하면 된다.

V. 결 론

제 2상 임상시험에서 사용되는 통계적 방법은 임상 시험을 설계할 때 필요한 환자수와 조기에 시험을 중단하기 위한 규칙을 정하고 시험이 종결되었을 때 관찰된 반응자수를 (혹은 독성을 보인 환자수)를 바탕

으로 실제반응률 (혹은 독성률)에 대한 추정을 하는 것이다. 일반적으로 제한된 환자수를 사용하여 다단계로 시행되는 2상시험의 특성상 지금까지 다른 분야에서 사용되지 않는 특별한 통계적 방법들이 사용되고 있고, 보다 효율적인 방법들이 계속 연구 개발되고 있다. 본 논문에서는 2상시험의 주 관심을 신약의 반응률인 경우에 국한하였지만 때로는 독성률(허용정도 이상의 독성을 보이는 환자수), 반응을 유지하는 기간(반응을 보인 시점부터 재발할 때까지 경과한 시간) 등을 사용하여 신약의 효과를 검정할 수도 있다. 여러 가지 통계적 방법들 중 어떠한 방법을 사용할 것인가는 시행하려는 2상시험의 주관심이나 그 주변 상황(연구비, 모집가능한 환자수, 연구기간 등)에 따라 적절히 정해져야 한다. 또한, 어떠한 2상시험을 계획할 때 지금까지 소개된 일반적 통계방법들이 적절치 않다면 기존의 방법을 적절히 변형하든지 주어진 상황에서 최적의 방법을 새로이 만들어 사용하여야 할 것이다.

VI. 요 약

임상시험이란 특정 치료법에 관한 정확히 정의된 의문사항들을 답하기 위해 잘 디자인된 전향적 의학 연구방법이라 정의될 수 있다. 전향적인 임상시험으로는 1상, 2상, 3상, 및 4상시험 등이 있으며, 보다 엄격한 정의에 의해 3상시험만을 임상시험이라고 부르기도 한다. 2상시험에서는 시험치료제의 치료효능과 안전성을 평가하여 시험치료약에 대해 본격적인 3상시험을 계속할 가치가 있는 것인가를 결정한다. 일반적으로 다단계 설계로 시행되는 2상시험의 설계 단계에서는 적절한 최적조건하에서 필요한 환자수를 결정한다. 2상시험이 끝나면 그 시험이 다단계로 행해졌다는 사실을 고려하여 신약 반응률(혹은 독성률)에 대한 정확한 신뢰구간을 구한다. 본 논문에서는 항암제를 대상으로 하는 2상시험에 필요한 다양한 통계적 방법들이 요약 소개된다.

참 고 문 헌

1. Rosner B. *Fundamentals of Biostatistics*. 3rd ed. PWS-KENT, 1990
2. Gehan EA. *The determination of the number of patients required in a follow-up trial of new chemotherapeutic agent*. *Journal of Chronic Diseases* 1961;13:346-353
3. Simon R. *Optimal two-stage design for phase II clinical trials*. *Controlled Clinical Trials* 1989;10:1-10
4. Chang MN, Therneau TM, Wieand HS, Cha SS. *Designs for group sequential phase II clinical trials*. *Biometrics* 1987;43:863-874
5. Herson J. *Predictive probability early termination plans for phase II clinical trials*. *Biometrics* 1979;35:775-783
6. Sylvester RJ, Staquet MM. *Design of phase II clinical trials in cancer using decision theory*. *Cancer Treatment Report* 1980;64:519-524
7. Lee YJ, Staquet MM, Simon R, Catane R, Muggia FD. *Two stage plans for patient accrual in phase II cancer clinical trials*. *Cancer treatment Report*. 1979;63:1721-1726
8. Jennison C, Turnbull BW. *Confidence intervals for a binomial parameter following a multi-stage test with application to MIL-STD 105D and medical trials*. *Technometrics* 1983;25:49-58
9. Duffy DE, Santner TJ. *Confidence intervals for a binomial parameter based on multi-stage tests*. *Biometrics* 1987;43:81-93