제 3상 임상시험에서 표본수 결정

연세의대 남정모

표본수를 결정하는 방법에는 크게 sequential design과 fixed sample size design이 있다. Fixed sample size design은 연구를 시행하기 전에 표본수를 합리적으로 결정하고 정해진 표본내에서 연구를 진행하는 방법이며, sequential design은 연구를 진행하면서 결과의 차이가 있는가 또는 없는가에 대해 미리 정해진 한계영역을 기준으로 계속적으로 연구대상을 추출하여 연구를 진행하는 방법이다. 여기서는 많이 사용되는 fixed sample size design에 대해서만 생각하기로 한다.

1. 표본수 결정요인

표본수를 결정하기 위해서는 연구시작전에 다음의 내용들을 검토하여야 한다.

1) 임상적으로 유의한 최소한의 차이(minimum treatment difference) : 연구자가 알아내고자 하는 실험군과 대조군의 결과에 대한 최소한의 차이를 말한다.

2) 제 1종의 오류(α)와 제 2종의 오류(β)

3) 통계적 적용방법 : 최종적으로 측정된 자료의 속성과 연구가설에 따라 적용되는 통계적 분석방법, 그리고 그 검정이 양측검정이냐 또는 단측검정이냐에 따라서도 계산하는 방법에 차이가 있다.

4) 할당비(allocation ratio) : 실험군과 대조군의 비가 동일한지의 여부에 따라 단형할당과 이형할당으로 구별할 수 있으며 할당비에 따라 표본수의 결정이 달라진다.

5) 연구를 진행하다 보면 여러가지 이유로 인해 피실험자가 중도탈락하는 경우가 발생할 수 있다. 따라서 이러한 영향을 보정하기 위해 중도탈락률(d)을 예상하여야 하며 최종적인 표본수는 산출된 표본수에 보정지수를 곱하여 계산되며 보정지수는 다음과 같다.

$$\text{보정지수} = \frac{1}{(1-d)}$$

2. 표본수 결정공식 및 예제

앞서 언급하여 듯이 표본수를 결정하는 방법은 최종적으로 적용되는 분석방법에

따라 달라지므로 여기서는 간단한 몇가지 통계방법에 국한하여 설명하기로 한다.

가. 최종 결과변수가 이분형인 경우

치료효과의 유무, 항체생성 여부와 같이 이분형으로 측정된 경우에 적용되는 방법이 여러가지 있으나 여기서는 χ^2-검정을 이용한 표본수 결정 방법에 대해서 알아본다.

1) χ^2-방법을 이용한 단형할당인 경우

$$n_c = \frac{(Z_{1-\alpha/2}\sqrt{2\overline{PQ}} + Z_{1-\beta}\sqrt{P_cQ_c + P_tQ_t})^2}{(P_t - P_c)^2}$$

여기서, P_t 와 P_c는 각각 실험군과 대조군의 치료에 대한 성공률을 나타내며, $\overline{P}$는 두 집단의 성공률의 평균, $\overline{Q} = 1 - \overline{P}$ 이다. Z_α는 표준정규분포에서 α 퍼센타일에 해당되는 값이며 많이 사용되는 값은 다음과 같다.

		양측검정	단측검정
유의수준 (α)	0.01	2.576	2.326
	0.05	1.960	1.645
	0.10	1.645	1.282
검정력 ($1-\beta$)	0.80	0.840	-
	0.90	1.282	
	0.95	1.645	
	0.99	2.326	

2) 가상적 연구설계

- 실험군과 대조군 : 2 집단
- 결과변수 : 사망여부
- 대립가설의 형태 : 단측검정
- 추적기간 : 5년
- 유의한 차의 정도 : $P_c = 0.40$ (5년 사망률), $\Delta_A = P_c - P_t = 0.10$
- $\alpha = 0.05$ $\beta = 0.05$
- 할당비 : 1:1
- 중도탈락률 : 20%

$$n_c = \frac{(1.645\sqrt{2(0.35)(0.65)} + 1.645\sqrt{(0.40)(0.60) + (0.30)(0.70)})^2}{(0.10)^2} = 490$$

따라서 대조군과 실험군 각각 490×(1/0.8) = 613명이 필요하다.

나. 최종 결과변수가 연속형인 경우

Baseline에서 측정한 값과 intervention 후 측정한 값의 변화가 두 군간(실험군과 대조군)에 차이가 있는가에 대한 경우 표본수 결정공식은 다음과 같다.

1) 정규분포 근사를 이용한 단형할당인 경우

$$n_c = \frac{2(Z_{1-\alpha/2}+Z_{1-\beta})^2\sigma_d^2}{(\mu_{dc}-\mu_{dt})^2}, \quad \sigma_d^2=2(1-\rho)\sigma^2$$

여기서, $\mu_{dc}=\mu_{1c}-\mu_{0c}$: 대조군의 baseline과 intervention 후 측정변수의 평균변화

$\mu_{dt}=\mu_{1t}-\mu_{0t}$: 실험군의 baseline과 intervention 후 측정변수의 평균변화

σ^2 : 측정변수의 분산(baseline)

ρ : baseline 값과 intervention 후 측정한 값과의 상관계수

2) 가상적 연구설계

- 실험군과 대조군 : 2 집단
- 결과변수 : 치료후 3년간 혈압의 변화
- 대립가설의 형태 : 양측검정
- 추적기간 : 3년
- 유의한 차의 정도 : 4mmHg
- $\sigma^2=100\text{mmHg}$, $\rho=0.3$
- $\alpha=0.05$ $\beta=0.05$
- 할당비 : 1:1
- 중도탈락률 : 30%

$$n_c = \frac{2(1.96+1.645)^2(140\text{mmHg})^2}{(4\text{mmHg})^2} = 228$$

따라서 대조군과 실험군 각각 228×(1/0.7) = 326명이 필요하다

3. 표본수의 현실적인 평가 및 표본수가 작은 경우의 문제

과학적인 근거(통계적인 방법)에 의해 표본수를 산출한 후 그 표본수가 현실에 비해 너무 크다면 어떻게 할 것인가? 만약 추정된 표본수를 무시하고 연구를 진행하면 어떤 문제가 야기될 수 있는가? 일반적인 내용을 정리하면 다음과 같다. 추정된 표본수가 큰 경우에는 먼저 참여할 것으로 예상되는 연구자들이 단위기간 동안 얼마나 많은 대상환자를 모을수 있는가(accrual rate)를 추정하고 이를 통해 연구대상을 수집하는 기간을 추정한다. 만약 추정된 기간이 너무 길다면 다음의 내용들을 생각해 볼 수 있다 (Pocock 1983).

첫째, accrual rate를 증가시킨다. 이 방법은 더 많은 연구자 또는 기관를 참여(multi-centre)시키거나 또는 환자 선정기준을 검토하여 임상적인 의미를 잃지 않는 범위내에서 선정기준을 완화하는 방법 등을 생각할 수 있다.

둘째, 표본수 결정에 대한 통계적인 기준을 검토하여 연구의 의미를 잃지 않는 범위내에서 minimum treatment difference를 증가시키거나, 제 1종 또는 제 2종의 오류를 증가시킨다. 그러나 제 1종의 오류는 0.1, 제 2종의 오류는 0.2를 넘지 않는 것이 일반적이다.

세째, 위의 두가지 방법이 불가능한 경우에는 연구를 중단하는 것이 최선이다.

추정된 표본수보다 작은 수를 가지고 연구를 진행하게 되는 경우의 문제는 제 2종의 오류가 커지는데 있으며 이는 실제로 실험약이 대조약에 비해 효과가 있지만 효과가 없는 것으로 잘못된 결론을 내릴 확률이 커진다는 것을 의미한다. 참고적으로 Freiman 등(1978)은 효과가 없다고 판정된 71개의 randomized control trial 논문을 대상으로 실험군의 실패률(주로, mortality rate, complication rate, no improvement rate)이 대조군에 비해 25%, 50% 정도 감소한 것으로 가정하였을 때에 연구된 표본수를 가지고 제 2종의 오류를 계산하였다. 만약 실험군의 실패률이 대조군에 비해 25% 감소를 가정하면, 71편의 논문들 중 과반수 이상은 70%가 넘는 제 2종의 오류가 있었고, 50%라고 가정하였 때에는 과반수 이상의 논문이 40%가 넘는 제 2종의 오류가 있었다(표 1).

표 1. 실험군과 대조군의 가정된 실패률의 차이에 따른 제 2종의 오류

제2종의 오류(%)	경우 1 (실험군이 25% 적은 경우)			경우 2 (실험군이 50% 적은 경우)		
	빈도	누적빈도	%	빈도	누적빈도	%
0-10	4	4	5.63	21	21	29.58
11-20	1	5	7.04	1	22	30.99
21-30	2	7	9.89	4	26	36.62
31-40	2	9	12.68	9	35	49.30
41-50	5	14	19.72	5	40	56.34
51-60	7	21	29.58	4	44	61.97
61-70	2	23	32.39	8	52	73.24
71-80	16	39	54.93	9	61	85.92
81-90	25	64	90.14	9	70	98.59
91-100	7	71	100.0	1	71	100.00

자료 : Freiman 등(1978)의 논문

따라서 71개의 연구에서 효과가 없다고 판정한 연구들의 상당수는 표본수를 작게 산출하여 생긴 결과(또는 제 2종의 오류를 크게 설정)임을 예상할 수 있다.

4. 우리나라 제 3상 임상시험에서 표본수 결정

최근(95년 10월 1일) 시행된 "의약품임상시험관리기준"에서 임상성적을 기술할 때 임상례수(계획된 수, 실제대상수, 완료된 수, 중도탈락자 수 및 이유)를 자세히 기술하도록 되어 있고 각 약효군별 임상시험 평가지침서에는 통계적인 방법에 의해 표본수를 결정하도록 되어 있다.

실제 우리나라 임상시험에서 표본수 결정은 어떻게 이루어졌는지 알아보기 위해 비록 이 안이 시행되기 이전이지만 지난 6개월 동안 ('95. 4 - '95. 9) 중앙약사 심의위원회에 제출된 제 3상 임상시험계획서의 표본수 결정에 대한 내용을 일부 검토하여 보았다. 그 결과 대부분의 계획서에는 객관성이 결여된 방법으로 표본수를 선정하였으며 통계적인 방법에 의해 표본수가 옳바르게 선정된 경우는 23건 중 3건에 지나지 않았고 나머지 대부분은 현실적인 여건을 지나치게 감안하여 제대로 계획하지 못하였다.

예를들면 정부의 규정에는 통계적으로 유의한 차이를 검증할 수 있는 표본수를 선정하도록 되어 있으나 30례 이상을 피험자수로 선정하여야 한다는 조항때문에 많은 계획서에서 30례만을 계획하였다. 또한 대조군의 선정을 과거대조군(historical control)으로 한 경우 실험군의 유효율을 자연치유율과 비교하므로서 표본수를 지나치게 과소추정한 경우가 있었으며, 희귀한 질병의 연구에서는 충분한 수의 피험자 선택이 불가능하므로 표본수를 임의로 작게 선정한 경우 등도 있었다(표 2).

한편 1990-1994년 동안 국내에서 시행된 임상시험 실시 실태를 분석한 "의약품 임상시험 관리기준 도입방안 연구"의 결과를 보면 표본수를 통계적인 것에 근거를 두고 산출한 연구의 비율은 허가용 임상시험인 경우 19.7%, 연구용 임상시험인 경우 3.4%로서 대부분은 정부 기준을 따랐다고 보고하였다(한국보건사회 연구원, 1994).

이러한 이유는 무엇보다도 현재 우리나라 현실에서 대상환자를 충분하게 확보한다는 것이 어려우며 또한 선진국에 비교할 때 연구비가 비교할 수 없을 정도로 적다는 것 등을 문제점으로 생각할 수 있다. 그러나 한편으로는 객관적인 연구 결과를 도출하는데 있어 임상시험 담당자나 의뢰자의 소극적인 태도 등도 크게 작용하는 것으로 보인다.

표 2. 우리나라 일부 임상시험계획서에 작성된 표본수 결정 방법에 대한 내용

표본수 결정방법의 내용	건수
표본수 결정에 대한 근거없이 보건복지부의 규정에 따라 결정	7
통계적인 방법에 의한 표본수가 많아 보건복지부의 규정에 따라 결정	4
과거대조군 선정시 자연치유율과 비교하여 표본수를 결정	4
표본수를 임의로 결정	1
표본수를 결정하기 위한 수식이나 계산의 착오로 표본수를 잘못 결정	4
표본수 결정에 문제가 없음	3

앞에서 언급하였지만 표본수가 작은 임상시험은 제 2종의 오류를 크게 하고 따라서 대조군에 비해 유효성에 차이가 없다는 결론을 얻기가 쉬어진다. 더욱 놀라운 것은 작은 수의 표본을 가지고 등가시험(equivalence trial)을 한다는데 있다. 작은 수의 표본으로는 통계학적으로 그 차이를 입증하기가 어렵기 때문에 실험군과 대조군간에 차이가 없다는 결론을 내리게 된다. 이러한 것들은 자칫 연구 전체를 왜곡할 수 있으며 표본수를 결정하는 것이 얼마나 중요한 것인지를 생각케 한다.

참고문헌

고웅린. 신약평가를 위한 임상시험과 자료분석. 서울, 신광출판사, 1994

한국보건사회연구원. 의약품 임상시험 관리기준 도입방안 연구. 1994

한국제약협회. 의약품 임상시험관리 업무지침서. 1995

Buyse ME, Staquet MJ, Sylvester RJ. Cancer Clinical Trials - Method and Pratice. Oxford, Oxford University Press, 1988

Fleiss JL. Statistical Methods for Rates and Proportions. New York, John Wiley & Sons, 1981

Freiman JA et al. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. N Engl J Med 1978; 299: 690-694

Meinert CL, Tonascia S. Clinical Trials - Design, Conduct, and Analysis. New York, Oxford University Press, 1986

Pocock SJ. Clinical Trials. New York, John Wiley & Sons, 1983