한국통계학회논문집 제2권 1호
1995년 4월 pp. 229-247

임상시험연구의 통계적 고찰

박 미 라, 이 재 원[1)]

요 약

신뢰할만한 임상시험연구결과를 얻기 위해서는 통계학자가 임상시험연구의 준비단계에서부터 실험계획 및 연구의 운용, 분석은 물론 결과의 보고에 이르기까지 총체적으로 관여해야 한다. 의학전문가들과의 원활한 공동연구를 위해서는 통계학자들이 의학분야에서의 용어나 관례, 임상시험이 이루어지는 절차등에 대한 명확한 이해가 필요하게 된다. 이 논문에서는 임상시험에서 진행되는 제반문제에 대해 통계학자가 숙지해야 할 사항들에 대해 논의하였다.

1. 서론

국내에서 시행되는 임상시험은 대부분 통계학자들의 도움없이 행해지고 있다. 의학연구자들이 어느 정도의 통계지식을 가지고 있다고 하더라도 의학연구의 결과에 대한 파급효과를 고려할 때 이러한 현상은 상당히 위험한 것이라고 할 수 있다. 의학연구자들의 상당수가 의학연구에서 통계학자가 관여할 부분은 이미 얻어진 자료의 통계분석에 국한되는 것으로 알고 있는 실정이다. 그러나 신뢰할 만한 임상연구결과를 얻기 위해서는 통계학자가 임상시험연구의 준비단계에서부터 실험계획 및 연구의 운용, 분석은 물론 결과의 보고에 이르기까지 총체적으로 관여해야 한다. 의학분야와 통계분야의 조우가 어려운 이유중의 하나는 통계학자들이 의학분야에서의 용어나 관례를 비롯하여 임상시험이 이루어지는 절차에 대한 올바른 이해가 부족하기 때문에 원활한 의사소통과 연구의 진전이 어려운 까닭도 있으리라 생각된다.

이 논문에서는 임상시험에서 진행되는 제반문제에 대해 통계학자가 숙지해야 할 사항들을 전반적으로 설명하고 아울러 통계적 사고방식에 근거하여 임상연구의 결과를 작성하는 기본 원칙에 대하여 논의해보겠다. 제 2 절에서는 의학연구의 유형과 임상시험연구의 단계 및 실험계획(experimental design)에 관한 통계적 방법을 설명하고, 제 3 절에서는 임상시험연구의 운용에 관한 일반적인 원칙과 중간분석에서 고려해야할 점을 소개하였다. 제 4 절에서는 임상연구가 끝나고 난 후 논문을 작성할 때나 혹은 이미 작성되어 있는 논문의 통계적인 평가를 위한 주요 사항을 의학논문이라는 특수성에 맞추어 논의하였다.

2.임상시험연구의 개념및 설계

2.1 연구의 유형

의학분야에서의 연구는 그 설계방법에 따라 크게 관측연구(observational study)와 실험연구

1) (136-701) 서울시 성북구 안암동 5-1, 고려대학교 통계학과.

(experimental study)로 분류할 수 있다.

관측연구란 연구대상자의 상태와 특성을 단순히 관찰하여 분석하는 것으로서 사례-대조 연구(case-control study), 코호트 연구(cohort study), 현황연구(cross-sectional study)등이 있다. 사례-대조 연구는 조사대상자들을 결과에 따라 사례군과 대조군으로 구분하고 각 집단에 대해 과거에 특정요인을 갖고 있었는지의 여부를 조사하여 분석하는 방법이다. 예컨대, 폐암환자와 비(非)폐암환자군들에 대해 그들이 흡연경력이 있었는지를 조사하여 흡연과 폐암과의 인과성에 대한 추론을 이끌어내는 방법이다. 이 방법은 연구시점에서 과거를 향해 조사해가는 방법이므로 후향연구(**後向研究**:retrospective study)이며, 희귀질병이나 경과기간이 긴 연구에서 효율적으로 쓰일 수 있다. 이와는 달리 코호트 연구는 연구방향이 앞으로 진행되는 전향연구(**前向研究**:prospective study)로서 예를 들어 비흡연자 집단과 흡연자 집단을 각각 20년간 추적하여 그들의 폐암발생여부를 조사하는 방법이다. 이 연구방법은 인과관계를 추론하는 데 있어서 관측연구중 가장 이상적인 방법이기는 하지만 시간과 비용의 소모가 크고 연구관리가 어렵다는 단점이 있다. 김일순(1986)은 코호트 연구의 장단점들을 소개한 바 있다. 이 두 연구설계가 시간간격을 두고 조사되는 반면 현황연구는 어느 특정 시점에서 대상자들을 조사(survey)하여 모은 자료를 분석하는 방법이다. 이는 상대적으로 단시일내 연구가 끝나는 장점이 있으나 시점에 따라 영향을 받는 연구의 경우에는 왜곡된 결과를 불러 일으킬 수 있다.

이들 연구방법중 어느 것을 이용했는가에 따라 결과의 해석에 많은 영향을 미칠 수가 있다. 다음의 <표 1>은 고혈압과 뇌질환의 관계를 알아보기 위하여 고혈압환자와 비고혈압환자 1000명씩을 각각 10년동안 추적조사한 코호트조사의 예이다(cf. Dawson-Saunders and Trapp,1994). 뇌질환을 일으킨 비율은 300대 100으로 고혈압환자에게서 많이 발생하며 이로 인한 사망비율만을 본다면 250대 20으로 고혈압이 뇌질환의 원인이 된다는 것을 알 수 있다. 그러나 만약 같은 데이타를 10년후 시점에서 거꾸로 조사하는 사례-대조 연구를 이용했다고 하자. 그렇다면 <표 1>에서 뇌졸중으로 사망한 사람들의 자료를 관찰할 수 없게 된다. 이 때 오즈비(odd-ratio)는 (50x900)/(80x700)=0.80으로 오히려 고혈압이 뇌질환발생을 억제하는 효과가 있다는 잘못된 결론이 내려지게 된다. 이것이 연구방법의 선택에 따른 편의(bias)이며 이외의 많은 편의들에 대한 논의가 있다. 특히 Sackett(1979)은 환자나 연구방법을 선택할 때 발생할 수 있는 35종의 편의를 지적하였다.

관측연구는 연구자에게 기술적인(descriptive) 측면에서 유용한 정보를 줄 수 있다. 예컨대 연구자는 이러한 연구의 결과로부터 임상시험에서 사용될 처리에 대한 정보를 얻을 수도 있다. 그러나 환자의 선정에 따른 편의가 발생할 소지가 크기 때문에 이 관측연구의 결과만으로 어떤 확정적인 결론을 내리는 데에는 한계가 있다. Green and Byar(1984)는 처리결과를 비교하기 위하여 관측연구를 사용할 때 발생할 수 있는 문제점을 논의한 바 있다.

<표 1> 코호트 연구의 예

코호트연구 결과 / 환자	뇌혈관 질환자	비(非)뇌혈관 질환자	뇌졸중원인 사망자
고혈압환자	50	700	250
비(非)고혈압환자	80	900	20

실험연구는 의학연구에 있어서 관측연구보다 일반적으로 사용되는 방법으로 환자를 대상으로 할 경우 임상시험(clinical trials)이라고도 한다. 실험연구에서는 연구대상에게 특정한 처리(treatment)를 한 후 그에 따른 효과를 관찰함으로써 처리와 반응간의 인과관계를 규명하고자 한다. 여기서는 반드시 실험의 대상이 되는 실험군과 비교하기 위한 대조군(control group)이 필요하게 된다. 대조군은 편의상 과거대조군(historical control)과 확률화대조군(randomized control)의 두 가지로 분류할 수 있다.

과거대조군은 과거에 실험된 환자들 중 현재 실험대상이 되는 환자군과 비슷한 유형의 사람들을 대조군으로 사용하는 것이다. 이 방법은 희귀한 질병의 연구시에 확률화 대조군을 사용하기에는 환자수가 부족할 때나 도의적인 문제를 일으킬 소지가 있을 때 확률화 대조군에 대한 대안으로 사용될 수 있으며, 시간과 비용이 적게 든다는 잇점이 있다. Gehan(1986)은 편의를 줄이는 방법과 함께 과거대조군이 확률화 대조군보다 적합한 경우에 대한 많은 예를 소개한 바 있다. 그러나 대조군으로 사용할 환자들은 다른 상황(예컨대, 병원, 의사, 간호 및 진단기법의 측면)에서 관리되었으므로 이러한 요인들이 실험결과에 영향을 미칠 위험이 따른다. Sacks et al.(1982)은 과거대조군을 사용한 연구에서의 결과가 확률화대조군을 사용한 연구에서의 결과에 비해 유의적인 차이를 보인 경우가 많았다는 보고를 한 바 있다. 과거대조군은 동일한 병원에서 동일한 연구자가 비슷한 유형의 환자들을 대상으로 계속해서 특정 처리에 대한 연구를 한 경우에 국한되는 것이 바람직하다고 하겠다.

확률화 대조군은 균등한 확률로 환자들을 각 처리방법에 할당하여 얻어지는 대조군을 말한다. 이는 처리의 배치에서 발생될 수 있는 편의(bias)를 감소시키는 가장 확실한 방법이며 따라서 가장 과학적인 접근방법이라고 할 수 있다. 대조군을 설정하는 문제에 관하여 많은 논의가 있어왔다.(cf.고응린,1977; 허명회, 1992). 확률화방법에 대해서는 2.3절에서 자세히 설명하기로 한다.

이상과 같이 연구설계를 유형별로 분류하여 보았으나 경우에 따라서는 이와 같은 분류가 적당하지 않을 수가 있다. 예를 들어 암 치료제의 새 촉매에 대한 반응률을 알아보기 위한 제 2 상(Phase II) 실험에서 연구자는 연구를 계속하기에 적당한 성공률이라고 생각되는 값을 미리 정해놓고 이와 비교를 하게 된다. 이러한 경우에 대조군은 과거의 경험에 근거하여 주관적으로 정한 것으로 편의상 과거대조군으로 구분할 수 있으나 과거대조군의 통상적인 정의에 합당한 경우는 아니다. 이외에도 한 병원에서 동시에 시행되었으나 확률화에 의한 처리의 할당을 하지 않은 경우가 있다. 이때 사용된 대조군을 동시적 비확률화 대조군(concurrent nonradomized controls)이라고도 하는데 이 역시 관측연구나 실험연구중의 어느 한 범주로 구분짓기는 어렵다. 허명회(1992)는 확률화의 기능과 비확률화연구의 필요성에 대하여 설명한 바 있다.

2.2 임상시험연구의 단계

임상시험단계는 편의상 4단계로 분류하여 단계별로 설정된 주요실험내용을 관찰하고 그 결과를 분석하게 된다. 첫 단계인 제 1 상(Phase I)은 환자들을 대상으로 독성이나 부작용등의 중요하고 제한된 반응만을 관찰하는 단계이다. 이 단계에서 얻은 결과를 이용하여 투약의 최대허용량(maximum tolerated dose;MTD)과 투약기간, 투약방법등이 선택되며 다음단계로 이행할 지의 여부도 결정된다. 제 1 상 실험에서는 동물실험에 근거해서 독성이 없다고 판단되는 용량수준에서 3명의 환자로 출발하는 것이 보편적이다. 여기서 유의할 만한 독성이 발견되지 않으면 용량을 한 단계 높여서 다시 3명에 대해 실험한다. 독성이 발견될 때까지 이러한 과정을 되풀이하다가 독성

이 발견되면 보다 많은 환자(보통 6명)를 대상으로 실험한다. 이런 방식으로 최대허용량(MTD:maximum tolerance dose)이 확인될 때까지 용량수준을 높여 나간다 (cf. Von Hoff et al., 1984).

제 2 상(Phase II)은 특정질환의 환자를 대상으로 임상적 효과를 처음 관측하게 되는 단계로서 새로 개발된 약(치료법)이 보다 정확한 비교실험으로 이행할 만한 가치가 있는 것인지를 결정하게 된다. 이 단계에서는 많은 경우 연구자가 종전의 경험을 바탕으로 새 약(치료법)의 효율성 또는 비효율성의 증거가 되는 기준을 임의로 선택한다. 이 단계에서는 일반적으로 효과있는 약이 기각되는 것(제 2 종 오류)이 효과가 없는 약을 채택하는 것(제 1종 오류)보다 더 치명적인 오류로 간주되는데 그 이유는 효과가 없는 약을 채택했을 경우에는 제 3 상의 체계적인 비교연구를 통해 검증할 기회가 있지만 한번 기각된 약은 다시 사용되지 않기 때문이다(cf. Schoenfeld, 1980). 이러한 이유로 새로운 치료제가 기존의 치료제보다 더 효과적으로 나타난 경우에 제 3 상으로 이행할 확률(1-β)을 되도록 크게(예컨대 90%)하고 새 치료제가 효과가 떨어지는 것으로 나타났을 때에는 적당한 확률(1-α, 예컨대 75%)로서 이를 기각하도록 표본의 크기를 결정한다.

제 2 상 실험에서 표본의 크기를 결정하는 가장 간단한 예를 들어 보기로 한다. 어떤 치료제가 r% 이상의 반응률을 보이면 제 3 상으로 이행할 가치가 있는 것으로 판단된다고 하자. 이항분포를 이용하여 제 2 종 오류가 각각 5%인 경우와 10%인 경우에 대해 반응률의 수준에 따른 표본의 크기를 계산한 것이 <표 2>에 있다 (cf. Gehan, 1961). 반응률과 제 2 종 오류를 크게 정할수록 요구되는 표본의 수가 줄어든다는 것을 알 수 있다. 만약 반응률을 30% 로 잡았다면 제 2 종 오류를 5% 로 할 때 9명의 환자를 대상으로 실험하게 된다. 이때 9명중 단 한명이라도 반응을 보이면 반응률이 30%인 것으로 본다. 이러한 계획은 제 1 종 오류가 발생할 확률을 높게 한다. 이에 대해서 제 2 상 실험의 목적이 단지 최소의 표본으로 효과가 매우 없는 치료제만을 여과하는 데 있을 뿐이라는 주장이 있다.

<표 2> 제 2 상에서의 표본수의 예

제 2 종 오류	반응률 r (%)									
(β)	5	10	15	20	25	30	35	40	45	50
5%	59	29	19	14	11	9	7	6	6	5
10%	45	22	15	11	9	7	6	5	4	4

제 3 상(Phase III)은 본격적인 실험단계로서 새로운 약 또는 치료법(experimental treatment)과 기존의 약 또는 치료법(standard treatment)을 비교하도록 계획되어진다. 제 3 상에서의 임상시험을 비교임상시험이라고도 한다. 여기서는 확률화에 의해 각 처리에 할당될 환자를 결정하고 이로부터 새로운 약(치료법)의 채택여부를 결정할 자료를 얻게 된다. 제 3 상 실험에서 대상환자의 수는 연구자가 원하는 차이의 정도(δ)와 유의수준(α), 검정력의 크기(1-β), 그리고 사용되어지는 통계기법에 따라 다르게 계산되어진다. 제 3 상에서 사용될 확률화의 방법들과 대상환자수를 결정하는 방법에 대해서는 2.3절과 2.4절에서 다루게 될 것이다.

제 4 상(Phase IV)은 일단 인정되어 시판(시행)되고 있는 약(치료법)에 대한 분석과정으로 유통과 치료과정을 거치면서 그 효과와 부작용등을 재평가하고 개선점을 찾기 위한 과정을 일컫는다. 제 1 상에서 제 4 상까지의 4가지 유형은 서로 완전히 독립적이지는 않으며 부분적으로 겹치거나

단계를 거슬러 올라오는 경우도 발생할 수 있다.

2.3 확률화(randomization)에 따른 실험계획법

임상시험의 제 3 상에서는 확률화에 의한 실험계획이 요구된다. 앞서 소개한 대로 확률화란 실험에서 각 처리가 배당될 확률이 균등하게 주어지도록 하는 통계적 방법을 말한다. 확률화를 하는 이유는 첫째, 확률화를 함으로써 환자들을 각 처리집단으로 할당할 때 (고의로 또는 우연히) 발생할 수 있는 편의(bias)를 감소시켜 처리 효과에 대한 보다 신뢰성 있는 결론을 내리는 데 도움이 되기 때문이다. 둘째로, 확률화는 통계적 추론(statistical inference)의 밑바탕이 되기 때문이다. 확률화 실험의 질을 높이기 위해서는 눈가림실험(blind experiment)이 보장되는 것이 좋다. 눈가림 실험이란 실험후 처리효과를 평가하는 데 있어서 환자나 실험자의 편견이나 심리적 영향이 발생할 소지를 막기 위하여 어느 종류의 치료법이 누구에게 적용되는가를 모르게 하는 방법이다. 환자측만 모르게 시행하는가 아니면 환자와 시험자 모두가 모르게 시행하는가에 따라 일방눈가림(one-sided blind) 또는 이중눈가림(double blind)으로 불리워 진다.

확률화에 의한 실험계획법은 대상환자수의 사전확정여부에 따라 고정설계법(fixed sample design)과 축차설계법(sequential design)으로 나뉘어지는데 이를 차례로 설명하기로 한다.

2.3.1 고정설계법(fixed sample design)

고정설계법은 표본의 크기를 사전에 정해놓는 방법으로 완전 확률화 계획법(simple randomization), 블럭 확률화 계획법(block randomization), 층화 확률화 계획법(stratified randomization), 확률화 동의 계획법(randomized consent design)을 비롯하여 요인실험법(factorial design), 교차계획법(cross-over design)등이 모두 이에 속한다.

완전 확률화 계획법은 가장 기본적인 확률화법으로 단순히 연구에 참여한 환자가 각각의 처리를 받을 확률이 같도록 만들어 주면 된다. 이 방법에서는 각 실험대상이 동질적(homogeneous)이라고 가정하는데 반해 블럭확률화계획법은 실험대상자들이 모두 동질적이지는 않으나 처리(treatment)의 수(또는 그의 배수)만큼 균일한 실험단위를 모을 수 있을 때 사용할 수 있는 방법이다. 이 때 한 블럭안의 실험단위들에 각 처리가 골고루 배치되도록 하면 된다. 층화확률화 계획법은 여러 계층별로 따로 확률화를 하는 방법이다. 예컨대 실험이 여러 병원으로 나뉘어져 실시된 경우에 "병원"이라는 요인의 효과를 제거하기 위해서는 각각의 병원내에서 A약과 B약의 투약대상을 확률화법에 따라 선정하는 것이다. 층화에 따르는 문제는 층이 여러 개 있을 때에는 요구되는 표본의 크기가 커져서 결과적으로 불균형 데이타를 만들게 될 가능성이 크다는 데 있다. 다음의 표는 층화확률화계획의 한 예로 세 병원에 대해 층화를 한 다음 각 병원에서 별도로 환자의 보행가능여부를 블럭으로 하는 블럭확률화법을 사용한 것이다(cf. Zelen, 1974). 이 계획에 의해 병원 α에 온 어떤 환자가 보행가능한 상태였다면 처리 A에, 다음 환자가 병원 γ로 왔고 보행불능이었다면 처리 B에 할당되며 그 다음 환자가 병원 γ로 왔고 역시 보행불능이었다면 그는 처리 A에 할당된다. 이러한 방식으로 각 병원에 실려오는 환자의 상태에 따라 미리 계획된 확률화방법에 의해 처리가 결정된다.

<표 3> 층화블럭확률화 계획법의 예

환자의 상태 (블럭)	보 행 가 능			보 행 불 능		
병원 (층)	α	β	γ	α	β	γ
처리(A,B)의 할당계획	A	B	A	B	A	B
	A	B	B	A	B	A
	B	A	A	B	B	A
	B	A	B	A	A	B

확률화 동의계획법(**確率化 同意計劃法**:randomized consent design)은 실제 문제에 있어서 동의를 받은 환자들만을 대상으로 확률화를 해야 하는 문제에 대한 대안으로 제시된 것이다(cf. Zelen, 1979). 예컨대 기존의 안전한 약 A와 새로 개발된 약 B를 비교하는 문제를 생각해 보자. 환자의 동의를 물을 경우 많은 환자들이 약 B의 투여를 꺼리게 될 것이므로 확률화 이전에 이미 많은 수의 표본을 잃게 되는 결과를 낳게 된다. 이 때 먼저 확률화를 하여 두 그룹으로 나누어 한 그룹에는 동의를 묻지 않고 약 A를 투여 하고 나머지 그룹의 환자들에게는 의사를 물어 처리를 선택하도록 함으로써 "확률화"와 "환자의 동의"라는 상충되는 문제를 표본의 손실없이 해결하려는 방법이다.

확률화를 하는 데 있어서 실험을 복잡하게 만드는 원인중의 하나는 여러가지 처리에 대한 연구를 동시에 수행하고자 하는 것이다. 예컨대 암환자들에게 영향을 미치는 요인 중 하나가 투여되는 약의 종류(A약, B약)이고, 다른 하나는 투약의 간격(x일, y일)라고 하자. 이 때 환자들에게 (A약, x일), (B약, x일), (A약, y일) ,(B약, y일)의 4가지 처리를 완전 확률화 또는 블럭 확률화에 의해 배치할 수 있다. 이와 같은 처리계획을 요인실험법(**要因實驗法**:factorial design)이라 한다. 두 요인간의 교호작용(interaction)이 없다면 두 약의 효과 또는 투약간격에 따른 효과를 비교하기 위해서는 자료를 통합하여 원하는 결과를 얻을 수 있다. 따라서 교호작용이 없는 경우에는 요인실험법을 사용함으로써 하나의 요인을 비교하는데 필요한 표본을 이용하여 두가지 목적을 달성하는 효과를 거둘 수 있게 된다. 그러나 교호작용이 있는 경우에 요인실험법을 사용하게 되면 동일한 검정력을 유지하기 위해 필요한 표본의 수가 상당히 커지게 된다. 교호작용이 없다는 가정에 대한 정당성은 연구되는 특정질병과 치료법의 특성에 크게 의존하지만 이 방법에 대한 논란의 여지는 있다. 일부 연구자들은 이러한 관점에서 임상시험에 있어서 요인실험법이 유용하지 못하다는 주장을 하기도 한다(cf.Peto et al.,1976; Peto, 1978).

또 다른 방법으로 교차계획법(**交叉計劃法**:cross-over design)이 있다. 이 방법은 환자들에게 어떤 처리를 하고 일정기간이 지나간 뒤에 다시 다른 처리를 적용하는 방법이다. 교차계획법에서의 대조군은 환자자신이 되므로 이를 자기대조군(self control)이라고도 한다 (cf. Dawson-Saunders and Trapp,1994). 다양한 찬반론이 있으나 교차계획법이 처리의 효과가 짧게 지속되는 연구에 보다 적합하다는 데는 대부분 동의하고 있다(cf. Brown, 1980; Louis et al., 1984). 이밖에 분할구계획법(split-plot design) 이나 지분계획법(nested design)등도 임상시험연구에 활발히 응용되고 있다 (cf. 송혜향 · 이홍준,1993).

2.3.2 축차설계법(sequential design)

축차설계법은 고정설계법과는 달리 환자의 수를 사전에 정하지 않고 실험의 결과를 관찰하여 결정하는 방법을 말한다. 여기서는 Armitage(1975)의 축차설계법과 Zelen(1969)의 승자원칙게임설계법(play the winner rule)에 대하여 설명하겠다.

Armitage의 축차설계법은 Wald(1947)에 의해 개발된 축차방법을 이용한 것으로 두 명의 환자를 한 쌍으로 간주하여 확률화를 통해 처리를 할당한 후 그 결과를 비교하여 우수한 효과를 보인 처리를 기록한다. 이런 방식으로 실험을 계속한다. 두 처리방법의 효과가 동일하다면 이 결과는 1/2의 확률을 가지는 이항분포를 보이게 되므로 이로부터 각 단계에서의 양측검정을 위한 한계치를 찾을 수 있다. 따라서 처리 A의 효과가 우수하거나, 처리 B의 효과가 우수하거나 또는 두 처리간의 차이가 없다는 세가지 형태의 결론을 내리게 되므로 이를 폐쇄형 축차설계(closed sequential design)라고도 한다. 이에 대해 두 처리중 어느 한 처리의 효과가 좋은 것으로 나타날 때까지 실험을 계속하는 방법을 개방형 축차설계(open sequential design)라고 한다.

Zelen 의 승자원칙게임설계법은 효과가 좋은 처리 방법에 더 많은 환자를 할당하는 원칙이다. 먼저 첫번째 환자에게 두 처리중 하나를 확률화를 통하여 할당한다. 그 결과가 성공이었다면 두번째. 환자에게도 같은 처리를 실시하게 되지만 실패였다면 다른 처리를 할당한다. n번째 환자에게 어떤 처리가 할당될 확률은 앞의 n-1명의 결과에서 그 처리의 성공비율로 된다. 예를 들어 첫번째 환자에게 처리 A를 실시하고 그 결과가 성공적이라 하자. 두번째 환자에게는 처리 A가 실시된다. 두번째 환자의 결과도 성공적이었다면 세번째 환자에게도 처리 A를 실시한다. 만약 세번째 환자의 결과가 실패였다면 네번째 환자가 처리 A를 할당받을 확률은 이제 2/3가 되고 처리 B를 할당받을 확률은 1/3이 된다. 이런 방식으로 누적되는 결과에 의거하여 처리를 할당하게 된다.

축차설계법은 비교되는 두 처리의 효과가 큰 차이를 보일 때 결정을 빨리 내릴 수 있다는 장점이 있으나 결과가 단시간내 파악될 수 없는 유형의 임상시험에는 부적당하며 경우에 따라서는 고정설계법에서보다 더 많은 수의 환자가 필요하게 되는 수도 있다.

2.4 표본크기의 결정

임상시험연구에 있어서 가장 중요한 사항중의 하나가 연구에 포함될 적정한 표본수를 결정하는 문제이다. 표본크기의 결정은 연구의 설계단계에서 미리 고려되어야하는 것임에도 불구하고 많은 연구에서 이러한 원칙이 무시되고 일정기간동안에 수집가능한 정도를 표본의 크기로 결정해 버리는 경우가 많다. 표본크기를 사전에 고려하지 않은 임상시험중에는 검정력이 떨어지거나 혹은 임상적으로 의미있는 효과를 찾아내지 못하는 경우가 많게 된다.

표본의 크기를 구하는 방법은 연구시작전에 미리 고정된 수를 할당하여 연구를 시행하는 고정설계법과 연구를 시행해가면서 나타나는 결과에 따라 표본의 수가 결정되는 축차설계법에서 각기 다르다. 또한 단 한번의 분석만을 할 것인지 연구도중에 여러번의 중간분석을 하게 될 것인지에 따라서도 달라지게 된다. 제 3절에서 설명할 집단축차검정(group sequential test)등을 이용한 중간분석(interim analysis)을 실시하도록 계획되어 있는 경우에는 필요한 표본의 크기가 증가하게 된다. 그러나 연구가 일찍 종료될 가능성도 있기 때문에 실제로 연구에 참가하는 환자의 수는 오히려 더 작아질 수도 있게 된다. 중간분석을 수행할 경우에 어떻게 표본크기를 결정하는가에 관한 많은 연구가 있었다(cf. Kim and DeMets, 1987; Jennison and Turnbull, 1989). 설계방법과 분석의 횟수가 정해지고 난 후에 요구되는 표본의 크기는 이 밖에 유의수준(α) 과 검정력의 크기($1-\beta$), 연구자가 원하는 차이의 정도(δ), 그리고 사용되어지는 통계기법에 의존하게 된다. 여기서는 고정

설계에 의한 연구에서 단 한번의 최종분석이 시행되는 경우에 한하여 표본의 크기를 구하는 몇가지 유용한 방법들을 소개하게 될 것이다.

먼저 두개의 비율을 비교하기 위해 필요한 표본크기를 결정하는 문제를 고려해보자. P_1과 P_2가 각각 그룹 1과 그룹 2에서 일정한 시간동안 발생한 사건(event)의 비율을 나타낸다고하고, N명의 환자가 그룹 1에 rN명의 환자가 그룹 2에 확률화를 통해 할당되었다고 하자. 귀무가설과 대립가설은 각각 다음과 같이 된다.

$$H_0: P_1 = P_2 \ (\text{즉}, \ \delta = P_1 - P_2 = 0)$$
$$H_A: P_1 > P_2 \ (\text{즉}, \ \delta = P_1 - P_2 > 0)$$

이때 필요한 표본크기는 다음과 같이 계산된다 (cf. Fleiss, 1973).

$$N = \frac{\{Z_\alpha \sqrt{(r+1)\bar{P}(1-\bar{P})} + Z_\beta \sqrt{rP_1(1-P_1) + P_2(1-P_2)}\}^2}{r(P_1 - P_2)^2}$$

여기서 $\bar{P} = (P_1 + rP_2)/(r+1)$이고, Z_α와 Z_β는 각각 α, β에 해당하는 표준정규분포의 임계치이다. 양측검정을 실행하는 경우 (즉, $H_A: P_1 \neq P_2$) 에는 Z_α 대신 $Z_{\alpha/2}$ 를사용한다. 일반적으로 위의 공식은 실제크기에 가까운 근사값을 주지만, 특히 표본크기가 작을때는 연속보정(continuity correction)이 있는 다음 공식이 더욱 좋은 근사값을 제공한다 (cf. Fleiss et al., 1980).

$$N' = \frac{N}{4}\left(1 + \sqrt{1 + \frac{2(r+1)}{rN\delta}}\right)^2.$$

각 그룹에서 $100\,p\,\%$ 의 환자가 다른곳으로 이동하는등의 이유로 추적불가능할때(loss to follow up) 같은 검정력을 얻기 위해서 필요한 환자의 수는 $N/(1-p)$가 된다. 검정력을 크게 하기 위해서는 두 그룹에 할당된 환자수가 같도록($r=1$) 하는 것이 바람직하나 윤리적 또는 경제적 이유등으로 서로 다른 수의 환자를 할당하는 것이 필요한 때도 있다. 환자그룹이 몇개의 층(strata)으로 구분되는 경우에는 다른 방법으로 표본크기를 계산한다 (cf. Gail,1973).

다음으로 처리효과의 동등성(equivalence)을 보이기 위한 연구에서의 표본크기에 대하여 설명하겠다. 어떤 질병에 대해서 좋은 효과가 있는 기존 치료법이 있다고 하자. 새로운 치료법이 기존치료법보다 훨씬 독성이 없고 시행하기 쉽거나 부작용이 적다고 한다면, 이 치료법이 기존 치료법만큼의 효과가 있다는 것을 보이기 위한 연구를 하고 싶을 것이다. P_1과 P_2를 각각 기존 치료법과 새로운 치료법의 성공율이라고 하자. 두 효과의 차이가 δ보다 작거나 같을때, 두 치료법의 효과가 같다고(equivalent) 본다. 이러한 경우에는 두 치료법의 효과가 동등하다는 것을 대립가설로

($H_A: P_1-P_2 \leq \delta$), 새로운 치료법의 효과가 기존 치료법보다 못하다는 것을 귀무가설로 ($H_0: P_1-P_2 > \delta$) 설정한다. 일반적으로 동등성검정에는 단측검정을 사용하는데, 그 이유는 새로운 치료법이 기존 치료법보다 더 좋다는 것을 보이는 것에는 관심이 없기 때문이다. N명의 환자를 기존치료법에 rN명의 환자를 새로운 치료법에 할당할때, N을 구하는 공식은 다음과 같다 (cf. Donner, 1984).

$$N=\frac{\{z_\alpha\sqrt{(r+1)\overline{P}(1-\overline{P})}+z_\beta\sqrt{rP_1(1-P_1)+P_2(1-P_{2)}}\}^2}{r(P_1-P_2-\delta)^2}$$

여기서, $\overline{P}=(P_1+rP_2)/(r+1)$ 이다. 실제로는 P_1과 P_2를 같게 하고 δ를 연구자가 무시해도 좋은 정도의 차이로 놓은 다음에 이 공식을 적용하는 것이 보통이다.

이번에는 병원에 입원한 기간, 콜레스테롤 수준, 혈압, 폐활량등 연속형 반응변수의 평균을 비교하고자 할 때에 표본크기를 구하는 방법을 알아보자. 반응변수 x가 평균이 μ이고, 분산이 σ^2인 정규분포를 따른다고 가정하고, 그룹 1과 그룹 2에 각각 N명과 rN명이 확률화를 통해 할당된다고 하자. 귀무가설과 대립가설을

$$H_0: \delta=\mu_1-\mu_2=0$$
$$H_A: \delta=\mu_1-\mu_2>0$$

라고 할 때 평균치의 차이가 정규분포를 따른다는 사실을 이용하면 표본의 크기를 쉽게 유도할 수 있다. $r=1$일때 N은 다음과 같이 구해진다.

$$N=\{2(Z_\alpha+Z_\beta)^2\sigma^2\}/\delta^2$$

비율을 비교하는 경우와 마찬가지로 양측검정일 경우에는 Z_α 대신 $Z_{\alpha/2}$ 를사용한다. 위의 공식에서 보듯이 표본의 크기는 관심있는 평균차이 (δ)가 작을수록 분산 (σ^2)이 증가할수록, 유의수준 (α)이 작을수록 검정력 ($1-\beta$)이 클수록 증가하게 된다.

두 그룹간의 평균변화율에 차이가 있는가를 알고 싶은 경우를 생각해보자. 각 환자에 대해 관측된 변수를 y라 할때 이를 회귀식 $y=a+bt+e$ 로 나타낼 수 있다. 여기서 a는 절편, b는 기울기, t는 시간을 나타내고 e는 오차항을 의미한다. e의 분산 $\sigma_e{}^2$이 각 환자마다 대략 일정하다고 가정했을때 단측검정의 경우 각 그룹에서의 표본크기는 다음과 같이 계산될 수 있다 (cf. Schlesslman,1973a).

$$N = \{2(Z_\alpha + Z_\beta)^2 \sigma_b^2\} / \delta^2$$

여기서 δ는 두 그룹간 변화율의 차이이며, ${\sigma_b}^2$은 기울기의 변동(variability)의 정도를 가리킨다. 그런데 ${\sigma_b}^2$ 은 다음과 같이 분할될 수 있다.

$${\sigma_b}^2 = {\sigma_B}^2 + \left(\frac{12(K-1)}{T^2 K(K+1)} \right) {\sigma_e}^2.$$

여기서 T는 환자 개개인에 대한 관측지속시간(time duration)이고, K는 같은시간간격의 관측횟수이며 ${\sigma_B}^2$는 기울기의 분산중 관측오류(measurement error)나 선형적합결여(lack of fit)등에 기인하지 않는 순수한 환자변동부분을 나타낸다. 여기서 T와 K가 커지면 필요한 표본의 크기가 작아지므로 이를 잘 조정함으로써 효율적인 계획을 세울 수 있다 (cf. Schlesslman,1973b).

마지막으로 두 그룹의 생존분포(survival distribution)의 비교에 관심이 있는 경우에 대하여 알아보자. 두 그룹에서의 위험율의 비(hazard rate ratio) λ_1/λ_2를 δ라고 했을때 귀무가설을 $H_0: \delta = 1$, 대립가설을 $H_1: \delta > 1$로 놓자. 연구에서 중도절단(censoring)이 없고 각 그룹에 동일한 수의 환자를 할당한다면 생존시간이 지수분포(exponential distribution)를 따른다고 가정했을 때 각 그룹에서 필요한 환자수는 다음과 같이 계산될 수 있다 (cf. George and Desu, 1974).

$$N = 2\{(Z_\alpha + Z_\beta)\}^2 \; / \; \{\ln(\delta)\}^2$$

Rubinstein et al.(1981)은 보다 현실적으로 중도절단이 있고 추가추적기간(additional follow-up period)이 있을 때 필요한 환자수를 계산하는 방법을 제안하였다. 각 그룹에서의 중도절단비율이 c_1, c_2 라고 하자. 연구기간이 T년, 추가추적기간이 τ년일때 매년 N명의 환자가 포아송과정(poisson process)에 따라 연구에 들어온다고 가정하면 한해에 필요한 환자수는 다음과 같이 구해진다.

$$N = \left(\frac{Z_\alpha + Z_\beta}{\log \delta} \right)^2 \sum_{i=1}^{2} \frac{2\lambda_i^{*2}}{T\lambda_i} \left\{ 1 - \frac{e^{-\tau\lambda_i^*}(1 - e^{-T\lambda_i^*})}{T\lambda_i^*} \right\}^{-1}$$

여기서 $\lambda_i^* = \lambda_i + c_i$ 이다. 이외에도 생존분포를 비교할 때 대상환자수를 계산하는 방법은 상황에 따라 다양하게 개발되어 있다(cf. Hsieh,1992; Xiang and Lee,1994).

3. 임상시험의 운용

3.1 연구의 점검 - 중간분석과 위원회의 운영

임상시험을 행하는데 있어 미리 계획된 적절한 시간에 단 한번의 분석을 함으로써 치료법의 효과등을 검정하는 것은 통계적 측면에서 볼 때 비교적 간단한 일이다. 그러나 실제로는 연구의 합리적인 운용과 환자의 안전을 위하여 정기적인 중간점검(interim monitoring)이 필요하게 된다. 예컨대 제 2 상 실험에서 새로운 치료약의 치명적인 독성이 여러 환자들에게서 발견되면 임상시험 중간에 투약의 감량을 고려하는 것이 바람직할 것이다. 또한 제 3 상 실험에서 중간분석을 행한 결과 어떤 처리가 다른 처리에 비해서 월등히 효과적인 것으로 드러났다면 실험을 일찍 종료하여 환자들에게 효과가 적은 약을 처방하지 않도록 하는 것이 필요하다. 따라서 연구자는 연구설계단계에서부터 이러한 중간점검을 고려하여 적절한 통계적방법을 채택하여야 한다. 이러한 중간점검과 관련한 임상적, 통계적 문제점들이 다양하게 논의된 바 있다(cf. Fleming et al., 1984; DeMets, 1984). 여기서는 제 3 상 실험에 촛점을 맞추어 중간분석의 방법에 대하여 살펴보기로 한다.

올바른 분석을 위해서는 중간분석의 각 단계에서 사용되는 유의수준이 조정되어야 한다. 각 중간검정마다 정해진 유의수준(예컨대 α=0.05)을 반복하여 사용하면 실제적으로 제 1 종 오류가 정해진 것보다 매우 큰 값을 갖게 된다. 예컨대 그런 검정이 2회 반복되면 제 1 종 오류가 약 8%가 되며, 5회의 검정은 약 14%의 제 1 종 오류를 낳게 된다(cf. Armitage et al., 1969). 원하는 수준의 제 1 종 오류를 얻기 위해서 많은 축차검정법(sequential test)들이 개발되었으나 이 방법들은 새로운 데이타가 얻어질 때마다 분석해야 하는 제약이 있어 임상시험에 실제로 적용시키는데 많은 어려움이 있다 (cf. Armitage, 1975).

보다 현실적인 방법으로 임상시험중 몇 차례 정해진 수의 계획된 분석을 실시하는 집단축차검정법(group sequential test)이 많이 개발되어 왔다(cf. Pocock, 1977; O'Brien and Fleming,1979; Lan and DeMets, 1983; Lee and DeMets, 1991, 1992; Lee,1994a,1994b). 이 방법의 실제적용은 특정한 임상시험에 관하여 분석의 시점, 횟수, 최종분석시기등의 다양한 결정을 요구한다. 무엇보다도 각각의 중간분석에 제 1 종 오류를 어떻게 분배할 것인가를 결정해야 한다. 많은 통계학자들은 변동에 불과한 처리간의 차이를 잘못 속단하여 연구를 조기종료하는 잘못을 범하지 않도록 하기 위해서 초기 분석에서의 유의수준은 매우 작게 하고 마지막 분석에서의 유의수준은 정해진 수준에 가깝도록 배정하는 방법을 선호한다.

그러나 집단축차검정법은 어디까지나 연구 종료시기의 결정을 위한 부분적인 지침에 불과하며 그러한 결정이 내려지기 전에 고려해야 할 다른 요인들이 많이 있다. 예를 들면 중요한 예후인자(prognostic factors)가 각 처리그룹에서의 균형을 이루는지, 결과가 통계적으로 유의할 때 그것이 임상적으로도 유의하며 다른 관련 연구와 일치하는지등을 고려해야 하며, 연구의 결과가 폭넓게 활용될 수 있는 것인지 아니면 특정 환자군에서는 상당히 달라질 수 있는 것인지등도 고려해야 할 것이다.

일단 연구의 점검에 대한 통계적인 기준이 결정되고 난 후에 해결해야 할 중요한 문제중의 하나는 누가 자료를 점검하고 여러가지 결정에 대한 책임을 지느냐하는 것이다. 미국 및 유럽에서는 임상시험을 행하는데 있어 연구점검위원회(study data monitoring committee)의 운영이 보편화되어 있으며, 이 위원회는 중간분석결과가 과학적으로 타당한 것인지를 점검하고 환자와 공공의 이익이 최대한 보장될 수 있는 결정을 내려야 한다. 위원회의 역할 및 운용원칙에 대해 많은 논의가

있어왔다(cf. Green et al., 1987 ; Geller, 1987).

위원회는 대체로 연구개발자(연구의 좌장, 통계학자등) 및 관련분야의 외부인사로 구성되어 있는데 외부인사를 참여시키는 이유는 연구개발자는 어느 정도 연구결과에 대해 이해관계를 갖고 있기 때문에 연구의 계속여부를 결정하는 문제에 있어서 완전한 객관성을 유지하지 못할 우려가 있기 때문이다. 예를 들어 미국의 아동암그룹(Childrens Cancer Group)에서는 연구의 좌장, 담당 통계학자, 그룹의 회장외에 미국국립보건연구원(National Institutes of Health:NIH)을 대표해서 연구관련 전문의학자와 통계학자, 윤리학자가 위원회에 참여하고 있다. 또한 프랑스에서는 연구의 좌장, 담당통계학자, 변호사, 스폰서 및 신부가 참여한다.

임상시험에 참여한 임상의가 중간분석결과를 알게 되면 연구수행에 영향을 미치게 되기 때문에 대규모 다기관 임상시험(multicenter clinical trial)에서는 위원회 외부의 참여자들에게는 그동안 참여한 환자의 수, 관찰된 독성 및 합병증에 대한 정보등의 극히 제한된 정보만을 제공하고 있다. 또한 중간결과를 알려줄 필요가 있을 때에도 치료법을 x, y등의 부호형식으로 표기하여 특정치료에 대한 개인적인 편견을 제거해야 하는데 이러한 방법이 완벽한 해결책이 되지는 못하므로 세심한 주의가 필요하다. 예를 들어서, 두 가지 약이 거의 비슷한 효과를 보이고 있다면 일부 참여자들은 환자의 참여나 실험의 지속이 결과에 영향을 미치지 않을 것이라고 속단하여 더 이상의 참여를 회피할 수도 있다. 이러한 이유에서 임상시험을 수행하는 데에 있어서 위원회의 역할이 더욱 강조되고 있다.

3.2 약정위배와 처리의향분석 (protocol violation and intent-to-treat analysis)

연구자들이 임상시험연구를 계획함에 있어서 가장 먼저 해야 할 일은 연구약정(study protocol)을 만드는 것이다. 약정서에는 일반적으로 임상시험의 목적과 실험설계, 수행방법및 일정등의 세부적인 계획등이 제시되어야 한다. 예를 들어 미국의 국립암연구소(NCI:National Cancer Institute)에서는 연구의 목적과 필요성, 환자의 적격심사기준(eligibility criteria), 처리의 계획, 약물에 대한 정보, 환자의 선정방법, 반응평가의 기준, 환자의 관리, 독성에 대한 투약용량의 조정계획, 처리의 종료나 환자의 제외에 대한 기준, 통계적 방법, 각종기록, 참여하는 임상의들의 역할등과 함께 다기관 실험의 경우 추가되는 몇가지 사항들을 약정서에 포함시키도록 요구하고 있다.

이러한 약정서대로 연구가 수행되어야 하는 것은 상당히 중요하다. 그러나 실제로는 많은 종류의 약정 위배(protocol violation)가 발생하게 되며 이것이 연구의 기본원칙을 위협할 만큼 중대한 것일 수도 있다. 연구시에 일어날 수 있는 모든 종류의 약정위배를 미리 예상하기는 어렵지만 연구의 설계자는 어떤 특정분야에서 약정위배의 가능성을 적절히 예상해야 할 필요가 있다. 예컨대 기술적으로 어렵고 경비가 많이 드는 실험이나 재료나 정보면에서 복잡한 논리적 전환이 필요한 실험을 행할 때에는 사전에 약정위배의 경우에 대한 고려가 있어야 한다. 그렇지 않으면 실험이 시작되자마자 이러한 문제들이 나타나게 되고 즉시 수정가능한 문제가 아니라면 연구의 전반적 가치를 제한할만큼의 많은 손실이 있을 수도 있다. 따라서 연구의 설계시점에서 가능한 주요 위배사항들을 미리 예측하고 가능한 한 이 문제를 그 시점에서 해결할 수 있도록 하는 것이 바람직하다.

이렇게 사전에 약정위배의 경우를 고려하여 조심스럽게 설계를 했다 하더라도 이의 발생 소지는 여전히 남아있다. 약정위배의 형태는 각 실험에서 독특한 형태로 나타나게 되므로 일반적으로 다루어질 수 있는 사항은 아니지만 모든 실험에서의 약정위배 형태를 아는 것은 이러한 문제의

통계적 접근방법에 도움을 줄 수 있을 것이다. 몇가지 예를 생각해 보자.

확률화실험에서 가장 기본적인 약정위배의 형태는 어떤 치료법에 임의로(randomly) 할당된 환자가 즉시 다른 치료법을 받기로 결정하는 경우이다. 그러한 문제가 발생하는 이유는 다양하며 모두 피하기는 어렵다. 임상의가 일단 확률화에 응하여 기회를 얻은 다음에 자신이 원하는 치료법으로 배정되지 않으면 즉시 약정위배를 범하게 되는 수도 있고 환자나 그의 가족이 마음을 바꾸는 경우도 있다. 이 외에도 일선병원에서 자료를 관리하는 사람의 연구자체에 대한 이해부족이나 실수등의 여러가지 다양한 돌발상황에 의해 약정위배가 발생할 수 있다. 앞서 설명한대로 정기적인 점검이 이런 문제에 도움이 될 수 있으며 특정한 상황에서의 반복적인 위배는 원인을 분석하여 제거해야 한다. 미국의 경우 대규모의 다기관 공동임상시험을 수행함에 있어 약정위배의 여부를 연구팀의 좌장과 담당통계학자가 상의해서 결정하며, 결정된 사항및 그에 따른 후속 조치는 각 임상의에게 즉시 통고하고 있다. 또한 통계학자를 주축으로 임상의 개개인의 약정위배여부를 검토하는 위원회(performance monitoring committee)가 있어서 일선병원의 자질과 재정지원정도를 정기적으로 심사하고 있다. 이와같이 대규모 임상시험에서는 약정위배를 철저히 관리하는 것이 필수적인 것이며 이를 위해서는 합리적이고 과학적인 통계적 사고방식이 필요한 만큼 통계학자의 역할이 강조되고 있다.

이제 이러한 경우의 분석방법에 대해 생각해 보자. 약정위배를 보이는 환자들을 통계분석의 첫 단계에서 제외시켜서는 안된다는 것이 원칙이다. 이런 환자들을 제외시키는 것은 비교가능한 처리집단을 얻기위한 확률화의 기본이념에 반하는 심각한 편의(bias)를 초래할 가능성이 있다. 예컨대 암 환자들이 모두 수술과 방사선치료를 받고 이후에 확률화에 의해 보조화학치료를 받는지 안받는지를 결정하는 연구를 생각해보자. 어떤 환자가 확률화된 이후에 즉시 (또는 얼마후에) 그 결정을 따르지 않기로 했을때 그를 분석에서 제외한다고 가정하자. 건강상태가 좋지 않은 환자가 보조화학치료를 받는 것으로 할당되었을때 환자나 주치의가 화학치료법의 부작용을 우려하여 그런 치료를 받지 않겠다고 결정할 수 있다. 그런 사람들이 많아지면 보조화학치료를 받는 집단에서는 건강상태가 나쁜 환자들이 제외된 결과가 되어 보조화학치료에 유리한 결론을 내리게 될 수 있다. 따라서 대부분의 통계학자들은 환자가 실제로 어떤 처리를 받았느냐에 관계없이 애초 할당된 처리집단으로 간주하고 분석해야 한다고 본다. 이것을 처리의향분석(intent-to-treat analysis)이라고 한다. 환자를 실제로는 그가 받지도 않은 처리의 집단에 포함시켜야 한다는 것이 임상의에게는 받아들이기 어려울 수도 있지만 앞서 설명한 바와 같이 이의 통계적인 개념은 매우 기본적인 것이다(cf. Peto et al.1976).

적용후 얼마 지나서 약정위배를 보이는 환자들에 대해서 그 시점을 중도절단시기로 보아 분석하는 전략이 쓰이기도 하는데, 단순히 기술적인 관점에서만 본다면 그러한 생각은 잘못된 것이 없으나 실제로는 환자들을 제외시킬 때와 마찬가지의 편의가 심각하게 발생한다.

어떤 처리법의 효율성에 대한 추정은 약정위배요인에 대한 적절한 보정에 실패함으로써 매우 부적절하게 될 수 있다. 소아암에 관한 연구에서 특정약물의 검사시 약정위배가 아동과 그의 부모의 개인적 특성등 많은 요인과 관련이 있다는 흥미있는 결과가 있다(cf. Tebbi et al., 1986). 이러한 약정위배는 앞으로 많은 연구가 요구되는 영역이다.

4. 임상시험연구결과의 보고 또는 평가

실험의 잘못된 계획이나 실험도중 발생한 여러 문제에 대한 적절한 대응이 없었던 연구는 신뢰할 수 없는 결과를 보여주게 된다. 한 예로 The British Medical Journal, The Journal of the American Medical Association, The New England Journal of Medicine, The Lancet 등 세계적으로 권위있는 30개 의학전문학술지에 발표된 4235편의 논문 중 정당한 기준을 만족시킨 것은 20% 정도에 불과하다는 연구결과가 있다(cf. Williamson et al., 1986). 또한 이들 중 일부를 선택하여 조사한 결과 불충분하고 계획이 잘못된 연구의 80%가 원하는 결과(positive findings)를 얻었으나 제대로 계획된 연구의 경우에는 단지 25%만이 원하는 결과를 얻을 수 있었다고 한다. 즉 원하는 결과는 오히려 잘못 계획된 연구에서 더 많이 얻어지고 있다는 결론이다. 따라서 이러한 연구가 제대로 여과되지 않고 발표되었을 때 미치는 악영향은 심각하다. 발표된 결과가 어느 정도 신빙성이 있는지를 엄밀하게 평가하기 위해서 통계학자의 역할이 필요하다. 한편 통계학자가 처음부터 관여하여 적절한 계획과 운용, 분석이 이루어진 연구라 할지라도 연구목적, 연구계획, 대상환자수, 연구과정, 통계분석 그리고 연구의 한계등에 대한 올바른 제시가 없이 발표된다면 그것이 잘못 이해되고 무분별하게 인용되어 부정적인 결과를 초래할 수도 있게 된다.

현재 대부분의 권위있는 의학전문학술지에는 통계학자들이 편집진으로 참여하여 기고되는 모든 논문을 심사하고 있으며, 따라서 의학논문의 통계부분을 임상시험에 처음부터 관여했던 통계학자가 작성하는 것은 필수적이다. 현실적으로 국내에서 시행되는 임상시험에서는 통계학자가 관여하는 부분이 자료의 통계분석에 국한되고 연구결과의 보고나 평가부분에서는 배제되고 있는 형편이나 이는 점차 개선되어야 할 부분이다. 이 절에서는 의학논문의 특성을 소개하고 결과보고서의 작성 또는 평가시에 유의할 점을 간략히 소개하겠다.

의학논문은 일반적으로 요약(Abstract or Summary), 서론(Introduction), 연구방법(Methods), 연구결과(Results) 그리고 토론및 결론(Discussion or Conclusions)의 다섯절(section)로 구성된다.

대부분의 의학학술회의에서 요약부분만이 발표되고 많은 전문학술지에도 이 부분만이 게재될 정도로 다른 분야의 연구논문과는 달리 의학논문의 요약부분은 중요하게 여겨진다. 여기서는 연구의 목적은 물론 연구대상자의 선택이나 연구방법등의 기본절차에 대한 소개, 주요결과(특정데이타와 통계적인 유의성), 그리고 주요결론이 실리게 된다.

서론에서는 연구의 목적과 대상자, 연구기간등이 소개된다.(이것은 "연구방법" 부분에 쓰여지기도 한다.) 주의해야 할 점 중의 하나는 연구목적이 데이타 수집전에 수립되어 있었는지, 아니면 데이타를 통해 나온 결과를 토대로 재구성한 것인지를 파악하는 것이다. 후자의 경우는 그 결과를 우연히 얻게된 경우로 간주되어야 한다. 또한 연구목적에 비추어 볼때 수행된 연구기간은 적절하였는가를 살펴보아야 한다.

연구방법부분에서는 연구설계방법과 대상환자의 적격기준, 표본의 크기, 사용된 통계적 기법등이 쓰여진다. 연구의 유형이나 사용된 실험계획법, 그리고 확률화의 방법이 구체적으로 소개되는 것이 바람직하다. 교차계획법을 사용한 경우에는 투약기간(run-in period)과 약효제거기간(wash-out period)을 명시하는 것도 필요하다. 비확률화실험에서는 특히 발생할 수 있는 주요 편의에 대한 언급이 있어야 하며 비확률화실험을 하게 된 이유를 설명해야 한다. 다음으로 대상환자의 적격기준(eligibility criteria)이 자세하게 쓰여져야 한다. 임상시험에서는 대개 특정환자군에 적합한 치료를 시도하게 되므로 실험의 결과가 올바로 적용되려면 이들 환자군에 대한 상세한 설명이 필요하게 된다. 또한 특이한 질병이나 병력, 기타 예후인자(prognostic factor)들을 가진 환자들

을 실험에 포함시키지 않기 위해서도 이러한 대상환자의 적격기준을 자세히 기술해야 할 필요가 있게 된다. 이때 적격기준을 좁혀서 정하면 변동을 줄이고 통계적 정도(precision)를 높일 수 있는 반면 참여하는 환자의 수가 적어지고 결과를 일반화시키는 데 제약이 있게 된다. 대상환자의 적격기준은 임상시험에서 미리 결정해야 할 중요한 요소중의 하나로서 일단 기준이 정해지면 이는 엄격하게 지켜져야 한다. 표본의 크기와 왜 표본의 수를 그렇게 정했는가에 대한 근거를 제시하는 것은 의학논문에서 필수적이다. 표본의 크기를 결정하는 문제는 연구의 검정력(power)을 결정하는 문제와 관련되며 특히 기대한 정도의 차이가 발견되지 않은 연구에서 검정력의 언급은 필수적이다. Frieman et al.(1978)은 발표된 논문중 기대한 정도의 차이나 관계가 발견되지 않은 연구(negative study) 71편을 조사한 결과 설령 50%의 차이가 있었더라도 표본의 크기가 너무 작아서 그러한 차이를 찾아낼 수 없었을 논문이 50편에 달했다고 보고한 바 있다. 다음으로 사용된 통계기법을 소개한다. 특히 통상적으로 쓰이지 않는 통계기법을 사용하였다면 그 이유를 명확하게 설명하는 것이 좋다.

연구결과부분에서는 자료에 대한 기술통계량(descriptive statistics)과 함께 통계검정의 결과가 쓰여진다. 통계검정의 결과를 제시할 때에는 검정통계량값과 함께 p값(p-value)을 제시하는 것이 바람직하다. 다른 분야의 연구도 마찬가지이나 특히 환자의 수를 충분히 확보하기 어려운 임상시험에서는 여러 연구에서의 결과를 병합하여 어떤 결론을 내리고자하는 메타분석(META analysis)이 유용하게 쓰여진다. 메타분석에서는 정확한 p값 또는 검정통계량을 이용하여 유의수준 또는 유효크기를 병합, 검정하게 된다 (cf. 송혜향, 1992; Hedges and Olkin, 1993). 또한 이 부분에서는 여러 그룹별로 중요한 위험인자(risk factor)에 대한 정보가 제공된다. 임상시험연구의 경우 환자의 특성이나 증상(예컨대 환자의 성별이나 나이 또는 수술시의 상태)등의 요인이 연구의 결과에 뜻하지 않은 영향을 미칠 소지가 있다. 이를 방지하기 위해 층화확률화계획법등의 확률화방법을 사용하지만 이러한 경우에도 연구자들은 종종 위험인자들이 각 그룹에서 차이가 없음을 보이기 위해 통계검정을 수행한다. 만일 차이가 있는 위험인자가 발견되면 그 인자에 대해서 보정(adjustment)하는 통계분석방법을 사용되어야 할 것이다.

토론 및 결론부분에서는 유의성 검정의 결과를 해석하게 된다. 주의해야 할 점은 통계적인 유의성과 임상적인 유의성을 동일시하지 말아야 한다는 점이다. 통계적으로 유의한 결과가 반드시 임상적인 유의성을 보장하지는 않으며 반대로 유의하지 않은 결과라 해서 치료의 효과가 없다는 것을 의미하는 것은 아니다. 따라서 p값을 해석하는 데에는 어느 정도의 융통성이 필요하다. 또한 앞에서 언급했듯이 표본의 크기가 결론을 좌우하는 경우도 있을 수 있다. 다른 분야에서도 마찬가지이지만 의학논문의 결과는 특히 확대해석(extrapolaion)되어서는 곤란하다. 다시 말해서 연구의 결과를 실제 연구에서 사용되지 않은 투약용량에 적용하여 언급한다든지, 또는 연구대상환자이외의 다른 그룹의 환자들에게 일반적으로 적용하여 해석하는 일이 없어야 한다. 마지막으로 연구의 설계와 수행에 있어서의 제약점과 그것이 연구의 결과나 해석에 영향을 미쳤을 가능성에 대해 살펴보는 것이 필요하다.

위에서 기술한 의학논문의 구성과 내용은 임상연구의 결과보고서를 작성, 평가하기 위한 조언을 다룬 많은 논문들에서 공통적으로 강조되는 것들이다 (cf.Altman et al., 1983; Zelen,1983). 마지막으로 결과보고서를 작성 또는 평가할 때 반드시 점검해야 할 사항을 다음과 같이 간추려보았다.

- 어떤 실험계획법이 사용되었는가?
- 연구결과를 적용할 대상은 누구이며 연구기간은 적절하였나?
- 연구대상자들은 선정방법(확률화, 또는 비확률화)은 무엇인가?

- 대조군은 있는가? 있다면 어떻게 선정되었는가?
- 연구대상자를 포함,제외시키는 기준(eligibility criteria)은 무엇인가?
- 표본의 크기나 검정력(power)에 대한 언급이 있는가?
- 연구에서 사용된 통계적 분석기법에 대한 설명이 있는가?
- 중요한 위험인자가 그룹간에 균등하게 배치되었는가?
- 결과가 통계적으로 유의할때 그것이 임상적으로도 유의한 수준인가?
- 통계적으로 유의하지 않다면, 표본의 수는 충분한가?
- 결론은 철저히 자료에 근거한 것인가? 확대해석은 없는가?
- 연구의 제약점과 향후연구를 위한 제안이 논의되었는가?

5. 맺음말

임상시험연구의 설계, 운용 및 분석을 올바로 행하고, 아울러 보다 분명하고 객관적인 방법으로 결과보고서를 작성 또는 평가하기 위해서 통계학자들이 반드시 숙지해야 할 사항들에 대해 전반적으로 간략하게 논의하였다. 이러한 논의를 계기로 통계학자들이 의학분야에 관심을 갖고 접근함으로써 의학연구자들과의 공동연구가 활발히 이루어지기를 기대한다.

참 고 문 헌

[1] 고응린 (1977). 『계량의학통론』, 신광출판사.
[2] 김일순 (1986). 『역학적 연구방법』, 대우학술총서 자연과학 41, 민음사.
[3] 송혜향 (1992). 『메타분석법』, 자유아카데미.
[4] 송혜향, 이홍준 (1993). 『의학실험계획법』, 자유아카데미.
[5] 이재원 (1994b). 집단축차검정법들에 관한 고찰, 『응용통계연구』, 제 7 권 제 2 호, 35-51.
[6] 허명회 (1992). 『비교연구를 위한 통계적 방법론』, 자유아카데미.
[7] Altman, D. G., Gore, S.M., Gardner, M. J. and Pocock, S. J. (1983). Statistical guidlines for contributors to medical journals. *British Medical Journal*, Vol. 286, 1489-1493.
[8] Armitage, P.(1975). *Sequential Medical Trials. 2nd edition.* John Wiley and Sons, New York.
[9] Armitage, P., McPherson, C. K. and Rowe, B. C. (1969). Repeated significance tests on accumulating data, *Journal of the Royal Statistical Society,* Ser. A, Vol.132, 235-244.
[10] Brown, B.(1980). The crossover experiment for clinical trials. *Biometrics,* Vol.36, 69-79.
[11] Dawson-Saunders, B. and Trapp, R. G. (1994). *Basic and Clinical Biostatistics.* 2nd edition. Prentice-Hall International Inc.
[12] DeMets, D.(1984). Can early stopping procedures impact significantly on the efficiency of clinical trials without serious loss of information?, *Statistics in Medicine,* Vol.3, 445-451.

[13] Donner, A.(1984) Approaches to sample size estimation in the design of clinical trials - A review, *Statistics in Medicine,* Vol3.,199-214.

[14] Fleiss, J. L. (1973). *Statistical methods for rates and proportions,* John Wiley & Sons, Inc., New York.

[15] Fleiss, J. L., Tytun, A., Ury, H. K. (1980). A simple approximation for calculation sample sizes for comparing independent proportions, *Biometrics*, Vol.36,343-346.

[16] Fleming, T., Green, S. and Harrington, D. (1984). Considerations for monitoring and evaluating treatment effects in clinical trials. *Controlled Clinical Trials*, Vol.5, 55-66.

[17] Frieman, J. A. et al. (1978). The importance of beta, the Type II error and sample size in the design and interpretation of the randomized control trial, *New England Journal of Medicine*,Vol.299, 690-694.

[18] Gail, M. (1973). The determination of sample size for trials involving sevral independent 2x2 tables, *Journal of Chronic disease,* Vol.26, 669-673.

[19] Gehan, E. (1961). The determination of the number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent. *Journal of Chronic disease,* Vol.13, 346-353.

[20] Gehan, E. (1986) Randomized or historical control groups in cancer clinical trials :Are historical controls valid? *Journal of Clinical Oncology*, Vol.4, 1024-1025.

[21] Geller, N.(1987). Planned interim analysis and its role in cancer clinical trials, *Journal of Clinical Oncology*, Vol.5, 485-1490.

[22] George, S. L. and Desu, M. M.(1974). Planning the size and duration of a clinical trial studying the time to some critical event, *Journal of Chronic disease,* Vol.27, 15-24.

[23] Green, S. and Byar, D. (1984). Using observational data from registries to compare treatments: The fallicy of ominmetrics. *Statistics in Medicine* Vol.3, 361-370.

[24] Green, S., Fleming, T. and O'Fallon J.(1987). Policies for study monitoring and interim reporting of results. *Journal of Clinical Oncology*, Vol.5, 1477-1484.

[25] Hedges, L. V. and Olkin, I. (1993). *Statistical methods for meta-analysis*, Academic Press, Inc.

[26] Hsieh, F. Y. (1992). Comparing sample size formulae for trials with unbalanced allocation using the logrank test, *Statistics in Medicine* Vol.11, 1091-1098.

[27] Jennison, C. and Turnbull, B. W. (1989). Interim analysis: The repeated confidence interval approach, *Journal of the Royal Statistical Society*, Ser.B, Vol.51, 305-361.

[28] Kim, K. and Demets, D. L. (1987). Design and Analysis of group sequential tests based on the Type I error spending rate function, *Biometrika*, Vol.74, 149-154.

[29] Lan, K. K. G. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials, *Biometrika*, Vol.70, 659-663.

[30] Lee, J. W.(1994a). Group sequential testing in clinical trials with multivariate observations: A review, *Statistics in Medicine*, Vol.13, 101-111.

[31] Lee, J. W. and DeMets, D. L. (1991). Sequential comparison of changes with repeated measurements data, *Journal of the American Statistical Association*, Vol.86, 757-762.

[32] Lee, J. W. and DeMets, D. L. (1992). Sequential rank tests with repeated measurements in clinical trials, *Journal of the American Statistical Association,* Vol.87, 136-142.

[33] Louis, T., Lavori, P., Bailar, J. et al.(1984). Crossover and selfcontrolled designs in clinical research, *New England Journal of Medicine* , Vol.310, 24-31.

[34] O'Brien, P. C. and Fleming, T. R.(1979). A multiple testing procedure for clinical trials, *Biometrics*, Vol.35, 549-556.

[35] Peto, R., Pike, M., Armitage, P. et al. (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *British Journal of Cancer,* Vol.34, 585-612.

[36] Peto, R. (1978). Clinical trial methodology, *Biomedicine*, Vol.28, 24-36.

[37] Pocock, S. J.(1977). Group sequential methods in design and analysis of clinical trials, *Biometrika,* Vol.64, 191-199.

[38] Rubinstein, L. V., Gail, M. H. and Santner, T. J. (1981). Planning the duration of a comparative clinic trial with loss to follow-up and a period of continued observation, *Journal of Chronic disease,* Vol.34, 469-479.

[39] Sacks, H.,Chalmers. T. C. and Smith, H.(1982) Randomized versus historical controls for clinical trials. *American Journal of Medicine.* Vol. 72, 233-240.

[40] Sackket, D. L. (1979). Bias in analytic research, *Journal of Chronic Disease,* Vol.32, 51-63.

[41] Schlesselman, J. J. (1973a). Planning a longitudinal study: I.Sample size determination. *Journal of Chronic Disease,* Vol.26, 553-560.

[42] Schlesselman, J. J. (1973b). Planning a longitudinal study: II. Frequency of mesurement and study duration. *Journal of Chronic Disease,* Vol.26, 561-570.

[43] Schoenfeld, D. A. (1980). Statistical consideration for pilot studies. *International Journal of Radiation Oncology, Biology and Physics*, Vol.6, 371-374.

[44] Tebbi, C., Cummings, K., Zevon, M. et al.(1983). Compliance of pediatric and adolescent cancer patients. Cancer, Vol.58, 1179-1184.

[45] Von Hoff, D. D., Kuhn, J. and Clark, G. M. (1984). Design and conduct of phase I trials, in Buyse, M. E., Staquet, M. D., Sylvester, R. D.(editors), *Cancer Clinical Trials:Methods and Practice.* Oxford University Press, 1984, 210-220.

[46] Wald, A. (1947). *Sequential Analysis*, John Wiley & Sons, New York.

[47] Williamson, J. W., Goldschmidt, P.G. and Colton, T. (1986). The quality of medical literature : An analysis of validation assessments. In: *Medical Uses of Statistics.* Bailar JC,Mosteller F(editors). Massachusetts Medical society.

[48] Xiang, A. H. and Lee, J. W. (1994). Sample size estimation in the design of randomized clinical trials:A review. Technical Report, Department of Preventive Medicine, University of Southern California.

[49] Zelen, M. (1969). Play the winner rule and the controlled clinical trial. *Journal of the American Statistical Association,* Vol. 64, 131-146.

[50] Zelen, M. (1974). The randomization and stratification of patients to clinical trials. *Journal*

of Chronic Disease, Vol.27, 365-375.

[51] Zelen, M.(1979). A new design for randomized clinical trials, *New England Journal of Medicine,* Vol.300, 1242-1245.

[52] Zelen, M, (1983). Guidelines for publishing papers on cancer clinical trials: Responsibilities of editors and authors. *Journal of Clinical Oncology*, Vol. 1, 164-169.