청렴韓세상

다중 평가변수를 사용하는 임상시험 가이드라인(안) [민원인 안내서]

2024. 4. 29.

식품의약품안전처

식품의약품안전평가원

의약품심사부 순환신경계약품과, 제품화지원팀

지침서·안내서 제·개정 점검표

명칭	다중 평가변수를 사용하는 임상시험 가이드라인(안)

아래에 해당하는 사항에 체크하여 주시기 바랍니다.

등록대상 여부	□ 이미 등록된 지침서·안내서 중 동일·유사한 내용의 지침서·안내서가 있습니까?	□ 예 ■ 아니오
	☞ 상기 질문에 '예'라고 답하신 경우 기존의 지침서·안내서의 개정을 우선적으로 고려하시기 바랍니다. 그럼에도 불구하고 동 지침서·안내서의 제정이 필요한 경우 그 사유를 아래에 기재해 주시기 바랍니다. (사유 :)	
	□ 법령(법·시행령·시행규칙) 또는 행정규칙(고시·훈령·예규)의 내용을 단순 편집 또는 나열한 것입니까?	□ 예 ■ 아니오
	□ 단순한 사실을 대외적으로 알리는 공고의 내용입니까?	□ 예 ■ 아니오
	□ 일회성 지시·명령에 해당하는 내용입니까?	□ 예 ■ 아니오
	□ 외국 규정을 단순 번역하거나 설명하는 내용입니까?	□ 예 ■ 아니오
	□ 신규 직원 교육을 위해 법령 또는 행정규칙을 알기 쉽게 정리한 자료입니까?	□ 예 ■ 아니오
	☞ 상기 사항 중 어느 하나라도 '예'에 해당되는 경우에 지침서·안내서 등록 대상이 아닙니다. 지침서·안내서 제·개정 절차를 적용하실 필요는 없습니다.	
지침서·안내서 구분	□ 행정사무의 통일을 기하기 위하여 내부적으로 행정사무의 세부기준이나 절차를 제시하는 것입니까? (공무원용)	□ 예(☞지침서) ■ 아니오
	□ 민원인들의 이해를 돕기 위하여 법령 또는 행정규칙을 알기 쉽게 설명하거나 특정 민원업무에 대한 행정기관의 대외적인 입장을 기술하는 것입니까? (민원인용)	■ 예(☞안내서) □ 아니오
기타 확인 사항	□ 상위 법령을 일탈하여 새로운 규제를 신설·강화하거나 민원인을 구속하는 내용이 있습니까?	□ 예 ■ 아니오
	☞ 상기 질문에 '예'라고 답하신 경우 상위법령 일탈 내용을 삭제하시고 지침서·안내서 제·개정 절차를 진행하시기 바랍니다.	

상기 사항에 대하여 확인하였음.

2024 년 4 월 29 일

담당자 김 송 이

확 인(부서장) 김 소 희

이 안내서는 다중 평가변수를 사용하는 임상시험에 대하여 알기 쉽게 설명하거나 식품의약품안전처의 입장을 기술한 것입니다.

본 안내서는 대외적으로 법적 효력을 가지는 것이 아니므로 본문의 기술방식('~하여야 한다' 등)에도 불구하고 참고로만 활용하시기 바랍니다. 또한, 본 안내서는 2024년 4월 29일 현재의 과학적·기술적 사실 및 유효한 법규를 토대로 작성되었으므로 이후 최신 개정 법규 내용 및 구체적인 사실관계 등에 따라 달리 적용될 수 있음을 알려드립니다.

※ "민원인 안내서"란 민원인들의 이해를 돕기 위하여 법령 또는 행정규칙을 알기 쉽게 설명하거나 특정 민원업무에 대한 행정기관의 대외적인 입장을 기술하는 것(식품의약품안전처 지침서등의 관리에 관한 규정 제2조)

※ 본 안내서에 대한 의견이나 문의사항이 있을 경우 식품의약품안전처 식품의약품안전평가원 의약품심사부 순환신경계약품과에 문의하시기 바랍니다.

전화번호: 043-719-3004, 3018
팩스번호: 043-719-3000

목 차

제·개정 이력

연번	제·개정번호	승인일자	주요내용
1			(안) 제정

다중 평가변수를 사용하는 임상시험 가이드라인(안)

1. 서론

의약품 개발을 위한 임상시험에서 다중 평가변수를 사용하는 경우가 많다. 다중 평가변수는 의약품의 치료효과를 평가하고 질병의 여러 특징에 미치는 의약품의 영향을 확인하기 위한 목적으로 주로 설정된다. 다만, 임상시험에서 하나 이상의 평가변수를 분석할 때 다중성(multiplicity)이 적절하게 보정되지 않는다면 해당 평가변수와 관련된 의약품의 효과에 대해 잘못된 결론을 도출할 가능성이 높아질 수 있다. 그러므로 다중 평가변수를 사용하는 임상시험을 계획, 결과 분석 및 해석하는 과정에서 다중성 문제를 고려하는 것은 매우 중요하다.

본 가이드라인은 제약업계 및 임상시험 관련 종사자들에게 임상시험 결과 분석 및 해석 과정에서 다중 평가변수로 인해 발생할 수 있는 문제점과 이러한 문제점을 해결하는 방법 등 다중 평가변수를 사용하는 임상시험에서 일반적으로 고려해야 할 사항을 제공하는 것을 목적으로 한다. 즉, 의약품의 치료효과를 분석하기 위해 다수의 평가변수들을 범주화하고 순서를 정하는 방법과 잘못된 결론을 도출할 가능성을 통제하기 위해 임상시험 내에서 다중성을 관리하는 통계학적 방법 등 다양한 전략을 안내하고자 한다.

2. 배경 및 적용 범위

유효성 평가변수는 의약품의 목표 치료효과를 확인할 수 있도록 설계되는 측도이다. 유효성 평가변수에는 임상적 사건(예: 사망, 뇌졸중, 폐 악화, 정맥 혈전색전증), 증상(예: 통증, 호흡곤란, 우울증 증상), 특정한 기능(예: 걷거나 운동하는 능력), 또는 최종 임상적 결과변수(clinical outcome)와 밀접한 관련이 있다고 입증된 대리 평가변수(surrogate endpoint) 등이 포함된다.

대부분의 질병의 경우, 하나 이상의 임상적 사건, 증상 및/또는 기능 이상을 유발할

가능성이 있으므로 대상 질환의 하나 이상의 측면에 대해 의약품의 치료효과를 확인할 수 있도록 임상시험을 설계하는 경우가 많다. 질병의 한 가지 측면만을 기반으로 의약품의 유효성을 확립하기 어려운 경우에는 질병의 여러 측면을 반영하는 단일 평가변수를 사용하거나 다수의 평가변수를 통해 치료효과를 입증해야 한다. 이와 달리, 여러 평가변수 중 어느 하나에 대한 치료효과가 품목허가를 뒷받침하기에 충분한 경우도 있을 수 있다.

임상시험에서 평가된 여러 평가변수에 대해 다중성을 충분하게 고려하지 못하면 의약품의 치료효과에 대하여 잘못된 결론을 내릴 가능성이 커진다. 다중성 문제는 의약품의 품목허가를 뒷받침하기 위한 핵심 임상시험을 평가할 때 주로 발생하지만, 의약품 전체 개발과정에서 수행되는 다른 임상시험에서도 중요하게 고려되어야 한다. 예를 들어, 안전성 결과변수가 가설 검정을 통해 평가되는 경우, 본 가이드라인에서 설명하는 다중성 관련 고려사항이 적용된다. 다만, 안전성 분석이 공식적인 통계 검정을 위해 사전에 계획된 가설의 일부로 설정되지 않은 경우, 안전성 분석에 대한 다중성 문제는 본 가이드라인에서 다루는 범위에 포함되지 않는다.

다음 항에서는 다중 평가변수의 문제점과 이를 해결하는 방법에 대해 안내하고자 하며, 본 가이드라인에서는 단일 임상시험 내에서 다중 평가변수의 분석 및 해석을 중점적으로 기술하고 있다.

2.1. 시험 목적(유효성)의 입증

의약품 품목허가를 뒷받침하기 위해서는 적절하고 잘 통제된 임상시험에서 의약품의 목표 치료효과를 입증하는 것이 필수적이다. 적절하고 잘 통제된 임상시험에서는 약물의 치료효과를 평가하기에 적합하도록 시험 결과를 분석하는 것이 무엇보다도 중요하며, 이는 치료효과에 영향을 줄 수 있는 다른 요인[예: 시간 경과에 따른 질병의 자연발생적 변화, 위약 효과, 측정치의 편향(bias) 등]으로부터 약물의 치료효과를 구분하기 위해 의약품의 안전성·유효성 자료를 수집하는 것을 목적으로 한다. 또한 임상시험에서 유효성에 대한 실질적 근거를 평가할 때 가설 검정 결과 외에 다른 요인들(예: 평가변수와 예상 효과의 임상적 관련성, 관련 외부 정보)도 중요하게 고려되어야 한다.

가설 검정은 설정한 평가변수에 대한 치료효과 평가 시 불확실성을 다루기 위하여 흔히 사용된다. 가설 검정은 설정된 평가변수에 대한 가설을 기술하는 것으로 시작된다. 임상시험의 목적이 단순히 시험약이 대조약보다 우월함을 입증하는 것인 경우, 임상시험을 실시하기 전에 평가변수에 대하여 상호 배타적인 두 가지 가설을 설정한다.

- 귀무가설 : 시험약과 대조약 간에 치료효과의 차이가 없다.
- 대립가설 : 시험약과 대조약 간에 치료효과의 차이가 있다

  (시험약에 적어도 일부 치료효과가 있다).

임상시험 결과가 귀무가설이 참이라는 가정하에서 발생할 가능성이 충분히 낮아서 귀무가설을 기각하고 대립가설을 지지할 수 있는지를 결정하기 위해 이 가설들을 사전에 정한 통계분석 방법에 따라 검정한다. 귀무가설이 기각되지 않는다고 하더라도 귀무가설이 반드시 참이라는 것을 의미하지는 않는다. 목표 시험대상자 수가 충분하지 않은 경우 등 귀무가설이 기각될 수 없는 다양한 가능성이 존재한다.

일부 경우(예: 백신 임상시험), 품목허가를 위해 최소한의 효과 크기에 대한 입증이 필요할 수 있다. 이 경우 유효성 입증을 위해 공식적인 통계 검정이 필요하다면 귀무가설은 임상적으로 의미있는 최소한의 효과를 고려하여 설정될 수도 있다.

본 가이드라인에서는 가설 검정에 기반한 통계학적인 접근방식을 중점적으로 다룬다. 특정 의약품 개발을 위하여 다른 접근방식[예: 베이지안(Bayesian) 접근법 등]을 사용하고자 하는 경우에는 개발 초기부터 규제기관과 논의하는 것이 바람직하다.

2.2. 제1종 오류

임상시험에서 귀무가설이 기각되면 치료군 간 차이가 있다는 결론을 내릴 수 있지만 귀무가설이 절대적으로 거짓이라는 것을 의미하지는 않는다. 귀무가설이 참임에도 불구하고 귀무가설을 잘못 기각할 가능성은 항상 존재한다. 이러한 잘못된 결론을 도출할 가능성을 제1종 오류라고 한다. 사실은 시험약에 의한 치료효과가 존재하지

않는데 귀무가설을 잘못 기각하여 평가변수에 대하여 치료효과가 존재한다고 결론을 내리는 가능성을 해당 평가변수에 대한 제1종 오류 확률(Type I error probability) 또는 제1종 오류율(Type I error rate)이라고 한다. 알파(α)로 표시되는 유의수준은 제1종 오류를 범할 확률을 통제하는 임계값이다. 귀무가설의 기각은 귀무가설이 참이라고 가정할 때 시험의 결과와 동일하거나 더 극단적인 결과를 관찰할 확률이 충분히 낮다(일반적으로 p값이 α보다 크지 않은 경우)는 결정을 기반으로 한다.

대립가설은 단측 또는 양측이 될 수 있으며 그에 따라 통계 검정이 수행된다. 가설에 대한 양측검정 시 제1종 오류 확률은 실제로 시험약과 대조약 간에 차이가 없지만 차이가 있다(시험약의 효과가 더 좋거나 나쁨)고 결론을 내릴 확률을 의미한다. 가설에 대한 단측검정 시 제1종 오류 확률은 실제로 시험약과 대조약 간에 차이가 없지만 시험약이 대조약보다 유익하다고 구체적으로 결론을 내릴 확률을 의미한다. 가장 널리 사용되는 유의수준(α)은 양측검정의 경우 0.05, 단측검정의 경우 0.025이다. 양측검정에서 유의수준(α) 0.05는 차이가 실제 존재하지 않는데 시험약이 대조약 대비 어느 한쪽 방향으로든 차이가 있다(시험약의 효과가 더 좋거나 나쁨)고 잘못된 결론을 도출할 확률이 5%를 넘지 않거나 20번 중 한 번인 경우를 의미한다. 단측검정에서 유의수준(α) 0.025는 시험약의 유익한 효과가 실제 존재하지 않는데 시험약이 유익한 효과가 있다고 잘못된 결론을 도출할 확률이 2.5%를 넘지 않거나 40번 중 한 번인 경우를 의미한다. 각 측면에 대칭적으로 유의수준을 배정하는 양측검정[유의수준(α) 0.05]을 사용하면 군 간 차이가 실제 존재하지 않을 때 효과가 있다고 잘못된 결론을 도출할 가능성이 약 2.5%를 넘지 않는다(40번 중 한 번). 통계 검정 방법이 적절하다면 이러한 제1종 오류율은 정확하다. 통계 검정에 문제가 있다면(예: 기본 가정이 유지되지 않는 경우 등), 제1종 오류율은 훨씬 더 커질 수도 있다.

규제기관에서는 제1종 오류 확률 통제와 관련하여 임상시험에서 일차 또는 이차 평가변수에 대해 잘못된 긍정적 결론이 도출될 가능성을 최소화하는 것을 중요하게 생각하며 이는 임상시험에서 효과가 확인되지 않은 평가변수의 종류나 개수에 관계없이 중요하다. 다중 평가변수를 검정할 때 발생할 수 있는 제1종 오류 확률을 전체 제1종 오류 확률(overall Type I error probability)이라고 한다. 이러한 확률을 통제하는 사유는 2.3.항을 참조한다. 하나 이상의 일차 또는 이차 평가변수에 대해 여러 가설을 평가할 때 전체 제1종 오류 확률(또는 오류율)이 사전에 계획된 수준에 비해 더 커지지 않도록 해야 한다. 임상시험 의뢰자는 제1종 오류율을 통제하기

위하여 다음을 사전에 기술해야 한다.

- 모든 일차 및 이차 평가변수 집단(3.1.항 참조)
- 일차 또는 이차 평가변수에 관계없이 사전에 지정된 평가변수에 대한 가설을 검정하기 위해 수행하는 모든 자료 분석

다중 평가변수를 사용하는 임상시험의 분석 계획에는 전체 제1종 오류율을 적절하게 통제하여 검정하고자 하는 가설들에 대한 검정 절차가 기술되어야 한다.

2.3. 다중성

하나의 평가변수에 대해 유의수준(α) 0.05로 양측검정하는 임상시험에서 모집단에 대한 시험약 및 대조약 간 차이가 실제로 존재하지 않는데 시험약에 유리한 군간 차이를 발견할 확률은 0.025(2.5% 가능성)이다. 즉, 해당 평가변수에 대하여 치료효과가 실제로 존재하지 않으면 유리한 효과를 발견하지 못할 가능성은 약 97.5%이다. 반대로, 2개의 독립적인 평가변수에 대하여 각각 유의수준(α) 0.05로 양측으로 검정하여 그중 하나의 평가변수만 유의해도 그 자체로 시험약이 효과가 있다고 인정하는 경우, 두 평가변수 모두에서 효과를 적절하게 확인하지 못할 확률은 0.975 × 0.975, 즉, 약 0.95이다. 적어도 하나의 평가변수에서 효과가 있다고 잘못된 결론을 내릴 가능성은 약 0.05이므로 전체 제1종 오류율은 2개의 독립적인 평가변수를 검정할 때 거의 2배가 된다. 이와 같이 다중성을 보정하지 않고 다중 검정을 진행할 때 전체 제1종 오류율이 의도한 것보다 높아지는 것을 다중성 문제라고 한다. 위의 사례와 같이 다중성을 보정하지 않으면 전체적으로 제1종 오류를 범할 가능성이 시험약에 유리하게 약 5%까지 높아질 수 있어 전체 제1종 오류율이 적절하게 통제되지 않는다. 2개가 넘는 평가변수를 고려할 때 다중성 문제는 더 심각해질 수 있다. 예를 들어, 3개의 독립적인 평가변수에 대한 제1종 오류율은 1 - (0.975 × 0.975 × 0.975)로 약 7%이다. 10개의 독립적인 평가변수에 대한 제1종 오류율은 약 22%이다. 다중 평가변수가 서로 관련성이 있는 경우, 전체 제1종 오류율이 더 증가될 수도 있지만 그 정도가 덜할 수 있다.

하나의 결과변수에 대해 평가하는 경우에도 분석 시 다중성 문제가 발생하여 제1종

오류율이 증가될 수 있다. 이러한 경우에는 해당 결과변수의 여러 측면(예: 여러 용량군, 여러 평가 시점, 인구통계학적 또는 다른 특성에 기반한 여러 하위군)을 분석하고 여러 분석 결과 중 어느 하나의 결과만을 기반으로 약물이 효과가 있다고 결론을 내리는 경우가 해당한다. 따라서 제1종 오류율이 증가함에 따라 다중성 문제는 시험 결과 해석에 불확실성을 초래하여 약물이 효과가 있다는 결론을 신뢰하기 어렵게 만든다. 전체 제1종 오류율을 2.5% 이하로 유지하기 위하여 전향적으로 계획하고 적용할 수 있는 다양한 접근방식이 있다.

다중성을 통제하기 위한 중요한 원칙은 먼저 평가변수, 평가 시점, 분석군, 용량 및 분석 방법 모두를 사전에 임상시험계획서에 명기하는 것이다. 일단 이 사항들이 정해지고 나면, 다중 평가변수를 적절하게 보정하고 분석하는 방법을 선택하여 임상시험계획서에 사전에 기술하고 적용할 수 있다. 추가적인 분석을 진행하려면 자료 분석 및 다중성 보정을 수행하기 이전에 분석 계획을 변경해야 한다. 그렇지 않으면 다중성 문제를 일으켜서 임상시험 결과 해석에 부정적인 영향을 미칠 수 있다. 통계분석 계획은 치료군 배정에 대한 맹검이 해제되고 통계분석이 진행된 이후에는 변경되어서는 안 된다.

본 가이드라인은 임상시험의 주요 결과들을 잘 뒷받침하고 시험약의 치료효과를 입증할 수 있도록 사전에 정한 평가변수들(즉, 일차 및 이차 평가변수)에 대한 제1종 오류율을 통제하는 데 초점을 둔다. 임상시험에서 시험약의 치료효과를 입증한 이후에 해당 평가변수에 대한 치료효과 특성을 설명할 수 있는 여러 분석 결과(약효 발현 시점, 대상 집단 내 효과 크기 분포, 하위군에서의 효과, 복합 평가변수의 개별 구성요소에 대한 효과 포함)를 제시할 수 있다. 이러한 분석은 시험약의 치료효과를 좀 더 세부적으로 이해하는 데 도움을 줄 수 있지만, 분석 결과는 모두 기술통계로만 제시하고 평가변수 범위 외로 효과를 확대하여 해석하지 않는다. 이 기술적 분석 자료는 의약품의 허가사항에 p값 제시 없이 기재될 수 있다.

임상시험의 주요 결과와 밀접하게 관련된 분석과 의약품의 추가 효과를 입증하는 분석은 항상 명확하게 구별되지는 않는다. 그러므로 명확한 결론을 도출하고자 한다면 이러한 분석은 임상시험계획서에 사전에 기술되어야 하며 다중 검정 전략에 적절하게 포함되어야 한다. 임상시험계획서의 다중 검정 전략에 사전에 포함되지 않은 기술적 분석을 통계적으로 확증된 결론을 암시하거나 임상시험에서 입증되지 않은 효과에

대하여 확신을 주는 방식으로 허가사항에 제시해서는 안 된다. 기술적 분석은 본 가이드라인의 적용 범위에 해당하지 않으므로 자세히 다루지 않는다.

3. 다중 평가변수 설정 시 일반적 고려사항

3.1. 평가변수 집단의 계층적 접근

적절하고 잘 통제된 임상시험에서 평가변수는 대개 임상적인 중요도에 따라 계층적으로 범주화되지만, 평가변수에 해당하는 사건의 예상되는 빈도 및 약물 효과도 함께 고려된다. 평가변수를 범주화할 때는 의약품 품목허가를 뒷받침하기 위해 유효성을 확립하기 위한 목적인지 또는 의미 있는 추가 효과를 입증하기 위한 목적인지 여부를 중요하게 고려하여 결정한다. 의약품 품목허가를 뒷받침하기 위해 유효성을 확립하는 데에 결정적인 평가변수는 주로 일차 평가변수로 설정된다. 이차 평가변수는 일차 평가변수를 뒷받침하고/하거나 임상적으로 중요한 추가 효과를 입증하는 데에 유용할 수 있다. 계층화된 세 번째 범주에는 다른 모든 평가변수가 포함되며 탐색적 평가변수로 간주된다. 탐색적 평가변수에는 학술적 목적 또는 새로운 가설 생성을 위한 목적의 평가변수도 포함될 수 있다. 계층화된 각 범주에는 단일 평가변수 또는 평가변수 집단이 포함될 수 있다.

3.1.1. 일차 평가변수 집단

약물의 유효성을 확립하고 임상시험이 목적을 달성했다는 결론을 내리는 데 근거가 되는 평가변수들은 일차 평가변수 집단으로 지정된다. 사전에 계획된 하나의 일차 평가변수를 통해 임상시험의 목적을 달성했는지 여부를 판단할 때는 다수의 평가변수 관련 다중성 문제는 발생하지 않는다.

다중 일차 평가변수는 세 가지 경우가 있을 수 있다(3.3.항 참조). 첫 번째는 일차 평가변수가 여러 개이고 각 평가변수만으로 약물의 유효성을 충분히 입증할 수 있는 경우이다. 이러한 다중 평가변수는 임상시험의 성공 여부를 여러 번 확인할 가능성이 있으므로 다중성을 보정하지 않으면 제1종 오류율이 증가하고 약물이 효과가 있다는 잘못된 결론을 도출할 수 있다. 두 번째는 일차 평가변수가 2개 이상 존재하고 일차 평가변수 모두에서 유의하면 약물이 효과가 있다고 결정하는 경우이다. 이러한 경우는

임상시험이 성공적인 결과로 이어지는 경로가 단 하나뿐이기 때문에 일차 평가변수와 관련된 다중성 문제는 발생하지 않으며 제1종 오류율이 증가할 우려가 없다. 세 번째 경우는 유효성과 관련된 중요 양상들을 단 하나의 복합 평가변수 또는 다른 여러 구성요소로 이루어진 복합 평가변수로 통합하여 평가함으로써 다중 평가변수로 인해 발생하는 다중성 문제를 피하게 되는 경우이다. 예를 들어, 많은 심혈관계 임상시험에서 여러 평가변수를 단 하나의 일차 복합 평가변수(예: 심혈관계 사망, 심근경색 및 뇌졸중)로 통합하여 평가하고 사망은 이차 평가변수로 평가하는 것이 일반적이다.

3.1.2. 이차 및 탐색적 평가변수 집단

일차 평가변수에 대한 치료효과가 입증된 후에는 이차 평가변수를 공식적으로 검정할 수 있다. 이때 이차 평가변수는 일차 평가변수와 관련된 임상적 효과로 해당 효과에 대하여 더 잘 이해할 수 있도록 보조적인 역할을 하거나(예: 심혈관계 약물이 일차 평가변수인 심부전으로 인한 입원에 대한 효과를 입증한 이후 이차 평가변수로 생존율에 대한 효과를 평가하는 경우), 일차 평가변수에서 나타난 효과와 구별되는 임상적 유익성을 뒷받침하는 근거(예: 다발성 경화증 치료제 임상시험에서 재발률에 대한 효과를 입증한 후 기능 이상에 대해 효과를 평가하는 경우)를 제공할 수도 있다. 일반적으로 대상 질환이나 상태에 대한 의약품의 추가적 효과를 확인하고자 한다면 이차 평가변수를 제1종 오류를 통제하는 계획에 포함하는 것이 중요하다.

일반적으로 이차 평가변수의 개수를 제한하는 것이 바람직할 수 있다. 다중성을 보정할 때 이차 평가변수 개수가 증가함에 따라 해당 평가변수에 대한 치료효과를 입증할 가능성이 점점 작아지거나 계층적으로 가설을 검정할 때 순서상 나중에 검정하는 중요한 가설들이 전혀 검정되지 않을 수 있기 때문이다.

탐색적 평가변수는 일반적으로 결론을 뒷받침하는 데 사용되지 않으므로 다중성 보정은 필요하지 않다.

3.1.3. 일차 및 이차 평가변수 집단에서 평가변수 설정 및 해석

이차 평가변수에서 나타나는 긍정적 결과는 일차 평가변수 집단에 대한 치료효과가

먼저 입증되는 경우에 한하여 해석될 수 있다. 전체 제1종 오류율은 일차 및 이차 평가변수 집단 모두에서 함께 통제되어야 한다.

임상시험에서 임상적으로 중요한 평가변수들(예: 사망률 또는 비가역적 이환율)의 발생 빈도가 너무 낮아서 적절한 검정력을 제공할 수 없을 것으로 예상되는 경우가 있다. 반면, 다른 임상적으로 중요한 평가변수가 더 자주 발생하거나 질병 진행 초기에 발생하여 검정력을 더 높이는 경우도 때때로 있다. 일반적으로 빈도가 낮아 충분한 검정력을 확보하기 어려운 평가변수는 이차 평가변수로, 검정력이 더 클 것으로 예상하는 평가변수는 일차 평가변수로 분류한다. 예를 들어, 일부 항암제 임상시험에서 무진행 생존시간(progression-free survival)을 일차 평가변수로 설정하고 전체 생존시간은 이차 평가변수로 설정하기도 한다. 이는 질병 진행에 미치는 치료효과가 임상적으로 중요하고 입증하기 좀 더 쉬우며 조기에 확인될 수 있기 때문이다. 또한 치료효과가 좀 더 크게 나타날 수도 있는데, 이는 전체 생존시간에 대해 관찰된 효과가 질병이 진행된 이후 후속 치료에 의해 영향을 받을 수 있기 때문이다.

3.2. 제2종 오류 및 시험대상자 수

규제기관은 약물이 실제 효과가 있는데 효과를 입증하는 데 실패할 확률인 제2종 오류(Type Ⅱ error)가 나타날 위험성도 중요하게 고려한다. 임상시험의 검정력이란 치료효과가 특정 크기로 실제 존재할 때 임상시험이 성공할 확률을 의미한다. 검정력은 특히 일차 평가변수를 입증하기 위해 필요한 시험대상자 수를 결정할 때 고려하는 중요한 요소이다.

임상시험에서 목표한 시험대상자 수는 일반적으로 일차 평가변수에 대한 치료효과가 특정 크기로 실제 존재할 때 그 치료효과를 입증할 만큼 충분하게 큰 검정력을 제공할 수 있도록 산출되어야 한다. 시험대상자 수 산출 시 다중성에 대한 제1종 오류율을 통제하기 위하여 통계적인 보정을 고려해야할 수도 있다. 예를 들어, 임상시험 평가변수에 대한 유의수준(α)을 좀 더 작게 설정한다면 원하는 통계적 검정력을 제공할 수 있도록 시험대상자 수를 조정해야 한다.

치료효과를 입증하기 위한 평가변수가 2개 이상 존재하고 개별 평가변수에 대한 유의성 입증으로 품목허가를 충분히 뒷받침할 수 있는 경우(공동 평가변수, 3.3.1.항 참조),

제2종 오류율이 증가하고 임상시험의 검정력이 감소한다. 예를 들어, 2개의 평가변수의 효과 크기가 동일하고 개개 평가변수의 유의성을 입증하기 위하여 80%의 검정력을 확보하도록 시험대상자 수를 산출한다고 가정해 보자. 평가변수들이 독립적이라면 평가변수 모두에서 유의성을 확인할 수 있는 검정력은 약 64%(0.8 × 0.8)가 된다. 즉, 약물의 효과가 실제 있는데 효과가 있다는 결론을 뒷받침하지 못하고 실패할 가능성(제2종 오류율)은 36%가 될 것이다. 임상시험의 검정력을 원하는 수준으로 유지하기 위하여 목표 시험대상자 수를 더 크게 설정하는 것이 권장되며 임상시험의 성공 가능성을 적어도 80% 이상 확보하기 위해서 개개 평가변수의 검정력은 약 90%로 설정할 수 있다. 평가변수들이 높은 양(positive)의 상관관계를 갖거나 각 평가변수의 검정력이 동일하지 않다면 다르게 산출될 수 있다.

3.3. 다중 평가변수의 유형

약물의 임상적 유익성을 판단할 때 하나 이상의 질병 양상 또는 결과변수에 대해 시험약의 효과를 입증하는 것이 중요하다면 다중 평가변수가 사용될 수 있다. 다중 평가변수는 다음 경우에도 사용할 수 있다. (1) 질병에 여러 가지 중요한 양상이 있거나 하나의 중요한 양상을 평가하는 방법이 여러 가지인 경우, (2) 어떤 양상이 약물 효과를 나타낼 가능성이 더 높은지 미리 알 수 없는 경우, (3) 어느 하나의 평가변수에 대한 치료효과만으로 품목허가를 뒷받침하기에 충분한 경우이다. 어떤 경우에는 질병의 여러 양상을 하나의 평가변수로 적절하게 통합할 수 있으나 약물의 효과를 잘 이해하기 위하여 평가변수의 각 구성요소나 질병 양상을 확인하는 후속 분석이 중요하다. 이러한 상황은 아래에서 더 자세히 설명하기로 한다.

3.3.1. 2개 이상의 평가변수에 대해 치료효과가 입증된 경우에만 시험약의 효과가 있다고 인정하는 경우(공동 일차 평가변수)

어떤 질병의 경우에는 임상적으로 중요하지만 서로 다른 두 가지 이상의 특징이 있고 이러한 질병의 모든 특징에 대해 유의성이 입증되지 않으면 시험약이 효과가 있다고 인정하지 않는 경우가 있다. 이러한 상황에서의 다중 평가변수는 공동 일차 평가변수(co-primary)라는 용어를 사용한다. 일차 평가변수가 2개 이상이고 모든

일차 유효성 평가변수가 유의한 경우에만 시험약의 효과가 있다고 인정하는 경우, 이러한 다중 평가변수는 공동 일차 평가변수가 된다.

대표적인 예로 편두통의 급성 치료제가 있다. 편두통의 가장 큰 특징은 통증이지만 눈부심(photophobia), 소리공포증(phonophobia), 및/또는 오심(nausea)도 주요 증상이며 임상적으로 모두 중요하다. 세 가지 중 어떤 증상이 임상적으로 가장 중요한지는 사람마다 다를 수 있다. 편두통의 급성 치료제에 대한 임상시험에서 약물 투여 후 2시간 시점에 두통이 나타나지 않는 시험대상자의 비율 및 가장 괴로운 증상이 나타나지 않는 시험대상자의 비율 모두에서 시험약에 의해 개선된 경우에 한하여 그 시험약이 편두통의 급성 치료에 효과적이라고 인정하는 경우가 있다. 또는 하나의 반응 평가변수에 대해 시험약의 치료효과를 평가하는 것도 한 가지 방법일 수 있다. 이때 반응이란 개개의 시험대상자 내에서 통증과 각 시험대상자별로 지정된 다른 증상이 모두 나타나지 않는 경우를 말한다. 이러한 접근 방법은 공동 일차 평가변수가 아닌 여러 구성요소로 이루어진 단일 평가변수로 분류된다.

공동 평가변수를 적용하는 대표적인 예시로 혼합백신이 있다. 혼합백신 임상시험은 일반적으로 백신을 통해 예방하고자 하는 개개의 병원체에 대한 유효성 평가변수 각각에 대하여 유의한 결과를 입증할 수 있도록 설계되고 충분한 검정력을 가져야 한다.

3.2.항에 기술한 바와 같이 개개 평가변수 모두에 대해 유의성을 입증하도록 임상시험이 설계되는 경우, 다중성 문제는 발생하지 않는다. 그러나 공동 일차 평가변수에 대한 검정으로 인해 제2종 오류율이 증가할 수 있다. 일반적으로 임상적으로 반드시 필요한 경우가 아니라면 3개 이상의 공동 일차 평가변수 설정은 검정력이 떨어질 수 있으므로 신중하게 결정한다.

공동 평가변수 모두에서 유의수준(α) 0.05로 유의성 입증이 필요할 때, 발생하는 통계적인 검정력 손실을 보전하기 위하여 각 공동 평가변수의 통계적 검정 기준을 높여달라는 요청이 있었다[예: 유의수준(α) 0.06 또는 0.07에서 검정]. 그러나 각 공동 일차 평가변수의 유의수준(α) 증가는 허용되지 않는다. 유의수준을 증가시키면 품목허가를 뒷받침하기 위해 시험약의 효과를 입증하는 데에 중요하다고 간주되는 각 질병 양상에 대한 치료효과를 해석하는 데에 제한이 있을 수 있다.

3.3.2. 여러 일차 평가변수 중 최소 하나 이상의 평가변수에 대한 치료효과 입증으로 충분한 경우

많은 질병에는 여러 양상이 있을 수 있는데 이 중 어느 하나에서 유의성이 입증되면 효과가 있다고 인정할 수 있다. 시험약의 반응이 나타나는 질병의 양상이나 치료효과를 더 잘 감지할 수 있는 평가 방법(임상시험 설계 당시에)이 사전에 알려져 있지 않은 경우에는 일차 평가변수를 단 하나만 설정하는 것이 어려울 수 있다. 이러한 상황에서는 여러 평가변수 중 어느 하나에서 유의성이 입증되면 효과가 있다고 결론을 내릴 수 있도록 임상시험이 설계될 수 있다. 이렇게 하면 일차 평가변수 집단이 생성된다. 예를 들어, 화상 상처 치료제의 경우 시험약이 상처 봉합 비율을 높여주는지 또는 흉터를 감소시키는지 알려지지 않았지만, 이 중 하나에 효과만 있어도 임상적으로 중요하다고 가정해 보자. 이런 경우 상처 봉합 비율과 흉터 크기는 모두 별개의 일차 평가변수로 설정될 수 있다.

이렇게 다중 평가변수를 사용하는 경우, 치료효과를 성공적으로 입증할 수 있는 여러 방법이 있을 수 있어 다중성 문제가 발생할 수 있다. 일차 평가변수 집단에 대한 제1종 오류율을 통제하는 것이 매우 중요하다. 이 다중성 문제를 해결하기 위해 다양한 접근방식이 사용될 수 있는데, 이러한 접근방식 중 몇 가지를 부록에서 설명한다.

3.3.3. 복합 평가변수

어떤 질병의 경우에는 임상시험에서 여러 임상적 결과가 중요하고 모든 임상적 결과가 시험약에 의해 영향을 받는다고 예상되는 경우가 있다. 이때 각각의 임상적 결과를 별도의 일차 평가변수(다중성 문제가 발생할 수 있음)로 설정하거나 일차 평가변수를 단 하나만 설정하고 다른 평가변수는 이차 평가변수로 설정하는 것보다는 이러한 임상적 결과들을 하나의 변수로 통합하여 평가하는 것이 적절할 수 있다. 이를 흔히 복합 평가변수라고 하며, 이는 사전에 명시된 구성요소 중 어느 하나가 시험대상자에서 발생하거나 실현되는 것으로 정의된다. 대표적인 예로 심혈관계 임상시험에서 주요 임상적 사건에 대한 복합 평가변수가 있다(예: 심근경색, 뇌졸중, 또는 사망의 복합 평가변수). 각 구성요소들이 특정 사건들에 해당하는 경우, 복합 평가변수는 각 구성요소 중 어느 하나라도 처음 발생할 때까지의 시간으로 평가된다.

해당 복합 평가변수에 대해 단 한 번의 통계적 검정이 수행된다면 다중성 문제는 발생하지 않는다.

복합 평가변수를 사용하는 중요한 이유는 각 구성요소들의 발생 빈도가 너무 낮아 적당한 규모의 임상시험으로는 적절한 검정력을 확보할 수 없기 때문이다. 이러한 경우에 복합 평가변수를 사용하면 전체 사건 발생 빈도가 각 구성요소들의 발생 빈도보다 실질적으로 더 많아지게 되어 적절한 수준의 목표 시험대상자 수와 시험기간으로 임상시험을 진행하여도 충분한 검정력을 확보할 수 있다. 치료의 목적이 여러 임상적으로 중요한 관련 사건 중 하나의 발생을 예방하거나 지연시키고자 하며, 그중 어떠한 사건이 영향을 받는지 명확하지 않을 때 복합 평가변수가 종종 사용된다. 예를 들면, 관상동맥질환에서 심근경색, 뇌졸중, 사망을 예방하기 위하여 항혈소판제를 사용하는 경우이다.

복합 평가변수의 구성요소는 신중하게 선택하여야 한다. 복합 평가변수의 각 구성요소들이 유사한 임상적 중요성을 갖는 경우, 복합 평가변수에 대한 시험약의 효과는 전체적인 임상적 효과로 해석될 수 있다. 그러나 구성요소들 간 임상적인 중요성에 상당한 차이가 있고 시험약의 효과가 주로 가장 덜 중요한 구성요소에 기인한다면 복합 평가변수에 대한 시험약의 효과는 모든 구성요소에 대한 효과를 합리적으로 나타내주지 못하거나 시험약의 유익성을 정확히 기술해주지 못하게 된다. 더욱 심각한 경우는 임상적으로 가장 중요한 구성요소에서는 시험약의 효과가 대조약보다 못한데 임상적으로 덜 중요한 구성요소에서는 시험약의 효과가 대조약보다 더 좋은 결과를 보여, 전체적인 결과는 양호한 통계 결과를 나타내지만 시험약의 임상적 가치에 대한 의문이 발생하는 경우이다. 이런 경우 비록 전체 통계분석 결과에서 시험약이 통계적으로 여전히 유의성이 있다고 하더라도 복합 평가변수의 구성 요소들을 자세하게 조사해보면 그러한 결론에 의문이 들 수 있다. 이러한 이유뿐만 아니더라도 시험약의 치료효과를 더 깊이 있게 이해하기 위하여 복합 평가변수의 각 구성요소들을 분석하는 것이 중요하며(3.4.항 참조), 그러한 분석 결과는 전체 시험 결과의 해석에 영향을 줄 수 있다. 복합 평가변수의 각 구성요소들에 대한 조사는 항상 필수적이지만 다중성 보정의 필요성은 임상시험의 목적에 따라 달라질 수 있다. 복합 평가변수에 대한 시험약의 효과를 더 잘 이해하기 위한 목적이라면 다중성 보정은 권장되지 않는다. 이 경우 시험약의 유익성이 임상적으로 의미있고 위해성을 상회하는지 여부와 이를 허가사항에 어떻게 반영할지에 대한 임상적 판단이 필요하다.

만약 시험약의 추가적인 효과를 확립하기 위한 목적이라면 다중성은 보정되어야 한다.

3.3.4. 다중 구성요소 평가변수

다중 구성요소 평가변수(multi-component endpoint)는 2개 또는 그 이상의 구성요소들을 시험대상자 내에서 조합하여 평가한다. 이 평가변수에서 각 개별 시험대상자에 대한 평가는 해당 시험대상자에서 얻어진 모든 구성요소들의 관측값에 의존한다. 그리고 그 모든 구성요소들에서 얻어진 관측값들을 기반으로 사전에 정해진 규칙에 따라 하나의 전반적인 평가 또는 상태로 결정된다.

각 시험대상자에서의 전반적인 평가는 각 구성요소(도메인) 점수의 평균(가중 평균 또는 가중치가 적용되지 않는 평균)으로 구할 수 있다. 이러한 다중 구성요소의 평가변수로는 조현병 연구에 사용되는 양성 및 음성 증후군 척도(Positive and Negative Syndrome Scale, PANSS) 등이 있다. 다중 구성요소 평가변수는 각 구성요소에서 별도로 지정된 기준을 만족하는 개별 시험대상자에 대한 이분법적인(반응) 평가변수일 수도 있다. 예를 들어, 동종 췌도세포의 제1형 당뇨병 임상시험에서 시험대상자가 이분법적인 반응 기준[예: 저혈당의 위험성 없이 당화혈색소(HbA1c)의 정상 범위의 도달]을 만족할 때만 반응자로 간주되고 일차 평가변수는 반응률이 될 수도 있다.

질병의 모든 특징은 아니더라도 여러 다른 특징들에 대해 어느 정도 효과가 있어야 각 개별 시험대상자에게 시험약이 효과가 있다고 인정하는 좀 더 복잡한 평가변수도 있다. 일례로, 류마티스 관절염을 가진 개개의 시험대상자에서의 반응은 미국 류마티스학회(American College of Rheumatology, ACR) 기준에 따라 질병의 다섯 가지 추가 특성 중 적어도 세 가지 이상에서 개선을 보이는 것과 함께 두 가지 특정 양상에서 어느 정도 개선을 보이는 것으로 정의될 수 있다.

각 개별 시험대상자 내에서 서로 다른 구성요소에 대한 치료효과가 전반적으로 동일한 방향으로 경향성을 보인다면 시험대상자 내에서 다중 구성요소를 사용하는 것이 효과적일 수 있다. 그러나 평가변수 간 경향성이 일치하지 않는다면 임상시험의 검정력은 떨어질 수 있다. 다중 구성요소의 평가변수가 공동 평가변수에 비해 좀 더

효율적일 수 있지만, 일반적으로 특정 시험대상자 내에서 다중 구성요소 평가변수를 통한 평가가 적절한지는 통계적으로 보다는 임상적으로 고려하여 판단한다. 3.3.3.항의 복합 평가변수의 개별 구성요소들을 평가하는 것과 같이 다중 구성요소의 평가변수를 이루는 구성요소를 평가하는 것이 중요할 수 있다. 다만, 특정한 구성요소에 대해 시험약의 효과를 입증하고 싶은 경우에는 임상시험계획서에 이를 명시해야 하고 다중성이 보정되어야 한다.

3.3.5. 임상적으로 중요하지만 일차 평가변수로 사용하기에 발생 빈도가 너무 낮은 평가변수

많은 중대한 질환에서 어떤 평가변수는 임상적으로 매우 중요하여 해당 평가변수의 자료 수집 및 분석이 반드시 필요한 경우가 있다. 대표적인 예로 사망률 또는 주요 이환율 사건(예: 뇌졸중, 골절, 폐 악화)을 확인하는 임상시험이 있다. 이러한 사건의 발생 빈도는 상대적으로 매우 낮지만 복합 평가변수에 포함될 수 있다(3.3.3.항 참조). 또한 이러한 사건을 임상시험계획서에 사전에 이차 평가변수로 지정하여 일차 복합 평가변수에 대한 시험약의 효과가 입증된 이후 개별 평가변수에 미치는 효과에 대한 결론을 뒷받침할 수도 있다.

3.4. 복합 및 다중 구성요소 평가변수의 개별 구성요소

3.4.1. 복합 평가변수의 결과 평가 및 보고

복합 평가변수의 구성요소가 사건을 기반으로 평가되는 경우, 사건은 일반적으로 사전에 정의된 구성요소에 해당하는 사건들 중 어느 하나가 처음 발생하는 것으로 정의한다. 이 구성요소는 임상시험이 종료된 후 군 간에 비율을 비교하거나 사건이 발생할 때까지의 시간으로 분석(time-to-event analyses)될 수 있다. 이러한 사건이 발생할 때까지의 시간으로 분석하는 방법은 임상시험 관찰 기간 내에 사건이 나타나지 않는(event-free) 기간이 임상적으로 의미가 있는 경우에 좀 더 흔하게 사용된다. 시험약이 복합 평가변수의 모든 구성요소에서 좋은 효과를 나타낼 것으로 기대되는 경우라도 실제로 시험약이 그러한 효과를 보일지는 확실하지 않다. 따라서 각 구성요소에 해당하는 사건에 대한 결과들은 개별적으로 검토되어야 하며 임상시험 결과보고서에

포함하여 제시되어야 한다. 이러한 분석 결과는 복합 일차 평가변수의 통계적 유의성에 대한 결론을 바꾸지는 않으나, 복합 평가변수의 결과 해석이 제한적일 수 있다(3.3.3.항 참조). 복합 평가변수의 하나 또는 그 이상의 구성요소들에 대해 시험약의 효과를 입증하기 위하여 해당 구성요소를 별개의 가설 검정으로 분석하고자 하는 경우에는 이러한 가설들을 분석할 때 수반되는 다중성 문제를 고려하여 통계분석 계획이 사전에 기술되어야 한다. 그러나 일반적으로 복합 평가변수 검정을 위해 목표 시험대상자 수 또는 사건의 총 발생 건수가 계획되기 때문에 개별 구성요소의 검정 시에는 검정력이 낮아질 가능성이 있다.

복합 평가변수의 사건이 비율 면에서 어떻게 구성되는지를 설명하기 위하여 처음으로 발생한 사건을 구성요소별로 세분화하여 분석한 결과를 제시하기도 한다. 예를 들어, RENAAL 임상시험(Brenner et al. 2001)에서 일차 유효성 평가변수는 혈청 크레아티닌 2배 증가, 말기 신질환으로 진행 또는 사망의 복합 평가변수 중 처음 발생한 사건이었다. 세부적으로 분석한 결과, 혈청 크레아티닌 2배 증가, 말기 신질환으로 진행 또는 사망의 복합 사건 중 처음 발생한 사건의 비율이 각각 52%, 19%, 29%였다. 그러나 시험대상자는 2건 이상 복수의 사건을 경험할 수 있고, 이러한 시험대상자에서 첫 번째 사건 후에 발생하는 사건(예: 혈청 크레아티닌 2배 증가 후 말기 신장질환으로 진행 또는 사망)은 분석에 포함되지 않을 수 있다. 그러므로 관심 있는 사건 유형에 해당하는 모든 사건(다른 사건 발생 이후 발생하는 사건까지 포함)을 분석에 포함하여 평가하는 것도 중요하다. 이러한 분석이 임상시험계획서에 사전에 기술되고 다중성이 적절하게 설명되어 그 결과를 해석할 수 있는 경우에는 시험약의 추가 효과를 입증할 수 있다.

3.4.2. 다른 다중 구성요소 평가변수의 결과 평가 및 보고

복합 평가변수와 마찬가지로 시험대상자 내에서 다중 구성요소 중 어떤 구성요소가 전체적인 통계적 유의성에 가장 크게 기여했는지를 파악하는 것은 시험약의 임상적 효과를 정확하게 이해하는데 중요하다. 그러므로 각 구성요소에 대한 시험 결과를 분석하는 것이 일반적으로 중요하다. 다만, 앞서 기술한 바와 같이 이러한 분석 결과는 통계적으로 확실한 결론을 암시하거나 해당 시험이 뒷받침하지 않는 효과에 대하여 확신을 주는 방식으로 허가사항에 제시되어서는 안 된다. 이러한 다중 구성요소

평가변수들은 많은 경우 전체 점수(overall score)가 포괄적이면서도 임상적으로 해석 가능한 것으로 간주된다. 반면, 각 개별 구성요소의 척도는 독립적으로 임상적인 해석이 가능할 수도 있지만 그렇지 않을 수도 있다. 임상적으로 해석이 가능하다면 특정 구성요소나 하위 도메인을 일차 및 이차 평가변수 집단의 설명 변수로 분석하는 것이 가능할 수 있다. 전체 다중 구성요소 평가변수 외에 특정 구성요소 또는 하위 도메인에 대하여 시험약의 효과를 입증하려고 한다면 적절하게 다중성을 통제하여 임상시험계획서에 평가변수로 사전에 명시하는 것이 권고된다.

4. 방법론적 고려사항

2절 및 3절에서는 다중성 문제가 발생할 수 있는 다양한 상황에 대하여 기술하였다. 평가변수 집단(3.1.항 참조)이 있는 경우, 다른 평가변수에서 치료효과가 있든 없든 관계없이 적어도 하나 이상의 평가변수에서 통계적으로 유의한 치료효과를 잘못 발견할 확률을 전체 제1종 오류율(overall Type I error rate)이라고 한다. 이 오류율은 일반적으로 0.05(단측검정인 경우, 0.025)로 유지된다. 이러한 오류율이 통계적인 방법을 통해 요구되는 수준으로 적절히 통제된다면 개개 평가변수에 대한 유효성 입증이 가능할 수 있다.

다중 평가변수와 관련하여 발생할 수 있는 다중성 문제를 다루는 여러 가지 일반적인 통계 방법들이 있다. 부록에서는 일반적으로 고려되는 몇 가지 방법을 제시하였다. 본페로니(Bonferroni) 방법, 홈(Holm) 절차, 학버그(혹버그, Hochberg) 절차는 검정하고자 하는 귀무가설들 사이에 계층 구조를 가정하지 않는다(즉, 다른 가설의 기각과 무관하게 개개 귀무가설은 기각될 수 있다). 부록에 제시된 그래픽 접근법 등과 같이 부분 유의수준 배정(partial alpha allocation) 및 계층 구조(hierarchies)를 조합하는 다른 방법도 적용할 수 있다. 평가변수 중 어느 하나에서 통계적으로 유의성이 입증되면 효과가 있다고 인정하는 경우, 평가변수 집단 전체에 걸쳐 다중성을 적절하게 보정하는 방법을 적용할 수도 있다.

그러나 평가변수들이 임상적 중요성 또는 논리적 관련성(입증할 가능성 포함)에 따라 순서가 정해지는 경우에는 다른 방법이 권장될 수 있다. 예를 들어, 일차 평가변수와 이차 평가변수 각각 한 개씩만 존재하는 단순한 경우에는 계층적 검정

방법을 사용할 수 있다. 평가변수 간에 좀 더 논리적/계층적 관계를 설명하기 위해 그래픽 접근법(graphical approach) 및 혼합 게이트키핑 방법(mixture gatekeeping procedures)과 같은 다른 방법들도 개발되었다. 그래픽 접근법은 순차적으로 검정하는 알고리즘이 있으며 그림을 통해 검정 절차를 시각화할 수 있다.

경우에 따라 일차 평가변수에 대한 비열등성(고정된 마진)을 검정한 다음 우월성을 검정할 수 있다. 하나의 평가변수만 유일하게 검정하는 경우에는 다중성 보정 없이 비열등성과 우월성이 검정될 수 있다. 다중성을 보정하지 않는 이유는 비열등성 및 우월성의 귀무가설이 당연하게 순차적으로 정렬되고, 두 검정이 해당 평가변수에서 하나의 계층으로만 적용되기 때문이다. 그러나 검정에 하나 이상의 평가변수가 더 포함된다면 다중성 문제가 발생하므로 전체 제1종 오류 확률을 통제하기 위하여 보정이 이루어져야 한다. 예를 들어, 일차 평가변수에 대한 우월성 가설 검정에 이어 추가적인 평가변수를 검정한다면 검정은 하나의 계층 구조로 정렬될 수 있다. 아니면 이러한 다중 가설 전체에 유의수준을 배정하여 일차 평가변수에 대한 우월성 가설과 추가 평가변수에 대한 가설 모두에 대한 검정을 진행할 수 있다. 이러한 유의수준 배정이 가능한 이유를 살펴보기 위해 다음과 같이 가정해 보자. 시험약이 일차 평가변수에 대해서는 활성대조약 대비 비열등하지만 활성대조약 대비 우월하지 않으며, 이차 평가변수에 대해 활성대조약 대비 비열등하지 않는 경우가 있을 수 있다. 따라서 이러한 가설 검정 중 하나에서 제1종 오류가 발생할 수 있다. 두 가지 모두에서 0.05에서 검정되는 경우, 둘 중 적어도 하나에 대해 결론을 잘못 내릴 확률은 0.05 보다 커질 수 있다. 따라서 어떤 방식으로든 적절한 통제가 있어야 한다 (예: 일차 평가변수의 우월성이 입증되는 경우에만 이차 평가변수를 검정하거나 두 검정 간 α를 분할한다). 이러한 특수 사례와 다른 방법론적인 고려사항은 부록에서 자세히 살펴보기로 한다.

5. 결론

품목허가를 뒷받침하는 임상시험에서 의약품의 유효성을 판단할 때 주요 우려사항은 위양성으로 결론(즉, 약물이 실제 효과가 없는데 효과가 있다고 잘못된 결론을 내리는 것)을 내리는 것이다. 일반적으로 효과 차이가 있다는 잘못된 결론에 대해 5% 미만 (1/20 확률) 또는 시험약이 효과가 더 좋다라는 위양성으로 잘못된 결론에 대해 2.5%

(1/40 확률) 미만으로 제1종 오류율을 통제하는 것이 일반적인 접근방법이다. 평가변수나 분석의 수가 증가함에 따라 다중성 문제로 인해 제1종 오류율은 2.5%보다 훨씬 더 커질 수 있다. 본 가이드라인에서 설명한 것처럼 다중 평가변수에 기반하여 시험약의 효과를 평가할 때 다중성 보정을 통해 제1종 오류율을 통제할 수 있다. 다중성을 보정할 수 있는 많은 전략 및 방법들이 있다. 이러한 방법에는 각각 장단점이 있으며 임상시험 계획 단계에서 적절한 전략과 방법을 선택하는 것이 어려울 수 있으나 통계적 전문지식을 활용하여 가장 적절한 접근법을 선택하는 것이 중요하다. 제1종 오류율을 적절하게 통제하지 못하면 시험약이 효과가 있다는 잘못된 결론 (위양성)을 내릴 위험이 커질 수 있다. 본 가이드라인은 그러한 잘못된 결론을 도출하지 않도록 다중 평가변수로 인해 발생하는 다중성 문제를 다루는 시기와 방법을 명확히 하기 위하여 마련되었다.

6. 참고문헌

1) Guidance for Industry: Multiple Endpoints in Clinical Trials. FDA, 2022.

2) 의약품 임상시험 통계 가이드라인. 식품의약품안전평가원, 2016.

3) 강승호. 신약개발에 필요한 임상통계학. 자유아카데미, 2019.

4) Brenner BM, ME Cooper, D de Zeeuw, WF Keane, et al. Effects of Losartan on Renal and Cardiovascular Outcomes in Patients with Type 2 Diabetes and Nephropathy. NEJM, 345:861-869, 2001.

5) Sakar S and CK Chang. Simes' method for multiple hypotheses testing with positively dependent test statistics. J Am Stat Assoc, 92:1601-1608, 1997.

6) Huque MF. Validity of the Hochberg procedure revisited for clinical trial applications. Stat Med, 35(1):5-20, 2016.

7) Moye LA. Alpha calculus in clinical trials: considerations and commentary for the new millennium. Stat Med, 19(6):767-779, 2000.

8) Hung HMJ and SJ Wang. Challenges to multiple testing in clinical trials. Biom J, 52(6):747-756, 2010.

9) Westfall PH and SS Young. Resampling Based Multiple Testing: Examples and Methods for P-value Adjustment, New York (NY): Wiley-Interscience, 1993.

10) Dmitrienko A, AC Tamhane, and BL Wiens. General Multistage Gatekeeping Procedures. Biom J, 50(5):667-677, 2008.

11) Dmitrienko A and RB D'Agostino. Tutorial in Biostatistics: Traditional multiplicity adjustment methods in clinical trials. Stat Med, 32(29):5172-5218, 2013.

부록 : 통계 방법

이 부록은 다중 평가변수에 대한 치료효과를 평가하는 대조군 임상시험에서 다중성 문제를 해결하기 위해 일반적으로 사용하는 몇 가지 통계적 방법과 접근법을 제시한다. 이 부록에 수록된 방법들이 다중성을 통제하기 위한 포괄적인 방법 목록은 아니며 특정 상황에서는 다른 접근법이 적합할 수 있다. 특정 임상시험에 사용할 방법의 선택은 해당 시험의 목적과 설계는 물론, 개발 중인 의약품에 관한 지식과 임상 환경에 따라 달라진다. 그러나 그 방법은 전향적으로 결정해야 한다. 다중성 보정 방법을 선택할 때 고려해야 할 사항은 복잡하고 개별 제품 개발 프로그램마다 다를 수 있기 때문에 본 가이드라인은 대부분 한 가지 방법을 다른 방법에 우선하여 권장하지 않는다. 의뢰자는 활용할 수 있는 다양한 방법을 고려하고 전향적 분석 계획에서 시험의 설계와 목적에 적합하고 제1종 오류율 통제를 유지하는 가장 검정력이 높은 방법을 선택해야 한다.

1. 본페로니(Bonferroni) 방법

본페로니 방법은 단순하고 폭넓게 적용할 수 있다는 점 때문에 흔히 사용하는 단일 단계 절차이다. 이 검정에 성공한 각 평가변수에 대하여 약물이 효과가 있다고 간주한다. 홈(Holm)과 학버그(혹버그, Hochberg) 방법(아래 참조)은 일차 평가변수에서 본페로니 방법보다 검정력이 더 뛰어나므로 많은 경우에 이 방법을 사용하는 것이 더 바람직하다. 그렇더라도 의뢰자는 이차 평가변수의 검정력을 극대화하기 위해서, 또는 학버그 방법의 가정이 정당화되지 않기 때문에 여전히 일차 평가변수에 본페로니 방법을 사용하고자 할 수 있다.

본페로니 방법의 가장 일반적인 형태는 유효한 전체 유의수준 α(일반적으로 양측 0.05)를 선택된 평가변수 간에 동일하게 나누는 것이다. 그런 다음 m개의 평가변수에서 p값이 α/m보다 작은 평가변수 각각에 대해 α 수준에서 치료효과가 유의하다는 결론을 내린다. 따라서 평가변수가 2개인 경우, 각 평가변수의 임계 α는 양측 0.025가 된다. 본페로니 검정은 상대 가중치의 합이 1.0(예: 평가변수 4개의 가중치가 0.4, 0.3, 0.2, 0.1)인 상태에서 평가변수에 할당된 다른 가중치로 수행할 수도 있다. 이러한 가중치는 평가변수의 임상적 중요성, 성공 가능성, 또는 기타 요인을 고려하여 시험 설계 시 미리 지정해야 한다.

2. 홈(Holm) 절차

홈 절차는 다단계 스텝다운 절차이며 조금이라도 상관관계가 있는 평가변수에 유용하다. 이 절차는 본페로니 방법보다 덜 보수적이다. 그 이유는 가장 작은 p값(본페로니 방법과 동일한 평가변수별 유의수준 값 α/m)으로 검정에 성공하면 본페로니 방법보다 더 큰 평가변수별 유의수준에서 다른 평가변수를 검정할 수 있기 때문이다. 이 검정을 수행하는 알고리즘은 다음과 같다.

완료된 시험에서 얻은 평가변수 p값은 먼저 가장 작은 값부터 가장 큰 값 순으로 정렬된다. 검정할 평가변수가 m개이고 $p_{(1)}$이 가장 작은 p값, $p_{(2)}$는 그다음으로 작은 p값, $p_{(3)}$는 세 번째로 작은 p값 등을 나타낸다고 하자.

가. 검정은 가장 작은 p값인 $p_{(1)}$을 균등 가중치 본페로니 보정에 사용된 동일한 임계값 α/m과 비교하는 것으로 시작한다. $p_{(1)}$이 α/m보다 작으면 이 p값과 관련된 평가변수에 대한 치료효과는 유의한 것으로 간주한다.

나. 다음으로 임상시험 전체 유의수준(α)을 미검정 평가변수의 수로 나눈 평가변수별 유의수준과 다음으로 작은 p값인 $p_{(2)}$를 비교한다[예: 두 번째로 가장 작은 p값에는 다소 덜 보수적인 유의수준인 $\alpha/(m-1)$ 적용]. 만약 $p_{(2)} < \alpha/(m-1)$인 경우, 이 $p_{(2)}$와 관련된 평가변수에 대한 치료효과도 유의한 것으로 간주한다.

다. 검정은 다음 순서의 p값인 $p_{(3)}$을 $\alpha/(m-2)$ 등의 값과 비교하는 작업을 반복하여 가장 마지막 p값(가장 큰 p값)을 α와 비교할 때까지 진행한다.

라. 그러나 각 단계에서 유의하지 않은 결과가 나올 때마다 절차는 중지된다. 정렬된 p값이 유의하지 않으면 나머지 더 큰 p값은 평가하지 않으며 통계적으로 유의하다고 간주할 수 없다.

개별 귀무가설에 불균등한 유의수준 할당을 허용하는 더 일반적인 홈 가중 버전도 있다.

3. 학버그(혹버그, Hochberg) 절차

학버그(혹버그, Hochberg) 절차는 스텝업 검정 절차이다. 이 절차는 홈 절차보다 검정력이 더 강하지만(즉, 치료효과가 홈 절차에서 유의하다면 학버그 절차에서도 해당 치료효과는 유의하다. 하지만 그 반대는 성립하지 않을 때도 있다.), 홈 절차와 달리 특정 가정 하에서만 전체 오류율을 통제한다. 학버그 절차에서는 p값을 홈 절차와 같이 α/m, α/(m-1), …, α의 유의수준 임계값과 비교하지만 홈 절차와 달리 학버그 절차는 스텝업 절차이다. 학버그 절차는 최소 p값으로 시작하지 않고 최대 p값을 먼저 최대 평가변수별 임계값(α)과 비교한다. 또한 기본적으로 홈 절차의 역순으로 가설의 첫 번째 검정이 통계적 유의성을 나타내지 않으면, 두 번째 큰 p값을 두 번째 큰 보정 유의수준 값인 α/2와 비교하는 검정을 진행한다. 이런 방식으로 학버그 절차는 어떤 평가변수에 대한 p값이 통계적으로 유의할 때까지 축차 검정을 계속하고 이를 기반으로 해당 평가변수와 더 작은 p값을 가진 모든 평가변수에 대해 통계적으로 유의한 치료효과가 있다고 결론 내린다. 예를 들면, 최대 p값이 전체 유의수준(α)보다 작은 경우, 이 방법은 모든 평가변수에 대해 유의한 치료효과가 있다는 결론을 내린다. 또 다른 상황에서 최대 p값은 α보다 작지 않지만 두 번째 큰 p값이 α/2보다 작다면, 최대 p값과 관련된 평가변수를 제외한 모든 평가변수에 대하여 치료효과가 입증되었다는 결론이 나온다.

본페로니와 홈 절차는 가정이 필요하지 않은 것으로 잘 알려져 있다. 이 두 가지 방법은 평가변수의 유형, 확률 분포, 상관관계의 유형과 관계없이 적용할 수 있다. 반면, 학버그 절차는 위의 가정과 무관하지 않다. 학버그 절차는 독립 평가변수 검정 또는 양의 상관관계를 가지는 평가변수들의 경우(즉, 검정통계량이 이변량 정규분포)에는 전체 유의수준 통제를 제공하는 것으로 알려져 있다. 학버그 절차는 특정 조건에 부합할 때 유효한 검정 절차이기도 하다. 일반적인 사례(예: 상관관계가 구조가 동일하지 않은 3개 이상의 평가 변수가 존재)에 대한 다양한 시뮬레이션 결과에 따르면 학버그 절차는 반드시 그렇다고는 할 수 없지만 일반적으로 양의 상관관계가 있는 평가변수의 전체 제1종 오류율을 제어한다. 그러나 음의 상관관계에 있는 검정일 때는 그렇지 못한 경우도 있다(Sarkar et al. 1997, Huque 2016).

4. 전향적 유의수준 배정 체계

전향적 유의수준 배정 체계(Prospective Alpha Allocation Scheme, PAAS)(Moye 2000)는 본페로니 방법보다 검정력이 약간 더 높은 단일 단계 방법이다. 이 방법은 모든 평가변수에 대해 균등하거나 불균등한 유의수준 배정을 허용하지만, 본페로니 방법과 마찬가지로 특정 평가변수는 각각 전체 유의수준의 특정 양을 전향적으로 배정받는다. 유의수준 배정은 다음 방정식을 충족해야 한다.

$$(1 - \alpha_1)(1 - \alpha_2) \cdots (1 - \alpha_k) \cdots (1 - \alpha_m) = (1 - \alpha)$$

이 방정식에서 각 요소$(1 - \alpha_k)$는 배정된 유의수준 α_k에서 검정할 때 k번째 평가변수에 대한 귀무가설을 정확하게 기각하지 않을 확률을 의미한다. 이 절차는 평가변수들이 독립적이거나 양의 상관관계에 있을 때 유효하지만, 평가변수들이 음의 상관관계에 있으면 제1종 오류율은 부풀려질 수 있다. 이 방정식은 각 m 개의 확률을 모두 곱하여 얻은 개별 귀무가설 전체를 정확하게 기각하지 않을 확률이 선택된 목표(예: 0.95)와 같아야 한다는 요건을 보여준다. 어떤 개별 평가변수 검정이든 임의로 유의수준 배정을 할 수 있지만 총 배정 조합은 위의 방정식을 항상 충족해야 한다. 일반적으로 어떤 평가변수에 임의의 유의수준 배정이 이루어진다면 적어도 최종 평가변수의 유의수준을 계산했을 때 전체 방정식을 충족해야 한다. 앞서 서술한 것처럼 본페로니 방법은 제한을 정의하는 유사한 방정식을 쓰지만 모든 개별 유의수준의 합이 전체 유의수준과 같아야 한다는 점은 다르다.

5. 고정 순서 방법

여러 시험에서 평가변수의 검정은 특정 순서대로 진행되며 종종 임상적 관련성이나 성공 가능성에 따라 순위를 매긴다. 고정 순서 통계 검정 절차에서는 미리 정해진 순서대로 모두 같은 유의수준(예: α = 0.05)에서 평가변수를 검정하며, 앞선 평가변수에서 검정이 성공해야만 다음 평가변수로 넘어가게 된다. 이러한 검정 절차에서는 (1) 검정 순서가 전향적으로 배정되어야 하고, (2) 해당 순서의 귀무가설이 기각되지 않으면 추가 검정은 없어야 한다. 즉, 유의수준(예: α = 0.05)에서 유의성이 보이지 않으면 즉시 추가 검정을 멈춘다.

고정 순서 검정 방법의 장점은 개별 검정의 유의수준 보정이 필요하지 않다는 점이다. 주요 단점은 순서에서 가설이 기각되지 않으면 후속 가설에 예정된 평가변수들에 대하여 p값이 극히 작은 경우에도 통계적 유의성을 달성할 수 없다는 것이다. 예를 들어, 어떤 시험에서 순서의 첫 번째 평가변수에 대한 p값이 0.59, 두 번째 평가변수의 p값이 0.001이라면, 두 번째 평가변수에서 매우 확실한 값이 나와도 그 결과는 통계적으로 유의하다고 간주하지 않는다. 첫 번째 평가변수의 결과를 무시하면 다중성 문제가 다시 발생하고 전체 제1종 오류율이 부풀려진다. 이 예시에서 본페로니 방법과 같은 제1종 오류를 제어하는 다른 방법을 적용했다면 두 번째 평가변수에서 효과가 나타났을 것이다.

따라서 고정 순서 방법의 경우, 가설의 검정 순서를 신중하게 선택하는 것이 중요하다. 검정 초기에 서열에서 통계적 유의성을 보여주지 못하는 경우, 평가변수의 나머지도 통계적으로 유의하지 않을 것이다. 검정에 가장 좋은 순서를 선험적으로 결정하는 것은 불가능할 때가 많고(Hung and Wang 2010), 다중성 문제를 해결할 수 있는 다른 방법들도 있다. 이는 다음 절(6절)에서 설명하고자 한다.

6. 리샘플링(Resampling) 기반 다중 검정 절차

다중 평가변수 사이에 상관관계가 있다면 리샘플링(Westfall and Young 1993)이 위에서 설명한 전체 제1종 오류율 제어를 유지하면서 실제 치료효과를 감지하기 위한 방법들보다 더 높은 검정력을 제공할 수 있는 일반적인 통계적 접근법이며 상관관계가 증가함에 따라 검정력도 증가한다. 이러한 방법을 사용하면 귀무가설 하에서 가능한 검정통계량의 분포가 임상시험의 관찰된 데이터를 기반으로 생성된다. 이 데이터 기반 분포는 대다수의 다른 방법과 같이 검정 통계의 이론적 분포(예: Z 점수의 정규분포 또는 t 점수의 t 분포)를 사용하는 대신 관찰된 시험 결과의 p값을 찾는 데 사용된다.

리샘플링 방법에는 다중 평가변수에 대한 부트스트랩 및 순열(permutation) 접근법이 포함되며, 중요하긴 하지만 평가변수의 실제 분포에 대한 가정은 거의 필요하지 않다. 그러나 이러한 방법에는 몇 가지 단점이 있다. 중요한 가정은 일반적으로 소규모 표본 크기 임상시험일 때 특히 입증하기 어렵다. 결과적으로 이러한 방법은 일반적으로 대규모 표본 크기 임상시험(특히 부트스트랩 방식)을 필요로 하며 제한된 시험 데이터에서 얻은 검정 통계량의 데이터 기반 분석을 적용하고, 적절한 제1종 오류율 제어를 보장하기 위해 종종 시뮬레이션해야 한다. 예를 들어, 비교 대상 치료 집단 간에 표본 분포의 모양이 다른 경우 제1종 오류율이 부풀려질 수 있다.

7. 게이트키핑 검정 전략

게이트키핑 절차(예: Dmitrienko et al. 2008, Dmitrienko and D'Agostino 2013)는 계층적으로 정렬된 귀무가설 집단의 검정 문제를 다룬다. 임상시험에서 집단은 보통 1차 및 2차 목적에 해당한다(3.1.항 참조). 각 집단에서 추론은 귀무가설 사이에 존재할 수 있는 논리적 관계와 일치하는 초기 집단에서 귀무가설을 채택하느냐 기각하느냐에 따라 달라진다. 이 관계는 일반적으로 관련 임상적 고려사항을 반영하며 일련의 논리적 제한조건을 사용하여 지정한다. 직렬 게이트키핑, 병렬 게이트키핑, 그리고 트리 구조 게이트키핑이라고 부르는 일반화를 포함하여 다양한 유형의 논리적 게이트키핑 제약조건이 연구되었다.

예를 들어, 일차 평가변수 집단(primary family)의 평가변수가 공동 일차 평가변수로 검정되는 시나리오에서 직렬 전략을 적용할 수 있다(3.3.항). 일차 평가변수 집단의 모든 평가변수가 유의수준에서 통계적으로 유의한 경우(예: α = 0.05), 이차 평가변수 집단의 평가변수를 검사한다. 이차 평가변수 집단에서 평가변수는 이차 평가변수 집단 내에서 제1종 오류율을 제어하는 사전 지정된 허용 가능한 방법(홈 절차, 고정 순서 방법, 또는 이 부록에 설명한 기타 방법)으로 전체 유의수준 수준에서 검정할 수 있다. 그러나 일차 평가변수 집단의 귀무가설 중 하나 이상이 기각되지 않으면 일차 평가변수 집단 기준을 충족하지 않으며 이차 평가변수 집단은 검정하지 않는다.

일차 평가변수 집단에서 평가변수가 모두 공동 일차 평가변수가 아닐 때 병렬 게이트키핑 전략을 적용하고, 일차 평가변수 집단에는 분리 가능한 검정 방법(예: 본페로니 방법 또는 절단 홈(truncated Holm) 방법)이 지정된다. 이 전략에서 일차 평가변수 집단의 평가변수 중 하나 이상이 통계적 유의성을 보였다면 이차 평가변수 집단을 검토한다.

일부 다중성 문제는 다차원적이다. 하나의 차원은 다중 평가변수에 해당할 수 있고, 두 번째 차원은(각 평가변수가 검정된) 여러 개의 용량 그룹이 존재할 수 있으며, 또 다른 차원은(각 용량과 각 평가변수에 대한) 비열등성과 우월성 검정과 같이 다중 가설에 해당할 수 있다. 다중성의 원인이 다수라면 가설 검정의 방법도 여러 가지가 될 수 있다. 예를 들어, 시험의 목적이 우월성뿐만 아니라 비열등성도 입증하는

것이라면 단일 경로의 순차 검정(sequential test)이 바람직하다. 하지만 일차 평가변수에서 우월성이 제대로 나타난 후 이차 평가변수의 비열등성을 분석하려고 한다고 가정해 보자. 이제 검정 경로는 이 초기 검정에서부터 두 가지 경로로 갈라진다(즉, 일차 평가변수의 우월성 검정과 이차 평가변수의 비열등성 검정).

다중 분기(multi-branched) 게이트키핑 절차에서는 앞선 검정이 성공하면 둘 이상의 평가변수를 검정하는 옵션을 가지고 검정 순서를 정할 수 있다. 여러 수준의 이런 축차 계층이 있고 분기가 여러 단계에 적용될 때, 평가변수 검정의 가능한 경로는 복잡한 다중 분기 구조가 된다.

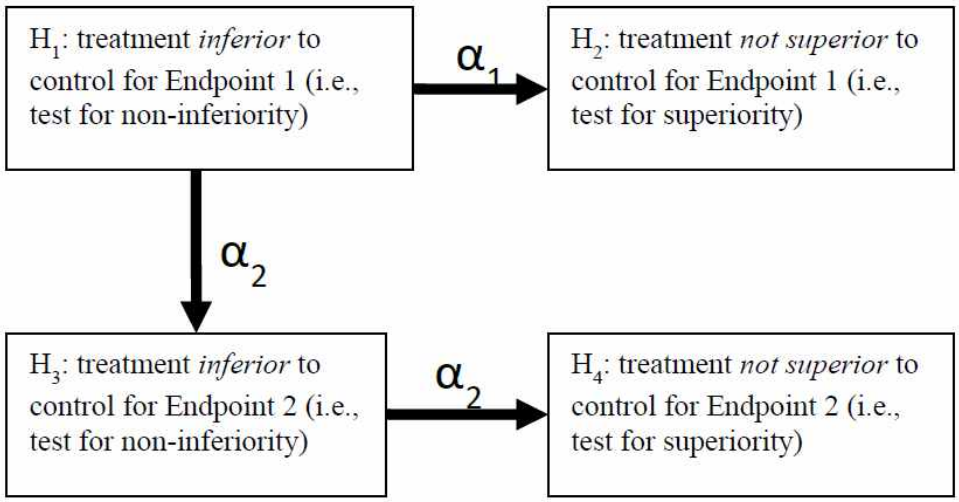
Figure 1에서 보듯이 시험군이 적어도 하나의 평가변수에 대해 대조군과의 비열등성 여부를 먼저 판단하기 위해 두 가지 일차 평가변수(평가변수 1과 평가변수 2)에서 시험군과 대조군을 비교하는 임상시험을 생각해보자. 두 평가변수 중 어느 하나에서 해당 시험군이 대조군보다 비열등하다고 밝혀지면 해당 평가변수에서 그 시험군이 대조군보다 우월한지 아닌지도 검정하고자 할 것이다. 따라서 임상시험의 분석 계획에 다음과 같은 논리적 제한조건을 설정한다.

가. 평가변수 1의 비열등성이 먼저 입증된 후에만 평가변수 2를 검정한다.

나. 평가변수의 비열등성이 먼저 결정된 후에만 해당 평가변수에서 우월성을 검정한다.

다음 다이어그램은 검정 전략의 결정 구조를 보여준다. 이 다이어그램에서 각 블록(또는 노드)은 검정 전략이 검정하는 귀무가설을 나타낸다.

따라서 위의 검정 전략에는 2차원 계층 구조가 있다. 하나는 2개의 서로 다른 평가변수에 해당하는 차원이고 다른 하나는 비열등성 및 우월성 검정에 해당하는 차원으로 위에서 보듯 논리적 제한조건이 있다. 이 절차 유형에서 다중 분기가 단일 노드에서 갈라져 나왔다면 유의수준은 다수의 분기에 걸쳐 분할되어야 한다.



H1: 평가변수 1의 대조군 대비 시험군 비열등성(즉, 비열등성 검정)
H2: 평가변수 1의 대조군 대비 시험군 우월성(즉, 우월성 검정)
H3: 평가변수 2의 대조군 대비 시험군 비열등성(즉, 비열등성 검정)
H4: 평가변수 2의 대조군 대비 시험군 우월성(즉, 우월성 검정)

Figure 1. 논리적 제약이 있는 임상시험의 평가변수 1과 평가변수 2에 대한 비열등성 및 우월성 검정의 흐름도 예시. 이 경우 α1 + α2 = α이다. 평가변수 1 및/또는 평가변수 2에 대한 우월성 검정을 위해 먼저 해당 평가변수에서 비열등성을 입증해야 한다.

8. 순차적 기각 검정에 기반한 그래픽 접근법

그래픽 접근 방법(예: Bretz et al. 2009)은 본페로니 기반 순차 기각 방법에 맞는 다수의 분석 전략을 개발하고 평가하기 위한 수단이다. 이 접근 방법은 제1종 오류율을 제어하고 대안적 검정 전략의 생성 및 평가에 도움이 되는 검정 전략을 매핑함으로써 평가변수 간의 관계뿐만 아니라 평가변수 중요도의 차이를 설명한다.

복잡한 분석 전략을 그래픽으로 나타내면 가설들의 평가변수 검정 간 모든 논리적 관계를 보여주어 제안된 계획을 설명하고 평가하는 데 도움이 될 수 있다.

그래픽 접근 방법의 기초: 꼭지점(노드)과 경로(순서 또는 방향)의 사용

그래픽 방법에서 검정 전략은 꼭지점(또는 검정 순서 경로의 교차점인 노드)에 위치한 각 가설(H1, H2, ..., Hm)을 보여주는 그림(그래프)으로 정의된다. 각 꼭지점(가설)에는 유의수준의 초기 양이 할당되며, 이 문서에서는(하나의 평가변수 검정은 하나의 가설 검정과 관련되며 그 반대도 마찬가지라는 이해를 토대로) 이를 평가변수별 유의수준으로 정의한다. 핵심 요건은 모든 평가변수별 유의수준의 합계가 시험에 사용할 수 있는 총 유의수준(전체 제1종 오류율)과 동일해야 한다는 것이다. 알고리즘의 각 단계에서 평가변수는 본페로니 절차를 사용하여 평가변수별 유의수준에서 검정한다.

그림(그래프)의 또 다른 특징은 방향성이 있는 간선 세트이다. 각 유향 간선(또는 화살표)은 2개의 가설을 연결하고 그 간선의 가중치라고 하는 0과 1 사이의 값이 할당되며 화살표 위에 표시된다. 이는 보존된 유의수준의 일부분(fraction)이 그 경로를 따라 후속 가설로 이동한다는 것을 보여주며, 이때 해당 경로의 말단에 있는 가설은 성공적이다(즉, 기각된다). 꼭지점에서 뻗어나가는 모든 경로에 걸쳐 가중치의 합은 1.0이 되어야 하며 그렇게 하여 보존된 앞 단계의 유의수준이 후속 가설을 검정하는 데 사용된다. 유효성에 대한 확고한 결론을 제시하도록 의도된 모든 시험 가설이 그래프에 표시된다.

이러한 다이어그램의 개념, 구성, 해석, 적용을 설명하는 데 도움이 되는 그래픽 방법의 몇 가지 예는 다음과 같다.

가. 고정 서열법

Figure 2에 표시된 고정 서열 검정 전략(부록 5절)은 세 가지 가설이 있는 그래픽 방법의 간단한 사례를 보여준다. 이 체계에서 평가변수(가설)는 순서가 지정된다. 첫 번째 평가변수에서 전체 유의수준(α)으로 검정을 시작하여 평가변수가 통계적으로 유의하지 않을 때까지만 순서에 따라 계속된다. 이 다이어그램을 보면 가설 H1, H2, H3과 관련된 평가변수별 유의수준이 처음에는 α, 0, 0으로 설정되어 있다. 고정 서열법의 경우, 화살표는 검정 순서를 나타내며 검정이 성공적이면 전체 유의수준이 다음 검정으로 넘어간다. 결과적으로 귀무가설 H1이 성공적으로 기각되면 H2에 대한 평가변수별 유의수준은 $0 + 1 \times \alpha = \alpha$가 되어 유의수준 α에서 H2를 검정할 수 있다. 그러나 H1의 검정이 실패하면 H2를 검정할 수 있도록 H2에 미리 할당된 0이 아닌 H1에서 넘어온 유의수준이 없으므로 검정은 중단된다.

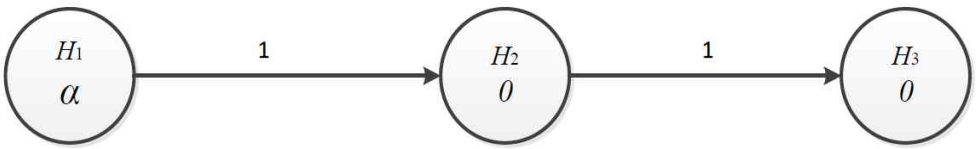


Figure 2. 가설이 셋인 고정 서열의 그래픽 예시

나. 양방향 재검정 가능성을 나타내는 루프백 특징

그래픽 방법의 또 다른 유용한 특징은 가용한 유의수준이 복수의 평가변수 사이에서 평가변수별 유의수준으로 분할될 때 나타난다. 이런 다이어그램은 평가변수별 유의수준의 루프백 통과 가능성을 보여준다.

홈 절차(부록 2절)는 그래픽 방법으로 절차와 그 근거를 간단하게 설명할 수 있는 루프백 특징을 가진 두 가지 가설의 구체적인 검정 사례이다. 홈 절차는 첫 번째 단계에서 더 작은 p값을 평가변수 특정 유의수준 = α/2에서 검정하고 성공한 경우에만 α(예: 0.05) 수준에서 더 큰 p값에 대한 검정을 진행하게 한다. 홈 절차는 유의수준을 균등하게 반으로 나누기 때문에 p값이 작은 가설의 검정이 유의하지 않다면 p값이 큰 검정도 유의하지 않을 게 분명하므로 해당 비교 수행은 불필요하다.

홈 절차에 대한 다이어그램(Figure 3)은 전체 유의수준 = 0.05의 요건을 충족하는 두 꼭지점과 관련 평가변수별 유의수준(각각 $\alpha 1 = 0.025$, $\alpha 2 = 0.025$)을 보여준다. 2개의 화살표는 유의수준이 H1에서 H2로, 또는 H2에서 H1으로 나아갈 수 있음을 보여준다. 첫 번째 검정에서 성공하면 0.025인 평가변수별 유의수준은 다른 가설로 완전히 이동하고 해당 가설에 이미 배정된 평가변수별 유의수준에 추가하여 총 유의수준 0.05를 제공한다. 두 가설 중 어느 것이든 하나가 먼저 검정될 수 있으므로 다이어그램은 루프백 구조를 나타낸다.

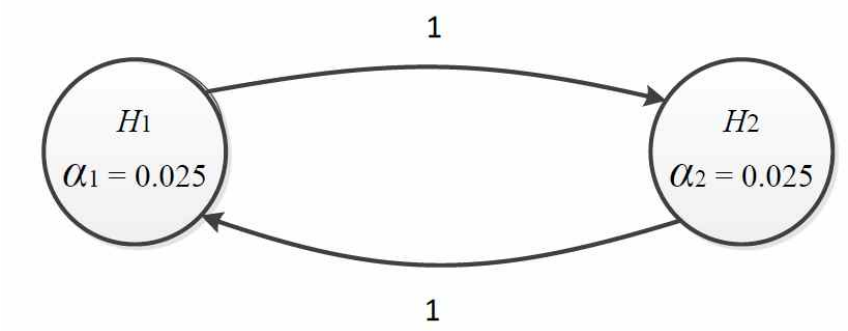


Figure 3. 가설이 둘인 홈 절차의 그래픽 예시

다이어그램 상에서 검정은 첫 다이어그램에서 유의수준이 0이 아닌 모든 꼭지점에서 시작할 수 있으며, 유의수준이 0이 아닌 모든 꼭지점은 검정에 성공한 꼭지점을 찾을 때까지(즉, 가설이 기각될 때까지) 검정할 수 있다. 그런 다음 해당 노드가 제거되고 기각된 가설에 할당된 유의수준이 다이어그램에 표시된 대로 화살표를 따라 다른 노드로 이동한다. 가설이 기각된 최종 결론과 기각되지 않은 최종 결론은 어떤 꼭지점을 먼저 검사했는지와 관계없이 동일할 것이다. 그래픽 방법을 사용하면 검정 경로 특징의 복잡한 유의수준 분할 및 분기를 분석 계획의 일부로 명확하게 식별하고 올바르게 구현할 수 있다.

다. 가설이 성공적으로 기각되었을 때 다이어그램의 점진적 업데이트

그래픽 접근 방법은 가설이 성공적으로 기각될 때마다 초기 그래프의 지속적인 업데이트를 통해 여러 가설의 계층적 검정을 안내한다. 초기 그래프는 전체 검정 전략

(모든 가설 포함)을 나타낸다. 각각의 새로운 그래프는 기각된 가설을 제거하고 검정 또는 재검정할 가설만 남겨 검정 전략의 진행 상황을 보여준다.

유의수준 분할이 복잡한 분석 전략을 고려하려는 경우, 다이어그램의 그래픽 방법과 점진적 업데이트를 통해 여러 가설 시나리오에 대한 다양한 전략의 의미를 이해하는데 도움이 될 수 있다. 이 점진적 업데이트는 최종 시험 통계분석 계획에 적용할 특정 전략을 정할 때 도움이 될 수 있다.

다중 평가변수를 사용하는 임상시험 가이드라인(안)

발 행 일	2024년 4 월 29 일
발 행 인	박 윤 주
편 집 위 원 장	김 영 림
편 집 위 원	식품의약품안전평가원 의약품심사부 순환신경계약품과 김소희, 주정흔, 서현옥, 우나리, 김정현, 배수영, 김송이, 유한빛, 변지영, 조혜영 식품의약품안전평가원 제품화지원팀 정지원, 김문신, 박봉서, 김건소, 정지원, 박애란, 김선희, 이민규
발 행 처	식품의약품안전평가원 의약품심사부 순환신경계약품과