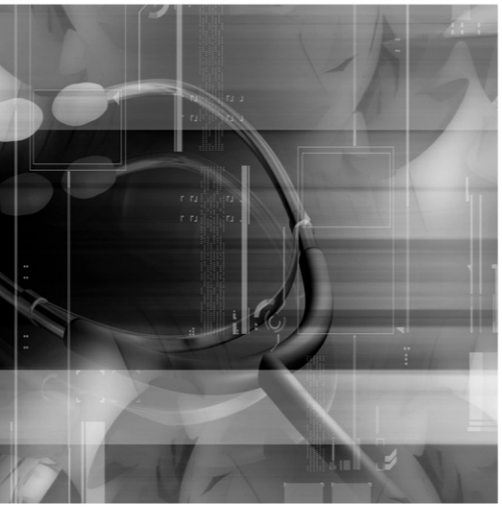
발 간 등 록 번 호
11-1470000-001891-01



임상시험에 있어 다중성에 대하여 고려할 점

Points to consider on multiplicity issues in clinical trials

2009. 6.

식품의약품안전청
Korea Food & Drug Administration
기획조정관 통상통계담당관

인 사 말

최근 한미 FTA 의약품 협상 및 약가제도 조정 등 제약사업의 환경변화와 함께 국내외 의약산업 규모가 점차 커지고 있고, 다국가 임상시험 및 국내임상시험 실시가 증가하고 있습니다. 이에 따라 국가적으로도 신약 개발에 지원을 적극 확대하여 국내 임상시험의 질적 향상 뿐 아니라 임상자료의 활용을 더욱 높여야 할 것입니다.

식품의약품안전청에서는 유럽 의약품국 (European Medicines Agency; EMEA)의 임상통계관련 가이드라인을 번역하여 국내 임상시험의 질적 향상을 높이는데 도움이 되고자 본 해설서를 작성하게 되었습니다. 그 두 번째로 임상시험에서 다중성의 고려 사항에 대한 유럽 의약품국의 “Points to consider on multiplicity issues in clinical trials” 가이드라인 해설서를 작성하였습니다.

본 해설서가 과학적이고 객관적인 임상시험의 수행에 도움이 되어, 국내 임상시험 및 신약개발에 있어 국제 경쟁력을 확보하는데 기여할 수 있기를 희망합니다.

2009년 6월

식품의약품안전청 기획조정관

왕 진 호

임상시험에 있어 다중성에 대하여 고려할 점

(Points to consider on multiplicity issues in clinical trials)

2002.9.19
CPMP/EWP/908/99

목 차

1. 서론

다중성 (multiplicity)은 사실상 모든 임상시험에서 존재한다. 만약 이러한 다중성 문제를 적절하게 다루지 않으면, 부적절한 결론이 나올 가능성이 높아지고 약물 효과에 대해 입증되지 않은 주장을 할 수 있다. 예를 들어, 5개 하위집단에 대한 비교검정이 유의수준 2.5%에서 독립적으로 수행된다면, 적어도 하나 이상의 위양성률은 12%로 증가한다.

이 예제는 다중성이 존재하는 분석에서 연구자가 가장 선호하는 결과를 선택한다면 시험약의 허가나 라벨링 (labelling)에 잘못된 결론을 초래하는 위양성률에 상당한 영향을 줄 수 있음을 보여준다. 그러나 연구자가 가장 선호하는 결과를 선택하지 않을 경우 이러한 영향은 없을 것이다. 이러한 상황에 대한 예제는 나중에 논의될 것이다. 시험별로 위양성률을 적절한 수준 α로 조절하는 것은 중요한 원칙이며, 확증 임상시험의 결과를 평가하는데 있어서 큰 의미가 있다.

위양성률을 조절하는 방법은 여러 가지가 있으며, 그 방법들은 각각의 상황에 따라서 결정된다. 본 문서에서 '제 1종 오류의 조절 (control of the family-wise type I error)'이라는 용어는 강한 의미에서 여러 귀무가설의 일부가 참인 것과 관계없이 최소한 하나의 참인 귀무가설을 기각할 확률을 조절한다는 '전체 제 1종 오류의 조절 (control of type I error)'의 축약어로 사용된다. 하나 이상의 확증시험이 필요한 경우, 제출 수준에 적합한 제 1종 오류 수준을 설정하는 문제에 대해서는 별도 지침서 (CPMP/2230/99 Points to Consider on Application, '1) 메타분석' 및 '2) 중추적 시험'으로 구성)에 다루었다.

본 문서에는 다중성의 모든 부분을 다루기보다, 최근 유럽에서 적용하고 있는 중요한 사항을 기술하였다.

- 다중성 보정 (adjustment of multiplicity) - 언제 필요하고, 언제 필요하지 않는가?
- 다중 이차 유효성 평가변수 (multiple secondary variables)의 유의성을 어떻게 해석하고, 이 중 하나에 대한 주장을 언제 할 수 있는가?
- 하위집단 분석 (subgroup analysis)의 결론을 언제 믿을 수 있는가?
- 하위집단의 허가 (licence to a subgroup)를 제한하는 것은 언제 적절한가?
- 원 변수와 함께 '반응자 (responder)' 분석을 어떻게 해석해야 하는가?
- 규정과 관련하여 복합 평가변수 (composite endpoint)를 통계적으로 어떻게 다룰 것인가?

위에서 언급된 사항 이외에도 임상시험에서의 다중성과 관련된 많은 분야가 있다. 예를 들어 중간분석 (효과가 있거나 혹은 효과가 없음을 중간에 확인하여 임상시험을 중단) 이나 단계적 설계 연구 (미래 시점에서 적응적 변화)에 관한 다양하고 복잡한 방법론이 이에 해당된다. 그러나 위와 같은 분야들은 중요한 문제이고 다룰 내용이 많기 때문에 각각을 개별적 문서로 다루는 것이 적절할 것이다.

반복적인 방문에 의한, 일차 유효성 평가변수의 해석은 반복 측정된 값에 대한 적절한 방법이 사전에 정의되어지거나, 효과가 나타나는데 요구되는 치료기간을 감안하여 미리 정해진 시점에서 일차평가가 이루어지므로 보통의 경우 다중성의 문제를 야기하지 않는다. 따라서 반복측정 분석의 잠재적인 다중성에 대한 문제는 이 문서에서 고려하지 않았다.

2. 다중성의 보정 - 언제 필요하고, 언제 필요하지 않는가?

하나의 일차 유효성 평가변수에 대해 두 처리군을 비교 (일차 유효성 평가변수와 관련하여 단 하나의 귀무가설을 미리 언급하고 중간분석이 없는 경우)하는 확증적 임상시험의 경우 제 1종 오류를 보정할 필요가 없다.

비록 많은 상황에서 다중성의 잠재적 효과에 주의를 기울여야 하지만, 예를 들어 일차 유효성 평가변수를 미리 정의하고 다른 모든 이차 평가변수를 보조적 요소로 규정하는 경우와 같이 다중성을 고려하지 않아도 되는 경우가 많다.

전체 제 1종 오류 (α: overall type I error)를 조절하는 방법은 때때로 '다중수준 α검정 (multiple-level-α-tests)'이라고 불린다. 전체 제 1종 오류를 조정하는 것은 보통 미리 정해진 전체 제 1종 오류 α를 분할하여, 분할된 α로 여러 귀무가설을 검정하여야 함을 의미한다. 이를 통상적으로 '제 1종 오류 수준의 보정 (adjusting the type I error level)'이라 한다. 이렇게 전체 제 1종 오류를 "나누는 (spend)" 방법을 정의하는 알고리즘은 다소 복잡하다. 이처럼 절차가 복잡해짐에 따라 연구 결과를 임상적으로 해석하는 것이 어려울 수 있다. 예를 들어 추정을 목적으로 할 때 추정값의 정확성을 평가하기 위해서 신뢰구간이 가장 중요할 수 있지만, 가설검정과 일치하는 신뢰구간을 만드는 방법은 제 1종 오류의 조절을 목적으로 하는 복잡한 다중수준 α검정 (혹은 더욱 일반적으로 닫힌검정) 상당수에서 가능하지 않다. 따라서 통계적 접근법을 선택할 때에는, 현재 사용되고 있는 통계적 절차를 이용할 경우 만족스러운 임상적 해석을 할 수 있는지 여부를 고려하도록 권고된다. 다중성을 고려하여 처리하는 대안적인 방법들은 종종 서로 다른 결론을 도출할 수 있기 때문에 선호되는 다중수준 α검정에 대한 사전정의가 필요하다. 해석상의 문제를 피하기 위하여 임상시험계획서나 통계분석 계획 (Statistical Analysis Plans:SAP)에 위 내용에 대한 자세한 절차가 포함되어야 한다.

만약 예측되지 않은 다중비교 상황이 발생한다면, 본페로니 (Bonferroni) 방법과 같은 보수적인 접근방법이 필요하다. 이 방법들은 본질적으로 검정력의 손실을 가져올 것이므로 다중검정 상황이 예측된다면 사용할 보정 방법을 사전에 정의하도록 권고된다.

본 문서에서는 임상시험에서 다중검정이 사용되는 적절한 상황과 제 1종 오류를 보정하는데 보편적으로 사용되는 방법을 논의한다.

2.1 다중 일차 유효성 평가변수 –공식적인 보정이 필요 없는 경우

임상시험에서의 통계학적 원리에 관한 ICH E9 지침에서는 일반적으로 하나의 일차 유효성 평가변수로 임상시험을 수행할 것을 권고한다. 만약 임상적 처리효과를 입증하는 것이 단일 변수만으로 가능하다면, 하나의 일차 유효성 평가변수만으로도 충분하다. 그러나 단일 변수로 임상적 처리효과를 입증하는 것이 충분하지 않다면, 하나 이상의 일차 유효성 평가변수를 사용해야 할 수도 있다. 때때로 관련성 있는 일련의 목적들이 일차 유효성 평가변수 각각을 가지고 동일한 시험에서 평가되기도 하고, 또는 일부 혹은 전부에 대한 유용성의 증거를 제공하기 위해 많은 일차 유효성 평가변수가 요구되기도 한다. 이러한 상황에서는 각각의 일차 유효성 평가변수에 대한 대립가설과 검정력의 한계가 정의되어야 하고 해당 목적에 맞도록 서로 균형이 잡혀야 하기 때문에 피험자 수 계산이 매우 복잡해진다.

하나 이상의 일차 유효성 평가변수를 가지는 임상시험을 상황별로 구분하여 다음 하위단락에 서술하고자 한다. 설명된 이 방법들을 사용할 경우, 임상적 해석이 가능하고 다중성에 대한 문제들을 만족스럽게 다루면서도 제 1종 오류율에 대하여 어떠한 공식적인 보정도 필요하지 않다. 사실 이 방법들은 전체 제 1종 오류를 조절하는 닫힌 검정 절차의 일부이다.

2.1.1 임상적 처리효과를 입증하기 위해 둘 이상의 일차 유효성 평가변수가 필요한 경우

통계적 유의성은 모든 일차 유효성 평가변수에서 필요하다. 따라서 다중성에 대한 공식적인 보정은 필요하지 않다.

이 경우, 임상적 처리효과를 충분히 입증하기 위해서는 모든 일차 유효성 평가변수에 대한 각각의 귀무가설이 동일한 유의수준 (예를 들어 0.05)에서 기각되어야 하기 때문에 결과의 해석이 가장 명확하다. 이러한 임상 상황의 예는「알츠하이머병의 처리에 대한 CPMP 지침」또는「만성 폐쇄성폐질환 피험자의 장기간 치료에 사용되는 의약품의 임상적 연구에 대한 CPMP 고려사항」을 참고하면 된다. 이러한 상황에서는 선호되는 몇 개의 일차 유효성 평가변수만을 선택할 수 없으므로 '각각의 제 1종 오류의 수준'이 '전체 제 1종 오류의 수준'과 동일하게 정해진다. 이러한 과정은 제 2종 오류 (적어도 하나의 귀무가설이 참이라는 잘못된 결론을 내릴 오류)를 증가시켜, 최악의 경우 제 2종 오류가 개별 가설들의 제 2종 오류의 합으로 나타나기도 한다. 이러한 제 2종 오류의 증가를 고려해야 임상시험의 표본 크기를 제대로 추정할 수 있다.

2.1.2 둘 이상의 일차 유효성 평가변수가 임상적 적절성에 따라 순위가 매겨진 경우

제 1종 오류의 공식적인 보정은 필요하지 않다. 그러나 귀무가설이 처음으로 기각되지 않은 경우에 해당 변수보다 순위가 낮거나 같은 변수에 근거해서는 확증적 주장을 할

수 없다.

때때로 동일한 임상시험 내에 관련성 있는 일련의 목적들이 존재하는 경우가 있고, 그 중 하나의 목적이 가장 중요하긴 하지만 다른 목적들과 관련된 명백한 결과들도 처리(treatment)의 가치를 분명히 높여줄 것이다. 전형적인 예로 (i) 우울증에서의 급성 효과 후 질병 진행의 예방 (ii) 급성 심근경색증에서의 사망 감소 후 중대한 부작용의 예방을 들 수 있다. 이러한 경우, 계층적 계획에 따라 가설을 검정할 수 있다 (그리고 신뢰구간을 산출할 수 있다) 계층적 순서는 자연스럽게 결정되거나 (예를 들어, 시간이나 고려되는 변수의 중요도에 따라 가설을 순서화한다) 혹은 연구자가 정한 우선순위에 따라 결정될 수 있으며, 유의수준 (α)을 축소하거나 분할할 필요가 없다. 그러나 귀무가설을 검정하기 위한 계층적 순서는 임상시험계획서에서 미리 정의되어야 한다. 그러한 절차의 효과는 귀무가설이 처음으로 기각되지 않은 변수보다 낮거나 같은 순위를 갖는 변수에 근거해서는 확증적 주장을 할 수 없다. 이러한 계층적 검정 절차와 일치성을 갖는 신뢰구간이 유도될 수 있다. 낮은 순위의 변수에 해당하는 가설의 경우 제 2종 오류가 증가하는 것이 명백하다. 이러한 유사한 방법은 이차 유효성 평가변수에 사용될 수 있다 (3.2절을 참고하라).

이러한 다중변수를 다루는 상황은 확증적 임상시험에서 드물게 발생하며, 이것에 대하여 논의하고 있는 문헌들이 다양하므로 본 문서에서는 다루지 않겠다. 그러나 이러한 방법을 적용하기 전 규제기관과의 사전상담이 권고된다.

2.2 분석 대상군

다중분석은 동일한 변수에서 피험자의 다양한 부분집합에 대해 수행될 수도 있다. ICH E9 지침서에서 지적하는 것과 같이 주분석 대상에 대한 피험자 자료의 집합은 임상시험계획서의 통계분석 계획에 정의되어야 한다. 이러한 피험자 부분집합 중 하나 (보통 완전한 집합)가 일차분석용으로 선택된다.
일반적으로 일차분석으로부터 결과의 민감성을 조사하는 목적으로 다양한 피험자의 부분집합 혹은 다양한 측도에 대하여 수행되는 다중분석은 단지 일차분석의 결과에 확신을 증가시키기 위한 것이므로 제 1종 오류의 보정이 요구되지는 않는다.

2.3 다중성 문제 해결을 위한 대안적인 통계적 방법

서로 다른 통계적 모형이나 통계적 기법 (예를 들어 모수적 방법 vs 비모수적 방법 혹은 윌콕슨 검정 vs 로그 순위검정)이 동일한 자료에서 시도되는 경우가 있다. 첫 번째 통계적 사전 검정 결과에 근거하여 처리효과를 비교하기 위한 특정 통계 기법을 선택하려는 목적으로 2단계 절차가 적용된다. 만약 그러한 절차가 피험자 할당 정보에 근거하여 선호되는 분석을 시행하고자 한 것이라면 다중성 문제가 즉시 발생한다. 공식적인 눈가림 검토 (Blind Review, ICH E9 지침서 참고)에 근거하여 최종 통계적 모델을 선택한다면 이러한

문제를 줄일 수 있을 것이다. 그러나 이러한 절차가 처리비교에 대한 비교정보를 사용하고 전체 제 1종 오류의 영향을 평가하기 어려운 경우 선택이 어려울 수 있다. 결국 사후분석에 근거하여 연구의 주요 특징을 변경한다면 연구의 신뢰성 및 결과의 로버스트성에 문제가 생길 수 있으며, 그 결과 추가 연구를 해야 하는 상황에 이를 수 있다. 그러므로 선택의 여지가 없다 하더라도 이러한 절차는 권고될 수 없다.

2.4 안전성 변수에서의 다중성 문제

안전성 변수가 연구의 승인이나 라벨링에서 중요한 역할을 하는 경우, 관측된 효과가 반대 방향으로 나타나거나 안전성의 문제를 일으키는 경우 (3.3절을 참고)를 제외하고는 일차 유효성 평가변수와 다르게 처리되지 않아야 한다. 이상반응의 경우, 관측된 유의확률(P값)과 상관없이 중대성, 중증도 혹은 결과에 따른 실질적 차이에 따라 우려가 제기되므로, 유의확률은 매우 제한적인 의미를 가진다.

이와 같이 수많은 통계적 검정 절차가 시험약에 의해 야기되는 잠재적 위험의 신호를 알리는 조건 장치 (flagging device)로써 사용되는 경우, 다중성 보정이 안전성에 역효과를 초래한다고 일반적으로 말할 수 있다. 이러한 경우 단일가설에 대한 제 1종 오류를 제어할 방법이 없음이 분명하며, 그러한 결과의 중요성 및 타당성은 약물의 약리학적 사전지식에 좌우될 것이다.

2.5 둘 이상의 처리군이 있는 임상시험에서의 다중성 문제

하나 이상의 일차 유효성 평가변수가 있고, 처리군이 두 개 이상인 임상시험의 경우 적절한 평가와 해석이 굉장히 복잡할 수 있다. 확증적 임상시험에서는 이러한 복잡한 설계를 적용하는 경우가 드물기 때문에 본 문서에서는 다중 처리군 관련 연구에 대한 모든 주제를 논의하지는 않을 것이다. 그러므로 다음의 논의는 보다 일반적이고 간단한 설계로 제한한다. 일반적인 규정에 따르면 넓은 의미의 전체 제 1종 오류의 조절 (즉, 닫힌 검정 절차의 적용)은 확증적 주장을 위한 최소한의 전제 조건이다. 처리군 간 차이에 대한 일반적인 신뢰구간은 몇 가지 경우를 제외하고는 닫힌 검정 절차와 일치하지 않고 일반적으로 매우 좁다는 점을 명심한다.

2.5.1 세 군의 '표준' 설계

일반적으로 널리 쓰이는 대조약이 존재하는 질병에 있어 신약의 유효성과 안전성을 입증하고자 할 경우, 윤리적 타당성을 전제로 대조약, 위약, 시험약 세 군을 사용하도록 권고한다. 이러한 연구의 목적은 보통 ① 시험약이 위약보다 효과가 있는가를 입증 (유효성 검증) ② 대조약이 위약보다 효과가 있는가를 입증 (민감도 평가 검증, ICH E10 지침서 2.5.1.1.1절을 참고하라) ③ 시험약이 최소한 위약 대비 대조약의 효과만큼은 되는가를 입증 (비열등성 검증) 하는 것 등으로 다양하다. 만약 이 모두가 목적이라면, 요구되는 유의수준

에서 세 가지 비교에 대한 통계적 유의성을 입증해야만 하며 제 1종 오류에 대한 보정은 필요하지 않다. 위약보다 시험약이 효과가 더 좋은지를 입증하는 것에 실패했다면 시험약이 효과가 없거나 (대조약이 위약보다 우월한 것이 입증된 경우), 민감도 평가가 부족 (시험약과 대조약이 위약보다 우월함을 입증하는데 실패한 경우)하기 때문이라고 설명될 수 있다.

2.5.2. 고정 조합에 대한 유효성의 검증

복합 의약품의 고정 조합에 대한 CPMP 지침서 (CPMP/EWP/240/95)를 보면 '고정 조합을 구성하는 각 물질의 복합 내 기여도를 문서화해야 한다.' 라고 되어 있다. 이에 따르면 두 (단일) 성분으로 이루어진 복합제의 경우, 두 성분에 대한 각각의 단독 처리군들과 조합으로 처리한 군, 세 군으로 이루어진 연구를 수행해야 한다. 만약 조합된 처리군이 두 단독 처리군보다 우월하다고 입증된다면 해당 연구는 성공적이라고 간주할 수 있다. 각각의 두 군간 비교에서 통계적으로 유의한 우월성이 나타나야 하기 때문에 전체 유의수준을 공식적으로 보정할 필요가 없다.

다중-용량 요인설계 (multiple-dose factorial design)는 복합제의 평가에 사용되며 그 목적은 ①복합 성분이 단일 성분보다 더욱 효과적이라는 확정적 증거를 제공하고 (ICH E4 지침서 참고), ②사용 대상 피험자 모집단에서 권장되는 (또는 유용한 용량 복합의 범위) 유효하고 안전한 용량 조합을 확인하는 것이다. 목적이 ①인 경우에는 보통 전반적인 검정 기법을 사용하는 반면, 목적이 ②인 경우에는 적절한 닫힌 검정 절차를 적용해야 한다.

2.5.3 용량-반응 연구

특정한 피험자 모집단에 권장되는 시험약에 대한 하나의 또는 여러 가지 용량을 확인하는 것을 목적으로 하는 용량-반응 연구를 수행하는 경우에는 반드시 전체 제 1종 오류를 조절해야 한다. 이러한 연구의 설계 특징, 가정 및 목적은 상당히 다양하기 때문에 (예를 들어 용량 증가 시 용량-반응 관계의 단조성에 대한 가정 여부, 사용된 설계의 제한 하에서 최소한의 효과적인 용량을 찾는 것, 대조약의 추천 용량과 동등 (비열등)한 용량을 찾는 것 등) 본 문서에서는 구체적인 권고사항을 다루지 않을 것이다. 특정 목적에 적용할 수 있고 제 1종 오류를 조절해야 하는 다중-용량 연구에 사용되는 닫힌 검정 절차에 대해서는, 관련 문헌에 다양한 방법이 공개되어 있다.

용량 (혹은 용량의 범위)의 유효성과 안전성을 확인하고 권고하는 것이 목적인 연구는 검정력이 충분하지 않지만, 용량이 증가함에 따라 임상적 효과와 용량의 양의 상관관계가 입증되는 경우가 있다 (ICH E4 지침서의 3.1절 참고하라). 단일 용량간의 쌍체 비교로부터 얻은 추정과 신뢰구간은 향후 연구 계획을 위한 탐색적 방법으로 사용된다. 이 경우 제 1종 오류의 보정은 필요 없다.

3. 다중 이차 유효성 평가변수에 대한 유의성을 어떻게 해석할 것이며, 이 중 하나를 토대로 주장할 수 있는 경우는 언제인가?

전통적으로 임상시험계획서에는 많은 이차 유효성 평가변수가 존재한다. 지금까지는 임상시험에서 이차 유효성 평가변수의 역할과 비중에 대해 공통적인 합의를 이룬 바가 없다.

3.1 유효성을 뒷받침하기 위해 사용되는 이차 유효성 평가변수

주장을 하려는 것은 아니다. 신뢰구간과 통계적 검정은 탐색적 성격을 띤다.

이차 유효성 평가변수는 처리효과에 대한 추가적인 임상적 특징을 제공할 수 있지만 그것만으로 시판허가나 추가적인 라벨링을 위한 주요한 증거가 될 수는 없다. 이차 유효성 평가변수는 일차목적과 관련된 연구자의 가설을 뒷받침하고자 하기 위함이므로 확증적 결론은 필요하지 않다. 이차 유효성 평가변수의 신뢰구간과 통계적 검정은 탐색적 성격일 뿐 주장을 의도하는 것은 아니기 때문이다.

3.2 추가적인 주장의 근거가 될 수 있는 이차 유효성 평가변수

이차 유효성 평가변수들의 유의한 효과는 임상시험의 일차목적이 입증된 후, 그리고 확증적 전략의 일부분인 경우에만 추가적인 주장으로 고려될 수 있다.

추가 주장의 근거가 되는 이차목적과 연관이 있는 이차 유효성 평가변수가 추가적인 가설을 뒷받침하기 위해서는 먼저 일차목적이 입증되어야 한다(2.1.2절 참고). 이러한 이차 유효성 평가변수를 처리하는 타당한 절차는 계층적으로 진행된다. 일차목적에 대한 귀무가설이 기각되면 (따라서 일차목적이 입증되면) 이차 유효성 평가변수가 하나 이상인 경우 이차 유효성 평가변수에 대한 추가적인 확증적 통계 검정을 계층적 순서에 의해 수행할 수 있다. 이러한 경우 일차 유효성 평가변수와 이차 유효성 평가변수는 가설의 계층적 구조의 위치만 다르며, 이는 두 변수가 연구에서 차지하는 상대적 중요성을 반영한 것이다. 일차 유효성 평가변수의 변화에 대한 직접적인 결과로 간주되는 이차 유효성 평가변수의 변화는 라벨링 주장의 부분이 될 수 없다는 점에 주의하여야 한다. 예를 들어 정신분열증 피험자들의 우울증 증상이 정신분열증상의 완화로 인해 사라진 경우, 치료의 항 우울성 작용 (anti-depressive action)에 대한 라벨링은 인정되지 않는다.

3.3 임상적 유용성을 나타내는 변수

임상적으로 유용성이 관찰되었으나 일차 목적을 달성하지 못했을 경우, 임상적으로 매우 중요한 변수 (사망률 등)가 일차 유효성 평가변수로 정의되지 않았다면, 이것에 대한 추가 연구가 필요하다.

유용성을 입증하기에는 계획하는 임상시험의 표본수가 너무 작다는 (따라서 검정력이 너무 낮다는) 선험적 확신이 있기 때문에, 중요한 임상적 유용성의 지표가 될 가능성이 있거나 다른 상황에서 중요한 안전성 문제를 나타낼 수도 있는 변수 (사망률 등)를 이차 유효성 평가변수로 하향 조정하는 경우가 있다. 이차 평가변수의 관찰된 유용성 효과가 예상보다는 훨씬 크지만, 임상시험의 일차목적이 입증되지 않은 경우, 추가 연구를 수행하여 관측된 유용성 효과를 뒷받침할 수 있는 정보를 확보해야 한다.

그러나 중요한 임상적 유용성의 지표가 될 수도 있는 변수의 효과가 반대 방향으로 나타난다면, 이것은 안전성 문제를 야기하게 된다. 이와 같은 변수가 확증적 계획에 포함되었는지 여부와 관계없이 허가가 거부될 수 있다.

4. 하위집단 분석에 대한 신뢰성 있는 결론과 허가의 제한

일반적으로 하위집단 분석을 통해 믿을만한 결론을 얻기 위해서는 하위집단의 대상을 사전에 구체화하고 적합한 통계 분석 기법을 사용해야 한다. 처리효과가 중요 부분모집단 간에 상당한 차이가 있을 때 그것을 설명하지 못하거나, 부분모집단 간에 처리효과가 상당히 다를 것이라고 분명히 예상되는 데도 불구하고 중요한 각 부분모집단에 대한 서로 다른 처리효과를 충분히 입증하지 못한 경우에는 허가가 제한될 수도 있다.

임상시험에서 하위집단의 치료효과를 조사하는 데에는 다양한 이유가 있다. 많은 연구에서 하위집단 분석은, 일차목적이 입증된 이후, 이를 확고히 하거나 탐색적인 역할로서 수행된다. 즉, 전반적인 임상적 효과의 유의성을 설명해주는 것이다. 특정 하위집단에서 임상적 효과에 대한 구체적인 주장을 하기 위해서는 이에 대한 가설 및 적절한 확증적 분석 전략이 사전에 구체화되어야 한다. 전체 모집단에서 유의한 효과가 없는 경우 하위집단 분석에 근거한 주장이 채택될 확률은 매우 낮다. 검정력이 계획서 상에서 기술되고 무작위배정은 일반적으로 층화를 사용하도록 한다.

모든 하위집단에서 치료효과가 일관성 있게 나타나는지를 평가하는 것은 일반적인 규제 관련 사항이다. 성별, 나이, 지역, 질병의 중증도, 인종, 신장의 손상 혹은 흡수나 신진대사의 차이와 같은 몇몇 요인은 처리효과에 대한 이질성의 원인으로 알려져 있다. (타당한 경우에는) 이런 중요한 하위집단을 분석하는 작업이 임상시험 평가의 통상적 부분이 되어야 하지만, 위와 같은 요인 중 하나 이상이 효과의 크기에 영향을 줄 것이라는 선험적 의심이 존재하지 않는다면 보통 탐색적으로 고려되어져야 한다. 그러나 하위집단 중 하나에서 치료의 역효과가 나타나는 강한 교호 작용이 발견되었고 이 현상을 납득시킬 설명이 없거나 다른 정보를 통해 교호작용의 가능성이 명확해진 경우, 추가적인 임상 자료가 얻어지기 전까지 해당 하위집단의 피험자에 대해서는 허가가 제외될 수도 있다. 또한 피험자의 특정 하위집단이 해당 약을 사용하여 효과를 보지 못할 것이라는 규제기관의 과거 증거가 있고 연구 결과가 이러한 믿음을 강하게 반박할 수 없는 경우에도 허가가 제한될 수

있다.

또한 하위집단이 매우 다양할 경우 이런 상황에 대처하는 적절한 계획을 계획서에 기술하지 않은 채로 연구를 수행한다면 특정 하위집단에 대한 허가가 제한될 수 있다. 규제기관의 입장에서는 각 하위집단에서 처리효과가 다르게 나타날 거라고 예상될만한 근거가 존재한다면 전체 연구 집단에서 (통계적 및 임상적으로) 전체적인 양성 결과가 나타났다 하더라도 모든 하위모집단에 대해 주장의 타당성을 인정할 수는 없을 것이다. 전체 연구 모집단에 대한 의미있는 정의가 결여된 경우에는, 대표성이 적절하고 통계적으로 유의하면서 임상적 적절성이 있는 결과가 관측된 하위모집단에 대해서만 허가가 내려질 수 있다.

5. 원변수를 반응과 비반응으로 분류하여 분석한 결과에 대한 해석

"반응/비반응" 분석이 일차 분석이 아닌 경우, 일차 유효성 평가변수의 평균수준에 대한 통계적 유의성을 입증한 후, "반응/비반응" 비율로 관측된 차이의 임상적 적절성을 입증하는 것이 사용될 수 있다. 그러나 원래의 일차 유효성 평가변수에 대한 분석이 반응/비반응 분석보다 검정력이 높음에 유의하여야 한다.

알츠하이머병이나 간질 등 많은 예시에서, 통계적으로 유의하긴 하지만 크기가 작은 개선 정도를 일차 유효성 평가변수의 평균을 이용하여 해석하기는 쉽지 않다. 이러한 이유에서 결과를 "반응"과 "비반응"으로 나누어 각 피험자에 대한 치료의 임상적 효과를 평가하는데 사용한다. "반응/비반응"을 정의하는 방법은 다양하게 존재한다. 그러나 사전에 임상시험 계획서에 반응/비반응에 대한 정의가 기술되어야 하며 임상적으로 설득력을 가져야 한다. 임상시험지침서들에 의하면 "반응/비반응" 분석은 단지 유효성과 안전성을 평가하는 보조적 역할로써 관측된 효과의 임상적 적절성을 입증하는데 사용되어야 한다고 언급되어 있다. 원 변수에 포함된 정보를 반응과 비반응으로 나누면서 어느 정도 정보가 손실된다는 점에 (따라서 통계적인 검정력도 손실된다는 점에) 유의해야 한다.

일부 상황에서는 "반응/비반응" 척도가 일차 유효성 평가변수가 될 수도 있다. (CPMP 파킨슨병 치료제 연구에 대한 임상시험 지침서를 참고하라). 이 경우에는 반응/비반응 척도를 이용한 귀무가설이 주요 검정으로 수행되어야한다. 그러나 본 절에서는 일차 유효성 평가변수의 평균을 이용하여 통계적으로 유의한 처리효과를 입증한 후 반응 비율을 이용하여 임상적인 적절성을 판단하고자 하는 상황을 주로 다룬다 (예를 들어 CPMP 비만에 사용되는 임상시험 지침서 혹은 알츠하이머병에 대한 임상시험 지침서). 이 경우 "반응/비반응" 분석 결과는 통계적으로 유의할 필요는 없지만 반응 비율은 임상적으로 적절한 효과가 있다는 것을 지지할 만큼의 차이를 보여야만 한다.

그러나 "반응/비반응" 분석이 일차 유효성 평가변수의 유의하지 않은 결과를 구제할 수 없다는 것을 알아야 한다.

6. 규제기관 입장에서의 복합변수에 대한 통계적 처리

복합변수가 일차 유효성 평가변수인 경우가 많이 있다. 이때 복합변수를 구성하는 모든 변수들은 개별적으로 분석되어야 한다. 만약 연구자의 주장이 복합변수를 구성하는 변수들의 부분집단에 근거한다면, 이를 사전에 구체화하고 타당한 확증적 분석 전략에 삽입해야 한다. 처리효과가 이를 구성하는 모든 변수에 대해 유용하게 나타나거나 혹은 최소한 임상적으로 보다 중요한 변수에서 부정적으로 나타나지 않아야 한다. 특정 적응증(indication)을 반영하는 변수들에 대한 처리효과는 자료에 의해 명백히 지지되어야 한다.

복합변수에는 두 가지 유형이 있다. 첫 번째는 평가 척도로서, 여러 임상 측정값들의 조합으로 만들어지는 경우이다. 이러한 복합변수는 특정 적응증 (예를 들면 정신질환이나 신경계장애)에서 오랜 동안에 걸쳐서 연구되고 있으므로 본 문서에서 더 이상 논의하지 않을 것이다.

다른 하나는 생존분석에서 발생한다. 여러 사건을 결합하여 하나의 복합결과가 된다. 미리 언급된 복합변수를 구성하는 모든 변수 (예를 들어 사망, 심근 경색 혹은 뇌졸중) 중 하나 이상의 사건이 피험자에게 발생한다면, 피험자는 임상적 결과를 갖는다고 할 수 있다. 결과가 발생할 때까지의 시간은 피험자를 무작위 배정한 시점부터 목록에 존재하는 어떤 사건이라도 최초로 발생한 시점까지의 시간으로 측정된다. 대개 각 변수는 상대적으로 드문 사건을 나타내므로 각 변수를 개별적으로 연구하려면 표본 크기가 엄청나게 커야 할 것이다. 따라서 복합변수는 임상적 결과에 도달하는 피험자의 비율을 증가시키고 그에 따라 연구의 검정력도 증가시키려는 목적으로 사용된다.

6.1 일차 유효성 평가변수로서의 복합변수

복합변수가 유효성을 입증하는 데 사용될 때에는 일반적으로 복합변수가 일차 유효성 평가변수가 될 것이다. 그러므로 일차 유효성 평가변수로 사용되는 복합변수는 단일 일차 유효성 평가변수의 요건을 충족시켜야 한다. 즉, 허가에 필요한 핵심적 유효성 근거를 제공할 수 있어야 한다. 설득력 있는 정보를 제공하기 위해서, 각각의 단일 변수들뿐 아니라 임상적으로 중요한 변수군들도 추가로 분석하도록 권고된다. 그러나 이때 일차 유효성 평가변수의 유의성이 입증되었다면 다중성의 보정은 필요 없다. 만약 유효성 평가가 변수의 하위집단에 근거하였다면 이를 사전에 구체화하고 타당한 확증적 분석 전략에 삽입해야 한다.

6.2 처리는 모든 변수에 동일한 방향으로 영향을 줄 것으로 예상되어야 한다.

복합변수를 정의하는 경우 처리가 각 변수에 유사하게 영향을 줄 것이라 가정되는 변수들만을 포함하도록 권고된다. 모든 변수에서 처리효과가 유사한 방향으로 이루어질 것이라는 가정은 유형이 비슷한 과거 연구 자료에 근거하여야 한다. 처리효과에 둔감할 것으로

예상되는 변수를 추가한다면, 처리 차에 대한 추정치의 불편성에 영향을 주지 않더라도 변동을 증가시킬 것이다. 그 직접적인 결과로 서로 다른 처리군 사이에서 우월성의 입증에 대한 민감성이 감소될 것이다. 분산의 증가는 비열등성이나 동등성 연구에서도 바람직하지 않은 성질이다. 만약 몇 개의 변수에서 효과가 반대 방향으로 관측되면, 비열등성 연구는 해석되기 힘들 것이다. 우월성을 입증하는 연구에서는 일반적인 변수를 사용하는 것이 가장 보수적인 분석이므로 일차 유효성 평가변수로 선호된다. 비열등성/동등성을 입증하는 연구에서는 보다 구체적인 변수 (예를 들어 사망과 관련된 질병 등)가 일차 유효성 평가변수로 선호된다.

6.3 임상적으로 중요하다고 여겨지는 변수는 부정적인 영향을 받지 않아야 한다.

임상시험에서 입원 기간이 유효성 평가변수일 때, 병원에 도달하기 전에 사망한 피험자를 중도 절단으로 처리하는 것은 일반적으로 적절하지 않다. 위 예제에서는 사망을 포함하여 중요한 임상적 사건을 변수로 포함하는 복합 평가변수를 사용하는 것이 더 적절하다. 그러나 규제적인 관점에서 복합 평가변수에 대해 우려되는 사항은, 몇몇 처치 군에서 복합변수를 구성하는 하나 이상의 변수에서 역효과가 존재하고 그러한 역효과가 복합변수의 결과에 의해 (예를 들어 나머지 변수 중 일부에서 보인 큰 효과에 의해) 가려지는 경우이다. 이러한 우려는 변수들이 질병의 중증도나 임상적 중요성에 서로 다른 정도로 관련이 있는 경우에 특히 관련된다. 예를 들어 모든 원인에 의한 사망이 변수일 경우, 이런 평가변수에 역효과가 없다는 것을 확인하기 위해서는 모든 원인에 의한 사망을 개별적으로 분석해야 한다. 통계적으로 유의성이 나쁜 (좋지 않은) 방향 쪽으로 얼마나 작게 나타나야 역효과를 의심하게 될지에 대한 일반적인 합의가 없기 때문에 일단 자료가 관측이 되면 '역효과가 존재하지 않음'을 지지할 수 있는 방법을 계획단계에서 언급해야 한다. 예를 들어 연구 계획은 처리 차이가 없다는 귀무가설 하에서 계획된 피험자 수에 대하여 충분히 큰 확률로 배제시킬 수 있는 중요한 변수의 역효과 위험도의 크기를 언급하여야 하고, 문서에는 각각을 비교한 추정치와 신뢰구간을 포함해야 한다.

6.4 적응증에 반영되는 변수에 대한 처리 효과는 반드시 자료에 근거하여야 한다.

복합 평가변수의 성공적인 일차분석 결과에 근거한 적법한 주장에 대해 고려해야 할 문제가 있다. 만약 그 주장이 복합 평가변수가 사용되었다는 사실을 제대로 반영하지 못한다면 즉, 복합 평가변수의 결과가 드물게 발생하는 하나의 변수와 명백하게 관련된 것이라면 문제가 발생한다. 예를 들어 복합 평가변수가 '사망 또는 간이식'이고 사망이 매우 적게 발생한다면, '사망률과 동시에 간이식의 필요성을 낮춘다'라는 주장은 사망률에 미치는 영향의 근거가 약하기 때문에 적절하지 않다. 그러나 이 복합변수로부터 사망을 제외해야 한다는 의미는 아니다. 왜냐하면 최소한 간이식만큼 중대한 질병과 관련된 결론을 동시에 고려하지 않는다면, '간이식의 필요성을 낮춘다'라는 주장은 불완전하게 될 것이기 때문이다. 적응증에 대하여 사망에 영향을 준다는 의미를 피해 다른 표현이 채택되야 함을 의미한다.

7. 결론

특정 질병에서 하나 이상 처리의 유효성 (혹은 안전성)에 대하여 하나 이상의 의문사항에 답변해야 하는 경우가 종종 있다. 이는 임상연구에서 신약개발 프로그램의 성공은 하나 이상의 의문사항에 하나의 긍정적 답변을 하는 것에 의존할 수 있기 때문이다. 통계적 다중검정에서 위양성이 증가되는 것을 조절하지 않는다면, 의문 사항이 많아질수록 위양성률이 증가한다는 것은 잘 알려진 사실이다. 이런 점에서 다중분석으로부터 선호하는 결과를 선택할 가능성이 우려된다. 따라서 다중성을 다루거나 피하기 위해 계획된 통계적 절차는, 적당하고 적합한 평가를 위해서 임상시험계획서나 통계분석 계획에 상세하게 언급되어야 한다.

위양성률을 조절하기 위한 다양한 방법이 개발되었다. 그러나 이러한 방법 모두가 임상적으로 해석할 수 있는 결과를 동일한 수준으로 제시하는 것은 아니므로 항상 이런 측면을 고려해야 한다. 처리효과의 추정이 대부분 중요한 사항이기 때문에, 신뢰구간을 제공하는지 여부가 선택 기준이 될 수도 있다.

만약 각 의문점들이 사전에 구체화되어 있고 통계분석 전략이 적절하다면, 이차 유효성 평가변수나 하위집단에 근거한 통계적 유의성과 임상적으로 적절한 조사결과에서의 추가적인 주장 (가설)은 임상시험의 일차목적을 달성한 후에 가능하다.

COMMITTEE FOR PROPRIETARY MEDICINAL PRODUCTS
(CPMP)

POINTS TO CONSIDER ON MULTIPLICITY ISSUES IN CLINICAL TRIALS

London, 19 September 2002
CPMP/EWP/908/99

DISCUSSION IN THE EFFICACY WORKING PARTY	January 2000
TRANSMISSION TO CPMP	July 2001
RELEASE FOR CONSULTATION	July 2001
DEADLINE FOR COMMENTS	October 2001
DISCUSSION IN THE EFFICACY WORKING PARTY	June 2002
TRANSMISSION TO CPMP	September 2002
ADOPTION BY CPMP	September 2002

TABLE OF CONTENTS

1. INTRODUCTION

Multiplicity of inferences is present in virtually all clinical trials. The usual concern with multiplicity is that, if it is not properly handled, unsubstantiated claims for the effectiveness of a drug may be made as a consequence of an inflated rate of false positive conclusions. For example, if statistical tests are performed on five subgroups, independently of each other and each at a significance level of 2.5% (one-sided directional hypotheses), the chance of finding at least one false positive statistically significant test increases to 12%.

This example shows that multiplicity can have a substantial influence on the rate of false positive conclusions which may affect approval and labelling of an investigational drug whenever there is an opportunity to choose the most favourable result from two or more analyses. If, however, there is no such choice, then there can be no influence. Examples of both situations will be discussed later. Control of the study-wise rate of false positive conclusions at an acceptable level < is an important principle and is often of great value in the assessment of the results of confirmatory clinical trials.

A number of methods are available for controlling the rate of false positive conclusions, the method of choice depending on the circumstances. Throughout this document the term ‘control of type I error’ rate will be used as an abbreviation for the control of the family-wise type I error in the strong sense, i.e., there is control on the probability to reject at least one true null hypothesis, regardless which subset of null hypotheses happens to be true. The issue of setting an appropriate type I error level on a submission level when this includes the need for more than one confirmatory trial is discussed in a separate Points-to-Consider document (CPMP/2330/99 Points to Consider on Application with 1.) Meta-analyses and 2.) One Pivotal study).

This document does not attempt to address all aspects of multiplicity but mainly considers issues that have been found to be of importance in recent European applications. These are:

- ☐ Adjustment of multiplicity - when is it necessary and when is it not?
- ☐ How to interpret significance with respect to multiple secondary variables and when can a claim be based on one of these?
- ☐ When can reliable conclusions be drawn from a subgroup analysis?

- When is it appropriate for CPMP to restrict licence to a subgroup?

- How should one interpret the analysis of "responders" in conjunction with the raw variables?

- How should composite endpoints be handled statistically with respect to regulatory claims?

There are further areas concerning multiplicity in clinical trials which, according the above list of issues, are not the focus of this document. For example, there is a rapid advance in methodological richness and complexity regarding interim analyses (with the possibility to stop early either for futility or with the claim of effectiveness) or stepwise designed studies (with the possibility for adaptive changes for the future steps). However, due to the importance of the problem and the amount of information specific to this issue it appears appropriate that a separate document may cover these aspects.

Interpretations of repeated evaluations of the primary efficacy variable at repeated visits usually do not cause multiplicity problems, because in the majority of situations either an appropriate summary measure has been pre-specified or according to the requirements on the duration of treatment endpoint, primary evaluations are made at pre-specified visits.
Therefore potential multiplicity issues concerning the analysis of repeated measurements are not considered in this document.

2. ADJUSTMENT FOR MULTIPLICITY – WHEN IS IT NECESSARY AND WHEN IS IT NOT?

A clinical study that requires no adjustment of the type I error is one that consists of two treatment groups, that uses a single primary variable, and has a confirmatory statistical strategy that pre-specifies just one single null hypothesis relating to the primary variable and no interim analysis. Although all other situations require attention to the potential effects of multiplicity, there are many situations where no multiplicity concern arises, for example, having predefined the primary variables and all secondary variables are declared supportive.

In the literature, methods to control the overall type I error < are sometimes called multiple-level-α-tests". Controlling type I error family-wise often (but not always) means that the accepted and pre-specified amount < of type I error has to be split, and that the various null hypotheses have to be tested at the resulting fraction of <. This is

usually referred to as 'adjusting the type I error level'. The algorithms that define how to "spend" < in this way are of different complexity. Often, for the more complex procedures, clinical interpretation of the findings can become difficult. For example, for the purpose of estimation and for the appraisal of the precision of estimates, confidence intervals are of paramount importance but methods for their construction that are consistent with the tests are not available for many of the more complex multiple-level-α-tests (or more generally closed tests) aiming at controlling the type I error. When choosing an approach, it is recommended to consider whether the existing valid statistical procedures allow a satisfactory clinical interpretation.

Because alternative methods to deal correctly with multiplicity are often available which may lead to different conclusions, pre-definition of the preferred multiple-level-α-test is necessary.
To avoid problems in interpretation, details of the procedure should be contained in the study protocol or the statistical analysis plan.

If a multiple test situation occurs which was not foreseen, a conservative approach will be necessary e.g. Bonferroni's or a related procedure. Inherently there will be a loss of power.
Therefore if a multiple test situation is foreseen pre-specification of the method use to deal with this is recommended

This document discusses situations with relevance for multiple testing in clinical trials and commonly practised and acknowledged methods for controlling (or adjusting) type I error.

2.1 Multiple primary variables - when no formal adjustment is needed.

The ICH E9 guideline on biostatistical principles in clinical trials recommends that generally clinical trials have one primary variable. A single primary variable is sufficient, if there is a general agreement that a treatment induced change in this variable demonstrates a clinical relevant treatment effect on its own. If, however, a single variable is not sufficient to capture the range of clinically relevant treatment benefits, the use of more than one primary variable may become necessary. Sometimes a series of related objectives is pursued in the same trial each with its own primary variable, and in other cases, a number of primary variables are investigated with the aim of providing convincing evidence of beneficial effects on some, or all of them. In these situations planning of the sample size becomes more complex because alternative hypotheses and limits for the power of the single primary variables have to be defined and balanced

against each other to give the study a solid basis to meet its objectives.

For trials with more than one primary variable the situations described in the following subsections can be distinguished. The methods described allow clinical interpretation, deal satisfactorily with the issue of multiplicity but avoid the need for any formal adjustment of type I error rates. Indeed the methods are members from the set of closed testing procedures that control the family-wise error rate.

2.1.1. Two or more primary variables are needed to describe clinically relevant treatment benefits

Statistical significance is needed for all primary variables. Therefore, no formal adjustment is necessary.

Here, interpretation of the results is most clear-cut because, in order to provide sufficient evidence of the clinically relevant treatment benefit, each null hypotheses on every primary variable has to be rejected at the same significance level (e.g. 0.05). For examples of this clinical situation, see CPMP Note for Guidance for the treatment of Alzheimer's disease, or CPMP Points to Consider on clinical investigation of medicinal products in the chronic treatment of patients with chronic obstructive pulmonary disease. In these situations, there is no intention or opportunity to select the most favourable result and, consequently, the individual type I error levels are set equal to the overall type I error level <, i.e. no reduction is necessary. This procedure inflates the relevant type II error (here: falsely accepting that at least one null hypothesis is true), which in the worst case scenario is the sum of the type II errors connected with the individual hypotheses. This inflation must be taken into account for a proper estimation of the sample size for the trial.

2.1.2. Two or more primary variables ranked according to clinical relevance

No formal adjustment is necessary. However, no confirmatory claims can be based on variables that have a rank lower than or equal to that variable whose null hypothesis was the first that could not be rejected.

Sometimes a series of related objectives is pursued in the same trial, where one objective is of greatest importance but convincing results in others would clearly add to the value of the treatment. Typical examples are (i) acute effects in depressive disorders followed by prevention of progression (ii) reduction of mortality in acute myocardial infarction followed by prevention of other serious events. In such cases the hypotheses

may be tested (and confidence intervals may be provided) according to a hierarchical strategy. The hierarchical order may be a natural one (e.g. hypotheses are ordered in time or with respect to the seriousness of the considered variables) or may result from the particular interests of the investigator. Again, no reduction or splitting of < is necessary. The hierarchical order for testing null hypotheses, however, has to be pre-specified in the study protocol. The effect of such a procedure is that no confirmatory claims can be based on variables that have a rank lower than or equal to that variable whose null hypothesis was the first that could not be rejected. Confidence intervals that are consistent with this hierarchical test procedure can be derived. Evidently, type II errors are inflated for hypotheses that correspond to variables with lower ranks. Note that a similar procedure can be used for dealing with secondary variables (see 3.2).

In the literature it is possible to find many methods of dealing with multiple variables that are of value for situations which may, however, be rarely met in confirmatory clinical trials, and which, therefore, are not discussed in this document. Before applying such methods regulatory dialogue is recommended.

2.2 Analysis sets

Multiple analyses may be performed on the same variable but with varying subsets of patient data. As is pointed out in the ICH E9 guideline on biostatistical principles for clinical trials, the set of subjects whose data are to be included in the main analyses should be defined in the statistical section of the study protocol. From these sets of subjects one (usually the full set) is selected for the primary analysis.

In general, multiple analyses on varying subsets of subjects or with varying measurements for the purpose of investigating the sensitivity of the conclusions drawn from the primary analysis should not be subjected to adjustment for type I error. The main purpose of such analyses is to increase confidence in the results obtained from the primary analysis

2.3 Alternative statistical methods – multiplicity concerns

Different statistical models or statistical techniques (e.g. parametric vs. non-parametric or Wilcoxon test versus log rank test) are sometimes tried on the same set of data. Sometimes a two step procedure is applied with the purpose of selecting a particular statistical technique for the main treatment comparison based on the outcome of the first statistical (pre-) test.

Multiplicity concerns would immediately arise, if such procedures offered obvious opportunities for selecting a favourable analysis strategy based on knowledge of the patients' assignment to treatments. There are situations, where selecting the final statistical model based on a formal Blind Review (see ICH E9) is exempted from such concerns. Opportunities for choice in such procedures are often subtle, when these procedures use comparative treatment information, and the influence on the overall type I error is difficult to assess.
Finally, the need to change important key features of a study on a post hoc basis may question the credibility of the study and the robustness of the results with the possible consequence that a further study will be necessary. Therefore, such procedures cannot be recommended even when it appears that there is no element of choice.

2.4 Multiplicity in safety variables

When a safety variable is part of the confirmatory strategy of a study and thus has a role in the approval or labelling claims, it should not be treated differently from the primary efficacy variables, except for the situation that the observed effects show in the opposite direction and may raise a safety concern (see also 3.3). In the case of adverse effects p-values are of very limited value as substantial differences (expressed as relative risk or risk differences) will raise concern, depending on seriousness, severity or outcome, irrespective of the p-value observed.

In those cases where a large number of statistical test procedures is used to serve as a flagging device to signal a potential risk caused by the investigational drug it can generally be stated that an adjustment for multiplicity is counterproductive for considerations of safety. It is clear that in this situation there is no control over the type I error for a single hypothesis and the importance and plausibility of such results will depend on prior knowledge of the pharmacology of the drug.

2.5 Multiplicity concerns in studies with more than two treatment arms

As for studies with more than one primary variable, the proper evaluation and interpretation of a study with more than two treatment arms can become quite complex. This document is not intended to provide an exhaustive discussion of every issue relating to studies with multiple treatment arms, only rarely have these more complex designs been applied in confirmatory clinical trials. Therefore, the following discussion is limited to the more common and simple designs. As a general rule it can be stated that control of the family-wise type I error in the strong sense (i.e. application of closed test procedures) is a minimal prerequisite for confirmatory claims. It should be remembered

that the usual confidence intervals for the pairwise differences between treatment groups are – except for a few instances – not consistent with the closed testing procedures, and are usually too narrow.

2.5.1 The three arm 'gold standard' design

For a disease, where a commonly acknowledged reference drug therapy exists, it is often recommended (when this can be justified on ethical grounds) to demonstrate the efficacy and safety of a new substance in a three arm study with three treatments: the reference drug, placebo and the investigational drug. Usually the aims of such a study are manifold: (1) to demonstrate superiority of the investigational drug over placebo (proof of efficacy); (2) to demonstrate superiority of the reference drug over placebo (proof of assay sensitivity, see ICH E10, section 2.5.1.1.1); and (3) to demonstrate that the investigational drug retains most of the efficacy of the reference drug as compared to placebo (proof of non-inferiority). If all of these are objectives , all three comparisons must show statistical significance at the required level, and no formal adjustment is necessary. A failure to show the investigational drug as superior to placebo could then be explained either as the investigational drug being not effective (when the reference drug showed superiority over placebo), or as lack of assay sensitivity (when test and reference drug failed to show superiority over placebo).

2.5.2 Proof of efficacy for a fixed combination

For fixed combination medicinal products the corresponding CPMP guideline (CPMP/EWP/240/95) requires that 'each substance of a fixed combination must have documented contribution within the combination'. For a combination with two (mono) components, this requirement has often been interpreted as the need to conduct a study with the two components as mono therapies and the combination therapy in a 3-arm study. Such a study is considered successful, if the combination is shown superior to both components. No formal adjustment of the overall significance level is necessary, because both pairwise comparisons must show statistically significant superiority.

Multiple-dose factorial designs are employed for the assessment of combination drugs for the purpose (1) to provide confirmatory evidence that the combination is more effective than either component drug alone (see ICH E4 Note for Guidance on Dose Response Information to support Drug Registration (CPMP/ICH/378/95)), and (2) to identify an effective and safe dose combination (or a range of useful dose combinations) for recommended use in the intended patient population. While (1) usually

is achieved using global test strategies, appropriate closed test procedures have to be applied for the purpose of achieving (2).

2.5.3 Dose-response studies

For therapeutic dose response studies that aim at identifying one or several doses of an investigational drug for its recommended use in a specific patient population, the control of the family-wise type I error in the strong sense is mandatory. Due to the large variety of design features, assumptions and aims in such studies (e.g. assuming or not assuming monotonicity of the dose response with increasing dose; finding the minimally effective dose under the constrains of the used design; finding a dose that is equivalent (non-inferior) to the recommended dose of a reference drug), specific recommendations are beyond the scope of this document. There are various methods published in the relevant literature on closed test procedures with relevance to multiple dose studies that can be adapted to the specific aims and that provide the necessary control on the type I error.

Sometimes a study is not powered sufficiently for the aim to identify and recommend a single effective and safe dose (or a dose range) but is successful only at demonstrating an overall positive correlation of the clinical effect with increasing dose. This is already a valuable achievement (see ICH E4, section 3.1). Estimates and confidence intervals from pairwise comparisons of single doses are then used in an exploratory manner for the planning of future studies. In this case, an adjustment of the type I error is not necessary.

3. HOW TO INTERPRET SIGNIFICANCE WITH RESPECT TO MULTIPLE SECONDARY VARIABLES AND WHEN CAN A CLAIM BE BASED ON ONE OF THESE?

Traditionally, in clinical trial protocols there will be a number of secondary variables for efficacy. Up to now there has been no common consent about the role and the weight of secondary endpoints in clinical trials.

3.1 Variables expressing supportive evidence

No claims are intended; confidence intervals and statistical tests are of exploratory nature.

Secondary endpoints may provide additional clinical characterisation of treatment effects

but are, by themselves, not sufficiently convincing to establish the main evidence in an application for a license or for an additional labelling claim. Here, the inclusion of secondary variables is intended to yield supportive evidence related to the primary objective, and no confirmatory conclusions are needed. Confidence intervals and statistical tests are of exploratory nature and no claims are intended.

3.2 Secondary variables which may become the basis for additional claims

Significant effects in these variables can be considered for an additional claim only after the primary objective of the clinical trial has been achieved, and if they were part of the confirmatory strategy

More importantly, secondary variables may be related to secondary objectives that become the basis for an additional claim, once the primary objective has been established (see 2.1.2).
A valid procedure, to deal with this kind of secondary variable is to proceed hierarchically.
Once the null hypothesis concerning the primary objective is rejected (and the primary objective thus established), further confirmatory statistical tests on secondary variables can be performed using a further hierarchical order for the secondary variables themselves if there is more than one. In this case, primary and secondary variables differ just in their place in the hierarchy of hypotheses which, of course, reflects their relative importance in the study. It is of note to mention that changes in secondary variables that are considered a direct consequence of the respective changes in the primary variables cannot be part of the labelling claims. For example, symptoms of depression in schizophrenic patients disappear as patients get into remission from schizophrenia. In this situation, a separate labelling claim on an anti-depressive action of the treatment cannot be made.

3.3 Variables indicative of clinical benefit

If not defined as primary variables, clinically very important variables (e.g. mortality) need further study when significant benefits are observed, but the primary objective has not been achieved.

Variables that have the potential of being indicative of a major clinical benefit or may in a different situation present an important safety issue (e.g. mortality) may be relegated to secondary variables because there is an a priori belief that the size of the planned trial is too small (and thus the power too low) to show a benefit. If, however, the

observed beneficial effect is much higher than expected but the study fell short of achieving its primary objective, this would be a typical situation where information from further studies would be needed which can be used in support of the observed beneficial effect.

If however, the same variable that may indicate a major clinical benefit exhibits treatment effects in the opposite direction this would give rise to concerns about the safety. A license may then well be refused, regardless of whether or not the variable was embedded in a confirmatory scheme.

4. RELIABLE CONCLUSIONS FROM A SUBGROUP ANALYSIS, AND RESTRICTION OF THE LICENSE TO A SUBGROUP

Reliable conclusions from subgroup analyses generally require pre-specification and appropriate statistical analysis strategies. A license may be restricted if unexplained strong heterogeneity is found in important sub-populations, or if heterogeneity of the treatment effect can reasonably be assumed but cannot be sufficiently evaluated for important sub-populations.

In clinical trials there are many reasons for examining treatment effects in subgroups. In many studies, subgroup analyses have a supportive or exploratory role after the primary objective has been accomplished, i.e. the demonstration of a significant overall clinical benefit. A specific claim of a beneficial effect in a particular subgroup requires pre-specification of the corresponding null hypothesis and an appropriate confirmatory analysis strategy. It is highly unlikely that claims based on subgroup analyses would be accepted in the absence of a significant effect for the overall study population. Considerations of power would be expected to be covered in the protocol, and randomisation would generally be stratified.

The evaluation of uniformity of treatment effects across subgroups is a general regulatory concern. Some factors are known to cause heterogeneity of treatment effects such as gender, age, region, severity of disease, ethnic origin, renal impairment, or differences in absorption or metabolism. Analyses of these important subgroups should be a regular part of the evaluation of a clinical study (when relevant), but should usually be considered exploratory, unless there is a priori suspicion that one or more of these factors may influence the size of effect. However, when a strong interaction is found that indicates an adverse effect of the treatment in one of the subgroups and no convincing explanation for this phenomenon is available or other information confirms

the likelihood of an interaction then patients from the respective sub-population may be excluded from the license until additional clinical data are available. This may also apply when there are historical reasons for regulators to believe that a certain sub-population of patients will not benefit from the drug and the results do not strongly contradict this believe.

Restriction of a license to certain subgroups is also possible, if a large variety of sub-populations are investigated without proper plans to deal with this situation in the protocol.
From the regulatory perspective an overall positive result (statistically and clinically) in the whole study population may not lead to valid claims for all sub-populations if there is a reason to expect heterogeneity of the treatment effect in the respective sub-populations. If a meaningful definition of the overall study population is lacking, licensing may be limited to sub-populations which are adequately represented and in which statistically significant and clinically relevant results were observed.

5. HOW SHOULD ONE INTERPRET THE ANALYSIS OF "RESPONDERS" IN CONJUNCTION WITH THE RAW VARIABLES?

If the "responder" analysis is not the primary analysis it may be used after statistical significance has been established on the mean level of the required primary variable(s), to establish the clinical relevance of the observed differences in the proportion of "responders".
When used in this manner, the test of the null hypothesis of no treatment effect is better carried out on the original primary variable than on the proportion of responders.

In a number of applications, for example those concerned with Alzheimer's disease or epileptic disorders, it is difficult to interpret small but statistically significant improvements in the mean level of the primary variables. For this reason the term "responder" (and "non-responder") is used to express the clinical benefit of the treatment to individual patients. There may be a number of ways to define a "responder"/"nonresponder". The definitions should be pre-specified in the protocol and should be clinically convincing. In clinical guidelines, it is stated that the "responder" analysis should be used in establishing the clinical relevance of the observed effect as an aid to assess efficacy and clinical safety. It should be noted that there is some loss of information (and hence loss of statistical power) connected with breaking down the information contained in the original variables into "responder" and "non-responder".

In some situations, the "responder" criterion may be the primary endpoint (e.g. CPMP guideline on clinical investigation of medicinal products in the treatment of Parkinson's disease). In this case it should be used to provide the main test of the null hypothesis. However, the situation that is primarily addressed here is when the "responder" analysis is used to allow a judgement on clinical relevance, once a statistically significant treatment effect on the mean level of the primary variable(s) has been established (e.g. CPMP Note for Guidance on clinical investigation of drugs used in weight control, or on the treatment of Alzheimer's disease). In this case, the results of the "responder" analysis need not be statistically significant but the difference in the proportions of responders should support a statement that the investigated treatment induces clinically relevant effects.

It should be noted that a "responder" analysis cannot rescue otherwise disappointing results on the primary variables.

6. HOW SHOULD COMPOSITE VARIABLES BE HANDLED STATISTICALLY WITH RESPECT TO REGULATORY CLAIMS?

Usually, the composite variable is primary. All components should be analysed separately. If claims are based on subgroups of components, this needs to be pre-specified and embedded in a valid confirmatory analysis strategy. Treatment should beneficially affect all components, or at least should the clinically more important components not be affected negatively. Any effect of the treatment in one of the components that is to be reflected in the indication should be clearly supported by the data.

There are two types of composite variables. The first type, namely the rating scale, arises as a combination of multiple clinical measurements. With this type there is a longstanding experience of its use in certain indications (e.g. psychiatric or neurological disorders). This type of composite variable is not discussed further in this guideline.

The other type of a composite variable arises in the context of survival analysis. Several events are combined to define a composite outcome. A patient is said to have the clinical outcome if s/he suffers from one or more events in a pre-specified list of components (e.g. death, myocardial infarction or disabling stroke). The time to outcome is measured as the time from randomisation of the patient to the first occurrence of any of the events in the list.
Usually, the components represent relatively rare events, and to study each component

separately would require unmanageably large sample sizes. Composite variables therefore present a means to increase the percentage of patients that reach the clinical outcome, and hence the power of the study.

6.1 The composite variable as the primary endpoint.

When a composite variable is used to show efficacy it will usually be the primary endpoint. Therefore, it must meet the requirements for a single primary endpoint, namely that it is capable of providing the key evidence of efficacy that is needed for a license. It is recommended to analyse in addition the single components and clinically relevant groups of components separately, to provide supportive information. There is, however, no need for an adjustment for multiplicity provided significance of the primary endpoint is achieved. If claims are to be based on subgroups of components, this needs to be pre-specified and embedded in a valid confirmatory analysis strategy.

6.2 Treatment should be expected to affect all components in a similar way.

When defining a composite variable it is recommended to include only components for which it can be assumed that treatment will influence them similarly. The assumption of similarly directed treatment effects on all components should be based on past experience with studies of similar type. Adding a component that foreseeably is insensitive to treatment effects will lead to an increase in variability, even if it does not affect unbiasedness of the estimation of the treatment difference. A direct consequence would be a decrease in sensitivity for demonstrating superiority between different treatment arms. An increased variance is also a undesirable property in non-inferiority or equivalence studies. Non-inferiority studies will be hard to interpret if negative effects on some components are observed. For studies aiming to show superiority the more general component is preferred as primary endpoint as this is the most conservative analysis. For non-inferiority/equivalence studies the more specific component (e.g. disease related mortality) are preferred as primary endpoint for the same reason.

6.3 The clinically more important components should at least not be affected negatively

If time to hospitalisation is an endpoint in a clinical study it is not generally appropriate to handle patients as censored who die before they reach the hospital. It is better practice to study a composite endpoint that includes all more important clinical events as components, including death in this example. One concern with composite outcome measures from a regulatory point of view is, however, the possibility that some of the treatments under study may have an adverse effect on one or more of the components,

and that this adverse effect is masked by the composite outcome, e.g. by a large beneficial effect on some of the remaining components. This concern is particularly relevant, if the components relate to different degrees of disease severity or clinical importance. For example, if all cause mortality is a component, a separate analysis of all cause mortality should be provided to ensure that there is no adverse effect on this endpoint. Since there is no general agreement how much less then statistical significance in the wrong direction will generate suspicion of an adverse effect, a way to create confidence in support of 'no adverse effect', once the data is observed, is to address this issue at the planning stage. For example, the study plan could address the size of the risk of an adverse effect on the more serious components that can be excluded (assuming no treatment difference under the null hypothesis) with a sufficiently high probability, given the planned sample size, and the study report should contain the respective comparative estimates and confidence intervals.

6.4 Any effect of the treatment on one of the components that is to be reflected in the indication should be clearly supported by the data.

An important issue for consideration is the claim that can legitimately be made based on a successful primary analysis of a composite endpoint. Difficulties arise if the claims do not properly reflect the fact that a composite endpoint was used, e.g. if a claim is made that explicitly involves a component with the low occurrence. For example, if the composite outcome is 'death or liver transplantation' and there are only a few deaths, a claim 'to reduce mortality and the necessity for liver transplantation' would not be satisfactory, because in this context the effect on mortality will have a weak basis. This does not mean that one should drop the component 'death' from the composite outcome, because the outcome 'liver transplantation' would be incomplete without simultaneously considering all disease related outcomes that are at least as serious as 'liver transplantation'. However, it does mean that different wording should be adopted for the indication, avoiding the implication of an effect on mortality.

7. CONCLUSION

In clinical studies it is often necessary to answer more than one question about the efficacy (or safety) of one or more treatments in a specific disease, because the success of a drug development program may depend on a positive answer to more than a single question. It is well known that the chance of a spurious positive chance finding increases with the number of questions posed, if no actions are taken to protect against the inflation of false positive findings from multiple statistical tests. In this context,

3concern is focused on the opportunity to choose favourable results from multiple analyses. It is therefore necessary that the statistical procedures planned to deal with, or to avoid, multiplicity are fully detailed in the study protocol or in the statistical analysis plan to allow an assessment of their suitability and appropriateness.

Various different methods have been developed to control the rate of false positive findings.
Not all of these methods, however, are equally successful at providing clinically interpretable results and this aspect of the procedure should always be considered. Since estimation of treatment effects is usually an important issue, the availability of confidence intervals connected with a particular procedure may be a criterion for its selection.

Additional claims on statistically significant and clinically relevant findings based on secondary variables or on subgroups are possible only after the primary objective of the clinical trial has been achieved, and if the respective questions were pre-specified, and were part of an appropriately planned statistical analysis strategy.

〈 임상시험에 있어 다중성에 대하여 고려할 점 〉 자문위원 명단	
강승호 (연세대학교)	김선우 (삼성의료원)
남정모 (연세대학교)	박용규 (가톨릭대학교)

임상시험에 있어 다중성에 대하여 고려할 점

발 행 일 : 2009년 6월

발 행 기 관 : 식품의약품안전청 기획조정관 통상통계담당관

발 행 인 : 왕진호

편 집 위 원 장 : 남봉현

편 집 위 원 : 김은희 장정훈 김현정 김문신 이석배

연 락 처 : 식품의약품안전청 기획조정관 통상통계담당관

전 화 번 호 : 02) 380-1661, 1662

팩 스 번 호 : 02) 356-2893