발 간 등 록 번 호
11-1470000-001888-01

비열등성 허용한계 설정 시 고려사항

Guideline on the choice of the non-inferiority margin

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 세 번째로 비열등성 시험 및 비열등성 허용한계 설정에 대한 유럽 의약품국의 “Guideline on the choice of the non-inferiority margin” 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

비열등성 허용한계 설정 시 고려사항

(Guideline on the choice of the non-inferiority margin)

2005.7.27
Dov.Ref.EMEA/EWP/2158/99

목 차

활성 대조약 (active comparator)과 시험약을 비교하는 대부분의 임상시험은 비열등성 시험 형태로 설계된다. '비열등성 (non-inferiority)'이라는 용어의 의미는 현재 확립된 상태이지만, 이 용어를 문자 그대로 사용한다면 오해의 소지가 높다. 비열등성 시험의 목표는 시험약이 대조약과 비교하여 효과가 떨어지지 않다는 사실을 입증하는 것이라고 때때로 진술된다. 하지만 이것은 오히려 우월성 시험 (superiority trial)으로만 입증될 수 있다. 실제로 비열등성 임상시험은 시험약이 사전에 정해진 어떤 작은 값 이상으로 대조약보다 떨어지지 않다는 것을 입증하는 것을 목표로 하며, 이 때 이 값을 비열등성 허용한계 또는 델타 (Δ)라고 명명한다.

비열등성 임상시험이 위약 대비 우월성 임상시험에 대립되거나 추가적으로 수행될 수 있는 아래와 같은 많은 상황이 존재한다.

- 개량신약이나 국소제제와 같이, 생물학적 동등성 연구가 불가능한 영역에서 주요 유사점에 근거한 응용 시험
- 표준약 대비 잠재 안정성이 뛰어난 약물에 대하여 위험-수익 평가를 실시할 때 표준약과의 유효성 비교가 필요한 경우
- 위험-수익 평가를 위해 활성 대조약과의 직접적인 비교가 필요한 경우
- 활성 대조약과 비교 시 유효성의 어떤 유의한 손실도 받아들여질 수 없는 경우
- 위약군의 사용이 불가능하여, 활성 대조임상시험이 시험약의 유효성을 입증하는데 사용되는 질병의 경우

만약 대조약과의 비교 시 우월성이 입증된다면, 위의 마지막 4가지 상황들에 대해서 비열등성 시험이 필요하지 않을 것이다.

비열등성을 입증하고자 하는 경우, 임상시험계획서에 허용한계에 대하여 사전에 언급할 것을 권고한다. 임상시험 종료 후, 두 약물간의 실제 차이에 대한 95% 양측 신뢰구간 (혹은 97.5% 단측 신뢰구간)이 설정될 것이다. 이 신뢰 구간은 허용한계의 큰 부분 (오른쪽 부분)에 전적으로 포함되어야 한다. 허용한계 (Δ)의 설정은 항상 임상적, 통계적 근거에 의해 정당화되어야 한다. 모든 임상상황에 적용되는 규칙은 존재하지 않으므로 허용한계는 특정 임상적 상황에 맞추어 설정되어야 한다. 그러나 어떤 기준들은 일반적 지침을 제공하는데 사용될 수 있다.

다음의 규제 지침들은 비열등성 혹은 동등성 허용한계를 설정할 때 참고될 수 있다. 본 문서와 함께 이 지침서들을 활용하기 바란다.

- ICH E9지침서 (임상시험에서의 통계적 원칙)
- ICH E10지침서 (대조군의 선택)
- CPMP 비열등성과 우월성 간 전환에 대한 고려사항

위의 지침서에서 비열등성 허용한계 설정 방법은 제한적으로 다루어지고 있다. 그러나 이들은 비열등성을 입증하기 위한 연구의 설계와 수행에 관한 세부적 설명들을 포함하고 있다. 이러한 사항들은 매우 중요하지만, 만약 임상시험이 적절한 기준에 의해 수행되지 않는다면 비열등성 허용한계에 대한 설정은 무의미할 수 있다.

본 문서는 두 가지 유형의 비열등성 임상시험 (시험약과 대조약의 두 군 임상시험, 시험약, 대조약, 위약의 세 군 임상시험)을 다룰 것이다.

임상시험 수행 시 고려해야 할 측면이 많이 있다. 대체적으로 유효성과 안전성이 관련되어야 하지만 이러한 전반적인 범주는 개개의 약물별관심에 따라 세분화될 수 있다. 어떤 특정 변수에 대해서는 비열등성으로, 나머지 변수들에 대해서는 우월성으로 임상시험 혹은 임상 프로그램이 계획될 수 있다. 본 문서에서 '비열등성'과 '우월성'이란 용어는 약물류의 전반적인 측면에서가 아니라 단일 결과변수의 측면과 관련시켜 사용되고 있다. 처리효과에 대한 측정이 가능하며, 그 측정은 바람직한 (양의) 효과와 바람직하지 않은 (음의) 효과로 구별이 가능하다고 가정한다. 또한 측정 변수에서는 양수 값이 클수록 더 큰 양의 효과를 나타내는 것으로 가정한다.

본 문서의 대부분은 위의 개념을 설명하기 위해 처리들 간 절대 값의 차이에 대한 예를 사용한다. 이 개념은 약간의 수정을 가한다면 상대효과를 고려하는 연구에서도 사용이 가능하다. 예를 들어, 상대효과 (relative effect)를 고려하는 임상시험에서 처리군 간 차이가 없다는 것은 점추정치가 0이 아닌 1이 된다.

비열등성 허용한계는 안전성 모수에 대해 사용될 수 있지만, 본 문서에 언급된 방법을 설명하기 위해 유효성 모수들을 사용한다. 하지만 3장에서 대부분 다루어질 유효성 평가변수에 대한 많은 논의는 안전성 임상시험에는 적용되지 않을 것이다.

1. 배경

비열등성 임상시험의 결과는 일반적으로 시험약 (T)과 대조약 (R) 간 실제 차이에 대한 믿을만한 범위를 나타내는 95% 양측 신뢰구간으로 평가된다. 여기서 특별히 고려해야 할 두 가지 측면이 있는데, 하나는 차이에 대한 점추정치 (즉, 시험약과 대조약 간 관찰 값의

차이)이며, 다른 하나는 신뢰구간의 하한이다. 점추정치는 실제 차이에 대한 최적의 추정값으로써 만약 이것이 양의 값을 나타내며, 존재하는 유일한 근거 자료일 경우, 시험약이 대조약보다 효과가 더 크다고 할 수 있으며 또한 그 역도 성립한다. 한편 신뢰구간의 아래쪽 한계는 신뢰구간의 하한을 의미하며 대체적으로 산출된 자료에 근거하여 제외될 수 있는 대조약에 대한 열등성의 정도로 해석된다. 물론 이것이 실제 하한은 아니며, 열등성의 정도는 더욱 커질 수 있다. 그러나 일반적으로 고려하여야 할 것은 실제 차이가 이러한 하한보다 더 커질 가능성은 작아야 한다는 점이다.

만약 시험약과 대조약의 효과가 동일하다면, 차이에 대한 점추정치는 피험자 수와 상관없이 양수가 될 확률이 50%, 음수가 될 확률이 50%가 된다. 그러므로 점추정치를 단독으로 상대적 유효성의 지표로 사용하기에는 충분하지 않다. 동일한 상황에서 신뢰하한은 피험자 수가 증가할수록 점점 0에 가까워질 것이며, 피험자 수를 크게 함으로써 열등성의 정도를 배제하는 것이 이론적으로 가능하다. 하지만 치료효과가 동일하다면 큰 피험자 수를 가진 시험이 요구되어지기 때문에 모든 열등의 정도를 배제하는 연구를 설계하는 것은 불가능하다.

따라서 시험약의 대조약에 대한 어떤 가능한 열등성도 허용될 수 없다면, 비열등성 시험으로 대조약과 동일한 효과를 갖는 약물의 개발이 불가능하다는 것이 시험 초기에 명백해져야 한다.

2. 일반적 고려사항

- 비열등성 허용한계의 설정은 통계적 관점과 임상적 판단에 근거한다.

- 시험약, 대조약, 위약의 세 군 임상시험은 시험 내에서 비열등성 허용한계 설정에 대한 타당성을 확인할 수 있으므로, 추천되는 시험 설계이다.

- 비열등성 허용한계의 적절한 설정은 시험약이 임상적으로 적절한 양의 유효성을 갖는다는 확신을 제공해야 한다. 이는 3장에서 논의된다.

- 비열등성 시험의 주된 관심사항은 단순히 시험약이 효과가 있음을 밝히는 것이 아니라 시험약과 대조약의 상대적 유효성을 평가하는 것이다. 이러한 경우 허용한계의 적절한 설정은 시험약이 효과가 있다는 것뿐만 아니라 시험약이 대조군보다 그다지 열등하지 않다는 것을 확신시켜 줄 것이며 결과적으로 보수적인 허용한계 (tighter margin)가 될 것이다. 이러한 측면에 대한 논의는 4장에서 다루어진다.

- 대부분의 비열등성 임상시험에서 비열등성 허용한계가 3장과 4장의 요구사항인 필요조건을 만족함이 입증되어야 한다. 비열등성 허용한계의 선택은 임상시험계획서에서

정당화되어야 하고, 이러한 정당성은 3장과 4장에서의 고려사항을 언급해야 하는 것이다.

- 비열등성 허용한계를 대조약과 위약 간의 차이에 대한 비율로 정의하는 것은 적절하지 않다. 이러한 생각들은 시험약이 (잠정적인) 위약보다 우월하다고 보이기 위해 설정되었지만, 오히려 이 우월성을 보이지 못할지도 모른다. 대조약의 효과가 위약의 효과보다 크다면, 큰 차이가 중요하지 않다는 것을 의미하는 것은 아니며 단지 대조약의 효과가 매우 뛰어나다는 것을 의미한다.

- 효과 크기 (effect size: 표준편차에 의해 나누어지는 처리의 차이)를 비열등성 허용한계에 대한 정당한 근거 자료로 사용하는 것은 적절하지 않다. 이 통계량은 차이를 인지하는 것이 얼마나 어려운 것인가에 대한 정보를 제공하지만, 차이의 임상적 적절성이 정당화될 수 없으며, 시험약이 위약보다 우월하다는 것을 확신할 수 없다.

- 비열등성 허용한계의 결정은 검정력과 무관해야 한다. 임상적으로 중요한 차이의 크기가 피험자 수에 의해 변경되어서는 안되는 것과 같이, 허용한계 설정은 피험자 수에 관한 문제에 근거하는 것이 아니라, 본 문서의 다음 장에서 다룰 임상적, 통계적 원리에 근거해야 한다. 피험자 수가 적은 임상시험은 보다 넓은 범위의 비열등성 허용한계를 정당화시킬 수 없다.

- 만약 적절한 허용한계를 선택했다면, 신뢰구간 전체가 $-\Delta$와 0사이에 포함된다는 것은 (즉, 시험약이 대조약보다 효과가 낮은 것으로 판단되나 비열등성 허용한계인 Δ보다 나쁘지는 않다) 비열등성을 입증하는데 적절하다. 만약 이 결과를 받아들일 수 없다면, 허용한계가 적절하지 않기 때문이다 (열등성 정도의 정당성을 입증하기 어려운 상황에 대한 논의를 위해 5장을 참고하라).

- 모든 상황에서 비열등성 시험을 수행할 수는 없다. 이 시험을 수행할지에 대한 결정은 치료 분야와 대조약의 특성을 고려하여 정당화되어야 한다.

- 효과가 입증된 약물이 위약 통제 시험 (예를 들어, 우울증 혹은 알레르기성 비염)에서 일관성 있는 우월성을 입증하지 못하는 경우가 많다. 민감도가 부족하다고 판단되는 시험의 경우, 위약을 포함하지 않은 비열등성 시험은 적절하지 않다. 민감도 평가에 대한 보다 자세한 설명을 위해 ICH E10을 참고하기 바란다.

- 대조약의 효과가, 비열등성 허용한계를 정의할 때 예상했던 효과와 매우 다르다면 설정된 허용한계는 더 이상 적절하지 않을 수도 있다. 이러한 내용은 시험 계획 단계에서 고려되어야 한다.

3. 유효성의 입증

비열등성 시험으로부터 나온 자료해석과 관련된 의사결정의 최소 요구사항은 위약 통제 시험이 수행되었다면 시험약이 효과가 있었음을 확신할 수 있어야 한다는 것이다. 이 장에서는 우월성 시험에 대한 자료를 비열등성 자료에 대한 최소 요구 사항을 평가하는 모델로 해석할 때 사용되는 방법에 대하여 논의한다.

위약 대비 시험약의 우월성을 보이기 위해 설계된 임상시험의 결과자료를 해석하는 경우에는 통계적 유의성과 임상적 적절성을 고려한 비공식적 2단계 절차를 사용한다. 이 절차는 비열등성 시험을 해석하는 데에도 사용될 수 있다. 우월성 시험에 있어 시험약이 위약보다 통계적으로 유의하다는 결론을 입증하는 것이 우선적으로 요구된다. 이것은 '통계적 사고와 임상적 판단' 모두가 결합되어 있는 ICH E10의 '통계적 사고' 단계와 관련이 있다. 일반적으로 통계적 유의성은 양측 유의수준 0.05 (혹은 단측 유의수준 0.025)를 사용하여 평가된다. 대안적인 방법은 시험약과 위약의 차이에 대한 양측 95% 신뢰구간 (혹은 단측 97.5% 신뢰구간)의 하한이 0보다 커야 한다는 것이다.

다음 단계는 위약과의 차이가 임상적으로 적절한가를 고려하는 것이다. 이것은 '통계적 사고와 임상적 판단' 모두가 결합되어 있는 ICH E10의 '임상적 판단' 단계와 관련이 있다.

시험약과 위약간의 차이에 대한 점추정치를 고려함으로써, 또는 원래의 척도를 사용하거나 혹은 반응률을 고려하여 임상적 적절성을 평가함으로써 위약에 대하여 임상적으로 적절한 유용성이 확립된다. 이것이 통계적으로 중요한 이슈는 아니지만 임상적 지식과 자료의 이해가 통합된 지적 사고를 필요로 한다. 통계적 유의성이 이미 입증되었다면, 실제적인 효과에 대한 입증 작업이 고려된다. 이와 같은 차이가 임상적으로 유용한지에 대한 판단이 반드시 필요하다. 이러한 판단은 일반적으로 위험-수익 평가를 통한 안전성 측면에서 이루어진다.

3.1 세 군 임상시험 : 시험약, 대조약 및 위약

세 군 임상시험 설계는 시험약과 대조약이 위약보다 더 좋은 유효성을 제공하는 지를 직접적으로 입증하게 한다. 이 경우 시험약의 유효성을 입증하기 위해 허용한계를 정의할 필요는 없지만, 임상 자료 해석시 다음의 사항들이 언급되어야 한다.

위약 통제 우월성 시험과 마찬가지로 위약보다 시험약의 효과가 통계적으로 유의한가를 입증해야 한다. 시험약과 위약간의 차이에 대한 95% 신뢰구간의 하한은 0보다 커야 한다.

세 군 임상시험에서 비록 시험약과 대조약이 모두 위약 대비 통계적 유의성을 입증하지

못한 경우, 이것이 민감도가 결여되었다는 것을 의미하는 것일지라도 이 시점에서 대조약의 효능은 중요한 관심이 아니다.

우월성 임상시험과 마찬가지로 임상적인 판단은 위약과의 관찰된 차이가 임상적으로 유의한가를 평가하기 위해 사용된다. 대조약을 활용하는 것이 이러한 임상적 판단에 도움이 된다. 만약 이러한 종류의 시험에서 사용되는 대조약이 임상적으로 적절한 효과를 가지는 허가된 시판 약물이라면, 이 시험에서 보여진 위약과 대조약의 차이는 위약과 시험약 간 차이에 대한 임상적 적절성을 평가하는데 활용될 수 있다. 예를 들어 임상시험에서 시험약이 대조약보다 더 큰 효과를 보인다면, 시험약의 유용성이 임상적으로 적절하다고 가정할 수 있을 것이다.

만약 대조약이 위약과 비교하여 통계적 유의성을 갖지 못하거나, 기대했던 것과 매우 다른 결과가 나타난다면 대조약의 효과에 대하여 의문이 제기될 것이다. 이 경우 대조약을 사용한 시험 결과들을 신뢰할 수 없을 것이고, 시험약에 대한 어떠한 긍정적 결과도 의미가 없을 것이다. 대조약의 예상치 못한 결과에 대해서는 적절한 이유가 논의되어야 한다.

3.2 두 군 임상시험 : 시험약, 대조약

두 군 임상시험에는 위약군이 없기 때문에, 약물의 유효성을 결정하는데 과거의 위약 대비 대조약의 비교연구가 약물의 위약에 대한 간접비교로서 약물의 유효성을 결정하는데 사용되어야 한다. 이는 본래 어려운 작업으로 비열등성 허용한계의 정의를 수반한다. 하지만 95% 신뢰구간의 하한을 활용하여 위약보다 유효성이 있는지 확인할 수 있다. 잠정적인 위약이라는 용어는 위약을 사용하지 않는 상황에서 종종 사용되어진다.

이런 상황에서 고려되는 위약 대비 대조약의 연구가 적절한 연구인지를 확인하기 위해서는 체계적 검토가 실행되어야 한다. 이런 검토는 계획된 피험자 모집단에서 대조군과 위약군간 처리효과 차이를 추정하기 위해 사용될 수 있다. 만약 그러한 추정이 불가능하거나, 혹은 적절한 검정력이 확보된 시험에서 대조약에 대한 위약의 우월성이 일관되지 못하다면, 대조약을 이용한 비열등성 시험의 민감도에 관한 의문이 야기될 수 있으므로 비교약물(대조약 혹은 위약)에 대해 시험약이 우월성을 갖는다고만 해석될 가능성이 있다. 만약 대조약이 유효성과 안전성을 모두 동일하게 갖추고 있다고 판단되는 약물군의 일부라면, 위약과 전반적인 약물군과의 차이를 활용하는 것이 허용될 수 있다.

임상 연구자들이 논의해야 될 필요가 있는 문헌검색과 관련된 이슈들은 다음과 같다 (더 많은 논의 사항에 대해서는 ICH E10을 참고하라).

1. 선택 편의: 어떤 연구들을 선택할지에 대한 기준을 철저히 기술하여, 연구를 선택할 때 가능한 편향성이 없도록 해야 한다.

2. 시간에 따른 임상시험 설계와 수행의 일관성: 임상시험 수행이 변경되었거나 대조약의 효과를 측정하는 기준이나 측정방법이 수정될 수 있기 때문에 어떤 연구들은 적절하지 않을 수도 있다. 현재의 임상시험을 설계할 때에는 이전 임상 시험과의 비교를 통해 처리결과에 영향을 줄 수 있는 변경 사항을 고려해야 한다. 예를 들어, 피험자 선정기준, 진단방법, 허용된 병용처리, 대조약의 용량, 측정된 평가변수, 그리고 평가시점 등이 고려되어야 한다. 가능하다면 현재 임상시험 설계가 위약과 대조약을 비교한 이전 임상시험과 밀접한 유사성이 있어야 한다. 임상시험 설계시, 불가피한 차이가 존재한다면 차이점을 주의 깊게 고려해야 하지만, 비열등성 허용한계를 명확하게 설정하는 것이 불가능할 수도 있다.

3. 시간에 따른 효과의 일관성: 시간에 따른 처리 차이의 변화를 고려해야 하여야 한다. 예를 들어 어떤 경우에는 시간이 지남에 따라 의료 서비스의 일반적인 개선으로 인해 발병률이 감소될 수도 있다. 이러한 경우 허용한계 설정 시에 최근 자료를 더 많이 포함하는 것이 적절하다. 만약 최근 시행된 임상시험부터 현재 임상시험까지의 효과의 일관성을 확신할 수 없다면, 비열등성 허용한계의 선택 시 보수적인 방법을 사용해야 한다.

4. 출판 편의: 유의한 결과를 가진 연구가 그렇지 아니한 연구보다 출판될 가능성이 더욱 크다. 만약 그러한 출판 편의가 존재한다고 판단될 경우, 대조약과 위약의 과거 신뢰구간을 산출하는데 보수적인 접근법이 사용되어야 한다.

만약 유용한 과거 자료의 활용이 가능하다면, 비열등성 허용한계를 산출하는데 사용될 수 있는 몇 가지 방법들이 있다. 이 방법들의 공통점은 과거 자료에 나타난 효과의 변동 및 크기를 현재 임상시험에서 기대되는 효과의 변동 및 크기와 결합하는 것이다. 또한 과거 자료를 사용하는 데 있어 모든 방법들은 공통적으로 본질적인 약점이 내재되어 있다.

과거 신뢰구간은 대조약과 위약을 비교한다 (R-P). 시험약과 대조약을 비교 하는 계획된 임상시험은 시험약과 대조약의 차이에 대한 신뢰구간을 산출한다 (T-R). 이 두 신뢰구간들을 결합할 경우, 시험약과 위약을 비교하기 위한 간접적인 신뢰구간이 산출될 수 있다 (T-P). 비열등성 허용한계는 T-P의 간접적 신뢰구간의 하한이 0보다 큼을 입증할 수 있는 T-R 신뢰구간의 하한으로 정의될 수 있다. 이때의 비교는 간접적인 방법을 사용하기 때문에, 이로써 제시되는 값보다 작은 값을 선택하는 보수적인 방법이 적절할 수도 있다.

이 경우 임상시험 결과문서 제출 시, 직접적 신뢰구간인 T-R과 간접적 신뢰구간인 T-P를 모두 제시해야 한다.

일단 통계적 유효성이 유의하다고 입증되면, 임상적 적절성을 고려해야 한다. 즉 T-P에 대한 간접적 신뢰구간으로부터의 점추정치에 주목하고, 그 차이가 임상적으로 적절한가를 고려해야 한다.

세 군 임상시험에서처럼 대조군은 몇 가지 결과를 제공한다. 만약 시험약이 대조약보다 뛰어난 효과를 보인다면, 효과 차이가 임상적으로 적절하다고 어느 정도는 확신을 가질 수 있다. 하지만, 분석시점에서 자료가 대조군의 효과에 대하여 의혹을 이끌어 내는 경우, 선택된 비열등성 허용한계는 부적절할 수 있으며 시험의 타당성이 의문시 될 수도 있다. 위약군이 없는 시험에서는 대조약의 효과에 대한 타당성을 판단하기가 더욱 어려우므로 과거 자료를 이용하여 현재의 임상시험 효과가 기대치와 일관성을 유지하고 있다는 것을 입증할 필요가 있다.

관찰된 차이가 임상적으로 유용한지를 입증하기 위해선 동일한 치료 영역에서 임상적으로 적절한 결과를 보이는 다른 임상시험 및 임상적 적절성을 보이지 않지만 마치 중요해 보이는 임상시험을 참조해야 한다.

3.3 위약에 대한 우월성이 신뢰성 있게 입증되지 않는 상황들

어떤 질병 영역에서는 위약 통제 임상시험이 비윤리적인 시험으로 간주되고, 아직까지 어떤 대조약도 위약보다 뛰어난 효과를 신뢰있게 입증하지 못하고 있다. 예를 들어, 종양학 조절 (oncology settings)이나 희귀한 징후 (orphan indications)가 나타나는 상황에서는, 위의 일반적인 사항을 고려하여 비열등성 허용한계를 설정하는 것은 어려운 작업이다. 하지만 대조약에 관한 활용 가능한 모든 자료를 동원하여, 위약과 대비되는 새로운 약물에 대해 간접적인 신뢰구간을 산출하도록 노력해야 한다. 만약 대조약에 관한 자료를 구할 수 없다면 대조약을 위약처럼 처리할 수 있고, 대조약 대비 우월성을 직접적으로 증명함으로써 그 유효성을 입증할 수 있다.

이러한 경우에 설정된 간접적인 신뢰구간은 위약 대비 우월성을 입증하지 못할 수도 있지만, 그렇다고 해서 이에 대한 긍정적인 결론 도출을 막을 필요는 없다. 3.2장에서 명시된 방법을 통하여 산출된 비열등성 허용한계는 위약보다 간접적으로 효과가 좋다는 것을 의미하는 값이므로 제외되어서는 안된다. 이러한 비열등성 허용한계를 사용한다는 것은 어떠한 결정 (긍정적 혹은 부정적)을 내릴 때에도 그 유효성을 입증하는데 자료에 한계가 있다는 것을 인정한다는 것과 같은 의미이다. 즉, 시험군이 위약군보다 뛰어나다고 확신할 수 없다. 임의로 비열등성 허용한계를 정하고, 그것을 사용하여 비열등성을 주장하는 것은 좋은 방법이 아니다. 이러한 접근은 약물의 유효성 연구에 대한 신뢰도를 낮추게 될 것이다. 따라서 비열등성 허용한계의 결정에 필요한 근거가 약하든 강하든 이를 명확히 하는 것이 중요하다.

3장의 지침을 따르게 되는 경우, 현실적인 규모의 피험자 수를 대상으로 한 임상시험에서 제외시킬 수 없는 작은 비열등성 허용한계를 산출하는 상황도 있을 수 있다. 다시 언급하지만 비열등성의 허용한계를 단순히 넓게 정의한 후 비열등성이라고 주장하는 것도 역시 좋은 방법은 아니다. 오히려 이러한 방법으로 위약 대비 우월성을 밝히는 것이 불가능하다는 것을 인정해야 한다. 어떤 분야에서는 유효성에 대한 명백한 입증 없이 단지

불충분한 양의 자료에 근거하여 의사결정이 이루어지기도 한다. 하지만 이것이 수행되고 있다는 것과, 수행 과정에 고려되어야 할 위험 사항을 충분히 인지한 상황에서 의사 결정이 이루어지고 있다는 것을 모두가 인식하는 것이 중요하다.

4. 대조약과 관련된 유효성 입증

새로운 약물이 위약 통제 임상시험에서 뛰어난 효과를 보이는 것을 입증하는 것이 필요하지만, 보통 그것만으로는 충분하지는 않다. 시험약과 대조약 간 비교는 종종 비교자체가 중요할 수 있다.

이러한 관점에서 볼 때, 임상시험 시작 전 연구의 목적을 정의하는 것은 중요하다. 비열등성 시험의 설계와 분석시, 임상시험을 통해 밝히고자 하는 가설을 고려해야 한다. 그러한 임상시험이 수행되는 많은 이유들이 있고, 이러한 임상시험의 수행 목적이 비열등성 허용한계의 선택에 영향을 주어야 한다. 임상시험의 목적이 대조약 대비 우월성을 보이는 것이 아니라면 임상시험은 대개 비열등성 임상시험이라고 간주된다. 그러나 비열등성 입증은 임상시험 목적을 자세하게 나타낸다고 할 수 없다. 보다 명확한 목적이 기술될수록 보다 많은 구체적인 결과들이 도출될 수 있다.
임상시험의 유일한 목적이, 새로운 약물이 위약보다 간접적으로 우월함을 증명하는 것이라면, 그 목적을 기술하고, 3장의 방법들만을 사용하여 비열등성 허용한계를 설정할 수 있다.

대안적으로 임상시험의 목적은, 만약 시험약이 대조약 대신으로 사용되었을 경우, 유효성의 주요 손실이 없음을 보이는 증거 자료를 제공하는 것으로 설정될 수 있다. 이는 비열등성 시험에 대한 가장 일반적인 목표로 간주된다. 이러한 목표 하에서는 과거에 실시된 위약 대비 대조약에 대한 임상시험들의 결과만을 이용하여 허용한계를 설정할 수는 없다. 그 동안은 비열등성 허용한계를 대조약과 위약 간의 기대 차이에 대한 백분율로 선택해 왔지만, 이것은 허용한계 설정을 위해 인정할 만한 것으로 고려되지 않는다. 이것은 이미 3장에서 언급하였듯이 원칙적으로 대조약이 위약보다 우월하다는 확신을 보장할 수 있도록 만들어져야 하기 때문이다. 비열등성 허용한계 (Δ)를 적절하게 선택하기 위해서는, 정보에 근거한 결정을 내려야 하며 특정 질병 영역에서 중요하지 않은 차이로 간주되는 사항들을 입증 자료로써 활용해야 한다.

만약 연구 중인 질병에 대해서도 적용 가능한 다양한 치료법이 이미 존재한다면, 그러한 모든 치료법에서 얻을 수 있는 정보의 활용도 가능하다. 이러한 약물들의 상대적 유효성에 대해 알려진 정보를 요약함으로써 비열등성 허용한계를 산출할 수도 있고, 새로운 약물의 상대적 유효성과 관련하여 비슷한 수준의 정보를 제공하기 위해 새로운 임상시험이 설계될 수 있다. 만약, 현재 하나의 약물만이 시장에 출시되어 있다면, 이러한 접근법은 가능하지 않을 것이다. 이 약물의 사용자가 누구인지를 고려하면서, 임상시험 수행자들이 중요

하지 않다고 생각하는 차이의 범위에 대해 조사하고, 각 조사 결과의 요약 통계량에 근거하여 비열등성 허용한계를 설정할 수 있다. 그러한 모든 조사는 응답자들이 의도한 큰 값 방향으로 편향성이 없음을 밝혀야 한다.

시험약이 대조약에 비해 안전성이 더 좋다고 기대하는 상황에서는, 비록 약물의 유효성이 증명될 것이 기대되어지더라도, 안전성의 확보를 위해 유효성 손실이 어느 정도 허용할 수 있기 때문에 보다 큰 허용한계가 정당화될 수 있다. 이러한 경우 하나는 안전성 평가변수의 우월성을, 다른 하나는 유효성 평가변수의 비열등성을 입증하는 다중 일차 유효성 평가변수를 이용할 수 있다. 고려되는 사항들을 확실하게 이행할 수 있는 기타의 상황으로는 보다 편리한 투약 경로 및 복용량, 이차 유효성 평가변수의 우월성 등을 포함한다.

본 장은 비열등성 시험의 목적과 비열등성 허용한계의 설정이 가능한 몇몇 방법들을 다루고 있다. 본 장의 주안점은 임상시험의 목적이 명확하게 정의되어야 한다는 것이다. 그 후 비열등성 허용한계의 선택은 증거자료에 의해 타당성이 입증되고, 명확한 목적에 근거하여야 한다. 그 근거는 과거에 실시된 위약과 비교하는 임상시험으로부터만 제시되지는 않는다. 물론 최종선택은 적어도 3장으로부터 산출된 값만큼 항상 작아야 한다. 임상시험의 결론은 '비열등성'이 입증되는 것이 아니라 임상시험의 목적을 반영한 진술을 보다 명확히 하는 것이어야 한다.

임상시험에서 대조약의 효과가 미리 예상했던 것과 매우 다르다면 시험약과 대조약 간의 차이에 대한 의미를 해석하는 것이 어려울 수도 있고, 미리 정해진 비열등성 허용한계는 더 이상 적절하지 않을 수도 있다. 이런 상황에서는 임상시험으로부터 좋은 결론을 도출하는 것이 불가능할 수도 있다. 위약을 포함한 세 군 임상시험은 위약과 비교한 대조약의 상대 유효성을 직접적으로 추정할 수 있기 때문에 이러한 문제를 해결할 수 있다.

5. 비열등성 허용한계값의 크기를 정당화하기 어려운 영역

연구 중인 치료법이 사망 혹은 치명적인 질병을 예방하기 위해 사용되며, 다른 대안 치료방법이 존재하지 않는 경우에는, 어떠한 크기의 비열등성 허용한계값이라도 정당화하는 것은 매우 어렵다. 이와 마찬가지로, 허용 가능한 추가적 사망자 수에 대해 논의를 한다는 것은 윤리적 관점에서 매우 어렵다. 그러나 허용한계에 대한 모든 선택을 허용하지 않는 것이 공중보건에서의 최대 관심사는 아니다. 처리효과의 차이가 통계적 유의성을 갖지 못한다면, 그 차이에 대한 신뢰구간은 시험약이 대조약보다 열등한 효과를 가질 수도 있고, 우월한 효과를 가질 수도 있다는 것을 의미한다. 따라서 어떤 수준의 열등성도 허용할 자세가 되어 있지 않다고 생각하더라도 현재 허가된 약물을 계속 사용함으로써 어느 수준의 열등성을 허용하고 있는 것이다. 그러므로 이러한 영역에서 비열등성 임상시험이 여전히 사용될 가능성이 있다는 것은 중요하다. 본 장에서는 이에 관한 이해를 돕는 몇 가지

접근법에 대하여 논의할 것이다.

5.1 증가된 유의수준을 사용한 우월성

본 문서에서 이미 언급한 바와 같이, 비열등성 허용한계를 인정하지 않으면(Δ=0, 특히 우월성 임상시험) 우연을 제외하고, 동등한 유효성을 갖는 약물에 대한 시험에서는 긍정적인 임상 시험 자료를 획득할 수 없게 된다. 작지만 임상적으로 유용성을 제공하는 약물은 피험자수가 많은 임상시험에서만 약물이 주는 혜택을 일관성 있게 나타낼 것이다.

이러한 특별한 상황에서 일차 유효성 평가변수에 관해서만 초점을 맞추었을 때, 대조약보다 높은 효과를 보이도록 설계된 우월성 임상시험보다 비열등성 임상시험이 더 많이 사용되는데, 중요한 이유는 다음의 세 가지와 같다.

① 비열등성 시험만이 사용 가능하며, 약물의 유효성이 실제로 동일하다.
② 시험약이 제공하는 효능이 작아 이를 입증하기 위해서는 비현실적으로 많은 피험자가 요구된다.
③ 약물에 단점이 있지만, 단점이 비열등성 허용한계보다 덜 중요하다.

②에 속하는 약물이 시험 성공을 더 많이 거두는 것은 공중보건 관점에서 매우 중요한 이슈이다. 그러나 ③의 경우와 같이 극단적인 경우에는 성공하지 않는 것이 더 낫다. 비열등성 시험은 어떠한 요구사항이라도 이러한 균형을 맞춰야한다. ①에 속하는 약물의 성공과 실패는 공중보건의 관점에서는 덜 중요하다.

이와 같이 중요한 영역에서는 0을 중심으로 잘못된 방향에 위치한 점추정치가 채택될 가능성은 거의 없다고 결론 내리게 한다. 그러한 자료는 ②번보다 ③번에서 가능성이 더욱 높다.

이 경우, 유의성 검정의 개념으로 비열등성 허용한계를 설정하는 것이 도움이 된다. 만약 95% 신뢰구간 전체가 0보다 큰 쪽에 위치한다면, 유의수준 5%에서 우월하다는 것이 입증된다. 95% 신뢰구간의 하한에 대한 특정한 값에 대응해서 동일한 자료에서 하한이 정확히 0이 되는 다른 확률 (coverage probability)을 갖는 또 다른 신뢰구간이 존재한다. 예를 들어 (95% 신뢰구간 보다 좁게 정의되는) 85%의 신뢰구간의 하한이 0을 포함하는 자료에 대해서, 95% 신뢰구간에서는 -5를 포함할 수도 있다. 만약 비열등성 허용한계가 -5로 정의되어 있다면, 위의 예에서 비열등성 시험을 입증하는 것은 유의수준 15%에서 우월성을 입증하는 것과 일치한다. 그러므로 덜 보수적인 신뢰수준에서 우월성 시험을 수행하거나 동시에 비열등성 시험을 수행할 수 있다.
극단적인 상황에서는 유효성을 보이는 약물을 기각하는 위험 (risk)과, 위양성 (false positive) 결과의 증가된 위험도를 비교 측정함으로써 0.05보다 덜 엄격한 유의수준을 사용하여 우월성 시험을 적용하는 것도 허용되는 접근법 중의 하나이다. 윤리적인 관점

에서는 임상적으로 무의미한 추가적인 사망자 수를 언급하는 것보다 약물의 우월성 입증에 필요한 신뢰수준을 구체적으로 설정하는 것이 보다 용인되며 또한 손쉬울 수 있다.

특정 비열등성 허용한계의 경우, 증가된 유의수준에서는 허용한계가 잠재되어 뚜렷이 나타나지 않는다는 사실을 인식해야 하지만 이 경우에 있어 발생 가능한 모든 열등성이 제외된다는 것을 의미하는 것은 아니라는 사실에도 주목해야 한다. 그러나 95% 신뢰수준과 5% 유의수준이 일반적으로 사용된다고 하더라도 그 유의수준 하에서도 위양성 결과가 존재한다는 것을 기억할 필요가 있다. 본 장에서는 위음성 (false negative)의 확률을 낮추기 위해서만 위양성 우월성 주장 확률에 대한 단순 증가 작업을 실시하고 있으며, 이 둘 사이에 존재하는 이상적인 균형점은 극단적인 상황에서 변동될 수 있음을 밝히는 바이다.

이러한 접근이 사용되는 경우에는 임상시험의 목적과 가설을 사전에 명확히 기술해야 한다. 예를 들어, 임상시험의 가설이 시험약이 효과가 있다는 것이라면, 이 가설을 명확히 기술해야 한다. 사전에 적절하게 정의된 가설 하에 시행된 임상시험 결과는 의외의 결과를 나타내는 임상시험의 결과보다 더욱 설득력이 있다.

일반적인 유의수준에서 유의성을 입증하지 못한 연구 결과를 얻자마자 이러한 접근 방법으로 다시 돌아가는 것은 허용되지 않는다. 유의수준을 높이고자 할 때에는 그 계획과 타당성이 임상시험계획서에 명확히 언급되어야 한다.

만약 이러한 접근법을 사용하게 된다면, 연구 중인 특정 과제에 대하여, CHMP가 그러한 접근법이 적절한가를 고려하고 있는지를, 임상시험계획서의 전문적인 검토 혹은 과학적 자문 수행을 통해 확인할 것이 권장된다.

5.2 다른 측면에서 이점을 갖는 약물

약물이 다른 측면에서 중요한 이점을 가지는 경우, 유효성에 대한 비열등성 허용한계를 정의하는 것이 가능할 수도 있다. 그렇다면, 0을 중심으로 잘못된 방향에 위치한 처리효과 간 차이의 점추정치가 인정될 수 있다. 이러한 경우에는 두 개의 일차 유효성 평가변수를 고려하여 유효성의 비열등성을 입증하거나 다른 중요 요인에서 우월성을 입증하도록 시험 설계할 것을 권장한다.

6. 결론

- 비열등성 허용한계를 설정할 때는 통계적 사고와 임상적 판단을 종합적으로 고려한다.

- 시험약, 대조약 및 위약을 모두 다루는 세 군 임상시험은 비열등성 허용한계의 선택에 대하여 임상시험 내 타당성 확인이 가능하므로 임상 시험 수행이 보다 용이하다. 이 방법은 권장되는 임상 시험 설계로써 어느 경우에나 사용이 가능하다.

- 비열등성 허용한계를 적절하게 선택함으로써 시험약이 위약보다 임상적으로 우월하다는 것을 입증할 수 있다. 이러한 비열등성 허용한계에 대한 측면들은 3장에서 논의되었다.

- 비열등성 시험의 주요관점은 단순히 시험약의 유효성을 입증하는 것이 아니라, 시험약과 대조약의 상대적 효과를 증명하는 것이다. 이러한 경우 허용한계의 적절한 선택은 약물이 효과가 있다는 것을 증명하는 것 외에도 시험약이 대조약보다 실질적으로 열등하지 않음을 입증할 수 있다. 이러한 비열등성 허용한계에 대한 측면은 4장에서 논의되었다.

- 대부분의 비열등성 시험에서 비열등성 허용한계가 3장과 4장에 명시된 요구 조건을 동시에 만족시키고 있다는 것을 입증해야 한다. 임상시험계획서에 비열등성 허용한계 선택에 대한 타당한 근거가 제시되어야 하며, 근거 자료에는 3장과 4장에서 다루고 있는 사항들이 만족되어야 한다.

- 모든 상황에서 비열등성 시험을 수행하는 것은 불가능하다. 따라서 비열등성 시험의 수행 여부는 치료 영역과 대조약의 측면 모두를 고려하여 결정되어야 한다.

- 시험약이 몇 가지 측면에서 이점을 가지고 있다면, 유효성에 대한 보다 넓은 비열등성 허용한계를 정의하는 것이 가능하다. 그러나 이러한 허용한계가 위약대비 우월성을 확증하지 못할 정도로 크지는 않아야 한다.

- 극단적인 경우에는 비열등성 허용한계를 정의하는 것 대신 0.05보다 큰 유의수준을 사용하는 우월성 시험을 수행하는 것이 인정될 수 있다.

COMMITTEE FOR MEDICINAL PRODUCTS FOR HUMAN USE
(CHMP)

GUIDELINE ON THE CHOICE OF THE NON-INFERIORITY MARGIN

London, 19 September 2002

CPMP/EWP/908/99

DRAFT AGREED BY THE EFFICACY WORKING PARTY	December 1999 - January 2004
ADOPTION BY COMMITTEE FOR RELEASE FOR CONSULTATION	February 2004
END OF CONSULTATION (DEADLINE FOR COMMENTS)	May 2004
AGREED BY WORKING PARTY	June 2004
ADOPTION BY COMMITTEE	July 2005
DATE FOR COMING INTO EFFECT	January 2006
ADOPTION BY CPMP	September 2002

TABLE OF CONTENTS

INTRODUCTION

Many clinical trials comparing a test product with an active comparator are designed as non-inferiority trials. The term 'non-inferiority' is now well established, but if taken literally could be misleading.
The objective of a non-inferiority trial is sometimes stated as being to demonstrate that the test product is not inferior to the comparator. However, only a superiority trial can demonstrate this. In fact a non-inferiority trial aims to demonstrate that the test product is not worse than the comparator by more than a pre-specified, small amount. This amount is known as the non-inferiority margin, or delta (Δ).

There are many situations where a non-inferiority trial might be performed as opposed to, or in addition to, a superiority trial over placebo. These include:

☐ Applications based upon essential similarity in areas where bioequivalence studies are not possible, e.g. modified release products or topical preparations;

☐ Products with a potential safety advantage over the standard might require an efficacy comparison to the standard to allow a risk-benefit assessment to be made;

☐ Cases where a direct comparison against the active comparator is needed to help assess risk-benefit;

☐ Cases where no important loss of efficacy compared to the active comparator would be acceptable;

☐ Disease areas where the use of a placebo arm is not possible and an active control trial is used to demonstrate the efficacy of the test product.

In the final 4 situations above a non-inferiority trial would not be necessary if superiority could be shown over the reference product.

In order to demonstrate non-inferiority, the recommended approach is to pre-specify a margin of non-inferiority in the protocol. After study completion, a two-sided 95% confidence interval (or one-sided 97.5% interval) for the true difference between the two agents will be constructed. This interval should lie entirely on the positive side of the non-inferiority margin. The choice of delta must always be justified on both clinical and statistical grounds. It always needs to be

tailored specifically to the particular clinical context and no rule can be provided that covers all clinical situations. However, certain principles can be used to provide general guidance.

The following regulatory guidelines make reference to the choice of the margin of non-inferiority or equivalence. They should be read in conjunction with this Guideline.

□ ICH Note for Guidance E9 (Statistical Principles for Clinical Trials);

□ ICH Note for Guidance E10 (Choice of Control Group);

□ CPMP Points to Consider on Switching Between Superiority and Non-inferiority.

In these documents the discussion of how to choose the margin of non-inferiority is limited. They do, however, make detailed comments regarding the design and conduct of studies designed to demonstrate non-inferiority. Such issues are extremely important, and if a trial has not been conducted to an appropriately high standard, the choice of delta can become an irrelevant issue.

This document will consider two types of non-inferiority trials: trials with two arms, the test product and a comparator; and three-armed trials with the test product, an active comparator and placebo.

There are many aspects of the performance of an experimental product to consider. Broadly these relate to efficacy and safety, but each of these broad categories can be broken down for an individual product into many points of interest. A clinical trial or clinical programme may plan to show non-inferiority for certain variables while superiority may be the objective for others. In this document the terms ‘non-inferiority’ and ‘superiority’ are used relating to single endpoints and not to the product profile as a whole.

It is assumed throughout that the effect of the treatments can be measured and that the measurements make it possible to distinguish between desired (positive) and undesired (negative) effects. It is further assumed that large positive values in the measured variable point to large positive effects.

The majority of the document uses the example of the absolute difference between treatments to illustrate the ideas. The discussion is also applicable to studies

considering a relative effect with a few modifications. For example in a trial considering relative effects, no difference between treatments is reflected by a point estimate of one, as opposed to a difference of zero.

Efficacy parameters are used to illustrate the methods mentioned in this document, although non-inferiority margins can be defined for safety parameters as well. However, a lot of the discussion for efficacy end-points will not apply to safety trials, most notably the whole of section III.

1. BACKGROUND

The outcome of a non-inferiority trial is usually assessed by a two-sided 95% confidence interval, showing a credible range for the true difference between the test product (test: T) and the active comparator (reference: R). There are two aspects of the results that should attract particular attention.
One is the point estimate of the difference, i.e. the observed difference between test and reference. The other is the lower limit of the confidence interval. The point estimate represents the best estimate of the true difference, so that if it is positive and if this is all the evidence available, it is more likely that the test product is better than the reference and vice-versa. The lower limit of the confidence interval, on the other hand, represents a lower bound and is usually interpreted as the degree of inferiority to the reference that can be excluded based on the data presented. Of course this is not an actual lower bound and the magnitude of inferiority could be greater. However it is generally considered that the chance of the true difference being worse than that suggested by this bound is acceptably small.

If T and R were equally efficacious, then the point estimate for the difference would have a 50% chance of being positive and a 50% chance of being negative, regardless of sample size. Hence the point estimate alone is not sufficient as an indicator of relative efficacy. The lower confidence limit for the difference, in the situation of true equality, would be expected to move closer to zero as the sample size increased, thus making it theoretically possible to rule out any desired degree of possible inferiority by using sufficiently large samples. However, if the treatments truly are equally efficacious, it is not possible to design a study to rule out all degrees of inferiority as this would require an infinitely large experiment.

Thus, it should be made clear at the outset that if no degree of possible inferiority

of T to R is acceptable, then the development of products with equal efficacy to a comparator by means of non- inferiority trials would become impossible.

2. GENERAL CONSIDERATIONS

□ The selection of the non-inferiority margin is based upon a combination of statistical reasoning and clinical judgement.

□ A three-armed trial with test, reference and placebo allows some within-trial validation of the choice of non-inferiority margin and is therefore the recommended design; it should be used wherever possible.

□ An appropriate choice of margin should provide assurance that the test drug has a clinically relevant effect greater than zero. This aspect of the choice of margin is discussed in section III.

□ Usually the primary focus of a non-inferiority trial is the relative efficacy of the test and reference products, not simply demonstration that the test product has an effect. In these cases an appropriate choice of margin will, in addition to proving that the product has an effect, also provide assurance that the test product is not substantially inferior to the reference, resulting in a tighter margin. This aspect of the choice of margin is discussed in section IV.

□ For the majority of non-inferiority trials it must be demonstrated that the margin satisfies both the requirements of section III and section IV. The choice of non-inferiority margin should be justified in the study protocol, and the justification should address the considerations of both sections.

□ It is not appropriate to define the non-inferiority margin as a proportion of the difference between active comparator and placebo. Such ideas were formulated with the aim of ensuring that the test product was superior to (a putative) placebo; however they may not achieve this purpose. If the reference product has a large advantage over placebo this does not mean that large differences are unimportant, it just means that the reference product is very efficacious.

□ It is not appropriate to use effect size (treatment difference divided by standard deviation) as justification for the choice of non-inferiority margin. This statistic provides information on how difficult a difference would be to detect, but does

not help justify the clinical relevance of the difference, and does not ensure that the test product is superior to placebo.

□ The choice of margin should be independent of considerations of power. It should be based upon the clinical and statistical principles noted in later sections of this document and not upon issues of sample size, as the size of the clinically important difference is not altered by the size of the study.
A small study is not a justification for a wider non-inferiority margin.

□ If an appropriate margin has been chosen, a confidence interval that lies entirely between $-\Delta$ and 0 (i.e. the test product is inferior to the reference, but not more than Δ worse) is still adequate to demonstrate non-inferiority. If this outcome does not seem acceptable, this demonstrates that Δ has not been chosen appropriately. (See also section V for discussion of situations where it is difficult to justify any amount of inferiority.)

□ It is not possible to perform a non-inferiority trial in all situations. The decision to perform a non-inferiority trial should be justified considering both the therapeutic area and the profile of the reference product.

□ There are many conditions where established effective agents do not consistently demonstrate superiority in placebo controlled trials (e.g. depression or allergic rhinitis). In areas where this lack of sensitivity exists, a non-inferiority trial which does not also include a placebo arm is not appropriate. See ICH E10 for a fuller discussion of assay sensitivity.

□ If the performance of the reference product in a trial is very different from what was assumed when defining the non-inferiority margin then the chosen margin may no longer be appropriate.
The implications of this should be considered at the planning stage.

3. DEMONSTRATING EFFICACY

A minimal requirement for the decision making process involved in interpreting data from a non-inferiority trial is that we must be confident that the test product would have been shown to beefficacious if a placebo controlled trial had been performed. The discussion in this section takes the methods commonly used when interpreting data from superiority trials as a model for assessing the minimal requirements for

data from non-inferiority trials.

When data from trials designed to show superiority of a test product over placebo are being interpreted, an informal two-stage procedure is employed involving the consideration of both statistical significance and clinical relevance. The same two-stage procedure can be used for interpreting non-inferiority trials. In a superiority trial, it would first be expected that the test product demonstrated a statistically significant advantage over placebo. This relates to the 'statistical reasoning' stage of the ICH E10 combination of 'both statistical reasoning and clinical judgement'. Statistical significance is generally assessed using the two-sided 0.05 level of significance (or one-sided 0.025). An alternative way of stating this requirement is that the lower bound of the two-sided 95% confidence interval (or one-sided 97.5% interval) for the difference between active and placebo should be above zero.

The next step in interpreting a superiority trial is to consider whether the difference from placebo is clinically relevant. This is the 'clinical judgement' stage of the ICH E10 combination of 'both statistical reasoning and clinical judgement'.

Establishing a clinically relevant benefit over placebo is accomplished by considering the point estimate of the difference between the test product and placebo and assessing its clinical relevance, either using the original scale or by considering responder rates. This is a not primarily a statistical issue, but does require an intelligent combination of clinical thinking and data comprehension.
Statistical significance has already been demonstrated, so the existence of an effect is considered to be established. A judgement must be made regarding whether the difference seen is clinically useful. This judgement is usually made in the context of the safety profile via an assessment of benefit/risk.

3.1 Three arm trials: test, reference and placebo

This trial design makes it possible to provide a direct demonstration of the superiority of the test and reference products over placebo. As such, it is not necessary to define a value for delta to establish that the test product has efficacy, however the following considerations should be addressed in the interpretation of the trial data.

As in the placebo controlled superiority trial, the test product must demonstrate a statistically significant advantage over placebo. The lower bound of the 95%

confidence interval for the difference between the test product and placebo should be above zero.

At this stage the performance of the reference arm is not the main consideration, although if the test and reference products both fail to demonstrate a statistically significant advantage over placebo this could suggest that the trial is insensitive, or lacks assay sensitivity.

As in a superiority trial, clinical judgement is then applied to assess whether the observed difference from placebo is clinically relevant. The existence of the reference arm can assist in making this judgement. If the reference product is a licensed agent that is known to regularly produce a clinically relevant effect in trials of this type, the reference product difference from placebo seen in this trial can be used to help assess the clinical relevance of the difference between placebo and the test product.
For example, if the test arm has performed better than the reference arm in the trial, it seems reasonable to assume that the test product's benefit is clinically relevant.

If the reference product has not demonstrated statistical significance over placebo, or has performed very differently to how experience would lead us to expect, questions could be raised about the performance of the reference product in this trial. In this situation the results from the reference arm could not provide a context, and any positive results from the test drug would have to stand alone.
Possible reasons for the unexpected results from the reference treatment should be discussed.

3.2 Two arm trials: test and reference

As there is no placebo arm in this type of trial, indirect comparisons to placebo via previous studies comparing reference to placebo must be used to establish that the product has efficacy. This presents inherent difficulties, and necessitates that a non-inferiority margin is defined. However, the lower-bound of a 95% confidence interval can still be used to establish an efficacy advantage over placebo.
The term 'putative placebo' is often used in this situation where no placebo has actually been used.

A systematic review should be conducted to identify studies relevant to the comparison of the reference treatment with placebo in the condition being

considered. These can be used for estimating the difference between the reference and placebo in the intended patient population. If such estimation is not possible, or if the comparator did not consistently demonstrate superiority over placebo in adequately powered trials, the sensitivity of a non-inferiority study using this comparator may be questioned and only superiority of the test product to a comparator (active or placebo) would be interpretable. If the reference product is part of a class where individual products are all felt to be equally effective and safe it might be acceptable to use the overall class difference from placebo.

There are several issues regarding the literature search that will need to be discussed by the applicant.(See also ICH E10 for further discussion):

1. Selection bias. The criteria used for selecting which of the available studies to include should be thoroughly documented so that it is clear that, as far as is possible, an unbiased selection of studies was made.

2. Constancy of trial design and clinical practice over time. Some of the studies may be of little relevance because clinical practice may have changed, or the criteria or methods for measuring the reference product's effect have been modified. Consideration should be given to the design of the current trial in comparison to the previous trials regarding changes that may affect treatment outcome. Examples include entry criteria, method of diagnosis, concomitant treatments allowed, dosing regime of reference product, endpoints measured, timing of assessments, etc. If possible the design of the current trial should closely match the well-designed previous trials comparing the reference with placebo. If there are unavoidable differences in trial design the implications of this should be carefully considered, and it may not be possible to formulate a non-inferiority margin.

3. Constancy of effects over time. Consideration should be given to changes in the treatment difference seen over time. For example in some conditions event rates may have decreased over time because of general improvements in healthcare. In this situation it might be appropriate to include only the more recent studies in the calculations. If constancy of effect from recent trials to the current trial cannot be assured then a conservative approach to selecting a margin should be considered.

4. Publication bias. It may be that studies with a 'positive' outcome are more likely to be published than those with disappointing results. If it seems possible that such publication bias exists, a conservative approach should be taken in producing the historical confidence interval for reference versus placebo.

If good historical data are available, several methods exist that can be used to provide a non-inferiority margin. Common to all methods is an attempt to combine the variability and size of effect from the historical data with those expected from the current trial. Also common to all the methods are the weaknesses inherent in using historical data.

The 'historical' confidence interval compares the reference product with placebo (r minus p). The planned trial comparing the test and reference products will also produce a confidence interval (for t minus r). If these intervals are combined, an indirect confidence interval comparing the test product and placebo can be obtained (t minus p). Delta can be defined as the lower bound of t minus r that ensures that the lower bound of the indirect confidence interval of t minus p will be above zero. As the comparison is indirect it might be wise to be conservative and select some value smaller than that suggested by this indirect calculation.

In a submission the applicant should present both the direct confidence interval T minus R and the indirect interval T minus P.

Once statistically significant efficacy has been established clinical relevance should be considered. The point estimate from the indirect confidence interval for t minus p should be noted and the clinical relevance of that difference considered.

As with 3-arm trials the reference arm can supply some context. If the test product outperforms the reference, this provides some assurance of the clinical relevance of the difference. However, if at the time of analysis the data lead to doubts about the performance of the reference arm in the trial, the non-inferiority margin selected may seem inappropriate and the validity of the trial may be questioned.
In the absence of a placebo arm it is more difficult to validate the performance of the reference treatment and historical data will be necessary to show that the performance in this trial is consistent with expectations.

Justification of whether the observed difference is considered to be clinically beneficial should include reference to other trials in the same therapeutic area

where clinically relevant results were seen, and just as importantly trials where the results were not considered clinically relevant.

3.3 Conditions where superiority over placebo has not been reliably established

In some disease areas placebo controlled trials may be considered unethical, yet no available comparator has reliably demonstrated efficacy over placebo. Examples include some oncology settings and some orphan indications. In such situations it will be difficult to specify delta using the considerations outlined above, however the best efforts should still be made to produce an indirect confidence interval for the new product against placebo using whatever data exist for the reference. If there are no data, the reference could be treated as placebo and efficacy could be established by demonstrating direct superiority over the reference.

It is likely that indirect confidence intervals constructed in these circumstances would fail to demonstrate superiority over placebo, but in such conditions this might not necessarily preclude a positive opinion. The delta derived using the methods of section III.2 should not be discarded, as this is the value which signifies indirect superiority over placebo. Using this value means that any decision (whether positive or negative) will be made acknowledging the limitations of the data for demonstrating efficacy, i.e. we cannot be sure the test treatment is superior to placebo. It would not be good practice to define an arbitrary achievable delta and use that to claim non-inferiority. Such an approach would create a false impression of the confidence we can have in the efficacy of the product.
It is important that the basis upon which a decision is being made is clear, whether that basis is weak or strong.

Similarly there may be situations where following the guidance of previous sections will lead to a small value of delta which cannot be excluded with a feasibly sized trial. Again it is not good practice to simply define a larger delta and then claim non-inferiority. Rather, it should be acknowledged that it has not been possible to demonstrate superiority to placebo. In some areas decisions are made based upon only small amounts of data without the demonstration of efficacy being clear, but it is important that everybody is aware that this is what is being done, and that decisions are made with full awareness of the risks being taken.

4. ESTABLISHING ACCEPTABLE EFFICACY RELATIVE TO THE ACTIVE COMPARATOR

Establishing that the new active compund would have been successful in a placebo-controlled trial is necessary but it will not usually be sufficient. The comparison between test and reference will often be of importance in its own right. In this respect it is important to define objectives before starting the trial. The design and analysis of a non-inferiority trial should reflect the question the trial is aiming to address. There are many different reasons why such a trial might be conducted, and the objective for running the trial should influence the choice of delta. Trials are generally labelled non-inferiority trials if they are not aiming to show superiority over the reference. However, 'demonstrating non-inferiority' is not considered to be a sufficiently detailed objective for a trial. A lot of clarity can be gained if more precise aims are described.

If the only objective is to show indirect superiority over placebo, this should be stated and delta can then be chosen using the methods of section III alone.

Alternatively the aim may be to provide data to show that there is no important loss of efficacy if the test product is used instead of the reference. This is probably the most common aim of non-inferiority trials. The choice of delta for such an objective cannot be obtained by only looking at past trials of the comparator against placebo. Ideas such as choosing delta to be a percentage of the expected difference between active and placebo have been advocated, but this is not considered an acceptable justification for the choice. Such ideas were principally formulated to ensure that the reference product was superior to placebo, but this has already been addressed in section III of this document. To adequately choose delta an informed decision must be taken, supported by evidence of what is considered an unimportant difference in the particular disease area.

If there are already many treatments being used interchangeably for the disease under consideration a possible approach might be to consider the information available from all of them. From this a delta may be constructed which summarises the information known about the relative efficacy of these products, and the new trial can be designed to provide a similar level of knowledge of the relative efficacy of the new product. This approach will not be possible if the market currently has only one product. In this situation, considering who will have to be persuaded to use the product after marketing authorisation, a possibility might be to survey practitioners on the range of differences that they consider to be unimportant, and choose delta based upon a summary statistic of the responses. Any such survey

should be phrased in a way that does not bias respondents towards nominating large values.

In the situation where the test product is anticipated to have a safety advantage over the reference it is likely that a larger delta could be justified as some loss of efficacy might be accepted in exchange for the safety benefits, although it would still be expected that superior efficacy to placebo should be demonstrated. In such situations it may be useful to specify co-primary endpoints, one to demonstrate superiority in terms of the safety endpoint, the other non-inferiority on the efficacy endpoint. Other circumstances which, may warrant such consideration include a more convenient route of administration, more convenient posology, superiority on a secondary efficacy endpoint, etc.

This section has only considered some of the aims of a non-inferiority trial, and some of the possible approaches to selecting delta. The main point is that the aim of the trial should be precisely defined.
Following that, a choice for delta should be made, supported by evidence, based upon the precise objectives. This evidence will not solely come from past trials of the comparator against placebo. Of course the final choice must always be at least as small as the value derived from the considerations of section III. The conclusions of the trial should not be that 'non-inferiority' has been demonstrated, but some more precise statement reflecting the objectives of the trial.

If the performance of the active comparator in the trial is very different to what was anticipated a priori there may be difficulty in interpreting the meaning of the differences between test and reference and the pre-defined delta may no longer seem appropriate. In this situation it may not be possible to draw positive conclusions from the trial. A three-arm trial including placebo provides a degree of protection against this problem as the relative effect of the active comparator compared to placebo is directly estimated in the trial.

5. AREAS WHERE IT IS DIFFICULT TO JUSTIFY A NON-INFERIORITY MARGIN OF ANY SIZE

Where the treatment under consideration is used for the prevention of death or irreversible morbidity and there is no second chance for treatment it can be very difficult to justify a non-inferiority margin of any size. Discussion of the number of extra deaths that are acceptable is ethically very difficult.

However it is not in the best interest of public health to reject all choices of margin. Unless a statistically significant difference has been found between treatments, the confidence interval for the difference will not only indicate that the test product has a possible inferiority to the reference, but it will also show that it has possible superiority. Hence even if we think we are not prepared to accept any possible level of inferiority we are accepting some, by continuing to use the currently authorised product. It is important therefore that non-inferiority trials should still be possible in these areas. This section discusses some approaches to help facilitate this.

5.1 Superiority using an increased significance level

As noted in section I.4, allowing no non-inferiority margin ($\Delta=0$, essentially a superiority trial) prevents equally efficacious products from producing positive trial data, except by chance. Even products with small but clinically useful advantages would only consistently demonstrate their benefit in huge trials.

Focusing, for now, on the main efficacy endpoint only, in this particular situation there are three main reasons why a non-inferiority trial might be run rather than a trial designed to show superiority over the reference:

1. The products truly are equally efficacious, leaving a non-inferiority trial as the only option.
2. The test product has a small advantage that would require such a large trial to detect as to be impractical.
3. The product has a disadvantage, but that disadvantage is smaller than a proposed non-inferiority margin.

Obviously it is important for public health that products falling into category 2 are able to pass the tests set up more often than they fail. However it would be better, in these extreme conditions, if those in category 3 did not succeed. Any requirements set up must find this balance. The success or failure of products in category 1 is less important from a public health perspective.

This leads us to conclude that in such critical areas, a point estimate on the wrong side of zero can rarely be acceptable. With such data we are more likely to be in category 3 than category 2.

It is helpful here to parallel the setting of a non-inferiority margin with the idea of

significance testing.

If the 95% confidence interval were entirely above zero, we have established superiority at the 5% level of significance. For each particular value for the lower bound of our 95% confidence interval there is another confidence interval with some other coverage probability that for the same data would have a lower bound of exactly zero. For example with a data-set where the lower bound of an 85% confidence interval (by definition narrower than a 95% interval) touches zero, it might be that the 95% interval touches -5. If delta had been defined to be -5 then achieving non-inferiority in this example would correspond to having demonstrated superiority at the 15% level of significance. Hence we can parallel running a non-inferiority trial to running a superiority trial at a less stringent significance level.

It might be an acceptable approach, in extreme situations, to run a superiority trial using a less stringent significance level than P=0.05, weighing up the increased risk of a false positive result against the risk of rejecting a drug with a valuable efficacy advantage. It might be more acceptable, and easier from an ethical perspective, to specify a level of confidence we require in the superiority of a drug, than to specify an extra number of deaths that is of no clinical importance.

It is recognised that a certain delta is implicitly hidden in the increased alpha level, and that this approach does not mean that all possible inferiority is ruled out. However, it is important to remember that, although 95% confidence and 5% significance have become commonly accepted, there is still a possibility of false positive results even using this significance level. Here we are merely increasing the chance of a false positive superiority claim to reduce the chance of a false negative, noting that the optimal balance between the two might be different in extreme situations.

If this approach is used the objectives and the hypothesis of the trial should be clearly stated in advance. If the hypothesis is that the test product has an advantage, this should be stated. A trial where the results support a sound pre-specified hypothesis is more persuasive than one where the results are surprising.

It would not be acceptable to switch to this approach retrospectively upon seeing study results which failed to achieve significance at conventional levels. The plan and justification for using an increased significance level would need to be clearly

stated in the study protocol.

If this approach is to be used it is strongly recommended that scientific advice or protocol assistance be sought on whether the CHMP consider it to be appropriate for the particular case under study.

5.2 Products with an advantage in another aspect

In some cases if the product has an advantage in another important facet of its profile, it may be possible to define a non-inferiority margin for efficacy. If this is the case a point estimate for the difference between treatments that is on the wrong side of zero could be acceptable. In this situation it would be advisable to have two primary endpoints and plan to show non-inferiority for efficacy and superiority for the other important factor.

6. CONCLUSIONS

□ The selection of the non-inferiority margin is based upon a combination of statistical reasoning and clinical judgement.

□ A three-armed trial with test, reference and placebo allows some within-trial validation of the choice of non-inferiority margin and is therefore associated with fewer difficulties. This is the recommended design and should be used wherever possible.

□ An appropriate choice of margin should provide assurance that the test drug has a clinically relevant superiority over placebo. This aspect of the choice of margin is discussed in section III.

□ Usually the primary focus of a non-inferiority trial is the relative efficacy of the test and reference products, not simply demonstration that the test product has an effect.. In these cases an appropriate choice of margin will, in addition to proving that the product has an effect, also provide assurance that the test product is not substantially inferior to the reference, resulting in a tighter margin. This aspect of the choice of margin is discussed in section IV.

□ For the majority of non-inferiority trials it must be demonstrated that the margin satisfies both the requirements of section III and section IV. The choice of

non-inferiority margin should be justified in the study protocol, and the justification should address the considerations of both sections.

□ It is not possible to perform a non-inferiority trial in all situations. The decision to perform a non-inferiority trial should be justified considering both the therapeutic area and the profile of the reference product.

□ It may be possible to justify a wider non-inferiority margin for efficacy if the product has an advantage in some other aspect of its profile. This margin should not, however, be so wide that superiority to placebo is left in doubt.

□ In some extreme situations it may be acceptable to run a superiority trial specifying a significance level greater than 0.05 as an alternative to defining a non-inferiority margin.

〈 비열등성 허용한계 설정 시 고려사항 〉
자문위원 명단

강승호 (연세대학교)	김선우 (삼성의료원)
남정모 (연세대학교)	박용규 (가톨릭대학교)

비열등성 허용한계 설정 시 고려사항

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