臨 床 藥 理 學 會 誌 第20卷 第1號 2012
J Korean Soc Clin Pharmacol Ther 2012;20(1):17-33

바이오시밀러의약품의 동등성한계 설정

18515 Fontana Lane, Gaithersburg, MD 20879, USA

이형기

=Abstract=

Equivalence Margin of the Biosimilar Product

Howard Lee

18515 Fontana Lane, Gaithersburg, MD 20879, USA

The equivalence margin is the largest difference that is clinically acceptable between the test (or experimental) drug and the active control (or reference) drug. This paper discusses the scientific principles, along with the regulatory issues, that need to be addressed when determining the equivalence margin for the biosimilar product. The concept of assay sensitivity is introduced, and the ways to ensure assay sensitivity in the equivalence trial are emphasized. A hypothetical example is presented to show how an equivalence margin is determined. The regulatory agency should carefully assess if the equivalence margin of the biosimilar product was determined using a scientifically valid and clinically relevant approach, not subject to selection bias. This is important because the consumer risk of erroneously declaring equivalence when in fact it is not must be controlled conservatively low in the approval of any biosimilar products.

Key words: Biosimilar product, Equivalence margin, Assay sensitivity, Selection bias

서 론

'바이오시밀러의약품(biosimilar products)' 또는 '동등생물의약품'은 규제기관의 엄밀한 심의를 거쳐, 오리지널 생물학적의약품과 비견(比肩)될 만하다(comparable)고 인정된 생물학적의약품을 가리킨다.[1] 최근 오리지널 생물학적의약품들 중에서 특허 만료를 앞 둔 제품들이 많아지면서 후발 제약사들이 개발하는 바이오시밀러의약품에 대한 관심이 증가하고 있다. 특허 만료를 겨냥한 후속 제품이라는 관점에서, 바이오시밀러의약품은 화합물의약품의 제네릭과 비슷한 개념이라고 할 수 있다. 하지만 생물학적의약품의 특성상, 오리지널 의약품과 완전히 똑같은 바이오시밀러의약품은 존재하지 않는다. 사소한 제조 공정의 차이가 완전히 다른 생물학적의약품을 만들어 낼 수 있기 때문이다.[2]

교신저자: 이형기
주 소: 18515 Fontana Lane, Gaithersburg, MD 20879, USA
전화번호: 1-240-994-8384, Fax: None, E-mail: leehwd@gmail.com
접수일자: 2012. 05. 15. 수정일: 2012. 06. 05. 게재확정일: 2012. 06. 08.

따라서 바이오시밀러의약품에 대해서는 '동일한(identical)'이나 '같은(the same)'이라는 용어보다 '비견될 만한(comparable)' 또는 '유사한(similar)' 등과 같은 수식어가 더 흔히 사용된다. 요컨대 바이오시밀러의약품에서 '동등성(equivalence)'은 오리지널 의약품과 '똑같다'는 의미가 아니다. 대신, 오리지널 생물학적의약품과 바이오시밀러의약품 사이에 존재하는 이화학적·생물학적·전임상적·임상적 특성들의 차이가, 그 이상을 넘어서면 유사 또는 비견될 만하다고 간주할 수 없는 어떤 한계보다 작은 경우를 가리켜 동등성이라는 개념으로 정의한다. 그리고 이러한 한계를 '동등성한계(equivalence margin)'라고 부른다. 결국 동등성한계란 '수용할 수 있는, 바이오시밀러의약품과 오리지널 의약품 사이의 가장 큰 차이'가 되는 셈이다.[3)]

동등성과 비슷한 개념에 '비열등성(non-inferiority)'이 있다. 한 마디로 비열등성이란 어떤 의약품이 기준이 되는 다른 의약품에 비해 어떤 한계 이상으로 더 못하지는 않다는 뜻이다.[3)] 따라서 바이오시밀러의약품과 오리지널 생물학적의약품의 관점에서 각각 따로 정의한 비열등성의 두 경우를 합하면 동등성이 된다. 통계학적으로는 비열등성이 단측(one-sided) 검정인데 반해, 동등성은 양측(two-sided), 보다 정확하게는 두 단측(two one-sided) 검정이 된다. 하지만 기타 접근 방법이나 원리는 비열등성과 동등성 사이에 서로 비슷하다.

바이오시밀러의약품을 허가 받으려면 비열등성보다 동등성을 입증하는 것이 더 중요하다. 식품의약품안전청(식의약청)이 2009년에 발표한 '동등생물의약품 평가 가이드라인'에도 이 사실이 분명하게 기술돼 있다. "대조약의 용법 용량을 그대로 적용하고, 유효성 자료를 대조약의 다른 적응증으로 외삽(다른 용량으로의 외삽 포함)하기 위해서는 일반적으로 비열등성보다는 동등성(equivalence)을 입증하여야 한다."[4)] 세계보건기구(WHO)[5)]는 물론, 미국의 Food and Drug Administration (FDA)[6)]와 유럽의 European Medicines Agency (EMA)[7)]가 발표한 바이오시밀러의약품의 허가 지침들도 동등성이 비열등성보다 중요하다는 입장을 취하고 있다.

따라서 동등성한계는 바이오시밀러의약품의 허가에 결정적인 영향을 끼치는 중요한 변수다. 뿐만 아니라, 동등성한계를 적절하게 설정했는지 따져 봄으로써, 규제기관이 과연 과학적으로 엄밀하고 임상적으로 타당한 기준에 의해 바이오시밀러의약품을 심의했는지 판단할 수 있다. 동등성한계와 개념적으로 유사한 '비열등성 한계(non-inferiority margin)'를 과학적·임상적 관점에서 타당하게 설정하는 방법에 대해 매우 상세히 설명하고 있는 FDA의 지침에서 "비열등성 한계를 결정하는 것이 비열등성시험의 고안·수행·해석에서 가장 큰 쟁점"이라고 강조한 것도 그 이유다.[8)]

그런데 바이오시밀러의약품이 규제기관의 관심을 받게 된 것은 최근의 일이므로 동등성한계를 구체적으로 결정하는 방법에 대해 아직까지 종합적인 지침이 정리되거나 발표된 바 없다. 그러나 약간의 변형을 거친다면, 전술한 것처럼 비열등성 한계를 설정하는 과학적, 임상적 원칙이 동등성한계 설정의 경우에도 그대로 적용될 수 있다.[9,10)] 예를 들어, 바이오시밀러의약품의 허가에 대한 일반 원칙을 상술한 FDA의 지침[6)]에서도 위에서 소개한 비열등성 한계 설정에 관한 지침[8)]을 참고할 것을 여러 차례 언급하고 있다.

이상의 배경에 따라, 본고에서는 동등성한계 설정에 관련된 각종 원리와 쟁점들, 특히 어떻게 하면 과학적으로 엄밀하며 임상적으로 타당한 방법을 사용해 동등성한계를 설정할 수 있는지 살펴보려고 한다. 이어 가상 자료를 이용하여 바이오

시밀러의약품의 동등성한계를 설정하는 과정을 구체적으로 예시하고자 한다. 이러한 과정을 통해 동등성한계 설정에 영향을 미치는 각종 쟁점들이 함께 검토될 것이다.

생동성시험에서 동등성한계 설정

바이오시밀러의약품의 동등성한계 설정 방법을 소개하기에 앞서, 생물학적동등성(bioequivalence, 생동성)시험에서 동등성한계가 갖는 의미를 먼저 따져 보는 것이 도움이 된다. 생동성시험이란 제네릭의약품의 허가를 얻기 위해, 이미 허가를 받은 기준의약품(reference drug)과 제네릭의약품을 대개 교차연구(crossover study)의 형태로 투여한 뒤, 혈중 농도로부터 구한 약동학적 경수(pharmacokinetic parameter)를 비교함으로써 동등성 유무를 판단하는 임상연구를 가리킨다.

생동성 판단에 사용하는 약동학적 경수는 전신혈 흡수(systemic absorption)의 속도와 정도를 각각 대변한다고 알려진 최고농도(C_{max})와 농도-시간곡선하면적(area under the concentration-time curve, AUC)이다. 적절한 분산분석모형을 사용해 이월효과(carryover effect) 등이 배제된 것을 확인한 뒤, C_{max}와 AUC의 제네릭의약품 대 기준의약품 기하평균비(geometric mean ratio)의 양측 90 % 신뢰구간이 [0.8, 1.25] 범위 안에 들면 동등하다고 간주한다. 동등성 영역의 하한측 한계치 0.8은 제네릭의약품 대 기준의약품의 기하평균비가 80 %(=0.8)라는 뜻이며, 기하평균비는 자연대수(natural logarithm, ln)를 취한 변수의 산술평균이 되므로 ln(0.8)과 절대값이 같은 ln(1.25), 즉 1.25를 동등성 영역의 상한 측 한계로 정한 것이다(Figure 1). 로그의 특성상, $\ln(A/B)=\ln(A)-\ln(B)=-[\ln(B)-\ln(A)]=-\ln(B/A)$, 즉 비와 그 비의 역수에 대한 로그는 절대값이 같은 양수와 음수의 짝이 된다. 요컨대 생동성시험에서 동등성한계를 20 %로 할 때 이는 로그변환 이전의 값에 대한 비가 0.8/1=0.8이 되는 경우이므로, 로그변환을 했을 때 0.8과 절대값이 같아지려면 0.8의 역수, 즉 1.25(=1/0.8)가 동등성한계 영역의 상한이 된다.

대상 약물의 약동학이나 임상적인 특성에 따라 동등성 영역을 다른 값으로 정하는 것이 더 적절할 수도 있다. 예를 들어 '치료영역이 좁은(narrow therapeutic index)' 약물의 동등성 영역은 [0.9, 1.11]로 축소하며, 이 때 양측 90 % 신뢰구간이 반드시 1 을 포함해야 한다. 반대로 '개체 변이가 큰 약물(highly variable drug)'의 동등성 영역은 [0.8, 1.25]보다 넓게 정하기도 한다.

요컨대 생동성시험에서는 다음과 같은 단계를 거쳐 생동성 유무를 평가하게 된다. 첫째, 동등성한계를 정한다. 앞에서 살펴 본 것처럼, 생동성시험에서는 특별한 경우를 제외하고 동등성한계는 [0.8, 1.25]로 고정돼 있다. 둘째, 적절한 비교지표를 정한다. 생동성시험에서는 기준의약품 대비 제네릭의약품의 C_{max}와 AUC의 기하평균비를 사용한다. 셋째, 비교지표의 점추정치(point estimate)를 구하고, 적절한 방법을 사용해서 해당 점추정치에 대한 신뢰구간을 추정한다. 생동성시험에서는 양측 90 % 신뢰구간을 구한다. 넷째, 신뢰구간이 동등성한계 영역 내 포함되는지 비교한다. 만일 신뢰구간 전체가 생동성의 동등성한계 영역인 [0.8, 1.25] 범위 안에 완전히 포함되면 생동성이 입증된 것으로 간주한다(Figure 1에서 위로부터 첫 번째 경우). 그러나 신뢰구간의 일부가 동등성한계치에 걸쳐 있거나(Figure 1에서 위로부터 두 번째 및 세 번째 경우), 아니면 아예 신뢰구간 전체가 동등성한계의 하한치(0.8) 보다 낮은 쪽

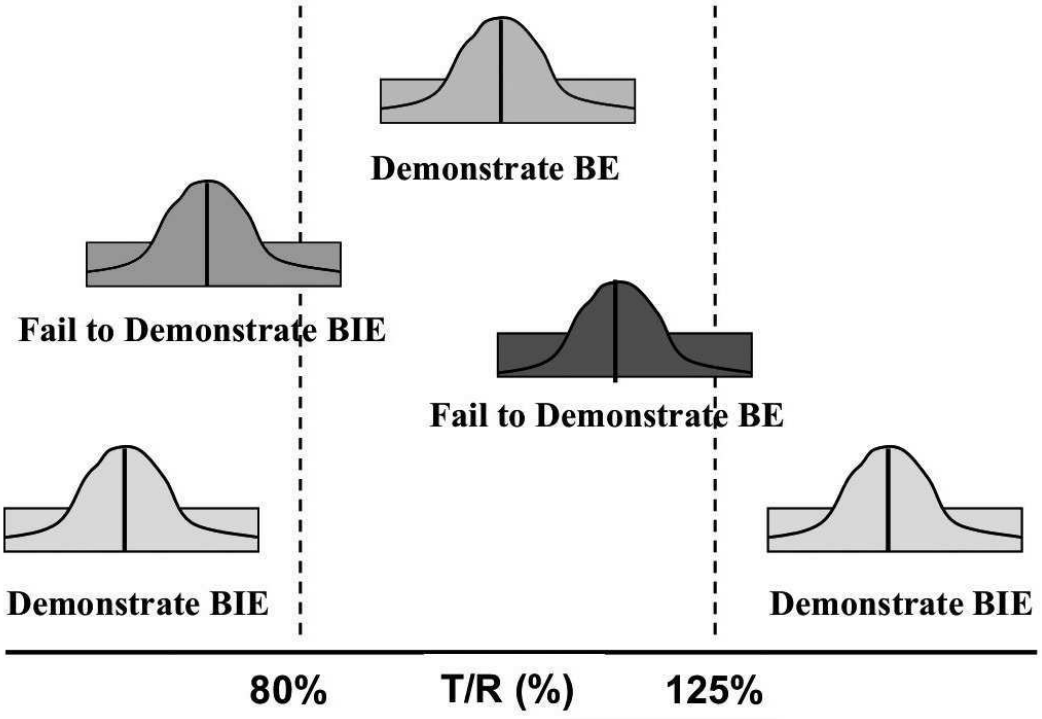

Figure 1. Assessment of bioequivalence of the generic product (T) against the reference product (R). Bioequivalence is declared if the two one sided 90 % confidence interval of the geometric mean ratio (T/R) falls entirely within the range of [0.8, 1.25]. BE and BIE represent bioequivalence and bioinequivalence, respectively. Adapted from Lawrence X. Yu, PhD., Deputy Director for Science and Chemistry, Office of Generic Drugs, FDA, "Approaches to Demonstrate Bioequivalence of Narrow Therapeutic Index Drugs", Advisory Committee for Pharmaceutical Science and Clinical Pharmacology July 26, 2011.

(Figure 1에서 위로부터 네 번째의 왼쪽) 또는 상한치(1.25) 보다 높은 쪽에 있으면(Figure 1에서 위로부터 네 번째의 오른쪽) 생동성 입증에 실패한 것이 된다.

본고의 뒷부분에서 더 분명하게 기술하겠지만, 바로 위에서 요약한, 생동성 입증에 필요한 각 단계 및 원리가 바이오시밀러의약품의 동등성 입증에도 거의 그대로 적용된다. 한편 생동성시험의 동등성 입증 원리로부터 다음과 같이 바이오시밀러의약품의 동등성 입증과 관련된 몇 가지 시사점을 도출할 수 있다.

첫째, 생동성시험에서는 동등성한계가 [0.8, 1.25]로 '고정'돼 있는데, 이는 혈중 농도와 같은 약동학적 변수가 임상적 결과변수보다는 단순하고, 측정 시점이 약물 투여에 가까워 약물 자체의 흡수 특성이 잘 반영되며, 무엇보다 동일한 비교지표-C_{max}와 AUC의 기하평균비-를 사용한다는 사실과 연관이 있다. 하지만 바이오시밀러의약품의 동등성 입증에 필요한 동등성한계는 약물의 종류나 대상 질환에 따라서 얼마든지 다를 수 있

다. 뒤에서 더 자세히 논의하겠지만, 위약 대비 기준의약품의 효과가 어느 정도인가에 따라 바이오시밀러의약품의 동등성한계는 가변적이며 또 당연히 가변적이어야 한다. 한편 동등성한계는 좌우 대칭인 경우가 대부분이지만, 타당한 이유가 있다면 비대칭으로 설정하는 것도 가능하다. 참고로 생동성시험에서 동등성한계인 [0.8, 1.25]는 좌우 비대칭처럼 보이지만, 실제로는 로그변환한 변수에 대해 대칭이다. 즉 |ln(0.8)|=| - ln(1.25)|이다.

둘째, 바이오시밀러의약품의 동등성을 입증할 때에도 적절한 비교지표를 선택하는 것이 중요하다. 예를 들어 생동성시험의 경우처럼 기준의약품 대비 바이오시밀러의약품의 효과에 대한 비(ratio)를 이용할 수 있다. 이러한 비의 종류에는 치료 효과의 비 이외에도 상대위험도(relative risk) 또는 위험비(risk ratio)·hazard ratio·교차비(odds ratio) 등이 있고, 이들 비교지표를 로그변환한 지표들도 사용할 수 있다.[8] 한편 비와 함께 '차이(difference)', 즉 치료 효과의 절대적인 차이도 동등성을 입증하기 위해 흔하게 사용되는 비교지표이다. 특히, 어떤 결과변수에 대해 분율(proportion)로 치료 효과를 평가하는 경우에는 분율의 차이가 가장 적절한 비교지표가 될 수 있다. 치료율·반응률·성공률·생존률 등이 좋은 예로, 항생제나 항암제의 동등성을 입증할 때에는 분율의 차이를 비교지표로 사용하는 게 일반적이다. 연속변수의 평균값으로 치료 효과를 판정하는 경우에도 차이는 적절한 비교지표가 된다. HbA_{1c}(당뇨), 총콜레스테롤(고지혈증) 등이 그 예이다.

셋째, 생동성의 경우처럼 바이오시밀러의약품 허가에서도 비교지표의 신뢰구간이 동등성한계 영역에 전적으로 내포되는 경우에 한해 동등성이 입증된 것으로 간주하는 '신뢰구간 방법'을 사용한다. 여기에서 신뢰구간은 비교지표의 점추정치에 내재된 통계적 불확실성에 대비하는 일종의 안전 장치라고 볼 수 있다. 예를 들어, 차이를 비교지표로 사용하는 경우에 비교지표의 점추정치가 0 즉 완벽하게 동등한 것처럼 보이는 경우--'비'를 동등성 검정의 비교지표로 사용한다면, 점추정치가 1인 경우--를 가정해 보자. 그런데 이 점추정치는 확률 변수이기 때문에 비록 이 점추정치가 참값이라고 하더라도 동일한 동등성시험을 반복해서 실시한다면 단순히 우연(확률)에 의해 점추정치가 0보다 작게 나올 경우가 50 %, 0보다 크게 나올 경우가 50 %이다. 따라서 점추정치 대신 점추정치의 신뢰구간이 동등성한계 영역 내에 완전히 포함되는지를 따져 봄으로써 참값이 신뢰구간의 영역 밖에 위치할 수도 있는 불확실성이 '100 - 신뢰도' %를 넘지 않도록 조정할 수 있게 된다. 이는 일반적인 우월성시험(superiority trial)에서 '100 - 신뢰도' %의 유의수준으로 통계 검정을 실시하는 것과 개념적으로 유사하다. 동등성시험에서는 일반적으로 양측 95 % 신뢰구간을 사용한다.

따라서 바이오시밀러의약품의 동등성을 입증하려는 후발 기업은 당연히 동등성한계를 크게 설정하고 싶어한다. 하지만 다음 항목에서 설명한 것처럼, 환자나 규제기관의 입장에서는 동등성한계를 타당한 범위 내에서 가능한 작게 유지하는 것이 보다 타당하다. 이러한 입장을 흔히 '보수적(conservative)'이라고 표현하는데, 이는 동등성 입증이 까다로워야 한다는 뜻이다.

넷째, 새로 개발하는 치료약의 우월성을 위약이나 기타 대조약과 직접 비교 - 입증하는 소위 우월성시험과는 달리, 동등성시험이나 비열등성시험에서는 치료약의 효과가 직접 검증되지 않는다. 왜냐하면, 동등성시험이나 비열등성시험은 과거에 입증됐던 위약(또는 대조약) 대비 기준의약품의 우월한

치료 효과가 그대로 나타났을 것이라고 '가정'한 상태에서-이를 assay sensitivity라고 한다-, 바이오시밀러의약품의 치료 효과가 특정 범위 안에서 기준의약품의 치료 효과보다 '못 하지도 더 낫지도 않다(동등성시험)'거나 '못 하지 않다(비열등성시험)'는 주장을 간접적으로 입증하려는 시도이기 때문이다. 이처럼, 동등성시험과 비열등성시험은 연구방법론의 특성상 우월성시험과는 비교할 수 없을 만큼 큰 불확실성을 갖고 있다. 따라서 이러한 불확실성으로부터 국민의 건강을 보호하기 위해서는 가능한 위양성오류(false-positive error), 즉 '바이오시밀러의약품이 오리지널 의약품과 동등하지 않은데도 불구하고 동등한 것으로 잘못 간주'할 가능성-이를 consumer risk라고 한다-을 최소화하는 것이 더 중요하다. 이미 유효성과 안전성이 입증된 오리지널 의약품이 허가를 받아 시장에 존재하기 때문이다. 요컨대 국민의 건강권 보호를 위해서 규제기관이 바이오시밀러의약품의 허가를 보수적으로 운영하는 것은 적절하고 필요하다.

바이오시밀러의약품의 동등성 평가 원리

앞에서 지적한 것처럼, 생동성 유무를 평가하는 원리가 동등성시험에 거의 그대로 적용될 수 있다.[9,10] Figure 2는 Figure 1과 유사한데, 바이오시밀러의약품(T)과 기준의약품(R)의 치료 효과의

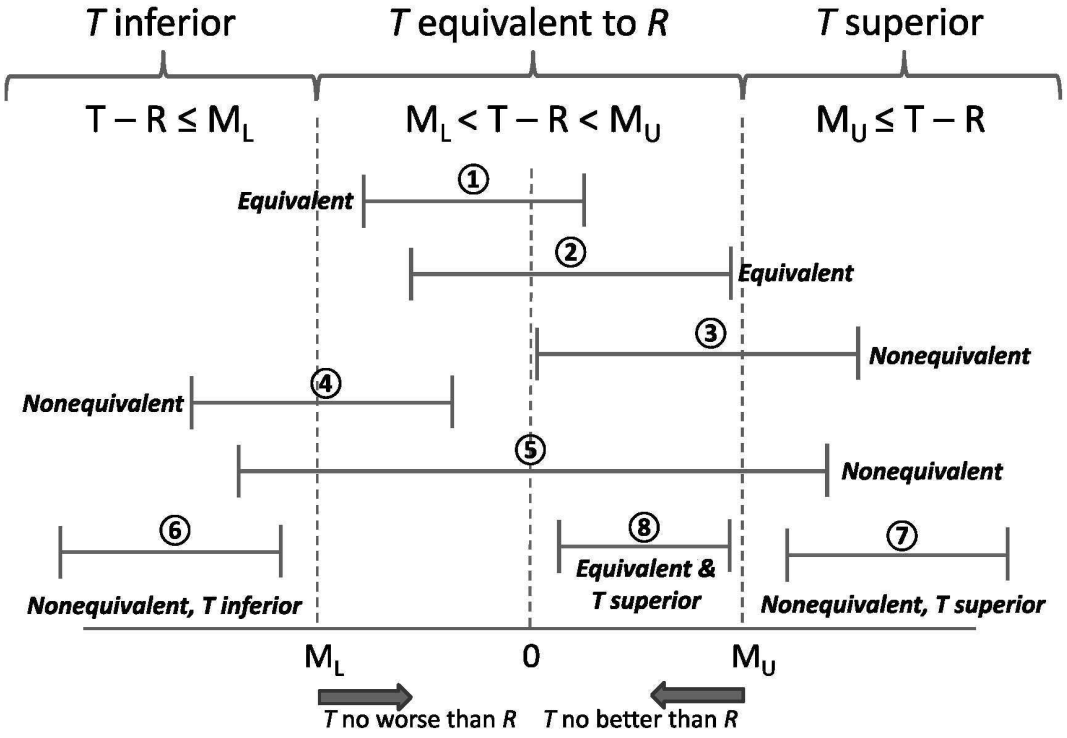

Figure 2. Assessment of equivalence of the biosimilar product (T) against the reference product (R) using the equivalence margin of [M_L, M_U]. Equivalence is declared if the confidence interval of the comparative index falls completely within the equivalence margin.

차이(T－R)를 비교지표로 사용해 동등성을 평가하는 방법을 모식적으로 보여 준다. 물론, 치료효과의 비나 다른 비교지표를 사용하더라도 기본적인 원리는 동일하다.

어떻게 동등성한계(M_L과 M_U)가 설정됐는지는 다음 항에서 자세히 다룰 것이므로 일단 논외로 하자. M_L은 바이오시밀러의약품이 기준의약품보다 더 나쁘지는 않다고 받아 들일 수 있는, 치료효과 차이의 가장 큰 값(즉 $T-R > M_L$), 그리고 M_U는 바이오시밀러의약품이 기준의약품보다 더 좋다고는 할 수 없는, 치료 효과 차이의 가장 큰 값이 된다(즉 $T-R < M_U$). 동등성한계를 좌우대칭으로 설정하는 게 일반적이므로, M_U가 임의의 양수(>0)라고 한다면, $M_L = -M_U = |M|$, 즉 M_U와 M_L은 절대값이 같으면서 부호는 다른, 임의의 한 숫자(M)가 될 것이다. 그러나 여기에서는 보다 일반적인 경우를 상정해서 동등성한계의 하한과 상한이 다른 절대값을 갖는 것으로 간주했다.

바이오시밀러의약품이 기준의약품과 '동등'하다는 것은 두 의약품의 치료 효과의 차이가 M_U보다는 작되, M_L보다는 큰 경우가 된다(즉 $M_L < T-R < M_U$). 그리고 생동성의 경우처럼 신뢰구간 방법을 사용해서 바이오시밀러의약품과 기준의약품의 동등성을 평가하므로, 치료 효과의 차이에 대한 양측 신뢰구간이 동등성한계, 즉 [M_L, M_U] 안에 완전히 내포될 때 동등성이 입증됐다고 선언할 수 있다(Figure 2에서 ①, ②). 하지만 신뢰구간이 동등성한계치에 걸쳐 있거나(Figure 2에서 ③, ④, ⑤), 아예 동등성한계 영역 밖에 위치하면(Figure 2에서 ⑥, ⑦) 두 의약품은 동등하지 않다고 간주한다. 흥미로운 것은 Figure 2에서 ⑧로 표시된 경우인데, 신뢰구간이 일단 동등성한계 영역 내에 위치하므로 동등하다고 말 할 수 있지만, 동시에 신뢰구간이 0을 포함하지 않아서 바이오시밀러의약품의 우월성도 입증되는 경우이다. 그러나 이런 경우는 매우 드물고, 또 해석상 많은 주의를 요한다.[8)]

바이오시밀러의약품에 대한 동등성한계 설정의 원리

제 2항에서 소개한 '보수적' 관점이야말로 동등성시험이나 비열등성시험의 고안·실시·자료 분석 과정에서, 그리고 그 결과에 따라 의약품의 허가를 결정해야 하는 규제기관이 가장 명심해야 하는 핵심 단어이다. 이 항의 뒤에서 다시 설명하겠지만, 동등성 또는 비열등성 한계는 기준의약품의 위약 대비 효과보다 절대로 클 수 없다. 따라서 과거 연구에서 추정한 기준의약품의 치료 효과를 낮은 쪽으로 상정해야 이것보다 당연히 작은 값을 가져야만 하는 동등성/비열등성 한계가 작아지게 되고, 결국 보수적으로 동등성/비열등성을 평가할 수 있다. 보다 구체적으로는, 기준의약품의 과거 위약 대조 연구에서 추정된 치료 효과의 차이에 대한 95 % 신뢰구간의 '하한값'을 동등성한계가 취할 수 있는 상한값으로 정하게 된다.

이러한 보수적 원리는 동등성한계를 설정할 때 가장 극명하게 나타난다. 비열등성 한계를 설정하는 방법에 대해 상술한 FDA의 지침이 assay sensitivity의 중요성을 언급하는 다음 문장에 이러한 보수적 원리의 중요성을 잘 표현하고 있다. 그리고 이 사실을 반영하듯, 관련 FDA의 지침에서는 'conservative'라는 용어가 무려 36 회나 사용되고 있다.

"[기준의약품의] 치료 효과의 크기에 대한 추정치를 보수적으로 선택하는 것(즉 과거 연구들이 보여 준 바 있는 치료 효과가 비열등성시험에서도 그대로 유지될 가능성이 높도록 하는 것)이 매

우 중요한데, 이는 비열등성시험을 관통하는 전체적인 원칙이, 활성이 있는 대조치료(기준의약품을 뜻함-역주)가 적어도 M1, 즉 비열등성 한계가 가질 수 있는 최대값과 같은 크기의 치료 효과를 비열등성시험에서 나타낼 것이라는 보장에 전적으로 근거하고 있기 때문이다."[8] (지침의 21-22 페이지)

제 3항에서 살펴 본 것처럼, 동등성한계가 설정되면 동등성 유무를 평가하는 것은 비교적 용이한 일이 된다. 그러나 동등성한계 자체를 설정하는 것은 결코 쉽지 않은 일이다. 왜냐 하면, 이 값은 동등성시험의 자료로부터 계산해 낼 수 있는 것이 아니기 때문이다. 대신 동등성한계, 또는 보다 정확하게는 동등성한계가 취할 수 있는 가장 큰 값은 과거에 위약 대조로 실시된 기준의약품의 우월성시험의 자료로부터 '추정'해 내야 한다. 따라서 동등성시험에서 얻어진 자료 해석 방법의 적합성은 이 추정 과정이 얼마나 과학적으로 엄밀하고, 임상적으로 타당했느냐에 달려 있다.

Figure 2에서 0을 기준으로 왼쪽, 즉 기준의약품이 바이오시밀러의약품보다 우월함을 보이는 영역(즉 $T-R < 0$)에 먼저 초점을 맞추어 보자. 위약(P) 대비 기준의약품의 치료 효과는 R-P로 나타내지만, 일반적으로 비열등성 또는 동등성 시험에서는 기준의약품 대비 바이오시밀러의약품의 치료 효과, 즉 T-R처럼 R의 위치가 바뀌게 된다. 따라서 동등성한계를 크게 설정한다는 것은 M_L의 절대값이 더 커지는 쪽으로, 또는 M_L이 더 낮은 음수값을 갖는 쪽으로 하향 조정한다는 뜻이다.

동등성한계를 위약 대비 기준의약품의 치료 효과보다도 크게 설정하는 경우를 생각해 보자. 예를 들어, 위약 대비 기준의약품은 단지 10만의 효과만을 갖는데, M_L을 -20으로 정하는 것 같은 경우가 여기에 해당한다. 이 경우 바이오시밀러의약품이 위약과 같은 정도의 효과를 갖는다고 해도 실제 동등성시험에서 T-R이 -10보다 작은 값을 가질, 또는 -10의 절대값보다 큰 음수값을 가질 가능성이 거의 없다. 다시 말해서, 이것은 바이오시밀러의약품이 위약 정도의 효과만을 갖는데도 불구하고 마치 기준의약품에 준하는 효과가 있는 것처럼 보이게 만드는 상황이다. 즉 바이오시밀러의약품이 기준의약품보다 열등하지만 열등하지 않은 것처럼 보이는 위양성오류가 증가되는 경우로, 앞에서 강조한 바 있는 비열등성 또는 동등성 시험의 보수적 접근 원칙에 위배된다. 그리고 이런 상황이 연출된 것은 동등성한계를 위약 대비 기준의약품의 치료 효과보다 크게 설정했기 때문이다. 요컨대 어떤 경우에도 M_L의 절대값이 위약 대비 기준의약품의 치료 효과보다 클 수는 없다는 사실이 분명해진다.

이번에는 Figure 2에서 0을 중심으로 오른쪽, 즉 바이오시밀러의약품이 기준의약품보다 우월함을 보이는 영역(즉 $T-R > 0$)을 살펴 보자. 앞에서 적용했던 원리를 그대로 따르면 M_U는 위약 대비 바이오시밀러의약품의 치료 효과보다 크게 설정할 수 없음을 쉽게 이해할 수 있다. 그런데 기준의약품과는 달리 바이오시밀러의약품의 경우에는 위약 대비 치료 효과를 쉽게 추정할 수 없다. 하지만 바이오시밀러의약품이 기준의약품과 동등하다면 바이오시밀러의약품의 위약 대비 치료 효과가 기준의약품의 위약 대비 치료 효과와 동일할 것으로 예상할 수 있다. 결국, M_U는 위약 대비 기준의약품의 치료 효과로 대신 추정할 수 있다. 이렇게 되면, 근사적으로 $|M_L| = |M_U|$가 성립하고, 따라서 동등성한계 영역은 $[M_L, -M_L]$로 주어진다. 그리고 동등성한계 영역의 상한과 하한에 해당하는 값들이 절대값이 같은 양수, 음수가

되므로 보다 간략하게는 [-M, M]이라고 표현해도 된다. 물론, 여기에서 M은 위약 대비 기준의약품의 치료 효과에 해당한다.

동등성한계가 취할 수 있는 가장 큰 값이 M이라는 것은 당연하지만, 그렇다고 동등성한계가 반드시 M만큼 될 필요는 없다. 왜냐 하면, 기준의약품 대비 바이오시밀러의약품의 치료 효과 차이(T-R)가 동등성한계의 상한인 M 또는 하한인 -M과 같다면, 이것은 각각 기준의약품 또는 바이오시밀러의약품이 위약 정도의 효과만을 갖고 있다는 뜻이 되기 때문이다. 따라서 M보다 작은 숫자를 택해 동등성한계로 정하는 것이 임상적으로 타당하고 합리적이다. 예를 들어, 기준의약품과 위약 사이에 치료 효과의 차이가 10이었다면, 바이오시밀러의약품은 10이라는 차이에서 적어도 70 %-또는 임상적 판단에 따라 어떤 %라도-이상을 갖고 있어야만 동등한 것으로 간주해야 임상적으로 타당하다. 이 예에서는 기준의약품과 바이오시밀러의약품 사이의 치료 효과가 30 (=100-70) % 미만이 되어야 한다. 즉 최종 동등성한계는 3(=10×0.3)이 되는 것이다. 이 둘-앞의 예에서 10과 3-을 구분하기 위해 위약 대비 기준의약품의 치료 효과는 M1(앞의 예에서 10)으로, 그리고 M1보다 작아야 하는 동등성한계는 M2(앞의 예에서 3)으로 지칭한다. 결국 M2는 위약 대비 기준의약품의 치료 효과(M1) 중에서 바이오시밀러의약품이 잃어 버려도 임상적으로 동등한 것으로 수용할 수 있는 가장 큰 값이다. 이러한 관계를 정리하면 다음과 같이 표현할 수 있다.

$$-M1 < -M2 < 0 < M2 < M1 \cdots\cdots (1)$$

이렇게 하면 바이오시밀러의약품의 치료 효과가 위약 대비 기준의약품의 치료 효과의 x %까지 갖고 있을 경우, 즉 기준의약품과 바이오시밀러의약품의 치료 효과 차이가 |(100-x)| %보다 작을 경우에 동등하다고 받아 들일 수 있다는 의미가 돼 동등성한계의 정의에도 잘 부합한다. 예를 들어, x가 각각 90 %, 120 %인 경우를 생각해 보자. 전자는 바이오시밀러의약품이 기준의약품보다 |100-90| %, 즉 10 % 작은 치료 효과를 갖고 있는 경우이며, 후자는 바이오시밀러의약품이 기준의약품보다 |100-120| %, 즉 20 % 큰 치료 효과를 갖게 되는 경우이다. 이들의 예에서 동등성한계 영역은 각각 다음과 같이 주어진다.

$$[-0.1\times M1,\ 0.1\times M1] \cdots\cdots (2)$$
$$[-0.2\times M1,\ 0.2\times M1] \cdots\cdots (3)$$

이상의 논의로부터, 동등성한계를 설정하려면 두 단계를 거쳐야 한다는 것을 알 수 있다. 첫 단계는 M1, 즉 위약 대비 기준의약품의 치료 효과를 추정하는 것이다. 이 단계에서는 과거에 실시된 기준의약품의 위약 대조 연구들을 조사하고, 각 연구에서 관찰된 치료 효과를 적절한 방법을 사용해 통합(pooling)한 뒤, 동등성시험에서 관찰될 가능성이 높은 M1을 신뢰성 있게 추정해 내야 한다. 과거 연구의 결과들로부터 통합된 치료 효과를 추정할 때에는 보통 메타분석법을 사용하며, 통합된 치료 효과는 점추정치와 신뢰구간으로 표시할 수 있다.

이 때 효과 측정 변수가 긍정적인 임상 결과이면-예를 들어, 치료 또는 반응- 신뢰구간의 하한값을, 그리고 부정적인 임상 결과이면-예를 들어, 사망이나 입원, 심근 경색 등과 같은 심각한 임상 성과(clinical outcomes)에 대한 위험(risk)- 신뢰구간의 상한값을 M1으로 정한다. 예를 들어, 반응률로 유효성을 평가한 여러 연구를 통합해 분석한 결과 위

약 대비 기준의약품의 치료 효과-즉 반응률의 차이-에 대한 점추정치는 40 percentage point, 그리고 95 % 신뢰구간은 [25, 55] percentage point와 같이 나왔다고 가정하자. 이 예에서 M1은 40 percentage point가 아니라, 25 percentage point가 돼야 한다. 왜냐 하면, 비록 점추정치인 40 percentage point가 참값이라고 하더라도 확률 변수의 특성상 동일한 동등성시험을 반복해서 실시한다고 가정할 때 이 중 절반에서는 기준의약품의 치료 효과가 40 percentage point보다 '낮은' 값을 갖게 될 것이기 때문이다. 그런데 만일 추정된 위약 대비 기준의약품의 치료 효과보다 낮은 효과가 동등성시험에서 나타났다면, assay sensitivity가 만족되지 못한 것이다. 어떤 동등성시험에서든 assay sensitivity가 만족되지 못 했다면 비록 바이오시밀러의약품이 기준의약품과 동등한 것처럼 평가됐다고 하더라도 그것은 아무런 의미가 없다. 요컨대 동등성시험에서 기준의약품의 치료 효과가 제대로 나타났다는 것이 전제돼야만 모든 동등성 관련 논의가 성립되기 때문이다.

따라서 신뢰구간의 하한값을 M1으로 선택함으로써, 동등성시험에서 위약 대비 기준의약품의 효과가 적어도 신뢰구간의 하한값 이상으로 나타났다는 것을 신뢰구간의 신뢰도-앞의 예에서는 95 %-만큼 담보할 수 있는 것이다. 결국, 이러한 논리는 가능한 동등성한계를 보수적으로 설정해야 한다는 대원칙에 부합한다. 위약 대비 기준의약품의 치료 효과를 작은 값으로 정함으로써 assay sensitivity를 확보해야만 동등하지 않은 바이오시밀러의약품을 기준의약품과 동등하다고 잘못 판단하는 위양성오류를 줄일 수 있기 때문이다.

첫 단계에서 M1이 정해졌다면, 그 다음에는 과연 M1의 몇 % 정도보다 작은 범위 내에서 기준의약품과 바이오시밀러의약품의 치료 효과가 차이가 나면 동등할 것으로 받아들일지, 즉 M2를 결정해야 한다. 따라서 M2는 M1보다 큰 값을 가질 수 없다는 제한 조건이 있다. 하지만 구체적으로 M2가 M1의 몇 % 정도가 되어야 할지는 질병의 특성, 기준의약품의 치료 효과, 바이오시밀러의약품의 기대 효과 등을 종합적으로 고려해 결정을 내려야 한다. 결국 M2의 결정은 임상적 판단에 많이 의존할 수 밖에 없다.

심혈관계질환의 경우, 일반적으로 M2는 M1의 50 %로 결정한다.[8] 그러나 만일 위약 대비 기준의약품의 치료 효과가 좋은 질병의 경우에는 50 %보다 훨씬 낮은 값으로 M2를 결정하는 것이 타당하다. 예를 들어, 급성감염에서 기준의약품인 항생제의 완치율이 90 %이고 위약은 10 %였다고 가정해 보자. 완치율 차이를 비교지표로 사용할 경우 M1은 80 %이므로 M2를 M1의 절반인 40 %로 정하는 것을 고려해 볼 수 있다. 하지만 질병이 그다지 위중하지 않으며 치료에 잘 반응한다는 급성감염 및 항생제의 특성을 고려할 때, 간신히 기준의약품의 절반에 달하는 치료 효과만을 갖는 바이오시밀러의약품이 기준의약품과 동등하다고 받아 들일 수 있을지 임상적으로 의문스럽다. 이처럼 기준의약품의 치료 효과가 큰 질병에서는 절대값 기준으로 10 - 15 %를 M2로 하는 게 일반적이다.[8] 실제로 2002년부터 2009년까지 비열등성시험의 결과에 의해 FDA로부터 허가를 받은 18개의 의약품에 대해 비열등성 한계를 조사한 미 정부의 보고서에 따르면, 대부분의 의약품은 항생제 또는 항균제였으며, 비열등성 한계가 치료율 절대값 기준으로 10 - 20 %에 달했다.[11]

이처럼 두 단계에 걸쳐 동등성한계를 정하는 방법을 '고정한계법(Fixed Margin Approach)'이라고 한다. 이 외에도 '합성법(Synthesis Approach)'을 사

용할 수 있다. 그러나, 고정한계법이 훨씬 직관적이고 계산이 간단하며, 무엇보다 합성법보다 더 보수적인 추정 방법이기 때문에, FDA는 이 방법을 권고하고 있다.[8)]

Assay Sensitivity

전 항에서 살펴 본 것처럼, 동등성한계를 설정할 때에는 먼저 M1을 추정하고, 이어 M1의 부분값으로 M2를 결정한다. 그런데 얼핏, 임상적 판단이 개입된 M2의 결정이 쉽지 않을 것처럼 보인다. 하지만 동등성시험의 타당성과 엄밀성을 보장하기 위해 더 중요한 것은 제대로 된 M1의 추정이다.

앞에서도 잠깐 언급한 것처럼, assay sensitivity는 기준의약품의 정상적인 치료 효과가 동등성시험에서 제대로 발현될 가능성·정도·능력을 의미한다. 다시 말해, 만일 위약 대조군이 동등성시험에 포함됐다면, 기준의약품과 위약 사이에 M1만큼 치료 효과의 차이가 나는 것을 적절한 신뢰도로 찾아낼 수 있어야만 'assay sensitivity가 확보됐다'고 주장할 수 있다.

Assay sensitivity가 중요한 이유는 위약 대비 기준의약품의 치료 효과를 동등성시험에서 직접 측정할 수 없는 반면, 바이오시밀러의약품의 효과는 기준의약품의 치료 효과에 비추어 간접적으로 추정해야 하기 때문이다. 따라서 과거의 연구 결과들로부터 추정한 기준의약품의 치료 효과가 동등성시험에서 여과 없이 발현됐다는 사실을 보장받는 것이 중요하다. 설령 기준의약품과 바이오시밀러의약품 사이에 동등하다고 받아 들일 수 없는 치료 효과의 차이가 존재하더라도, 기준의약품의 assay sensitivity가 적절한 신뢰도로 보장되지 않은 동등성시험에서는 이러한 치료 효과의 차이를 제대로 찾아 낼 수가 없게 된다. 거꾸로, 아무리 동등성시험에서 기준의약품의 치료 효과가 원래 의도했던 대로 잘 발현됐다고 하더라도, M1을 터무니없이 크게 추정했다면 assay sensitivity는 상대적으로 감소하게 된다. 어떤 경우든, assay sensitivity가 확보되지 않으면, 동등하지 않은 바이오시밀러의약품을 동등한 것으로 간주하는 위양성오류가 커지게 된다. 그리고 이것은 보수적으로 동등성 기준을 적용해야 한다는 대원칙에 위배된다.

과학적으로 엄밀한 방법을 사용해 M1을 추정하고, 동등성시험에서 assay sensitivity를 제대로 확보하기 위해서는 다음 세 가지 조건이 필요하다.[8)] 첫째, 위약과 비교해 우월한 기준의약품의 치료 효과가 과거에 실시된 연구들에서 일관되고 규칙적으로 나타났어야 한다. 이래야만 재현성이 있는 M1의 추정이 가능해진다. 증상을 완화시키는 것이 주 치료 목적인 상황에서는 위약 대비 기준의약품의 우월한 치료 효과가 들쑥날쑥한 게 일반적이다. 따라서 이런 경우에는 비열등성 또는 동등성 시험 자체가 좋은 연구 방법이 아닐 수 있다. 이러한 인식을 반영하듯, FDA는 전신성홍반성낭창(systemic lupus erythematosus)이나 루프스신염(lupus nephritis) 같은 질환은 비열등성 시험으로 새로운 의약품의 유효성을 입증하기가 어렵다고 판단하고 있다.[11)] 우울증 등의 신경정신과 질환들도 동등성시험이 적절하지 않은 경우에 해당한다.

또, 각 연구마다 똑 같은 치료 효과를 관찰할 수는 없으므로 이러한 연구 간의 변이가 M1의 추정에 고려돼야 한다. 메타분석법은 이러한 상황에 적절한 분석 방법인데, 다음 항에서 구체적으로 M1을 추정할 때 여기에 대해서 좀 더 자세히 기술하겠다.

Table 1. Summary of the efficacy of the reference product vs. placebo (hypothetical data)

Study	Study A (1999)			Study B (2008)		
Treatment	Reference	Placebo	Diff.	Reference	Placebo	Diff.
Response rate	42/83 (=50.0 %*)	17/84 (=20.0 %)	30.0	98/165 (=59.4 %*)	46/110 (=41.8 %)	17.6

Diff. is the difference (percentage points) in the proportions of patients meeting the response criteria between the reference product and placebo (i.e., reference - placebo). *: $p<0.01$ vs. placebo.

둘째, 과거에 기준의약품의 우월성을 입증한 위약 대조 연구의 디자인 특성이 동등성연구의 디자인 특성과 충분히 유사해야 하는데, 이것을 '항구성 가정(constancy assumption)'이라고 부른다. 예를 들어, 대상 환자군, 포함 및 제외 기준, 병용요법, 유효성 평가 변수의 측정 방법, 용량, 용법, 자료분석법 등이 모두 중요한 디자인 특성에 해당한다. 한 마디로 과거 연구에서 발현된 기준의약품의 치료 효과가 동등성시험에서도 그대로 나타나려면(assay sensitivity), 연구를 실시한 조건이 서로 유사해야 한다는 것이다. 그래서 만일 이러한 항구성 가정이 의심되는 경우에는 소위 '할인(discounting)'이라고 해서 추정된 M1의 값을 줄이기도 -예를 들어, 25 % 또는 50 % 만큼- 한다. 요컨대 assay sensitivity가 충분히 확보되지 않았기 때문에 M1을 보수적으로 조정함으로써 동등성 입증이 더 까다로운 방향으로 보정하는 것이다.

마지막으로, 동등성시험의 질적 수준이 높아야 한다. 일반적으로 어떤 의약품이 대조약 또는 위약과 비교해 더 낫다는 것을 보이려는 우월성 연구의 경우, 연구의 질적 수준이 떨어지면-프로토콜 위반, 낮은 순응도, 높은 중도탈락 등- 대조약(위약) 대비 해당 의약품의 우월성을 입증하는 것이 어려워진다. 이러한 것을 '귀무가설 쪽으로의 편의(bias toward the null)'라고 하는데, 연구의 질적 수준이 떨어지면 연구자가 의도했던 것과는 달리 귀무가설의 기각이 어려워진다는 뜻이다. 하지만 동등성시험에서는 낮은 질적 수준이 오히려 기준의약품의 치료 효과가 제대로 발현되는 것을 방해함으로써(즉 낮은 assay sensitivity), 바이오시밀러의약품과 마치 동등한 것처럼 보이게 만드는 위양성오류율을 높인다. 이러한 상황을 '대립가설 쪽으로의 편의(bias toward the alternative)'라고 부르며, 이것은 동등성 판정을 보수적으로 해야 하는 동등성시험의 대원칙에 어긋난다.

바이오시밀러의약품의 동등성한계 설정 사례

지금까지 논의한 원칙에 따라, 어떻게 하면 과학적으로 엄밀하고 임상적으로 타당한 방법을 사용해 바이오시밀러의약품의 동등성한계를 설정할 수 있는지 가상의 예를 통해 구체적으로 살펴 보자.

제일 먼저, 동등성시험과 유사한 연구대상(질환 및 환자), 방법론을 사용해, 동등성시험에서 투여된 바이오시밀러의약품과 동일한 용량의 기준의약품을 위약과 비교한 연구 논문들을 검색해야 한다. 그 결과, Table 1에 요약된 것처럼 A와 B라는 두 개의 연구 결과를 찾아 냈다고 가정하자. 연구 A와 B에서 기준의약품을 투여 받은 환자들의 가상 기저치 특성을 Table 2에 요약했으며, 비교를 위해 바이오시밀러의약품의 동등성시험에

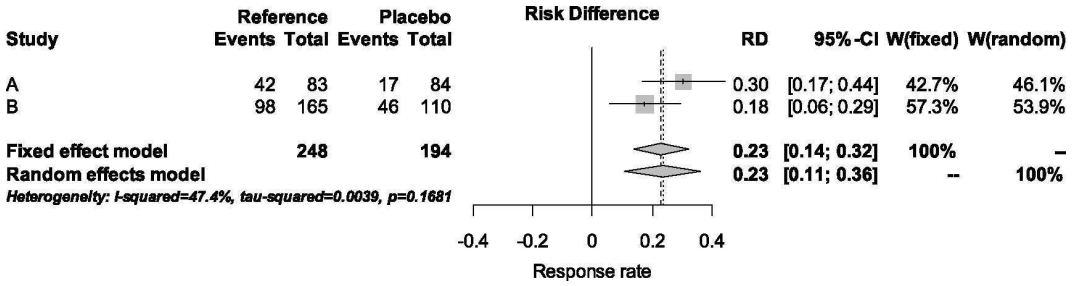

Figure 3. Forrest plot of the differences in the proportions of patients meeting the response criteria between the reference product and placebo (hypothetical data). 'Events' denotes the number of patients who met the response criteria in each treatment group. 'RD' represents risk difference, where risk means meeting the response criteria. 'W' is relative weight used to combine the results of different studies to determine the pooled estimate.

등재된 환자들의 가상 기저치 특성을 표의 마지막 열에 병기했다. Table 1과 2로부터 M1의 추정과 관련, 다음과 같은 논의를 이끌어 낼 수 있다.

우선, 기준의약품은 연구 A와 B 모두에서 위약 대비 유의한 반응률(response rate)의 차이를 나타냈다($p<0.01$, Table 1). 따라서 전항에서 M1을 추정할 때 assay sensitivity와 관련해 강조했던 첫 번째 조건, 즉 '위약과 비교해 우월한 기준의약품의 치료 효과가 과거에 실시된 연구들에서 일관되게 나타났어야 한다'는 조건이 일단 만족됨을 알 수 있다.

그러나 위약 대비 기준의약품의 치료 효과 크기는 연구마다 달랐다. 예를 들어, 연구 A에 비해 연구 B에서는 위약의 효과가 더 커서, 위약 대비 기준의약품의 치료 효과가 상대적으로 낮았다(17.6 vs. 30.0 percentage points, Table 1). 따라서 후발 제약사의 입장에서는 M1을 추정할 때 연구 B의 결과를 제외하는 것이 유리하다. 하지만 연구방법론이나 연구대상의 특성이 이들 두 연구 사이에 큰 차이가 없었다는 사실(Table 2)은 연구 B에서 관찰된 위약 효과의 증가가 확률 변수의 변동에 의한 것임을 의미한다.

이를 확인하는 하나의 방법은, 연구 A나 B와 유사한 환자군을 대상으로 기준의약품과 유사한 다른 치료제와 위약을 비교한 연구 논문에서 위약의 치료 효과를 살펴 보는 것이다. 그 결과, 위약군의 치료 효과가 30 - 40 % 가까웠거나 심지어 40 %가 넘었다면 연구 B에서 관찰된 위약 효과가 확률 변수의 변동에 의한 것이었을 가능성이 커진다. 또 다른 방법은, 다음의 Figure 3에도 요약돼 있지만, 연구 A와 B에서 위약 대비 기준의약품의 치료 효과에 대한 점추정치가 이들 연구를 통합해 추정한 M1의 신뢰구간 내에 모두 포함되는지 살펴 보는 것이다. 만일 어느 연구에서 얻어진 점추정치가 통합 추정치의 신뢰구간을 벗어난다면 해당 연구를 통합 추정에 포함시키는 것이 적절한지 잘 따져 보아야 한다.

그러나 위의 예에서는, 연구 B에서 상대적으로 크게 관찰된 위약의 효과가 확률 변수의 변동에 의한 것일 가능성이 크며, 치료 효과의 study-to-study 변이가 있어도, 연구 A와 B 모두 위약 대비 기준의약품의 치료 효과를 일관되게 보여 주고 있기 때문에 이 둘 중 어느 하나를 임의로 제외하고 M1을 추정하는 것은 옳지 않다. 이처럼

M1을 추정할 때, 임의로 어떤 연구의 결과를 제외하는 것이 타당하지 않음은 EMA나 FDA의 지침에서도 분명히 강조하고 있다. 왜냐 하면, 선택비뚤림(selection bias)의 문제가 발생하고,[12] 특히 연구 B처럼 상대적으로 기준의약품의 치료 효과가 낮게 나타난 연구를 제외하면 M1이 왜곡된 큰 값(falsely high)으로 추정돼, 결국은 보수적이어야 하는 동등성 평가의 원칙에도 위배되기 때문이다.[8]

한편 Table 2는 전항에서 assay sensitivity와 관련해 강조한 두 번째 조건, 즉 '항구성 가정'이 적절한지 판단하는 데 도움이 된다. 전술한 것처럼, M1을 추정하기 위해 사용된 두 편의 연구들은 치료 예후에 영향을 미칠 수 있는 여러 가지 기저치 특성에서 큰 차이를 보이지 않았다. 이러한 경우에는, 항구성 가정이 만족된 것으로 간주해 특별히 M1에 대한 할인을 고려하지 않아도 된다. 하지만 어떤 경우에는 연구가 실시됐던 시간에 많은 차이가 나서, 그 동안 의료 행태나 보조적인 치료 방법이 바뀌었을 수 있다. 이러한 경우에는 M1에 대한 할인이 필요하다. 즉 M1의 값을 25 % 또는 50 % 정도, 낮추어야 한다. 이는 동등성 판정을 보수적으로 할 수 있는 안전 장치라고 볼 수 있다.

위약 대비 기준의약품의 치료 효과, 즉 M1을 타당하게 추정하기 위해서는 이러한 연구들 사이의 '변이'를 적절한 방법으로 보정해야 한다. 메타분석법, 그 중에서도 random-effects analysis는 이처럼 연구들 사이의 변이를 고려해 전체적인 통합 치료 효과를 추정하는 데 적절한 방법으로 알려져 있다.[13-15] 통계프로그램인 R (version 2.15.0)에서 메타분석을 실시할 수 있는 *meta* 및 *rmeta* 라이브러리를 이용해 Table 1에 요약된 각 연구의 결과로부터 다음과 같이 M1을 추정했다. 먼저, 위약 대비 기준의약품의 치료 효과의 차이에 대한 통합 추정치를 Inverse Variance Method 방법을 사용해 구했고, 이 때 연구들 사이의 변이를 보정하기 위해 DerSimonian and Laird가 제안한 random-effects model을 사용했다.[14] Figure 3은 메타분석의 결과를 요약한 것이다. Figure 3의 결과로부터 random-effects model에 의해 얻어진 통합된 치료효과의 점추정치 및 95 % 신뢰구간은 23 [11, 36] percentage points가 된다.

따라서 만약 위약군이 동등성시험에 포함됐다면 위약 대비 기준의약품의 치료 효과가 제대로 나타났을 것이라는 기대(즉 assay sensitivity)가

Table 2. Baseline characteristics (hypothetical data)

Baseline characteristics	Study A (1999)	Study B (2008)	Equivalence trial C
Age (year)	56 (25 - 74)*	49.1±12.0	18 - 75
Female (%)	81	82.4	80
Variable 1	8.4 (0.7 - 45)*	7.3±6.2	8.0±6.0
Variable 2	15±2.5	16.3±3.6	15.5±3.0
Variable 3	32±15	31.7±14.5	30.7±16
Variable 4	19±8.5	20.3±8.0	20.0±9.0
Variable 5	1.8±0.5	1.7±0.7	1.9±0.6
Variable 6	3.1±2.0	3.3±3.2	3.2±3.0

Data are shown in median ± semi-interquartile range (study A) or mean ± standard deviation (study B and equivalence trial C). *: range.

Table 3. Equivalence margins* in various settings (sensitivity analysis)

Significance level for M1	Clinically acceptable proportion for M2	Study A only	Studies A & B
90 %	50 %	9.4	6.5
	70 %	5.6	3.9
95 %	50 %	8.3	5.5
	70 %	5.0	3.3
99 %	50 %	6.1	3.5
	70 %	3.7	2.1

* percentage point.

신뢰성을 갖추어 보장되기 위해 95 % 신뢰구간의 하한치인 11 percentage points를 M1으로 정해야 한다. 또한, 앞에서 기술한 것처럼, 본 예에서는 항구성 가정이 충족돼 추가적인 M1의 할인은 필요 없는 것으로 간주했다. 동시에 동등성시험이 연구계획서대로 잘 실시돼 질적 수준도 만족된 것으로 가정하자.

다음 단계에서는, 위약 대비 기준의약품의 치료효과에서 과연 어느 정도를 바이오시밀러의약품이 유지해야 임상적으로 동등한 것으로 간주할 수 있는지, 즉 M2를 결정해야 한다. 여기에서는 심혈관계질환 치료제에 일반적으로 적용되는 50 %를 사용하도록 하자. 사실 이것은 임상가의 입장에서는 매우 느슨한 기준일 수도 있다. 왜냐 하면, 기준의약품의 치료 효과의 50 %만을 갖고 있는 바이오시밀러의약품을 동등하다고 하는 것이 임상적으로는 지나치게 허용적이기 때문이다. 어쨌든 50 % 기준을 적용하면, M2는 11 percentage points의 절반인 5.5 percentage points가 될 것이다. 즉 동등성한계는 ± 5.5 또는 [-5.5, 5.5] percentage points가 된다.

M1을 추정할 때 특정 연구-위에 예에서 연구 B-를 임의로 제외하는 것은 선택비뚤림 등의 문제를 일으키기 때문에 타당한 방법이 아님은 이미 앞에서 설명한 바 있다. 그런데 만일 후발 제약사가 이를 고집하면 반드시 이에 대한 타당한 근거를 제시해야 한다. 만일 연구 B를 M1 추정에서 제외시킨다면 연구 A 하나만 남게 된다. 그런데 이처럼 단 하나의 연구 결과만을 이용해 M1을 추정하는 경우, study-to-study variability를 평가할 수 있는 방법이 없기 때문에 신뢰도의 수준을 95 %가 아닌 99 %로 올려야 한다고 FDA의 지침은 강조하고 있다.[8] 요컨대 연구들 사이의 변이에 대한 불확실성이 증가한 것에 대해 신뢰도 수준을 증가시켜 보수적인 방법으로 동등성을 평가함으로써 균형을 맞추어야 한다는 것이다.

따라서 M1을 추정할 때 사용하는 신뢰도(90 %, 95 % 및 99 %), M2를 설정할 때 기준의약품 대비 바이오시밀러의약품이 유지해야만, 임상적으로 동등하다고 받아 들일 수 있는 치료 효과의 분율(50 % 및 70 %), 그리고 M1 추정에 사용하는 과거 연구(연구 A만 또는 연구 A와 B 모두) 등, 각각의 조합에 따라 동등성한계가 어떻게 달라지는지 sensitivity analysis를 실시할 수 있다 (Table 3).

연구 A의 결과만을 이용해 M1을 추정할 때에는 95 % 대신 99 % 신뢰구간이 권장되므로, 기준의약품 대비 바이오시밀러의약품의 치료 효과의 비가 50 %는 돼야 동등하다고 인정할 수 있으려면, Table 3으로부터 동등성한계는 6.1

percentage points가 된다. 이 값은 앞에서 구한 5.5 percentage points보다는 약간 크지만 대체로 비슷한 값이다. 한편 후발 제약사에게 가장 유리한 조합은 M1 추정에 대한 신뢰도가 90 %, M2를 설정할 때 임상적 효과가 기준의약품의 50 %를 유지, 그리고 M1을 추정할 때 연구 A의 결과만을 이용하는 경우라고 할 수 있다. 이 경우 동등성한계는 9.4 percentage points가 된다. 물론, 이러한 조합은 과학적 엄밀성과 임상적 타당성을 만족시키지 못 한다.

다만, 동등성한계가 작아지면 동등성시험에 필요한 환자 수가 크게 증가하므로, 임상시험 수행의 현실적인 면을 고려할 필요가 있다. 이러한 관점에서, M1 추정의 신뢰도를 95 %로 하고, M2를 M1의 50 %로 하되, 각각 연구 A로부터 또는 연구 A와 B로부터 추정한 동등성한계인 8.3과 5.5의 평균값인 6.9 또는 이의 근사값인 7.0 percentage points를 최종 동등성한계로 정하는 것이 현실적인 절충안이 될 수는 있다.

요약 및 결론

본고는 바이오시밀러의약품의 허가와 관련, 동등성한계를 설정하는 근간이 되는 과학적, 임상적, 규제과학적 원리들을 소개했다. 또한 이 과정에서 다루어야 하는 중요한 쟁점들을 검토했고, 가상자료를 이용해 바이오시밀러의약품의 동등성한계를 설정하는 과정을 예시했다. 모든 동등성시험은 사전에 어떤 자료에 근거해, 그리고 어떤 과정을 거쳐 과학적으로 엄밀하고, 임상적으로 타당하게 동등성한계를 설정했는지 면밀히 검토돼야 한다. 이것은 동등성 평가를 보수적으로 운영함으로써, 소비자에게 가해질 위험(consumer risk), 즉 동등하지 않은 바이오시밀러의약품을 기준의약품과 동등한 것으로 잘못 간주하는 오류를 타당한 수준 이하로 통제해야 할 필요가 있는 규제기관에게 중요한 의미를 갖는다.

참고문헌

1. McCamish M, Woollett G. The state of the art in the development of biosimilars. *Clin Pharmacol Ther*, 2012;91(3):405-417.
2. Mellstedt H, Niederwieser D, Ludwig H. The challenge of biosimilars. *Ann Oncol*, 2008;19(3):411-419.
3. Njue C. Statistical considerations for confirmatory clinical trials for similar biotherapeutic products. *Biologicals*, 2011;39(5):266-269.
4. 식품의약품안전청, 식품의약품안전평가원. 동등생물의약품 평가 가이드라인. Seoul (Korea) 2009.
5. World Health Organization. Guidelines on evaluation of similar biotherapeutic products (SBPs). 2012.
6. Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER). Guidance for industry: Scientific considerations in demonstrating biosimilarity to a reference product (Draft). 2012.
7. European Medicines Agency. Guideline on similar biological medicinal products containing monoclonal antibodies (EMA/CHMP/BMWP/403543/2010). http://wwwemaeuropaeu/docs/en_GB/document_library/Scientific_guideline/2010/11/WC500099361pdf. 10/2010, 2010. [Online] (last visited on Feb 13 2012).
8. Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER). Guidance for Industry: Non-inferiority clinical trials (Draft). March 2010.
9. Greene CJ, Morland LA, Durkalski VL, Frueh BC. Noninferiority and equivalence designs:

issues and implications for mental health research. *J Traum Stress*, 2008;21(5):433-439.
10. Hwang IK, Morikawa T. Design Issues in Noninferiority/Equivalence Trials. *Drug Inf J*, 1999;33(4):1205-1218.
11. United States Government Accountability Office. New drug approval: FDA's consideration of evidence from certain clinical trials. Washington, DC, USA: United States Government Accountability Office. 2010.
12. Committee for Medicinal Products for Human Use (CHMP). Guideline on the Choice of the Non-inferiority margin. EMEA/CPMP/EWP/2158/99. London (UK) July 27, 2005.
13. Higgins JP, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *J R Stat Soc Ser A*, 2009;172(1):137-159.
14. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials*, 1986;7(3):177-188.
15. The Cochrane Collaboration. 9.5.4 Incorporating heterogeneity into random-effects models. In: Higgins J, Green S, eds. *Cochrane Handbook for Systematic Reviews of Interventions 5.1.0* [updated March 2011] 2011.