The Korean Journal of Applied Statistics (2012) DOI: http://dx.doi.org/10.5351/KJAS.2012.25.4.575
25(4), 575–587

Statistical Methods in Non-Inferiority Trials - A Focus on US FDA Guidelines -

Seung-Ho Kang[1] · So-Young Wang[2]

[1]Department of Applied Statistics, Yonsei University
[2]Department of Statistics, Ewha Women's University

(Received April 26, 2012; Revised May 16, 2012; Accepted June 18, 2012)

Abstract

The effect of a new treatment is proven through the comparison of a new treatment with placebo; however, the number of parent non-inferiority trials tends to grow proportionally to the number of active controls. In a non-inferiority trial a new treatment is approved by proof that the new treatment is not inferior to an active control; however, both additional assumptions and historical trials are needed to show (through the comparison of the new treatment with the active control in a non-inferiority trial) that the new treatment is more efficacious than a putative placebo. The two different methods of using the historical data: frequentist principle method and meta-analytic method. This paper discusses the statistical methods and different Type I error rates obtained through the different methods employed.

Keywords: Assay sensitivity, constancy assumption, Type I error.

1. 서론

최근 시험약과 활성대조약(active control)을 투여받는 두 그룹만이 존재하는 비열등성 임상시험(non-inferiority clinical trial)이 신약 개발에 매우 중요한 역할을 차지하고 있다 (D'Agostino 등, 2003; Fleming, 2008; Hung 등, 2003; Hung 등, 2007; Hung 등, 2009; Tsong 등, 2003; Tsong, 2007; Wang과 Hung, 2003). 활성대조약이란 과거의 임상시험을 통하여 위약(placebo)보다 효과가 우월함이 명확히 입증되어 시판허가된 약물을 의미한다. 마찬가지로 시험약의 효과를 입증하여 시판허가를 받기 위해서는 질병이나 증상을 치료하는데 있어 시험약이 위약에 비하여 우월한 효과가 있음을 증명하여야 한다.

그러나 이미 활성대조약이 개발되어 널리 사용되고 있는 상황에서 활성대조약을 사용하지 않고 위약을 사용하는 임상시험을 실시할 경우, 질환이 있는 환자에게 치료제가 아닌 위약이 투여될 수 있으며, 이로 인해 질병이 악화될 수 있어 비윤리적인 측면이 있다. 이러한 사유로 위약 대신 기존에 시판되고 있는 활성대조약을 사용하여, 활성대조약에 비해 시험약의 효과가 비열등함을 입증하는 비열등성 임상시험이 증가하고 있다.

This research was supported by the Basic Science Research Program of the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(No. 2010-0009224).
[1]Corresponding author: Professor, Department of Applied Statistics, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 120-749, Korea. E-mail: seungho@yonsei.ac.kr

시험약과 위약의 효과를 직접적으로 비교하는 우월성 임상시험과 달리, 비열등성 임상시험에는 위약을 투여 받는 위약군이 존재하지 않기 때문에, 시험약의 효과가 위약보다 우월함을 입증하기 위해서는 위약과 대조약의 효과를 비교한 과거의 임상시험 자료가 활용되며, 이러한 과정에는 많은 통계적 이슈가 발생하게 된다. 본 논문에서는 비열등성 임상시험에서의 주요 통계적 이슈들과 이에 관련된 최근의 논쟁들에 대해 고찰하고자 한다.

2. 비열등성 임상시험의 목적

우선 비열등성 임상시험의 목적을 명확하게 구별하는 것이 중요한다. 비열등성 임상시험의 목적은 크게 다음의 세 가지로 구분된다 (Tsong 등, 2003). 설명의 편의상 주평가변수에 대한 측정값이 클수록 더 좋은 효과를 나타낸다고 가정한다. 비열등성 임상시험에서 가장 흔히 설정되는 가설은 다음과 같다.

$$H_0 : \mu_T - \mu_C \leq -\delta \quad vs. \quad H_1 : \mu_T - \mu_C > -\delta, \tag{2.1}$$

여기서 μ_T는 주평가변수에 대한 시험약의 모평균, μ_C는 활성대조약의 모평균, $\delta\,(>0)$는 비열등성 마진을 의미하며, 질병이나 증상에 따라 다른 값이 사용된다. 비열등성 임상시험의 첫 번째 목적과 두 번째 목적은 가설 (2.1)을 검정하는 것과 연관되어 있다.

비열등성 임상시험의 첫 번째 목적은 가설 (2.1)을 검정함으로써 시험약의 효과가 위약의 효과보다 우월함을 보이는 것으로, 시험약의 시판허가를 목적으로 실시하는 대다수의 임상시험이 여기에 해당된다 (Hasselblad과 Kong, 2001). 다시 말해, 첫 번째 목적에서 검정하고 싶은 가설을 보다 명확히 기술하면 다음과 같다.

$$H_0 : \mu_T \leq \mu_{P,put} \quad vs. \quad H_1 : \mu_T > \mu_{P,put}, \tag{2.2}$$

여기서 $\mu_{P,put}$는 현재의 비열등성 임상시험에서 위약을 투여 받는 환자들이 존재한다고 가정시, 가상의 위약군 환자들에 대한 주평가변수의 모평균이다. 하지만 실제 현재 비열등성 임상시험에서는 위약을 받는 그룹이 존재하지 않으므로, $\mu_{P,put}$은 가상의 값이다. 아래첨자 put는 실제로는 존재하지 않는 가상적(putative)이라는 것을 표시하기 위하여 사용되었다. $\mu_{P,put}$는 현재의 비열등성 임상시험 자료만으로는 추정이 불가능한 모수이므로, 가설 (2.2)는 엄밀한 의미에서 검정이 불가능한 가설이다. 다시 말해 비열등성 임상시험의 첫 번째 목적은 위약군은 존재하지 않고 시험약과 활성대조약을 투여 받는 두 그룹만이 존재하는 임상시험에서, 가설 (2.1)을 검정함으로써 가설 (2.2)를 검정하여 시험약이 위약보다 우월함을 보이고자 한다는 것이다. 이런 목적을 달성하기 위해서는, 시험약과 활성대조약을 비교하는 현재의 비열등성 임상시험 자료와 활성대조약과 위약을 비교한 과거의 임상시험 자료를 합쳐 시험약과 위약을 비교하는 통계적 추론을 하게 된다. 여기에는 assay sensitivity와 constancy assumption과 같은 가정이 필요하게 되는데, 이 가정들에 대해서는 다음 절에 더 상세하게 살펴볼 것이다.

두 번째 목적은 시험약의 효과가 활성대조약의 효과보다 $\delta\,(>0)$ 범위 안에서 비열등함을 보이는 것으로, 가설은 (2.1)처럼 주어진다 (Blackwelder, 1982). 이러한 비열등성 임상시험의 목적은 가상의 위약 효과인 $\mu_{P,put}$와 관련이 없어, 위약 대비 시험약의 우월성 입증을 요구하는 시판허가용 임상시험자료의 기본적 요건을 충족하지 못하게 된다. 두 번째 목적에 해당하는 임상시험의 예로 이미 시판허가를 받은 두 약의 효과는 거의 유사(비열등)하나, 대조약에 비해 시험약의 독성이 감소하였거나, 복용의 편리성이 증대되었거나, 비용이 감소되는 점 등을 평가하는 4상 임상시험을 들 수 있다.

세 번째 목적은 시험약의 효과가 위약보다 우월할 뿐만 아니라, 시험약과 위약의 효과 차이가 활성대조약과 위약의 효과 차이의 일정 비율 (예를 들면, $100\lambda\%$, $0 \leq \lambda \leq 1$) 보다 크다는 것을 보이는 것이다

(Holmgren, 1999; Wang과 Hung, 2002). 이러한 목적은 활성대조약이 위약보다 월등하게 효과가 좋은 경우에 적용할 수 있다. 단순히 시험약이 위약보다 효과가 우월한 것만으로는 시험약의 시판허가를 승인하기에 충분하지 않을 수 있다. 예를 들어 어떤 임의의 단위를 사용하여 활성대조약의 효과는 80, 위약의 효과는 20이라고 가정해보자. 시험약이 시판허가를 받기 위한 최소한의 요건은 위약에 비하여 우월함을 증명하는 것이므로, 시험약은 위약의 효과인 20보다만 크면 시판허가를 받을 수 있다 (단, 위약보다 우월한 시험약의 약효 크기가 임상적으로 유의하다고 가정하자). 하지만 활성대조약의 효과가 매우 좋은 경우, 심사기관은 경우에 따라 시험약의 효과가 단지 위약에 비해 우월한 것 뿐만 아니라, 위약 대비 시험약의 효과가 활성대조약과 위약의 효과 차이의 일정 비율 (예를 들면, $100\lambda\%$, $0 \le \lambda \le 1$) 보다 더 우월한 효과의 입증을 요구할 수도 있다. 예를 들어 심사기관으로부터 시험약의 효과가 대조약 효과의 50% ($\lambda = 0.5$)를 초과함을 입증할 것을 요구받았다고 가정하자. 위약 대비 활성대조약 효과 크기의 50%는 30 ($0.5 \times (80 - 20) = 30$)이므로, 시험약의 효과가 50 ($20 + 30$)보다 커야만, 위약 대비 시험약의 효과가 위약 대비 활성대조약 효과의 50%보다 크다는 조건을 만족하게 된다. 이와 같은 목적을 기술한 가설은 다음과 같이 주어진다.

$$H_0 : (\mu_T - \mu_{P,put}) \le \lambda(\mu_C - \mu_{P,put}) \quad vs. \quad H_1 : (\mu_T - \mu_{P,put}) > \lambda(\mu_C - \mu_{P,put}). \tag{2.3}$$

가설 (2.3)에 주어진 대립가설의 좌변은 위약 대비 시험약의 효과를 나타내고, 우변은 위약 대비 활성대조약의 효과를 나타낸다. 여기서 λ는 0과 1 사이의 미리 지정된 상수이다. 만일 $\lambda = 0$이면 대립가설은 시험약이 가상의 위약보다 효과가 우월함 ($\mu_T > \mu_{P,put}$)을 나타내고, 만일 $\lambda = 1$이면 대립가설은 시험약이 활성대조약보다 효과가 우월함 ($\mu_T > \mu_C$)을 나타낸다. 가설 (2.3)은 다음과 같이 정리된다.

$$H_0 : (\mu_T - \mu_C) \le (\lambda - 1)(\mu_C - \mu_{P,put}) \quad vs. \quad H_1 : (\mu_T - \mu_C) > (\lambda - 1)(\mu_C - \mu_{P,put}). \tag{2.4}$$

또한 가설 (2.3)의 대립가설이 $\lambda \ge 0$인 경우에 증명된다면, $\mu_C - \mu_{P,put} > 0$이므로

$$(\mu_T - \mu_{P,put}) > \lambda(\mu_C - \mu_{P,put}) \ge 0 \Rightarrow \mu_T > \mu_{P,put}$$

이 되어, 시험약이 가상의 위약보다 효과가 우월함을 증명할 수 있다. 그러므로 세 번째 목적은 첫 번째 목적을 포함하는 더 일반적인 목적으로 간주할 수 있다.

3. Assay Sensitivity와 Constancy Assumption

2절에서 살펴본 비열등성 임상시험의 여러 목적을 수행하는데 있어 assay sensitivity와 constancy assumption 가정이 필요하게 되며, 3절에서는 이러한 사항에 대해 설명하고자 한다.

활성대조약이란 과거의 임상시험을 통해 위약보다 효과가 우월함이 입증된 약이다. 하지만 이렇게 과거의 임상시험을 통해 활성대조약이 위약보다 효과가 우월함을 보였다고 하더라도, 만일 현재의 임상시험에서 그 활성대조약과 위약을 다시 비교하는 임상시험을 수행해보면 활성대조약이 위약보다 효과가 우월함을 다시 증명하지 못할 수 있으며, 실제 그러한 사례도 있다 (Leber, 1989; Temple과 Ellenberg, 2000). 이러한 일이 발생하는 이유는 여러 가지인데, 임상시험에 참여하는 피험자들이 모집단을 잘 대표하는 임의표본(random sample)이 아닐 수 있으며, 현재와 과거의 두 임상시험에서 피험자 관리, 주평가변수의 평가방법 등이 동일하지 않을 수 있기 때문이다. 그 외에도 위약의 효과가 매우 크고 변동이 심하거나, 활성대조약의 크기가 아주 작으며 변동이 심한 경우에도 대조약의 효과가 재현되지 않을 수 있다.

만일 현재의 비열등성 임상시험에서 활성대조약의 효과가 위약과 같다면, 시험약의 효과가 활성대조약보다 비열등하다고 해도 (즉 거의 비슷하다고 해도), 시험약의 효과가 위약의 효과보다 우월함을 증명

하지 못하게 되어, 2절에서 설명한 비열등성 임상시험의 첫 번째 목적을 만족시키지 못하게 된다. 그러므로 과거의 임상시험에서 활성대조약의 효과가 위약보다 우월했던 것처럼, 현재의 비열등성 임상시험에서도 (비록 위약군이 존재하지는 않지만) 활성대조약의 효과가 위약보다 우월함을 확인하는 것이 중요한데, 이러한 개념을 나타내주는 것이 바로 assay sensitivity이다. 비열등성 임상시험에서 assay sensitivity를 수식으로 표현하면 $\mu_C > \mu_{P,put}$이 된다. 하지만 현재의 비열등성 임상시험에는 위약군이 존재하지 않아 $\mu_{P,put}$를 추정할 수 없으므로 assay sensitivity를 직접적으로 확인해 볼 수 있는 방법은 없다. 하지만 유사하게 설계된 임상시험에서 활성대조약이 위약보다 항상 우월함을 보인 다수의 과거 임상시험자료, 현재의 비열등성 임상시험과 과거의 임상시험의 유사성, 그리고 현재의 비열등성 임상시험이 높은 수준으로 관리되었는지 여부 등을 종합적으로 고려하여 assay sensitivity를 간접적으로 확인해볼 수 있다 (Kang, 2010, Chapter 9; ICH E10, 2001; US FDA, 2010).

Constancy assumption이란 위약 대비 활성대조약의 효과 크기가 현재의 비열등성 임상시험에서와 과거의 임상시험에서 동일하다는 것이다 (Fleming, 2008). $\mu_{C|H}$와 $\mu_{P|H}$를 과거의 임상시험에서 활성대조약과 위약을 투여 받은 피험자들의 주평가변수의 모평균이라고 하면, constancy assumption의 의미를 다음의 수식으로 표현할 수 있다.

$$\mu_C - \mu_{P,put} = \mu_{C|H} - \mu_{P|H}. \tag{3.1}$$

식 (3.1)의 왼쪽은 현재의 비열등성 임상시험에서 (가상의) 위약 대비 활성대조약의 효과 크기이고, 식 (3.1)의 오른쪽은 과거의 임상시험에서 위약 대비 활성대조약의 효과 크기이다. Constancy assumption에 대한 보다 자세한 내용은 Kang (2010)의 9장에 설명되어 있다.

4. 비열등성 임상시험의 통계검정시 필요한 기호들의 소개

비열등성 임상시험을 분석하는데 사용되는 통계방법은 fixed margin method, synthesis method, two confidence interval method가 있다 (Hung 등, 2009; Tsong, 2007). 이 방법들은 과거 임상시험 자료를 활용하는 방법으로서 빈도론적인 원칙(frequentist principle)을 사용하는 경우와 메타분석 형태의 방법을 사용하는 경우로 나누어질 수 있다. 이 세 가지 방법들에 대해서는 5–7절에서 자세하게 살펴볼 것이다. 우선 논의에 필요한 기호들을 정의하면 다음과 같다. 여기서 정의되는 주평가변수들의 모평균은 높은 값일수록 더 좋은 효과를 나타낸다고 가정한다.

- μ_T : 현재의 비열등성 임상시험에서 시험약 T를 투여 받는 환자들의 주평가변수의 모평균.
- μ_C : 현재의 비열등성 임상시험에서 활성대조약 C를 투여 받는 환자들의 주평가변수의 모평균.
- $\mu_{P,put}$: 만일 현재의 비열등성 임상시험에서 위약 P_{put}를 투여 받는 환자들이 존재한다고 가정시, (가상의) 위약 을 투여 받는 환자들의 주평가변수의 모평균. 그러나 현재 비열등성 임상시험에서는 실제로 위약을 받는 환자군이 존재하지 않으므로, $\mu_{P,put}$은 가상의 값이다.
- $\mu_{C|H}$: 과거의 임상시험에서 활성대조약 C를 투여 받은 환자들의 주평가변수의 모평균.
- $\mu_{P|H}$: 과거의 임상시험에서 위약을 투여 받은 환자들의 주평가변수의 모평균.

여기서 주의할 점은 μ_C와 $\mu_{C|H}$가 다를 수 있고, $\mu_{P,put}$와 $\mu_{P|H}$가 다를 수 있다는 점이다. 그 이유는 현재의 비열등성 임상시험에 참여한 피험자 모집단이 과거의 임상시험에 참여한 피험자 모집단과 다를 수 있고, 그 외 병용용법이나 기타 두 임상시험의 특징이 다를 수 있기 때문이다.

자료의 형태를 표시할 기호들도 소개할 필요가 있다. 현재의 비열등성 임상시험에는 시험약을 투여 받는 그룹과 활성대조약을 투여 받는 그룹이 존재한다고 가정시, 시험약을 투여 받는 그룹에서 얻어

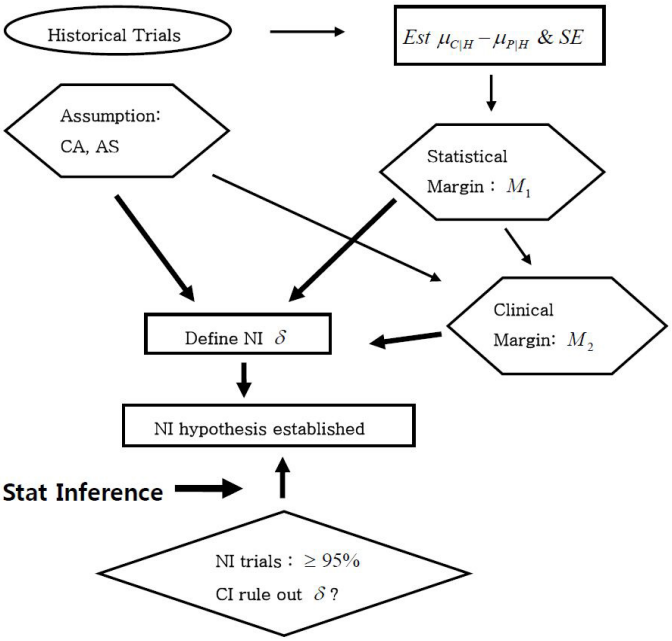

Figure 5.1. Fixed margin method

진 관측치를 X_{Ti}, $i = 1, 2, \ldots, n$으로, 활성대조약을 투여 받는 그룹에서 얻어진 관측치를 X_{Ci}, $i = 1, 2, \ldots, n$으로 표기한다. 설명의 편의상 두 그룹의 표본크기는 n으로 동일하다고 가정하고, $E(X_{Ti}) = \mu_T$, $E(X_{Ci}) = \mu_C$, $\mathrm{Var}(X_{Ti}) = \mathrm{Var}(X_{Ci}) = \sigma^2 < \infty$라고 가정한다.

과거의 임상시험에는 활성대조약과 위약을 투여 받는 그룹이 존재한다고 가정시, 활성대조약을 투여 받은 그룹에서 얻어진 관측치를 $X_{C|H_i}$, $i = 1, 2, \ldots, n_H$으로, 위약을 투여 받은 그룹에서 얻어진 관측치를 $X_{P|H_i}$, $i = 1, 2, \ldots, n_H$으로 표기한다. 설명의 편의상 두 그룹의 표본크기는 n_H으로 동일하다고 가정하고, $E(X_{C|H_i}) = \mu_{C|H}$, $E(X_{P|H_i}) = \mu_{P|H}$, $\mathrm{Var}(X_{C|H_i}) = \mathrm{Var}(X_{P|H_i}) = \sigma_H^2 < \infty$라고 가정한다. 또한 본 논문에서는 중심극한정리를 이용하기 위하여 표본크기 n과 n_H는 충분히 크다고 가정한다.

5. Fixed Margin Method

Fixed margin method의 가장 중요한 특징은 빈도론적인 원칙(frequentist principle)을 사용한다는 점이다 (Hung 등, 2009). 이것의 의미는 과거의 임상시험에서 얻어진 정보는 단지 상수값으로만 요약되며, 통계적 추론 과정에 포함되지 않는다는 것이다. Fixed margin method는 Figure 5.1에서 보듯이 두 단계로 이루어진다 (Hung 등, 2009). 첫 번째 단계에서는 과거의 임상시험 자료, assay sensitivity, constancy assumption, 그리고 임상적 경험을 이용하여 최종 비열등성 마진을 결정하게 된다. 비록 비열등성 마진을 결정하는데 과거의 임상시험 자료를 활용하나, fixed margin method는 빈도론적인 원칙을 따르는 방법이므로 가설검정을 위한 첫 번째 단계인 비열등성 마진을 정하는 단계는 통계적 추론의 대상에 포함되지 않는다. 두 번째 단계는 비열등성 마진이 정해진 후부터 시작된다. 통계적 추론은 이 두 번째 단계에서부터 시작되게 되고, 현재의 비열등성 임상시험에 있는 자료들만을 사용하여 통계적 추론을 하게 된다.

이러한 fixed margin method의 각 단계를 좀 더 자세하게 살펴보자. Fixed margin method는 미국 FDA 비열등성 임상시험 가이드라인에서 권고되는 방법이다 (US FDA, 2010).이 방법에서는 비열등성 마진 δ를 결정하는데, 통계적 마진 M_1과 임상적 마진 M_2를 모두 고려하여 결정하게 된다. 우선 통계적 마진은 현재의 비열등성 임상시험에서 가상의 위약 대비 활성대조약의 예측 효과 보다 크지 않게 설정하여야 한다 ($\mu_C - \mu_{P,put} \geq M_1$). 이에 대해 ICH E10은 다음처럼 기술하고 있다 (ICH E10, 2001).

"The margin chosen for a non-inferiority trial cannot be greater than the smallest effect size that the active drug would be reliably expected to have compared with placebo in the setting of the planned trial."

하지만 $\mu_{P,put}$는 알 수 없는 가상의 값이므로, constancy assumption인 $\mu_C - \mu_{P,put} = \mu_{C|H} - \mu_{P|H}$를 이용하여 $\mu_{C|H} - \mu_{P|H} \geq M_1$을 만족하도록 통계적 마진 M_1을 설정하게 된다. 활성대조약이 갖추어야할 조건 중의 하나는 유사하게 설계된 여러 과거의 임상시험에서 대조약의 효과가 비교적 일정했다는 증거이다. 이 조건이 바로 assay sensitivity인 것이다. 이러한 사유로 활성대조약이 위약에 비해 우월함을 보인 여러 개의 과거 임상시험 자료가 필요하게 된다. 여러 개의 과거 임상시험 자료를 사용하여 $\mu_{C|H} - \mu_{P|H} \geq M_1$를 만족하는 M_1의 추정치를 얻는 방법으로는 메타분석이 사용되는데, 문제는 fixed effect model을 쓰는 경우와 random effect model을 쓰는 경우에 따라 M_1의 추정치가 달라져 이에 대해 전문가들 사이에 명확한 의견 일치가 존재하지 않는다는 점이다.

어떤 모형을 사용하던 $\mu_{C|H} - \mu_{P|H} \geq M_1$을 만족하는 M_1의 추정치를 얻는 가장 쉬운 방법은 $\mu_{C|H} - \mu_{P|H}$의 점추정치를 사용하는 것이지만, 원래 목적이 $\mu_C - \mu_{P,put} \geq M_1$을 만족하는 M_1의 추정치를 얻는 것이고, $(\mu_C - \mu_{P,put}) = (\mu_{C|H} - \mu_{P|H}) \geq M_1$이라는 조건은 확인하기 어려운 두 가정인 assay sensitivity와 constancy assumption으로부터 얻어지는 것이기 때문에, 보수적인 방법을 사용하여 $\mu_{C|H} - \mu_{P|H}$의 95% 또는 99% 신뢰구간의 하한을 M_1으로 설정하게 된다.

통계적 마진 M_1을 정한 이후에는 임상적 마진 M_2를 정해야 하는데, 그 이유는 통계학적으로 시험약이 활성대조약에 비하여 M_1만큼 비열등하다는 것이 받아들여질 수 있다고 하더라도, 임상적으로는 받아들여질 수 없는 경우가 생길 수도 있기 때문이다. 임상적 마진 M_2은 통계학적인 고려없이 임상적 판단에 의해서만 결정된다. 미국 FDA 가이드라인에 의하면 이렇게 정한 통계적 마진 M_1과 임상적 마진 M_2 중 최소값으로 비열등성 마진(δ)을 설정할 것을 권고하고 있다 (US FDA, 2010). 즉 $\delta = \min(M_1, M_2)$이 된다. 또한 ICH E10에는 비열등성 마진을 구하는데 있어 통계적인 사고와 임상적인 판단 모두가 고려되어야 함이 기술되어 있다 (ICH E10, 2001).

"The determination of margin in a non-inferiority trial is based on both statistical reasoning and clinical judgement, should reflect uncertainties in the evidence on which the choice is based, and should be suitably conservative."

이렇게 구한 비열등성 마진(δ)을 상수로 간주할지 아니면 확률변수로 간주할지는 통계학적으로 매우 중요한 이슈이다. 그 이유는 어느 방법을 선택하느냐에 따라 모수에 대한 추정량과 그 추정량의 표준오차가 달라져 결국 검정통계량이 달라지기 때문이다. $\mu_{C|H} - \mu_{P|H}$의 95% 또는 99% 신뢰구간의 하한을 M_1으로 추정시 확률변수가 되므로 당연히 비열등성 마진(δ)도 확률변수로 간주되어야 한다고 주장할 수 있다. 만일 비열등성 마진(δ)이 확률변수로 간주된다면 비열등성 임상시험의 가설인 아래의 식 (5.1)에 확률변수가 포함되게 되는데, 이는 통계적 가설에 현재나 과거에서 관측된 어떠한 확률변수도 포함될 수 없다는 빈도론적인 원칙(frequentist principle)에 어긋나게 된다.

$$H_0 : \mu_T \leq \mu_C - \delta \quad vs. \quad H_1 : \mu_T > \mu_C - \delta. \tag{5.1}$$

결국 fixed margin method는 빈도론적인 원칙을 따르는 방법이므로, 비록 비열등성 마진(δ) 결정시 $\mu_{C|H} - \mu_{P|H}$의 신뢰구간 하한이 사용되었다고 하더라도, 결정된 비열등성 마진(δ)은 확률변수가 아닌 상수로 간주하게 된다. 또한 비열등성 마진(δ)을 결정한 이후 과정부터 우리가 검정할 가설이 정해지는 것이기 때문에, 비열등성 마진(δ)을 정하는 과정 중에 수행된 일은 통계적 추론 과정에 전혀 포함되지 않는다. 물론 이에 동의하지 않는 통계전문가들도 있을 것이다. 하지만 빈도론적인 원칙을 따르는 한 이는 어쩔 수 없는 일인 것이다. 이처럼 비열등성 마진(δ)을 상수로 간주할지 아니면 확률변수로 간주할지는 비열등성 임상시험의 통계방법에 대한 핵심적인 논쟁거리 중의 하나이다.

이러한 일련의 과정은 Figure 5.1에 잘 요약되어 있다. Figure 5.1을 다시 설명하면 과거 임상시험에서 얻어진 자료와 임상적 판단, assay sensitivity, constancy assumption 등을 이용하여 비열등성 마진(δ)을 결정하는 일은 머리 속에서 관념적으로만 이루어지는 행위이고 통계적 추론 과정에 포함되지 않으며, 이러한 모든 정보는 하나의 상수로 요약된다는 것이다 (Hung 등, 2009).

빈도론적인 원칙에 따르면 비열등성 마진(δ)이 결정되어 가설 (5.1)이 정해진 후에야 비로소 통계적 추론 과정이 시작되게 된다. 하지만 정작 통계적 추론 과정은 간단하다. $\mu_T - \mu_C$의 95% 신뢰구간의 하한이 $-\delta$보다 크다면, 대립가설을 채택하여 시험약은 활성대조약에 비하여 비열등하다고 결론 내리게 된다. $\mu_T - \mu_C$의 95% 신뢰구간의 하한이 $-\delta$보다 크다는 것을 수식으로 표현하면 다음과 같다.

$$\left(\overline{X_T} - \overline{X_C}\right) - 1.96\sqrt{\frac{2S^2}{n}} > -\delta, \tag{5.2}$$

여기서 $\overline{X_T}$, $\overline{X_C}$ 그리고 S^2은 다음처럼 주어진다.

$$\overline{X_T} = \frac{1}{n}\sum_{i=1}^{n} X_{T,i}, \qquad \overline{X_C} = \frac{1}{n}\sum_{i=1}^{n} X_{C,i}, \qquad S^2 = \frac{\sum_{i=1}^{n}\left(X_{T,i} - \overline{X_T}\right)^2 + \sum_{i=1}^{n}\left(X_{C,i} - \overline{X_C}\right)^2}{2n-2}.$$

식 (5.2)는 검정통계량을 사용하여 다음과 같이 나타낼 수 있다.

$$T_f = \frac{\overline{X_T} - \overline{X_C} + \delta}{\sqrt{2S^2/n}} > 1.96.$$

다시 한 번 강조하지만 이 검정통계량의 분자 중에 $\overline{X_T} - \overline{X_C}$만 확률변수로 간주되고, 비열등성 마진($\delta$)은 상수로 간주된다. 그러므로 검정통계량의 분모에는 $\overline{X_T} - \overline{X_C}$의 표준오차만 존재하게 되고 비열등성 마진($\delta$)은 검정통계량의 분모에 아무런 영향을 미치지 않게 된다.

빈도론적 원칙에 따라 통계적 추론시 시험약이 활성대조약에 비해 '비열등함'만을 입증할 수 있으며, 그 이상으로 확장하여 해석할 수 없다. 다시 말해 빈도론적인 원칙에 의해서는 시험약이 활성대조약에 비하여 비열등하다는 것까지만 알 수 있고, 시험약이 가상의 위약보다 우월하다는 ($\mu_T - \mu_{P,put} > 0$) 결론은 내릴 수 없다.

이와 같은 빈도론적 원칙에 따른 통계적 추론을 모두 마치고 나서, 시험약이 가상의 위약보다 우월하다는 것을 보이기 위해 다음의 추론을 하게 되는데, 이런 과정 역시 통계적 추론의 과정이 아닌 (머리 속에서 이루어지는) 관념적인 추론 과정에 해당된다.

$$\begin{aligned}
&\mu_T - \mu_C > -\delta \geq -(\mu_{C|H} - \mu_{P|H}) \\
&\Rightarrow (\mu_T - \mu_C) + (\mu_{C|H} - \mu_{P|H}) > 0 \\
&\Rightarrow (\mu_T - \mu_C) + (\mu_T - \mu_{P,put}) > 0 \text{ (under constancy assumption)} \\
&\Rightarrow \mu_T - \mu_{P,put} > 0.
\end{aligned}$$

이런 모든 과정을 마친 후에 1종의 오류 확률을 평가해보는 일은 매우 의미있는 일이다. Fixed margin method는 당연히 빈도론적 원칙에 의해서만 평가되는 것이므로, 여기서 평가되는 fixed margin method의 1종의 오류 확률은 비열등성 마진(δ)이 주어졌을 때, 시험약과 대조약을 비교하는 동일한 비열등성 임상시험을 무수히 반복하는 경우, 비록 시험약이 대조약에 비하여 비열등하지 않음에도 불구하고 ($H_0 : \mu_T \leq \mu_C - \delta$), 시험약이 대조약에 비하여 비열등하다고 결론 ($H_1 : \mu_T > \mu_C - \delta$)을 내리게 되는 확률을 의미하며 이를 수식으로 표현하면 다음과 같이 쓸 수 있다 (Hung 등, 2007).

$$P(T_f > 1.96|H_0) = P\left(\frac{\overline{X_T} - \overline{X_C} + \delta}{\sqrt{2S^2/n}} > 1.96|H_0\right)$$

다시 말해, Fixed margin method에 의한 통계적 추론시 비열등성 마진(δ)은 상수이므로, 1종의 오류 확률은 오직 나머지 확률변수들인 $\overline{X_T}$, $\overline{X_C}$, S^2만을 사용하여 계산되어진다. 이러한 확률변수들은 오직 현재의 비열등성 임상시험에서 얻어지는 확률변수들이므로, 여기서 1종의 오류 확률의 의미는 상수인 비열등성 마진(δ)에 대하여 현재의 비열등성 임상시험을 무수히 반복하는 경우에 비록 시험약이 대조약에 비하여 비열등하지 않음(다시 말해 열등할 수 있음)에도 불구하고 ($H_0 : \mu_T \leq \mu_C - \delta$), 시험약이 대조약에 비하여 비열등하다고 결론 ($H_1 : \mu_T > \mu_C - \delta$)을 내리게 되는 확률을 의미한다. 결국 이 1종의 오류 확률을 평가하는 데는 현재 비열등성 임상시험만을 고려하며 과거의 임상시험은 고려하지 않게 되므로, 이러한 1종의 오류 확률은 within-trial Type I error rate이라고 불리기도 한다 (Hung 등, 2009). 여기서의 1종의 오류란 고정된 비열등성 마진(δ)에 대하여 현재의 비열등성 임상시험에서 세운 가설인 귀무가설 ($H_0 : \mu_T \leq \mu_C - \delta$)과 대립가설 ($H_1 : \mu_T > \mu_C - \delta$)에 대한 것이지, 그것을 넘어서서 비열등성 마진을 결정하거나 시험약이 가상의 위약보다 우월하다는 결론에 대한 것은 아니다. 결국 시험약의 효과가 가상의 위약보다 우월하다는 ($\mu_T - \mu_{P,put} > 0$) 결론은, 빈도론적 원칙에 의해 얻어지는 통계적 추론 과정이 아니기 때문에, 그 오류 확률이 얼마나 될지도 역시 평가 불가능한 일인 것이다.

6. Synthesis Method

Synthesis method는 fixed margin method가 사용한 빈도론적 원칙이 아닌 메타분석 형태의 방법을 사용한다 (Hung 등, 2009). 즉 synthesis method에서는 현재의 비열등성 임상시험과 과거의 임상시험을 마치 동일한 시점에 동일한 임상시험계획서에 따라 수행한 하나의 동일한 임상시험에서 얻어진 자료라고 가정하여 분석하는 방법이다. 물론 이러한 가정은 사실이 아니므로 synthesis method의 단점이 된다.

Synthesis method를 사용하여 검정하려고 하는 가설은 식 (2.3)과 식 (2.4)로 주어진다. 식 (2.4)는 constancy assumption을 사용하여 다음과 같이 표현될 수 있다.

$$H_0 : (\mu_T - \mu_C) \leq (\lambda - 1)(\mu_{C|H} - \mu_{P|H}) \quad vs. \quad H_1 : (\mu_T - \mu_C) > (\lambda - 1)(\mu_{C|H} - \mu_{P|H}). \quad (6.1)$$

식 (6.1)로 표현되는 가설의 왼쪽에 있는 모수들은 현재의 비열등성 임상시험에서 추정할 수 있고, 오른쪽에 있는 모수들은 과거의 임상시험에서 추정할 수 있다.

Synthesis method에서는 현재의 비열등성 임상시험과 과거의 임상시험을 마치 동일한 시점에서 동일한 임상시험계획서에 따라 수행한 하나의 동일한 임상시험에서 얻어진 자료라고 가정하여 분석하므로, 가설 (6.1)에 대한 검정통계량은 다음처럼 주어진다 (Kang과 Tsong, 2010).

$$T_s = \frac{(\overline{X_T} - \overline{X_C}) - (\lambda - 1)(\overline{X_{C|H}} - \overline{X_{P|H}})}{\sqrt{2S^2/n + (\lambda - 1)^2 2S_H^2/n_H}},$$

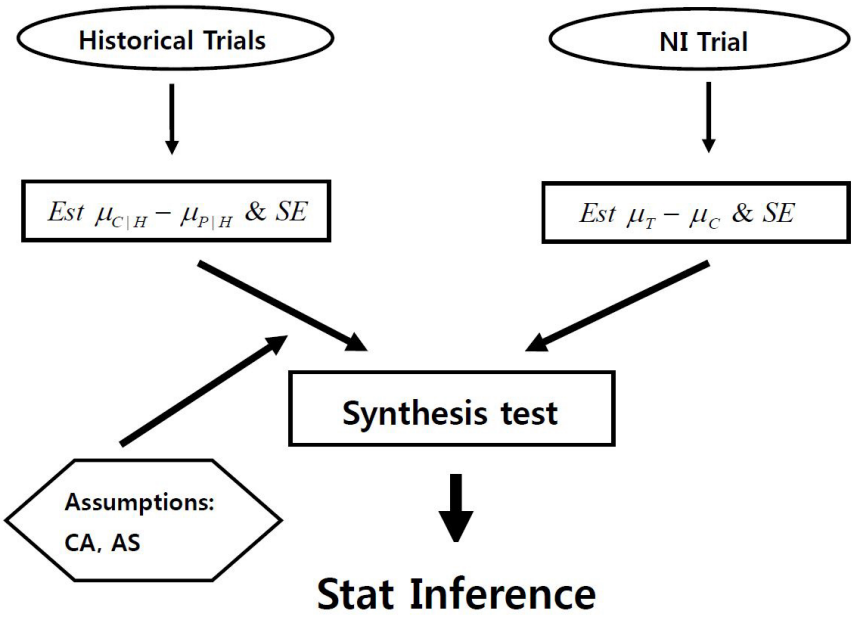

Figure 6.1. Synthesis method

여기서 $\overline{X_{C|H}}$, $\overline{X_{P|H}}$ 그리고 S_H^2은 다음처럼 주어진다.

$$\overline{X_{C|H}} = \frac{1}{n_H}\sum_{i=1}^{n} X_{C|H,i}, \qquad \overline{X_{P|H}} = \frac{1}{n_H}\sum_{i=1}^{n} X_{P|H,i},$$

$$S_H^2 = \frac{\sum_{i=1}^{n_H}\left(X_{C|H,i} - \overline{X_{C|H}}\right)^2 + \sum_{i=1}^{n_H}\left(X_{P|H,i} - \overline{X_{P|H}}\right)^2}{2n_H - 2}.$$

만일 $T_s > -1.96$이면 귀무가설을 기각하게 된다.

Figure 6.1은 synthesis method가 어떤 방법인지를 잘 설명해주고 있다 (Hung 등, 2009). 즉 현재의 비열등성 임상시험과 과거의 임상시험을 마치 하나의 임상시험에서 얻어진 자료인 것처럼 간주하여 분석하는 것이다.

Synthesis method에서 적절한 1종 오류 확률은 현재의 비열등성 임상시험과 과거의 임상시험을 모두 동시에 무수히 반복했을 때, 식 (6.1)에 있는 귀무가설이 참임에도 불구하고 대립가설을 채택하게 되는 오류 확률로, 이를 수식으로 표현하면 다음과 같다 (Hung 등, 2007).

$$P(T_s > 1.96|H_0) = P\left(\frac{(\overline{X_T} - \overline{X_C}) - (\lambda-1)(\overline{X_{C|H}} - \overline{X_{P|H}})}{\sqrt{2S^2/n + (\lambda-1)^2 2S_H^2/n_H}} > 1.96|H_0\right),$$

여기서 1종 오류의 확률은 현재의 비열등성 임상시험에서 얻어지는 $\overline{X_T}$와 $\overline{X_C}$뿐만 아니라 과거의 비열등성 임상시험에서 관측되는 $\overline{X_{C|H}}$와 $\overline{X_{P|H}}$가 모두 확률변수로 간주되는 것으로, 이러한 사실은 검정통계량의 분모에 이러한 확률변수들의 표준오차가 포함됨을 통해서도 알 수 있다. 그렇기 때문에 이 1종의 오류가 의미하는 바는 현재의 비열등성 임상시험과 과거의 임상시험을 모두 동시에 무수히 반복했을

때, 식 (6.1)에 있는 귀무가설이 참임에도 불구하고 식 (6.1)에 있는 대립가설을 채택하게 되는 오류 확률을 나타내게 된다. 하지만 과거의 임상시험은 이미 과거에 수행된 임상시험이라 무수히 반복한다는 가정 자체가 논리에 맞지 않는 단점이 있다.

Synthesis method의 또 다른 단점으로는 1종 오류가 constancy assumption에 매우 민감하게 달라진다는 점이다 (Wang과 Hung, 2003). 즉 constancy assumption이 만족되는 경우에는 1종 오류가 유의수준과 동일하게 되지만, constancy assumption이 만족되지 않는 경우에는 다른 방법보다 1종 오류의 증가가 더 심할 수 있다.

그 외에도 synthesis method의 문제점으로는 만일 서로 두 제약회사가 같은 적응증을 갖는 두 다른 약을 개발하는데 동일한 활성대조군을 사용하는 경우, 두 개의 검정통계량에 $\overline{X}_{C|H} - \overline{X}_{P|H}$이 동일하게 포함되므로, 두 검정통계량은 독립이 아니게 된다. 즉 서로 다른 두 개의 비열등성 임상시험임에도 불구하고 동일한 활성대조군을 사용했다는 사실만으로 두 비열등성 임상시험에서의 비열등성 가설 검정은 통계적으로 독립이 아니게 되며, 이는 한 임상시험의 결과가 다른 임상시험의 결과에 영향을 미치게 됨을 의미한다. 이는 결국 1종의 오류 증가로 이어지게 된다 (Kang과 Tsong, 2010; Kang과 Ryu, 2011).

7. Two Confidence Interval Method

Two confidence interval method(이하 TCI)는 두 가지 의미로 사용되어져 왔기 때문에 많은 혼란을 일으켰고 그래서 그 두 가지 의미를 명확하게 구분하는 것이 매우 중요하다 (Wang과 Kang, 2012).

TCI가 갖는 첫 번째 의미는 비열등성 마진(δ)이 상수인 TCI이다. 상수의 비열등성 마진을 갖는 TCI에서 검정하고자 하는 비열등성 가설은 fixed margin method가 검정하고자 하는 가설인 식 (2.1)과 동일하다. 또한 비열등성 마진(δ)도 fixed margin method에서 사용하는 동일한 방법을 사용하여 설정한다. 뿐만 아니라 비열등성을 검정하는 방법도 동일하여 $\mu_T - \mu_C$에 대한 95% 신뢰구간의 하한이 $-\delta$보다 크면 귀무가설을 기각하게 된다. 결국 요약하면 상수의 비열등성 마진을 갖는 TCI는 fixed margin method와 동일한 방법이다. 다만, 그 이름에 two confidence interval이라는 이름이 들어가게 된 이유는, 비열등성 마진을 결정할 때 과거의 임상시험 자료를 활용하여 $\mu_{C|H} - \mu_{P|H}$에 대한 신뢰구간을 구하게 되고, 현재의 비열등성 시험에서 비열등성을 검정할 때 $\mu_T - \mu_C$에 대한 신뢰구간을 구하게 되어, 총 두 개의 신뢰구간을 구하는 데서 붙여진 이름이다.

TCI가 갖는 두 번째 의미는 비열등성 마진(δ)이 확률변수인 TCI이다 (Hung 등, 2003). Figure 7.1은 synthesis method가 어떤 방법인지를 잘 설명해주고 있다 (Wang과 Kang, 2012). 비열등성 마진(δ)이 확률변수인 TCI의 특징은 비열등성 마진을 추정하는데 발생하는 불확실성을 추론에 반영한다는 것이다. 비열등성 마진을 설정하는 방법은 위약 대비 시험약의 효과가 위약 대비 활성대조약 효과에 대해 적어도 $100\lambda\%$ $(0 \le \lambda \le 1)$ 를 보존하도록 하게 하는 것이다. 이를 수식으로 표현하면 다음과 같다.

$$\frac{(\mu_T - \mu_{P,put})}{(\mu_C - \mu_{P,put})} > \lambda. \tag{7.1}$$

식 (7.1)을 좌변에 $\mu_T - \mu_C$만 남도록 적절히 변환하면, 식 (7.1)은 다음과 같이 정리할 수 있다.

$$\mu_T - \mu_C > -(1-\lambda)(\mu_C - \mu_{P,put}).$$

그러므로 비열등성 마진(δ)이 확률변수인 TCI에서 검정할 가설은 다음과 같이 주어진다.

$$H_0 : \mu_T - \mu_C \le -\delta \quad vs. \quad H_1 : \mu_T - \mu_C > -\delta, \quad \delta = (1-\lambda)(\mu_C - \mu_{P,put}), \tag{7.2}$$

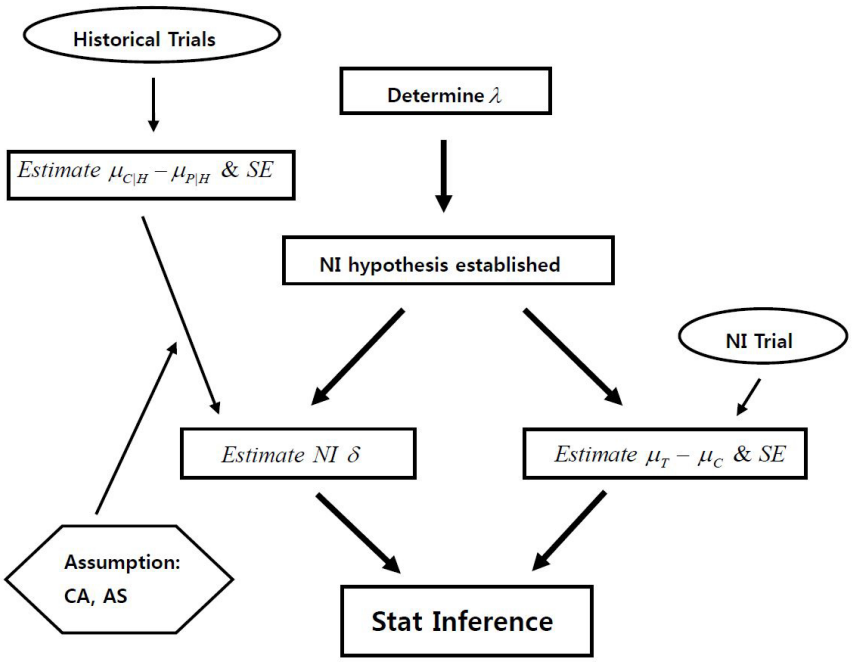

Figure 7.1. TCI with random margin

여기에서 문제점은 비열등성 마진(δ) 안에 $(\mu_C - \mu_{P,put})$가 포함되어 있으나, 현재의 비열등성 임상시험에는 위약군이 존재하지 않아 $\mu_{P,put}$에 대한 추정치를 얻을 수 없다는 점이다. 그래서 constancy assumption $(\mu_C - \mu_{P,put} = \mu_{C|H} - \mu_{P|H})$을 가정하여, 과거의 임상시험 자료를 이용하여 $\mu_{C|H} - \mu_{P|H}$의 신뢰구간의 하한을 구하고, 이를 $\mu_C - \mu_{P,put}$의 추정치로 삼게 되는 것이다. $\mu_{C|H} - \mu_{P|H}$의 신뢰구간의 하한에 $1-\lambda$를 곱하면 비열등성 마진 δ의 추정치를 얻게 된다. $\mu_T - \mu_C$에 대한 95% 신뢰구간의 하한이 $-\delta$보다 크게 되면 식 (7.2)에 주어진 귀무가설을 기각하여, 위약 대비 시험약의 효과는 위약 대비 활성대조약 효과 중 적어도 $100\lambda\%$ $(0 \le \lambda \le 1)$를 보존한다고 결론내리게 된다.

확률변수의 비열등성 마진(δ)을 갖는 TCI의 가장 큰 특징은 과거의 임상시험 자료를 이용하여 얻은 $\mu_{C|H} - \mu_{P|H}$의 신뢰구간의 하한이 여전히 상수가 아닌 확률변수로 간주한다는 점이다 (Wang과 Kang, 2012). 그러므로 확률변수의 비열등성 마진을 갖는 TCI에서 사용되고 있는 통계적 추론 방법은 synthesis method에서 사용된 통계적 추론 방법처럼, 현재의 비열등성 임상시험과 과거의 비열등성 임상시험이 마치 하나의 동일한 임상시험계획서에 따라 동일한 시점에서 수행된 하나의 임상시험 결과로 간주한다.

다음으로 확률변수의 비열등성 마진을 갖는 TCI와 synthesis method의 차이점을 이해하기 위해 확률변수의 비열등성 마진을 갖는 TCI를 수식으로 표현하면 다음과 같다. 앞에서 설명한 바와 같이 확률변수의 비열등성 마진을 갖는 TCI는 다음을 만족시킬 때 식 (7.2)에 주어진 귀무가설을 기각하게 된다.

$$\left(\overline{X_T} - \overline{X_C}\right) - 1.96\sqrt{2S^2/n} > -(1-\lambda)\left[\left(\overline{X_{C|H}} - \overline{X_{P|H}}\right) - 1.96\sqrt{2S_H^2/n_H}\right]$$

$$\Rightarrow \left(\overline{X_T} - \overline{X_C}\right) - (\lambda - 1)\left(\overline{X_{C|H}} - \overline{X_{P|H}}\right) > 1.96\left(\sqrt{2S^2/n} + (1-\lambda)\sqrt{2S_H^2/n_H}\right)$$

$$\Rightarrow \frac{(\overline{X_T} - \overline{X_C}) - (\lambda - 1)(\overline{X_{C|H}} - \overline{X_{P|H}})}{\sqrt{2S^2/n} + \sqrt{(\lambda - 1)^2 2S_H^2/n_H}} > 1.96$$

위 마지막 부등식의 좌변은 synthesis method에서 사용한 검정통계량 T_s와 매우 유사한다. 보다 정확하게 비교하면 두 식에서 분자는 완벽하게 동일하고, 분모만 약간 다르다. Synthesis method에서는 분모 전체에 루트가 씌워져서 분자에 있는 통계량에 대한 표준오차가 바로 분모에 있는 반면, 확률변수의 비열등성 마진을 갖는 TCI에서는 루트가 각각 따로 씌워져 있기 때문에, 분모에 있는 통계량이 분자에 있는 통계량의 표준오차가 되지 못한다. 그러므로 synthesis method가 확률변수의 비열등성 마진을 갖는 TCI보다 더 높은 검정력을 가지게 되는데, 그 이유는 synthesis method는 1종이 오류가 2.5%인 반면, 확률변수의 비열등성 마진을 갖는 TCI는 1종의 오류가 2.5%보다 작아지기 때문이다.

8. 결론

본 논문에서는 비열등성 임상시험에서 얻어진 자료를 분석하는데 주로 사용되는 세 가지 통계적 검정방법의 특징을 살펴보았다. 문제의 핵심은 과거의 임상시험에서 얻어진 자료들을 어떻게 활용할 것인가 하는 점이다. 빈도론적 원칙에 의하면 과거의 자료는 어떻게 활용하고 요약하던 현재의 통계적 추론에서는 상수로 간주되어야 한다는 사실이다. 이렇게 되면 과거 임상시험에서 얻어진 자료들을 요약하는 과정에서 생기는 변동 (또는 분산)이 반영되지 않는다는 문제점이 있다. 반면에 이러한 변동 (또는 분산)을 반영할 경우 과거의 임상시험을 마치 현재에 수행된 임상시험처럼 간주해야 하는 모순이 발생하게 된다. 이러한 이슈들은 결국 임상시험 자료를 분석하는데 어떠한 통계적 철학을 가지고 접근할 것인가에 해당하는 문제로 귀착된다. 현재 임상시험에서는 빈도론적 원칙을 사용하는 방법과 메타분석 형태의 방법이 모두 사용되고 있으며 이는 비열등성 임상시험에서 사용되는 통계방법에 대한 가장 큰 논쟁거리 중 하나이다.

References

Blackwelder, W. C. (1982). "Proving the null hypothesis" in clinical trials, *Controlled Clinical Trials*, **3**, 345–353.

D'Agostino, R. B., Massaro, J. M. and Sullivan, L. M. (2003). Non-inferiority trials: Design concepts and issues - the encounters of academic consultants in statistics, *Statistics in Medicine*, **22**, 169–186.

Fleming, T. R. (2008). Current issues in non-inferiority trials, *Statistics in Medicine*, **27**, 317–332.

Hasselblad, V. and Kong, D. F. (2001). Statistical methods for comparison to placebo in active-control trials, *Drug Information Journal*, **35**, 435–449.

Holmgren, E. B. (1999). Establishing equivalence by showing that a specified percentage of the effect of the active control over placebo in maintained, *Journal of Biopharmaceutical Statistics*, **9**, 651–659.

Hung, H.-M., Wang, S.-J. and O'Neill, R. (2007). Issues with statistical risks for testing methods in non-inferiority trial without a placebo arm, *Journal of Biopharmaceutical Statistics*, **17**, 201–213.

Hung, H. M., Wang, S. J. and O'Neill, R. (2009). Challenges and regulatory experiences with non-inferiority trial design without placebo arm, *Biometrical Journal*, **51**, 324–334.

Hung, H.-M., Wang, S.-J.,Tsong, Y., Lawrence, J. and O'Neill, R. (2003). Some fundamental issues for noninferiority testing in active controlled trials, *Statistics in Medicine*, **22**, 213–225.

ICH E10 (2001). *Choice of control group and related issues in clinical trials.*

Kang, S. H. (2010). *Biostatistical Methods for New Drug Development*, Freeacademy, Seoul.

Kang, S. H. and Ryu, Y. (2011). The adjustment of the type I error rate in non-inferiority trials with -margin approach: Each of two different new drugs is approved with two independent trials with the same active control, *Journal of Biopharmaceutical Statistics*, **21**, 498–510.

Kang, S. H. and Tsong, Y. (2010). Strength of evidence of non-inferiority trials - the adjustment of the type I error rate in non-inferiority trials with the synthesis method, *Accepted by Statistics in Medicine*, **29**, 1477–1487.

Leber, P. D. (1989). Hazards of inference: The active control interpretation, *Epilepsia*, **30**, S57–S63.

Temple, R. and Ellenberg, S. S. (2000). Placebo-controlled trials and active- controlled trials in the evaluation of new treatment, part I: ethical and scientific issues, *Annals of Interna Medicine*, **133**, 455–463.

Tsong, Y. (2007). The utility of active-controlled noninferiority / equivalence trials in drug development, *International Journal of Pharmaceutical Medicine*, **21**, 225–233.

Tsong, Y., Wang, S.-J., Hung, H.-M. and Cui, L. (2003). Statistical issues on objective, design and analysis of noninferiority active-controlled clinical trial, *Journal of Biopharmaceutical Statistics*, **13**, 29–41.

U.S. FDA (2010). *Guidance for Industry: Non-Inferiority Clinical Trials (draft guidance)*

Wang, S.-J. and Hung, H.-M. (2002). Utility and pitfalls of some statistical methods in active controlled clinical trials, *Controlled Clinical Trials*, **23**, 15–28.

Wang, S.-J. and Hung, H.-M. (2003). Assessing treatment efficacy in non-inferiority trials, *Controlled Clinical Trials*, **24**, 147–155.

Wang, S. Y. and Kang, S. H. (2012). Strength of evidence of non-inferiority trials with the two confidence interval method with random margin, *Journal of Biopharmaceutical Statistics*, **In press**