The Korean Journal of Applied Statistics (2019) DOI: https://doi.org/10.5351/KJAS.2019.32.1.001
32(1), 1–13

A comparison of multiple hypothesis testing methods and combination methods in seamless Phase II/III clinical trials

Song Han[a] · Hanna Yoo[b] · Jae Won Lee[c,1]

[a]LG Chem, Ltd; [b]Department of Computer Software, Busan University of Foreign Studies;
[c]Department of Statistics, Korea University

(Received May 11, 2018; Revised December 2, 2018; Accepted December 5, 2018)

Abstract

Abstract An adaptive seamless Phase II/III clinical trial design enables a reduction in the sample size (in comparison to a conventional design) that also shortens the clinical development time. It is also very effective in clinical trials since it can have higher statistical power than Phase III alone. In this study, we use extensive simulation studies to compare several multiple hypothesis testing methods that can help select the best doses in a Phase II study along with several methods to combine p-values of the Phase II and Phase III study.

Keywords: adaptive design, seamless design, multiple testing, combination tests, Phase II design, Phase III design

1. 서론

적응적 심리스 제 2상/제 3상 디자인(adaptive seamless Phase II/III design)은 따로 진행하는 임상시험의 전통적인 방법을 단일 연구로 통합하는 원리로써, 이는 기존의 디자인들과 비교하여 피험자수를 줄일 수 있을 뿐만 아니라 임상 개발 시간이 단축되므로 더욱 효율적이다. 즉, 이는 제 2상 연구의 끝 시점과 제 3상 연구의 시작시점의 상당한 시간적 차이를 없앨 수 있을 뿐 아니라, 제 3상 시험을 단독으로 진행 하였을 때보다 더 높은 검정력을 갖는다.

본 논문에서는 적응적 심리스 제 2상/제 3상 디자인을 이용하여 제 2상에서의 최고효과 용량군을 선택하고, 각 단계에서의 유의확률이 주어졌을 때 두 단계를 결합 하는 과정들을 고려해 보고, 최고효과 용량군을 선택하기 위해서 제 2상에서의 각 용량군에 대한 다중가설 검정을 이용하여 살펴볼 것이다. 다중가설에 관한 기존의 연구로는 family wise error rate (FWER) 방법인 Holm (1979), Simes (1986), Hochberg (1988), 그리고 Hommel (1988) 등이 있고, false discovery rate (FDR) 방법인 Benjamini와

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No.2017R1D1A1B03028279 and the Korea government (MEST) (No.2017R1C1B5076671).
[1]Corresponding author: Department of Statistics, Korea University, 145, Anam-Ro, SeongBuk-Gu, Seoul 02841, Korea. E-mail: jael@korea.ac.kr

Hochberg (1995)와 Benjamini와 Yekutieli (2001) 등이 있다. 두 단계를 결합하는 방법에 관한 기존의 연구로는 Bauer와 Kohne (1994), Fisher (1998), Cui 등 (1999), Lehmacher와 Wassmer (1999), 그리고 Denne (2001) 등이 있다. 특히 Bauer와 Kohne (1994)는 결합하는 방법에 있어서 역카이제곱법(inverse χ^2 method) (Fisher, 1932)을 이용하였고, Fisher (1998), Cui 등 (1999), Lehmacher과 Wassmer (1999), Denne (2001)는 가중역정규방법(weighted inverse normal method) (Mosteller와 Bush, 1954)을 이용하였고, George (1977)이 제안한 로짓방법(logit method) 등 여러 방법들이 있다.

두 단계로 이루어진 설계를 검정하는 방법에 관한 연구는 지금까지 많이 진행되어 왔다. 특히 Thall 등 (1988)는 임상시험에 있어서 단계를 나누어 최고효과 용량군을 선택하여 위약군과 비교하는 연구를 진행하였다. Jennison과 Turnbull (2007)은 지금까지 적응적 심리스 디자인에 있어서 제 2상과 제 3상을 결합하는 방법을 비교하였다. 최근에도 다양한 결합 검정 방법들이 제안되고 있다. Kunz 등 (2015)에서는 심리스 제 2상/3상 디자인에서 처리 선택(treatment selection) 방법과 관련하여 Stallard (2010)과 Fried 등 (2011)의 두 방법을 비교 분석하였다. 두 방법들은 중간분석의 의사결정에 있어서 short-term endpoint 데이터의 사용을 가능하게 하는 비교적 최근에 제안된 방법들이다. Fried 등 (2011)에서는 첫 번째 단계의 검정통계량에 Dunnett 수정방법을 이용하여 중간분석에서 관찰한 개체들과 두 번째 단계의 마지막에 관찰한 개체들을 결합하는 결합 검정 방법을 제안하였다. Stallard (2010)는 각 중간 분석에서 계산되는 long-term 처리 효과의 최대우도추정량에 기반하여 처리 선택하는 방법을 제안하였고 Fried 등 (2011) 방법과는 다르게 중간분석에서 적어도 몇 개의 long-term endpoint 데이터가 존재해야 한다는 차이점이 있다.

이 논문에서는 제 2상과 제 3상 두 단계를 결합하는 방법 중에서 메타분석방법인 역카이제곱법, 가중역정규방법 그리고 Thall 등 (1988) 논문에서 제시한 Thall, Simon, 그리고 Ellenberg (TSE) 방법을 비교하였다. 본 연구에서는 제 2상에서 최고효과 용량군을 선택하기 위해 각 치료군에 대한 다중가설 검정을 비교하고, 두 단계를 결합하는 5가지 방법들을 비교해 보고자 한다.

본 논문의 구성은 다음과 같다. 2절에서는 적응적 심리스 디자인에 대해 개념을 설명하고 3절에서는 다중가설검정방법들에 대하여 소개할 것이며, 4절에서는 두 단계를 결합하는 방법과 제 3상의 데이터만 이용하는 기존의 방법에 관하여 설명할 것이다. 5절에서는 모의실험 결과를 토대로 다중가설검정방법을 이용하여 제 2상에서의 최고효과 용량군을 선택하고 제 2상과 제 3상을 결합하는 과정에 있어서 여러 가지 검정방법들을 통해 어떤 분석 방법을 사용하는 것이 적절한 것인가에 관하여 설명하였다. 마지막으로 6절에서는 본 연구에 대한 결론과 전체적인 내용에 대하여 기술하였다.

2. Adaptive seamless design

임상시험에서 제 2상b는 제 3상에서 사용될 투약용량을 정하고, 그 효율성을 평가하기 위해 고안된 것이고 제 3상은 제 2상b 시험을 통해 효과가 있다고 판단된 신약과 기존의 약을 충분한 수의 환자들을 대상으로 비교하도록 설계된다. 적응적 심리스 제 2상b/제 3상 디자인은 제 2상b를 단계 1로 가정하고 제 3상을 단계 2로 가정하여 단계 1에서 l개 그룹의 용량군과 위약군을 두고 가장 효과적인 용량군을 선택하여 단계 2에서 다시 위약군과 실험을 진행한다. 여기서 중요한 점은 기존의 방법과 달리 제 2상b와 제 3상 사이의 시간적 공백이 없이 두 단계의 데이터를 결합하여 한 번의 실험으로 진행한다는 점이다. 따라서 두 단계에의 데이터를 결합하는 과정에서 결합검정(combination test)을 이용하여 편향(bias)를 줄이고, 유의수준 α하에서 familywise 제1종 오류를 통제하여야 한다. 설계의 절차는 다음과 같으며, Jennison과 Turnbull (2007)에서 정의된 표현을 사용하기로 한다. i번째 용량군과 위약군의 실제 유효크기를 θ_i라 하자 (단, $i = 1, \ldots, l$).

Table 3.1. Contigency table of hypothesis test

참	H_0 채택 (유의하지 않음으로 판정)	H_1 채택 (유의하다고 판정)	총
H_0	U	V	m_0
H_1	T	S	m_1
	$m-R$	R	m

- 단계 1. (제 2상b) - 예측된 치료효과를 $\hat{\theta}_{1,i}$라 하면 (단, $i=1,\ldots,l$) l개의 용량군 중 가장 효과 있는 i^*번째 용량군을 제 3상으로 진행시킨다.
- 단계 2. (제 3상) - 단계 1에서 선정된 용량군과 위약군의 유효성 확증 시험을 실시하고, 시험 종료 후 단계 1의 데이터와 단계 2의 데이터를 통합하여 최종 검정을 실시한다.

즉, 가설 $H_{i^*}:\theta_{i^*}\le 0$를 기각하기 위해 유의수준 α에서의 $i^*\in I$인 각 교호가설(intersection hypothesis) H_I를 기각시킨다. 여기서 $H_I=\bigcap_{i\in I}H_i$는 모든 $i\in I$에서의 $\theta_i\le 0$인 가설들의 교호가설을 말한다. 직관적으로 볼 때, 용량군 i^*가 제 2상b에서 최고효과 용량군으로 선택되고, 제 3상이 진행된 후 최종 분석을 실시하기 위해 두 단계의 데이터를 결합할 때 l개의 용량군중 i^*를 최고효과 용량군으로 선택했을 때의 효과를 고려하여야 한다.

이를 위해 closure 원리를 이용한 결합검정을 이용한다. 여기에는 다음 두 가지 요소가 필요하다.

(a) 교호가설 검정

(b) 제 2상b, 제 3상 데이터의 결합

(a)를 해결하기 위해 3절에 소개된 다중가설 검정을 이용하고 (b)를 해결하기 위해 4절에 소개된 결합검정을 이용한다.

3. 다중 가설 검정

용량군의 개수가 늘어날수록 제 1종 오류가 증가하기 때문에 한 개의 용량군을 검정할 때 정한 유의수준을 그대로 사용하게 되면 최고효과 용량군을 결정할 때 잘못 판단할 우려가 있다. 이러한 다중검정 문제를 해결하기 위해 FWER과 FDR을 이용하게 된다. 가설과 가설 검정 결과에 따른 분할표는 아래의 Table 3.1과 같다.

총 m개의 검정에서 귀무가설 H_0을 만족하는 가설의 수가 m_0개, 대립가설 H_1을 만족하는 가설의 수를 m_1개라고 하자 가설검정의 결과로 제 1종 오류를 범한 것이 V개 이고, 제 2종 오류를 범한 것이 T개 있음을 알 수 있다. FWER은 적어도 하나의 가설을 잘못 판정할 확률이 유의수준보다 작도록, 즉 $P(V\ge 1)\le\alpha$이 되도록 조정하는 것이다. FDR은 유의하게 판정한 검정결과 중에서 잘못된 검정 비율을 조정하는 방법이다. 즉, $E(V/R)\le q$를 만족되도록 보정하는 방법이다.

3.1. Bonferroni

Dunn (1961)은 Bonferroni 부등식을 이용하여 다중가설검정을 하는 평균의 다중비교 검정 방법을 제시하였다. Bonferroni 부등식은 가설검정을 할 때 유의수준 α를 할당하는 일반적인 방법으로 이 부등식이 다중검정에 적용될 경우 용량군의 제 1종 오류가 설정된 유의수준 α를 초과하지 않도록 통제하는 역할을 한다. 고전 Bonferroni 검정에서는 집합 내의 모든 가설이 동일한 유의수준 α/m에서 검정된다.

3.2. Holm

Holm (1979)은 m개의 가설 $H_1, \ldots, H_m$에 대하여 적절한 검정통계량을 이용하여 각 가설의 유의확률을 얻는 방법을 제안하였다. 여기서 얻어진 유의확률을 가장 작은 것에서부터 순서대로 나열하고, 그 나열된 P값의 순서에 따라 가설을 재배열한다. 재배열된 유의확률 $P_{(1)}, P_{(2)}, \ldots, P_{(m)}$으로 가설을 다시 정의하면 $H_{0(1)}, H_{0(2)}, \ldots, H_{0(m)}$와 같이 나타낼 수 있다. $H_{0(1)}, H_{0(2)}, \ldots, H_{0(m)}$은 최소의 집합이 되는데, 어느 가설이 다른 가설의 부분집합인 경우, 그 가설은 최소의 집합에 포함될 수 없게 된다. Holm (1979) 방법에서 각 가설들은 한 번에 한 가설씩 연속적으로 증가된 유의수준 하에서 검정된다. Holm (1979) 방법의 절차는 가장 작은 유의확률에 상응하는 첫 번째 가설 $H_{0(1)}$을 검정하는 것으로 시작된다. P_1은 유의수준 α/m에서 검정되는데, 여기에서 α는 제 1종 오류이고 m은 가설의 개수이다. 첫 번째 가설이 α/m에서 기각되면, 두 번째 가설은 $\alpha/(m-1)$에서 검정되고, 두 번째 가설도 기각되면 세 번째 가설은 $\alpha/(m-2)$에서 검정한다. 이와 같은 방법으로 현재 검정하는 가설이 기각되면 다음 순서의 가설이 단계적으로 증가하는 유의수준에서 검정된다. 일반적으로 Holm 방법에서의 가설은 α/k에서 검정되는데 $k = m, m-1, \ldots, 1$로 1씩 감소한다. 이 단계적 과정은 어느 한 가설이 기각되지 않거나 마지막 가설 $H_{0(m)}$이 기각될 때까지 계속된다. 만약 $P_{(i)} \leq \alpha/(m-i+1)$이면 첫 번째 가설부터 $i-1$번째 가설 $H_{0(1)}, H_{0(2)}, \ldots, H_{0(i-1)}$까지는 기각하고, i번째 가설을 검정하는 단계 (단계 i)에서 i번째 가설 $H_{0(i)}$과 그 이후에 있는 모든 가설 $H_{0(i+1)}, H_{0(i+2)}, \ldots, H_{0(m)}$은 검정조차 시도하지 않고 Holm 방법은 그 검정절차가 끝난다.

3.3. Simes

Simes (1986)은 고전 Bonferroni 방법을 개선하여 다음과 같은 방법을 제안하였다. m개의 가설에 따라 각각의 유의확률을 구하고 유의확률을 낮은 것부터 순서화 한 것을 $P_{(1)}, \ldots, P_{(m)}$로 하고 그에 따른 가설을 $H_{(1)}, \ldots, H_{(m)}$이라고 하면 $P_{(j)} \leq j\alpha/m$일 때 그에 대한 가설 $H_{(i)}$을 기각하는 방법이다 (단, $j = 1, \ldots, m$).

3.4. Hochberg

Hochberg (1988)는 최초로 단계오름검정(step-up)을 제안하였다. Hochberg 검정은 Holm (1979)의 검정과 동일한 임계값 즉 동일한 유의수준을 사용하지만 가장 큰 유의확률을 갖는 가설부터 검정을 시작한다. 가장 큰 유의확률은 유의수준 $\alpha' = \alpha/1$과 비교되고 $P_1 > \alpha/1$이면 두 번째 가설을 유의수준 $\alpha/2$에서 검정하고 $P_2 > \alpha/2$이면 세 번째 가설을 검정하는 절차를 반복하여 단계오름검정을 한다. i번째 가설을 검정할 때 $P_i < \alpha/i$이 성립하게 되면 i번째 가설을 포함하여 다음 순서에 존재하는 가설들을 모두 기각하게 된다.

3.5. Hommel

Hommel (1986, 1988)은 Simes (1986) 검정에 기초하여 closed testing 방법을 제시하였다. m개의 가설 $H_1, \ldots, H_m$에 대하여 $j = \max\{i \in i = 1, \ldots, m : P_{(m-i+k)} > k\alpha/i \text{ for } k = 1, \ldots, i\}$를 Simes (1986) 검정에서 유의하지 않은 m개 가설들의 가장 큰 부분집합의 크기로 정의한다. 이때 j가 존재하지 않으면 모든 $H_i\,(i = 1, \ldots, m)$들이 기각이 되고 j가 존재하면 $P_i \leq \alpha/j$인 H_i들을 기각하는 방법이다. Hommel 방법은 Holm 방법보다 검정력은 좋지만, 전체 제 1종 오류를 α로 보장할 수 없다는 단점이 있다. Hommel 방법은 단지 각각의 검정이 독립이거나 양의 관계로 의존되어 있을 때만 familywise error rate α를 보장할 수 있다 (Sarkar, 1998; Sarkar와 Chang, 1997).

3.6. Benjamini and Hochberg (BH)

Benjamini Hochberg (1995) 방법은 최초로 FDR을 이용한 검정법을 소개하였다. 예를 들어 m개의 검정에서 가설 $H_{(1)}, \ldots, H_{(m)}$에 대한 유의확률이 각각 $P_{(1)}, \ldots, P_{(m)}$의 순으로 나타났다면 $P_j \leq (j/m)q$, $j = 1, \ldots, m$를 만족하는 가장 큰 j가 있을 때 $H_{(1)}, \ldots, H_{(m)}$를 기각하는 방법이다.

4. 결합 검정

4.1. 결합 가설

적응적 절차에서 중요한 부분은 메타 분석의 방법론에서부터 시작된 결합 검정이다. K 단계까지 고려한 실험에서, θ_k를 각 단계에서의 치료효과라 가정하면 각 단계에서 가설은 다음과 같다.

$$H_{0k} : \theta_k \leq 0 \quad \text{vs.} \quad H_{1k} : \theta_k > 0.$$

또한 결합 검정에서의 복합가설은 다음과 같다.

$$H_0 : \bigcap_{k=1}^{k} H_{0k},$$

$$H_A : \text{적어도 하나의 } H_{0k}\text{는 기각 한다.}$$

4.2. 유의확률의 결합

4.2.1. Inverse χ^2 method Fisher (1932)의 역카이제곱법은 균일분포와 카이제곱분포와의 관계를 이용한 방법이다. 복합 귀무가설 H_0하에서 P_i는 균일분포를 따르고 $-2\log(P_1, \ldots, P_k)$는 자유도 $2k$인 카이제곱분포를 따르게 된다. 따라서 각 단계에서 유의확률이 주어졌을 때

$$-2\log(P_1, \ldots, P_k) > \chi^2_{2k}(\alpha)$$

인 경우 귀무가설을 기각하게 된다. 즉, 다중가설 검정을 이용하여 제 2상b에서 H_i를 검정한 유의확률을 $P_{1,i}$라고 하고 제 3상에서 H_I를 검정한 유의확률을 $P_{2,i}$라고 하면

$$-2\log(P_{1i}, P_{2i}) > \chi^2_4(\alpha)$$

인 경우 귀무가설 H_i를 기각 한다.

4.2.2. Weighted inverse normal method 가중 역정규법은 각각의 유의확률 P_i를 정규점수 z_i로 치환한 후 이를 평균화 하여 얻은 통계량을 이용하여 검정하는 방법이다. 표준정규분포의 누적분포 함수를 $\Phi(x)$라 하면 H_{0i}하에서 $z_i = \Phi^{-1}(P_i)$는 표준정규분포를 따르게 된다. 따라서 각 단계에서 유의확률이 주어졌을 때

$$w_1 Z_1 + \cdots + w_k Z_k > z(\alpha)$$

인 경우 귀무가설을 기각하게 된다. 즉, 다중가설 검정을 이용하여 제 2상b에서 H_i를 검정한 유의확률을 $P_{1,i}$라고 하고 제 3상에서 H_i를 검정한 유의확률을 $P_{2,i}$라고 하면

$$w_1 Z_{1i} + w_2 Z_{2i} > z(\alpha)$$

인 경우 귀무가설 H_i를 기각 한다.

4.2.3. Logit method 실험된 결과의 유의성검정을 위한 방법으로 George (1977)는 다음과 같은 통계량을 제안하였다.

$$L = \sum_{i=1}^{k} \log\left[\frac{P_i}{1-P_i}\right].$$

복합귀무가설 H_0가 사실일 때

$$L^* = L\sqrt{\frac{3(5K+4)}{\pi^2 k(5k+2)}}$$

는 근사적으로 자유도가 $5k+4$인 t-분포를 따르게 된다. 따라서 각 단계에서 유의확률이 주어졌을 때 L^*가 다음을 만족하면 귀무가설을 기각하게 된다.

$$L^* = L\sqrt{\frac{3(5K+4)}{\pi^2 k(5k+2)}} > t_{(1-\alpha)}(5k+4).$$

즉, 다중가설 검정을 이용하여 제 2상b에서 H_I를 검정한 유의확률을 $P_{1,i}$라고 하고 제 3상에서 H_i를 검정한 유의확률을 $P_{2,i}$라고 하면

$$L\sqrt{\frac{42}{\pi^2 \times 24}} > t_{(1-\alpha)}(14) \quad \left(\text{단}, L = \log\frac{P_1}{1-P_1} + \log\frac{P_2}{1-P_2}\right)$$

인 경우 귀무가설 H_i를 기각 한다.

4.3. Thall, Simon, and Ellenberg (TSE) method

TSE 방법은 앞에서 제시된 유의확률 결합의 방법과 달리 closure 원리를 이용하지 않고 단계 2까지 설계된 실험에서 제 2상b에서 처리효과가 가장 높은 용량군 i^*를 선택하고 그 용량군을 제 3상으로 진행하여 검정하는 방법이다. 설계의 절차는 다음과 같다.

- 단계 1 (제 2상b) - 각 용량군과 위약군에 m_1명씩 배정한다. 위약군에 대한 용량군 i의 효과를 추정하여 $\hat{\theta}_{1,i}$라 하고 이 추정량의 최대값을 $\hat{\theta}_{1,i^*}$라 하자. 만약 $\hat{\theta}_{1,i^*} < C_1$이면 단계 1 (제 2상b)에서 귀무가설 ($H_0 : \theta_1 = \cdots = \theta_l = 0$)을 기각하지 않고 실험을 종료한다. 즉, 용량군과 위약군의 처리효과의 차이는 없다는 것을 의미한다. 만약 $\hat{\theta}_{1,i^*} \geq C_1$이면 용량군 i^*를 선택하고 단계 2 (제 3상)으로 진행한다.
- 단계 2 (제 3상) - i^*번째 용량군과 위약군에 각각 m_2명씩 배정한다. 통계량 $T_{i^*} = (m_1\hat{\theta}_{1,i^*} + m_2\hat{\theta}_{2,i^*})/(m_1+m_2)$을 이용하여 두 그룹의 데이터를 결합한다. 만약 $T_{i*} < C_2$이면 귀무가설 H_0을 기각하지 않는다. 만약 $T_{i*} \geq C_2$이면 귀무가설 H_0를 기각하고 이는 $\theta_{i^*} > 0$를 의미하게 된다한다. 제 1종 오류율은 $H_0 : \theta_1 = \cdots = \theta_l = 0$ 하에서 어떠한 용량군이 선택될 확률이므로, 단계 1에서 선택된 용량군이 선택 되었다면 그 용량군은 i^*라는 것을 알 수 있다.

4.4. Conventional

본 논문에서는 앞에서 제시한 방법들과 임상시험의 전통적(conventional) 설계방법을 함께 비교할 것이다. 전통적 방법은 제 2상b와 제 3상을 따로 분리하여 실험을 실시하고 제 3상의 데이터만 이용하여 결론을 내는 방법이다.

Table 5.1. Sample size of effect size and standard deviation

용량군-위약군					$\sigma = 3$		$\sigma = 5$	
	trt1	trt2	trt3	trt4	n_1	n_2	n_1	n_2
치료 효과	0	0	0	1	50	500	50	500
					100	100	100	100
					100	500	100	500
	0.5	0.5	0.5	1	50	500	50	500
					100	100	100	100
					100	500	100	500

5. 모의실험

5.1. 모의실험 구조

모의실험을 위하여 자료는 용량군 3집단과 위약군의 치료효과 비교를 위해 분산(σ^2)이 동일한 정규분포를 따른다고 가정하였으며 치료 효과 θ_i는 용량군과 위약군의 평균차이이고 추정된 치료 효과 $\hat{\theta_i}$는 평균 θ_i, 분산 $2\sigma^2/m_1$을 갖는 정규분포를 따르게 된다. 최고효과 용량군이 음의 효과를 보일 때 조기종료를 시행한다. 본 연구의 모의실험 과정은 다음과 같다.

Table 5.1에서 정의된 각각의 치료효과와 표준편차를 따르는 정규분포로부터 난수를 생성한다. 표본의 크기는 Table 5.1에서의 n_1, n_2로 설정하고, 위의 제 2상b, 제 3상 과정을 1,000번 반복 시행한다. 이렇게 생성된 난수로부터 제 2상b에서 Bonferroni 방법, Hochberg 방법, Simes 방법, Hommel 방법, Holm 방법, BH 방법에 대하여 각각 유의확률을 구한 후, 제 3상에서 inverse χ^2 방법, weighted inverse normal 방법, logit 방법을 이용하여 제 2상b, 제 3상을 결합한다. 그 후 4번째 치료군이 제 2상b에서 최고효과 용량군으로 선택되고 이 용량군이 제 3상에서 효과가 없다는 귀무가설을 기각시킬 확률을 검정력으로 가지는 conventional 방법, Tall 등 (1988)가 제시한 TSE 방법을 1,000번 반복 시행한다. 이렇게 제 2상b에서의 5가지 방법과 제 3상에서의 3가지 방법을 조합한 15가지 방법과, conventional 방법, TSE 방법 총 17가지 방법에 대하여 유의수준 2.5%에서 검정력을 계산하였다.

5.2. 모의실험 결과

각 표적 집단(집단 1)과 비표적 집단(집단 2)의 서로 다른 표본 크기의 조합에 대해서 제 2상에서 사용된 Bonferroni 방법, Hochberg 방법, Simes 방법, Hommel 방법, Holm 방법, BH 방법과 제 3상에서 사용된 inverse χ^2 방법, weighted inverse normal 방법, logit 방법의 조합인 18가지 방법과 conventional 방법, TSE 방법 총 20가지 방법을 유의 수준 $\alpha = 0.025$일 때, 처리효과에 대한 검정력을 비교하였다.

5.2.1. 치료효과가 $(0, 0, 0, 1)$이고 표준편차를 3으로 정의한 경우 (Table 5.2) 제 2상b/제 3상 표본의 차이가 큰 경우인 (50, 500)일 때는 logit-Simes 방법의 검정력이 가장 높게 나타났다. 구체적으로 살펴보면 제 2상b/제 3상을 inverse χ^2 방법으로 결합하는 경우는 Simes 방법이 검정력이 가장 높았고 제 2상b/제 3상을 weighted inverse normal 방법으로 결합하는 경우 또한 Simes 방법의 검정력이 가장 높게 나타났고았고, 제 2상b/제 3상을 logit 방법으로 결합하는 경우는 Simes 방법의 검정력이 가장 높았다. 또한 제 2상b/제 3상 표본의 크기가 같은 경우인 $(100, 100)$일 때는 logit-Holm 방법의 검정력이 가장 높게 나타났다. 구체적으로 살펴보면 제 2상b/제 3상을 inverse χ^2 방법으로 결합하는 경우는 BH 방법이 검정력이 가장 높았고 제 2상b/제 3상을 weighted inverse normal 방법으로 결합하는 경우는 BH 방법의 검정력이 가장 높았고, 제 2상b/제 3상을 logit 방법으로 결합하는 경우 또한 Holm 방법

Table 5.2. Comparison of power ($\alpha = 0.025, \sigma = 3$) under effect size $(0, 0, 0, 1)$

n_1	n_2		Power				
			iv.chi	wiv.nor	logit	CON	TSE
50	500	Bonferroni	0.871	0.838	0.936	0.861	0.863
		Hochberg	0.884	0.856	0.911		
		Simes	0.911	0.871	0.948		
		Hommel	0.891	0.861	0.923		
		Holm	0.879	0.849	0.942		
		BH	0.899	0.865	0.931		
100	100	Bonferroni	0.633	0.721	0.764	0.686	0.758
		Hochberg	0.652	0.717	0.679		
		Simes	0.649	0.689	0.685		
		Hommel	0.658	0.743	0.688		
		Holm	0.648	0.730	0.775		
		BH	0.665	0.752	0.701		
100	500	Bonferroni	0.893	0.927	0.886	0.759	0.891
		Hochberg	0.901	0.916	0.892		
		Simes	0.907	0.920	0.897		
		Hommel	0.908	0.921	0.875		
		Holm	0.896	0.929	0.890		
		BH	0.911	0.925	0.888		

iv.chi = inverse 방법; wiv.nor = weighted inverse normal 방법; logit = logit 방법; CON = conventional, TSE = Thall, Simon, and Ellenberg 방법

의 검정력이 가장 높게 나타났다.

또한 제 2상b/제 3상 표본의 차이가 작은 경우인 (100, 500)일 때는 logit-Simes 방법의 검정력이 가장 높게 나타났다. 구체적으로 살펴보면 제 2상b/제 3상을 inverse χ^2 방법으로 결합하는 경우는 BH 방법이 검정력이 가장 높았고 제 2상b/제 3상을 weighted inverse normal 방법으로 결합하는 경우는 Holm 방법의 검정력이 가장 높았고, 제 2상b/제 3상을 logit 방법으로 결합하는 경우 또한 Simes 방법의 검정력이 가장 높았음을 알 수 있다.

5.2.2. 치료효과가 $(0.5, 0.5, 0.5, 1)$이고 표준편차를 3으로 정의한 경우 (Table 5.3) 제 2상b/제 3상 표본의 차이가 큰 경우인 $(50, 500)$일 때는 weighted inverse normal-BH 방법의 검정력이 가장 크게 나타남을 확인 할 수 있었다. 구체적으로 살펴보면 제 2상b/제 3상을 inverse χ^2 방법으로 결합하는 경우는 Simes 방법이 검정력이 가장 높았고 제 2상b/제 3상을 weighted inverse normal 방법으로 결합하는 경우 BH 방법의 검정력이 가장 높았고, 제 2상b/제 3상을 logit 방법으로 결합하는 경우는 Holm 방법의 검정력이 가장 높게 나타났다.

또한 제 2상b/제 3상 표본의 크기가 같은 경우인 (100, 100)일 때는 TSE 방법의 검정력이 가장 높게 나타났다. 구체적으로 살펴보면 제 2상b/제 3상을 inverse χ^2 방법으로 결합하는 경우는 Simes 방법이 검정력이 가장 높았고 제 2상b/제 3상을 weighted inverse normal 방법으로 결합하는 경우는 BH 방법의 검정력이 가장 높았고, 제 2상b/제 3상을 logit 방법으로 결합하는 경우 또한 Simes 방법의 검정력이 가장 높게 나타났다.

또한 제 2상b/제 3상 표본의 차이가 작은 경우인 (100, 500)일 때는 inver normal-Simes 방법의 검정력이 가장 높았음을 알 수 있다. 구체적으로 살펴보면 제 2상b/제 3상을 inverse χ^2 방법으로 결합하는 경

Table 5.3. Comparison of power ($\alpha = 0.025, \sigma = 3$) under effect size $(0.5, 0.5, 0.5, 1)$

n_1	n_2		Power				
			iv.chi	wiv.nor	logit	CON	TSE
50	500	Bonferroni	0.654	0.759	0.884	0.602	0.883
		Hochberg	0.679	0.870	0.722		
		Simes	0.730	0.765	0.719		
		Hommel	0.691	0.905	0.849		
		Holm	0.672	0.809	0.891		
		BH	0.715	0.913	0.865		
100	100	Bonferroni	0.539	0.708	0.680	0.528	0.773
		Hochberg	0.575	0.727	0.645		
		Simes	0.634	0.687	0.702		
		Hommel	0.602	0.745	0.666		
		Holm	0.554	0.718	0.629		
		BH	0.610	0.768	0.683		
100	500	Bonferroni	0.750	0.797	0.851	0.746	0.900
		Hochberg	0.768	0.869	0.786		
		Simes	0.800	0.905	0.874		
		Hommel	0.771	0.886	0.828		
		Holm	0.763	0.815	0.859		
		BH	0.775	0.893	0.846		

iv.chi = inverse 방법; wiv.nor = weighted inverse normal 방법; logit = logit 방법; CON = conventional, TSE = Thall, Simon, and Ellenberg 방법

우는 Simes 방법이 검정력이 가장 높았고 제 2상b/제 3상을 weighted inverse normal방법으로 결합하는 경우는 Simes 방법의 검정력이 가장 높았고, 제 2상b/제 3상을 logit방법으로 결합하는 경우 Simes 방법의 검정력이 가장 높게 나타났다.

5.2.3. 치료효과가$(0, 0, 0, 1)$이고 표준편차를 5로 정의한 경우 (Table 5.4) 제 2상b/제 3상 표본의 차이가 큰 경우인 $(50, 500)$일 때는 TSE 방법의 검정력이 가장 높게 나타났다. 구체적으로 살펴보면 제 2상b/제 3상을 inverse χ^2 방법, weighted inverse normal 방법으로 결합하는 경우 모두 Simes 방법의 검정력이 가장 높았고 logit 방법으로 결합하는 경우는 Bonferroni 방법의 검정력이 가장 높았다.

또한 제 2상b/제 3상 표본의 크기가 같은 경우인 (100, 100)일 때는 logit-Holm 방법의 검정력이 가장 높게 나타났다. 구체적으로 살펴보면 제 2상b/제 3상을 inverse χ^2 방법으로 결합하는 경우는 BH 방법이 검정력이 가장 높았고 제 2상b/제 3상을 weighted inverse normal 방법, logit 방법으로 결합하는 경우 모두 holm 방법의 검정력이 가장 높게 나타났다.

또한 제 2상b/제 3상 표본의 차이가 작은 경우인 (100, 500)일 때는 TSE 방법의 검정력이 가장 높았다. 구체적으로 살펴보면 제 2상b/제 3상을 inverse χ^2 방법으로 결합하는 경우는 holm 방법이 검정력이 가장 높았고 제 2상b/제 3상을 weighted inverse normal 방법으로 결합하는 경우는 Simes 방법의 검정력이 가장 높았고, 제 2상b/제 3상을 logit 방법으로 결합하는 경우 또한 Simes 방법의 검정력이 가장 높게 나타났다.

5.2.4. 치료효과가$(0.5, 0.5, 0.5, 1)$이고 표준편차를 5로 정의한 경우 (Table 5.5) 제 2상b/제 3상 표본의 차이가 큰 경우인 $(50, 500)$일 때는 weighted inverse normal-BH 방법의 검정력이 가장 높게 나

Table 5.4. Comparison of power ($\alpha = 0.025, \sigma = 5$) under effect size $(0, 0, 0, 1)$

n_1	n_2		Power				
			iv.chi	wiv.nor	logit	CON	TSE
50	500	Bonferroni	0.557	0.623	0.616	0.594	0.693
		Hochberg	0.589	0.620	0.566		
		Simes	0.606	0.642	0.513		
		Hommel	0.596	0.627	0.582		
		Holm	0.556	0.625	0.602		
		BH	0.599	0.639	0.580		
100	100	Bonferroni	0.433	0.627	0.673	0.592	0.673
		Hochberg	0.486	0.522	0.555		
		Simes	0.498	0.524	0.570		
		Hommel	0.504	0.530	0.592		
		Holm	0.449	0.641	0.687		
		BH	0.519	0.540	0.614		
100	500	Bonferroni	0.678	0.742	0.785	0.744	0.829
		Hochberg	0.723	0.786	0.791		
		Simes	0.761	0.797	0.812		
		Hommel	0.726	0.750	0.792		
		Holm	0.782	0.769	0.786		
		BH	0.685	0.774	0.777		

iv.chi = inverse 방법; wiv.nor = weighted inverse normal 방법; logit = logit 방법; CON = conventional, TSE = Thall, Simon, and Ellenberg 방법

타났다. 구체적으로 살펴보면 제 2상b/제 3상을 inverse χ^2 방법, weighted inverse normal 방법으로 결합하는 경우 모두 BH방법의 검정력이 가장 높았고, 제 2상b/제 3상을 logit 방법으로 결합하는 경우는 Simes 방법의 검정력이 가장 높게 나타났다.

또한 제 2상b/제 3상 표본의 크기가 같은 경우인 $(100, 100)$일 때는 TSE 방법의 검정력이 가장 높게 나타났다. 구체적으로 살펴보면 제 2상b/제 3상을 inverse χ^2 방법으로 결합하는 경우는 Simes 방법이 검정력이 가장 높았고 제 2상b/제 3상을 weighted inverse normal 방법으로 결합하는 경우는 BH 방법의 검정력이 가장 높았고, 제 2상b/제 3상을 logit 방법으로 결합하는 경우는 holm 방법의 검정력이 가장 높았음을 알 수 있다.

또한 제 2상b/제 3상 표본의 차이가 작은 경우인 $(100, 500)$일 때는 weighted inverse normal-Simes 방법의 검정력이 가장 높았음을 알 수 있다. 구체적으로 살펴보면 제 2상b/제 3상을 inverse χ^2 방법, weighted inverse normal 방법으로 결합하는 경우 모두 Simes 방법의 검정력이 가장 높았고 logit 방법으로 결합하는 경우는 Bonferroni 방법의 검정력이 가장 높게 나타났다.

6. 결론

임상시험에서 표본의 크기를 결정하는 일은 임상 연구에 있어서 매우 중요한 일이다. 모의실험 결과를 보면 알 수 있듯이 제 2상b/제 3상 표본의 크기 조합에 따라 그리고 여러 분산의 크기에 따라 검정력 결과에 차이가 있었음을 알 수 있다. 즉, 각 상황에 따라 검정력이 높은 방법들을 선택하게 되면, 같은 검정력으로 적은 환자수를 이용하여 임상시험을 진행할 수 있을 것이다. 실제 임상시험을 진행 할 때 표본의 차이가 클 때에는 본 논문 모의실험의 $(50, 500)$의 경우를, 표본의 크기가 같을 때에는 $(100, 100)$의

Table 5.5. Comparison of power ($\alpha = 0.025, \sigma = 5$) under effect size $(0.5, 0.5, 0.5, 1)$

n_1	n_2		Power				
			iv.chi	wiv.nor	logit	CON	TSE
50	500	Bonferroni	0.617	0.690	0.677	0.654	0.733
		Hochberg	0.638	0.705	0.643		
		Simes	0.630	0.733	0.713		
		Hommel	0.626	0.714	0.657		
		Holm	0.623	0.694	0.685		
		BH	0.644	0.735	0.670		
100	100	Bonferroni	0.543	0.572	0.647	0.551	0.653
		Hochberg	0.552	0.600	0.590		
		Simes	0.573	0.601	0.630		
		Hommel	0.559	0.618	0.627		
		Holm	0.551	0.582	0.652		
		BH	0.560	0.621	0.606		
100	500	Bonferroni	0.502	0.718	0.694	0.516	0.763
		Hochberg	0.548	0.746	0.659		
		Simes	0.601	0.834	0.682		
		Hommel	0.568	0.768	0.678		
		Holm	0.524	0.733	0.609		
		BH	0.582	0.782	0.635		

iv.chi = inverse 방법; wiv.nor = weighted inverse normal 방법; logit = logit 방법; CON = conventional, TSE = Thall, Simon, and Ellenberg 방법

경우를, 표본의 차이가 작을 때에는 $(100, 500)$의 경우를 따라가면 될 것이다. 분산을 같이 고려하는 경우에는, 분산이 작은 경우에는 $\sigma = 3$를 분산이 큰 경우에는 본 논문 모의실험의 $\sigma = 5$를 따라가면 될 것이다. 최고효과 용량군 이외의 다른 용량군의 효과를 없다고 가정했을 때에는 $(0, 0, 0, 1)$의 경우를, 최고효과 용량군 이외의 다른 용량군의 효과도 나타날 것을 가정 했을 때에는 $(0.5, 0.5, 0.5, 1)$의 경우를 따라가게 되면 이에 따른 방법들을 선택했을 때 같은 검정력으로 적은 피험자수를 이용하여 임상시험을 진행할 수 있을 것이다. 본 논문에서는 제시되지 않았지만 다양한 치료효과에 따라서, 또한 자료가 정규성을 만족하지 못한 경우에는 본 논문에서 비교한 20가지 방법 간에 검정력이 다르게 나타날 수 있다. 따라서 이들에 대한 연구가 이루어진다면 더욱 체계적인 가이드라인을 제시할 수 있을 것이다.

현재 국내 임상에서 적응적 심리스 디자인 방법은 많이 사용되고 있지 않다. 하지만 적응적 디자인의 한 가지 방법인 적응적 심리스 디자인 뿐 아니라 적응적 디자인에 대해 더욱 연구하여 국내 임상에 적용시킨다면 적은 환자수를 이용하여 비용을 줄일 수 있을 뿐 아니라, 임상시험을 진행함에 있어서 시간이 상당히 단축될 수 있을 것이다.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society*, **57**, 289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency, *Annals of Statistics*, **29**, 1165–1188.

Bauer, P. and Kohne, K. (1994). Evaluation of experiments with adaptive interim analyses, *Biometrics*, **50**, 1029–1041.

Cui, L., Hung, H. M. J., and Wang, S. J. (1999). Modification of sample size in group sequential clinical trials, *Biometrics*, **55**, 853–857.

Denne, J. S. (2001). Sample size recalculation using conditional power, *Statistics in Medicine*, **20**, 2645–2660.

Dunn, O. J. (1961). Multiple comparisons among means, *Journal of the American Statistical Association*, **56**, 54–62.

Fisher, L. D. (1998). Self-designing clinical trials, *Statistics in Medicine*, **17**, 1551–1562.

Fisher, R. A. (1932). *Statistical Methods for Research Workers* (4th ed), Oliver and Boyd, London.

Friede, T., Parsons, N., Stallard, N., Todd, S., Valdes-Marquez, E., Chataway, J., and Nicholas, R. (2011). Designing a seamless phase Ⅱ/Ⅲ clinical trial using early outcomes for treatment selection: an application in multiple sclerosis, *Statistics in Medicine*, **30**, 1528–1540.

George, E. O. (1977). Combining Independent One-sided and Two-sided Statistical Tests - Some Theory and Applications. Unpublished Doctoral Dissertation, University of Rochester.

Hochberg, Y. (1988) A shaper Bonferroni procedure for multiple tests of significance, *Biometrika*, **75**, 800–803

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure, *Scandinavian Journal of Statistics*, **6**, 65–70.

Hommel, G. (1986). Multiple test procedures for arbitrary dependence structures, *Metrika*, **33**, 321–336.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test, *Biometrika*, **75**, 383–386.

Jennison, C. and Turnbull, B. W. (2007). Adaptive seamless designs: selection and prospective testing of hypotheses, *Journal of Biopharmaceutical Statistics*, **17**, 1135–1161

Kunz, C., Friede, T., Parsons, N., Todd, S., and Stallard, N. (2015). A comparison of methods for treatment selection in seamless Phase II/III clinical trials incorporating information on short-term endpoints, *Journal of Biopharmaceutical Statistics*, **25**, 170–189.

Lehmacher, W. and Wassmer, G. (1999). Adaptive sample size calculation in group sequential trials, *Biometrics*, **55**, 1286–1290.

Mosteller, F. and Bush, R. R. (1954). Selected quantitative techniques. In: Lindzey, G., ed. Handbook of Social Psychology, Vol. 1. Cambridge, MA: Addison-Wesley, 289–334.

Sarkar, S. (1998). Some probability inequalities for ordered MTP2 random variables: a proof of the Simes conjecture, *Annals of statistics*, **26**, 494–504.

Sarkar, S. and Chang, C. (1997). The Simes method for multiple hypothesis testing with positively dependent test statistics, *Journal of American Statistical Association*, **92**, 1601–1608.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance, *Biometrika*, **79**, 751–754.

Stallard, N. (2010). A confirmatory seamless phase Ⅱ/Ⅲ clinical trial design incorporating short-term endpoint information, *Statistics in Medicine*, **29**, 959–971.

Thall, P. F., Simon, R., and Ellenberg, S. S. (1988). Two-stage selection and testing designs for comparative clinical trials, *Biometrika*, **75**, 303–310.

심리스 제2상/제3상 임상시험에서 다중가설검정방법과 결합검정방법의 비교연구

한송[a] · 유한나[a] · 이재원[a,1]

[a]LG화학㈜, [b]부산외국어대학교 컴퓨터소프트웨어학부, [c]고려대학교 통계학과

(2018년 5월 11일 접수, 2018년 12월 2일 수정, 2018년 12월 5일 채택)

요 약

최근에 제안된 심리스(seamless) 제 2상/제 3상 임상시험 디자인은 기존의 임상시험 디자인들과 비교하여 피험자수를 줄일 수 있을 뿐만 아니라 임상 개발 시간을 단축시킬 수 있다는 장점을 가지고 있어 임상시험연구자들의 많은 관심을 끌고 있다. 또한 제 3상 시험을 단독으로 진행 하였을 때보다 더 높은 검정력을 가질 수 있으므로 임상시험에서 매우 효율적이라 말할 수 있다. 본 논문에서는 제 2상에서 최고효과 용량군을 선정하기 위한 여러 가지 다중가설검정방법들을 제시하고 제 2상에서 최고효과 용량군을 선정한 후에 제 2상과 제 3상을 결합하는 여러 가지 유의확률 결합검정방법들을 제시하였다. 또한 모의실험을 통해서 심리스 제 2상/제 3상 임상설계가 적용되었을 때 여러 가지 방법들을 비교함으로써, 제 2상/제 3상 표본의 크기 조합이나 분산의 크기가 다른 여러 가지 상황에서 가장 적절한 방법을 선택하는 가이드라인을 제시하고자 한다.

주요용어: 적응적 디자인, 심리스 디자인, 다중가설검정, 결합검정, 제 2상시험, 제 3상시험

이 논문은 2017년도 정부(교육부)와 교육과학기술부 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2017R1D1A1B03028279, No.2017R1C1B5076671).

[1]교신저자: (02841) 서울시 성북구 안암로 145, 고려대학교 통계학과. E-mail: jael@korea.ac.kr