응용통계연구 제 **21**권 *1*호, *2008*년, *pp. 35–51*

제 3상 임상시험에서 여러 형태 반응변수의 다변량 검정법인 P값 병합법

김수영[1] 송혜향[2]

요 약

제 3상 임상시험에서 치료효과가 여러 반응변수(endpoints)로 측정될 때, 이들 반응변수가 대등하게 중요하여 주요 반응변수(primary endpoint)를 선택할 수 없는 상황이 발생할 수 있다. O′Brien (1984)은 이들 반응변수 모두를 종합하여 치료효과에 대한 단측검정(one-tailed testing) 통계량으로서 반응변수가 연속형(continuous) 자료로 측정되었을 때 Ordinary Least Square(OLS)와 Generalized Least Square(GLS) 검정 통계량을 제시하였다. Pocock 등 (1987)은 여러 형태, 즉 연속형, 이산형(binary), 생존(survival) 자료의 반응변수를 함께 분석할 수 있음을 언급하고 있으나 실제로 이와 같이 여러 형태의 반응변수 병합에 대한 문제점을 설명하거나 구체적으로 모의실험으로서 이러한 경우의 OLS와 GLS 통계량의 효율성을 알아보지는 않았다. 본 논문에서는 특히 여러 형태의 반응변수를 종합하여 치료효과에 대한 결론을 내리는데 P값의 병합 통계량을 제안하며, 이때 각 반응변수의 치료효과에 대한 검정 결과인 P값은 서로 상관성이 존재하는 P값이다. OLS 및 GLS 검정 통계량보다 장점을 지닌 P값의 병합방법 중, 방법 F와 G는 제 1종 오류가 유의수준보다 커서 검정의 결론이 잘못 내려질 수 있는 경우가 있고 방법 B는 제 1종의 오류가 잘 통제되고 또한 효율성이 높은 것으로 나타났다.

주요용어: 다중 반응변수, 독립된 P값의 병합, 연관된 P값의 병합, OLS, GLS.

1. 서론

독립된 두 군의 치료효과를 비교하는 임상시험의 현장에서 각 환자에게서 둘 이상의 반응변수가 측정되어 다중 반응변수(multivariate endpoints)를 분석해야 하는 경우가 흔치 않게 발생한다. 예를 들어서, 미국 식약청(U.S. FDA)이 전립선 비대증(prostatic hyperplasia) 환자에 대한 치료로 위약(placebo)과 비교한 적정량 5mg의 finasteride를 승인했던 임상시험을 살펴보면, 총 배뇨증상점수(urinary symptom score), 요속검사(urinary-flow rate test), 전립선 전체부피(total prostate volume), 잔뇨량(residual volume), 전립선 특이항원(prostate specific antigen) 등의 반응변수로 치료효과를 평가하였다. 특히, 처음 세 반응변수의 통계적 유의성을 근거로 그 효과를 증명하였다(Chow와 Liu, 2003). 이러한 다중 반응변수에 관

1) (137-701) 서울시 서초구 반포동 505, 가톨릭대학교 의학통계학과, 석사과정.
E-mail: kimsuyoung@catholic.ac.kr
2) (137-701) 교신저자. 서울시 서초구 반포동 505, 가톨릭대학교 의학통계학과, 교수.
E-mail: hhsong@catholic.ac.kr

한 분석 시 가장 쉬운 접근은 주관적 판단 혹은 연구목적 등에 비추어 보아 주요 반응변수(primary endpoint)와 보충 반응변수(subsidiary endpoint)를 우선적으로 선택하여 각 변수에 대해 단변량 분석(univariate analysis)을 실시하는 것이다. 그러나 Pocock 등 (1987)이 언급하였듯이 유의하지 않은 반응변수를 제외함으로써 인위적으로 유의한 결론을 얻는 경우도 발생하며, 특히 자료의 부분적인 결론을 제시하는데 그치게 되는 단점이 있다. 이러한 맥락에서 여러 반응변수가 대등한 중요성으로 분석되는 방법이 고려되어야 한다.

Hotelling의 T^2 통계량은 다중 반응변수 모두에 근거한 치료효과를 검정하지만, O′Brien (1984)과 Meier (1975)가 지적하였듯이 총 K개의 반응변수의 각 반응변수 효과크기(effect sizes)인 $\delta_k = \mu_{1k} - \mu_{2k}$가 0과 다른가에 대한 대립가설의 검정이므로 치료효과를 의미하는 $\delta_k > 0$뿐만 아니라 치료가 유해한 경우를 의미하는 $\delta_k < 0$도 함께 기각하여 검정력이 낮을 수 밖에 없다.

여러 반응변수의 효과크기의 양방향성 문제를 피할 수 있는 접근으로서 단측 검정 후 Bonferroni 수정법이 있다. K개의 반응변수가 서로 독립이라 가정할 때, 각 반응변수의 단측 검정의 결과로 얻은 P값 중 가장 작은 값에 K를 곱한 값을 유의수준 α와 비교하여 검정결론을 내리는 방법이다. 적용이 간단하고 이해하기 쉬운 이 방법은 각 반응변수에 대해 단변량 분석을 실시하게 되어 여러 횟수의 유의성 검정으로 증가되는 제 1종 오류를 보정한다는 장점이 있지만 반응변수의 상관성(ρ)이 증가할수록, 특히 $\rho \geq 0.5$인 경우에 Bonferroni 수정법이 매우 보수적임을 Pocock 등 (1987)은 논문의 표 1에서 밝히고 있다. 이는 Gupta (1963)가 제시한 다변량 정규(multivariate normal) 분포의 표준화된 정규편차 중, 최대값의 확률 분포에 근거한 α'값과 Bonferroni 수정에 의한 $\alpha'' = \alpha/K$값을 비교함으로써 증명하였다. Bonferroni 수정법의 또 다른 단점은 여러 반응변수에 해당하는 P값 중 가장 작은 P값에 근거하여 결론을 내린다는 것이며, 따라서 K개의 반응변수 중 어떤 하나에서 뚜렷한 치료효과가 존재한다는 대립가설 하에서 Bonferroni 수정법은 검정력이 높지만, 모든 반응변수에서 뚜렷하지는 않으나 일률적으로 어느 정도의 치료효과가 존재하는 대립가설 하에서는 검정력이 높지 않다.

O′Brien (1984)은 모든 반응변수를 종합하여 두 군의 치료효과를 비교하는 방법으로서 OLS와 GLS 검정 통계량을 제안하였다. 이 검정 통계량은 방향성을 갖는 대립가설 하에서 검정력이 높은 장점을 지니는 반면, 저자 등은 연속형 자료에 대한 적용만을 고려하였다. 이에 Pocock 등 (1987)은 연속형, 이산형, 생존 자료와 같이 여러 자료 형태의 반응변수가 함께 제시된 경우에 GLS 통계량이 더욱 적절함을 언급하였다. GLS 통계량은 구체적으로 각 단변량 분석에서 계산된 검정 통계량을 여러 반응변수간의 공분산으로 가중시켜 합산한 통계량이다. 그러나 현실적으로 형태가 서로 다른 반응변수의 단변량 분석을 통해 얻은 검정 통계량은 매우 다양하여 병합에 있어 어려움이 있다. 본 논문에서는 여러 형태의 다중 반응변수의 분석방법으로서 앞에서 언급한 여러 단점과 어려움이 없다고 판단되는 단변량 분석의 결과인 P값을 병합하는 방법을 제안한다. 임상시험의 결론이 하나의 P값으로 요약됨이 중요하며, 여러 반응변수에 의한 서로 상관성이 존재하는 P값의 병합방법으로서 여러 이론에 근거한 방법이 제시될 수 있다. 앞에서 언급한 OLS와 GLS의 방법보다 검정력이 높은 P값 병합방법을 본 논문에서 제안할 것이다.

2. 방법

O′Brien (1984)과 Pocock 등 (1987)의 OLS와 GLS 통계량을 설명하며, 다음으로 상관성이 있는 P값을 병합하는 방법들을 제안한다. 이미 독립적인 P값의 병합방법으로 Fisher (1950)의 방법을 비롯하여 매우 많은 방법들이 제안되었으나, 상관성이 있는 P값의 병합방법은 Brown (1975)과 Zaykin 등 (2002)에 불과하다. 이들의 방법과 다른, 임상시험의 경우에 합당한 P값의 병합방법을 설명할 것이다.

2.1. OLS와 GLS 검정 통계량

2.1.1. OLS

OLS 통계량을 설명키 위해 연속형 자료로 출발하며 이 연속형 자료 $Y_{ijk}(i = 1, 2,\ j = 1, \ldots, n_i,\ k = 1, \ldots, K)$는 i번째 집단에 속한 j번째 개체의 k번째 반응변수를 나타낸다. 서로 다른 개체의 자료는 독립이지만 동일 개체의 여러 반응변수는 연관되어 있고 다변량 정규 분포하며 동일 공분산 행렬을 가졌음을 가정한다. 즉, $\boldsymbol{Y}_{ij} \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$이며, 여기서 $\boldsymbol{Y}_{ij} = (Y_{ij1}, \ldots, Y_{ijk})^T$, $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{ik})^T$ $(i = 1, 2)$이고 $\Sigma_{kk} = \sigma_k^2$로서 각 반응변수의 분산은 서로 다르며 변수간 공분산 $\Sigma_{kk'}$이 존재함을 가정한다. OLS 통계량은 각 반응변수를 우선 표준화시키는 것으로 출발한다. 즉, 표준화된 반응변수는 다음과 같다.

$$Y_{ijk}^* = \frac{Y_{ijk} - \bar{Y}_{..k}}{S_{..k}}. \tag{2.1}$$

여기서 $\bar{Y}_{..k}$와 $S_{..k}^2$는 각 반응변수에서의 평균과 합병분산(pooled variance)의 추정값으로 다음과 같다.

$$\bar{Y}_{..k} = \frac{n_1\bar{Y}_{1.k} + n_2\bar{Y}_{2.k}}{n_1 + n_2}, \tag{2.2}$$

$$S_{..k}^2 = \frac{(n_1 - 1)S_{1.k}^2 + (n_2 - 1)S_{2.k}^2}{n_1 + n_2 - 2}. \tag{2.3}$$

여기서 $\bar{Y}_{i.k}$는 각 군의 평균으로서 $\bar{Y}_{i.k} = (1/n_i)\sum_{j=1}^{n_i} Y_{ijk}$로 정의된다.

이제 표준화된 Y_{ijk}^*에 근거하여 각 반응변수에 대해 치료효과가 있는가에 대한 t 통계량을 구할 수 있으며 Lehmacher 등 (1991)의 기호를 따라서 이를 편의상 Z로 표현한다.

$$Z_k = \frac{\bar{Y}_{1.k}^* - \bar{Y}_{2.k}^*}{\sqrt{(1/n_1) + (1/n_2)}}. \tag{2.4}$$

여기서 $\bar{Y}_{i.k}^* = (1/n_i)\sum_{j=1}^{n_i} Y_{ijk}^*$이다.

OLS 통계량은 이와 같이 구한 각 반응변수의 Z_k를 선형 합산시켜 구한 것이다. 이제 위에서 언급한 K개의 Z_k들로 이루어진 $K \times 1$벡터를 $\mathbf{Z} = (Z_1, \ldots, Z_k)^T$로, Y_{ijk}^*로부터 추정된 $K \times K$ 의 상관행렬(correlation matrix)을 $\hat{\mathbf{R}}$으로, 각 원소가 1로 이루어진 $K \times 1$ 벡터를 $\mathbf{J} = (1, \ldots, 1)^T$로 정의한다. OLS 통계량은 다음과 같다.

$$T_{\mathrm{OLS}} = \frac{\mathbf{J}^T\mathbf{Z}}{\sqrt{\mathbf{J}^T\hat{\mathbf{R}}\mathbf{J}}}. \tag{2.5}$$

이 통계량 T_{OLS}의 분자를 살펴보면 Z_k를 단순 합산(unweighted sums)시키고 있어 후에 설명하는 GLS의 가중 합산(weighted sums)과 다르다.

OLS 통계량 T_{OLS}는 각 반응변수별로 자료로부터 추정한 합병분산으로 표준화시킨 Y^*_{ijk}에 근거한 Z_k를 합산시키므로 귀무가설 하에서도 일반적으로 정확히 t 분포하지 않으며, 또한 대립가설 하에서도 정확히 비중심 t 분포하지 않는다. 귀무가설 하에서 특별한 경우에 한하여 T_{OLS}는 정확히 t 분포하는데 이러한 경우란 모든 반응변수가 동일 측정값 범위(same scale)를 가지면서 동일 분산(즉, $\sigma_k = \sigma$, $k = 1, \ldots, K$)일 때 Y_{ijk}를 $S_{\cdot\cdot k}$가 아닌 상수인 $S_{\cdots}$로 나누어 표준화시키는 경우이다. O′Brien (1984)에서 귀무가설 하에서 T_{OLS}가 자유도 $n_1 + n_2 - 2$의 근사 t 분포함을 이용하여 검정하고 있으며 본 논문에서도 이를 그대로 채택한다.

단순하게 계산되는 OLS 통계량은 수치적으로 안정하다는 장점이 있지만 대부분의 경우에 여러 반응변수간에 동일 상관성을 지니고 있지 않기 때문에 이러한 상관성을 가중으로 두고 합산한 GLS 통계량이 더 큰 검정력을 가지게 된다.

2.1.2. GLS

GLS 통계량은 각 표준화된 정규편차 Z_k에 표준화된 Y^*_{ijk}로부터 추정된 상관행렬의 역(inverse)을 가중으로 합산한 통계량이며, 따라서 각 반응변수의 기여 정도가 다르게 된다. 즉, GLS 통계량은 다음과 같다.

$$T_{\mathrm{GLS}} = \frac{\mathbf{J}^T \hat{\mathbf{R}}^{-1} \mathbf{Z}}{\sqrt{\mathbf{J}^T \hat{\mathbf{R}}^{-1} \mathbf{J}}}. \tag{2.6}$$

이 T_{GLS} 역시 O′Brien (1984)은 귀무가설 하에서 자유도 $n_1 + n_2 - 2$의 근사 t 분포함을 이용하여 검정하고 있으며 본 논문에서도 이를 그대로 채택한다. 반응변수의 수가 단지 두 개인 경우에는 OLS 통계량과 GLS 통계량이 일치하게 되지만, 반응변수가 셋 이상인 경우에는 GLS 통계량의 검정력이 더욱 크다고 알려져 있다 (Lehmacher 등, 1991).

OLS와 GLS 통계량은 표준화된 반응변수에 근거하여 계산되는데 여러 반응변수가 다양한 자료 형태인 경우, 특히 이산형이나 순서형(ordinal), 생존 자료를 동일 범위를 가지도록 표준화시키는 일이 자명치 않다. 여러 반응변수를 각 반응변수별 순위로 변환한다고 해도 문제는 해결되지 않는다. 따라서 이러한 다양한 반응변수의 분석에는 다음 절에서 소개하는 P값의 병합이 더욱 바람직하다.

2.2. P값의 병합

2.2.1. 여러 형태 반응변수로부터 P값의 계산 과정

임상시험에서 수집되는 반응변수의 형태로는 연속형 자료로서 정규 또는 비정규 분포 자료가 있고, 이외에도 이산형, 순서형 및 생존 자료가 있다. 치료효과가 존재한다는 대립가설 하에서 각 형태의 반응변수에 대해 단측 검정의 P값을 구하는 과정을 되도록 짧게 서술한다.

우선 k번째 반응변수의 연속형 자료가 정규 또는 근사 정규 분포하는 경우 다음의 T통계량이 자유도 n_1+n_2-2인 t 분포함을 이용하여 검정한다.

$$T_k = \frac{\bar{Y}_{1.k} - \bar{Y}_{2.k}}{S_{..k}\sqrt{(1/n_1)+(1/n_2)}}. \tag{2.7}$$

여기서 $S^2_{..k}$는 두 군의 합병추정량이며, P값은 $P_k = \Pr(T \geq T_k)$에 의해 구한다.

반응변수가 정규 분포하지 않는 경우, 비모수 방법인 윌콕슨(Wilcoxon) 순위합 검정으로 두 군을 비교한다. 윌콕슨 순위합 검정은 두 군 자료를 순위로 변환한 후 한 군의 순위합 W_k를 통계량으로 사용하며 표본수 n_1이나 n_2가 20보다 크면 다음의 통계량이 정규 근사함을 이용하여 검정한다.

$$Z_k = \frac{W_k - E(W_k)}{\mathrm{Var}(W_k)}. \tag{2.8}$$

여기서 $E(W_k)$와 $\mathrm{Var}(W_k)$는 각각 W_k의 평균과 분산으로서 다음과 같다.

$$E(W_k) = \frac{n_1(n_1+n_2+1)}{2}, \qquad \mathrm{Var}(W_k) = \frac{n_1 n_2}{12}(n_1+n_2+1). \tag{2.9}$$

동점 자료가 있다면, 위의 분산 공식은 보정 되어야 한다 (Hollander와 Wolfe, 1999). P값은 $P_k = \Pr(Z \geq Z_k)$에 의해 구한다.

이산형 반응변수인 경우에는 비율차 검정으로 표본수가 작지 않다면 비율차의 근사 정규 분포함을 이용한다. $\hat{P}_{1k}$와 $\hat{P}_{2k}$가 두 군의 사건 발생의 확률이고, $\hat{P}_p$는 모비율의 합병 추정량일 때, 비율차 검정 통계량은

$$Z_k = \frac{\hat{P}_{1k} - \hat{P}_{2k}}{\sqrt{\hat{P}_p(1-\hat{P}_p)(\frac{1}{n_1}+\frac{1}{n_2})}} \tag{2.10}$$

이며, P값은 $P_k = \Pr(Z \geq Z_k)$에 의해 구한다.

절단 자료(censored data)가 있는 생존 자료의 경우는 Gehan (1965)의 통계량

$$W_s = \sum_{j=1}^{n_1} |u_j| \tag{2.11}$$

을 이용한다. 여기서 u_j는 Mantel (1966)의 스코아로서 첫번째 집단의 j번째 관측값보다 큰 두번째 집단의 관측값에서 첫번째 집단의 j번째 관측값보다 작은 두번째 집단의 관측값을 뺀 값이다. W_s의 기대값은 0이고, 분산은

$$\mathrm{Var}(W_s) = \frac{n_1 n_2}{(n_1+n_2)(n_1+n_2-1)} \sum_{j=1}^{n_1+n_2} (u_j^*)^2 \tag{2.12}$$

이며, u_j^*는 혼합 표본의 j번째 관측값보다 큰 관측값에서 j번째 관측값보다 작은 관측값을 뺀 값으로 정의한다. n_1, n_2가 충분히 크다면 통계량 $Z_k = W_s/\sqrt{\mathrm{Var}(W_s)}$이 근사 정규 분포함을 이용하여 검정하며, P값은 $P_k = \Pr(Z \geq Z_k)$에 의해 구한다.

이와 같이 여러 형태 반응변수로부터 구한 P값은 상관성이 있는 P값으로서 $K \times 1$ 벡터 $\mathbf{P} = (P_1, \ldots, P_k)^T$로 표현한다. 이제, 여러 P값을 병합하여 하나의 종합된 P값(global P-value)을 제시하는 방법을 설명한다.

2.2.2. 독립된 P값의 병합

독립된 P값의 병합으로 여러 가지 방법들이 제시되었으며 본 논문에서는 가장 널리 쓰이는 Fisher (1950)의 방법과 Good (1958)의 조화평균(harmonic mean)을 이용한 방법을 고려한다. 본 논문의 주 목적은 독립된 P값의 병합이 아니라 연관된 P값의 병합으로서 이를 곧 설명하게 된다.

Fisher (1950)의 P값 병합방법은 Littell과 Folk (1971, 1973)가 지적하였듯이 독립된 검정으로부터 구해진 P값의 병합방법 가운데 효율성이 높다고 알려져 있다. P값은 일양 분포하므로, $\Pr(-2\log_e P > t) = \Pr(P < e^{-t/2}) = e^{-t/2}$로부터 자유도 2의 카이제곱 분포를 따름을 알 수 있다. 따라서 독립된 여러 P값의 합은 자유도 $2K$인 카이제곱 분포를 따른다.

$$X^2 = \sum_{k=1}^{K} -2\log_e P_k. \tag{2.13}$$

그러므로 종합된 P값은 $P_{\text{Fisher}} = \Pr(\chi^2_{(2K)} \geq X^2)$이 된다. 여기서 $\chi^2_{(2K)}$는 자유도 $2K$의 카이제곱 분포이다.

Good (1958)은 다음과 같이 반응변수의 수를 독립된 P값의 역수의 합으로 나눈 값을 종합된 P값으로 정한다.

$$P_{\text{Good}} = \frac{K}{\sum_{k=1}^{K} P_k^{-1}}. \tag{2.14}$$

2.2.3. 상관성이 있는 P값의 보정된 자유도를 이용한 병합

Brown (1975)은 상관성이 있는 P값의 병합방법으로 앞에서 언급한 Fisher (1950)의 방법을 수정하여 적용한다. 상관성이 있는 P값을 병합하는 경우 위의 식 (2.13)에 제시된 X^2의 평균과 분산을 구해보면 다음과 같다.

$$E(X^2) = 2K, \tag{2.15}$$

$$\begin{aligned} \text{Var}(X^2) &= \sum_{i=1}^{K} \text{Var}(-2\log_e P_i) + 2\sum_{i<j}^{K}\sum^{K} \text{Cov}(-2\log_e P_i, -2\log_e P_j) \\ &= 4K + 2\sum_{i<j}^{K}\sum^{K} \text{Cov}(-2\log_e P_i, -2\log_e P_j). \end{aligned} \tag{2.16}$$

Brown (1975)은 분산의 오른쪽 항에 있는 $\text{Cov}(-2\log_e P_i, -2\log_e P_j)$를 i번째와 j번째 반응변수의 상관계수 ρ_{ij}의 함수로 표현하여, ρ_{ij}가 양수일 때 $\rho_{ij}(3.25 + 0.75 \times \rho_{ij})$로, 음수일 때 $\rho_{ij}(3.27 + 0.71 \times \rho_{ij})$로 제시하였다. 그러므로 과거 연구로부터 ρ_{ij}의 값을 알게 되면 $\text{Var}(X^2)$를 구할 수 있다. 이제 X^2의 평균과 분산이 구해졌고, 카이제곱 분포와 연관 짓는 일이 남아 있다.

$\chi^2_{(f)}$가 자유도 f의 카이제곱 분포일 때 식 (2.13)의 X^2를 $c\chi^2_{(f)}$로 근사시킬 수 있다고 가정하면, X^2의 평균과 분산은 다음과 같다.

$$E(X^2) = cf, \qquad \mathrm{Var}(X^2) = 2c^2 f. \tag{2.17}$$

그러므로 식 (2.15), (2.16)과 (2.17)을 이용하면 c와 f는 다음과 같이 정리된다.

$$c = \frac{\mathrm{Var}(X^2)}{2E(X^2)}, \qquad f = \frac{2[E(X^2)]^2}{\mathrm{Var}(X^2)}. \tag{2.18}$$

이제 $E(X^2)$와 $\mathrm{Var}(X^2)$를 이용하여 c와 f의 값을 구하고, 이로부터 자유도 f의 카이제곱 분포를 따르는 다음의 통계량을 얻는다.

$$\frac{X^2}{c} \approx \chi^2_{(f)} \tag{2.19}$$

따라서 종합된 P값은 $P_{\mathrm{Brown}} = \Pr(\chi^2_{(f)} \geq X^2/c)$이다. 이와 같이 상관성이 있는 여러 P값을 병합하여 P_{Brown}을 유도하는 방법을 우리는 편의상 '방법 B'라 부르겠다.

2.2.4. 상관성이 있는 P값의 독립 변환 후 병합

상관성이 있는 P값의 병합보다는 독립된 P값의 병합에 대한 이론들이 더욱 많이 제시되었음이 사실이다. 그러므로 상관성이 있는 P값을 독립된 P값으로 변환시켜 이를 병합하는 방법도 또한 가능하다. Zaykin 등 (2002)은 유전학의 상황, 즉 수만개의 유전자 정보를 분석해야 하는 경우에 매우 적절한 Wilkinson (1951)의 어느 값 이하의 P값만을 곱하여 종합하는 Truncation Product 방법을 제안하는 과정에서 또한 독립된 P값으로의 변환을 설명하였다. 본 논문에서는 임상시험에의 적용으로 상관성이 있는 P값을 독립된 P값으로 변환시킨 후에 효율성이 높다고 밝혀진 Fisher (1950)의 방법과 Good (1958)의 방법을 접목시킨다.

여러 반응변수의 상관행렬 $\mathbf{R}$이 양정치 행렬(positive definite matrix)일 때, $\mathbf{R} = \mathbf{M}\mathbf{M}^T$을 만족하는 콜레스키 요소(Cholesky factor)인 $K \times K$행렬 $\mathbf{M}$이 존재한다. 독립된 K개의 P값으로 이루어진 $K \times 1$벡터를 $\mathbf{P}_I$로 정의하면, 이에 상응하는 표준 정규 분포의 Z값 벡터는 $\mathbf{Z}_I = \Phi^{-1}(\mathbf{J} - \mathbf{P}_I)$이 된다. 여기서 $\Phi(\cdot)$는 표준 정규 분포의 확률분포 함수이며, $\mathbf{J}$는 1절에서 정의하였듯이 $K \times 1$ 벡터 $\mathbf{J} = (1, \ldots, 1)^T$이다. 이제 상관성이 있는 P값 벡터를 $\mathbf{P}$로 정의하면 이는 우선 $\mathbf{M}\mathbf{Z}_I$가 평균이 $\mathbf{0}$이고, 분산이 $\mathrm{Var}(\mathbf{M}\mathbf{Z}_I) = \mathbf{M}\mathrm{Var}(\mathbf{Z}_I)\mathbf{M}^T = \mathbf{M}\mathbf{M}^T = \mathbf{R}$인 다항 정규 분포를 따른다는 것을 이용하여 구한다.

$$\begin{aligned} \mathbf{P} &= \mathbf{J} - \Phi\{\mathbf{M}\mathbf{Z}_I\} \\ &= \mathbf{J} - \Phi\{\mathbf{M}\Phi^{-1}(\mathbf{J} - \mathbf{P}_I)\}. \end{aligned} \tag{2.20}$$

여기서 P값에 상응하는 Z값 벡터의 상관행렬 $\mathrm{Var}(\mathbf{Z}_I)$는 여러 반응변수의 상관행렬과 동일하다고 가정하였는데, 이는 상관계수는 단조 변환(monotone transformation)에 의해 근사적으로 불변하다는 사실, 즉, $\mathrm{Corr}\{g(P_i), g(P_j)\} \approx \mathrm{Corr}\{P_i, P_j\}$을 이용한 것이다 (Zaykin 등, 2002).

이로부터 독립된 P값 벡터 $\mathbf{P}_I$는 다음과 같이 구해진다.

$$\mathbf{P}_I = \mathbf{J} - \Phi\{\mathbf{M}^{-1}\Phi^{-1}(\mathbf{J} - \mathbf{P})\}. \tag{2.21}$$

이와 같이 상관성이 있는 P값의 벡터 $\mathbf{P}$를 변환하여 독립된 P값의 벡터 $\mathbf{P}_I$를 구하였고 이제 $\mathbf{P}_I$의 각 요소인 독립된 P값의 병합으로서 앞에서 소개한 Fisher (1950)의 P값 병합과 Good (1958)이 제시한 조화평균 방법을 적용한다. 각각의 방법에 의한 종합된 P값의 병합을 편의상 '방법 F'와 '방법 G'라 부르겠다.

3. 적용사례

여러 형태의 반응변수로부터 종합된 결론을 내려야 하는 경우가 흔히 발생하며 다음의 예제가 그 한가지 경우이다. 의학적 기술발전 및 국민의 고령화 등으로 인하여 의료비용이 크게 급증하면서 이에 대한 검토가 연구자들에 의해 여러 방면에서 이루어지고 있다. 특히 큰 지출항목으로 병원 입원치료가 지적되면서 퇴원의 불필요한 지연, 비효율적인 치료 연계 시스템으로 인한 입원일수의 증가 등, 부적정 입원일수를 줄임으로써 비용을 줄이고자 하는 노력이 부각되고 있다.

본 논문에서 제시하는 자료는 2002년 두 달 동안 대학산하 병원에서 수집되었으며, 50-59세 남성 환자 중 63명의 I과 환자와 47명의 S과 환자의 입원치료를 분석하여 두 진료과를 비교할 목적으로 부적정 입원일수를 수집하였다 (Kim, 2004). 부적정 입원일은 Gertman과 Restuccia (1981)에 의해 개발된 Appropriate Evaluation Protocol(AEP)을 이용하여 정의하며, 11개의 의료 서비스 및 진행에 관련된 항목, 7개의 간호 등의 보조 서비스에 관련된 항목, 8개의 환자의 임상적 특징에 관련된 항목으로 구성된 총 27개 항목에 근거하여 적정 입원일을 평가한다. 외국과 다른 한국 의료시스템의 특징을 반영하기 위해 국내 의료 연구자들에 의해 몇 차례 수정을 거친 도구로 입원일이 위의 항목 중 어느 것에도 부합되지 않을 때를 부적정일로 정의하였다. 환자의 부적정일 자료로부터 여러 부적정성 평가척도가 산출되며, 가장 흔히 쓰이는 척도는 각 환자의 총 입원일수 중 부적정 입원 일수의 비율이다. 그러나 연속변수인 이 부적정율은 대략적인 척도로서 충분히 세부적이지 못하다. 부적정일이 되기 쉬운 입원 첫날과 퇴원 전날은 의료 관리자들에게 특별한 의미를 가지며 이 각각은 이산형 변수이다. 처음 부적정일이 시작된 이후로 다음 부적정일이 나타나기 전까지를 하나의 런 통계량(run statistics)으로 정의하여 런수(run number)를 나타내는 순서형 변수가 또한 의미가 있다. 한 예로서, 입원기간이 총 21일이 되는 환자의 자료에서 0을 적정일로, 1을 부적정일로 표현하여 0-0-0-1-1-1-1-1-1-1-0-0-0-1-1-1-1-1-0-1-0-0-0와 같다고 하자. 총 11일의 부적정 일수로 입원일에 대한 부적정율은 11/21=0.524이 되며, 런수는 4이다. 또한 입원 첫날과 퇴원 전날 모두 적정한 치료를 받았다. 여러 변수가 함께 부적정한 정도를 나타낸다.

표 3.1에 I과와 S과 환자의 각 변수에 대한 요약 수치가 제시되었으며, 여러 척도를 종합하여서 두 진료과를 비교하고자 한다. 입원 첫날, 퇴원 전날은 전체 환자 중에서 부적정일을 가진 환자수가 제시되었고, 런수의 4는 그 이상을 포함하는 범주로서 런수에 제시된

표 3.1: 각 변수에 대한 요약 수치

	환자수	입원 첫날	퇴원 전날	런수				부적정율
				1	2	3	4	
S과	47	25 (53%)	27 (57%)	11 (23%)	13 (28%)	16 (34%)	7 (15%)	0.22
I과	63	15 (24%)	25 (40%)	30 (48%)	19 (30%)	7 (11%)	7 (11%)	0.20
	110	40 (36%)	52 (47%)	41 (37%)	32 (29%)	23 (21%)	14 (13%)	0.21

수는 각 범주에 속하는 환자수를 나타낸다. 이산형과 순서형 변수의 경우에 괄호 안에 비율을 제시하였다. S과에서 입원 첫날 부적정일의 비율이 높으며 또한 런수가 큰 경우의 비율도 높다. 그러나 부적정율은 S과가 I과보다 부적정한 정도가 높다는 사실을 반영하지 못하고 있다.

각 변수별로 S과가 I과보다 부적정 척도가 높다는 것을 검정할 목적으로 실시한 단변량 분석에서 입원 첫날과 퇴원 전날 변수는 비율차 검정으로 Z값이 각각 3.169과 1.846이고, 런수는 윌콕슨 순위합 검정으로 Z값이 2.797이다. 연속형인 부적정율은 모평균 비교로 t 검정을 실시하고 대표본 근사 Z통계량으로 0.39이 구해졌다. 유의수준 5%에서 부적정율은 S과의 부적정한 치료에 대한 유의한 근거가 되지 못하는 반면, 나머지 세 변수는 유의한 근거를 제시한다.

본 논문에서 설명한 다중 반응변수를 종합한 검정방법을 적용해 본다. 각 변수간 상관행렬은 과거 연구로부터 구하는 것이 최적이지만 이 진료과 분석에서는 표본에서 추정한 것을 사용한다.

$$\mathbf{R} = \begin{pmatrix} 1 & 0.2306 & 0.4970 & 0.5612 \\ & 1 & 0.5298 & 0.5387 \\ & & 1 & 0.4111 \\ & & & 1 \end{pmatrix}.$$

상관성 있는 P값 병합방법인 방법 B에서 구한 X^2 및 평균 $E(X^2)$과 분산 $\mathrm{Var}(X^2)$는 다음과 같다.

$$\begin{aligned} X^2 &= \sum_{k=1}^{K} -2\log_e P_k = 14.352 + 6.857 + 11.92 + 2.106 = 35.235, \\ E(X^2) &= 2K = 8, \\ \mathrm{Var}(X^2) &= 4K + 2\sum_{i<j}^{K}\sum^{K} \mathrm{Cov}(-2\log_e P_i, -2\log_e P_j) \\ &= 16 + 2 \times (0.791 + 1.801 + 2.06 + 1.933 + 1.97 + 1.462) \\ &= 36.026. \end{aligned}$$

표 3.2: 종합된 분석 결과

	OLS	GLS	방법 B	방법 G	방법 F
검정 통계량	2.656	2.957	15.648	-	26.185
종합된 P값	0.004	0.0016	0.0023	0.0004	0.001

따라서 $c = \mathrm{Var}(X^2)/2E(X^2) = 36.026/(2 \times 8) = 2.252$, $f = 2E(X^2)^2/\mathrm{Var}(X^2) = (2 \times 8^2)/36.026 = 3.553$을 이용하여 얻은 $X^2/c = 15.648$는 자유도 f인 카이제곱 분포를 따른다. 이 때 P_{Brown}값은 0.0023으로, 유의수준 5%에서 I과와 S과 입원환자의 부적정한 정도에 차이가 없다는 귀무가설을 기각한다.

다음은 방법 F와 G의 계산에서 요구되는 $\mathbf{M}$과 $\mathbf{M}^{-1}$이다.

$$\mathbf{M} = \begin{pmatrix} 1 & 0.2306 & 0.4970 & 0.5612 \\ 0 & 0.9731 & 0.4267 & 0.4206 \\ 0 & 0 & 0.7556 & -0.0625 \\ 0 & 0 & 0 & 0.7101 \end{pmatrix},$$

$$\mathbf{M}^{-1} = \begin{pmatrix} 1 & -0.2369 & -0.5240 & -0.6960 \\ 0 & 1.0277 & -0.5803 & -0.6599 \\ 0 & 0 & 1.3234 & 0.1165 \\ 0 & 0 & 0 & 1.4082 \end{pmatrix}.$$

이제 단변량으로부터 구해진 P값 벡터 $\mathbf{P} = (0.0008, 0.0324, 0.0026, 0.3488)^T$과 이에 해당하는 Z값 벡터 $\Phi^{-1}(\mathbf{J}-\mathbf{P}) = (3.1690, 1.8461, 2.7968, 0.3885)^T$에 의해 $\Phi\{\mathbf{M}^{-1}\Phi^{-1}(\mathbf{J}-\mathbf{P})\} = (0.8403, 0.5071, 0.9999, 0.7079)^T$이 계산되어, 독립 변환된 P값 벡터 $\mathbf{P}_I = (0.1597, 0.4929, 0.0001, 0.2921)^T$를 유도한다.

$\mathbf{P}_I$를 이용하여 $\sum P_k^{-1} = 6.2617 + 2.0288 + 10000 + 3.4235 = 10011.71$을 분모로 하는 $P_{\mathrm{Good}} = 4/10011.71 = 0.0004$이 계산된다. 그리고 자유도 8의 카이제곱 분포를 따르는 $X^2_{\mathrm{Fisher}} = 3.6692 + 1.4148 + 18.6401 + 2.4610 = 26.1851$의 P_{Fisher}값은 0.0010이다. 방법 G와 F 또한 유의수준 5%에서 I과와 S과의 부적정한 정도에 차이가 없다는 귀무가설을 기각한다.

OLS와 GLS 통계량은 단변량 분석 결과 유도한 검정 통계량 $\mathbf{Z}$벡터로부터 $T_{\mathrm{OLS}} = \mathbf{J}^T\mathbf{Z}/\sqrt{\mathbf{J}^T\hat{\mathbf{R}}\mathbf{J}} = 8.2013/3.0881 = 2.656$과 $T_{\mathrm{GLS}} = \mathbf{J}^T\hat{\mathbf{R}}^{-1}\mathbf{Z}/\sqrt{\mathbf{J}^T\hat{\mathbf{R}}^{-1}\mathbf{J}} = 3.8574/1.3045 = 2.957$을 각각 계산한다. 모든 방법에서 종합된 P값이 유의수준 5%하에서 유의하다. 표 3.2에 OLS, GLS 및 방법 B, G, F의 결과를 요약하였다.

4. 모의실험

4.1. 모의실험 계획

여러 형태의 다변량 반응변수로 치료효과가 측정된 경우, 독립된 두 군의 비교에 대한 검

정법으로 본 논문에서 여러 방법이 제시되었고, 이를 비교하는 모의실험을 실시하였다. 상관성이 있는 4개의 다변량 반응변수를 생성하며 각 군에서 동일 표본수 $n(=n_1=n_2)$=20, 50인 경우를 고려한다. 두 집단 중 치료군(treatment group)은 T로, 대조군(control group)은 C로 표시하고, T군의 반응값이 C군보다 크다는 대립가설 하에서 단측 검정을 1000번 반복하여 제 1종 오류와 검정력을 비교한다.

본 논문에서 제안한 P값의 병합방법이 OLS와 GLS 통계량에 비하여 타당한지 알아내기 위해 우선 모든 방법을 연속형 반응변수에 적용하여 비교한다. 연속형으로는 다변량 정규와 비정규 분포를 생성하며 비정규 분포로 오염된 정규(contaminated normal) 분포와 로그 정규(log-normal) 분포를 고려한다.

다변량 정규 분포를 N이라 지칭하고, N은 $Y_{ijk}=cX_{ij0}+\sqrt{1-c^2}X_{ijk}$에 기초하여 랜덤하게 생성한다. 여기서 X_{ij0}과 X_{ijk}는 각각 독립된 표준 정규 변량이다. 다변량 정규 분포의 상관행렬 $\mathbf{R}$은 다음과 같으며, 모의실험에서 $c^2=1/2$을 가정한다.

$$\mathbf{R}=\begin{pmatrix} 1 & c^2 & c^2 & c^2 \\ & 1 & c^2 & c^2 \\ & & 1 & c^2 \\ & & & 1 \end{pmatrix}.$$

오염된 정규 분포의 생성은 표본수의 95%는 $N(0,1)$에서, 나머지 5%는 $N(0,25)$에서 생성한다. 4개 변수 중 2개는 정규 분포로, 나머지 2개 변수는 오염된 정규 분포로 생성된 경우를 CN이라 지칭한다. 대립가설의 검정은 T군의 정규 분포의 평균을 0.5로 두고 실시하였다. 비정규 분포로서 로그 정규 분포의 생성은 N과 같은 방법으로 생성된 자료에 지수변환($\exp(Y_{ijk})$)을 거쳐 얻는다. 다변량 정규 분포에서 3개 변수가 로그 정규 분포로 대치된 경우를 LN이라 지칭한다.

연속형 자료에 모든 검정법, 즉, OLS와 GLS 통계량과 방법 B, F, G를 비교하기 위해 우선적으로 변수별로 표준화를 실시한다. 표준화된 자료에 OLS와 GLS 통계량은 t 통계량을 구하여 병합하고, 방법 B, F, G는 각 변수 모두 t 검정을 실시하여 구한 P값을 병합한다. 표본에서 계산된 상관계수 행렬을 공분산 행렬 $\mathbf{R}$로 사용하여 P값을 병합한다.

여러 형태 반응변수의 자료 생성은 다음과 같다. 4개의 변수를 순서대로 위에서 언급한 연속형 분포 N, CN, LN으로 생성한 후, 세 번째 변수는 이산형으로, 네 번째 변수는 순서형으로 변환한다. 이산형 변수는 T군에서 미리 지정한 이산 비율(binary property)에 상응하는 분위수를 절단점(cutoff point)으로 하여, T군과 C군의 자료가 절단점과 같거나 작을 경우를 0으로, 큰 경우를 1로 변환한다. 이산 비율을 0.3, 0.5, 0.8로 증가시켜 모의실험을 실시한다. 순서형 자료는 범주수인 m을 미리 정한 후, 두 군 전체 자료를 $1/m$씩으로 나누어 순위를 부여함으로서 변환한다. 범주수를 3, 5, 7개로 증가시켜 모의실험을 실시한다.

P값의 변환 시에 요구되는 상관행렬은 원래의 연속형 자료로부터 추정된 것을 사용한다.

표 4.1: 4차원 연속형 분포에서의 제 1종 오류

	분포	표본수	OLS	GLS	방법 B	방법 F	방법 G
$\alpha = 0.05$	다변량 정규 분포	20	0.051	0.050	0.053	0.070	0.103
		50	0.056	0.057	0.056	0.070	0.098
	오염된 정규 분포	20	0.052	0.052	0.053	0.071	0.109
		50	0.056	0.055	0.059	0.067	0.094
	로그 정규 분포	20	0.055	0.058	0.055	0.072	0.100
		50	0.052	0.051	0.050	0.057	0.080
$\alpha = 0.01$	다변량 정규 분포	20	0.013	0.012	0.015	0.027	0.040
		50	0.013	0.013	0.013	0.025	0.026
	오염된 정규 분포	20	0.010	0.010	0.010	0.027	0.038
		50	0.013	0.013	0.013	0.020	0.028
	로그 정규 분포	20	0.010	0.010	0.011	0.021	0.031
		50	0.011	0.010	0.010	0.013	0.018

로그 정규 분포 : 한 변수는 정규 분포, 세 변수는 로그 정규 분포.

4.2. 모의실험 결과

표 4.1은 연속형 자료 N, CN, LN 각각에서 유의수준과 표본의 크기를 달리하여, 치료효과가 없다는 귀무가설 하에서의 제 1종 오류를 나타낸 것이다. OLS와 GLS의 제 1종 오류는 정해진 유의수준에 비교적 근사하고 P값의 병합방법 중 방법 B가 이 두 방법과 비슷한 결과를 보이지만, 방법 F와 G는 유의수준보다 큰 수치를 보이고 있어 바람직하지 않다.

표 4.2는 연속형 자료 N, CN, LN 각각에서 단측 대립가설 하에서의 검정력이다. 모든 검정방법에서 분포가 N과 CN일 때 검정력이 대체로 높다. 특히 1개의 변수만이 정규 분포하고 나머지 3개의 변수가 로그 정규 분포하는 LN에서 가장 낮은 검정력을 보인다. 다변량 정규 분포에서 2개의 변수만이 로그 정규 분포로 대치되어 비대칭의 변수가 덜 섞인 경우도 모의실험을 실시하였으나 앞에서의 로그 정규 분포보다 검정력이 약간 높을 뿐이어서 제시치 않는다. 이러한 연속형 자료에서의 각 방법의 검정력을 일부 변수를 이산형과 순서형 자료로 변환한 후 적용한 P값 병합방법의 검정력과 비교할 것이다.

이제 N, CN, LN 자료에서 세번째 변수를 이산형으로, 네번째 변수를 순서형으로 변환시킨 경우의 제 1종 오류가 표 4.3에 제시되었다. 대칭 분포인 N과 CN분포에서 표본수가 50인 경우에 방법 B가 유의수준에 가까운 값을 보이고, 비대칭 분포인 변수가 섞인 LN의 경우에 표본수가 20으로 작더라도 방법 B의 1종 오류는 유의수준에 가깝다. 그러나 방법 F와 G는 제 1종 오류가 유의수준보다 크며, 특히 방법 G는 지정된 유의수준의 2배에 가까운 큰 값을 가진다. 또한, 대칭인 N과 CN분포에서 방법 F와 G의 제 1종 오류는 표본수에 따라 변함없이 여전히 크며, LN분포에서는 작은 표본수 20인 경우에 큰 값을 가진다. 이와 같이 이산형, 순서형으로 변환시킨 경우의 여러 방법의 검정력이 표 4.4에 제시되었다. 모든 경우에 걸쳐 이산 비율이 증가할수록, 그리고 범주수가 증가할수록 검정력이 증가하고 있다.

표 4.4의 P값 병합의 검정력을 표 4.2에 제시된 연속형 분포의 경우와 비교해 보면 대칭

표 4.2: 4차원 연속형 분포에서의 검정력

	분포	표본수	OLS	GLS	방법 B	방법 F	방법 G
$\alpha = 0.05$	다변량 정규 분포	20	0.647	0.637	0.649	0.677	0.662
		50	0.938	0.934	0.935	0.938	0.911
	오염된 정규 분포	20	0.637	0.632	0.639	0.665	0.658
		50	0.936	0.933	0.932	0.934	0.910
	로그 정규 분포	20	0.436	0.421	0.455	0.484	0.471
		50	0.695	0.682	0.724	0.732	0.647
$\alpha = 0.01$	다변량 정규 분포	20	0.381	0.373	0.384	0.488	0.442
		50	0.790	0.789	0.792	0.836	0.769
	오염된 정규 분포	20	0.375	0.369	0.384	0.492	0.453
		50	0.798	0.793	0.802	0.834	0.749
	로그 정규 분포	20	0.203	0.196	0.218	0.274	0.239
		50	0.450	0.435	0.486	0.495	0.349

로그 정규 분포 : 한 변수는 정규 분포, 세 변수는 로그 정규 분포.

표 4.3: 4차원 연속형 반응변수를 여러 형태로 변환 후의 제 1종 오류

표본	범주	이산	정규 분포			오염된 정규 분포			로그정규 분포		
수	수	비율	방법B	방법F	방법G	방법B	방법F	방법G	방법B	방법F	방법G
20	3	0.3	0.042	0.064	0.103	0.041	0.064	0.108	0.046	0.074	0.113
		0.5	0.043	0.066	0.107	0.041	0.063	0.106	0.048	0.077	0.116
		0.8	0.043	0.065	0.103	0.038	0.063	0.105	0.042	0.078	0.116
	5	0.3	0.044	0.064	0.106	0.039	0.061	0.105	0.050	0.077	0.113
		0.5	0.045	0.065	0.104	0.042	0.062	0.103	0.052	0.080	0.112
		0.8	0.043	0.062	0.106	0.039	0.059	0.100	0.047	0.079	0.110
	7	0.3	0.044	0.067	0.105	0.042	0.066	0.105	0.048	0.077	0.111
		0.5	0.046	0.068	0.108	0.042	0.066	0.101	0.050	0.079	0.112
		0.8	0.042	0.066	0.105	0.040	0.063	0.099	0.049	0.076	0.110
50	3	0.3	0.046	0.066	0.106	0.047	0.065	0.098	0.044	0.061	0.088
		0.5	0.048	0.068	0.105	0.048	0.064	0.100	0.045	0.060	0.092
		0.8	0.044	0.067	0.110	0.044	0.065	0.100	0.042	0.062	0.096
	5	0.3	0.049	0.068	0.110	0.046	0.063	0.096	0.049	0.064	0.085
		0.5	0.051	0.069	0.109	0.050	0.064	0.099	0.051	0.065	0.087
		0.8	0.048	0.066	0.113	0.047	0.062	0.095	0.048	0.066	0.090
	7	0.3	0.047	0.066	0.105	0.049	0.063	0.097	0.051	0.063	0.086
		0.5	0.050	0.068	0.104	0.051	0.065	0.098	0.051	0.065	0.083
		0.8	0.048	0.066	0.108	0.047	0.063	0.093	0.049	0.066	0.089

분포인 N과 CN분포의 경우 표 4.2의 검정력에는 못 미치지만 범주수가 커지면서 매우 가까운 값을 나타내 보인다. 이와 반대로 비대칭 분포인 LN분포에서는 오히려 P값 병합의 검정력이 표 4.2의 연속형 분포의 OLS와 GLS의 검정력보다 높다. 이는 이산 비율과 각 범주의 비율에 따라 각각의 통계량에 의한 결과가 더욱 유의하고 따라서 P값 병합방법이 더

표 4.4: 4차원 연속형 반응변수를 여러 형태로 변환 후의 검정력

표본 수	범주 수	이산 비율	정규 분포			오염된 정규 분포			로그정규 분포		
			방법B	방법F	방법G	방법B	방법F	방법G	방법B	방법F	방법G
20	3	0.3	0.554	0.597	0.573	0.448	0.487	0.444	0.483	0.543	0.500
		0.5	0.566	0.608	0.582	0.499	0.528	0.470	0.526	0.568	0.518
		0.8	0.557	0.601	0.577	0.616	0.679	0.644	0.646	0.706	0.663
	5	0.3	0.567	0.609	0.593	0.514	0.549	0.510	0.546	0.591	0.557
		0.5	0.582	0.620	0.603	0.559	0.585	0.530	0.585	0.616	0.569
		0.8	0.574	0.620	0.604	0.666	0.700	0.659	0.689	0.731	0.684
	7	0.3	0.574	0.613	0.597	0.532	0.562	0.525	0.569	0.613	0.575
		0.5	0.583	0.621	0.607	0.581	0.598	0.544	0.605	0.636	0.584
		0.8	0.574	0.621	0.605	0.682	0.708	0.664	0.711	0.745	0.693
50	3	0.3	0.896	0.897	0.841	0.895	0.898	0.841	0.822	0.855	0.794
		0.5	0.902	0.904	0.853	0.903	0.902	0.849	0.863	0.883	0.807
		0.8	0.894	0.898	0.853	0.896	0.900	0.848	0.951	0.967	0.951
	5	0.3	0.907	0.909	0.868	0.901	0.905	0.862	0.858	0.873	0.810
		0.5	0.912	0.913	0.876	0.907	0.911	0.866	0.892	0.891	0.820
		0.8	0.904	0.910	0.869	0.903	0.909	0.866	0.964	0.971	0.950
	7	0.3	0.907	0.915	0.871	0.904	0.911	0.871	0.890	0.898	0.845
		0.5	0.915	0.917	0.878	0.911	0.914	0.875	0.914	0.912	0.853
		0.8	0.909	0.912	0.870	0.906	0.912	0.875	0.970	0.974	0.956

욱 유의할 수 있음을 말해주고 있다. P값의 병합방법 중 방법 G의 검정력이 떨어지고 방법 F의 검정력이 방법 B보다 약간 우월한 경우가 있지만 우선 방법 F의 1종 오류가 보장되지 않으므로 방법 F가 효율적이라고 말할 수 없으며 오히려 방법 B가 가장 효율적이라고 결론내릴 수 있다. 모의실험을 $\alpha = 0.01$인 경우도 시행하였으나 $\alpha = 0.05$의 경우와 비교하여 여러 검정법의 우위에 있어 변함이 없으므로 제시치 않는다.

5. 결론

임상 시험에서 종종 다중 반응변수의 종합된 분석으로 결론을 내려야하는 현실에 부딪히게 되는데, 이때 다중 반응변수는 연속형 뿐만 아니라, 이산형, 순서형 등 여러 형태를 포함한다. 이러한 상황에도 불구하고 다중 반응변수의 종합적 검정 방법에 대한 연구가 부족한 것이 사실이다.

검정에서 널리 이용되는 Bonferroni 수정법은 그 적용에 있어서 매우 간편한 장점이 있다. 그러나 여러 반응변수에 대한 P값 중 가장 작은 값을 이용하기 때문에 다음과 같은 문제가 발생할 수도 있다. 예를 들어서 두 반응변수의 P값이 0.01과 0.8로 이루어진 경우가 모두 0.2로 이루어진 경우보다 Bonferroni 수정법에 의해 유의하게 결론 내리게 되는데, 치료효과에 대한 설득력은 후자의 경우가 더욱 크다하겠다.

이에 대한 대체안으로 Pocock 등 (1987)이 GLS 통계량을 확장하여 연속형 뿐만 아니라, 이산형, 생존형 자료에까지 적용하였다. 이 방법은 공분산 구조에 있어서 불규칙성을 갖는

경우, 즉, 어떤 한 변수에서 공분산이 음수인 경우에 여러 반응변수를 합하는 GLS 통계량의 유의성 검정 결과에 영향을 미치기 때문에 그 적용 및 해석에 주의가 필요하게 된다.

이에 본 논문에서는 연속형, 이산형, 순서형 반응변수의 종합적인 분석방법으로 단측 검정의 P값의 병합방법을 제안하였다. 일반적으로 P값의 병합방법은 반응변수의 형태나 반응변수의 검정 통계량의 형태에 상관없이 적용할 수 있기 때문에, 그 병합에 있어서 제한이 없다는 장점을 지닌다. P값의 병합방법으로 크게 상관성이 있는 P값의 보정된 자유도를 이용한 병합과 독립된 P값으로 변환 후의 병합방법을 고려하였다. 결과적으로 독립된 P값으로 변환 후 Fisher (1950)방법과 Good (1958)의 조화 평균을 이용한 병합방법보다는 Brown (1975)방법에 의한 상관성이 있는 P값의 병합이 더욱 효율적으로 나타났다. Fisher (1950)방법과 Good (1958)의 방법은 공분산의 구조가 음수를 취할 경우 콜레스키 요소로 분해할 수 없어 GLS통계량의 경우와 마찬가지로 적용이 제한된다. 결과적으로 이러한 문제가 발생되지 않는 Brown (1975)방법이 더욱 바람직하다하겠다.

Zaykin 등 (2002)이 소개한 독립된 P값으로 변환 후 Wilkinson (1951)의 방법을 적용한 Truncation Product 방법은 유전학 분야로 적용의 폭이 다양해졌다는 점에서 흥미롭다. 수만개의 유전자에 대한 검정을 종합하게 되는 유전학 분야 뿐만 아니라 여러 분야에서 효율성이 높은 종합적 검정 통계량이 제안되고 더욱 연구되어져야 한다.

참고문헌

Brown, M. B. (1975). 400: A method for combining non-independent, one-sided tests of significance, *Biometrics*, **31**, 987–992.

Chow, S. C. and Liu, J. P. (2003). *Design and Analysis of Clinical Trials: Concepts and Methodologies*, 2nd ed., John Wiley & Sons, New Jersey.

Fisher, R. A. (1950). *Statistical Methods for Research Workers*, 11th ed., Oliver and Boyd, London.

Gehan, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly censored samples, *Biometrika*, **52**, 203–223.

Gertman, P. M. and Restuccia, J. D. (1981). The appropriateness evaluation protocol: A technique for assessing unnecessary days of hospital care, *Medical Care*, **14**, 855–871.

Good, I. J. (1958). Significance test in parallel and in series, *Journal of the American Statistical Association*, **53**, 799–813.

Gupta, S. S. (1963). Probability integrals of multivariate normal and multivariate t, *The Annals of Mathematical Statistics*, **34**, 792–828.

Hollander, M. and Wolfe, D. A. (1999). *Nonparametric Statistical Methods*, 2nd ed., John Wiley & Sons, New York.

Kim, J. H. (2004). *Assessing Inappropriate Hospital Days with Random Coefficient Regression Models*, Catholic University, Seoul, Korea.

Lehmacher, W., Wassmer, G. and Reitmeir, P. (1991). Procedures for two-sample comparisons with multiple endpoints controlling the experimentwise error rate, *Biometrics*, **47**, 511–521.

Littell, R. C. and Folks, J. L. (1971). Asymptotic optimality of Fisher's method of combining independent tests, *Journal of the American Statistical Association*, **66**, 802–806.

Littell, R. C. and Folks, J. L. (1973). Asymptotic optimality of Fisher's method of combining independent tests Ⅱ, *Journal of the American Statistical Association*, **68**, 193–194.
Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration, *Cancer Chemotherapy Reports*, **50**, 163–170.
Meier, P. (1975). Statistics and medical experimentation, *Biometrics*, **31**, 511–529.
O´Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints, *Biometrics*, **40**, 1079–1087.
Pocock, S. J., Geller, N. L. and Tsiatis, A. A. (1987). The analysis of multiple endpoints in clinical trials, *Biometrics*, **43**, 487–498.
Wilkinson, B. (1951). A statistical consideration in psychological research, *Psychological Bulletin*, **48**, 156–158.
Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H. and Weir, B. S. (2002). Truncated product method for combining P-values, *Genetic Epidemiology*, **22**, 170–185.

[2007년 8월 접수, 2007년 10월 채택]

Methods of Combining P-values for Multiple Endpoints of Various Data Types

Su-Young Kim[1] Hae-Hiang Song[2]

Abstract

Comparative studies in Phase III clinical trials quite often involve two or more equally important endpoints, and one cannot select primary endpoint from them. O´Brien(1984) proposed for continuous endpoints the OLS and GLS statistics as miltivariate test statistics. Pocock *et al.* (1987) mentioned the possibility of analyzing a mixture of data types, such as quantitative, binary and survival data types, with the OLS and GLS statistics, but the authors did not explore problems in combining several endpoints of different types. Furthermore, they did not perform a simulation study to assess the efficiencies of the OLS and GLS statistics for endpoints of a mixture of data types. In this paper, we propose the combining methods of correlated P-values for the analysis of multiple endpoints, and compare the efficiencies of this method with those of OLS and GLS statistics for a mixture of data types with a simulation study. Among the several methods of combining P-values that are more advantageous than combining of OLS and GLS statistics, method B maintains nominal significance levels and is more efficient, while method F and G have type I error rates that are larger than the specified significance levels, which might occasionally lead to a wrong conclusion.

Keywords: Multiple endpoints, combining independent P-values, combining correlated P-values, OLS, GLS.

1) Graduate Student, Dept. of Biostatistics, The Catholic University of Korea, Seoul 137-701, Korea. E-mail: kimsuyoung@catholic.ac.kr

2) Corresponding author. Professor, Dept. of Biostatistics, The Catholic University of Korea, Seoul 137-701, Korea.
E-mail: hhsong@catholic.ac.kr