Review
Korean J Otorhinolaryngol-Head Neck Surg 2020;63(3):97-100 / eISSN 2092-6529
https://doi.org/10.3342/kjorl-hns.2020.00164

Biostatistics for Multiple Testing

Jeong-Seok Choi
Department of Otorhinolaryngology-Head and Neck Surgery, Inha University School of Medicine, Incheon, Korea

다중 검정을 위한 생물 통계학

최정석
인하대학교 의과대학 이비인후과학교실

Received February 7, 2020
Revised February 28, 2020
Accepted February 28, 2020
Address for correspondence
Jeong-Seok Choi, MD, PhD
Department of Otorhinolaryngology-Head and Neck Surgery,
Inha University School of Medicine,
27 Inhang-ro, Jung-gu,
Incheon 22332, Korea
Tel +82-32-890-2438
Fax +82-32-890-3580
E-mail jschoi@inha.ac.kr

Multiple testings are instances that contain simultaneous tests for more than one hypothesis. When multiple testings are conducted at the same time, it is more likely that the null hypothesis is rejected, even if the null hypothesis is correct. If individual hypothesis decisions are based on unadjusted *p*-values, it is usually more likely that some of the true null hypotheses will be rejected. In order to solve the multiple testing problems, various studies have attempted to increase the power by taking into account the family-wise error rate or false discovery rate and statistics required for testing hypotheses. This article discuss methods that account for the multiplicity issue and introduces various statistical techniques.
Korean J Otorhinolaryngol-Head Neck Surg 2020;63(3):97-100

Key Words Bioinformatics · Biostatistics.

서 론

최근에는 인체의 염기서열을 쉽게 읽어내는 기술(next-generation sequencing)이 발달함으로 인해 대량의 유전체에 관한 정보가 증가하고 있으며, 이러한 다양한 유전적인 정보들을 해석하고 진단 및 치료 등에 이용하고자 하는 생물통계 및 생물정보학(biostatistics and bioinformatics)이라는 새로운 학문 분야가 탄생하게 되었다.[1] 염기서열을 통해 분석이 가능한 유전자의 개수는 통상 수백 개에서 수만 개에 이르는 것으로 알려져 있다.

예를 들어, 일반적인 방법을 통해 두경부 암을 가진 환자군과 정상군 간에 만 개의 유전자 발현 차이를 분석하여 두경부 암과 관련이 있는 유전자를 찾기 위한 연구를 수행한다고 가정을 해 보면, 두 군에서 다르게 발현이 되는 유전자를 찾기 위해 각 유전자마다 연속형 변수에 준한 t 검정(t-test)을 통해 두 군 간 차이가 있는 유전자들을 선별해 낼 수가 있다. 그러나, 이렇게 선택된 유전자들이 두경부암과 연관성이 있다고 결론을 내기에는 무리가 따른다. 물론 각각의 검정이 특정 유의 수준하에서 검정한 것이기 때문에, 여기에서 선택된 유전자들이 두경부 암과 연관성이 있을 가능성이 높은 것이 사실이다. 하지만 만 가지의 유전자에 대한 검정을 하나의 통계로 볼 때, 모든 유전자가 연관성이 실제로 없더라도 대략적으로 수십에서 수백여 개의 잘못된 유전자를 선택하게 되는 결과를 얻게 될 가능성이 높다. 가령, 만 개의 모든 유전자가 두 군에서 모두 다르게 발현되지 않는다고 가정을 했을 때, 유의 수준을 5%(α=0.05)로 가정한다면 본 통계에서는 만 개의 5%인 500개(10000×0.05)의 유전자가 두 군 간에 의미가 있다고 잘못 선택이 될 수 있다. 만약 이러한 잘못된 결론에 의한 두경부 암의 유전자 연구는 오히려 그 분야의 연구 발전을 저해할 수 있는 과학적 근거가 될 수 있다. 설사 500개의 유전자가 정말 의미가 있다고 가정을 하더라도 이를 통한 후속 연구를 하기에는 선택된 유전자가 너무 많다는 것도 또 하나의 문제가 될 수 있다. 결국, t 검정을 여러 번 반복하게 되

면 제1종 오류(type 1 error, 귀무가설이 사실인데 그 가설을 기각함으로써 발생하는 오류)가 매우 커질 수밖에 없다. 따라서, 연구의 오류를 줄이는 방법이 필요하며, 이러한 연구에 있어 잘못된 결과를 줄이는 것이 다중 검정의 보정(multiple test adjustment)의 핵심이 된다. 어느 유전자가 실제로 의미가 있는가를 따지기 위해서는 이러한 p 값(p value)을 보정해야 하는 방법이 매우 중요하다.[2] 본 논문에서는 다중검정 통해 p 값을 보정하는 다양한 생물통계학적 기법에 대해 소개하고자 한다.

본 론

제1종 오류를 조절하여 다중 검정을 보정하는 통계학적 방법은 크게 두 가지로 나눌 수 있는데, 집단별 오류율(family-wise error rate, FWER)과 위발견율(false discovery rate, FDR)을 조절하는 것이다.[3,4] 집단별 오류율을 통한 다중 검정 방법은 Bonferroni test, Single-step procedures(SSP), step down procedure(SDP) 등이 있으며,[5-7] 위발견율을 통해 다중 검정 보정을 하는 방법은 Benjamini and Hochberg의 방법과 Storey의 방법 등이 있다.[8,9]

검정해야 할 유전자가 m개 있고 이중에서 m_1개의 유전자가 두 군 간에 차이를 보이는 유전자라 가정하고 $m_0(=m-m_1)$의 유전자가 두 군 간의 차이를 보이지 않는 유전자라고 가정을 해보자. 각각의 유전자에 대해서 우리는 통계적 가설을 설정할 수가 있다. 즉, 하나의 유전자 j에 대해서, 다음과 같은 두 개의 통계적 가설(귀무가설, 대립가설)을 세울 수 있다.

귀무가설: 유전자 j는 두 군에서 동등하게 발현을 한다.

대립가설: 유전자 j는 두 군에서 다르게 발현을 한다.

m개의 유전자에 대해서 통계학적 검정을 m번 실시하게 되면, m_0개의 유전자에 대해서는 귀무가설이 참이고 m_1개의 유전자에 대해서는 대립가설이 참이 된다. 이러한 데이터를 가지고 검정을 하게 되면, 어떤 유전자는 귀무가설을 받아들이게 되고(귀무가설을 기각할 수 없게 되고), 어떤 유전자는 대립가설을 받아들이게 된다(귀무가설을 기각하게 된다). 대립가설을 받아들인 유전자의 수를 R이라고 하면, 이는 R개의 유전자를 선택한 것이 된다. 이 R개의 유전자 중에서 귀무가설이 참인데, 대립가설을 참으로 선택한 경우를 R_0라 정의한다면 이는 잘못된 선택이 되므로 잘못된 발견(false discovery)에 해당된다. 나머지의 유전자를 $R_1(=R-R_0)$이라고 정의한다면, 이는 대립가설이 참인데, 대립가설을 받아들인 경우가 되므로 올바른 발견(true discovery)에 해당이 된다(Table 1). 여기에서 집단별 오류율의 정의는 다음과 같다.[10]

Table 1. Probable outcomes when testing multiple null hypotheses

	Number not rejected	Number rejected	Total
True null hypotheses	A_0	R_0	m_0
Non-true null hypotheses	A_1	R_1	m_1
Total	A	R	m

$$\text{Family – Wise Error Rate (FWER)} = P(R>0 \mid m_1=0)$$

이를 풀어서 설명하면, m_1은 대립가설이 참인 유전자의 수를 의미한다. $m_1=0$이라는 뜻은 m개의 유전자들이 두 군을 감별하는 데 아무런 영향을 주지 않는다고 가정을 할 때 적어도 하나 이상의 유전자가 선택이 될 확률을 말한다. 위발견율은 선택된 유전자 중에서 잘못 선택된 유전자의 개수에 대한 기대값을 의미하며 아래와 같이 정의한다.[11]

$$\text{False Discovery Rate (FDR): } FDR = E\left(\frac{Ro}{R}\right) = E\left(\frac{Ro}{R0+R1}\right)$$

집단별 오류율을 통한 보정 방법 중에 가장 전통적인 Bonferroni의 방법을 설명하고자 한다. 집단별 오류율의 정의에 의하여 모든 유전자가 두 군 간에 차이가 없다고 가정을 했을 때 어느 하나의 유전자라도 t 검정 통계의 절대값이 특정 변수 c보다 클 경우 그 유전자를 선택하는 확률을 의미하게 되고 이는 아래 와 같이 수식으로 표시가 될 수 있다.[12]

$$\text{Family – Wise Error Rate (FWER)} = P(|T_1|> c \text{ or, ..., or } |T_m| > c \mid \cap_{j=1}^{m} Hj)$$

Bonferroni의 부등식의 정의는 아래와 같이 정의가 되는데, 모든 유전자 상호 간에 교집합이 없다고 가정을 하게 되면 두 식은 등식이 성립하게 된다. 하지만 실제로는 유전자들은 서로 상호 관계를 가지며 발현하는 경우가 대부분이므로 $T_1 \sim T_m$ 사이에는 매우 복잡한 상관관계를 가질 수밖에 없다. 이를 감안하여 확률을 계산하는 것은 불가능에 가까우므로 Bonferroni는 아래의 부등식을 이용하여 값을 추정하고자 하였다.[12]

Bonferroni의 부등식:

$$FWER = P(|T_1|> c \text{ or, ..., or } |T_m| > c \mid \cap_{j=1}^{m} Hj) \leq \sum_{j=1}^{m} P(|T_j|>c \mid H_j)$$

집단별 오류율의 값을 α로 정하고자 한다면, m개의 유전자에 대해 각 유전자를 선택하는 확률의 합보다 작은 수를 선택하면 되므로 각각의 t 검정에서 사용하는 제1종 오류율(marginal type I error rate)을 얼마로 정하는 것이 집단별 오류율의 값이 α가 되는가의 문제로 귀결이 된다. t 검정은 샘플

수가 크면 정규분포를 따르게 되므로, 제1종 오류율은 결국 α/m보다 작아야 한다. 이러한 Bonferroni의 방법은 많은 문제가 있는데, 모든 유전자 상호 간에 교집합이 없다고 가정하는 것과 유전자의 수(m)가 매우 많은 현실적인 측면을 감안하면 두 등식의 차이가 매우 커지게 된다. 이는 Bonferroni의 방법에 의해 결정된 값이 매우 보수적인 값이 되며, 이는 실제 두 군 간의 차이가 있는 유전자가 선택이 되지 않을 수도 있다. 만약 m개의 유전자가 완벽하게 교집합을 이루고 있다고 가정하면 marginal type I error rate는 그대로 α가 된다. 결국 우리가 찾고자 하는 정확한 marginal type I error rate는 α/m와 α 사이의 값임을 추정할 수 있다. m이 매우 큰 숫자라고 한다면 α/m와 α 사이도 매우 큰 차이가 나게 되므로 Bonferroni의 방법을 응용하기에는 현실적으로 무리가 따른다. 보수적인 Bonferroni의 방법은 귀무가설을 잘 기각하지 않음으로 해서 위양성(false positive)은 줄일 수 있는 장점이 있지만 위음성(false negative)은 많아질 수밖에 없다. 물론 위양성이 적은 것이 좋기는 하지만 위음성이 많은 것은 또 다른 문제가 생기므로, 다중 비교 검정의 핵심은 얼마나 위양성을 줄이면서, 위음성도 같이 줄일 수 있느냐에 관건이 있다. 따라서, Bonferroni의 방법에 의한 위양성의 수준을 잘 유지하면서, 위음성의 정도를 줄일 수 있는 통계적 방법이 필요한데, 이것이 multi-step을 이용한 보정 방법이다. Multi-step을 이용한 보정 방법은 step-down procedure(Holm's procedure)과 step-up procedure(Hochberg procedure)으로 나눌 수 있으며, 집단 오류(family wise error)는 그대로 두면서 위음성을 줄이는 방법으로 알려져 있으며, 고전적인 Bonferroni의 방법보다 더 효율적이다. Multi-step 보정 방법은 모든 검정에서 나온 p 값을 정렬(sorting)한 후, 각 검정마다 각기 다른 p 값을 적용시키는 방법이다. 한편, Westfall과 Young 등은 permutation method를 이용한 새로운 방법을 제시하였는데, 이는 유전자 발현 정도를 각 군과의 연관성을 끊어버린 후 무작위로 섞은 후 임의로 두 군에 배정을 한 후 검정을 하는 방식이다.[13] Multi-step 보정법 중 이와 비슷한 step-down procedure에 대해서 소개를 하면, 주어진 m개의 유전자에 대한 t 검정를 시행하고, 이러한 t 검정의 값을 차례로 나열을 하게 되고, 이에 따라 귀무가설도 t 검정의 값에 따라 아래와 같이 재배치한다.[14]

Step-Down Procedure : $T_{r1} \geq \ldots \geq T_{rm}$

여기에 permutation method를 이용하는데, b번째의 permutation을(b=1, …, B) 시행된 데이터에서 $U_{bj} = \max_{j' = j,\ldots m} t$ 라고 가정하면,

$P_{rj} \approx \frac{\sum_b^B I(U_{bj} \geq t_{rj})}{B}$ 으로 정의가 된다.

집단별 오류율이 아닌 위발견율을 줄이는 방법도 최근 많이 사용되고 있는데, 이는 앞에서 설명한 다중 검정에서 위양성, 위음성의 오류를 줄이는 것이 아니라 귀무가설을 기각한 검정 중 틀린 것(선택된 유전자 중에서 잘못 선택된 유전자의 개수)의 비율을 줄이는 데 개념을 두고 있다. 최근 위발견율 통제에 많이 쓰이는 방법 중 하나인 Storey가 제안한 방법에 대해서 살펴보면, 대립가설을 받아들인 유전자의 수를 R이라고 하면, 이는 R개의 유전자를 선택한 것이 되며, R_0는 귀무가설이 참이면서 검정 결과 귀무가설이 기각된 유전자의 개수를 말한다. m개의 유전자의 상관관계가 거의 없다고 한다면 위음성율은 $E\left(\frac{Ro}{R}\right) = E\left(\frac{Ro}{R0+R1}\right)$ 확률로 근사될 수 있다.[9] 이 확률은 H_j가 참일 확률과 H_j가 참일 때 H_i를 기각할 확률의 곱으로 나타나게 된다. H_j가 참일 확률의 합은 m_0이고, H_j가 참일 때 H_i를 기각할 확률은 α가 되므로

$$R_0 = \sum_{j=1}^{m} I\left(Hj\ true, Hj\ rejected\right),\ for\ large\ m,\ is\ approximated\ by$$

$$\sum_{j=1}^{m} Pr\left(Hj\ true, Hj\ rejected\right)$$

$$= \sum_{j=1}^{m} Pr\left(Hj\ true\right)\Pr(Hj\ rejected\ |Hj\) = m_0\alpha$$

$FDR(\alpha) \approx \frac{m_0\alpha}{R(\alpha)}$로 추정할 수 있다. m_0를 추정하기 위해서는 p value distribution에서 0.5보다 큰 임의의 p 값을 λ라고 하면 λ보다 큰 p 값은 귀무가설이 참인 유전자들일 가능성이 많다. λ는 0과 1의 값을 가지게 되고 λ보다 큰 p 값의 개수(#)는 $m_0(1-\lambda)$가 된다. 여기서 m_0는 $\frac{\#(pj > \lambda)}{1-\lambda}$로 추정이 된다.

따라서 $FDR(\alpha) = \frac{m_0\alpha}{R(\alpha)} = \frac{\#(pj > \lambda)\,\alpha}{(1-\lambda)R(\alpha)}$가 됨을 확인할 수 있다.

결 론

한 유전자에 대한 가설 검정을 동시에 여러 개 수행할 경우 귀무가설이 참인데도 불구하고 귀무가설을 기각할 확률이 증가할 수 있다. 최근에는 많은 유전자 분석, 임상 시험, 신약 개발 등의 연구에 있어 다양한 다중 검정법이 사용되는데, 이는 다중 비교 분석을 통해 가설의 검정에 필요한 제1종 오류을 보정하여 검정력을 높이려는 다양한 방법들이 제시되고 있으므로 이러한 통계적 접근에 대한 이해가 필요하다.

Acknowledgments

None.

ORCID

Jeong-Seok Choi https://orcid.org/0000-0001-9669-2141

REFERENCES

1) Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286(5439):531-7.
2) Kirkham EM, Weaver EM. A review of multiple hypothesis testing in otolaryngology literature. Laryngoscope 2015;125(3):599-603.
3) Lawrence J. Familywise and per-family error rates of multiple comparison procedures. Stat Med 2019;38(19):3586-98.
4) Jung SH, Jang W. How accurately can we control the FDR in analyzing microarray data? Bioinformatics 2006;22(14):1730-6.
5) Armstrong RA. When to use the Bonferroni correction. Ophthalmic Physiol Opt 2014;34(5):502-8.
6) Dudoit S, van der Laan MJ, Pollard KS. Multiple testing. Part I. Single-step procedures for control of general type I error rates. Stat Appl Genet Mol Biol 2004;3:Article13.
7) van der Laan MJ, Dudoit S, Pollard KS. Multiple testing. Part II. Step-down procedures for control of the family-wise error rate. Stat Appl Genet Mol Biol. 2004;3:Article14.
8) Solari A, Goeman JJ. Minimally adaptive BH: A tiny but uniform improvement of the procedure of Benjamini and Hochberg. Biom J 2017;59(4):776-80.
9) Chen X, Robinson DG, Storey JD. The functional false discovery rate with applications to genomics. Biostatistics 2019:kxz010.
10) Finch R. Multiple testing problems in pharmaceutical statistics. Pharm Stat 2014:14;153-4.
11) Glickman ME, Rao SR, Schultz MR. False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies. J Clin Epidemiol 2014;67(8):850-7.
12) Bland JM, Altman DG. Multiple significance tests: The Bonferroni method. BMJ 1995;310(6973):170.
13) Troendle JF, Westfall PH. Permutational multiple testing adjustments with multivariate multiple group data. J Stat Plan Inference 2011;141(6):2021-9.
14) Lin D, Shkedy Z, Yekutieli D, Burzykowski T, Göhlmann HW, De Bondt A, et al. Testing for trends in dose-response microarray experiments: A comparison of several testing procedures, multiplicity and resampling-based inference. Stat Appl Genet Mol Biol 2007;6:Article26.

정답 및 해설

답 ①

해 설 ① 우선 굴곡형 내시경을 이용해서 외래 조직 검사를 시행하고 어려울 경우 전신마취하에서 조직 검사를 시행한다. ② 다른 염증성 질환이나 양성 질환이 배제되지 않은 상태에서 치료로 들어가는 것은 바람직하지 않다. ③ 기도가 충분해 보이더라고 전신마취 유도 시에 호흡곤란이 올 수 있고, 항암 치료나 방사선 치료 중에도 호흡곤란이 올 수 있으므로 기도 유지에 신경써야 하고 필요 시 기관절개 수술을 적절한 시점에 시행해야 한다. ④ 종양의 갑상연골의 침투로 T4a로 생각되지만 항암방사선 치료를 초치료로 선택하고 필요 시 후두전절제 수술은 구제 수술로 시행 가능하다. ⑤ 방사선 치료만으로는 충분한 치료 효과를 기대할 수 없고 동시항암방사선 치료나 후두전절제 수술이 필요하다.
참고 문헌: 대한이비인후과학회. 이비인후과학:두경부. 파주: 군자출판사;2018. p.489–506.