Special Report

J Korean Soc Radiol 2023;84(4):866-878
https://doi.org/10.3348/jksr.2022.0108
eISSN 2951-0805

JOURNAL of THE KOREAN SOCIETY of RADIOLOGY

Statistical Mistakes Commonly Made When Writing Medical Articles

의학 논문 작성 시 발생하는 흔한 통계적 오류

Soyoung Jeon, PhD, Juyeon Yang, MS, Hye Sun Lee, PhD*

Biostatistics Collaboration Unit, Yonsei University College of Medicine, Seoul, Korea

ORCID iDs
Soyoung Jeon https://orcid.org/0000-0002-9916-1917
Juyeon Yang https://orcid.org/0000-0002-7621-5150
Hye Sun Lee https://orcid.org/0000-0001-6328-6948

Received July 29, 2022
Revised December 2, 2022
Accepted February 26, 2023

***Corresponding author**
Hye Sun Lee, PhD
Biostatistics Collaboration Unit,
Yonsei University
College of Medicine,
20 Eonju-ro 63-gil, Gangnam-gu,
Seoul 06229, Korea.

Tel 82-2-2019-5401
Fax 82-2-2019-5210
E-mail hslee1@yuhs.ac

Statistical analysis is an essential component of the medical writing process for research-related articles. Although the importance of statistical testing is emphasized, statistical mistakes continue to appear in journal articles. Major statistical mistakes can occur in any of the three different stages of medical writing, including in the design stage, analysis stage, and interpretation stage. In the design stage, mistakes occur if there is a lack of specificity regarding the research hypothesis or data collection and analysis plans. Discrepancies in the analysis stage occur if the purpose of the study and characteristics of the data are not sufficiently considered, or when an inappropriate analytic procedure is followed. After performing the analysis, the results are interpreted, and an article is written. Statistical analysis mistakes can occur if the underlying methods are incorrectly written or if the results are misinterpreted. In this paper, we describe the statistical mistakes that commonly occur in medical research-related articles and provide advice with the aim to help readers reduce, resolve, and avoid these mistakes in the future.

Index terms Statistical Mistake; Study Design; Statistical Analysis; Result Interpretation

서론

대부분의 의학 논문에는 통계학이 포함되어 있고 연구의 설계와 분석 그리고 결과의 해석까지 모든 과정에서 활용되고 있다. 논문에서 통계 분석을 수행하는 이유는 표본의 결과를 토대로 모집단의 특성을 일반화하여 추론하는 것이고 다른 연구자가 동일한 방법으로 연구를 진행했을 때 같은 결과가 나올 수 있도록 하는 것이다(1). 또한 통계 분석을 통하여 도출된 객관적인 수치를 이용하면 정확한 내용의 논문을 작성할 수 있다.

만약 적합하지 않은 방법으로 연구를 설계하고 분석하면 그 논문의 내용을 신뢰할 수 없기 때문에 논문의 질이 떨어지게 된다(2). 따라서 통계학적으로 정확한 방법을 사용하여 논문을 작성하는 것이 중요하다. 많은 임상시험 및 의약품 관련 기관에서는 통계학의 중요성을 인지하고 통계 가이드라인을 발표하여 올바른 방법으로 연구가 수행되도록 권고하고 있다.

의학 논문에서의 통계학에 대한 중요성은 계속 언급되어 왔지만 여전히 많은 연구에서 통계적 오류가 발생하고 있다. 출판된 논문도 통계적인 부분에 대해 충분한 검토가 이루어지지 않았다면 오류가 있을 수 있다. 의학 논문에서 흔히 발생할 수 있는 통계적 오류는 설계 단계에서의 오류, 분석 단계에서의 오류, 작성과 해석 단계에서의 오류로 분류할 수 있다(3). 본 논문에서는 세 가지 오류에 대하여 고찰하고, 의학 논문의 통계적 오류를 줄이는데 기여하고자 한다.

설계 단계에서 발생하는 오류

불명확한 가설

통계적 설계는 연구의 주제에 맞는 자료를 수집하여 적합한 통계 분석을 통해 객관적인 결과를 도출할 수 있도록 설계하는 것을 말한다(4). 연구의 주요 목적에 맞게 연구를 수행하려면 가설이 명확해야 하고, 가설이 명확하지 않으면 연구를 통해 주장하고자 하는 내용을 파악하기 어렵기 때문에 독자들에게 혼란을 야기할 수 있다. 하나의 연구 주제에 여러 개의 가설이 포함되어 있는 경우에도 혼란을 줄 수 있다. 따라서 연구를 통해 입증하고자 하는 가설을 명확하게 세우고 여러 개의 가설이 있다면 주된 가설과 그 외 가설로 구분해야 한다.

적합하지 않은 표본수 계산

설계 단계에서 발생할 수 있는 오류 중 가장 흔하게 발생하는 것은 표본수에 대한 오류이다. 대부분의 연구에서는 전체 대상자를 모두 모집하기 어렵기 때문에 전체 대상자를 대표할 수 있는 표본을 설정해야 한다(5).

표본수가 적으면 검정력이 떨어지게 되고 실제 모집단에서 가설 검정 결과가 유의하더라도 표본에서는 유의하지 않을 수 있다(6). 예를 들어, 100명의 표본집단과 20명의 표본집단에서 독립된 두 군간 유병률의 차이를 검정하는 상황을 가정해 보겠다. 100명의 표본집단에서 한 군이 총 75명이었고 그중 10명이 질병을 가지고 있으므로 유병률이 13.3%였다. 다른 군은 총 25명이었고 그중 10명이 질병을 가지고 있으므로 유병률이 40%였다. 이 상황에서 카이제곱 검정을 이용해 두 군간 유병률이 차이가 있는지 검정해 보면 p값은 0.004로 통계적으로 유의한 차이를 보였다. 20명의 표본집단에서는 한 군이 총 15명이었고 그중 2명이 질병을 가지고 있어서 유병률이 13.3%였다. 다른 군은 총 5명이었고 그중 2명이 질병을 가지고 있어서 유병률이 40%였다. 이 상황에서 피셔의 정확한 검정을 이용해 두 군간 유병률이 차이가 있는지 검정해 보면 p값은 0.249로 통계적으로 유의한 차이를 보이지 않았다(Table 1). 이와 같이 각 군의 비율이 동일하더라도 표본의 수가 다르면 두 집단에서 다른 결과를 얻을 수 있다. 따라서 표본을 이용하여 정확한 결과를 얻기 위해서는 적당한 수를 계산해야 한다.

Table 1. Incorrect Result Owing to Insufficient Sample Size

	Sample 1			Sample 2		
	Treatment Group	Control Group	*p*-Value	Treatment Group	Control Group	*p*-Value
Disease	10 (13.3)	10 (40.0)	0.004	2 (13.3)	2 (40.0)	0.249
Non-disease	65 (86.7)	15 (60.0)		13 (86.7)	3 (60.0)	

Data are number of patients with percentages in parentheses.

표본수를 결정할 때 많이 하는 실수는 선행 연구의 표본수를 그대로 사용하는 것이지만 연구마다 주장하고자 하는 가설과 자료의 특성 등이 다르기 때문에 이렇게 결정한 표본수는 적합하지 않다. 표본수는 일반적으로 연구의 일차 유효성 평가변수에 근거하여 산출한다. 일차 유효성 평가변수의 속성을 파악하여 통계 분석법을 결정하고 그에 맞는 방법 및 공식을 활용하여 표본수를 계산해야 한다. 표본수를 계산할 때 필요한 사항에는 제1종 오류, 검정력, 효과 크기가 있다. 제1종 오류는 일반적으로 0.05로 설정하고 검정력은 0.8 또는 0.9로 설정한다. 효과 크기는 연구자가 밝히고자 하는 최소한의 유의한 차이의 정도를 의미하고 일반적으로 선행 연구의 결과로 나타난 추정치를 이용하여 계산한다. 이 외에도 연구의 상황에 따라서 각 군의 할당비, 추적 기간, 중도 탈락률 등이 고려될 수 있다.

표본수에 영향을 미치는 요인은 여러 가지가 있다. 두 군의 일차 유효성 평가변수에 대한 차이를 작게 가정하거나 표준편차를 크게 가정할수록 효과 크기도 작아지기 때문에 표본수는 증가한다. 또한 제1종 오류를 작게 설정하거나 검정력을 크게 설정할수록 표본수는 증가한다. 이런 특성을 고려하여 연구자는 선행 연구를 토대로 연구자의 가설을 입증할 수 있는 표본수를 명확하게 계산해야 한다.

측정 도구의 변경

의학 연구에서는 일차 유효성 평가변수를 측정하는 도구를 사전에 결정하여 연구를 진행한다. 만약 하나의 측정 도구로 변수를 측정하는 것이 아니라 중간에 도구를 변경하는 경우 변수가 불명확하게 측정될 수 있다. 어떤 도구를 사용하는지에 따라서 측정값이 달라질 수 있기 때문에 계획 단계에서 측정 도구를 명확하게 결정하고, 결정된 도구만 이용하여 변수를 측정해야 한다. 불가피하게 여러 도구를 활용하게 된다면 도구 간의 관련성, 일치도 등에 대한 설명이 제시되어야 한다.

무작위배정 및 눈가림법의 불명확화

무작위배정은 연구 대상자에게 치료 방법을 배정할 때 연구자의 의도가 개입되지 않도록 무작위로 배정하는 방법이다(7). 적합하지 않은 방법으로 무작위배정을 실시하면 의도적 혹은 비의도적으로 선택바이어스(Selection bias)가 발생하여 비뚤림이 발생한다. 또한 눈가림법이 적합하게 수행되지 않으면 두 군의 특성에 불균형이 발생하거나, 하나의 치료 방법에 대상자가 상대적으로 더 많이 배정되어 비뚤림이 발생한다.

임상 시험에서 연구자와 대상자의 주관에 의한 비뚤림을 배제하기 위해 시험이 종료될 때까지

대상자가 어떤 치료 방법을 받았는지 알리지 않는 것을 눈가림법이라 한다. 눈가림법이 적합하지 않거나 잘 수행되지 않는 경우에 연구자의 주관이 개입되어 비뚤림이 발생하여 신뢰성이 떨어지게 된다. 그러므로 사전에 무작위배정과 눈가림법을 적합하게 계획하고 수행해야 한다.

적절한 계획 없이 중간 분석 시행

일반적으로 임상시험에서는 시험이 종료된 후에 수집된 자료를 이용하여 분석을 수행하지만 사전에 중간 분석을 하도록 계획된 경우 시험 중간에 분석을 수행할 수 있다. 만약 중간 분석을 세 번 실시하게 되면 최종 분석까지 총 네 번의 가설 검정을 수행하는 것이므로 제1종 오류가 증가하게 되어 잘못된 결과를 얻을 수 있다(8). 따라서 중간 분석은 시점이나 횟수 등을 사전에 계획해야 하고, 계획된 대로 정확히 실시해야 한다.

분석 단계에서 발생하는 오류

자료를 잘못 처리

분석 단계에서 흔히 발생하는 오류는 자료를 적합하지 않은 방법으로 처리하는 것이다. 양적 자료를 명확한 이유 없이 단지 통계학적 유의성을 찾기 위해 이분형 또는 순위형 자료로 변환하여 분석하면 오류가 생길 수 있다. 특히 일반적으로 양적 자료보다 이분형 자료로 얻은 결론의 정보 손실이 더 크다고 알려져 있기 때문에 유의해야 한다(9).

검정에 필요한 가정을 보여주지 않거나 언급하지 않는 오류도 종종 발생한다. 선형 회귀 분석은 오차항의 선형성, 정규성, 독립성, 등분산성을 만족하는지 확인하고 분석해야 한다. 콕스 비례위험모형(Cox proportional hazard model)은 비례위험 가정을 검토해야 하고, 반복측정 분산분석은 구형성 가정을 만족하는지 확인하고 분석해야 한다.

기저 시점 특성 비교에서 고려 사항

무작위 대조군 연구에서 군간 기저 시점 특성의 차이를 분석하는 것도 분석 단계에서 발생하는 오류 중 하나이다. 무작위 대조군 연구는 대상자들에게 무작위로 군을 배정하여, 연구 결과에 영향을 줄 수 있는 요인에 대해 군간 차이가 없도록 설계된 연구이다(10). 이러한 연구에서 군간 기저 시점 특성을 비교하는 이유는 무작위배정이 제대로 되었는지 확인하는 것이다. 하지만 무작위배정이 적합하게 수행되었다면 동일한 모집단에서 두 군을 추출한 것과 같기 때문에 군간 차이를 통계적으로 검정하는 것은 불필요하다(11). 두 군의 차이를 검정한다는 것은 군간 차이가 있을 것으로 가정하는 것이고, 동일한 모집단에서 추출된 두 군은 차이가 없을 것이기 때문이다.

기존에 수집된 자료를 가지고 중재효과를 비교하는 성향점수 분석 연구는 성향점수 매칭(propensity score matching) 후에 군간 기저 시점 특성의 차이를 통계적으로 검정하는 것이다. 관찰연구에서 특정 질병을 가진 환자군과 질병이 없는 대조군을 비교한다면 두 군의 기저 시점 특성이 다를 가능성이 크고 이것이 연구 결과에도 영향을 줄 수 있다. 두 군에서 일차 유효성 평가변수에 차이가 있는 것으로 결과가 얻어지더라도 이 결과가 정말 일차 유효성 평가변수의 차이인지 또는

기저 시점 특성의 차이에 기인한 것인지 확신할 수 없다. 따라서 군간 기저 시점 특성의 차이를 줄여서 일차 유효성 평가변수를 직접적으로 비교할 수 있는 분석 방법들이 고안되었고 그중 많이 이용되는 방법이 성향점수 매칭이다. 성향점수는 기저 시점 특성에 의해 환자군에 속하게 될 확률을 의미하고, 두 군에서 성향점수가 동일하거나 비슷한 대상자들을 짝짓는 것을 성향점수 매칭이라 한다(12). 매칭 후에 두 군에서 기저 시점 특성에 차이가 없는지 알아보기 위해 통계적으로 검정을 하는 것은 적합한 방법이 아니다. 통계적 검정은 일반적으로 *p*값을 기준으로 통계학적 유의성을 판단한다. 여기서 표본수가 감소하면 검정력이 작아지므로 *p*값이 커지게 된다. 성향점수 매칭을 하면 짝 지어지지 않은 자료는 제외되기 때문에 표본수가 감소하여 *p*값이 커지므로 두 군간 기저 시점 특성에 유의한 차이가 존재하지 않는 것으로 결과가 얻어질 수 있다(13). 따라서 매칭 후에 두 군간 기저 시점 특성의 차이를 확인하기 위해서는 표본수에 영향을 받지 않는 방법을 이용해야 한다.

무작위 대조군 연구와 성향점수 매칭을 이용한 연구에서 표본수에 영향을 받지 않고 두 군의 기저 시점 특성을 비교할 때 많이 사용되는 방법은 표준화된 평균 차이(standardized difference; 이하 SMD)와 표준화된 비율 차이(standardized proportion difference; 이하 SPD)를 계산하는 것이다. 표준화된 차이는 표본수와 무관하게 계산이 가능하고 두 군간 기저 시점 특성에 대한 차이의 정도를 확인할 수 있다(14). 양적변수에 대한 표준화된 평균 차이는 두 군의 분산이 다르다고 가정할 경우 각 군의 평균과 표준편차로 계산할 수 있다[1]. 두 군의 분산이 같다고 가정할 경우 각 군의 평균, 표준편차, 표본수로 계산할 수 있다[2].

$$SMD = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2 + s_2^2}{2}}} \quad [1]$$

$$SMD = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}} \quad [2]$$

여기서 $\bar{x}_1$과 $\bar{x}_2$는 두 군의 평균, s_1과 s_2는 두 군의 표준편차를 의미한다. 질적변수에 대한 표준화된 비율 차이는 각 군의 비율로 계산할 수 있다[3].

$$SPD = \frac{(p_1 - p_2)}{\sqrt{\frac{p_1(1 - p_1) + p_2(1 - p_2)}{2}}} \quad [3]$$

여기서 p_1과 p_2는 두 군의 비율을 의미한다. 표준화된 차이가 0.1보다 작으면 두 군의 균형이 맞는 것으로 해석이 가능하다.

적합하지 않은 분석 방법

연구마다 분석하고자 하는 변수의 특성과 목적이 모두 다르기 때문에 연구 내용에 따라서 적합한 방법을 선택하여 분석해야 한다. 만약 적합하지 않은 방법으로 분석을 하게 되면 잘못된 연구 결과가 도출될 수 있다. 이러한 오류 중에서 많이 발생하는 오류는 독립적인 자료에 대한 분석 방

법과 짝지은 자료에 대한 분석 방법을 혼동하는 것이다. 자료가 독립적인지 짝 지어져 있는지에 따라서 분석하는 방법이 달라지므로 자료의 특성을 잘 판단하여 분석해야 한다.

또한 모수적 방법과 비모수적 방법을 혼동하는 경우가 있는데 자료가 정규성을 만족하면 모수적 방법으로 분석해야 하고 정규성을 만족하지 않으면 비모수적 방법으로 분석해야 한다. 추가적으로 자료의 개수가 충분히 크면 중심극한 정리와 대수의 법칙에 의해 모수적 방법으로 분석이 가능하고, 자료의 개수가 충분하지 않고 치우쳐진 형태의 분포를 가지면 비모수 방법으로 분석해야 한다. 일반적으로 모수적 방법보다 비모수적 방법의 검정력이 떨어지기 때문에 분석 방법을 선택할 때 유의해야 한다. 따라서 분석의 목적, 자료의 특성, 정규성 만족 여부 등을 고려하여 적합한 분석 방법을 선택하는 것이 중요하다(Table 2).

무작위 대조군 연구에서 특정 치료제에 대한 치료 효과를 확인할 때 각 군내에서 치료 전과 후의 일차 유효성 평가변수를 비교하는 경우가 있다. 이때 시험군에서는 치료 전보다 후의 일차 유효성 평가변수가 유의하게 향상되었고 대조군에서는 치료 전과 후의 차이가 없다고 해서 시험군이 대조군보다 치료 효과가 더 크다고 할 수 없다. 이렇게 분석하는 것은 두 군을 비교한 것이 아니고 시험군 내에서의 치료 효과만을 분석한 것이기 때문이다. 시험군과 대조군을 비교하기 위해서는 두 군의 치료 전과 후의 변화량을 직접적으로 비교해야 한다(3).

자료 분석군과 결측치, 이상치 처리에서 고려 사항

임상시험에서 모든 대상자가 계획된 내용에 맞게 끝까지 시험에 참여할 수 있다면 결측치나 이상치가 없는 완전한 자료를 수집할 수 있다. 하지만 현실적으로 이런 경우는 드물기 때문에 연구 대상자 중에서 분석에 이용할 분석군을 결정하고 결측치와 이상치에 대한 처리 방법을 결정해야 한다. 같은 자료를 이용하여 분석하더라도 분석군과 결측치, 이상치를 어떻게 처리하느냐에 따라 다른 결과가 도출될 수 있고 그로 인한 오류가 발생할 수 있다.

분석군을 결정하는 방법 중에서 많이 사용되는 방법은 intent to treat (이하 ITT), per protocol (이하 PP), full analysis set (이하 FAS)이다. ITT는 사전에 배정된 대상자를 모두 분석에 포함하는 방법이고, PP는 계획서의 내용에 맞게 끝까지 시험을 완료한 대상자만 분석에 포함하고 시험 중에 탈락하거나 규칙을 위반한 대상자는 제외하는 방법이다(15). FAS는 ITT 원칙을 최대한 지키고,

Table 2. Analysis Methods Used in Medical Papers According to Normality Assumption and Analysis Purpose

Purpose of Analysis	Analysis Methods	
	Satisfy Normality Assumption	Not Satisfy Normality Assumption
Comparison of means or medians of two independent groups	Two sample *t*-test	Mann Whitney U test, Wilcoxon rank sum test
Comparison of means or medians of two paired groups	Paired *t*-test	Wilcoxon signed rank test
Comparison of means or medians of more than 2 independent groups	One way ANOVA	Kruskal-Wallis test
Correlation of two continuous variables	Pearson correlation	Spearman correlation
Comparison of proportions of two independent groups	Chi-square test (Fisher's exact test)	
Comparison of proportions of two paired groups	McNemar test	

타당한 이유가 있는 대상자만 제외하는 방법이다. 분석군에 따라서 분석에 사용하는 자료와 분석 결과가 달라지기 때문에 이러한 방법들을 이용하여 적합한 분석군을 선정해야 한다.

결측치 또는 이상치는 측정자의 실수나 측정 기기의 오작동 등 여러 가지 이유로 발생되고, 자료에서 삭제하거나 다른 값으로 대체하는 방법으로 처리할 수 있다. 처리하는 방법에 따라서 자료의 값이 달라지므로 어떤 방법을 사용할지 사전에 정하여 분석해야 한다.

잘못된 다중 검정

통계적 가설 검정을 할 때 우연에 의하여 귀무가설을 잘못 기각하는 것을 제1종 오류라고 하고, 제1종 오류를 범할 확률의 최대 허용치를 유의수준이라 한다(16). 일반적으로 유의수준은 5%로 지정하고 p값이 유의수준보다 작으면 증명하고자 하는 대립 가설을 채택한다. 하지만 많은 가설 검정을 동시에 수행하게 되면 제1종 오류가 생길 확률은 검정의 개수가 증가함에 따라 급격하게 커지게 된다(17). 즉, 검정의 수가 많아질수록 그만큼 우연에 의하여 잘못된 결과가 발생할 확률이 높아지게 된다.

유의한 결과를 얻기 위해 여러 가지 방식으로 변수를 변환하여 분석하고 그중 원하는 결과만 사용하는 것도 다중 검정에 해당한다. 예를 들어, 나이가 질병 발생에 영향을 주는지를 확인할 때 상대적으로 나이가 적은 군과 많은 군으로 나누어 질병 발생률을 비교할 수 있다. 이때 20세, 30세, 40세 등 여러 가지 값을 기준으로 두 군을 나누고, 모두 분석을 해본 후 발생률의 차이가 가장 큰 결과를 선택하면 제1종 오류가 높아지므로 결과에 대한 신뢰도가 떨어지게 된다.

다중 검정을 할 경우에 유의수준을 조정하여 제1종 오류를 제어할 수 있는 몇 가지 방법이 있다. 이런 방법 중 대표적인 방법 중 하나는 본페로니(Bonferroni) 방법이다. 본페로니 방법은 유의수준을 가설 검정의 횟수로 나누어 제1종 오류를 유의수준 이하로 통제하는 방법이다. 다소 보수적인 방법으로 이를 완화한 홈-본페로니(Holm-Bonferroni), 호치버그-본페로니(Hochberg-Bonferroni)도 종종 이용된다.

해석 단계에서의 오류

잘못된 분석 방법 및 분포 기술

적합한 방법으로 연구의 설계와 분석을 수행하는 것만큼 결과에 대해 정확히 해석하고 서술하는 것도 중요하다. 논문을 작성할 때 분석의 결과를 서술하기 전에 먼저 분석 방법에 대해 기술한다. 사용하지 않은 분석 방법을 기술하거나 사용된 분석 방법에 대한 서술이 충분하지 않으면 분석 결과가 어떻게 도출되었는지 명확하게 알 수 없다.

자료의 분포에 대한 서술에서 평균 또는 중위수 등의 대표값만 제시하고 표준편차와 사분위 범위 같은 산포 또는 중심위치의 측도와 관련된 값을 제시하지 않으면 분포를 예측하기 어렵다. 비대칭인 분포를 가진 자료에서는 평균과 표준편차로 분포를 나타내는 것은 적합하지 않고 중위수와 사분위수로 표현하는 것이 더 적합하다. 또한 평균과 함께 표준편차를 나타내지 않고, 값이 작다는 이유로 표준오차를 나타내는 경우가 있는데 이는 적합한 방법이 아니다. 표준편차는 표본자

료의 값들이 표본평균으로부터 어느 정도 떨어져 있는지를 측정하는 척도이고, 표준오차는 모수 추정량의 정확성을 나타내주는 척도이다. 목적에 맞게 두 값을 구분하여 서술해야 한다. 이 외에도 분석 방법과 결과를 서술할 때는 자세하고 명확하게 해야 한다.

연관성과 인과 관계, 연관성과 예측력 고려 사항

해석 단계에서 발생할 수 있는 오류 중 하나는 연관성과 인과 관계를 혼동하여 해석하는 것이다. 두 변수가 연관성이 있는지를 알아보기 위해 상관 분석을 했다면 이 결과는 연관성에 대한 것만 해석할 수 있고 한 변수가 다른 변수의 원인이 되는지는 확인할 수 없다(18). 인과 관계에 대해 해석하고 싶을 때는 회귀 분석 등의 다른 방법으로 분석해야 한다.

의학 논문에서 질병과의 연관성에 대해 분석할 때 로지스틱 회귀분석(Logistic regression)이 많이 이용된다. 로지스틱 회귀분석을 통해 오즈비(Odds ratio)를 구하여 연관성, 인과성을 판단하고, 이것을 예측력이 높은 것으로 해석하는 오류도 발생한다. 임의의 독립변수와 질병 발생의 연관성을 알아보기 위해 로지스틱 회귀 모형을 이용하여 오즈비를 산출했다고 가정하겠다. 오즈비는 1.5, 25, 350일 때 모두 유의하여 연관성이 있는 것으로 나타났지만 실제 자료의 분포를 보면 오즈비가 350이 되어야 질병 발생군과 아닌 군을 잘 판별할 수 있다(Fig. 1) (19). 따라서 오즈비가 유의하다는 것은 연관성이 있다는 것으로 해석할 수 있지만 예측력이 높다는 것으로 해석하기는 어렵다. 예측력을 측정하고 싶은 경우 receiver operating characteristic 곡선을 이용할 수 있고, 민감도와 특이도를 통해서도 판단할 수 있다.

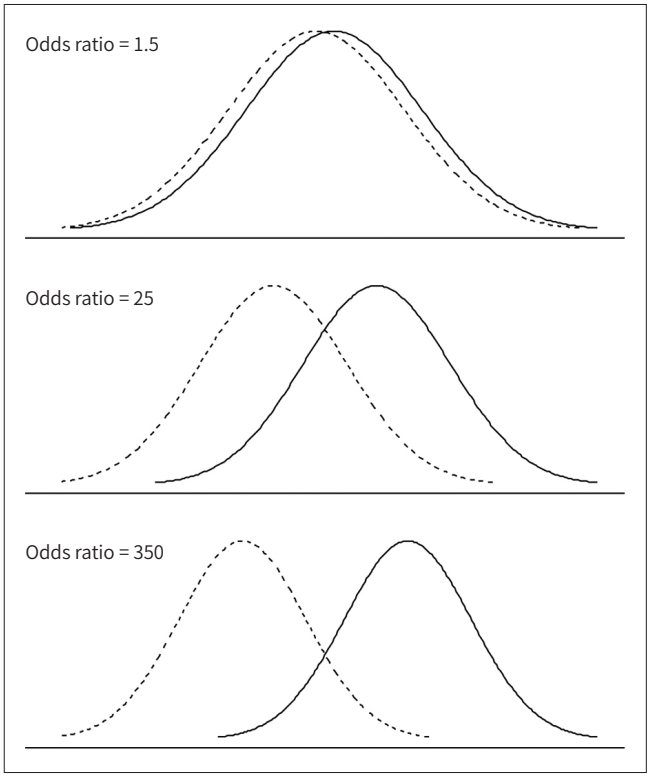

Fig. 1. Probability distributions of independent variable in disease group (solid curves) and control group (dashed curves) consistent with the logistic regression model.

양성예측도(PPV)와 음성예측도(NPV)

양성예측도(positive predictive value; 이하 PPV)와 음성예측도(negative predictive value; 이하 NPV)를 오남용하는 오류도 발생한다. 두 예측도는 진단의 정확도를 나타내는 지표이고 영상의학 연구에서 많이 이용된다. 양성예측도는 어떤 질병에 대한 진단 결과가 양성인 대상자들 중 실제로 질병이 있는 대상자들의 비율이다. 환자-대조군 연구에서는 일반적으로 질병의 유병률이 모집단보다 높게 나타나고 이것이 잘못된 양성예측도를 얻게 되는 원인이 된다. 유병률이 높다는 것은 실제로 질병이 있는 대상자가 더 많다는 것이고 그로 인해 양성예측도가 모집단에 비해 높아지게 된다(20).

예를 들어, 영상의학과에 내원하여 진단을 받은 총 600명의 환자 중 100명이 질병이 있고, 질병이 있는 군에서 0.5배 그리고 질병이 없는 군에서 0.1배를 표본 추출하는 상황을 가정해 보겠다. 만약 모집단에서 질병이 있다고 진단받은 환자가 총 190명이고 이 중 90명은 실제로 질병이 있고 100명은 질병이 없다면 양성예측도는 90/(90 + 100) = 0.47이다. 반면 표본집단에서는 질병이 있는 군과 없는 군에서 추출된 비율이 다르기 때문에 양성예측도를 구해보면 (0.5 × 90)/(0.5 × 90 + 0.1 × 100) = 0.82로 모집단과 차이가 있다(Table 3).

음성예측도는 진단 결과가 음성인 대상자들 중 실제로 질병이 없는 대상자들의 비율이고 양성예측도와 마찬가지로 유병률에 영향을 받는다. 위 예시의 모집단에서 질병이 없다고 진단받은 환자가 총 410명이고 이 중 400명은 실제로 질병이 없고 10명은 질병이 있다면 음성예측도는 400/(400 + 10) = 0.98이다. 반면 표본집단에서는 질병이 없는 군과 있는 군에서 추출된 비율이 다르기 때문에 음성예측도를 구해보면 (0.1 × 400)/(0.1 × 400 + 0.5 × 10) = 0.89로 모집단과 차이가 있다(Table 3).

표본집단의 유병률이 아닌 모집단을 대표할 수 있는 외부 자료의 유병률을 이용하여 양성예측도와 음성예측도를 구하면 이와 같은 오류를 줄일 수 있다. 두 예측도는 베이즈 정리에 의하여 민감도, 특이도, 유병률로 계산이 가능하다[4, 5] (21).

$$PPV = \frac{\text{prevalence} \times \text{sensitivity}}{\text{prevalence} \times \text{sensitivity} + (1 - \text{prevalence}) \times (1 - \text{specificity})} \quad [4]$$

$$NPV = \frac{(1 - \text{prevalence}) \times \text{specificity}}{\text{prevalence} \times (1 - \text{sensitivity}) + (1 - \text{prevalence}) \times \text{specificity}} \quad [5]$$

위 예시의 표본집단에서 민감도는 실제 질병이 있는 사람 중 질병이 있다고 진단받은 사람의 비율이므로 45/50 = 0.9이고 특이도는 실제 질병이 없는 사람 중 질병이 없다고 진단받은 사람의 비

Table 3. Incorrect PPV and NPV in Case-Control Study

	Population			Sample		
	Disease	Non-Disease	PPV, NPV	Disease	Non-Disease	PPV, NPV
Positive test result	90	100	0.47, 0.98	0.5 × 90 = 45	0.1 × 100 = 10	0.82, 0.89
Negative test result	10	400		0.5 × 10 = 5	0.1 × 400 = 40	

NPV = negative predictive value, PPV = positive predictive value

율이므로 40/50 = 0.8이다. 모집단의 전체 대상자는 총 600명이고 이 중 실제 질병이 있는 대상자가 100명이기 때문에 유병률은 100/600 = 0.167이다. 이를 식에 대입하여 계산하면 양성예측도는 0.167 × 0.9/(0.167 × 0.9 + [1 - 0.167] × [1 - 0.8]) = 0.47이고 음성예측도는 (1 - 0.167) × 0.8/(0.167 × [1 - 0.9] + [1 - 0.167] × 0.8) = 0.98로 모집단과 같은 값을 산출할 수 있다. 그러므로 선행 연구를 통해 유병률을 모집단과 유사한 값으로 대입하면 모집단과 비슷한 양성예측도와 음성예측도를 추정할 수 있다.

통계학적 유의성 해석 고려 사항

가설 검정의 통계학적 유의성에 대한 해석을 잘못하는 오류도 발생할 수 있다. 환자-대조군 연구에서 두 군간 질병 발생률에 차이가 있는지를 검정하여 *p*값이 유의수준보다 크게 얻어졌다고 가정해 보겠다. 이때 두 군의 차이가 없으므로 두 군의 발생률이 같다고 잘못 해석하는 경우가 있다. 두 군간 질병 발생률에 차이가 없다는 가설이 기각되지 않았지만 그렇다고 두 군의 발생률이 같다고 해석할 수 없다. 두 군의 차이가 있지 않다는 것이 두 군이 같다는 것과 동일하지 않기 때문이다. 따라서 가설 검정에서 구한 *p*값을 해석할 때는 가설을 기준으로 정확하게 해석해야 한다.

의학 논문을 작성할 때 *p*값을 구체적으로 제시하지 않고 *p*값이 유의수준보다 크거나 작다는 것을 기호로만 표시하는 경우가 있다. *p*값이 유의하다면 값의 크기를 기준으로 귀무가설을 기각할 수 있는 정도를 파악할 수 있기 때문에 값을 정확하게 명시해야 한다. Efron과 Tibshirani (22)는 *p*값의 크기에 따라서 귀무가설을 기각할 수 있는 정도에 대한 해석 방법을 제안했다(Table 4).

통계학적으로 *p*값이 의미가 있다고 하더라도 이는 통계학적 의미이므로 임상적으로는 의미 있는 수치가 아닐 수 있다. 예를 들어, 두 군간 차이가 있는지를 확인하는 연구에서 *p*값은 유의하지만, 결과 값의 차이가 작아서 임상적인 기준 하에 의미가 있다고 하기 어려운 경우가 있다. 그러므로 단순히 *p*값을 통해 통계학적 의미만 해석하기보다는 통계학적인 측면과 임상적인 측면 모두를 고려해서 해석해야 한다.

결론

대부분의 의학 논문에는 통계학이 포함되어 있고 연구의 설계부터 결과의 해석까지 전체적인 과정에서 사용되고 있다. 논문의 신뢰성을 높이기 위해서는 통계적인 오류 없이 적합한 방법으로 연구를 설계 및 분석하고 정학하게 결과를 해석해야 한다. 통계학의 중요성은 연구자들에게 널리

Table 4. Interpretation of *p*-Value

p-Value	Evidence Against Null Hypothesis
< 0.1	Borderline
< 0.05	Reasonably strong
< 0.025	Strong
< 0.01	Very strong

Table 5. Type of Mistakes in Medical Articles

Stage	Type of Mistakes
Design stage	- Unclear hypothesis
	- Incorrect calculation of sample size
	- Change research instrument
	- Unclear randomization and blinding
	- Omission of interim analysis plan
Analysis stage	- Incorrect data processing
	- Unnecessary comparison of baseline characteristics
	- Incorrect analysis method
	- Incorrect processing of analysis set, missing and outlier
	- Multiple testing problem
Interpretation stage	- Unclear writing of analysis method
	- Confusion between association and causation, association and prediction
	- Incorrect calculation of positive predictive value and negative predictive value
	- Unclear interpretation of significance

알려져 있지만 여전히 많은 의학 논문에서 통계적 오류가 발생하고 있다. 의학 논문을 작성하기 전에 먼저 연구의 목적에 맞는 올바른 연구 설계가 이루어져야 한다. 일차 유효성 평가변수의 속성과 연구의 가설을 고려하여 적당한 표본수를 산출하고 적합한 통계 분석 방법을 선택하여 분석을 수행해야 한다. 분석을 수행한 후에는 분석에 사용된 방법들과 결과에 대해 논문에 명확하게 기술하는 것도 중요하다. 만약 통계적인 내용에 대하여 충분히 고려하지 않고 연구를 진행하면 오류가 발생하여 논문의 신뢰성이 떨어지기 때문에 주의해야 한다.

본 논문에서는 의학 논문에서 흔히 발생하는 통계적 오류에 대해 고찰했다(Table 5). 독자들이 본 논문의 내용을 통하여 논문을 작성할 때 통계적 오류를 줄이는데 도움이 되었으면 한다.

Author Contributions

Conceptualization, L.H.S.; data curation, all authors; formal analysis, all authors; investigation, L.H.S.; methodology, J.S., L.H.S.; visualization, J.S., L.H.S.; writing—original draft, J.S., L.H.S.; and writing—review & editing, Y.J., L.H.S.

Conflicts of Interest

The authors have no potential conflicts of interest to disclose.

Funding

None

REFERENCES

1. Winters R, Winters A, Amedee RG. Statistics: a brief overview. *Ochsner J* 2010;10:213-216
2. Ali Z, Bhaskar SB. Basic statistical tools in research and data analysis. *Indian J Anaesth* 2016;60:662-669
3. Petra G. Common statistical errors in medical publications. Available at. https://www.slideshare.net/AustralianNationalDataService/common-statistical-errors-in-medical-publications. Published 2017. Accessed June 15, 2022
4. Montgomery DC. *Design and analysis of experiments.* 10th ed. Hoboken, NJ: John Wiley & Sons Inc 2017:14-15
5. Shin IH, Yim HW. *Calculation of sample size using excel in clinical research.* 1st ed. Paju: Koonja 2009:13-14
6. Altman DG, Bland JM. Absence of evidence is not evidence of absence. *BMJ* 1995;311:485

7. Yoo KY, Kang DH, Ko KP, Kwak J, Kim YJ, Kim Y, et al. *Research methodology in medicine.* Seoul: Seoul National University Publishing & Cultural Center 2014:206-207
8. Kang SH. *Medical statistics necessary for new drug development.* 2nd ed. Paju: Free Academy 2013:198-199
9. Harrison E, Pius R. Should I convert a continuous variable to a categorical variable? Available at. https://argoshare.is.ed.ac.uk/healthyr_book/should-i-convert-a-continuous-variable-to-a-categorical-variable.html. Published 2020. Accessed January 15, 2021
10. Sibbald B, Roland M. Understanding controlled trials. Why are randomised controlled trials important? *BMJ* 1998;316:201
11. de Boer MR, Waterlander WE, Kuijper LD, Steenhuis IH, Twisk JW. Testing for baseline differences in randomized controlled trials: an unhealthy research behavior that is hard to eradicate. *Int J Behav Nutr Phys Act* 2015;12:4
12. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res* 2011;46:399-424
13. Linden A. Graphical displays for assessing covariate balance in matching studies. *J Eval Clin Pract* 2015;21:242-247
14. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat Med* 2009;28:3083-3107
15. Ranganathan P, Pramesh CS, Aggarwal R. Common pitfalls in statistical analysis: intention-to-treat versus per-protocol analysis. *Perspect Clin Res* 2016;7:144-146
16. Kirkham EM, Weaver EM. A review of multiple hypothesis testing in otolaryngology literature. *Laryngoscope* 2015;125:599-603
17. Shaffer JP. Multiple hypothesis testing. *Annu Rev Psychol* 1995;46:561-584
18. Makin TR, Orban de Xivry JJ. Ten common statistical mistakes to watch out for when writing or reviewing a manuscript. *Elife* 2019;8:e48175
19. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol* 2004;159:882-890
20. Naeger DM, Kohi MP, Webb EM, Phelps A, Ordovas KG, Newman TB. Correctly using sensitivity, specificity, and predictive values in clinical practice: how to avoid three common pitfalls. *AJR Am J Roentgenol* 2013;200:W566-W570
21. Khamis HJ. An application of Bayes' rule to diagnostic test evaluation. *J Diagn Med Sonogr* 1990;6:212-218
22. Efron B, Tibshirani R. *An introduction to the bootstrap.* London: Chapman & Hall 1993:203-204

JOURNAL of
THE KOREAN SOCIETY of
RADIOLOGY

의학 논문 작성 시 발생하는 흔한 통계적 오류

전소영 · 양주연 · 이혜선*

의학 논문을 작성할 때 통계학은 필수적인 요소로 알려져 있고 중요성이 강조되고 있지만 많은 논문에서 통계적 오류가 발생하고 있다. 의학 논문에서 발생할 수 있는 통계적 오류는 설계 단계에서의 오류, 분석 단계에서의 오류, 작성과 해석 단계에서의 오류로 분류할 수 있다. 설계 단계에서는 연구의 가설이나 자료의 수집 및 분석 계획이 명확하지 않으면 오류가 발생한다. 분석 단계에서는 연구의 목적과 자료의 특성을 충분히 고려하지 않고 올바른 분석 방법을 적용하지 않으면 오류가 발생한다. 분석을 수행한 후에는 결과를 해석하여 논문을 작성하게 되고, 이 단계에서 분석 방법을 잘못 작성하거나 결과를 올바르게 해석하지 못하면 오류가 발생한다. 본 논문에서는 의학 논문에서 흔히 발생하는 통계적 오류에 대해 고찰하고 오류를 줄이는데 기여하고자 한다.

연세대학교 의과대학 의학통계실