

학술지 편집인이 알아야 하는 의학통계 리뷰에서의 고려사항

안형진 교수
고려대학교 의과대학
의학통계학교실

Introduction

- 통계학이란
 - 연구의 결론을 객관적으로 도출하기 위하여 자료 (data)를 수집, 처리, 해석하는 학문
- 올바르지 않은 설계 및 통계분석법으로 도출된 결과
 - 신뢰도 하락 → 논문의 질 하락
 - 사회적 비용의 증가
 - 예: 신약과 위약의 비교 임상시험
 - Real efficacy but not significant: 새로운 치료의 기회 박탈
 - No efficacy but significant: 잠재적 부작용에 노출, 후속연구의 진행으로 인한 사회적 비용증가

Introduction

- 의학 학술지에서의 통계학
 - 통계방법을 수행한 경우 통계 리뷰과정이 필요함 (적어도 두번째 리뷰 과정에서라도)
 - 통계리뷰 과정은 의학통계 전공의 전문가가 수행하는 것이 바람직함
 - 통계리뷰 과정에 대한 가이드라인을 학술지에 소개 (체크리스트 포함)

연구설계

- 통계학의 중요한 역할
 - 궁극적으로 연구가 비교할만하고(comparable) 일반화할 만한(generalizable)한지 증명
 - 연구계획 및 설계가 매우 중요
- 연구는 인과관계 추론에서 발생할 수 있는 여러 가지 편향(bias)을 최소화할 수 있도록 설계해야 함.
 - 전향적 무작위 중재연구 (best, 일반화의 문제 있을 수 있음.)
 - 관찰연구 (confounding bias 통제 필요, 규모가 큰 연구가 가능, 일반화의 측면에서 좋을 수 있음)
- 잠재적 교란변수(potential confounders)를 미리 수집

연구설계에서의 오류/고려사항

- RCT에서 무작위 배정의 오류
- 적절하지 않은 대조군의 이용
- 교차설계에서 carry-over effect의 존재
- 적절한 표본수 및 근거
- 응답률의 적절성

연구의 크기

- 표본수 산정은 연구설계 시 필요
- 좋은 연구계획은 임상적 유의성과 통계적 유의성을 동시에 보일 수 있도록 충분히 큰 규모의 연구이어야 하나 임상적 유의성이 없음에도 통계적 유의성을 보일 정도의 너무 큰 연구는 지양하여야 함.
- 표본수 계산에 필요한 요소
 - 통계적 분석방법
 - 유의수준 (일반적으로 0.05)
 - 검정력 (일반적으로 0.8 이상)
 - 연구효과의 크기 (예: 군간 비율차이, 군간 평균차이)
 - 다른 관련 모수 (예: 표준편차 등)
- 준비연구 필요

분석에서의 오류/고려사항

- 분석한 방법의 명확하고 정확한 기술
 - “통계 분석”부분을 기술할 때 독자가 이 부분을 읽고 원자료가 있다면 논문의 결과를 다시 재생산할 정도로 자세하게 기술하는 것이 일반적 원칙임.
- 연구결과의 질은 얼마나 많은 통계분석방법을 사용하였느냐 또는 얼마나 어려운 통계분석방법을 사용하였느냐에 결정되는 것이 아니라 얼마나 적절한 방법을 사용하였느냐에 의존함.

분석에서의 오류/고려사항

- 반복측정 자료에서 상관을 고려하지 않은 분석 (예: unpaired t-test for paired data)
- t-test를 이용한 생존시간 비교 (생존시간은 분포가 skewe이고 중도절단이 있음)
- P-value의 적절하지 않은 해석
- 잘못된 분석 단위(units)

분석에서의 오류/고려사항

- 교란변수(confounders)의 선택 및 보정방법
- 연속형 변수의 범주화
- 제 1종의 오류 통제 어려움
 - 다중비교 방법 사용하지 않음
 - 적절하지 않은 사후-부집단 분석
- 사용한 분석의 가정을 검토하지 않음

분석에서의 오류/고려사항

- 모수적 방법의 적절성
- 분석에 사용한 모든 통계분석법은 명확하고 정확하게 서술해야 함.
- 올바른 통계분석법의 이름을 사용해야 함.
(multivariate analysis for multiple regression, impaired t-test X)
- 일반적이지 않거나 독자에게 익숙하지 않은 통계분석법을 사용한 경우에는 명확한 설명이나 참고문헌을 제시하여야 함.

분석에서의 오류/고려사항

- 적절한 통계방법의 선택 기준
 - 연구의 목적 (가설)
 - 연구설계방법
 - 분석 변수의 수
 - 비교하고자 하는 군의 수
 - 자료의 종류 (연속형, 이분형, 범주형, 생존기간 등)

논문 표현에서의 오류/고려사항

- 자료의 분포적 특징
 - 기술통계 (그림 또는 요약숫자)
 - 연속형 자료: 평균 \pm 표준편차, 중앙값 \pm 사분위범위
 - 범주형 자료: 절대빈도 및 상대빈도 (%)
 - 이상값 및 특이값 검출
 - 특이값의 제외 시 그 이유를 논문에 기술
- 방법(method)부분에 연구설계와 자료수집과정을 자세하게 기술
 - 통계분석방법은 방법의 이름만을 나열하는 것이 아니라 어떤 연구가설을 보이기 위하여 어떤 방법을 사용하였는지 자세히 기술
 - 통계적 유의수준 지정
 - 분석에 사용한 통계프로그램의 명시

논문 표현에서의 오류/고려사항

- 결과는 표에 제시된 숫자를 다시 반복하는 것이 아니라 가능하면 표에 제시된 값들의 질적인 표현에 중점을 둬
- 자료의 분포에 대해 기술하는 경우(일반적으로 표 1)는 SE보다 SD를 제시함.
- 원그림(pie-chart)는 논문에서 사용하지 않음.
- 단지 p-value만 기술함. (적절하지 않음.)
- 군간 비교의 경우 군별 추정치 및 CI뿐만 아니라 군간 차이(또는 OR 등)의 추정치 및 CI를 제시함.

논문 표현에서의 오류/고려사항

- Table or Figure?
- 수행한 분석 가정의 만족(goodness of fit)을 기술해야 하나?
- 너무 적거나 또는 많은 소수점
- 부적절한 P-value의 표현 ($p=NS$, $p=0.000$, p-value 없는 *, ** 등)
- 정확한 숫자 확인 (예: 표에서 총합이 맞지 않는 경우 등)
- 그림으로 결과를 표현하는 경우에는 x축과 y축의 단위를 포함한 변수 설명을 명확히 제시하고 범례(legend)를 구체적으로 표시
- 표에서 약어를 사용하는 경우 주석에 풀어서 제시
- 그림과 표의 핵심은 표나 그림만으로 저자가 표현하고 자 하는 바를 독자가 이해할 수 있도록 함.

결과해석에서의 오류/고려사항

- 유의하지 않음을 차이가 없음이나 영향이 없음으로 해석하는 경우
- 연구 자료 분석결과와 관련 없는 결론
- 제약점을 결론의 시작에 기술하는 것이 아니라 연구의 강점, 논문이 주는 시사점을 먼저 설명한 후 마지막에 연구의 제약점을 기술

결과해석에서의 오류/고려사항

- 유의하지 않은 결과해석 시 신뢰구간을 고려하지 않는 경우 (특히, 연구의 크기가 작은 경우)
- Association vs. Causation
- 연구 전반에 걸쳐 편향의 요소를 충분히 고려하였나?
- P-값의 잘못된 해석 및 over-reliance

Frequently used statistical methods for independent continuous response

Situation	Statistical Methods
One sample	<ol style="list-style-type: none"> 1. Parametric: One sample t-test 2. Non-parametric: Wilcoxon signed rank test
Independent two-sample comparison	<ol style="list-style-type: none"> 1. Parametric: Independent two-sample t-test 2. Non-parametric: Wilcoxon rank sum test (A.K.A. Mann-Whitney U test)
Three or more group Comparison	<ol style="list-style-type: none"> 1. Parametric: analysis of variance(ANOVA) with multiple comparison 2. Non-parametric: Kruskal-Wallis test
Relationship with continuous predictors	<ol style="list-style-type: none"> 1. One predictor: simple linear regression 2. Multiple predictors: multiple linear regression
Relationship with continuous and categorical predictors	<ol style="list-style-type: none"> 1. Analysis of covariance (ANCOVA) or 2. General linear models (GLM)

Frequently used statistical methods for correlated continuous response

Situation	Statistical Methods
Paired data	<ol style="list-style-type: none"><li data-bbox="696 730 1279 778">1. Parametric: paired t-test<li data-bbox="696 847 1957 895">2. Non-parametric: Wilcoxon signed rank test on differences
Repeated measures	<ol style="list-style-type: none"><li data-bbox="696 1050 1704 1098">1. Repeated measures ANOVA (RM-ANOVA) or<li data-bbox="696 1166 1346 1214">2. Linear mixed model (LMM)

Frequently used statistical methods for independent categorical response

Situation	Statistical Methods
Relationship with one categorical predictor	<ol style="list-style-type: none"><li data-bbox="772 676 1532 724">1. Chi-square test with large sample<li data-bbox="772 791 1592 839">2. Fisher's exact test with small sample<li data-bbox="772 906 1491 954">3. Simple linear logistic regression
Relationship with categorical and continuous predictor	<ol style="list-style-type: none"><li data-bbox="772 1070 1809 1118">1. Multiple logistic regression for binary response<li data-bbox="772 1185 2020 1294">2. Multinomial or ordinal logistic regression for categorical response (more than two levels)

Frequently used statistical methods for correlated categorical response

Situation	Statistical Methods
Paired categorical data analysis	1. McNemar's test
Repeated categorical responses	1. Marginal model using generalized estimating equation (GEE) 2. Generalized linear mixed model (GLMM)

Frequently used statistical methods

Other methods

Situation	Statistical Methods
Correlation between two continuous variables	<ol style="list-style-type: none"> 1. Pearson correlation coefficient or 2. Spearman's correlation coefficient
Survival data	<ol style="list-style-type: none"> 1. Kaplan-Meier survival curve 2. Log-rank test 3. Cox's proportional hazard model
일치도 분석	<ol style="list-style-type: none"> 1. Sensitivity, Specificity 2. Kappa measure 3. Concordant correlation coefficient) 4. ROC Analysis including AUROC curve 5. Bland-Altman analysis

Common Statistical Errors

(Tom Lang)

- Reporting measurements with unnecessary precision
- Dividing continuous data into ordinal categories without explaining why or how
- Reporting group means for paired data without reporting within-pair changes
- Not Defining Each Variable in Measurable Terms (e.g., operational definition)
- Not Providing the Level of Measurement of Each Variable

Common Statistical Errors

(Tom Lang)

- Dividing Continuous Data into Ordinal Categories Without Explaining Why or How the Categories Were Created
- Using the Mean and Standard Deviation to Describe Continuous Data That Are Not Normally Distributed
- Using the Standard Error of the Mean (SEM) As a Descriptive Statistic
- Reporting Only P Values for Results
- Not Confirming That the Assumptions of Statistical Tests Were Met

Common Statistical Errors

(Tom Lang)

- Interpreting Nonstatistically Significant Results As “Negative” When They Are, in Fact, Inconclusive
- Not Reporting Whether or How Adjustments Were Made for Multiple Hypothesis Tests
- Confusing Statistical Significance with Biologic Importance
- Not Confirming That the Data Met the Assumptions of Analysis
- Not Identifying the Procedure Used to Adjust for Multiple Comparisons in ANOVA
- Not Testing the Explanatory Variables for Interaction or Colinearity

Common Statistical Errors

(Tom Lang)

- Not Indicating the Goodness-of-Fit of the Model to the Data
- Not Reporting Whether and How the Model Was Validated
- Not Reporting Confidence Intervals with Estimates
- Reporting Only Relative Differences and Not Absolute Ones
- Not Differentiating Between Unit of Observation and the Number of Patients Improved
- Confusing Post-hoc Analyses with Planned Analyses

Common Statistical Errors

(Tom Lang)

- Using linear regression analysis without establishing that the relationship is, in fact, linear
- Not defining “normal” or “abnormal” when reporting diagnostic test results
- Not explaining how uncertain (equivocal) diagnostic test results were treated when calculating the test’s characteristics (such as sensitivity and specificity)
- Using a chart or graph in which the visual message does not support the message of the data on which it is based

Common Statistical Errors

(Tom Lang)

- Confusing the “units of observation” when reporting and interpreting results
- Interpreting studies with nonsignificant results and low statistical power as “negative,” when they are, in fact, inconclusive
- Not distinguishing between “pragmatic” (effectiveness) and “explanatory” (efficacy) studies when designing and interpreting biomedical research
- Not reporting results in clinically useful units
- Confusing statistical significance with clinical importance

출판이 거절되는 흔한 이유

- 주제가 임상적으로 중요하지 않음
- 고유한 연구가 아님
- 실제로 저자의 가설을 검증한 연구가 아님
- 연구설계의 문제
- 연구계획대로 하지 못한 연구
- 표본의 크기가 작은 연구
- 대조군이 없거나 선정에 문제가 있는 연구
- 부적절하거나 잘못된 통계분석
- 자료에 근거하지 않은 결론을 유도
- 이해관계의 상충의 의심
- 이해하기 힘들 정도로 글이 엉망인 경우
- 심사위원을 잘못 만난 경우 (?)

통계 오용의 예

- P-값이 0.05보다 작게 나올 때까지 갖가지 방법을 사용
- 교란변수를 통제하지 않고 분석하고 결론을 내림
- 사용한 통계분석의 가정을 확인하지 않음 (가장 많은 오용)
- 중도탈락자와 무응답자를 무시함.
- 인과관계와 연관관계의 혼용
- 분석결과가 좋지 않으면 몇 몇 값을 자료에서 제외 (특히, 특이값)
- 보이고자 하는 결과만 보임 (6개월을 예상한 연구에서 4개월째 유의한 결과를 보이면 연구를 중단하고 논문작성, 6개월의 결과가 좋지 않으면 임의로 연구를 6개월 더 연장)
- 특정 집단을 계속 나누어 유의한 결과를 보일 때까지 분석

결론

- 의학연구를 수행하고 타당한 결과를 도출하여 논문으로 출판하기 위해서는 먼저 명확하고 의미 있는 연구주제를 확립하고 이 연구주제에 맞는 올바른 연구설계를 하여야 함.
- 연구설계대로 자료를 수집하고 적절한 통계분석법을 이용하여 결과를 내고 논문에 명확하게 통계방법과 결과를 기술함.
- 이 때 기준은 같은 자료가 있다면 독자들이 같은 통계방법을 시행할 수 있을 정도로 명확하게 기술함.
- Curran-Everett와 Benos(2004)가 제시한 의학저널에 통계를 보고하는 10가지 가이드라인.

Guideline 예

1. If in doubt, consult a statistician when you plan your study.
2. Define and justify a critical significance level α appropriate to the goals of your study.
3. Identify your statistical methods, and cite them using textbooks or review papers.
4. Control for multiple comparisons.
5. Report variability using a standard deviation.
6. Report uncertainty about scientific importance using a confidence interval.
7. Report a precise P value.
8. Report a quantity so the number of digits is commensurate with scientific relevance.
9. In the Abstract, report a confidence interval and a precise P-value for each main result.
10. Interpret each main result by assessing the numerical bounds of the confidence interval and by considering the precise P-value.