Data Analysis 단계에서의 고려사항

- 분석군 및 결측자료의 처리 -

2019. 4. 18.

Prof. Sang Gyu Kwak

대구가톨릭대학교 의과대학 의학통계학교실

Harvard University Multi-Regional Clinical Trial, Qualified Member

CDISC, Platinum Member & Certified Member(SDTM, CDASH, ADaM)

DCUMC Medical Statistics & Informatics R&D Institute, Deputy Director

목차

배경

- 무작위 배정(***Randomization***)은 처치군간 비교성을 보장한다.

- 무작위 배정과 시험종료 사이에 시험대상자가
  - 무작위 배정된 처치를 따르지 않거나
  - 다른 처치로 바꾸거나
  - 연구에서 중도탈락(drop-out)하는 경우

- 이런 시험대상자를 제외하고 분석하는 경우 ***bias***가 발생할 수 있다.

분석 대상군

- ***ITT (Intent-to-treat)***
  - 분석대상 : 임상시험 계획서에 따라 등록된 연구대상자
  - 처리의향 분석
  - Randomization 된 모든 연구대상자를 분석하는 방법
  - 중도탈락된 연구대상자는 결측치 처리 후 분석 시행
- ***FAS (Full analysis set) or mITT (modified ITT)***
  - 분석대상 : 임상시험 계획서에 따라 등록된 연구대상자 중 Randomization 이후 치료를 한번도 받지 않고 탈락된 대상자를 제외
  - ITT의 엄격함 해결
- ***PP (Per protocol)***
  - 분석대상 : 임상시험 계획서에 순응하여 임상시험을 완료한 연구대상자
  - 중도 탈락한 경우 제외하고 분석하는 방법
  - Protocol에 따라 임상시험을 성공적으로 마친 환자만 분석

Guidance for ITT, FAS, PP

ICH E9 (1988): 'Note for Guidance on Statistical Principles for Clinical Trials)

> ***'The ITT principle implies that the primary analysis should include all randomized subjects. Compliance with this principle would necessitate complete follow-up of all randomized subjects for study outcomes. In practice this ideal may be difficult to achieve, for reasons to be described. In this document the term 'full analysis set' is used to describe the analysis set which is as complete as possible and as close as possible to the ITT ideal of including all randomized subjects'***

Guidance for ITT, FAS, PP

CPMP (1997): 'Note for Guidance on Evaluation of New Anti-Bacterial Medicinal Products'

'The modified ITT, where unqualified patients are excluded, and patients with clinically and/or microbiologically documented infections (as stated in the protocol), who have received at least one dose of the investigated drug thus addressed, is particularly valid for regulatory purposes.'

Guidance for ITT, FAS, PP

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'The 'per-protocol' set of subjects, sometimes described as the 'valid cases', the 'efficacy' sample or the 'evaluable subjects' sample, defines a subset of the subjects in the full analysis set who are more compliant with the protocol …'

ITT의 특징

- 연구참여자를 무작위 배정된 대로 비교 (***as randomized***)
- 처치(therapy)의 완전한 잠재적 이익을 포착하지 못하므로 ITT가 최적의 방법이 아닌 것으로 생각될 수도 있음
- ITT 분석은 무작위배정의 강점을 보유 (***bias minimization***)
- ITT는 PP보다 실제(real-life)에서 일어날 수 있는 상황을 더 고려함

ITT의 특징

- ITT분석은 처치에 있어 실제적인 정보를 제공
- ITT는 처치에 순응한 개체, 비순응 개체, 처치를 바꾼 개체 모두 포함
- ITT는 PP처럼 처치의 최대의 잠재적 효과를 보려는데 목적이 있지 않음
- ITT는 PP의 결과보다 그 효과가 통계적으로 유의하지 않거나 효과의 크기가 작은 경우가 많음
- 결측자료의 처리가 쉽지 않음

FAS의 특징

- 규제기간에서는 어느 정도의 타협을 허용함
- FAS는 ITT의 원칙을 최대한 유지하여야 함
- Bias없이 개체를 제외할 수 있는 가능한 경우의 예
  - **포함/제외 기준에 맞지 않는 시험대상자**
  - **시험약/대조약을 한번도 복용하지 않은 시험대상자**
  - **베이스라인 측정값 이후 측정값이 없는 시험대상자**
- Full analysis set (FAS)는 ITT와 다름
- 하지만 연구자들은 FAS를 ITT라고 사용하는 경우가 흔함

FAS의 특징

- FAS 대신 ***modified ITT*** (***mITT***) 라는 용어를 사용하기도 함.
- 우월성 검정에서는 거의 모든 임상시험에서 주분석군으로 FAS를 사용함
- 규제기관에서 FAS군을 선호하는 이유는 처치의 차이 비교에서 보수적인 결과를 주기 때문임
- 만약, FAS군을 이용한 분석에서 처치의 차이가 통계적으로 유의한 경우 규제기관은 처치의 유효성이 있다고 판단함
- 비열등성이나 동등성 임상시험에서는 FAS군의 결과가 비보수적(anti-conservative)일 수 있음

PP의 특징

- FAS 분석군과 마찬가지로 PP 분석군도 프로토콜에서 분명히 명시하여야 함
- PP 분석은 bias된 결과를 가져올 수 있으며 처치효과가 과대평가될 수 있음
- 이런 이유로 우월성 시험에서는 PP군은 일반적으로 부분석(secondary analysis)에 사용됨

ITT와 PP의 비교

1. PP 유의함 > ITT 유의함

➜ 비순응의 비율이 매우 높음을 시사함

2. PP 유의하지 않음 & ITT 유의함

➜ 처치간의 차이라기 보다는 다른 변수에 의한 차이일 가능성 있음

3. PP 유의함 & ITT 유의하지 않음

➜ 상당한 비율의 연구참여자들이 한 방향으로 처치를 바꾸었기 때문일 가능성이 있음(예: 위약에서 신약으로)

임상시험에서의 결측자료

- 결측자료는 RCT에서 매우 흔하게 발생

  중도탈락 (Drop-out)

  중간방문 놓침 (Miss visits)

  자료의 손실 (Lost data)

- 결측자료가 있는 경우의 문제점

  편향의 발생 (Introduction of bias)

  효율의 감소 (Loss of efficiency) → 검정력 약화

  자료의 처리와 분석이 어려워짐

KJA
Korean Journal of Anesthesiology

Statistical Round
pISSN 2005-6419 · eISSN 2005-7563

Statistical data preparation: management of missing values and outliers

Sang Kyu Kwak[1] and Jong Hae Kim[2]

Departments of [1]Medical Statistics, [2]Anesthesiology and Pain Medicine, School of Medicine, Catholic University of Daegu, Daegu, Korea

Missing values and outliers are frequently encountered while collecting data. The presence of missing values reduces the data available to be analyzed, compromising the statistical power of the study, and eventually the reliability of its results. In addition, it causes a significant bias in the results and degrades the efficiency of the data. Outliers significantly affect the process of estimating statistics (*e.g.*, the average and standard deviation of a sample), resulting in overestimated or underestimated values. Therefore, the results of data analysis are considerably dependent on the ways in which the missing values and outliers are processed. In this regard, this review discusses the types of missing values, ways of identifying outliers, and dealing with the two.

Key Words: Bias, Data collection, Data interpretation, Statistics.

임상시험에서의 결측자료

- 결측자료에 대한 고려할 사항

1. 결측자료의 발생을 최소화할 수 있는 연구설계를 고려함

2. 결측자료로 인하여 결과에 미칠 잠재적인 영향을 보정하고 평가함

임상시험에서의 결측자료

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'Missing values represent a potential source of bias in a clinical trial. Hence, every effort should be undertaken to fulfill all the requirements of the protocol concerning the collection and management of data. In reality, however, there will almost always be some missing data. A trial may be regarded as valid, nonetheless, provided the methods of dealing with missing values are sensible, and particularly if those methods are pre-defined in the protocol. Definition of methods may be refined by updating this aspect in the statistical analysis plan during the blind review. Unfortunately, no universally applicable methods of handling missing values can be recommended. An investigation should be made concerning the sensitivity of the results of analysis to the method of handling missing values, especially if the number of missing values is substantial.'

임상시험에서의 결측자료

- 임상시험에서는 피험자가 연구종료 이전에 중도탈락(***drop out*** or ***withdraw***)하는 경우가 흔함
- RCT에서는 피험자를 각 처치군에 무작위로 배정하고 여러 시점($t = 1,\ldots,T$)에 걸쳐 측정하게 됨
- 어떤 피험자는 임상시험 종료시점(T visit) 이전의 시점(t visit) 이전에 중도탈락함 ($t < T$). 또 어떤 피험자는 중도탈락은 아니지만 정해진 방문기간을 놓침. (***intermittent missing***).

임상시험에서의 결측자료

- 이런 중도탈락은 아마도 임상시험과 관련이 있거나 (adverse event, death, lack of improvement, etc.) 관련이 없을 수도 있음 (moving away, study unrelated disease)
- 문제는 중도탈락한 피험자와 임상시험을 종료한 피험자의 특성이 서로 다를 가능성이 높음 (**무작위 배정의 손상**)
- 따라서 결측자료는 처치비교에 편향을 발생 시킬수 있고 (***biased treatment comparisons***) 전체 연구의 통계적 검정력에 영향을 미침(***overall statistical power of the study***)

임상시험에서의 결측자료

- 임상시험에서 결측자료를 처리하는 두가지 방법

  (1) design stage

  (2) statistical analysis stage

- Design stage

- 교육 등의 노력을 통하여 중도탈락을 최소화함
- 시험적 연구에서는 결측이 발생할 것 같은 피험자는 시험에서 제외하는 방법 (추천되지 않음)
- 베이스라인에서 가능한 많은 정보를 모아야 함. 이는 이후 통계적 분석을 통한 결측자료 보정에 큰 도움을 줌

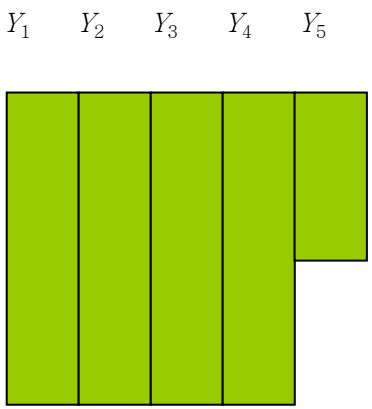
결측자료 패턴

The missing data ***pattern***: 데이터 행렬에서 어떤 값이 관측되었는 지 아니면 결측인 지 나타내는 모양

Example:

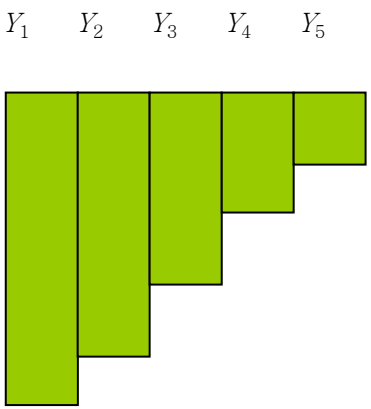
Univariate

(일변량)



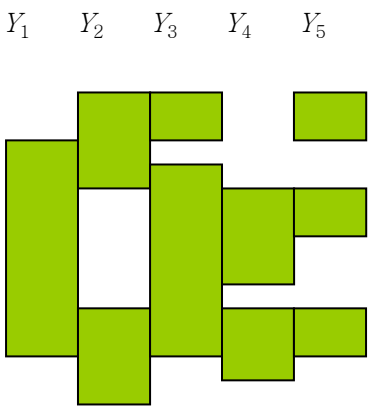
Monotone

(단조)



General

(일반)



결측값의 종류

- ***완전무작위 결측 발생 (MCAR: Missing completely at random)***
- 결측값이 발생한 경우가 다른 값에 영향을 받지 않고 완전히 랜덤하게 발생
- 완전무작위 결측값은 갑자기 연구대상자가 측정시점에 나타나지 않아 총 측정시점 중간에 발생할 수도 있고, 연구에서 중도탈락을 하여 특정 측정시점부터 연구종료시까지 발생할 수도 있음

결측값의 종류

- ***임의적 결측 발생 (MAR: Missing at random)***
- 특정 시점에서 연구대상자가 참여한 연구 성과에 만족하지 못할 시 발생할 수 있음
- 임의적 결측값은 완전무작위 결측값보다 임상연구에서 훨씬 보편적으로 발생하는 결측 유형이라고 할 수 있다.

결측값의 종류

- ***비임의적 결측 발생(NMAR: Not missing at random)***
- 특정 시점에서 연구대상자가 참여한 연구 성과에 만족하지 못하는 것과 동시에 방문 전 개인적으로 측정 한 경우 발생 할 수 있음

결측 자료 분석

- 결측 자료 분석은 결측치의 종류에 따라서 달라짐
- 결측치 종류 구분 어려움
- 결측 자료 분석에서는 **민감도 분석**(***sensitivity analysis***)이 중요함
- 민감도 분석이란 다양한 결측치 종류의 가정과 모형 가정하에 분석결과가 어떻게 다른 지를 알아보는 방법

Listwise deletion analysis

- 모든 변수들이 관측된 대상자 데이터만 이용하여 분석
- 단 하나의 변수에서 결측값이 있어도 그 대상자는 분석에서 제외
- 대부분의 통계프로그램에서 이 방법 사용

구분	장점	단점
내용	• 간편성 • 일변량 통계량들의 비교 가능 • 완전 무작위 결측 o → 모수 추정치 편의 x	• 많은 표본수의 감소 • 정보의 손실 • 검정력의 약화 • 완전 무작위 결측 x → 모수 추정치 편의 o

Pairwise deletion analysis

- 각 각의 분석 단계에서 사용 가능한 대상자 데이터를 이용.

구분	장점	단점
내용	• 표본수는 complete case analysis보다 많다. • 완전 무작위 결측 O → 모수 추정치 편의 X	• 표본의 기준 데이터가 분석마다 변함 • 실용적이지 못함 • 완전 무작위 결측 X → 모수 추정치 편의 O

단일대체 (Single Imputation)

- 각 결측치를 하나의 예측값으로 대체하여 완전한 데이터 행렬로 만들고 이 대체된 값들을 실제로 관측한 값으로 여기고 분석을 한다.
- 이 때 통계적 모형을 이용하여 대체값(imputed value)을 구한다.

단일대체 (Single Imputation)

Unconditional Mean Imputation

각 결측값을 관측된 값들의 평균으로 대체한다.

y_1	y_2
0.419	0.556
1.235	2.282
0.756	1.102
0.422	0.480
1.909	1.867
-0.929	-0.572
-0.378	0.427
-1.321	-1.575
-0.074	?
0.905	?

$$\frac{\sum_{i=1}^{8} y_{i1}}{8} = 0.571$$

0.571

0.571

단일대체 (Single Imputation)

Conditional Mean Imputation (Regression Mean Imputation)

결측값을 가진 변수를 다른 변수들과 회귀분석한 후 결측값의 예측값으로 대체한다.

y_1	y_2
0.419	0.556
1.235	2.282
0.756	1.102
0.422	0.480
1.909	1.867
-0.929	-0.572
-0.378	0.427
-1.321	-1.575
-0.074	?
0.905	?

$$y_{i2} = 0.287 + 1.073 \times y_{i1}$$

$$0.208 = 0.287 + 1.073 \times (-0.074)$$

$$1.259 = 0.287 + 1.073 \times (0.905)$$

단일대체 (Single Imputation)

Conditional Mean Imputation(Stochastic Regression Imputation)

결측값을 회귀식의 예측값과 임의로 추출한 오차를 합하여 대체한다.

(예측값의 불확실성 고려)

y_1	y_2
0.419	0.556
1.235	2.282
0.756	1.102
0.422	0.480
1.909	1.867
-0.929	-0.572
-0.378	0.427
-1.321	-1.575
-0.074	?
0.905	?

$$y_{i2} = 0.287 + 1.073 \times y_{i1} + \epsilon$$

$$\epsilon \sim N(0, \sigma^2)$$

$$0.355 = 0.287 + 1.073 \times (-0.074) + (0.147)$$

$$0.594 = 0.287 + 1.073 \times (0.905) + (-0.665)$$

단일대체 (Single Imputation)

- Last Observation Carried Forward (LOCF)
- 경시적 자료(longitudinal data)에서 **각 개체 내 결측값은 마지막으로 관측된 값으로 대체한다.** 대체된 값을 관측치로 생각하고 분석을 행한다.

단일대체 (Single Imputation)

- Last Observation Carried Forward (LOCF)

	Time					
Subject	1	2	3	4	5	6
1	2.3	3.2	4.5	? ← 4.5	? ← 4.5	? ← 4.5
2	1.3	1.5	2.4	1.5	? ← 1.5	? ← 1.5
3	2.1	2.0	3.3	3.5	2.9	3.5

- 쉽다.
- random variability를 고려하지 않음
- **현재 임상시험 자료분석에서 가장 많이 사용되는 방법**

Discussion

- 결측치 분석방법은 피험자의 중도탈락이 예상되는 경우 프로토콜에 미리 기술
- 높은 비율의 중도탈락이 시험에서 발생한 경우에는 통계적 결과의 해석이 불분명한 경우가 있음
- 결과를 해석하는 경우 중도탈락의 비율이 항상 고려
- 중도탈락 자료에 대한 가장 좋은 하나의 방법은 없음

Conclusion

- 결측자료를 다루는 경우, 특히 결측률이 높은 경우 조심스럽게 분석하여야 함
- 항상 좀 더 많은 정보(자료)를 모을 수 있도록 노력
- 왜 결측이 일어났는 지 탐색하고 그에 따른 적절한 분석법을 선택
- 결측자료 분석으로 한 가지 방법만 고집하지 말 것
- 민감도 분석(sensitivity analysis)을 실시