중재연구에서 인과추론 고려사항들

Causal Inference Considerations in Intervention Research

Aug. 2023

Dae Ryong Kang, Ph.D.

Department of Precision Medicine & Biostatistics

Yonsei University, Wonju College of Medicine

- ✓ **Association (연관성 聯關性)**
  동질성/독립성
- ✓ **Correlation (상관성 相關性)**
- ✓ **Probability (개연성 蓋然性)**
- ✓ **Causality (인과성 因果性)**

Study Designs

Observational study

Unit of study

Descriptive study
Analytical study

Hypothesis ?

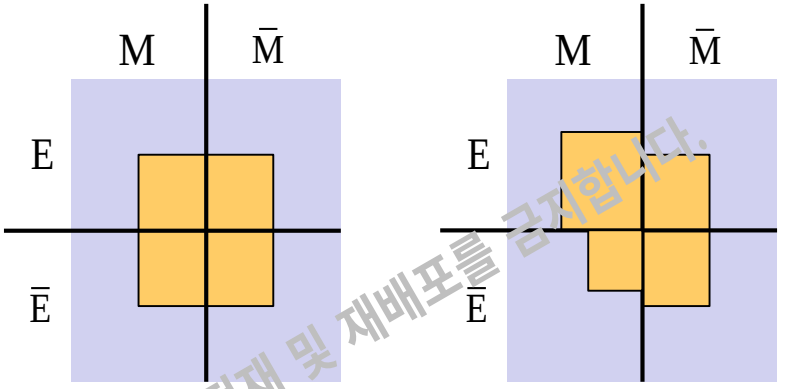
Ecological study (Correlation study) — Population
Cross-sectional study (Prevalence study) — Individuals
Case-control study (Case-reference study) — Individuals
Cohort study (Follow-up study, Prospective study) — Individuals
Hybrid designs : Nested case-control design, Case-cohort design
Case-crossover design, Case-time-control design

Randomization ?
Intervention ?

Experimental study, Quasi-experimental study

Experiment
Randomized controlled trials (Clinical study) — Patients
Field trials — Healthy people
Community trials (Community intervention study) — Community

at the research design stage bias



① Selection bias

- ✓ sampling frame bias : admission rate bias (*Berksonian* bias)
- ✓ non random sampling bias : detection bias
- ✓ non-converge bias : loss to follow-up bias, withdrawal bias

② Non comparability bias

- ✓ lead time bias, length bias, historical control bias
- ✓ immortal time bias : Landmark method, td Cox model

③ Sample size bias

in the data collection process information bias

① Instrument bias

② Data source bias

③ Observer bias

- ✓ diagnostic suspicion bias
- ✓ exposure suspicion bias
- ✓ therapeutic bias (➔ Blinding)

④ Subject bias

- ✓ proxy respondent bias
- ✓ recall bias
- ✓ attention bias ("Hawthorne effect")

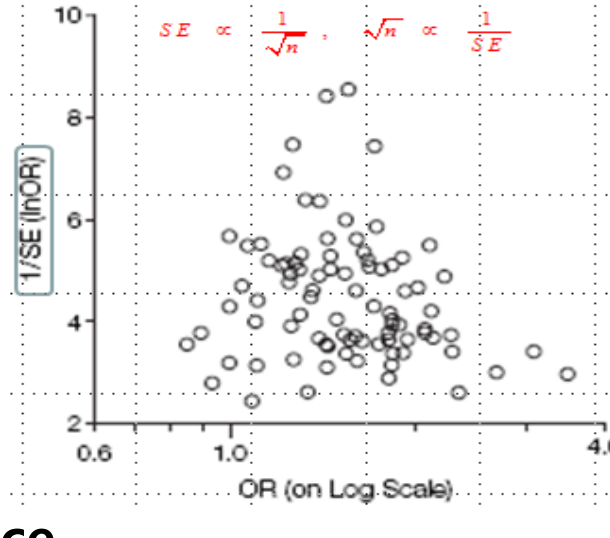
in the process of analysis & interpretation of results bias

① **Confounding bias**

② **Analysis strategy bias**

: **missing data handling, outlier handling, unit of analysis**

③ **Post-hoc analysis bias** (← data dredging bias)



④ **Assumption bias**

⑤ **Generalization bias** (← lack of external validity)

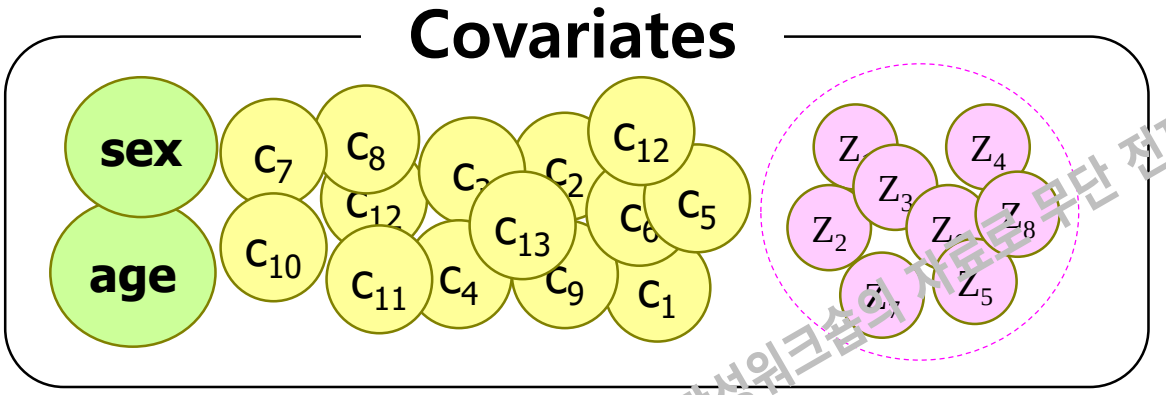
⑥ **Significance bias**

: **statistical significance vs. biological significance**

⑦ **Publication bias** (by Funnel plot, Egger's regression asymmetry test)

Statistical Methods for Causal Inference

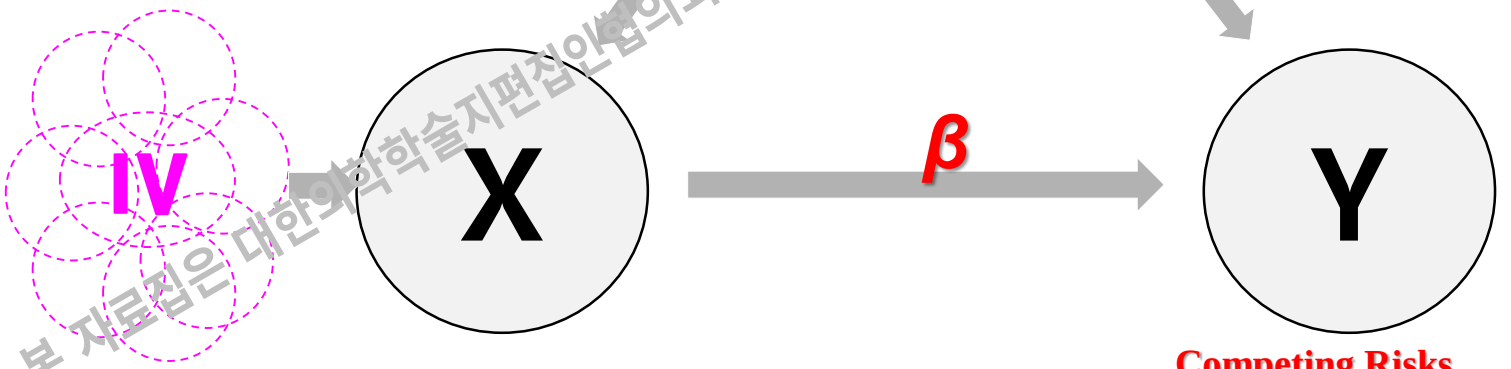
- confounders variables
- unmeasured/unknown confounders
- stratification variables
- intermediate variables
- effect modifier / interaction effect



+ 'Time'을 어떻게 보정할 것인가?

Propensity Score Matching
Coarsened Exact Matching

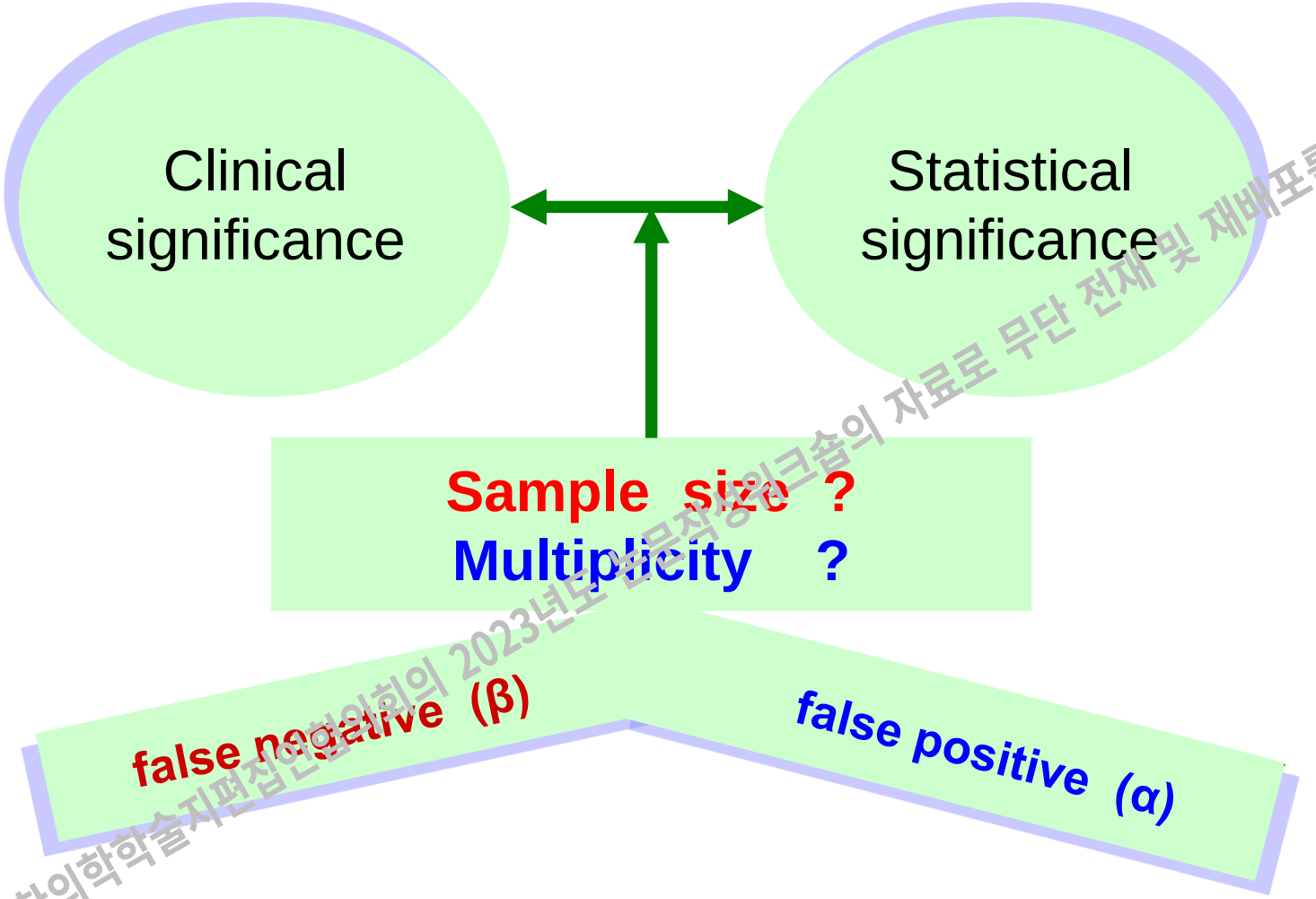
Multiple Regression Adjustment



Mendelian Randomization Analysis

Competing Risks

- Multiple Regression Analysis
- Logistic Regression Analysis
- Poisson Regression Analysis
- Cox's PHM
- Linear Mixed Model (**LMM**)
- Generalized Estimating Equation (**GEE**)



- ✓ Type III error : 연구가설 정립 ?
- ✓ Bias 줄이기 위한 노력 + 시간

임상시험에서의 합리적 연구대상자수

➔ 어떤 원칙으로 연구대상자수를 결정할까 ?

Sawada et al., *European Heart Journal* 2009;30(20):2461-69.

Treatment	Event		
	Y	N	
Valsartan	83(5.5%)	1434	**1517**
Non-ARB	155(10.2%)	1359	**1514**
	238	2793	**3031**

p-value **< .0001**

10% random sampling

Treatment	Event		
	Y	N	
Valsartan	9(5.9%)	144	**153**
Non-ARB	16(10.5%)	136	**152**
	25	280	**305**

p-value **= 0.139**

너무 많은 연구대상자수로 RCT를 진행하면,

- ✓ '시간'과 '비용' 낭비이다
- ✓ 안전성 + 유효성이 입증되지 않은 의약품/의료기기라면 윤리적으로 문제가 있다.

너무 적은 연구대상자수로 RCT를 진행하면,

- ✓ 실제 존재하는 의미있는 효과를 찾아내지 못한다.
- ✓ (1-β)이 떨어지므로 유효효과가 있는데도 이를 찾아내지 못한다.

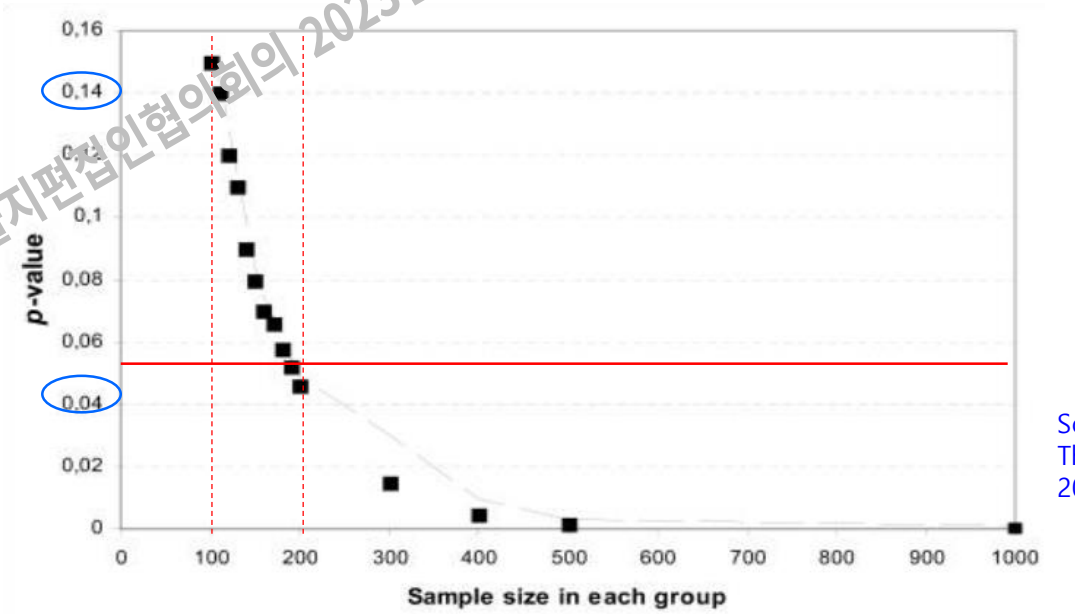
⇒ 연구대상자수에 영향을 주는 (α, 1-β, 치료효과 크기와 Δ 변이, . . .)를 고려하여 합리적인 피험자수를 산출/결정 하여야 한다.

연구대상자수 결정에 고려해야 할 사항들

- ✓ **Level of significance $\alpha \leq 0.05$, Power $(1-\beta) \geq 0.80$**
- ✓ **Effect Size Δ / (previous research + researcher's intention)** ➔ **'pilot study'**
  - **minimum treatment difference**
  - **true mean/proportion difference**
  - **superiority/non-inferiority/equivalence 'limit'**
- ✓ **Type of comparison : superiority(inequality), non-inferiority, equivalence**
- ✓ **Design configuration : parallel, crossover**
- ✓ **Number of interim analyses**
- ✓ **Other :**
  - **number / allocation rate of Experimental-Control**
  - **follow-up period (~ inverse correlation)**
  - **drop-out rate d ($n' = n/(1-d) = 100/(1-0.2) = 125$)**
  - **compliance rate c ($n' = n/c^2 = 100/0.8^2 = 156$)**

p-value in Biomedical Research

- ***Is the p-value of 0.05 an absolute criterion for determining 'significance' ?***
- ***Statistical significance interpretation based on p-value has limitation.***
  - ➔ p-value is an important criterion in decision, but has some limitations.
- ***The p-value doesn't explain the degree of importance of the observed effect.***
  - ➔ p-value is small does not necessarily mean that the association is strong.
- ***The p-value is closely related to the sample size.***
  - ➔ Interpreting the statistical results to rely only on p-value would be difficult to prove the improvement of clinical usefulness because the target sample size is not achieved, especially when dealing with "*rare diseases*".



Source: DB Panagiotakos,
The Open Cardiovascular Medicine Journal,
2008, 2, 97-99.

임상연구의 자료분석에서 '다중검정'을 하게 되는 경우

① **그룹 수가 3개 이상인 경우**

② **Primary endpoint가 2개 이상인 경우**

③ **중간분석 (Interim Analysis)**을 허락하도록 디자인된 임상시험은, 자료가 어느 정도 모아진 중간단계에서, 약효에 대한 통계적 검정을 실시하고, 만일 그것이 유의하면 일찍 임상시험을 종료하는 것을 허락한다. 이런 디자인에서 만일 2번의 중간분석을 허락한다면, 최종분석과 함께 총 3회의 가설검정을 하게 되는 것이다.

④ **Subgroup Analysis :** 남자와 여자를 따로 분석

⑤ **여러 Analysis Sets에 대한 검정 :** FAS 분석과 PP 분석

⑥ **Sensitivity Analysis (missing value) :** 결측치 처리 방법의 민감도를 알아보기 위하여, 여러 방법으로 결측치를 처리하여 유효성 변수를 검정하는 경우

⑦ **여러 개의 안전성 변수들에 대한 분석 :**
유효성 변수 분석 후, 여러 안전성 변수들에 대한 검정

다중검정 multiple comparison

여러 개의 (독립적인) 각 가설검정을 5% 유의수준으로 검정하면 family-wise type I error (모든 H_0가 참임에도 불구하고 적어도 하나의 H_0 을 기각하게 되는 오류) 가 증가하게 된다.

k 번의 다중검정을 실시하는 경우 family-wise type I error

$$FWE = 1 - (1-0.05)^k$$

➔ 그러나 사실상 *k* 개의 검정이 엄밀하게 '독립적'이라고 가정하기가 어렵다.

K	FWE
1	0.05
2	0.098
3	0.143
4	0.186
5	0.250
10	0.40

다중비교(multiple comparison) 방법들

검정법	비교시기	비교집단	표본 수	비고
Tukey HSD 正直有意差檢定 (Honestly Significant Difference)	사후	Pairwise	**Equal**	필요이상으로 보수적-성직적 이다
Student-Newman-Keuls의 검정	사후	Pairwise	Equal	
Duncan의 다중범위검정(MRT)	사후	Pairwise	Unequal	
Scheffé의 다중비교절차	사후	가능한 모든 조합	**Unequal**	여러 처리평균의 결합(combination)으로 이루어진 대비(contrast) 적용
Dunnett	사전 (Planed)	대조군과 비교	Unequal	선별절차 **Gupta & Sobel** (1958)
Bonferroni t-검정 **Sidak** t-검정	사전 (Planed)	계획된 조합	Unequal	**Holm, Hommel Hochberg, FDR**

➔ Scheffé 방법이 가장 보수적이며 가장 융통성이 있는 방법이므로 의학에서는 주로 Scheffé 방법을 사용함.

※ Fisher의 보호/비보호 최소유의차 검정 (protected least significant difference, **LSD**)

※ 자료가 정규분포를 따르고 (오차항이 정규분포를 따르고), 특정한 모형 가정이 가능 할 때 적용하는 방법과 자료분포 상관없이 어떠한 raw p-value에도 적용할 수 있는 방법

Bonferroni 방법

- family-wise type I error를 5% 이하로 통제하기 위하여, 각 개별 가설검정을 0.05/**k** 유의수준에서 검정하는 방법이다 (k = 다중검정 횟수).

- 가장 보수적인 방법이다.

Bonferroni 방법을 사용하여 분석한 결과가 유의하게 나오면 다른 어떤 방법 (Hochberg 방법, Holm 방법)을 사용해도 유의하게 나온다. ➔ 심사기관 선호함.

- 단점, 어떠한 경우에도 fwe를 5% 이하로 통제하다 보니, 실제 범하는 fwe는 5%보다 매우 낮다는 것이다. ➔ 제2종 오류 β가 크다는 말이고 이는 검정력(1-β)이 낮아진다는 의미이다. ➔ 연구자에는 불리함.

- Bonferroni 방법보다 더 높은 검정력을 주는 Holm's step-down 방법, Hochberg step-up 방법, Tukey 방법, Hommel 방법 등을 개발.

```
/* http://ftp.sas.com/samples/A56648 */

data pValue;
   input  Variable $  raw_P  @@;

cards;
Delusions                  0.811
Hallucinations             0.743
Disinhibition              0.446
Irritability               0.272
Agitation                  0.146
Anxiety                    0.072
Appetite_changes           0.047
Nighttime_behavior         0.022
Aberrant_motor_behavior    0.021
Depression                 0.013
Apathy                     0.003
;
run;

proc multtest pdata=pValue  bon  stepbon  sid  stepsid  holm  hoc  fdr  out=outP; run;

proc sort data=outP;  by raw_P; run;
proc print data=outP; run;
```

The Multtest Procedure

P-Value Adjustment Information

P-Value Adjustment	Bonferroni
P-Value Adjustment	Stepdown Bonferroni
P-Value Adjustment	Sidak
P-Value Adjustment	Stepdown Sidak
P-Value Adjustment	Hochberg
P-Value Adjustment	False Discovery Rate

Benjamini and Hochberg's Method

p-Values

Test	Raw	Bonferroni	Stepdown Bonferroni	Sidak	Stepdown Sidak	Hochberg	False Discovery Rate
1	0.8110	1.0000	1.0000	1.0000	0.9340	0.8110	0.8110
2	0.7430	1.0000	1.0000	1.0000	0.9340	0.8110	0.8110
3	0.4460	1.0000	1.0000	0.9985	0.8300	0.8110	0.5451
4	0.2720	1.0000	1.0000	0.9696	0.7191	0.8110	0.3740
5	0.1460	1.0000	0.7300	0.8238	0.5458	0.7300	0.2294
6	0.0660	0.7920	0.4320	0.5604	0.3613	0.4320	0.1320
7	0.0470	0.5170	0.3290	0.4111	0.2861	0.3290	0.1034
8	0.0220	0.2420	0.1890	0.2171	0.1739	0.1760	0.0605
9	0.0210	0.2310	0.1890	0.2082	0.1739	0.1760	0.0605
10	0.0130	0.1430	0.1300	0.1341	0.1227	0.1300	0.0605
11	0.0030	0.0330	0.0330	0.0325	0.0325	0.0330	0.0330

Subgroup Check List

Guyatt et al., User's guide to the medical literature. The EBM working group. 2008 JAMA evidence

family-wise type I error 증가 문제를 보정 ➔ Bonferroni 방법 적용하여

① Subgroup들이 a priori하게 계획되었는가?

② Subgroup들이 existing trial이나 biological data를 근거로 정의되었는가?

③ Subgroup들이 Pre-randomization Characteristic들에 의해 정의되었는가?

④ Subgroup 효과의 예상 방향성(expected direction)이 a priori하게 명시되었는가?

⑤ Subgroup analysis에서 patient misallocation의 효과는 얼마나 되는가?

⑥ Subgroup analysis에서 intention-to-treat population이 사용 되었는가?

⑦ 임상시험은 Key subgroup questions에 대해 적절한 크기의 검정력이 유지되도록 Design 되었는가?

임상연구에서의 결측자료

- 임상시험에서 한가지 문제점은 피험자가 시험이 종료되기 전에 중도 탈락(censoring)한다는 점이다. (intermittent missing value, drop out)

- 이런 중도탈락의 이유는 부작용(adverse event), 사망, 병세의 비호전 등 시험과 관련된 경우도 있을 수 있고. 이사, 시험과 관련 없는 병 등 시험과 관련이 없을 수도 있다.

- 연구에 끝까지 참여한 집단은 아마도 original sample의 random sub-sample이 아닐 것이다. 즉, 무작위 배정의 손상이 발생할 수 있다.

- 그러므로 결측은 처치의 효과에 '편향'을 발생시킬 수 있으며 시험의 전체적인 통계적 검정력에 영향을 미칠 수 있다.

Pattern of missing data :

- Missing Completely at Random (MCAR) 결측여부가 관측치+결측치 모두와 완전 독립
- Missing at Random (MAR) 결측여부가 관측치에는 의존하고, 결측치에는 의존하지 않는 경우
- Missing Not at Random (MNAR) 결측여부가 관측치에도 의존하고, 결측치에도 의존하는 경우

- **Deletion** techniques are widely criticized because they assume that the data are **MCAR**, pose a risk for bias, and lead to reduction of sample size and power. (← "complete case analysis")

Missing Data Mechanism

- 결측자료 분석은 이런 결측 메케니즘에 따라서 달라진다.
- 만일, 우리가 가진 결측자료가 어떤 메케니즘인지 안다면 좋겠으나 우리가 가진 자료로는 MCAR, MAR 또는 NMAR인지 구분할 수 없다.
- 일반적으로 임상시험에서 중도탈락은 결과변수와 관련이 있는 경우가 많으므로 메커니즘은 **MCAR이 아닌 경우가 많다**.
- 임상시험처럼 잘 조절된 연구에서는 주효과변수를 관찰하기 위한 노력을 기울이고 또한 주효과변수와 관련된 요인들을 수집하므로 **MNAR**에 의한 편향을 최소화할 수 있다.
- 이러한 이유로 임상시험에서의 결측 메커니즘은 **MAR로 가정하는 경우**가 많다.

Single Imputation

Single Imputation :

- Sample and Group mean substitution
- Case mean substitution
- Hot / Cold - deck imputation
- Regression imputation
- Maximum likelihood (ML)
- Expectation maximization (EM)

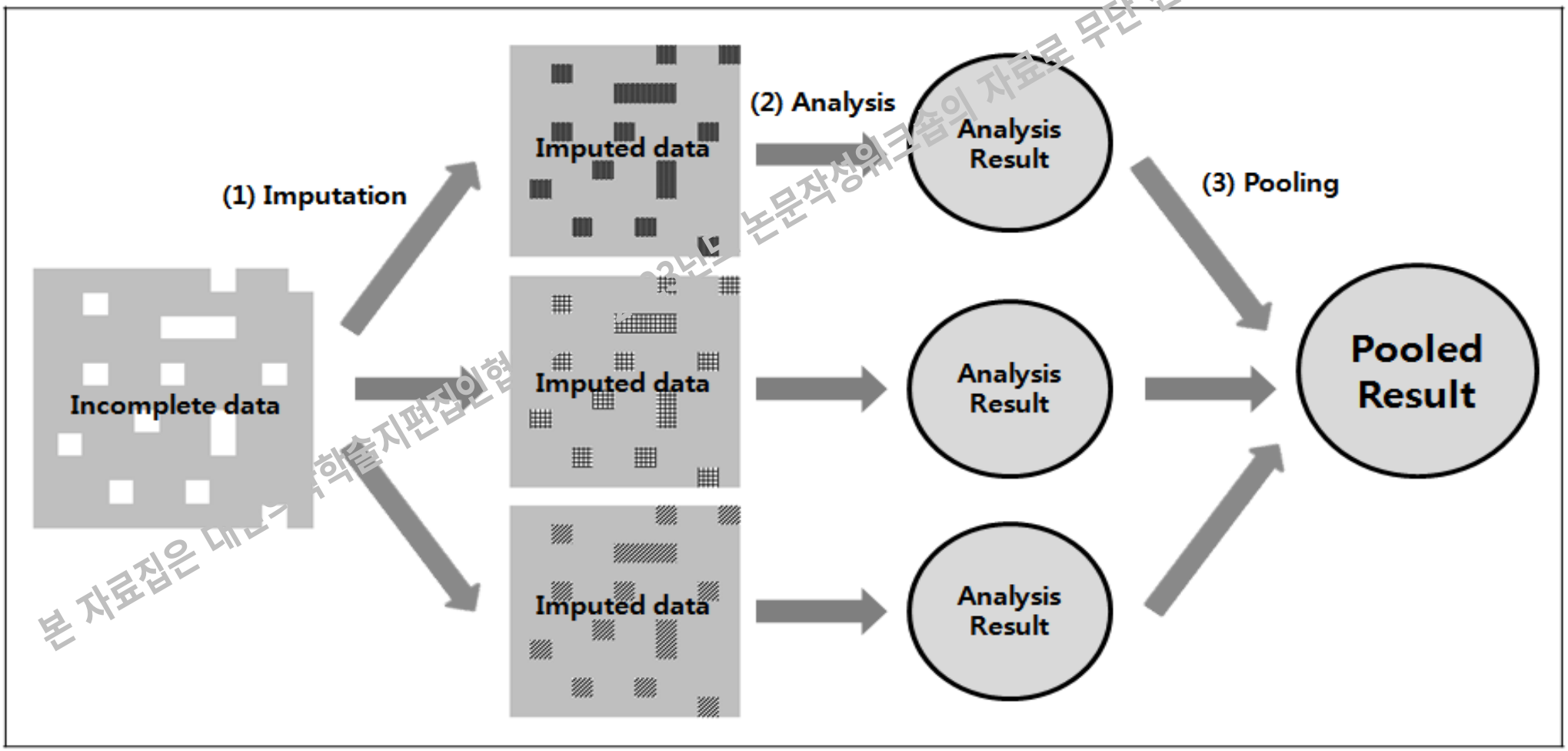
Multiple Imputation

- 결측값이 모두 대체된 경우, 일반적인 **통계방법**과 **통계프로그램**을 이용할 수 있기 때문에 "imputation"을 많이 사용하고 있다.
- 하지만 대체값을 생성하기 위한 통계모형이나 **결측자료 메커니즘**에 의해 편향(bias)이 발생할 수 도 있다.
- 또한, single imputation은 추정치의 **표준편차**가 **과소 추정**된다. 왜냐하면, 대체된 값을 실제로 관측된 값처럼 이용하였기 때문에 실제 표본수 보다 더 많은 표본수를 사용하였기 때문이다.
- 다시 말해, 결측값 때문에 발생하는 **불확실성 (uncertainty)**을 고려하지 않았기 때문이다.
- 이런 불확실성을 고려하는 방법으로는 Maximum Likelihood를 이용한 추정법, Resampling 방법, Multiple Imputation 등이 있다.

Multiple Imputation

MI는 결측으로 인한 **불확실성**을 고려하기 위한 single imputation의 확장이다.

Imputation ➔ Analysis ➔ Pooling (Rubin's)



Multiple Imputation

- SI 방법에서의 한가지 중요한 문제는 결측으로 인한 **불확실성**을 고려할 수 있는 **표준오차**의 추정이다.

- MI 방법을 이용하면 거의 모든 환경에서 결측으로 인한 불확실성을 고려할 수 있다.
- 이 방법은 표본조사론의 무응답 자료 분석을 위해 **Rubin**이 고안하였으나 지금은 일반적인 결측자료분석에 이용되고 있다.

- SAS에서 **PROC MI**를 이용하여 결측값을 대치하여 여러 개의 대체된 데이터셋을 만들 수 있다.

- 또한 **PROC MIANALYZE**를 이용해 ***m***개의 분석 결과를 **Rubin**이 제시한 결합방법을 이용하여 합치고 마지막 결과를 얻는다.

- Mixed-effects Model for Repeated Measures (**MMRM**)
- Imputation and Variance Estimation Software (**IVEware**)

Thank you for listening.

Q & A

iNHCBCR
국민건강빅데이터임상연구소
National Health BigData Clinical Research Institute