논문작성시 고려해야할 통계적 오류

2019. 2. 16.

연세대학교 원주의과대학

정밀의학과 · 의학통계학과

강 대 용

Assessment on the adequacy of the statistical methods

평가항목	항 목 설 명
1	적절한 통계방법을 사용하였음*
2	통계가 필요하나 사용하지 않았음
3	p-value를 제시하고 있으나 통계방법 언급 없음
4	통계방법을 나열하고 있으나 어느 분석에 적용한 것인지 알 수 없음
5	**틀린 방법 적용** 1) 자료의 성격과 전혀 맞지 않는 방법 사용 2) ANOVA를 *t*-test로 분석 3) 비모수를 모수로 분석 4) 빈도 비교에서 기대치가 **5** 이하인 칸이 있음에도 Fisher's exact test를 쓰지 않았거나, 범주를 줄여 그 칸을 없애려는 시도를 하지 않음 5) 상관관계를 회귀분석으로 분석 6) 통계용어를 잘못 사용 7) 기타 경우 †
6	**적용통계는 맞게 언급했으나 결과를 제대로 제시하지 못함** 1) 어느 군과 어느 군의 비교 시에 나온 p-value인지 불명확 함 2) p-value 또는 confidence interval을 제시하지 않고 유의하다고 주장 3) 양측 / 단측 언급 없음, 유의수준 언급 없음 4) 통계량이나 자유도를 제시하지 않음 5) 계산이 틀렸을 가능성이 농후한 경우 (유의하지 않을 것 같은데 유의 하다고 주장) 6) 결론을 잘못 유도 (통계결과를 확장하여 결론을 내림) 7) 기타 경우 ‡

대한의학학술지 편집인협의회
(KoreaMed, 2004. 10)

* 아래 2~6 항목에 해당사항 없을 때

† 교란변수의 존재로 logistic analysis가 필요한 경우에 χ^2-pearson 으로만 분석

‡ 4 X 2 table에 Fisher's exact test를 사용했다고 쓰고 p-value 제시 없이 결론지음

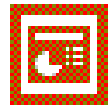
What makes it difficult for Medical Research?

- ✓ **the research target is 'Human'**
  - ethical problems
  - limit of study design
  - problems caused from the limit of study design
- ✓ **distortion of research results occurs when we have no enough time**
  - need for comparing several analytic results
  - lack of reflections in the discussion part
- ✓ **data noise**
  data incomplete
  - outliers
  - missing value
  - there is no data without 'noise'
- ✓ **when we use inappropriate statistical methods in data analysis**

Study design Study design Study design Study design **Data collection** **Data analysis** Interpretation & Discussion


EXCEL


MSACCESS


POWERPNT


Hwp

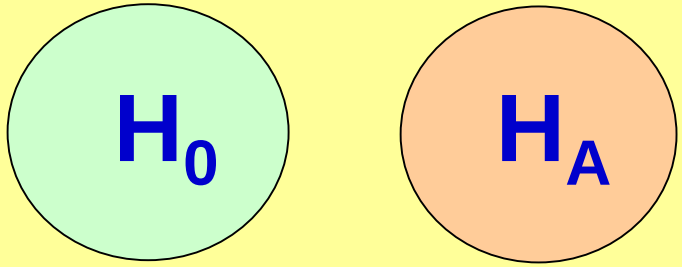
1. Descriptive Statistics :

we work on **'data cleaning'** while calculating **DS** of each observed variable
e.g., n, missing value, checking outlier, category regrouping, …

2. Statistical Testing (検定) :

Statistical inference
(1:1, 1:k, sub-group)

under significant level $\alpha = 5\%$



decision making with 'p-value'

3. Statistical Interpretation :

(highly) significant, limit significant, borderline significant, not significant
(difference, association, correlation, influence with adjustment, …)

4. Interpretation through 'Medicine' and 'Public Health'

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 ?

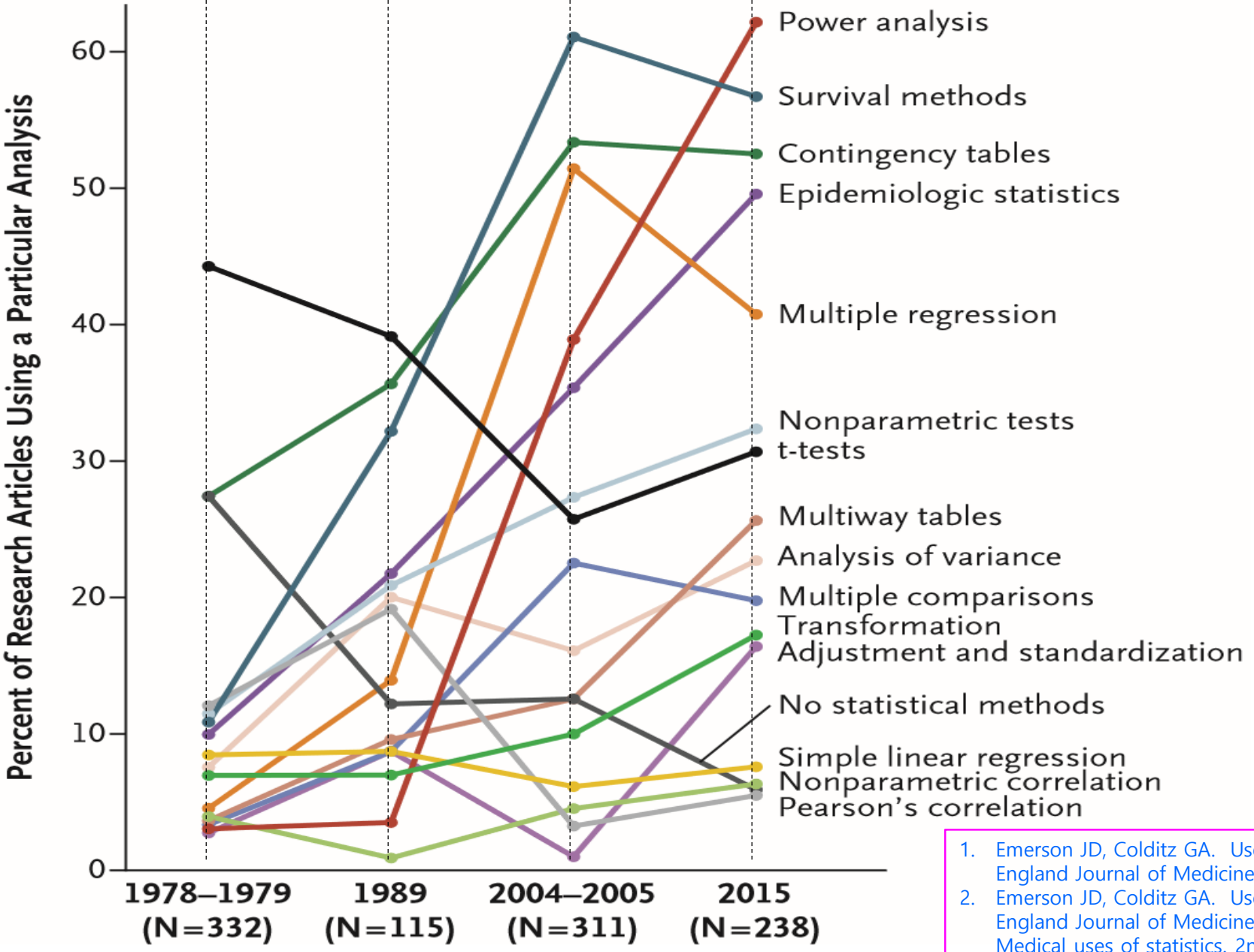
Experimental study, Quasi-experimental study

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

Categories of statistical procedures used to assess the statistical content in the articles

자료 성격	권고 통계분석방법
사례보고, 임상연구, 치료결과분석 등	No statistical method or Descriptive study
진단능력평가, 참고치 정하기	Sensitivity, Specificity, ROC curve
짝을 이룬 두 그룹간 평균비교	**Paired t-test, Wilcoxon signed rank test***
독립적인 두 그룹간 평균비교	**t-test, Wilcoxon rank sum test*, Mann-Whitney U test***
독립적인 세 그룹 이상 평균비교 (또는 군간비교)	**ANOVA (with multiple comparison), Kruskal-Wallis test***
동일인에 대한 3회 이상 반복측정자료의 평균비교	**Repeated measures of ANOVA, Friedman test***
두 그룹 또는 세 그룹 이상 빈도 비교	Chi-squared test*, Fisher's exact test*
동일인에 대한 반복측정 빈도 비교	McNemar's test*
두 연속변수간 상관관계 분석	**Pearson's correlation, Spearman's rho***
두 개 이상 독립변수와 종속변수와의 관계 분석	**Simple linear regression, Multiple (logistic) regression**
생존율 추정, 생존율 비교 생존형 자료의 회귀분석	Life table, Kaplan-Meier method Log-rank test, Cox's proportional hazard model (**HR**)
역학적 통계량 분석	Incidence, Prevalence, Risk ratio (**RR**), Odds ratio (**OR**)

Source : Emerson JD, Colditz GA. Use of Statistical Analysis in The New England Journal of Medicine. *N Engl J Med* 1983; 309: 709-713.



1. Emerson JD, Colditz GA. Use of statistical analysis in The New England Journal of Medicine. NEJM **1983**; 309: 709-13.
2. Emerson JD, Colditz GA. Use of statistical analysis in the New England Journal of Medicine. In: Bailar JC III, Mosteller F, eds. Medical uses of statistics. 2nd ed. Waltham, MA: NEJM Books, **1992**: 45-57.
3. Horton NJ, Switzer SS. Statistical methods in the Journal. NEJM **2005**; 353: 1977-9.
4. Sato Y, Gosho M. Statistical Methods in the Journal - An Update. NEJM **2017**; 376: 1086-7.

측정 수준에 따른 확률변수의 구분 :

① 범주형 변수 (categorical variable)

- **명목척도(nominal scale) :** 단지 범주로만 의미가 있다.

  ex) 성별, 혈액형, 치료방법 등

- **순위척도(ordinal scale) :** 명목 + 대소 관계를 나타냄.
  가감승제와 같은 수학적 계산은 무의미하다.

  ex) 증상의 정도, 학력수준, 사회경제적 수준 등

② 연속형 변수 (continuous variable)

- **구간척도(interval scale) :** 측정치 간의 **'간격'**에 의미가 있는 경우

  ex) IQ, 온도의 경우 10℃~15 ℃, 20℃~25 ℃의 5 ℃ 는 본질적으로 같다.
  가감은 가능하나, 승제는 불가능함 (➔ 比(ratio)의 개념은 갖지 못함)

  ex) 100℃ / 50℃ ≠ 212℉ / 122℉ (∵ 절대 0℃, 0℉ 이 아니라 인위적으로 정한 것임)

- **비율척도(ratio scale) :** 'age'

  절대 영점을 가지게 되므로 수학적으로 가장 완벽한 형태의 변수 (가감승제 모두 가능함)

  ex) 80세는 20세에 비해 60살 더 많고(구간), 4배(비율) 더 살았다.

Data Entry 유의사항

- **범주형 자료**
  - 적절한 형태의 숫자 코드 할당 (조사지 위에 함께 적어 놓는 것을 추천!!)
  - 이진수 자료의 경우 **0/1**을 사용 추천 (주로, '예'=1 / '아니오'=0)

- **연속형 자료**
  - 해당 자료를 측정한 **그대로** 기록 (크기를 줄여서 입력 지양)
  - 측정 단위는 일관성을 유지

- 한 사람 당 여러 개의 형식을 사용하는 경우
  - **id** 부여 (자료의 결합을 위해) – 중요!!

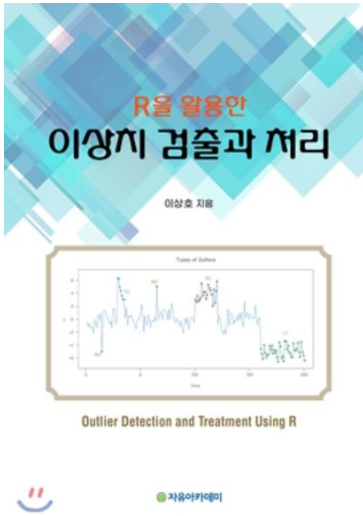
- 날짜와 시간의 문제 : 조사 / 입력 형식의 통일
- 결측치(**missing value**)의 입력 : 가능한 한 default value (**.** 또는 공백) 사용

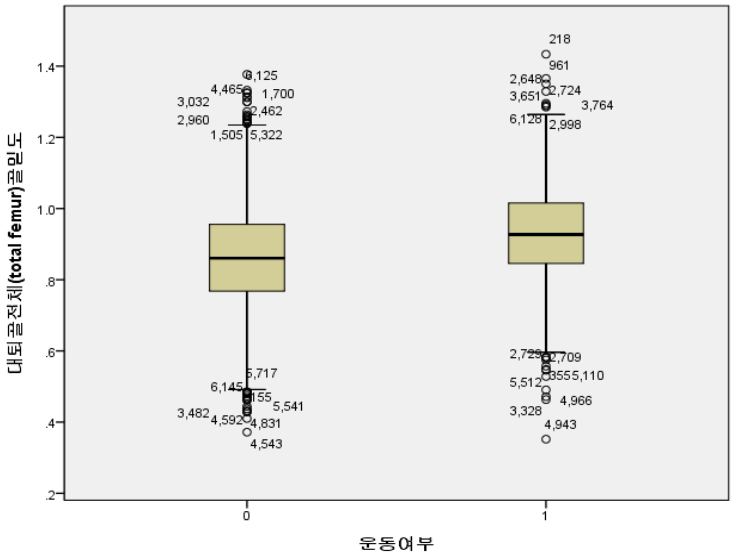
오류검토

- 원본 대조 / **double entry**
  - 둘 다 완벽한 것은 아님. 그러나 오류를 최소화 할 수 있다.
- 오류 검토
  - 범주형 자료
    - 검토가 비교적 쉬움
    - 나타나서는 안 되는 범주의 값이 입력되어 있다면 분명히 오타
    - 빈도표(frequency table)의 활용이 효과적
  - 연속형 자료
    - 오타발생 가능성 높음. 찾아내기는 어려움 (예: 소수점의 문제 등)
    - 범위(range) 검토 (Max – Min check)가 효과적
  - 날짜 자료
    - 쉬운 문제는 아님
    - 이상값의 확인 (예: 20070230)
    - 논리적인 검토 (예: 생년과 나이의 일치, 연구시작일 검토 등)
- 자료의 수정
  - 해당 자료가 잘 못 입력된 자료라는 **명백한 증거**가 있는 경우에 한함
  - 단지 이상한 값이 있다는 이유만으로 자료 수정은 곤란, 위험 함

(extreme) outliers (극)이상치

- 자료의 전반적인 값들과 구별되는 값 / 다른 값들과 병립될 수 없는 값
- 실제 관찰 값일 수도 있고 / 잘못 입력된 값일 수도 있음
  - 예, 키가 210인 여성
  - 만일 연구 대상이 일반인/초등학생이면? 만일 체중과 함께 검토한다면?
- 통계분석방법의 종류에 따라 결과에 심각한 영향을 미칠 수도
  - 예, Student's t-test vs. Wilcoxon's test
  - 따라서 자료 내에 이상치가 있는지 검토하는 것은 매우 중요
- **이상치에 대한 검토 !!!**
  - Range check이 효과적 / Graphical method (histogram, scatter plot 등) 사용도 효과적
  - 특정 통계모형 내에서도 가능 (예: 회귀분석 등)
- 이상치의 처리
  - 무분별한 삭제는 곤란
  - 최선의 방법 : **with / without analysis**
    - 분석결과가 서로 비슷하면 ok
    - 결과가 서로 상이하면 이상치에 영향을 많이 받지 않는 분석법 (예 : 자료의 변환, 비모수적 방법 등)





정규성 검정 (normality test) :

주어진 자료가 정규 모집단에서 랜덤하게 뽑힌 것인지 확인하고자 할 때 사용한다. 정규 모집단이라는 것이 확인되면 **모수적 검정(ex. Independent t-test)**를 하고, 아니면 **비모수적 검정(ex. Mann-Whitney U-test)**를 수행한다. 히스토그램을 통해 자료의 분포가 정규분포와 유사한지 확인해 볼 수 있고, 통계적 검정방법으로는 콜로고로프-스미르노프 검정과 샤피로-윌크 검정을 주로 사용한다.

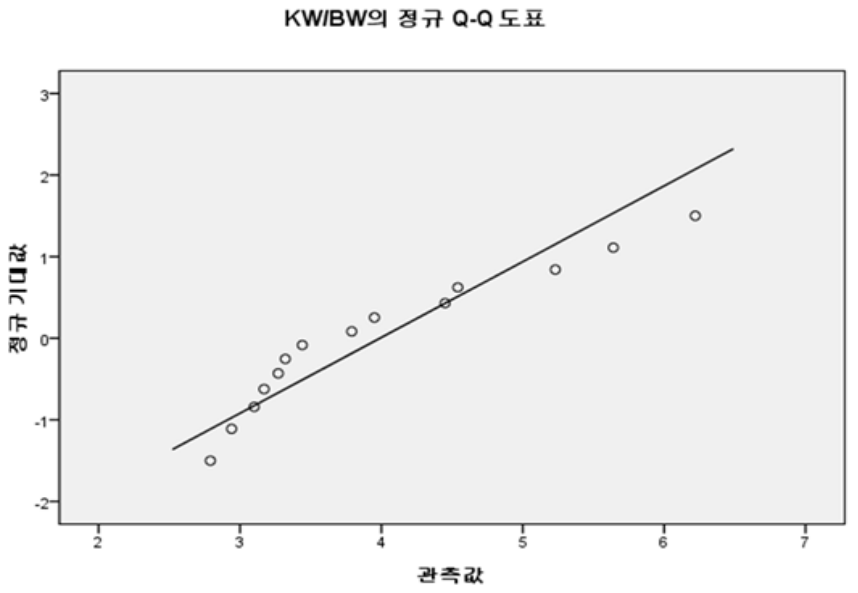
- **콜모고로프-스미르노프 검정 (Kolmogorov-Smirnov test) :** 자료의 가장 작은 값부터 가장 큰 값까지의 누적상대빈도가 이론적 정규분포에서의 누적상대빈도와 얼마나 다른가를 측정하여 검정하는 방법이다.

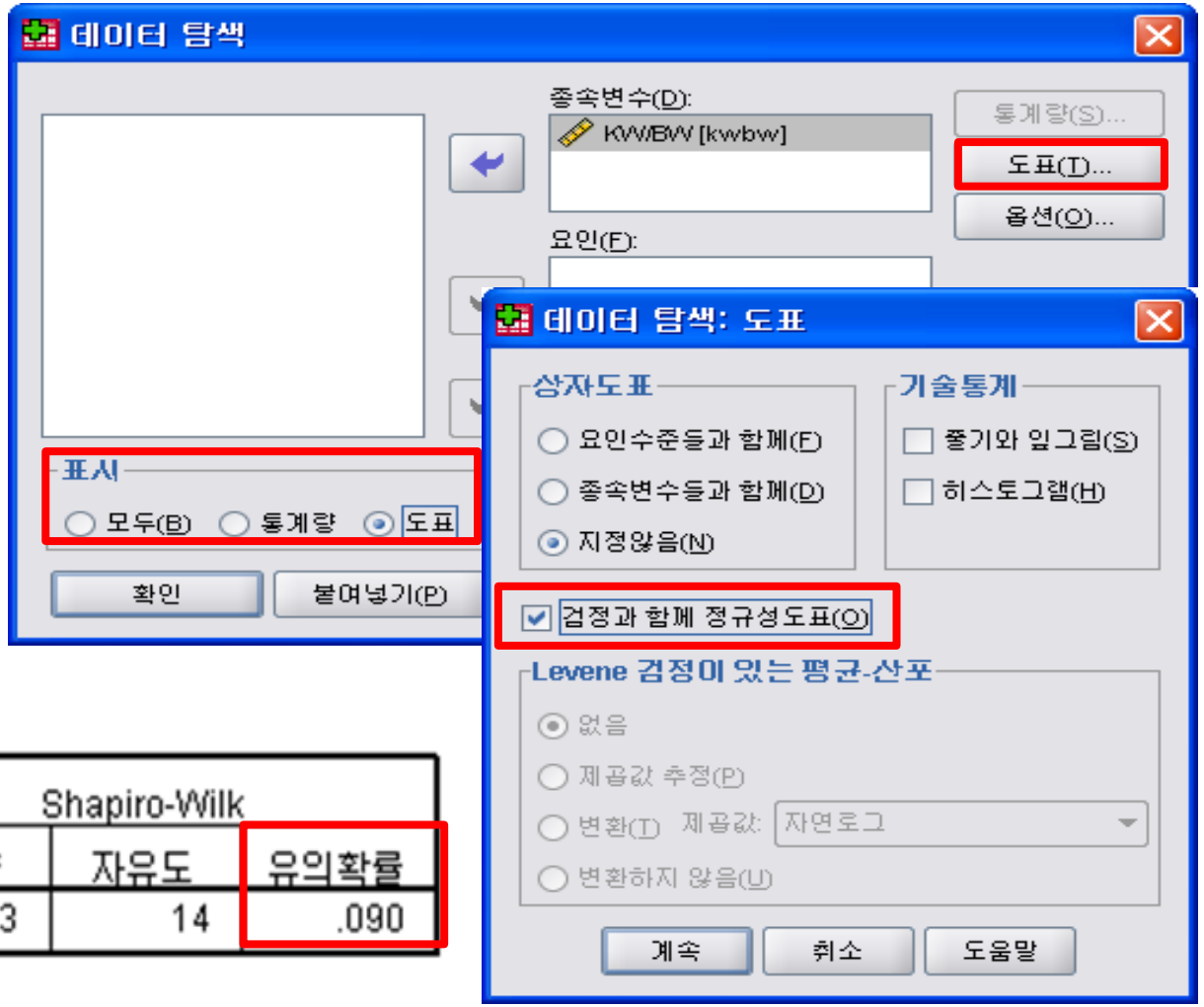
- **샤피로-윌크 검정 (Shapiro-Wilk test) :** 자료 값들과 표준정규점수와의 선형상관관계를 측정하여 검정하는 방법이다.

두 검정법의 가설은 "주어진 자료가 정규분포를 따른다"이므로 유의확률이 0.05 보다 클 때 정규성 가정을 만족한다고 할 수 있다. 표본크기(sample size)가 충분히 클 때, 콜모고로프-스미르노프 검정법을 사용하고, 작을 때는 샤피로-윌크 검정법을 사용한다. 이 외에 **크래머-본 미제스(Cramér–von Mises), 앤더스-달링(Anderson-Darling) 검정** 등이 있다.

생체신장이식 과정 중 이식신장의 실제무게를 전향적으로 측정하여 공수여자간의 체격불일치 척도로 공여신장무게와 신장수여자체중 간의 비율(KW/BW) kidney weight(g) / recipient body weight(kg)을 측정하였다. 14명의 비율값이 정규분포를 따른다고 할 수 있는가?

KW/BW	3.79	2.79	2.94	3.95	4.45	4.54	5.64	6.22	5.23	3.10	3.17	3.44	3.32	3.27





정규성 검정

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	통계량	자유도	유의확률	통계량	자유도	유의확률
KW/BW	.195	14	.155	.893	14	.090

비모수적 통계분석 (nonparametric analysis)

모수적(parametric) 통계분석은 기본적으로 모집단의 정규성, 등분산성, 측정치의 연속성을 가정하거나 조건으로 하는 방법이다. 하지만 실제 임상연구 자료에서는 정규분포의 가정을 만족하지 않아도 큰 문제가 되지 않는 경우가 많고, 표본수가 작아서 정규분포의 가정을 만족하지 않는 경우의 연구도 많다. 따라서 **비모수적 방법**은 표본수가 크지 않고, 모집단의 분포가 정규분포를 따르지 않는 경우에 효율적인 통계분석 방법이 될 수 있다.

➔ **비모수적 통계분석의 기본은**

- **모집단은 연속성이어야 하며, 그 분포에 대한 가정은 필요 없다.**
- **측정치 보다는 그들의 상대 순위(rank) 혹은 서열(order)에 기반한다.**
- **평균(mean)을 비교하는 것이 아니라 중위수(median)을 비교한다.**
- **변이 정도는 표준편차(sd)가 아니라 범위(range)나 사분위수(IQR)로 표기한다.**

자료의 특성에 따른 통계분석	모수적 방법	비모수적 방법
짝을 이룬 두 그룹간의 비교	Paired t-test	**Wilcoxon signed rank test**
독립적인 두 그룹간의 비교	Independent two-sample t-test	**Wilcoxon rank sum test** **Mann-Whitney U test**
독립적인 세 그룹 이상의 비교	One-way ANOVA Two-way ANOVA	**Kruskal-Wallis test** **Friedman test**
두 연속형 변수간의 상관분석	Pearson's correlation	**Spearman's rank correlation** **Kendall's tau**

Association (연관성, 관련성) :

두 변수 간의 일반적인 관계

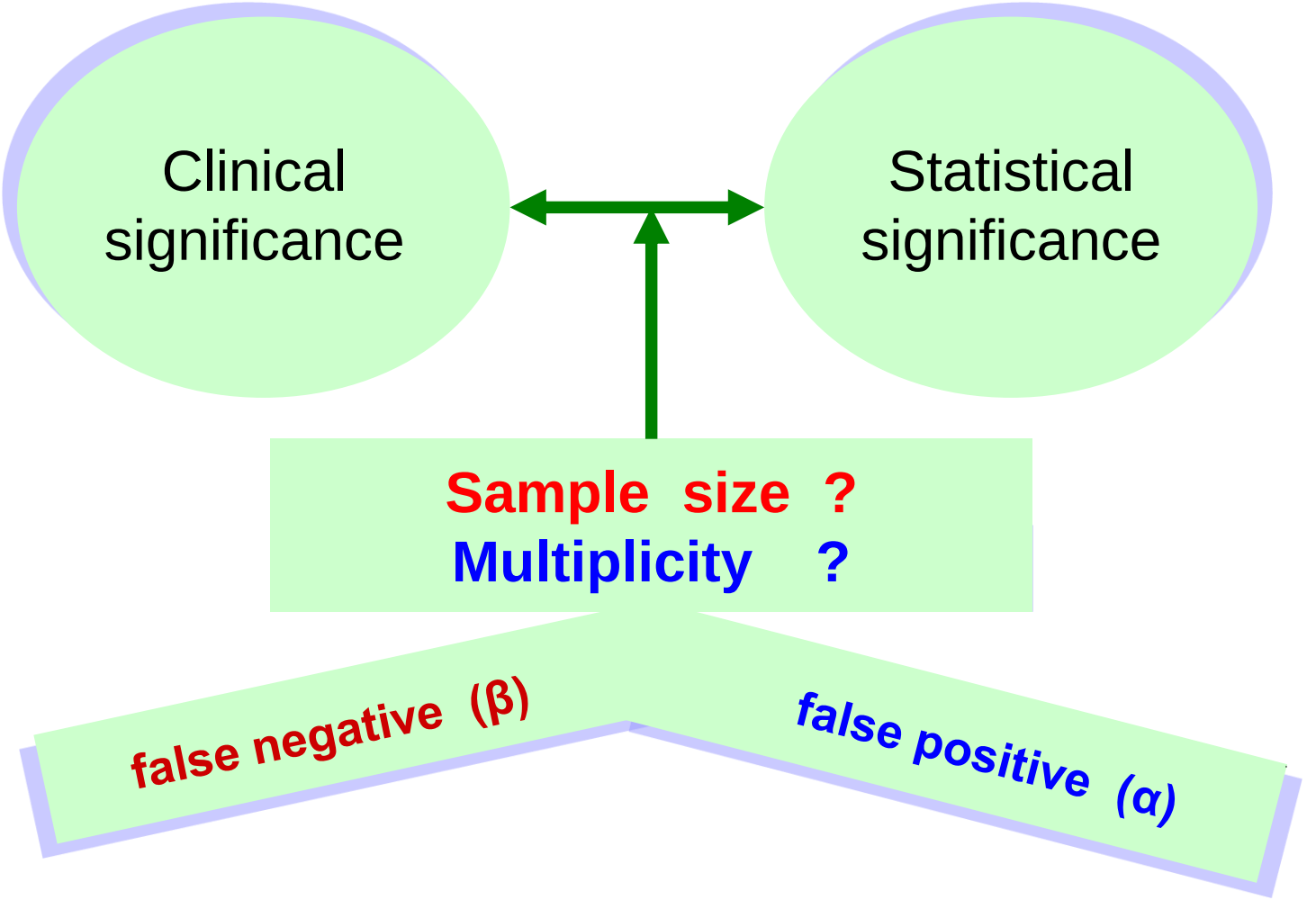
- 동질성 검정 (homogeneity test)
- 독립성 검정 (independence test)

Correlation (상관성) :

연관성의 한 종류로 선형성을 정량화

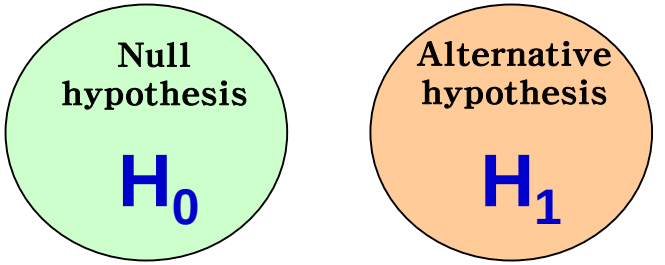
Causation (인과성) :

방향성이 포함된 원인-결과 관계



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

Errors in Hypothesis Testing



H_0 : No difference between effects of two drugs

H_1 : Not H_0

Decision \ True	H_0 is True	H_0 is False
Fail to Reject H_0	☺ Ex. = Control	☹ β Type II error Ex. = Control
Reject H_0	☹ ☹ α Type I error Ex. ≠ Control	☺ ☺ Ex. ≠ Control "Power (1-β)"

Type I error = P(Positive | H_0:True) = "False Positive"

Type II error = P(Negative | H_0:False) = "False Negative"

▪ **유의수준, level of significance α=5%,** 이는 **우연**에 의해 연구결과가 참이 되어버리는 것(false positive)을 5%까지 허용하겠다는 연구자의 의지이다.

▪ **유의확률, p-value,** 내 데이터를 이용한 가설검정의 결과를 p-value로 요약됨. H_0이 참인 연구결과가 **우연**에 의해 얻어지는 확률을 의미한다.

➔ H_0가 참인 확률(p-value)이 연구자가 미리 정해놓은 유의수준 α=5%보다 작으면 H_0이 거짓이다. H_0를 기각한다 ('차이가 없다는 것이 아니다'). 반증적으로 연구자가 원하는 H_1을 증명한다. (all or none)

(Frequentist)귀무가설이 참/거짓 판단 vs. (Bayesian)연구가설이 참이 되는 확률 계산

➔ Bayesian 입장 에서는 *all uncertainty is measured by probability* 이며 *continuous learning process* 여서 임상연구를 중간에 조기종료 / 수정이 가능할 수 있다. 이러한 Bayesian concept이 임상시험에 많이 도입되면서 adaptive design이 가능하게 되었다.

➔ **Baye's Theorem :** $$P(\theta \mid data) = \frac{P(data \mid \theta)}{P(data)} \times P(\theta)$$

p-value, Statistical significance & Clinical significance

$0.01 \leq p < 0.05$	➔	significant
$0.001 \leq p < 0.01$	➔	highly significant
$p < 0.001$	➔	very highly significant
$p > 0.05$	➔	not statistically significant
$0.05 \leq p < 0.10$	➔	a trend toward statistical significance is sometimes noted

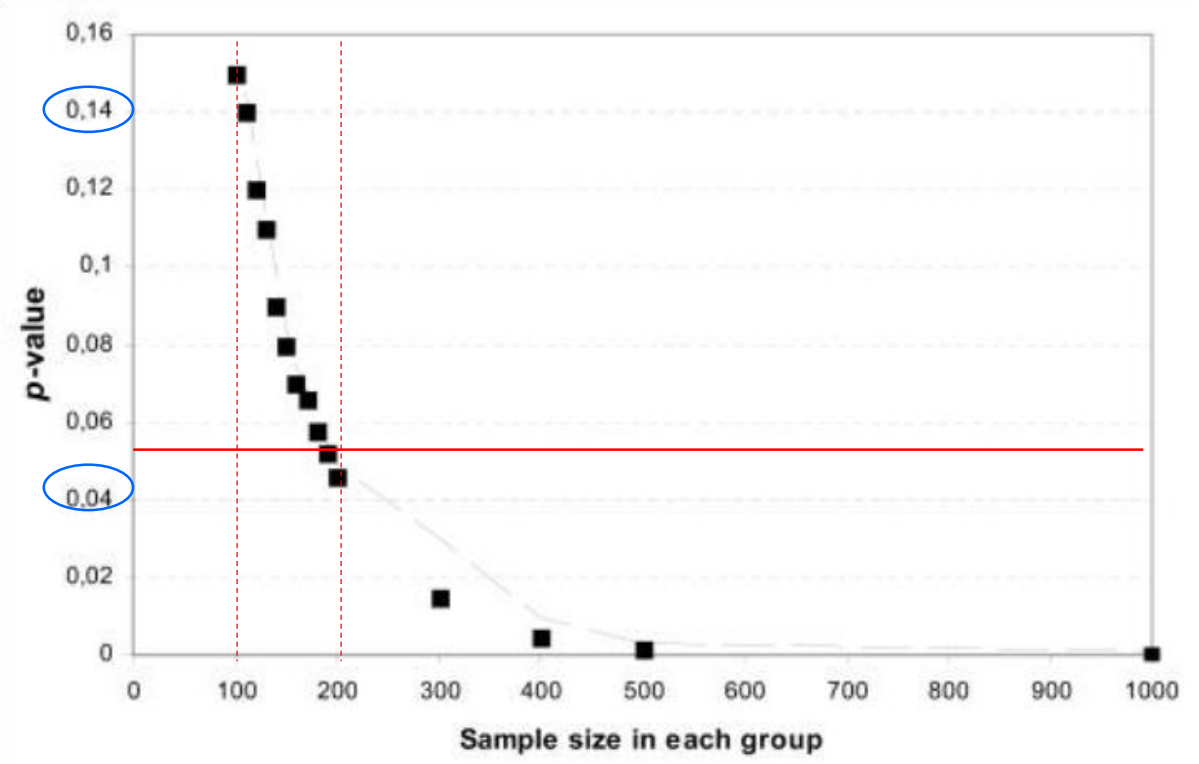
❖ Analyses were performed with SAS software, version 9.1 (SAS Institute, Cary, NC). A 2-sided probability value 0.05 was considered statistically significant.

❖There was a borderline significant improvement in survival for the experimental arm with a median time to death of 63.2 months compared with 52.2 months in the standard cisplatin/paclitaxel arm (log-rank ***P = 0.05***, *one-tail)*

❖ The multiplicative interaction term between vitamin D deficiency and hypertension had a borderline statistical significance (***P = 0.08*** *for both 2- and 3-category 25-OH D* models).

The Value of *p*-value in Biomedical Research

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



- ***p*-value 0.05가 '유의함(significance)'을 판단하는 절대적인 기준인가?**
- ***p*-value에 의존한 통계적 유의성 해석의 한계 有** (동일한 효과에서도 표본수에 따라 ***p*-value**가 달라진다)
  - ➔ 의사결정에서 ***p*-value**는 중요한 기준은 되지만 이것의 해석에는 한계가 있다.
- ***p*-value는 관찰된 효과의 중요성 정도는 설명하지 못한다.** 즉, *p*-value가 작다고 **association**이 강하다는 의미는 아니다.
- ***p*-value의 수준은 sample size와 밀접한 관련이 있다.**
  - ➔ *p-value*에 의한 해석에 의존한다면, 상대적으로 **N**수가 적은 '희귀질환'의 경우, 해당 의약품의 효과를 입증하기 위한 sample size에 도달하지 못하여 임상적 유용성 개선의 입증이 어렵게 된다.

Table 1. Key Questions to Ask When the Primary Outcome Is Positive.

Does a P value of <0.05 provide strong enough evidence?
What is the magnitude of the treatment benefit?
Is the primary outcome clinically important (and internally consistent)?
Are secondary outcomes supportive?
Are the principal findings consistent across important subgroups?
Is the trial large enough to be convincing?
Was the trial stopped early?
Do concerns about safety counterbalance positive efficacy?
Is the efficacy–safety balance patient-specific?
Are there flaws in trial design and conduct?
Do the findings apply to my patients?

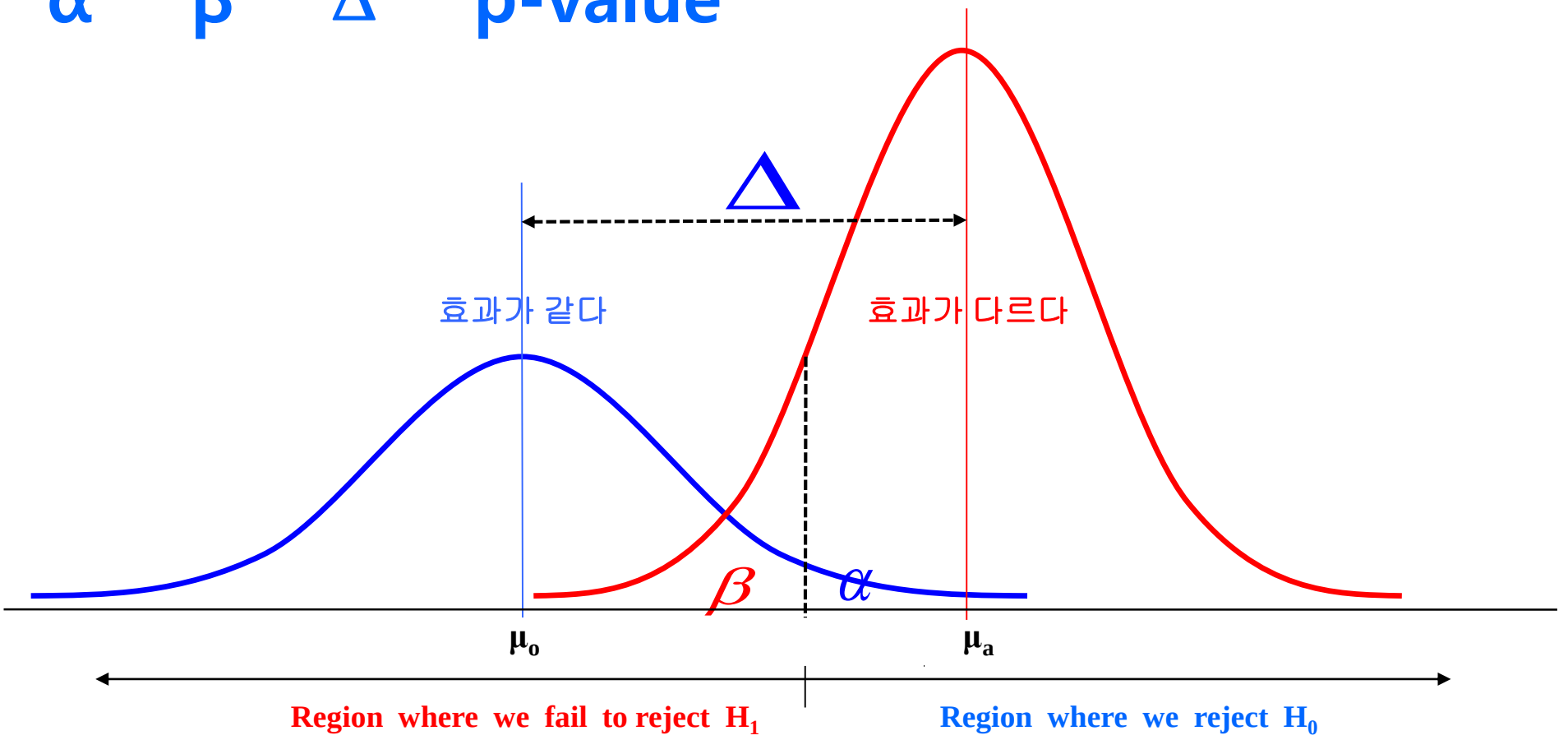
Stuart et al., N Engl J Med 2016;375:861-70.

Stuart et al., N Engl J Med 2016;375:971-9.

Table 1. Questions to Ask When the Primary Outcome Fails.

Is there some indication of potential benefit?
Was the trial underpowered?
Was the primary outcome appropriate (or accurately defined)?
Was the population appropriate?
Was the treatment regimen appropriate?
Were there deficiencies in trial conduct?
Is a claim of noninferiority of value?
Do subgroup findings elicit positive signals?
Do secondary outcomes reveal positive findings?
Can alternative analyses help?
Does more positive external evidence exist?
Is there a strong biologic rationale that favors the treatment?

α β Δ p-value



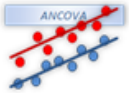
"검정력"	0.80	0.90	0.95	0.99
Z_β	0.840	1.282	1.645	2.326

"유의수준"	0.01	0.05	0.10
Z_α	2.326	1.645	1.282
$Z_{\alpha/2}$	2.576	1.960	1.645

Considerations for sample size determination

- ✓ **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* ($n' = n/c^2 = 100/0.8^2 = 156$)**

Acceptance Sampling for Attributes



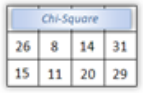
Analysis of Covariance

Balanced Incomplete Block Designs

Bartlett Test of Variances (Simulation)

Brown-Forsythe Test of Variances (Simulation)

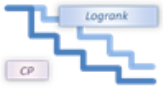
Chi-Square Effect Size Estimator



Chi-Square Tests

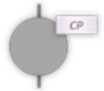
Cochran-Armitage Test for Trend in Proportions

Conditional Power of 2x2 Cross-Over Designs



Conditional Power of Logrank Tests

Conditional Power of One Proportion Tests



Conditional Power of One-Sample T-Tests

Conditional Power of Paired T-Tests

Conditional Power of Two Proportions Tests

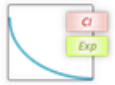
Conditional Power of Two-Sample T-Tests

Confidence Interval for Pearson's Correlation

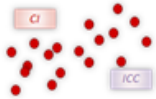
Confidence Intervals for Coefficient Alpha

Confidence Intervals for Cp

Confidence Intervals for Cpk



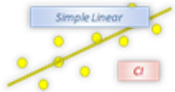
Confidence Intervals for Exponential Reliability



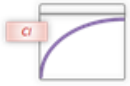
Confidence Intervals for Intraclass Correlation

Confidence Intervals for Kappa

Confidence Intervals for Kendall's Tau-b Correlation



Confidence Intervals for Linear Regression Slope



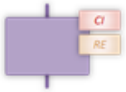
Confidence Intervals for Michaelis-Menten Parameters

Confidence Intervals for One Mean

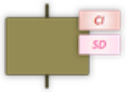
Confidence Intervals for One Mean with Tolerance Probability

Confidence Intervals for One Proportion

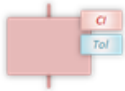
Confidence Intervals for One Proportion from a Finite Population



Confidence Intervals for One Standard Deviation using Relative Error

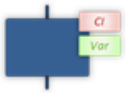


Confidence Intervals for One Standard Deviation using Standard Deviation

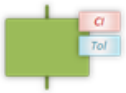


Confidence Intervals for One Standard Deviation with Tolerance Probability

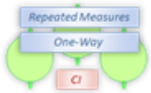
Confidence Intervals for One Variance using Relative Error



Confidence Intervals for One Variance using Variance



Confidence Intervals for One Variance with Tolerance Probability



Confidence Intervals for One-Way Repeated Measures Contrasts

PASS 14 Citation: http://www.ncss.com

PASS 14 Power Analysis and Sample Size Software (2015). NCSS, LLC. Kaysville, Utah, USA, ncss.com/software/pass.

PASS 14

Power Analysis and Sample Size Software from NCSS

두 집단의 유효율 비교 : Superiority Trial

- 실험군(Cisplatin+Topotecan) vs. 대조군(Cisplatin)
- 일차결과변수 : complete response rate
- 대립가설의 형태 : 양측검정
- 최소 유의한 차이 정도 : $\Delta = P_B - P_A = 15\%$ ($P_A = 15\%$)
- α=0.05, β=0.20
- 할당비 : 1:1

$$n_A = \frac{\left[z_{\alpha/2}\sqrt{2\bar{p}\bar{q}} + z_{\beta}\sqrt{p_A q_A + p_B q_B} \right]^2}{(p_A - p_B)^2}$$

여기서, $\bar{p} = (p_A + p_B)/2, \quad \bar{q} = 1 - \bar{p}$

$$n_A = \frac{\left[1.96\sqrt{2\times 0.225\times 0.775} + 0.84\sqrt{0.15\times 0.85 + 0.3\times 0.7} \right]^2}{(0.30 - 0.15)^2}$$

$$\approx 121$$

Solve For: Sample Size

1 = Treatment
2 = Control

Power Calculation

Power Calculation Method: Normal Approximation

Test

Alternative Hypothesis: Two-Sided

Test Type: Z-Test (Pooled)

Power and Alpha

Power: 0.8

Alpha: 0.05

Sample Size

Group Allocation: Equal (N1 = N2)

Effect Size

Input Type: Proportions

P1 (Group 1 Proportion|H1): 0.20 0.25 0.26 0.27 0.28 0.29 0.30 0.35

P2 (Group 2 Proportion): 0.15

D1 = the actual difference under H1 = P1 - P2

"complete response rate"에 따른
최소 필요한 각 군당 연구대상자수

Numeric Results for Testing Two Proportions using the Z-Test with Pooled Variance
H0: P1 - P2 = 0. H1: P1 - P2 = D1 ≠ 0.

Target Power	Actual Power*	N1	N2	N	P1	P2	Diff D1	Alpha
0.80	0.80028	906	906	1812	0.2000	0.1500	0.0500	0.0500
0.80	0.80003	250	250	500	0.2500	0.1500	0.1000	0.0500
0.80	0.80140	211	211	422	0.2600	0.1500	0.1100	0.0500
0.80	0.80073	180	180	360	0.2700	0.1500	0.1200	0.0500
0.80	0.80106	156	156	312	0.2800	0.1500	0.1300	0.0500
0.80	0.80217	137	137	274	0.2900	0.1500	0.1400	0.0500
0.80	0.80173	121	121	242	0.3000	0.1500	0.1500	0.0500
0.80	0.80333	73	73	146	0.3500	0.1500	0.2000	0.0500

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

multiple comparison, post-hoc analysis

독립적인 여러 개의 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 가정하기 어렵다.

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

multiple comparison

검정법	비교시기	비교집단	표본 수	비고
Tukey HSD 正直有意差検定 (Honestly Significant Difference)	사후	Pairwise	Equal	필요이상으로 보수적-정직적 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, Hochberg, Hommel

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

※ Fisher의 보호/비보호 최소유의차검정 (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

The Multtest Procedure

P-Value Adjustment Information

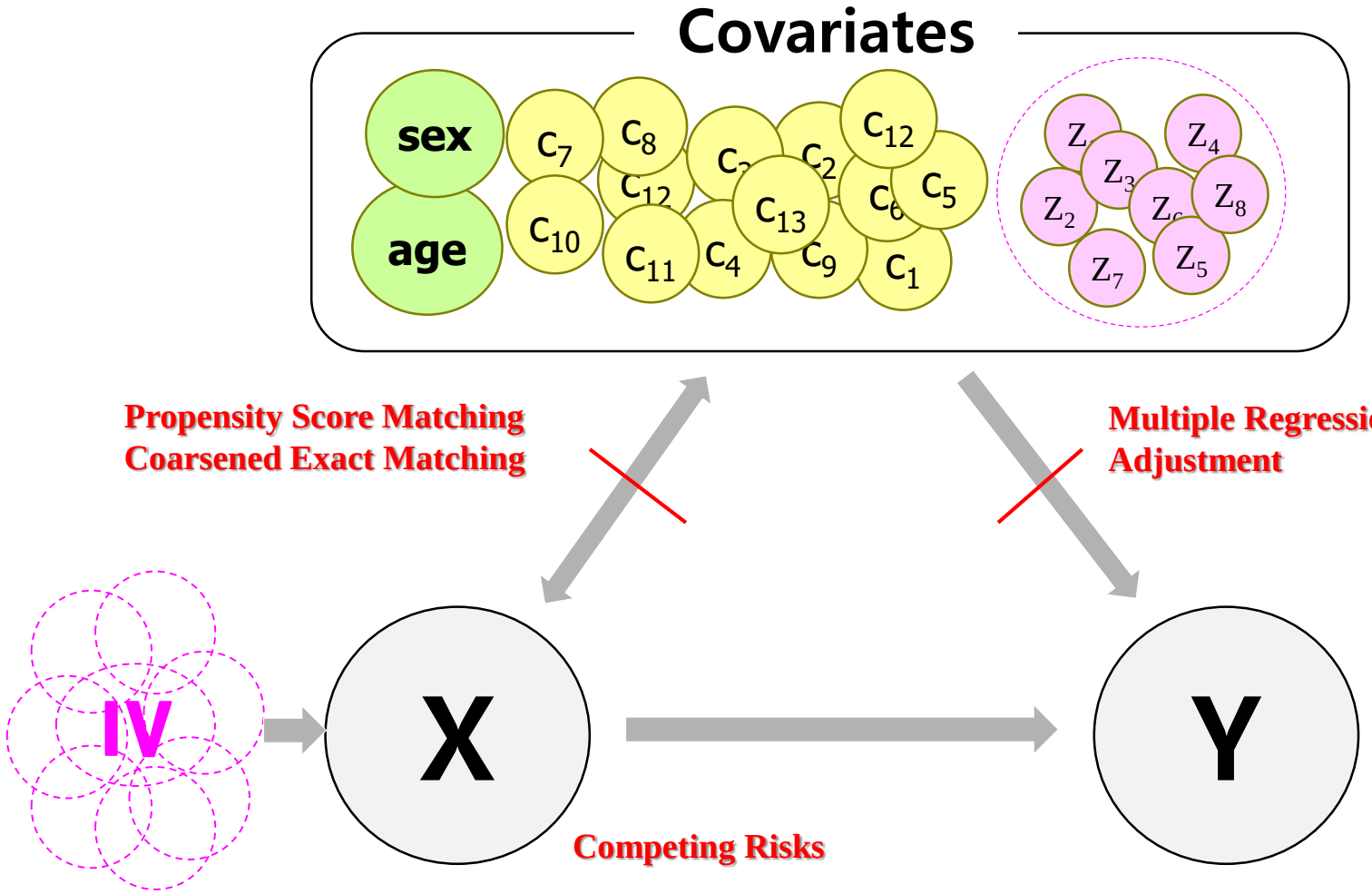
P-Value Adjustment	Bonferroni
P-Value Adjustment	Stepdown Bonferroni
P-Value Adjustment	Sidak
P-Value Adjustment	Stepdown Sidak
P-Value Adjustment	Hochberg
P-Value Adjustment	False Discovery Rate

(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	0.0720	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

Statistical Methods for Causal Inference



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

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

Table 3. Selected Baseline and Exercise Characteristics According to Aspirin Use in Propensity-Matched Patients*

2310 3864

Variable	Aspirin (n = 1351)	No Aspirin (n = 1351)	*P* Value
Demographics			
Age, mean (SD), y	60 (11)	61 (11)	.16
Men, No. (%)	951 (70)	974 (72)	.33
Clinical history			
Diabetes, No. (%)	203 (15)	207 (15)	.83
Hypertension, No. (%)	679 (50)	698 (52)	.46
Tobacco use, No. (%)	161 (12)	162 (12)	.95
Cardiac variables			
Prior coronary artery disease, No. (%)	652 (48)	659 (49)	.79
Prior coronary artery bypass graft, No. (%)	251 (19)	235 (17)	.42
Prior percutaneous coronary intervention, No. (%)	166 (12)	147 (11)	.25
Prior Q-wave MI, No. (%)	194 (14)	206 (15)	.52
Atrial fibrillation, No. (%)	21 (2)	24 (2)	.65
Congestive heart failure, No. (%)	79 (6)	89 (7)	.43
Medication use			
Digoxin use, No. (%)	115 (9)	114 (9)	.94
β-Blocker use, No (%)	352 (26)	358 (26)	.79
Diltiazem/verapamil use, No. (%)	223 (17)	223 (17)	>.99
Nifedipine use, No. (%)	127 (9)	144 (11)	.28
Lipid-lowering therapy, No. (%)	281 (21)	271 (20)	.63
ACE inhibitor use, No. (%)	209 (15)	214 (16)	.79
Cardiovascular assessment and exercise capacity			
Body mass index, mean (SD), kg/m^2	29 (6)	29 (6)	.83
Ejection fraction, mean (SD), %	51 (8)	51 (9)	.65
Resting heart rate, mean (SD), beats/min	77 (13)	76 (14)	.13
Resting blood pressure, mean (SD), mm Hg			
Systolic	141 (21)	141 (21)	.68
Diastolic	85 (11)	86 (11)	.57
Purpose of test to evaluate chest pain, No. (%)	153 (11)	159 (12)	.72
Mayo Risk Index ≥1, No. (%)†	1108 (82)	1110 (82)	.92
Peak exercise capacity, mean (SD), METs			
Men	8.7 (2.5)	8.3 (2.5)	.01
Women	6.5 (2.0)	6.7 (2.0)	.13
Heart rate recovery, mean (SD), beats/min	28 (12)	28 (11)	.82
Ischemic ECG changes with stress, No. (%)	231 (22)	223 (21)	.64
Echocardiographic left ventricular ejection fraction ≤40%, No. (%)	147 (11)	156 (12)	.50
Stress-induced ischemia on echocardiography, No. (%)	239 (18)	259 (19)	.32
Fair or poor physical fitness for age and sex,[13] No. (%)	445 (33)	459 (34)	.57

*MI indicates myocardial infarction; ACE, angiotensin-converting enzyme; MET, metabolic equivalent task; and ECG, electrocardiogram.
†The Mayo Risk Index is described in the "Methods" section.

Figure 1. Kaplan-Meier Curve Relating Aspirin Use to Time to Death Among Propensity-Matched Patients

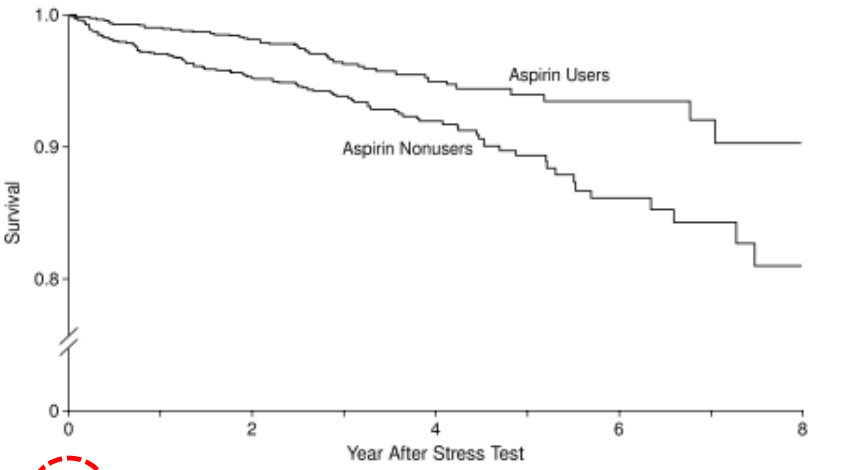


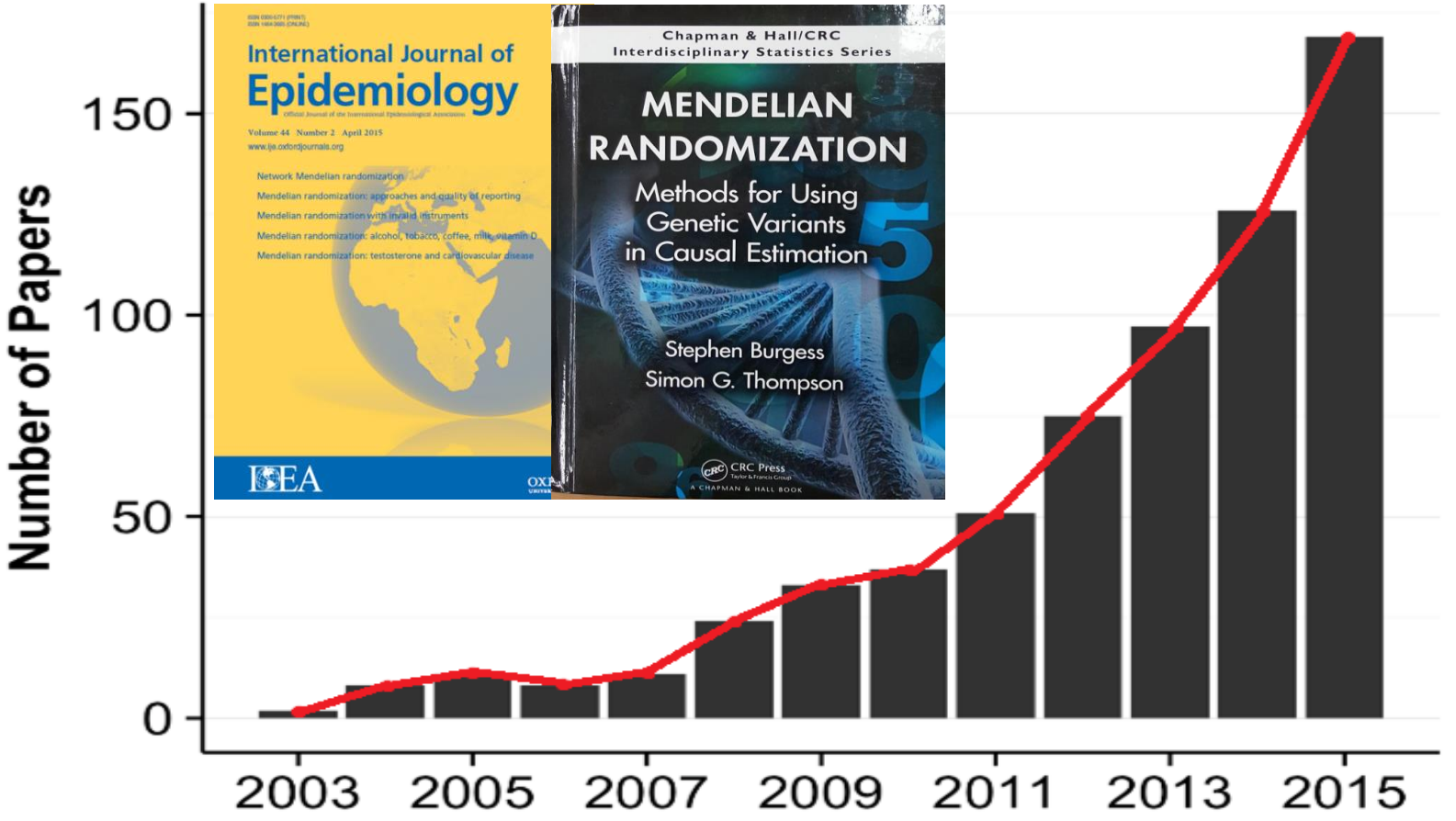
Table 4. Cox Proportional Hazards Analyses of Aspirin Use and Mortality Among Propensity-Matched Patients (n = 2702)*

Model	Hazard Ratio (95% CI)	*P* Value
Unadjusted	0.53 (0.38-0.74)	.002
Adjusted for propensity	0.53 (0.38-0.74)	<.001
Adjusted for propensity and selected variables†	0.59 (0.42-0.83)	.002
Adjusted for propensity and all covariates‡	0.56 (0.40-0.78)	<.001

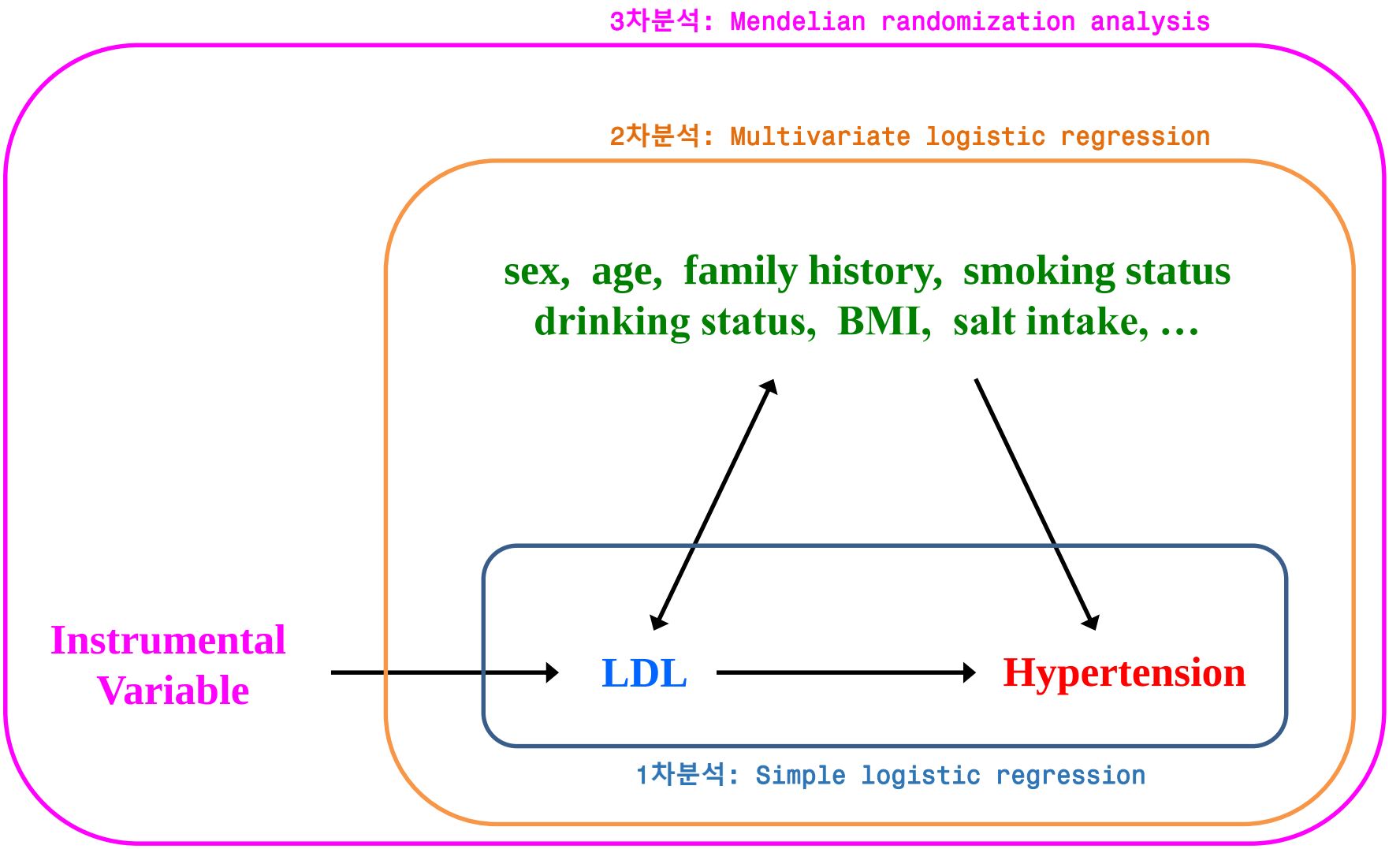
*CI indicates confidence interval.
†Selected variables included prior coronary artery disease, prior coronary artery bypass grafting, prior percutaneous intervention, and ejection fraction ≤40%.
‡For a list of covariates, see Table 2 footnote (†).

Gum et al., JAMA 2001;286(10):1187-1194

Number of Publication Using MR approach (2003~2015)



GROWTH OF THE NUMBER OR **MR** STUDIES AS ESTIMATED BY A PUBMED SEARCH OF **"MENDELIAN RANDOMISATION"** OR **"MENDELIAN RANDOMIZATION"** ON THE 7TH OF DECEMBER 2015 (US NATIONAL LIBRARY OF MEDICINE 2015.)



KoGES (AIE chip, K-chip) :

rs10903129, rs11206510, rs2479409, rs505151, rs12130333, rs629301, rs599839, rs174547, rs174570, rs7953249, rs2259816, rs4942486, rs9989419, rs314253, rs10401969, rs16996148, rs753381

※ Affy 6.0, Illumina Omni, Exome chip에 공통적으로 있는 SNP 중, **GLGC** (Global Lipids Genetics Consortium)에 있는 것을 선정

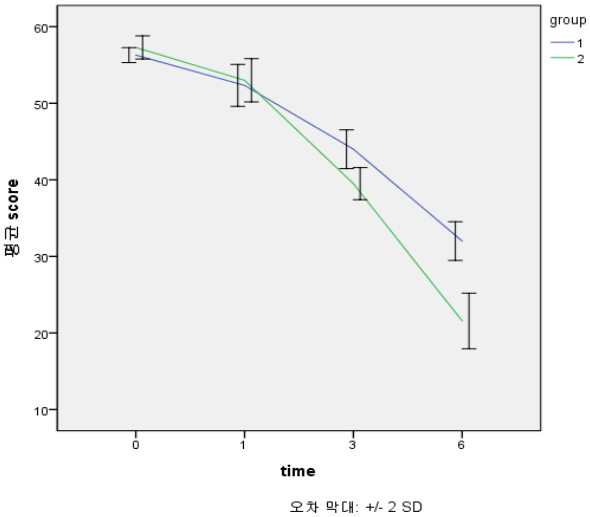
반복측정 - 추적조사 자료분석

종속변수	독립변수	통계분석법
연속형	범주형(3개 이상)	ANOVA **Repeated Measures ANOVA**
연속형	연속형 + 범주형	회귀분석 General LM / **LMM** Linear Mixed Model
이분형	연속형 + 범주형	로지스틱 회귀분석 HGLM / **GEE** Generalized Estimating Equations
생존시간	연속형 + 범주형	**Cox PH 모형** Frailty 모형

Time을 '범주형' 으로 고려

	Group1 (n=7) Estimated Mean(SE)	Group2 (n=7) Estimated Mean(SE)	p-value
month0	58.286(0.240)	57.286(0.240)	group: <0.001 time: <0.001 group*time: <0.001
month1	52.184(0.558)	52.963(0.603)	
month3	43.948(0.453)	39.516(0.453)	
month6	31.959(0.621)	21.571(0.591)	

그래프



- LMM으로 분석
- 시간의 흐름에 따라 두 군의 변화 패턴이 다름을 알 수 있음.
- 1번 군에 비해 2번 군이 여드름의 중증도가 더 감소함을 알 수 있음.
- 특히 3개월째부터 두 군간 차이가 도드라짐.

Group (x4) post-hoc p-value		Time (x6) post-hoc p-value			GroupxTime (x6) post-hoc p-value	
	Group 1 vs. 2		Group=1	Group=2		Group 1 vs. 2
mo0	0.012	mo0 vs. mo1	<0.001	<0.001	mo0 vs. mo1	0.801
mo1	0.367	mo0 vs. mo3	<0.001	<0.001	mo0 vs. mo3	<0.001
mo3	<0.001	mo0 vs. mo6	<0.001	<0.001	mo0 vs. mo6	<0.001
mo6	<0.001	mo0 vs. mo1	<0.001	<0.001	mo1 vs. mo3	0.001
		mo0 vs. mo3	<0.001	<0.001	mo1 vs. mo6	<0.001
		mo0 vs. mo6	<0.001	<0.001	mo3 vs. mo6	0.001

※ 보수적으로는 Bonferroni correction을 위해 나온 p-value에 비교횟수만큼 곱해줌

Clinical Research

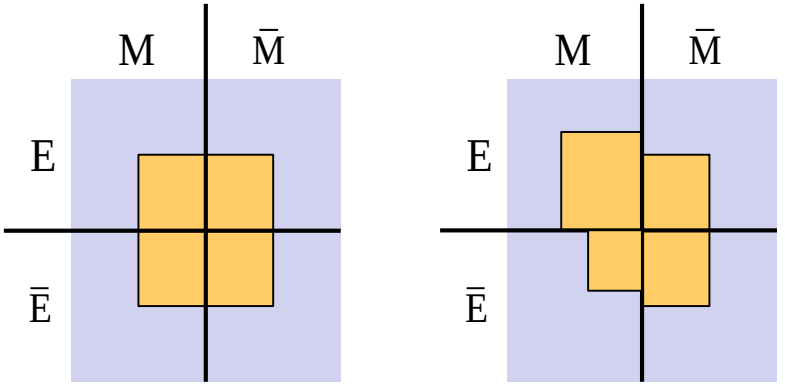
Random error / Systematic error

Bias

편향(偏向) 편의(偏倚)

뒤틀림 비뚤림 치우침

연구설계 단계 bias



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

Non comparability bias

- ✓ lead time bias, length bias, historical control bias

Sample size bias

자료수집 과정에서의 information bias

instrument bias

data source bias

observer bias

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

subject bias

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

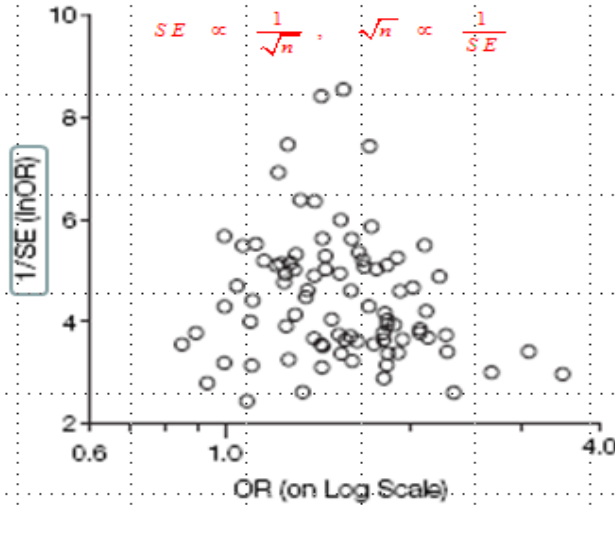
분석 & 결과 해석 과정에서의 bias

confounding bias

analysis strategy bias

: missing data handling, outlier handling, unit of analysis

post-hoc analysis bias (← data dredging bias)



assumption bias

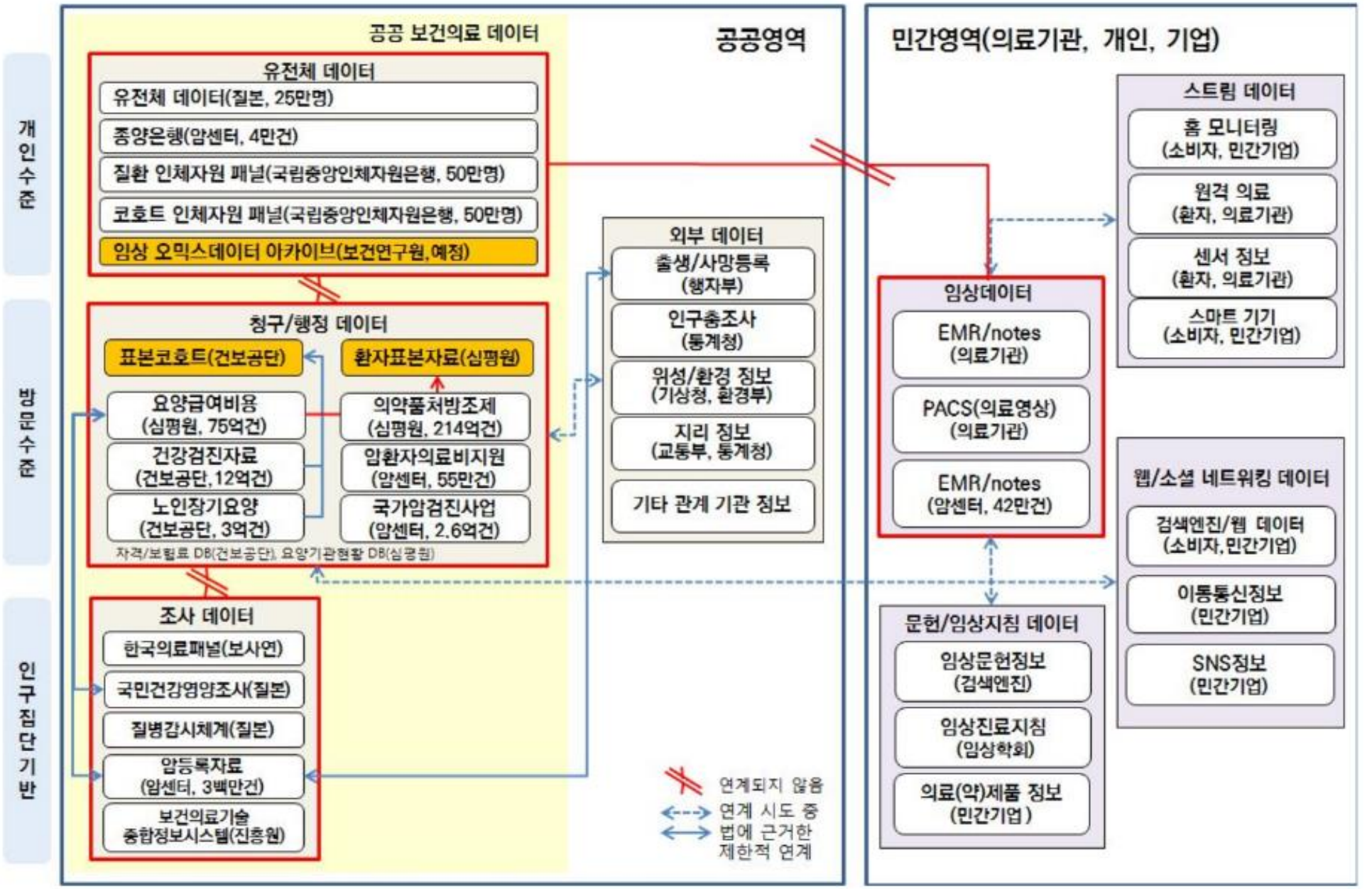
generalization bias (← lack of external validity)

significance bias

: statistical significance vs. biological significance

Publication bias (by Funnel plot, Egger의 회귀비대칭성 검정)

Possibility of Connection for Korea's Healthcare Bigdata



자료: 강희정 등(2015), 보건의료 빅데이터 활용을 위한 기본계획 수립 연구