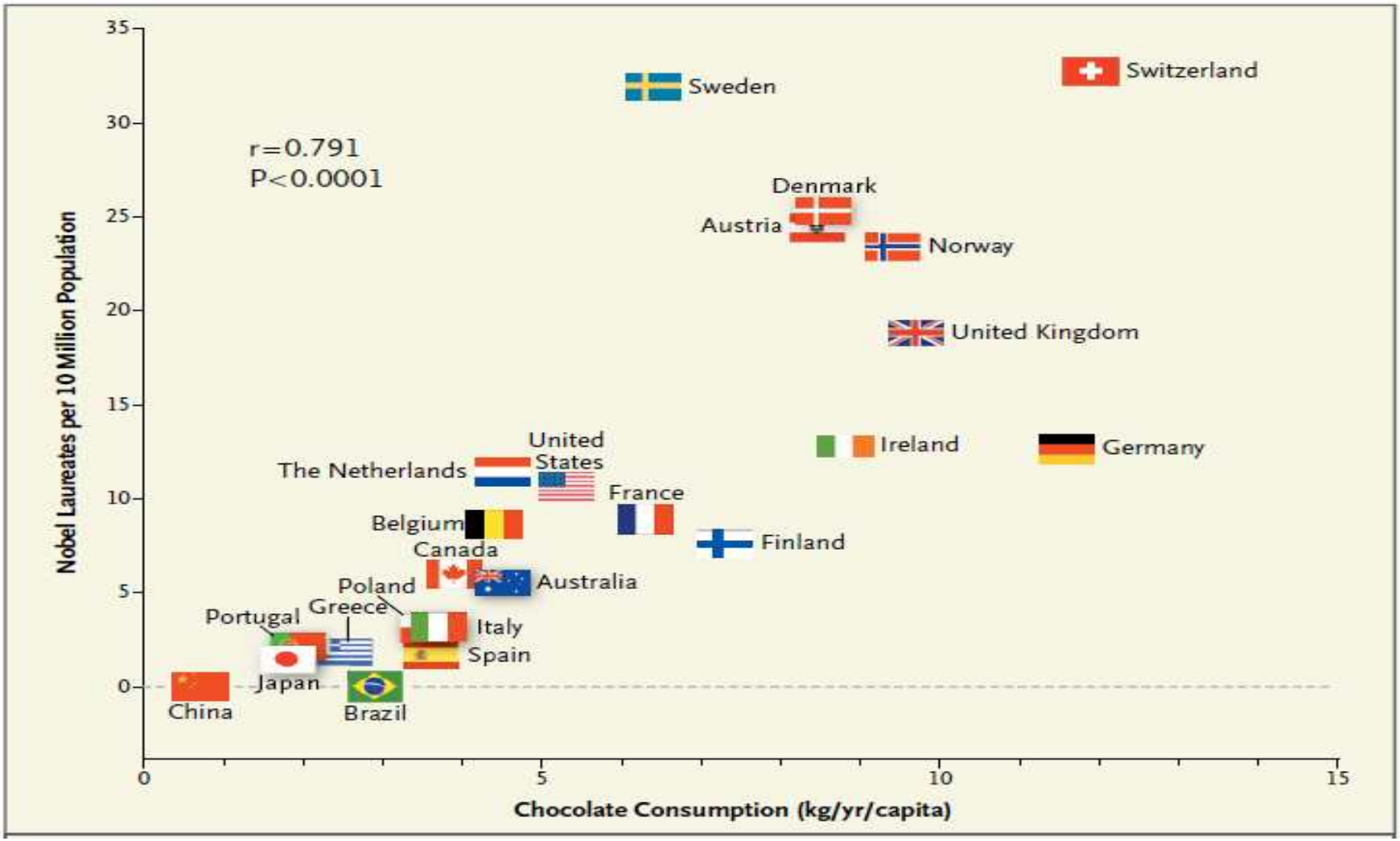
# 인과성 추론 강화를 위한 통계분석 방법들

2019. 8. 17.

연세대학교 원주의과대학

정밀의학과 · 의학통계학과

강 대 용




Franz H. Messerli, ***N Engl J Med*** 2012 Oct; 367(16):1562-4.

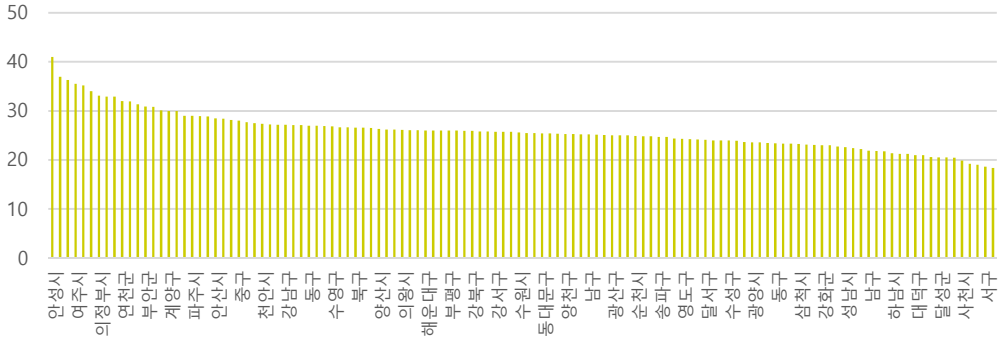
**Chocolate Consumption** — **Nobel Laureates**

**Cognitive Function**

# PM2.5 ∝ 연령표준화사망률

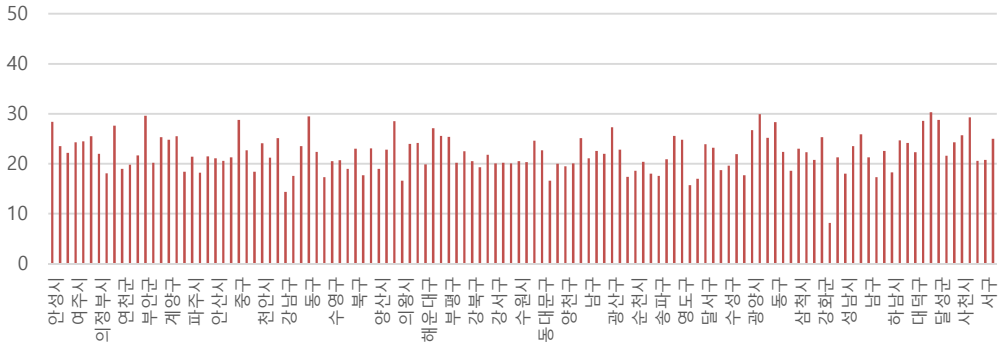
**PM2.5, WHO guideline :**
- 10 μg/m$^3$ / annual mean
- 25 μg/m$^3$ / 24-hour mean

PM2.5 연평균 농도 / 130 측정소 (2016, 통계청)

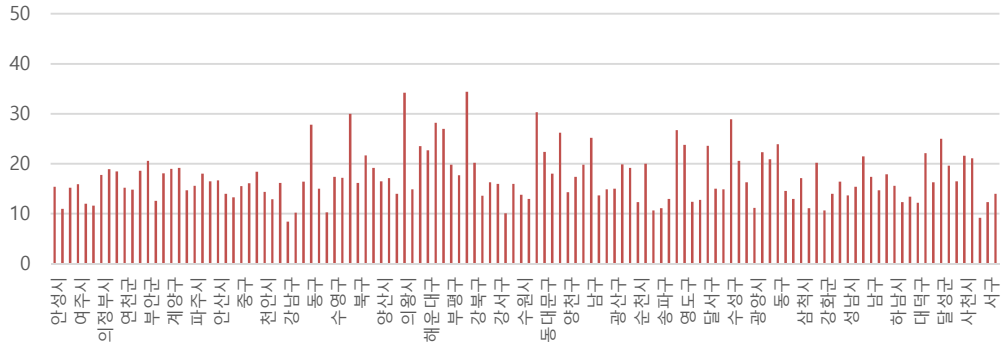



생태학적 연구, GIS 접근 한계 !!
개인의 특성, 생활패턴, 이사 지역/횟수, 주거/직장 환경
"개인/집단/지역 맞춤형 노출평가"

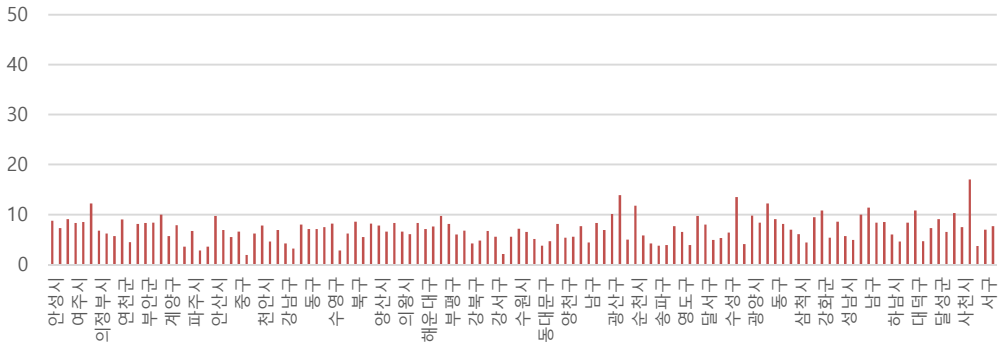
기관, 기관지 및 폐의 악성 신생물 (C33-C34) (2016, 통계청)



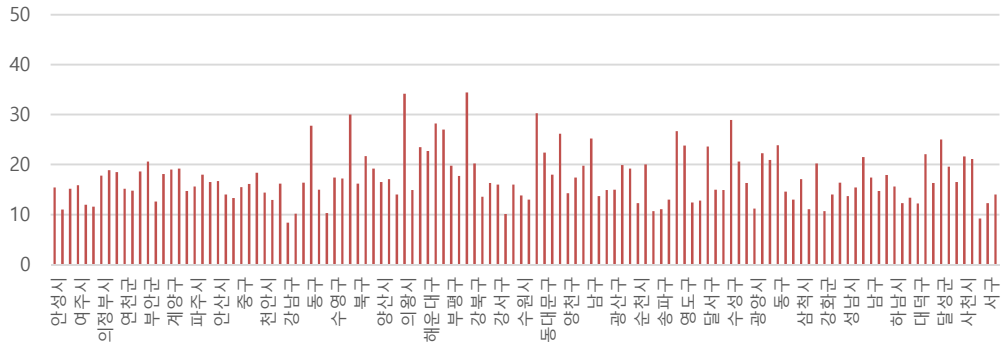

허혈성 심장 질환 (I20-I25) (2016, 통계청)




만성 하기도 질환 (J40-J47) (2016, 통계청)




뇌혈관 질환 (I60-I69) (2016, 통계청)



# Nursing Research

# Random Error / Systematic Error

## Bias

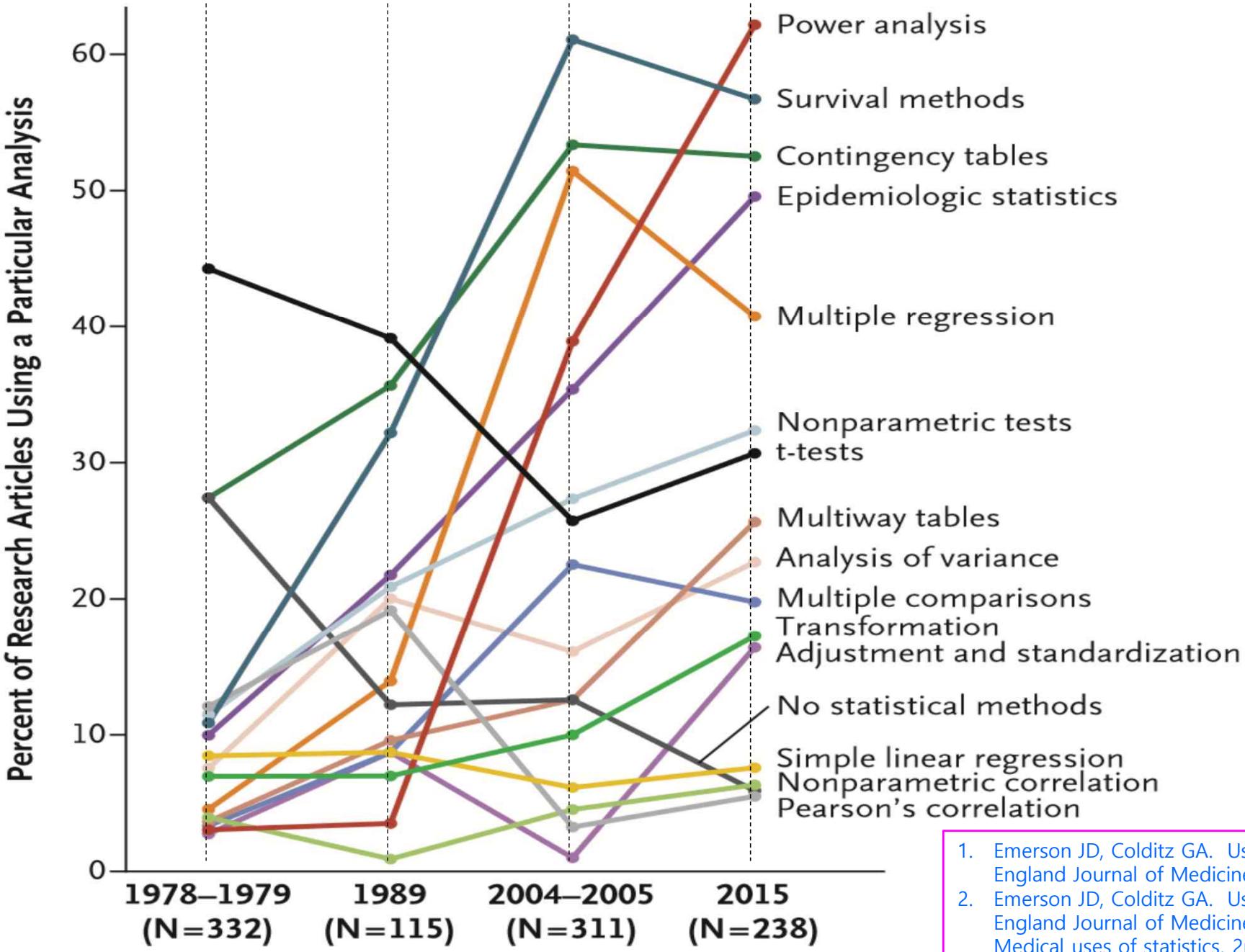
편향(偏向) 편의(偏倚)

뒤틀림 비뚤림 치우침

# 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<br>생존형 자료의 회귀분석 | Life table, Kaplan-Meier method<br>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.

# Study Designs

Randomization ?
Intervention ?

## Observational study

Unit of study

Descriptive study
Analytical study

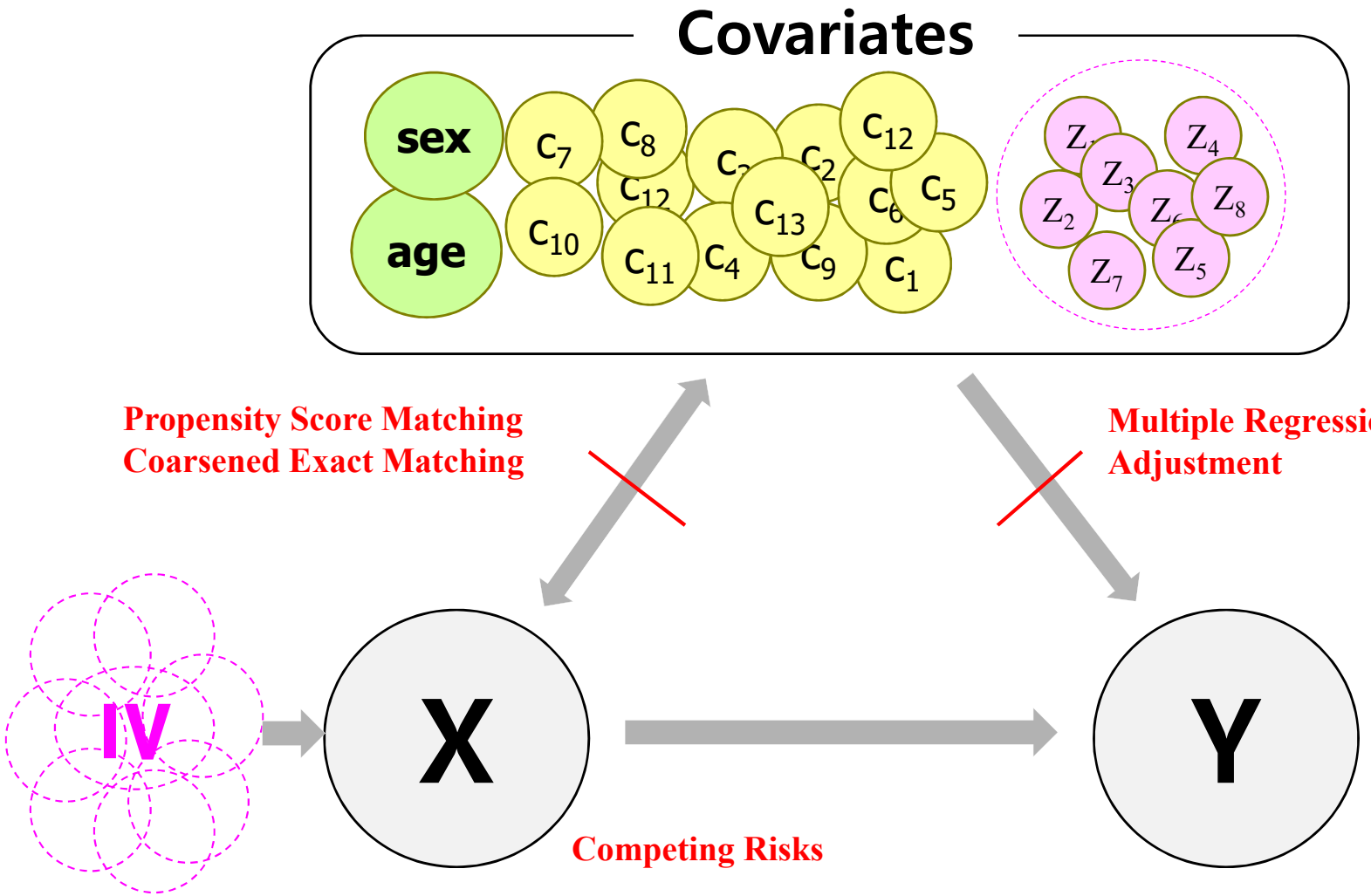
Hypothesis ?

- 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

## Experimental study, Quasi-experimental study

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

# 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

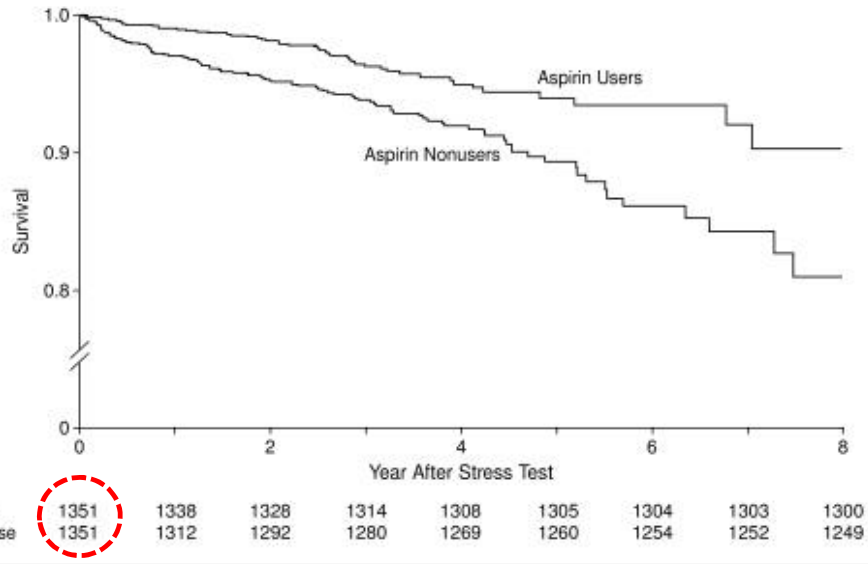



**Table 4.** Cox Proportional Hazards Analyses of Aspirin Use and Mortality Among Propensity-Matched Patients (n = 2702)*

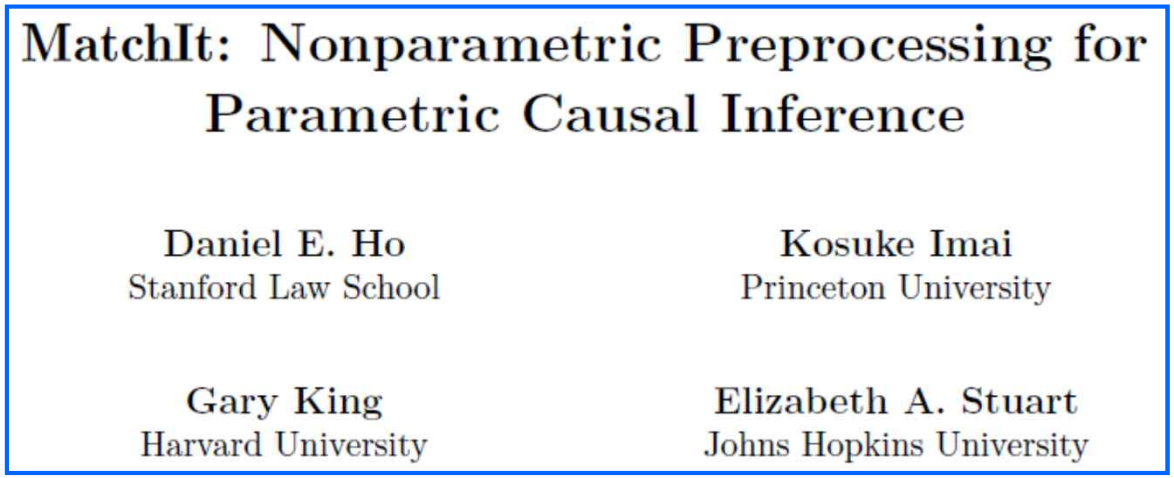
| Model | Hazard Ratio (95% CI) | *P* Value |
|---|---|---|
| Unadjusted | 0.53 (0.38-0.74) | .002 |
| Adjusted for propensity | 0.53 (0.38-0.74) | <.001 |
| Adjusted for propensity and selected variables† | 0.59 (0.42-0.83) | .002 |
| Adjusted for propensity and all covariates‡ | 0.56 (0.40-0.78) |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

## Propensity Score Computational Statistical Packages

- MatchIt in R (Ho, Imai, King, and Stuart, 2007)
- PSMATCH2 algorithm in STATA (Leuven & Sianesi, 2004)
- %PSMatching "GREEDY" Macro in SAS (D'Agostino, 1998)



MatchIt: Nonparametric Preprocessing for Parametric Causal Inference


Daniel E. Ho
Stanford Law School

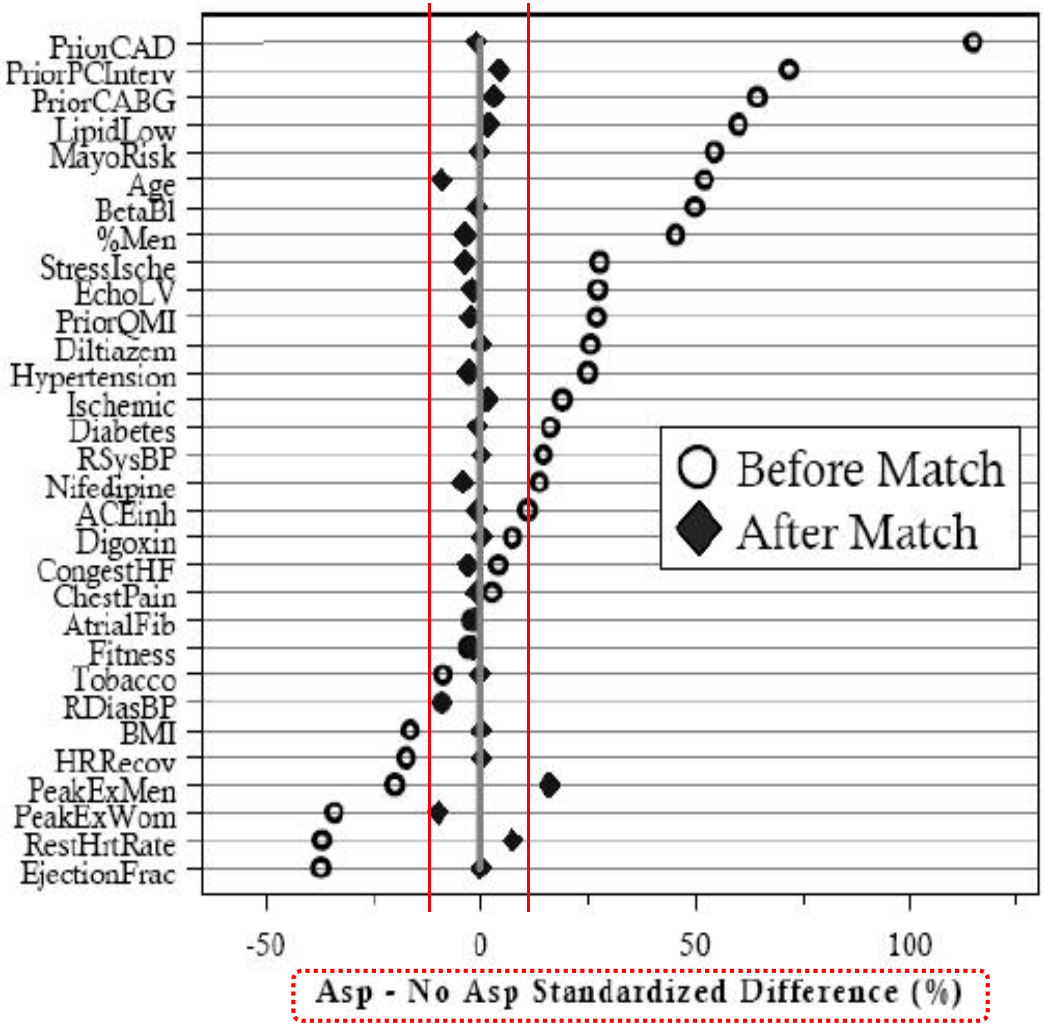
Kosuke Imai
Princeton University

Gary King
Harvard University

Elizabeth A. Stuart
Johns Hopkins University



**Journal of Statistical Software** May 2011, Volume 42, Issue 8. http://www.jstatsoft.org/




Covariate Balance for Aspirin Study
**(Love, 2004)**

$$C \perp X / \pi(C)$$

➔ greater than **10** percent
Represents meaningful imbalance

**for continuous variables**

$$d = \frac{100 \cdot (\bar{x}_{\text{treatment}} - \bar{x}_{\text{control}})}{\sqrt{(s^2_{\text{treatment}} + s^2_{\text{control}})/2}}$$

**for categorical variables**

$$d = \frac{100 \cdot (p_T - p_C)}{\sqrt{(p_T(1-p_T) + p_C(1-p_C))/2}}$$

**Some ways of assessing balance (Rubin, 2001) :**

**- The standardized difference in the mean propensity score in the two groups should be near zero** ($d < .25$)

**- The ratio of the variance of the propensity score in the two groups should be near 1 one, (preferably between 0.80 and 1.25)**

**Warning**

| |
|---|
| |
| |
| No warnings in estimation or matching procedure |

**Sample Sizes**

| | Control | Treated |
|---|---|---|
| All | 927 | 244 |
| Matched | 244 | 244 |
| Unmatched | 683 | 0 |
| Discarded | 0 | 0 |

← 비교군과 대조군이
각각 244명으로 1:1 매칭됨

**Overall balance test (Hansen & Bowers, 2010)**

| | chisquare | df | p.value |
|---|---|---|---|
| Overall | 2.598 | 6.000 | .857 |

**Hansen & Bowers (2010), $d^2$**
실험군/대조군 간의 전반적인 불균형 파악을 위한 검정통계량이다. 공변량과 공변량의 선형결합이 매칭 후에도 unbalance 한지를 평가하는 것으로 귀무가설이 기각되지 않으면 두 집단 간의 구조가 유사하여 매칭이 잘 된다고 할 수 있다.
본 통계량은 개체 N수에 민감하다. 실제 유의한데 N수가 너무 적어서 유의하게 나오지 않을 수도 있으므로 다른 통계량들과 함께 참조하여 의사결정을 해야 한다.

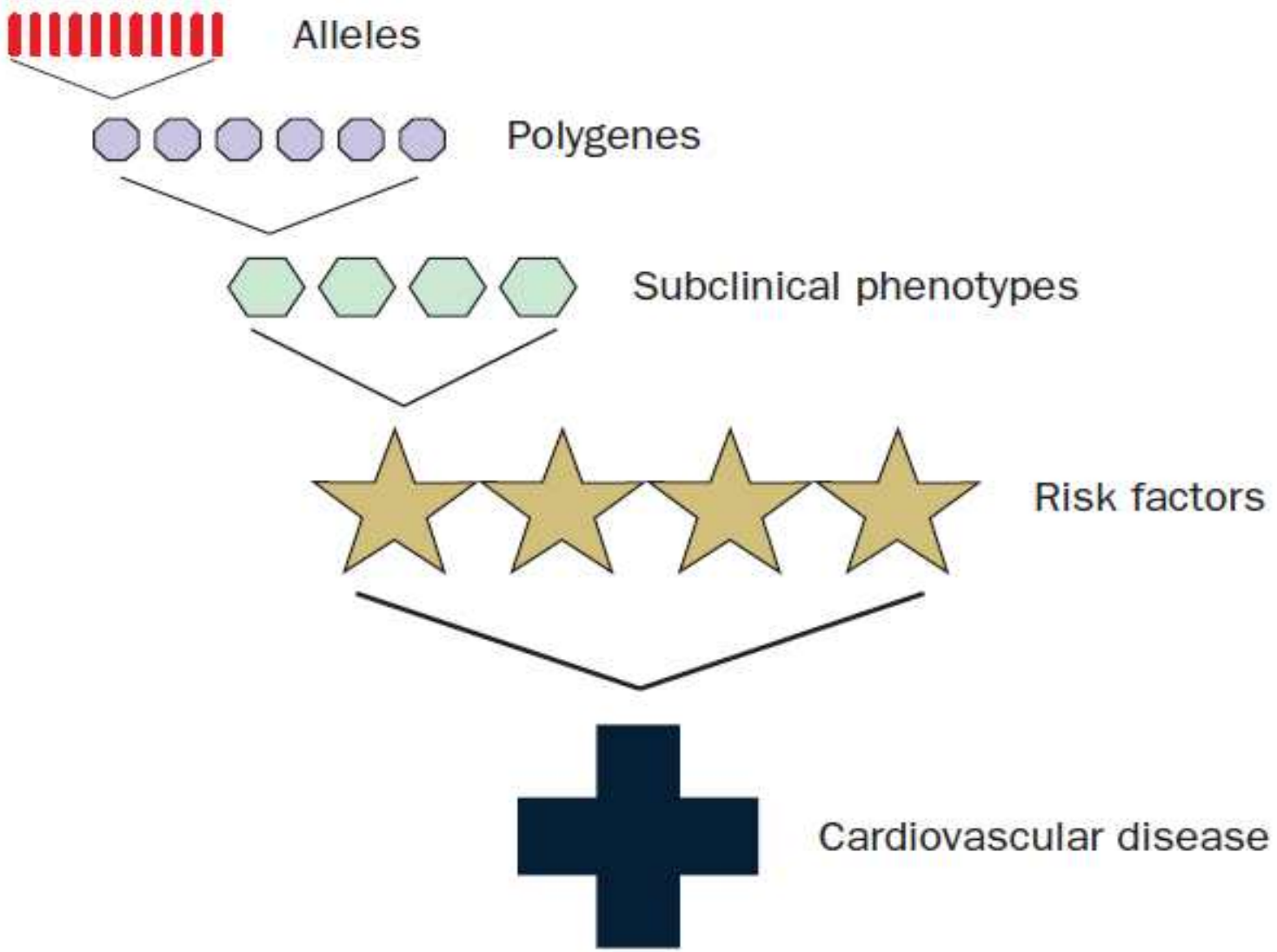
**Relative multivariate imbalance L1 (Iacus, King, & Porro, 2010)**

| | Before matching | After matching |
|---|---|---|
| Multivariate imbalance measure L1 | .493 | .250 |

**Iacus, King, Porro (2010), L1**
매칭 균형에 대한 Multivariate Imbalance Measure이다. 이 지표의 해석은
**0**에 가까울수록 두 군의 분포 균형이 좋음을 뜻하고, **1**에 가까울수록 점점 불균형이 강화된다고 할 수 있다.

**Summary of unbalanced covariates (|d| > .25)**

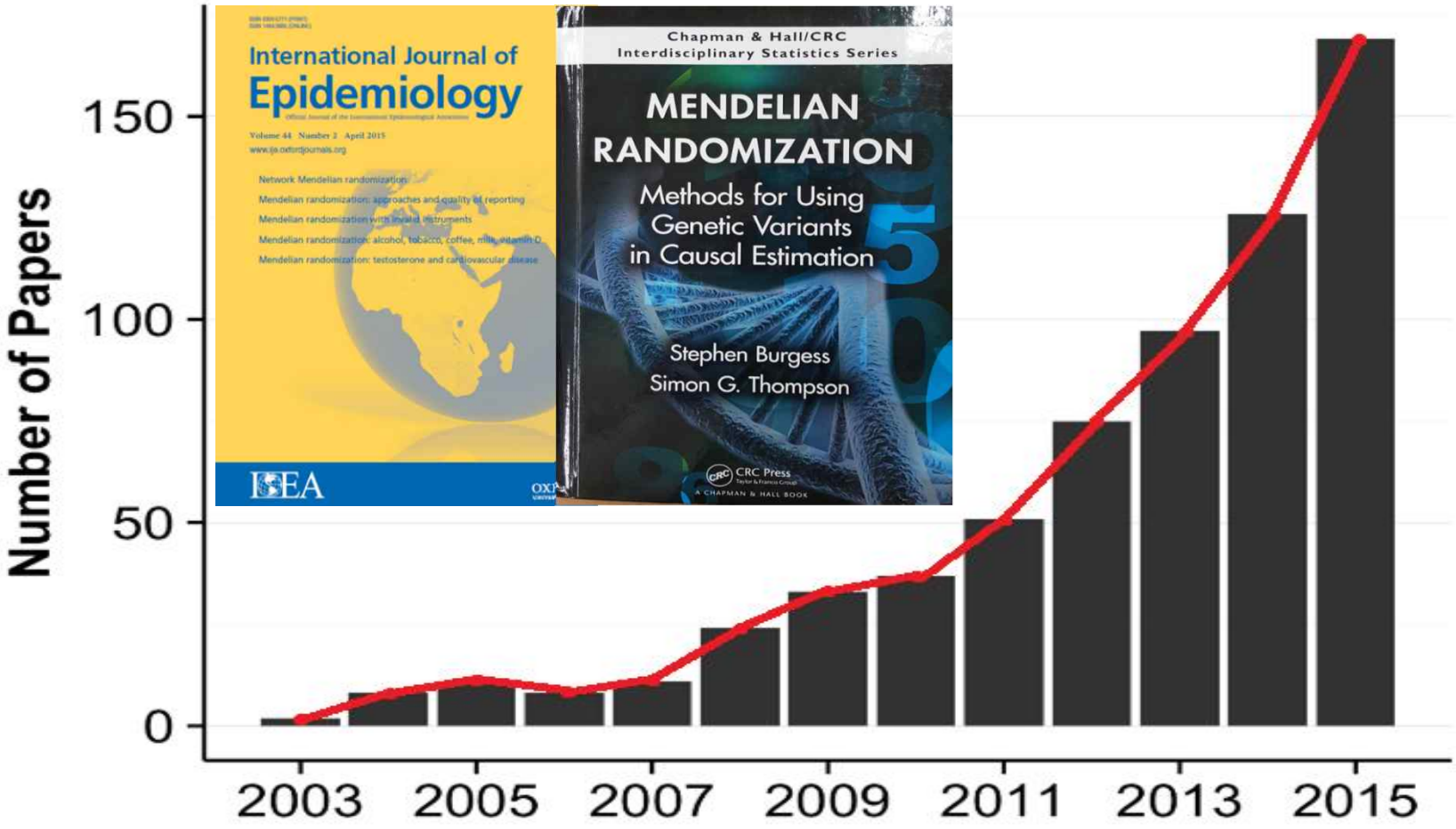
| |
|---|
| No covariate exhibits a large imbalance (\|d\| > .25). |




**Hierarchy in genetics of cardiovascular disease**

Source: Harrap et al., *Lancet* 2003;361:2149-51.

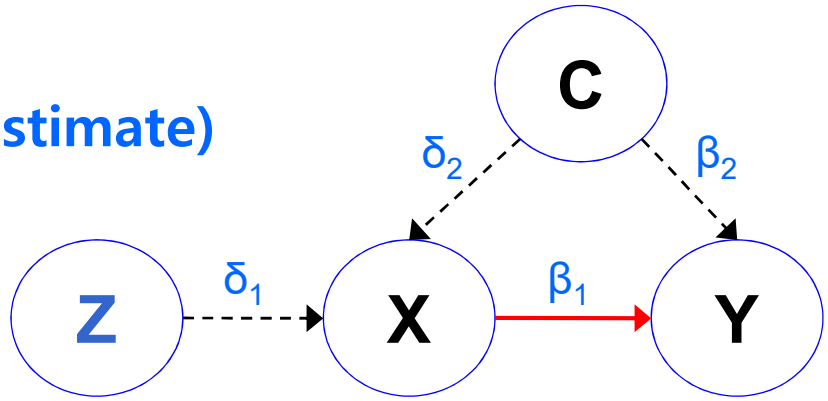
# Number of Publication Using MR approach (2003~2015)




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.)

# Estimation (2 Stage Least Squares Estimate)

$$Y = \beta_0 + \beta_1 X + \beta_2 C + \varepsilon_Y$$
$$X = \delta_0 + \delta_1 Z + \delta_2 C + \varepsilon_X$$




**Stage 1 :** regress of the X on the Z

$$\hat{X} = Z(Z'Z)^{-1}Z'X = P_Z X, \qquad P_Z = Z(Z'Z)^{-1}Z' \quad and \quad P_Z^2 = P_Z$$

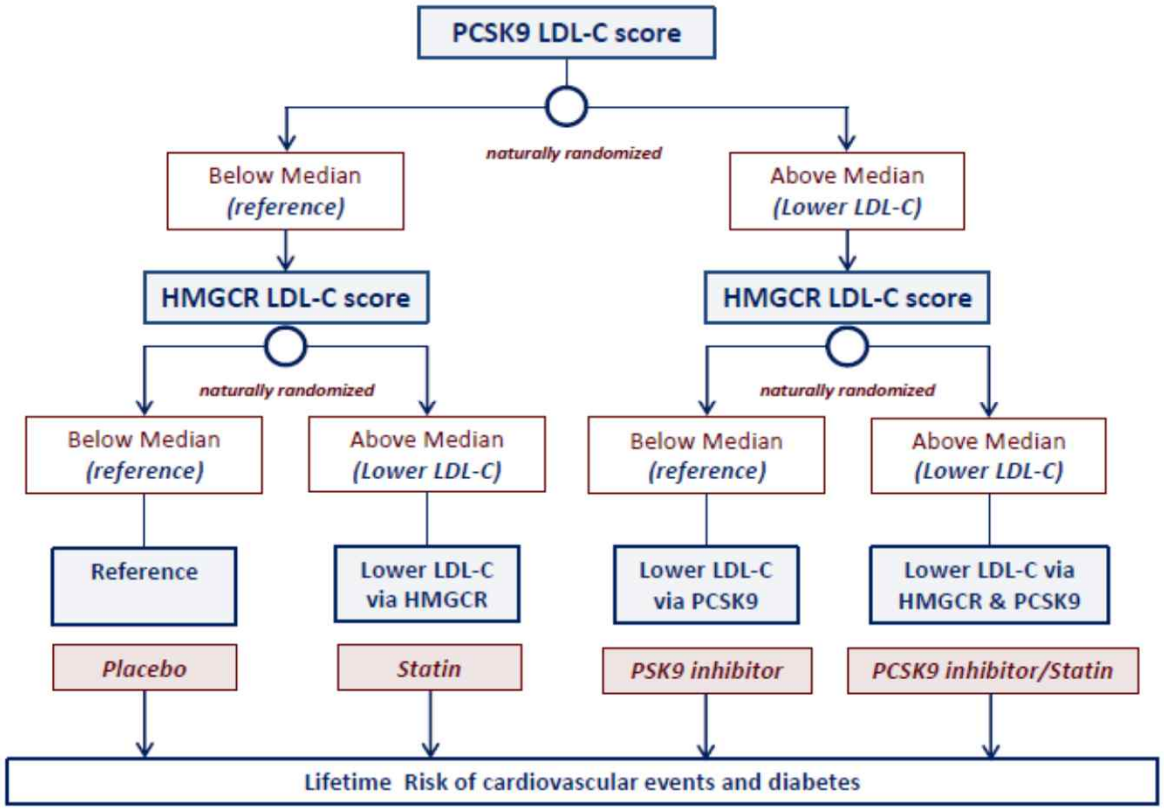
**Stage 2 :** regress of Y on the fitted X-values from stage 1.
i.e. only the variation in X that is explained by Z is used in stage 2.

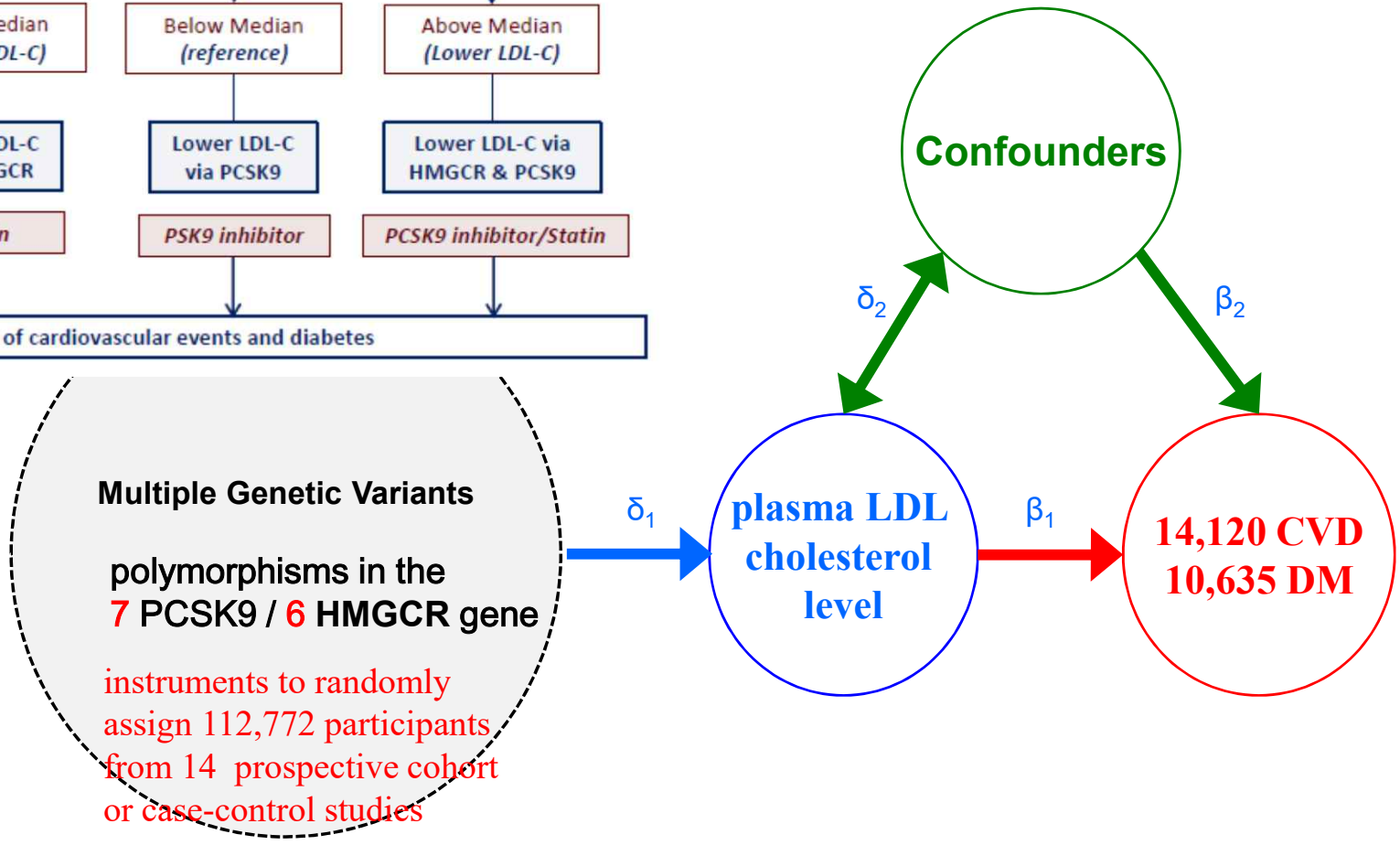
$$\hat{\beta}_{IV} = [(P_Z X)'(P_Z X)]^{-1}(P_Z X)'Y$$
$$= (X'P_Z'P_Z X)^{-1}X'P_Z'Y$$
$$= (X'P_Z X)^{-1}X'P_Z Y$$

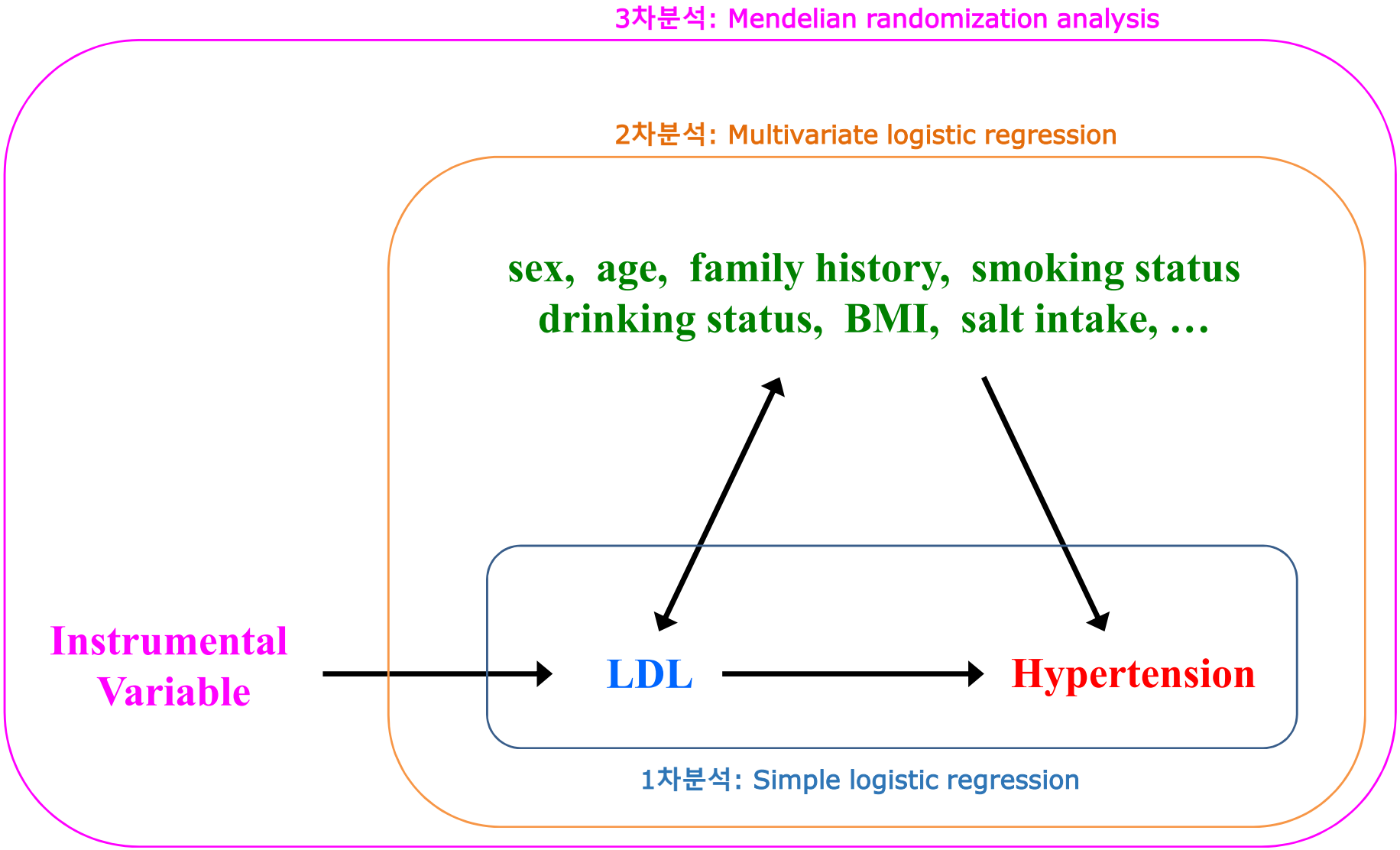
```
/*** IV analysis in SAS ***/
proc syslin data=in 2SLS;
    endogenous x;
    instruments z;
    model y = x;
run;
/* 2SLS can be replaced by
LIML or FIML as appropriates */
```

# Variation in *PCSK9* and *HMGCR* and Risk of Cardiovascular Disease and Diabetes

Ference et al., ***N Engl J Med*** 2016;375:2144-53.





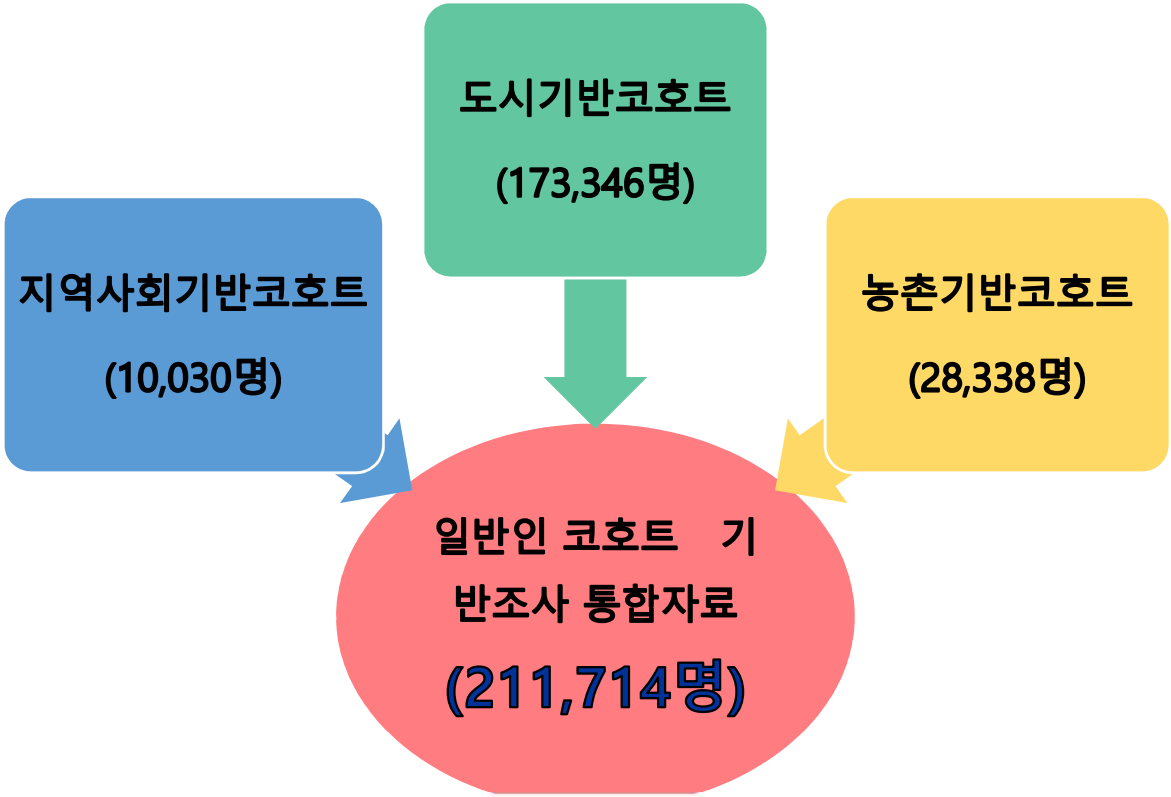

**KoGES (AIE chip, K-chip) :**

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

# KoGES 일반인 코호트 기반조사(2001- ) 통합자료 구성

한국인유전체역학조사사업 (**Ko**rean **G**enome and **E**pidemiology **S**tudy)




- 코호트 대상자수를 크게하고 폭로요인의 변이 수준을 넓혀 결과 변수와의 연관성 규명에 보다 큰 **검정력**을 갖게함

- 수집된 코호트의 기반조사 자료 중 공통으로 조사된 항목을 중심으로 통합기준을 마련하여 **총 통합변수는 201개**

## 한국인 맞춤형 만성질환 유전체 연구를 한눈에

– 한국인 만성질환 유전적 원인 규명을 위한 '한국인칩사업' 백서 발간 –

| 일정 | 공개 버전 | 샘플 수 | 누계 |
|---|---|---|---|
| '17.01 | KCHIP data Release v1 | 8,000 | 8,000 |
| '17.09 | KCHIP data Release v2 | 27,000 | 35,000 |
| '18.07(예정) | KCHIP data Release v3 | 39,000 | 74,000 |

# 반복측정 - 추적조사 자료분석

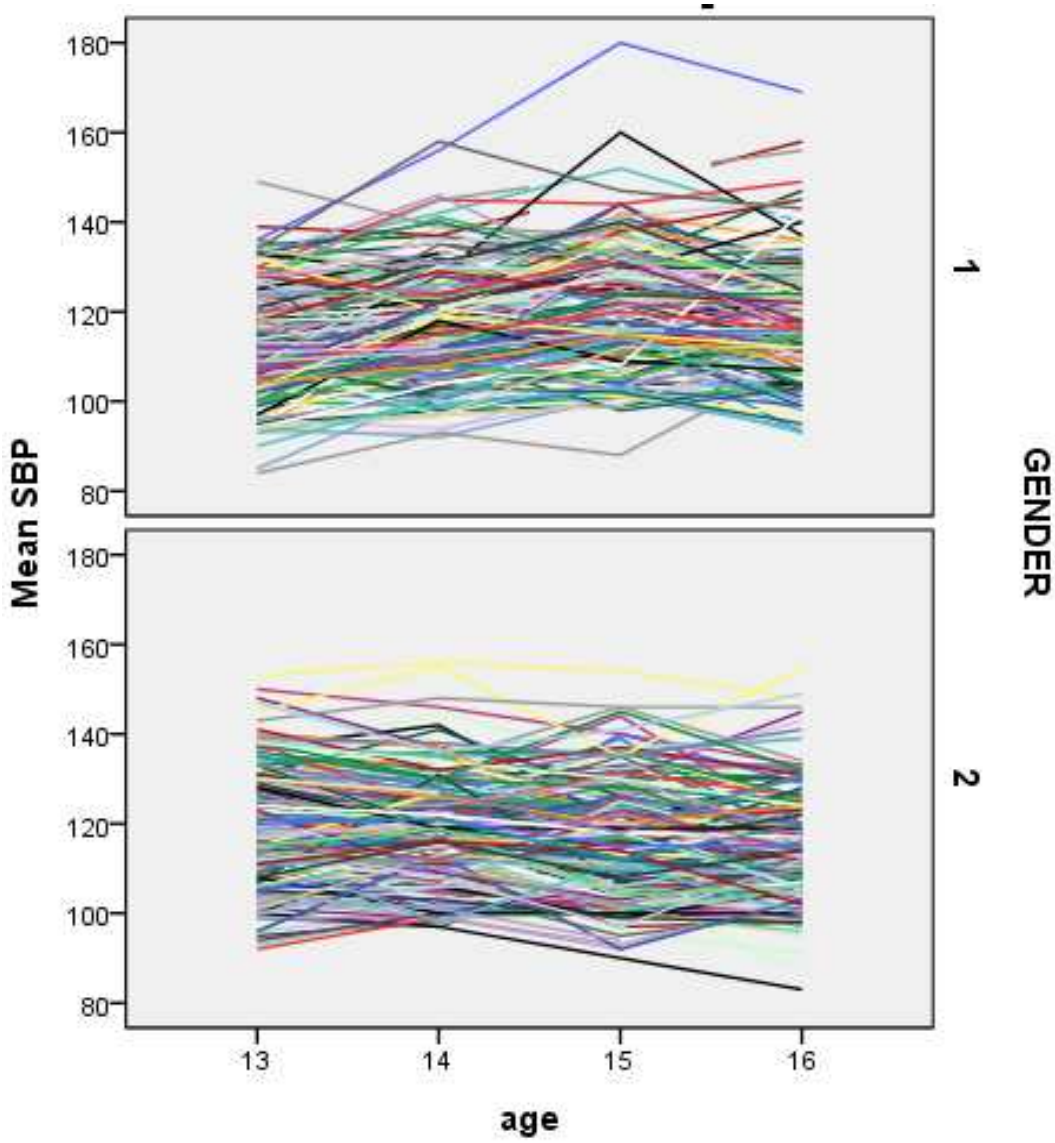
| 종속변수 | 독립변수 | 통계분석법 |
|---|---|---|
| 연속형 | 범주형(3개 이상) | ANOVA<br>**Repeated Measures ANOVA** |
| 연속형 | 연속형 + 범주형 | 회귀분석<br>General LM / **LMM**<br>Linear Mixed Model |
| 이분형 | 연속형 + 범주형 | 로지스틱 회귀분석<br>HGLM / **GEE**<br>Generalized Estimating Equations |
| 생존시간 | 연속형 + 범주형 | **Cox PH 모형**<br>Frailty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

Gender별 각 age에서의
SBP Descriptive statistics

SBP

| GENDER | age | N | Mean | Std. Deviation | Minimum | Maximum |
|---|---|---|---|---|---|---|
| 1 | 13 | 252 | 111.93 | 11.168 | 84 | 149 |
| | 14 | 241 | 116.20 | 11.302 | 92 | 158 |
| | 15 | 206 | 120.63 | 12.125 | 88 | 180 |
| | 16 | 206 | 118.81 | 11.831 | 93 | 169 |
| | Total | 905 | 116.61 | 12.023 | 84 | 180 |
| 2 | 13 | 256 | 117.73 | 11.521 | 92 | 153 |
| | 14 | 253 | 117.66 | 10.831 | 97 | 156 |
| | 15 | 240 | 114.92 | 11.792 | 90 | 154 |
| | 16 | 235 | 113.77 | 11.140 | 83 | 155 |
| | Total | 984 | 116.08 | 11.437 | 83 | 156 |
| Total | 13 | 508 | 114.86 | 11.701 | 84 | 153 |
| | 14 | 494 | 116.95 | 11.076 | 92 | 158 |
| | 15 | 446 | 117.56 | 12.269 | 88 | 180 |
| | 16 | 441 | 116.12 | 11.727 | 83 | 169 |
| | Total | 1889 | 116.34 | 11.721 | 83 | 180 |

Gender별 spaghetti plot

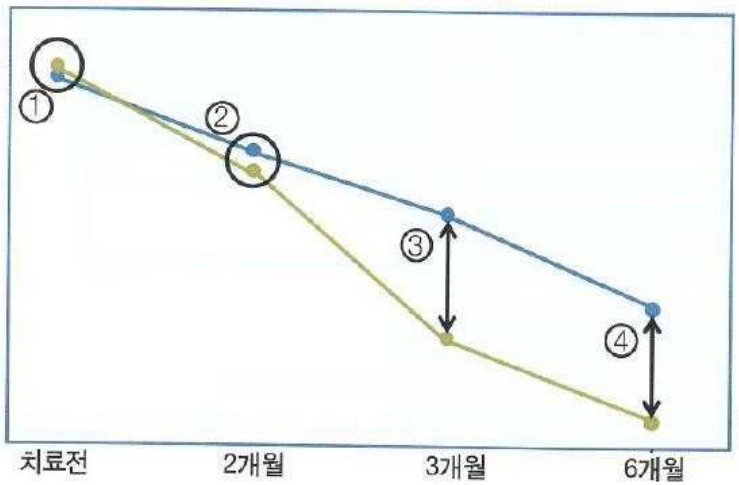


# 반복측정 분산분석 에서의 세 가지 검정

| | | |
|---|---|---|
| 군 | 두 군 간에 차이가 있는가 ? | 개체**간** 검정 |
| 시간 | 종속변수가 시간에 따라 변하는가 ? | 개체내 검정 |
| 시간x군 | 시간에 따른 변화는 군 간 차이가 있는가 ? | |
| 사후검정 | 어느 시점에서 군 간에 차이가 나는가 ? | |

- 반복측정 분산분석의 결과를 해석하는 방법

➔ 시간과 군의 '교호작용'이 통계적으로 유의한지를 검정하는 것이 최우선적 목적




보정된 유의수준=0.05/4=0.0125
by Bonferroni's method

# 반복측정 분산분석의 가장 큰 단점

## Repeated Measures ANOVA

- 결측치가 하나도 없는 완전무결한 자료만을 대상으로 함.
- 실제 임상 연구에서는 환자가 제때에 방문하지 않는 경우가 많음.

- 아래와 같은 자료의 경우 정보 손실이 많음. (6명이 빠짐)

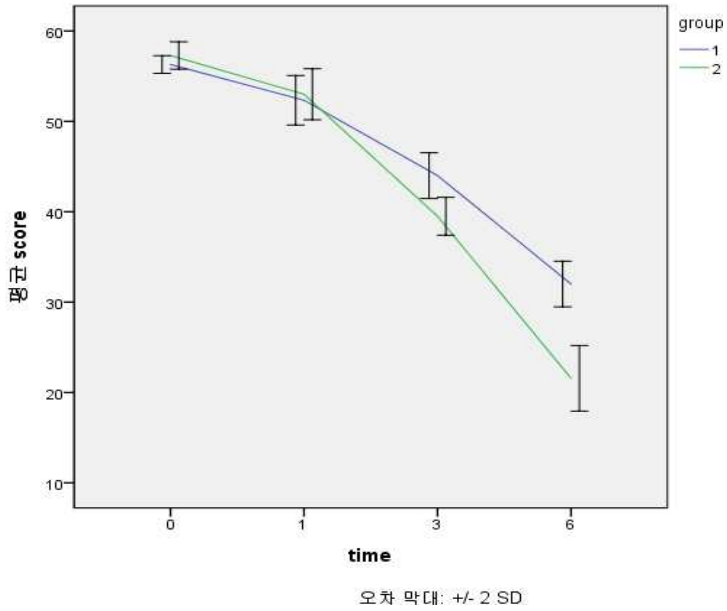
| id | group | sex | baseline | month1 | month3 | month6 |
|---|---|---|---|---|---|---|
| 1 | 1 | F | 60 | 결측치 | 25 | 16 |
| 2 | 1 | F | 52 | 38 | 23 | 12 |
| 3 | 1 | F | 62 | 36 | 22 | 14 |
| 4 | 1 | F | 58 | 34 | 21 | 13 |
| 5 | 1 | M | 65 | 34 | 28 | 18 |
| 6 | 1 | M | 58 | 42 | 26 | 결측치 |
| 7 | 1 | M | 53 | 38 | 결측치 | 21 |
| 8 | 2 | F | 55 | 42 | 33 | 22 |
| 9 | 2 | M | 55 | 54 | 46 | 26 |
| 10 | 2 | M | 60 | 55 | 46 | 23 |
| 11 | 2 | M | 63 | 45 | 결측치 | 25 |
| 12 | 2 | M | 52 | 결측치 | 35 | 22 |
| 13 | 2 | F | 61 | 38 | 32 | 18 |
| 14 | 2 | F | 58 | 결측치 | 39 | 21 |

- 극복 방법
  - 혼합모형 LMM
  - 일반화추정방정식 GEE

- 분석의 대상
  - 반복측정자료: 개체
  - 혼합모형 & 일반화추정방정식: 개별 관측치

# 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<br>time: <0.001<br>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) | |




- 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

# 실습 (Time을 '연속형' 으로 고려)

Table & Figure

| | B(SE) | P-value |
|---|---|---|
| intercept | 57.600(0.363) | <0.001 |
| group | | |
| 1 | -1.617(0.427) | 0.001 |
| 2 | Ref(0) | |
| sex | | |
| 1(M) | 0.592(0.334) | 0.105 |
| 2(F) | Ref(0) | |
| Time | -6.057(0.077) | <0.001 |
| group=1 x time | 1.998(0.111) | <0.001 |
| group=2 x time | Ref(0) | |

-LMM으로 분석
-시간의 흐름에 따라 두 군의 변화 패턴이 다름을 알 수 있음.
-2번 군에 비해 1번 군이 1개월 지남에 따라 중증도 점수가 1.998 더 증가함을 알 수 있음.
-즉, 1번 군에 비해 2번 군이 더 감소함을 알 수 있음.

EX>
* 모형식
=57.600-
1.617x[group=1]+0x[group=2]+0.592x[sex=1]+0x[sex=2]-
6.057xtime+1.998x[group=1][time]+0x[group=2]xtime

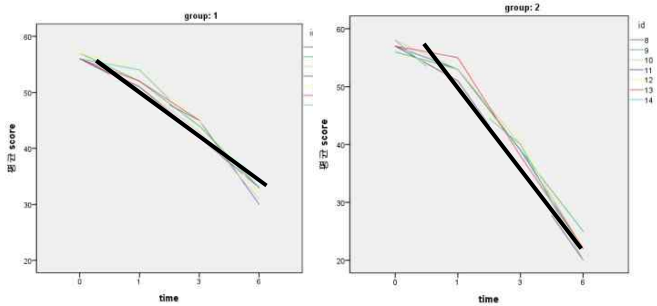
- 여자일 때 group=1의 중증도 점수
=57.600 -1.617+6.057xtime+1.998xtime
=(57.600-1.617)+(-6.057+1.998)xtime

- 여자일 때 group=2의 중증도 점수
=57.600 -6.057xtime

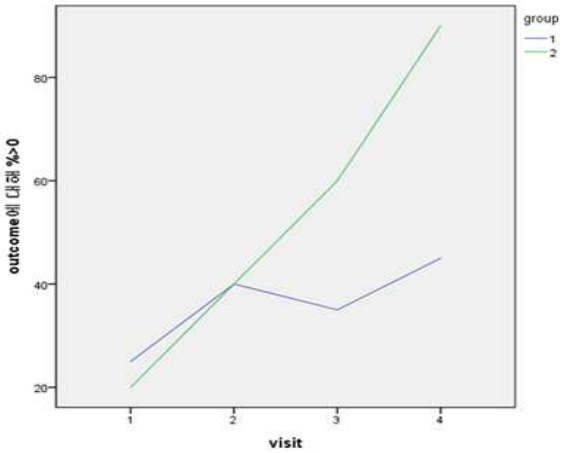
➔ group=2일 때에는 1개월 지남에 따라 6.057만큼 중증도 점수가 감소하는 반면, group=1일 때에는 1개월 지남에 따라 4.059만큼 중증도 점수가 감소. 즉 기울기의 차이가 1.998로 p-value<0.001로 유의한 차이가 존재함.

| | Group=1 | Group=2 | P-value |
|---|---|---|---|
| | Slope(SE) | Slope(SE) | P-value |
| score | -4.059(0.080) | -6.057(0.077) | <0.001 |

그래프



# (Ratio기반) GEE 결과




## GEE 분석

시간의 흐름에 따라 두 군의 변화 패턴이 다름을 알 수 있음.
1번 군에 비해 2번 군이 시간이 지남에 따라 호전될 가능성이 높음.
특히 처음시점에 비해 6개월 시점에 group1에 비해 group2가 호전될 가능성이 15.683배 높게 나타남.

| | OR (95% CI) | p-value |
|---|---|---|
| intercept | 1.854(0.273-12.581) | 0.527 |
| group | | |
| 1 | Ref(1) | |
| 2 | 0.840(0.204-3.455) | 0.809 |
| Sex | | |
| 1(M) | Ref(1) | |
| 2(F) | 0.682(0.332-1.404) | 0.299 |
| Visit | | |
| 1 | Ref(1) | |
| 2 | 2.049(0.512-8.191) | 0.310 |
| 3 | 1.642(0.634-4.252) | 0.307 |
| 4 | 2.536(0.880-7.314) | 0.085 |
| age | 0.951(0.904-0.999) | 0.047 |
| Group x visit | | |
| group=1 x visit1 | Ref(1) | |
| group=1 x visit2 | Ref(1) | |
| group=1 x visit3 | Ref(1) | |
| group=1 x visit4 | Ref(1) | |
| group=1 x visit1 | Ref(1) | |
| group=1 x visit2 | 1.340(0.189-9.504) | 0.769 |
| group=1 x visit3 | 3.875(0.648-23.171) | 0.138 |
| group=1 x visit4 | 15.683(2.056-119.658) | 0.008 |

# 일반적 통계분석 방법과 생존분석의 비교

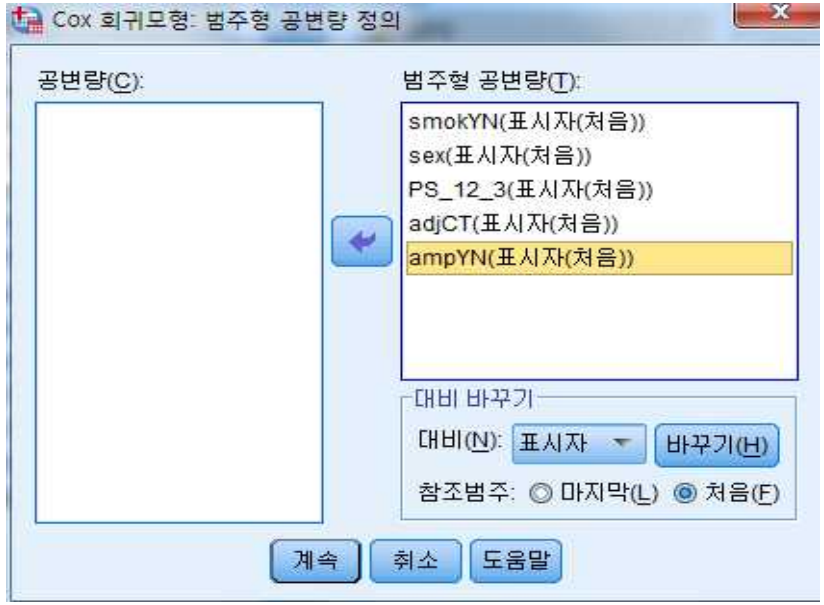
| | 일반적인 방법 | 생존분석 |
|---|---|---|
| **Demographic Graph** | - Mean±SD (bar graph)<br>- Median (min, max) or Median (Q1, Q3)<br>- Box plot | - Kaplan-Meier method (1958) |
| **1:1의 관계** | - Independent two sample t-test, ANOVA<br>- Mann-Whitney U test, Kruskal-Wallis test<br>- Chi-square test (Fisher's exact test) | - Log-rank test (Mantel-Haenszel, 1959) |
| **1:N의 관계** | - Linear regression<br>- Logistic regression | - Cox's PH regression (1972) |
| **Predictive ability** | - ROC curve, AUC | - Harrell's C, tdAUC, iAUC |

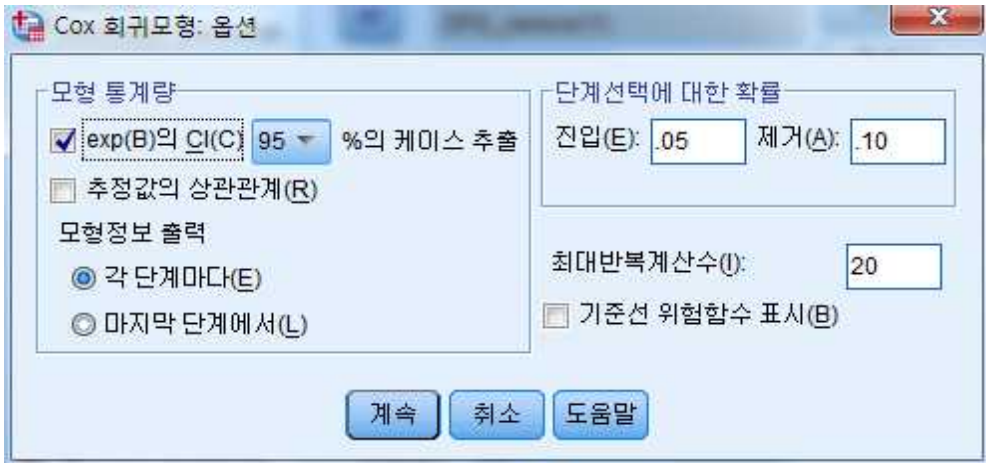
# Cox Proportional Hazards Model (1972)

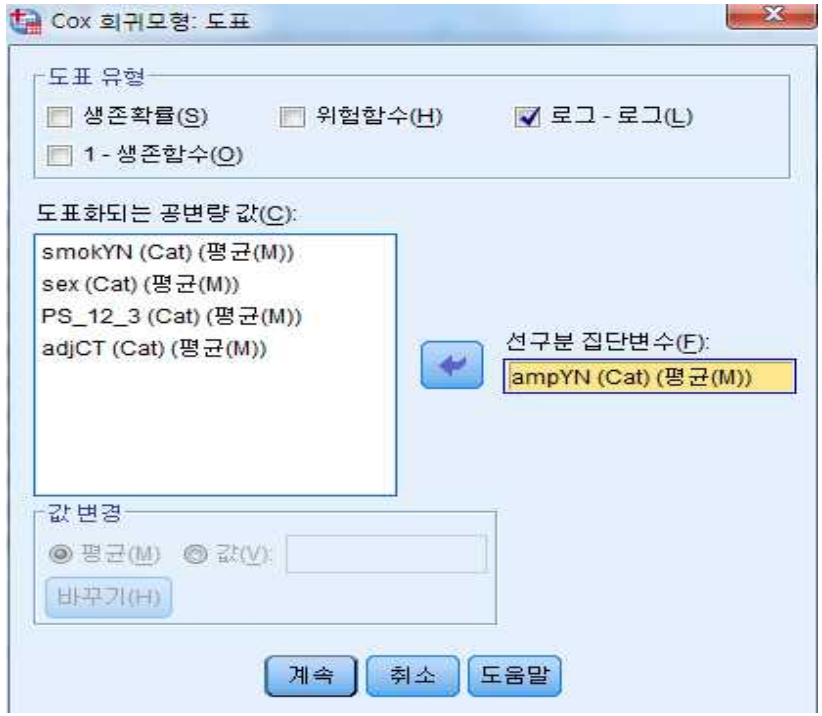
- 생존곡선에 영향을 주는 '위험요인'과의 관련성을 모형화하는 것이 목적
- 다른 변수들의 효과를 보정한 후 치료효과를 볼 수 있는 대표적인 통계모형

| | |
|---|---|
| Baseline | $h_0(t)$ : 모든 독립변수가 0일 때의 위험함수 |
| Model<br>Semi-parametric model | t시점에서 p개의 독립변수가 $x_1, x_{2,\ldots,} x_p$ 일 때의 위험함수<br>$h(t) = h_0(t)\exp(\beta_1 x_1 + \ldots + \beta_k x_k)$<br>$\log h(t,x) = \log h_0(t) + \beta_1 x_1 + \ldots + \beta_k x_k$ |
| Assumption | j번째 환자와 j번째 환자의 위험비가 시간과 무관하게 상수가 됨<br>$h_i(t)/h_j(t) = \exp(\beta_1(x_{i1} - x_{j1}) + \ldots + \beta_k(x_{ik} - x_{jk}))$ |
| Model check | 독립변수의 서로 다른 값에서 $\log(-\log S(t))$와 t는 시점에 관계없이 일정함(비례위험) |


Cox 회귀모형
ID
sex [sex]
Age [Age]
packyears [packyear...
gPY
smokStatus
PS_1_2_3
PS_12_3
adjCT
ampYN
amp123
OS [OS]
OS_censor
생존함수 [SUR_1]
흡연여부 [smokYN]
시간(I):
DFS
상태변수(S):
DFS_censor(1)
사건 정의(D)...
범주형(C)...
도표(T)...
저장(S)...
옵션(O)...
붓스트랩(B)...
블록1대상1
이전(V)
다음(N)
공변량(A):
smokYN(Cat)
adjCT(Cat)
ampYN(Cat)
>a*b(A)>
방법(M):
입력
계층변수(A):
확인
붙여넣기(P)
재설정(R)
취소
도움말


Cox 회귀모형: 범주형 공변량 정의
공변량(C):
범주형 공변량(T):
smokYN(표시자(처음))
sex(표시자(처음))
PS_12_3(표시자(처음))
adjCT(표시자(처음))
ampYN(표시자(처음))
대비 바꾸기
대비(N):
표시자
바꾸기(H)
참조범주:
마지막(L)
처음(F)
계속
취소
도움말

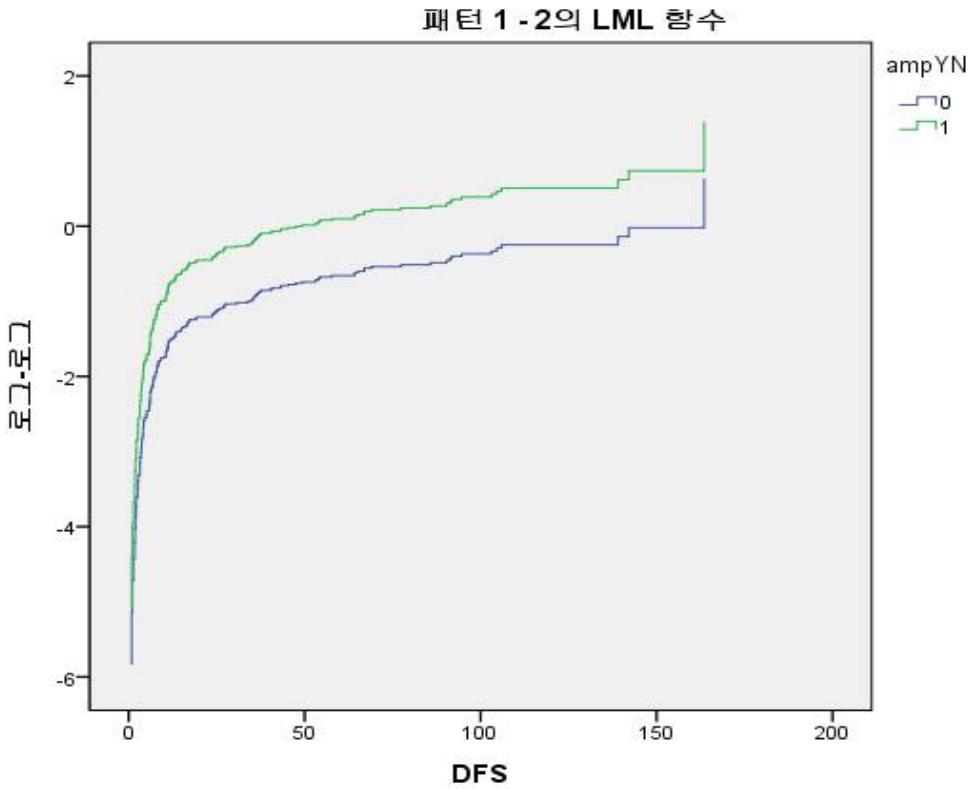

Cox 회귀모형: 옵션
모형 통계량
exp(B)의 CI(C)
95
%의 케이스 추출
추정값의 상관관계(R)
모형정보 출력
각 단계마다(E)
마지막 단계에서(L)
단계선택에 대한 확률
진입(E):
.05
제거(A):
.10
최대반복계산수(I):
20
기준선 위험함수 표시(B)
계속
취소
도움말


Cox 회귀모형: 도표
도표 유형
생존확률(S)
위험함수(H)
로그 - 로그(L)
1 - 생존함수(O)
도표화되는 공변량 값(C):
smokYN (Cat) (평균(M))
sex (Cat) (평균(M))
PS_12_3 (Cat) (평균(M))
adjCT (Cat) (평균(M))
선구분 집단변수(F):
ampYN (Cat) (평균(M))
값 변경
평균(M)
값(V):
바꾸기(H)
계속
취소
도움말

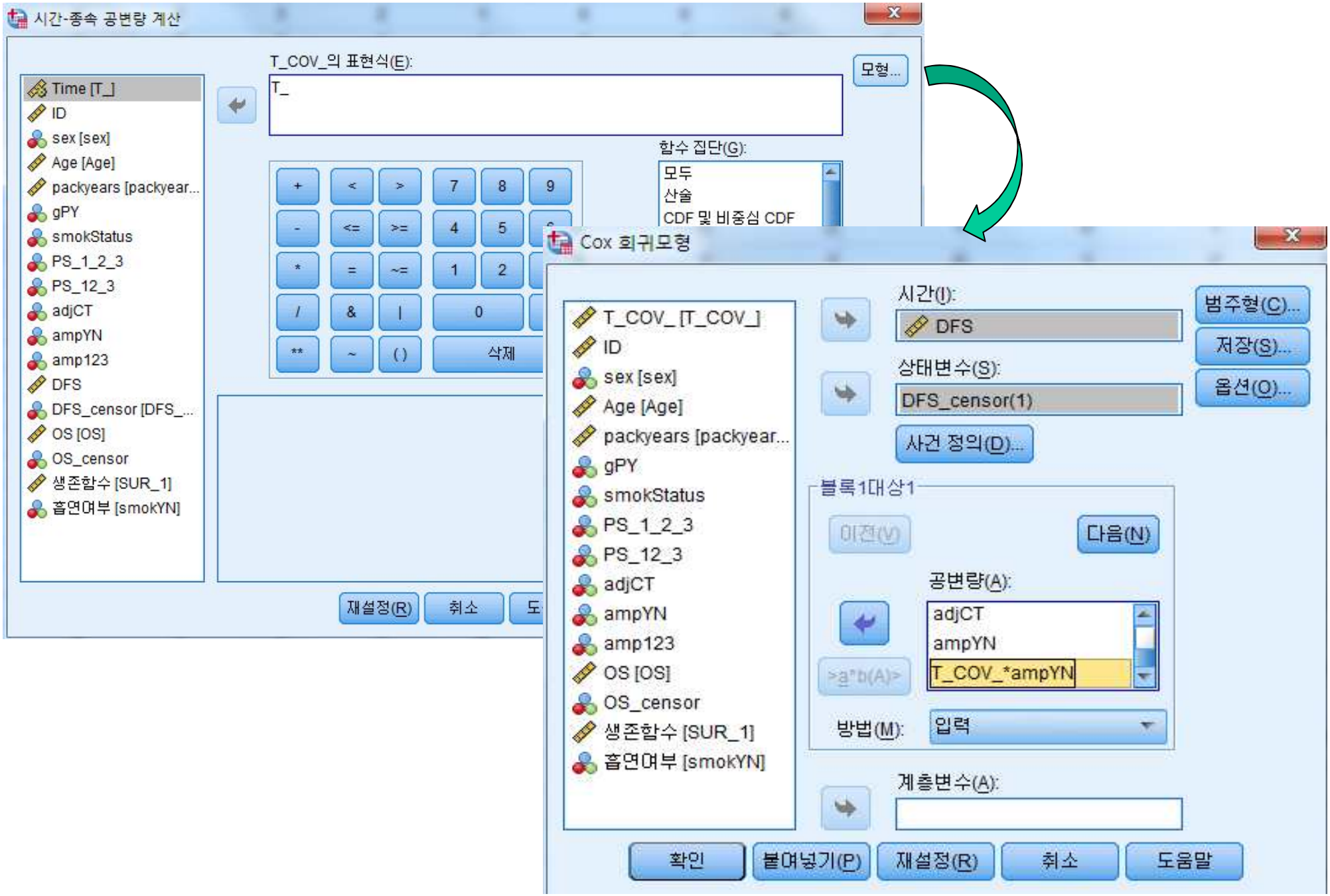
## 회귀계수의 유의성 검정

| | B | 표준오차 | Wald | 자유도 | 유의확률 | Exp(B) | Exp(B)에 대한 95.0% CI | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | 하한 | 상한 |
| sex | -.379 | .479 | .628 | 1 | .428 | .684 | .268 | 1.749 |
| PS_12_3 | .807 | .221 | 13.347 | 1 | .000 | 2.241 | 1.454 | 3.455 |
| smokYN | .471 | .329 | 2.048 | 1 | .152 | 1.602 | .840 | 3.056 |
| adjCT | .125 | .217 | .333 | 1 | .564 | 1.134 | .741 | 1.736 |
| ampYN | .756 | .216 | 12.264 | 1 | .000 | 2.131 | 1.395 | 3.254 |

HR (Hazard Ratio)




➔ sex, pathologic stage, smoking status, adjuvant chemotherapy 등의 효과를 보정한 상태에서 FGFR1 amp-에 비해 amp+인 환자가 폐암 수술 후, 재발할 위험비는 2.13배로 통계적으로 유의하게 높다 (p<.0001).


시간-종속 공변량 계산
T_COV_의 표현식(E):
T_
모형...
Time [T_]
ID
sex [sex]
Age [Age]
packyears [packyear...
gPY
smokStatus
PS_1_2_3
PS_12_3
adjCT
ampYN
amp123
DFS
DFS_censor [DFS_...
OS [OS]
OS_censor
생존함수 [SUR_1]
흡연여부 [smokYN]
함수 집단(G):
모두
산술
CDF 및 비중심 CDF
+ < > 7 8 9
- <= >= 4 5
* = ~= 1 2
/ & | 0
** ~ ( ) 삭제
재설정(R)
취소
Cox 회귀모형
T_COV_ [T_COV_]
ID
sex [sex]
Age [Age]
packyears [packyear...
gPY
smokStatus
PS_1_2_3
PS_12_3
adjCT
ampYN
amp123
OS [OS]
OS_censor
생존함수 [SUR_1]
흡연여부 [smokYN]
시간(I):
DFS
상태변수(S):
DFS_censor(1)
사건 정의(D)...
범주형(C)...
저장(S)...
옵션(O)...
블록1대상1
이전(V)
다음(N)
공변량(A):
adjCT
ampYN
T_COV_*ampYN
>a*b(A)>
방법(M):
입력
계층변수(A):
확인
붙여넣기(P)
재설정(R)
취소
도움말

# 시간종속 Cox 회귀모형

| | B | 표준오차 | Wald | 자유도 | 유의확률 | Exp(B) | Exp(B)에 대한 95.0% CI | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | 하한 | 상한 |
| sex | -.375 | .479 | .613 | 1 | .434 | .688 | .269 | 1.757 |
| PS_12_3 | .802 | .221 | 13.149 | 1 | .000 | 2.230 | 1.446 | 3.439 |
| smokYN | .469 | .329 | 2.023 | 1 | .155 | 1.598 | .838 | 3.047 |
| adjCT | .130 | .218 | .359 | 1 | .549 | 1.139 | .744 | 1.745 |
| ampYN | .582 | .279 | 4.348 | 1 | .037 | 1.790 | 1.036 | 3.095 |
| T_COV_*ampYN | .007 | .007 | 1.075 | 1 | .300 | 1.007 | .994 | 1.020 |

T_COV_*ampYN 는 유의확률 0.300으로 유의수준 0.05에서 통계적으로 유의하지 않다. 따라서 FGFR1 유전자 증폭여부에 따른 위험함수의 비는 시간에 따라 변한다고 할 수 없다 (비례위험의 가정 만족).


**[질문]** 리뷰어의 중요한 지적으로, 제안 콕스비례위험 회귀모형 내에서 혼란변수로 'adjuvant chemotherapy' 영향을 보정하는데 **'effect modifier'**로 작용할지 모르니 확인을 요청했습니다.

**[답변]** 두 변수 'FGFR1 유전자 증폭여부'와 'adjuvant chemotherapy' 간에 **상호작용(interaction effect)**이 존재하지 않음을 확인하였다 (p=0.174).

| | B | 표준오차 | Wald | 자유도 | 유의확률 | Exp(B) | Exp(B)에 대한 95.0% CI | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | 하한 | 상한 |
| sex | -.402 | .478 | .706 | 1 | .401 | .669 | .262 | 1.708 |
| PS_12_3 | .768 | .222 | 11.919 | 1 | .001 | 2.155 | 1.394 | 3.333 |
| smokYN | .492 | .329 | 2.235 | 1 | .135 | 1.636 | .858 | 3.118 |
| adjCT | .016 | .230 | .005 | 1 | .944 | 1.016 | .648 | 1.594 |
| ampYN | .400 | .354 | 1.274 | 1 | .259 | 1.491 | .745 | 2.986 |
| adjCT*ampYN | .603 | .444 | 1.845 | 1 | .174 | 1.827 | .766 | 4.362 |

# Basic concept： competing risks

- Subdistribution for an event of type ***i*** ($i$ = 1, 2, …, p)

$$F_i(t) = P(T \le t,\ C = i)$$

- Subhazard

$$\tilde{h}_i(t) = \lim_{\delta t \to 0} \left\{ \frac{P(t < T \le t + \delta t,\ C = i \mid T > t)}{\delta t} \right\} = \frac{f_i(t)}{S(t)}$$

- Hazard of the subdistribution

$$\gamma_i(t) = \lim_{\delta t \to 0} \left\{ \frac{P(t < T \le t + \delta t, C = i \mid T > t \quad or \quad T \le t \,\&\, C \ne i)}{\delta t} \right\} = \frac{f_i(t)}{1 - F_i(t)}$$

➔ the probability of observing the event of interest, type $i$, at the time $t$ while knowing that either the event of interest did not happen until then or that the competing risks event was observed.

# Model

- Cox's proportional hazard model, Cox (1972)

Partial likelihood

$$L(\beta_1, \beta_2, \ldots, \beta_m) = \prod_{j=1}^{r} \left( \frac{\exp\{\beta_1 x_{1j} + \ldots + \beta_m x_{mj}\}}{\sum_{i \in R_j} \exp\{\beta_1 x_{1i} + \ldots + \beta_m x_{mi}\}} \right)$$

- Model for the hazard of CIF, Fine & Gray (1999)

$$L(\beta_1, \beta_2, \ldots, \beta_m) = \prod_{j=1}^{r} \left( \frac{\exp\{\beta_1 x_{1j} + \ldots + \beta_m x_{mj}\}}{\sum_{i \in R_j} w_{ji} \exp\{\beta_1 x_{1i} + \ldots + \beta_m x_{mi}\}} \right)$$

$$w_{ji} = \frac{\hat{G}(t_j)}{\hat{G}(\min(t_j, t_i))}$$

$\hat{G}(\cdot)$ : K-M estimate of the survivor function of the censoring distribution

$R_j(t) = \{i \; ; \; T_i \geq t$ or ( $T_i \leq t$ and the subject experienced a competing risk event)}

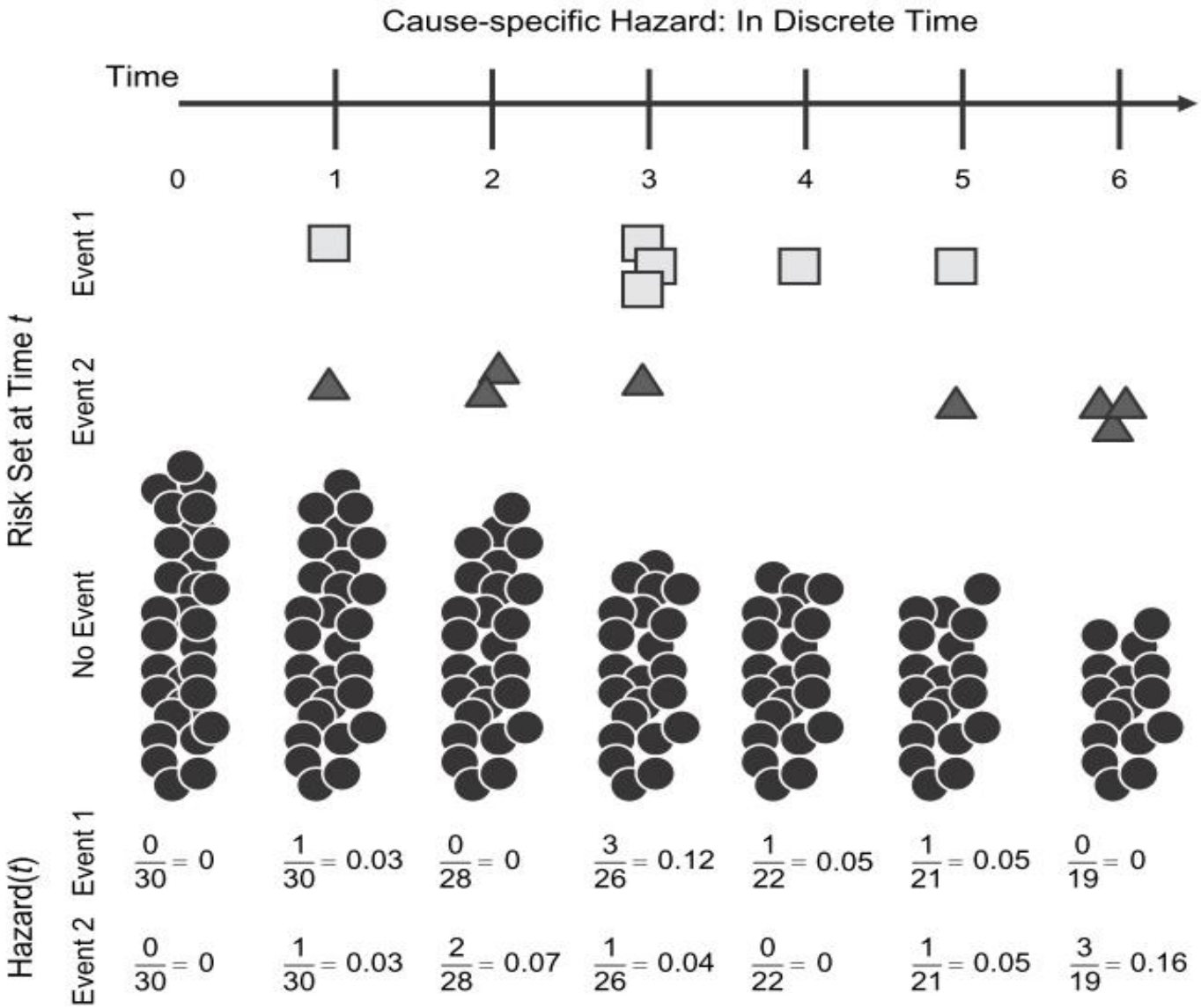



**Figure 1.** **Cause-specific hazard** schematic. The risk set starts with **30** individuals (solid circles). Over time, individuals have either event 1 (square) or event 2 (triangle). As individuals have either event, they are removed from the remaining risk sets. The calculation for the cause-specific hazard is given at the bottom of the figure.

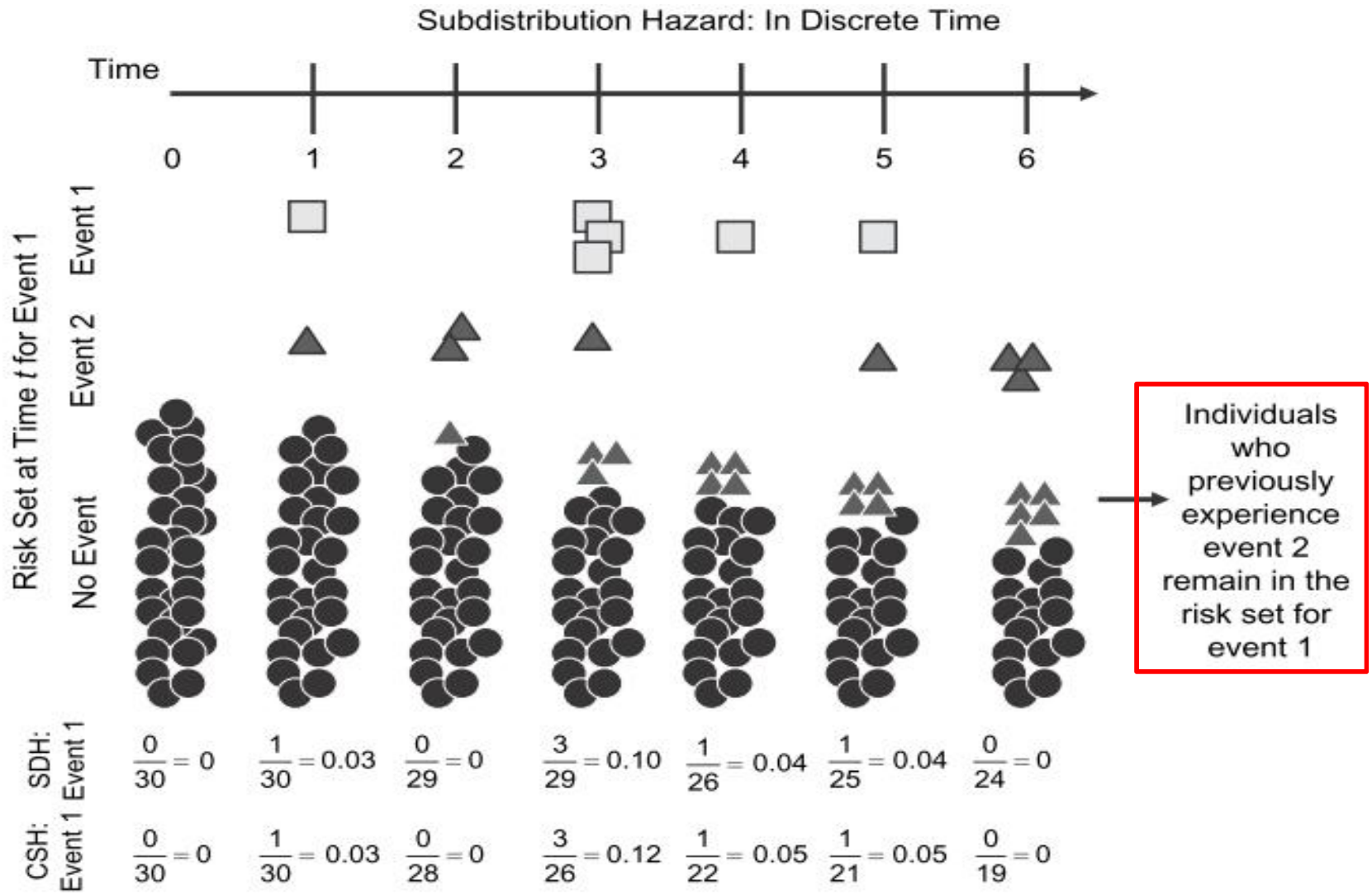



**Figure 2. Subdistribution hazard(SDH)** schematic. The risk set starts with **30** individuals (solid circles). Over time, individuals have either event 1 (square) or event 2 (triangle). As individuals have the competing event (event 2, triangle), they are maintained in the risk set as triangles. Thus, over time, a greater proportion of the risk set becomes full of triangles that are individuals who have had the competing event prior to that time. The SDH for event 1 is given near the bottom of the figure along with the cause-specific hazard (CSH) for event 1 for comparison. Note that, because individuals are maintained in the risk set, the SDH tends to be lower than the CSH.

# New-onset atrial fibrillation predicts long-term newly developed atrial fibrillation after coronary artery bypass graft

*Am Heart J* 2014;167(4):593-600.e1.

Therefore, we set a competing risk of death **without long-term AF recurrence** and compared the cumulative incidence(CIF) curves by **Gray's method** with a 1-KM method. Competing risk regression by **Fine and Gray's model** was also performed. Finally, we assessed the 2 HRs of Cox proportional hazard regression and competing risk regression using death as a competing event.

Ref. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. J Am Stat Assoc 1999;94:496-509.

**Table III.** Competing risk and Cox proportional hazard models of time to long-term AF of POAF group compared with no-POAF group

| | Long-term AF | |
|---|---|---|
| **Regression models*** | **HR (95% CI)** | ***P*** |
| Cox proportional hazard regression | 5.25 (1.75-15.77) | .003 |
| Competing risk regression | 4.99 (1.68-14.84) | .004 |

*Adjusted for age, sex, covariates, and propensity score.

# 4 Major Bigdata in Bio-Healthcare

병원/개인 진료정보
EMR/EHR/PHR

GENOME

Multi-OMICS
KoGES

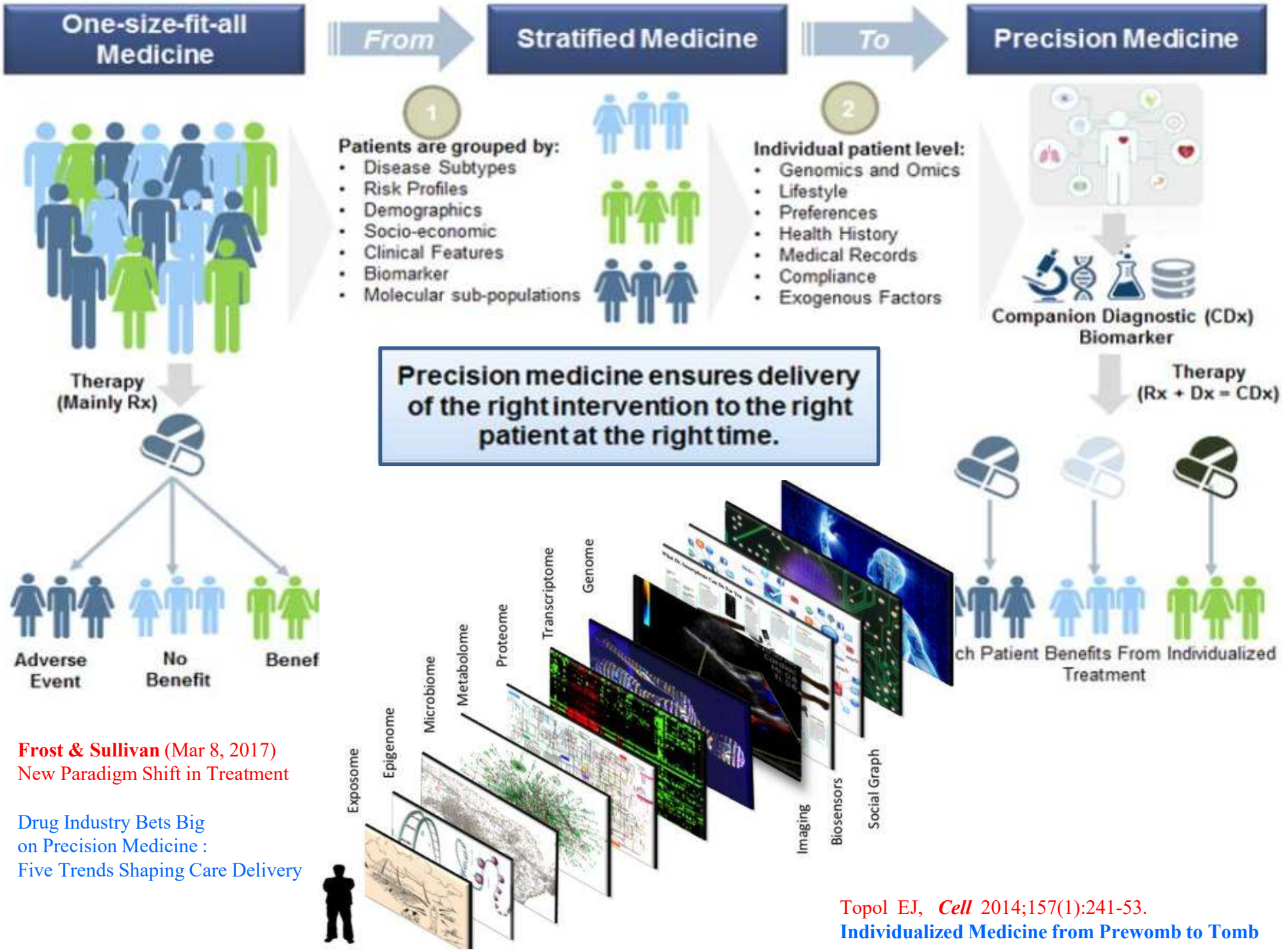


Lifelog Data
Wearable technology
Mobile devices

h·well 국민건강보험 National Health Insurance Service
건강보험심사평가원 HEALTH INSURANCE REVIEW & ASSESSMENT SERVICE
질병관리본부
기상청
식품의약품안전처

## Data

Information ➡ Knowledge ➡ Theory & Expertise




**Frost & Sullivan** (Mar 8, 2017)
New Paradigm Shift in Treatment

Drug Industry Bets Big
on Precision Medicine :
Five Trends Shaping Care Delivery

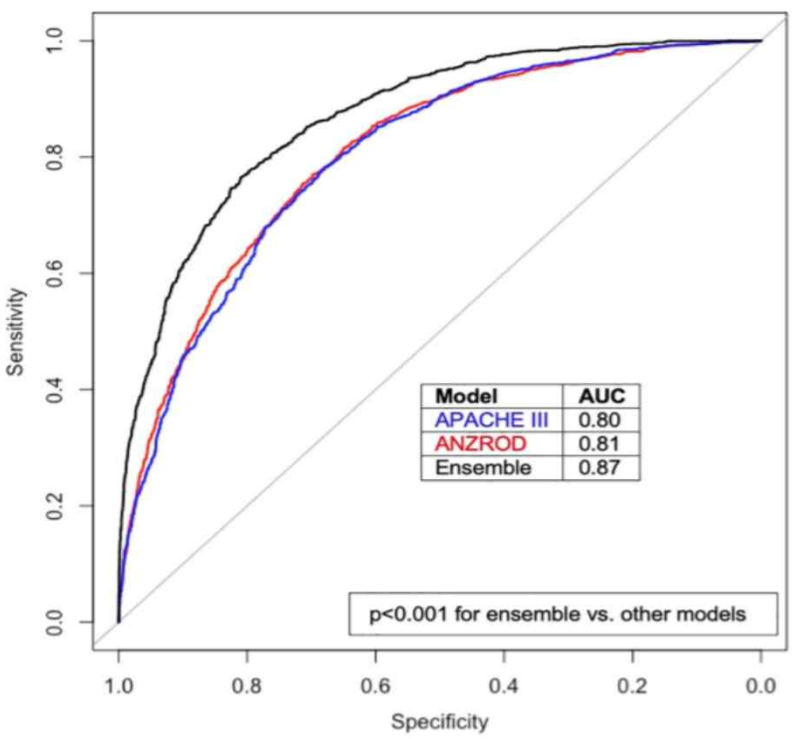
Topol EJ, ***Cell*** 2014;157(1):241-53.
**Individualized Medicine from Prewomb to Tomb**

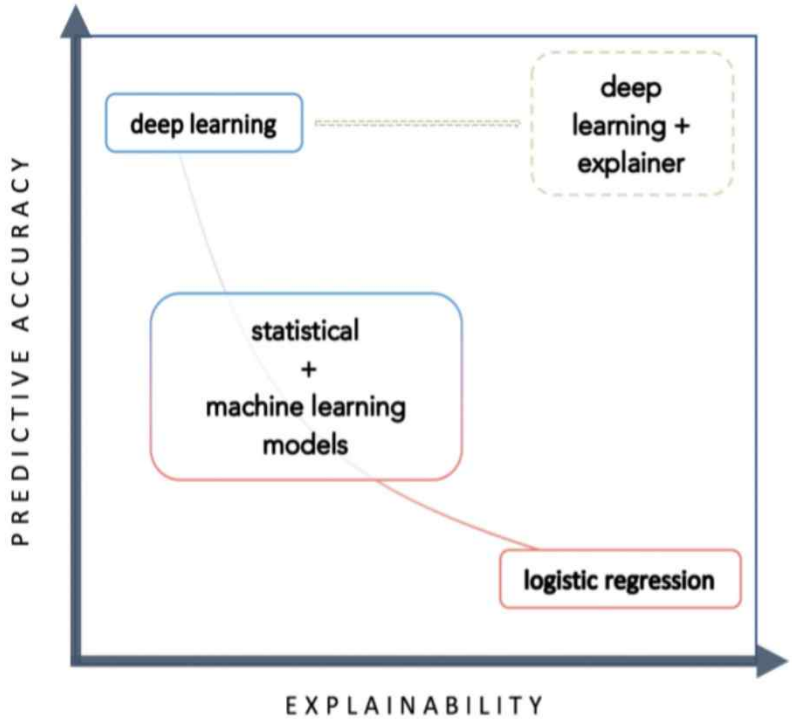
PLOS | MEDICINE

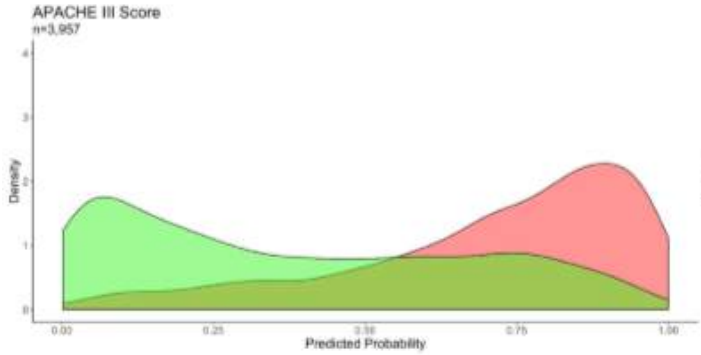
November 2018

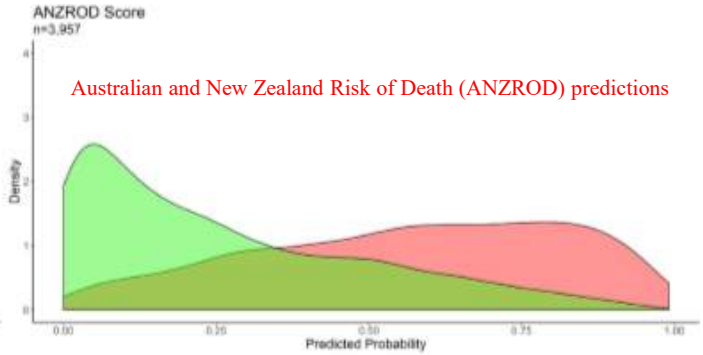
# Advancing the beneficial use of machine learning in healthcare and medicine

- Deep learning for lung cancer prognostication A retrospective multi-cohort radiomics study
- Characterising risk of in-hospital mortality following cardiac arrest using machine learning A retrospective international registry study
  Nanayakkara et al., Survivors (n = 21,547) + Non-survivors (n = 18,019) / Australian and New Zealand Intensive Care Society (ANZICS)
- Mobile detection of autism through machine learning on home video A development and prospective validation study
- Enhancing the prediction of acute kidney injury risk after percutaneous coronary intervention using machine learning techniques A retrospective cohort study
- Development and validation of machine learning models to identify high-risk surgical patients using automatically curated electronic health record data (Pythia) A
- Deep-learning-assisted diagnosis for knee magnetic resonance imaging Development and retrospective validation of MRNet
- Predicting the risk of emergency admission with machine learning Development and validation using linked electronic health records
- Machine learning to identify pairwise interactions between specific IgE antibodies and their association with asthma A crosssectional analysis within a population-b
- Deep learning for chest radiograph diagnosis A retrospective comparison of the CheXNeXt algorithm to practicing radiologists
- Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs A cross-sectional study
- Prediction of myopia development among Chinese school-aged children using refraction data from electronic medical records A retrospective, multicentre machi
- Machine learning in population health Opportunities and threats
- Deep learning and artificial intelligence in radiology Current applications and future directions
- The use of machine learning to understand the relationship between IgE to specific allergens and asthma
- Transforming health policy through machine learning
- Machine learning in medicine Addressing ethical challenges
- Machine learning assessment of myocardial ischemia using angiography Development and retrospective validation
- Advancing the beneficial use of machine learning in health care and medicine Toward a community understanding
- Automated detection of moderate and large pneumothorax on frontal chest X-rays using deep convolutional neural networks A retrospective study

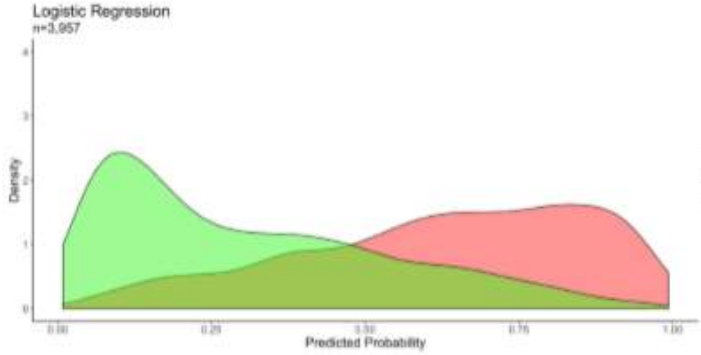


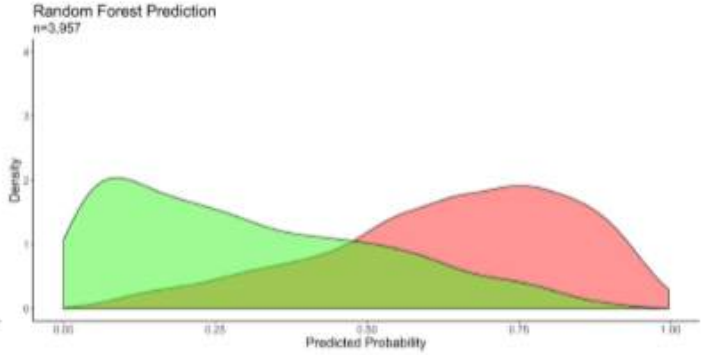



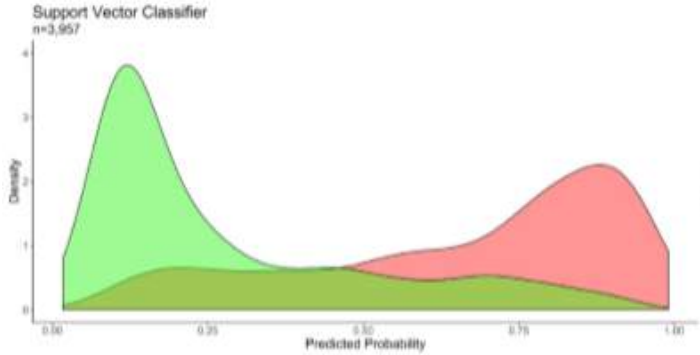



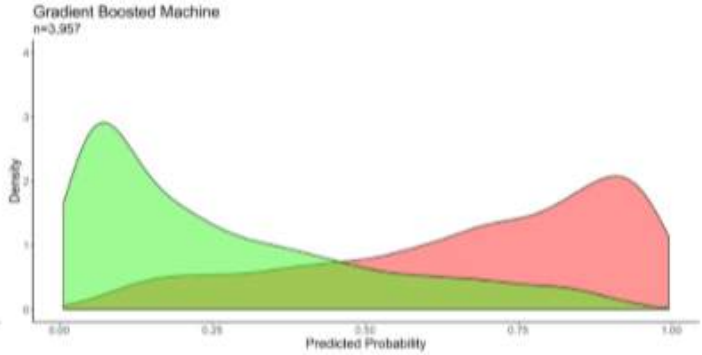



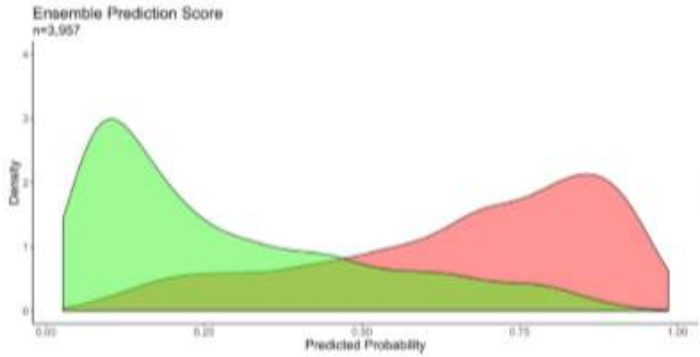

Australian and New Zealand Risk of Death (ANZROD) predictions

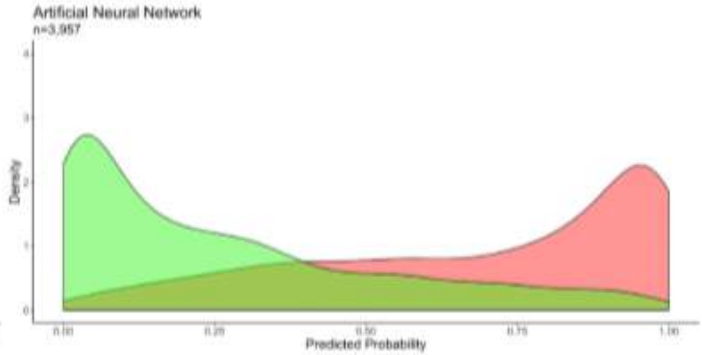


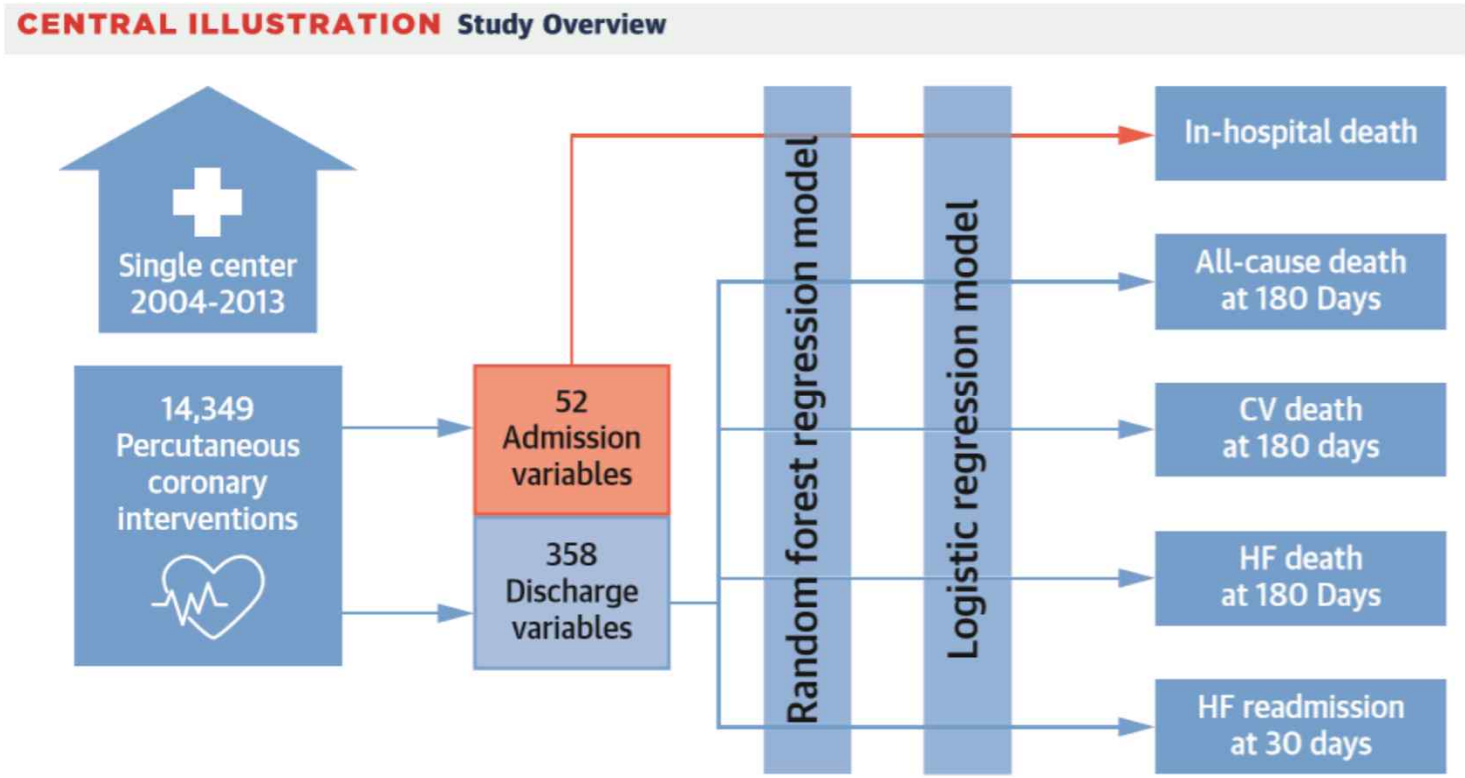



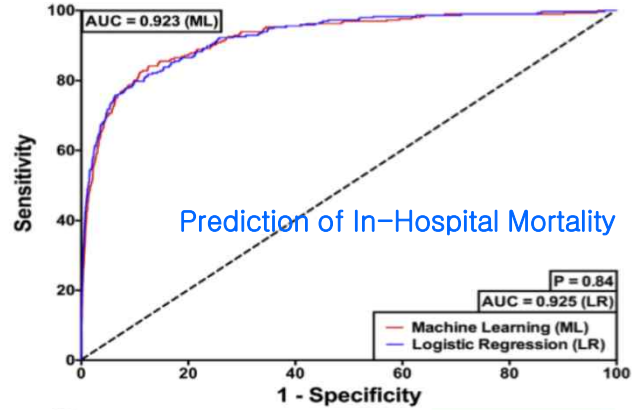



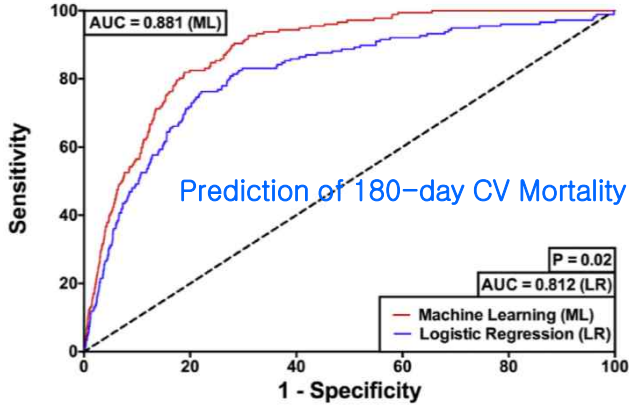



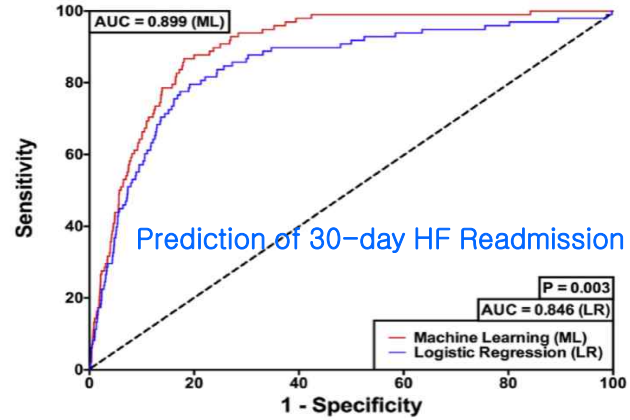





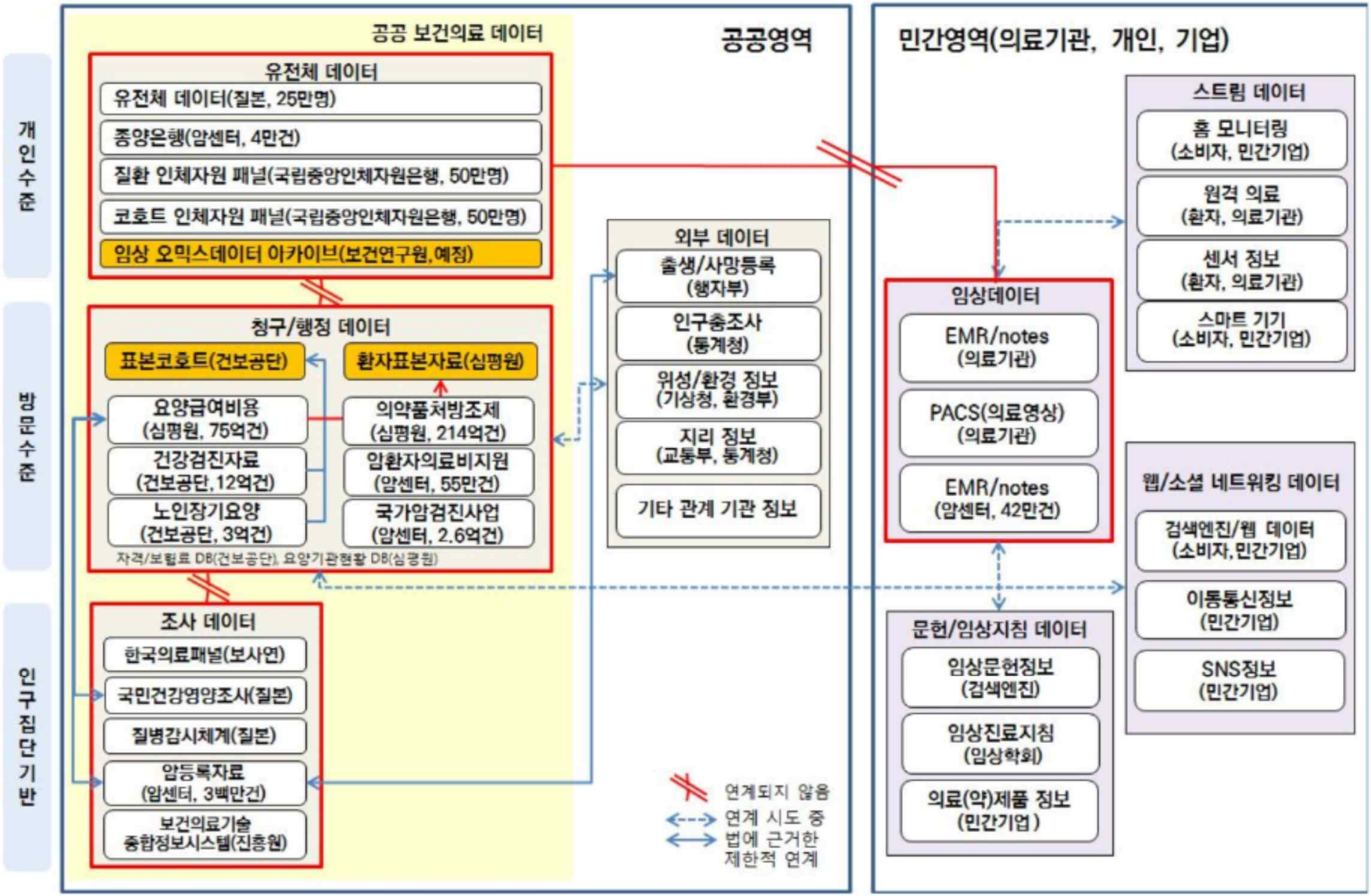




## Leveraging Machine Learning Techniques to Forecast Patient Prognosis After Percutaneous Coronary Intervention (PCI)

Mayo Clinic, PCI registry, Zack, C.J. et al., *JACC* (J Am Coll Cardiol Intv) 2019

This study sought to determine whether machine learning can be used to better identify patients at risk for death or congestive heart failure (CHF) re-hospitalization after PCI

# Possibility of Connection for Korea's Healthcare Bigdata




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