Statistical Analysis Methods for Reinforcing Causal Inference

2021. 11. 14.

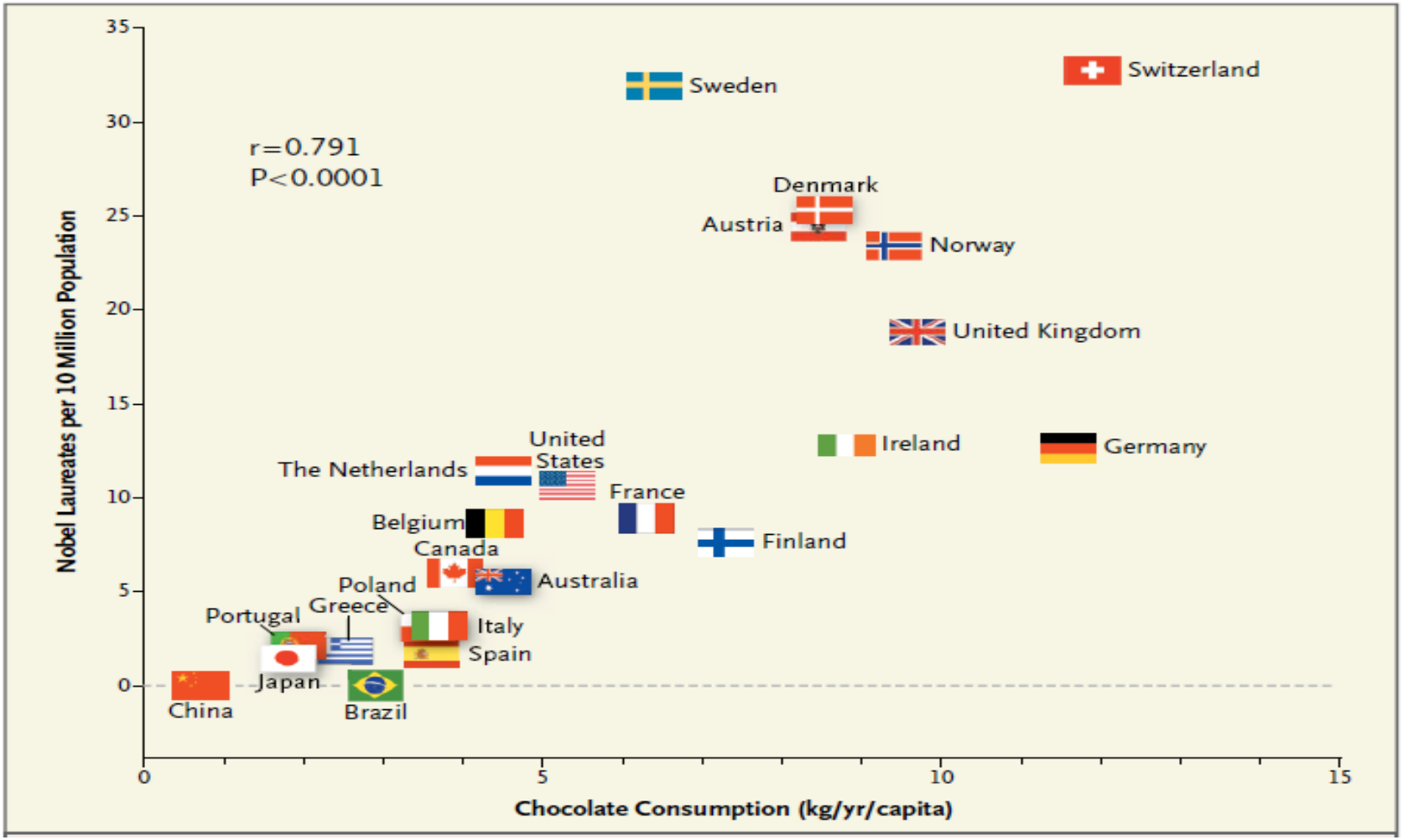
국민건강빅데이터임상연구소
National Health BigData Clinical Research Institute

연세대학교 원주의과대학 정밀의학과·의학통계학과

강 대 용

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

노벨위원회는 "데이비드 카드가 노동경제학에 대한 실증적 기여로 경제학상 수상자로 선정됐다"고 밝혔다. 또 "조슈아 앵그리스트와 귀도 임벤스는 인과관계 분석에 대한 방법론적 기여를 인정받았다"고 전했다.



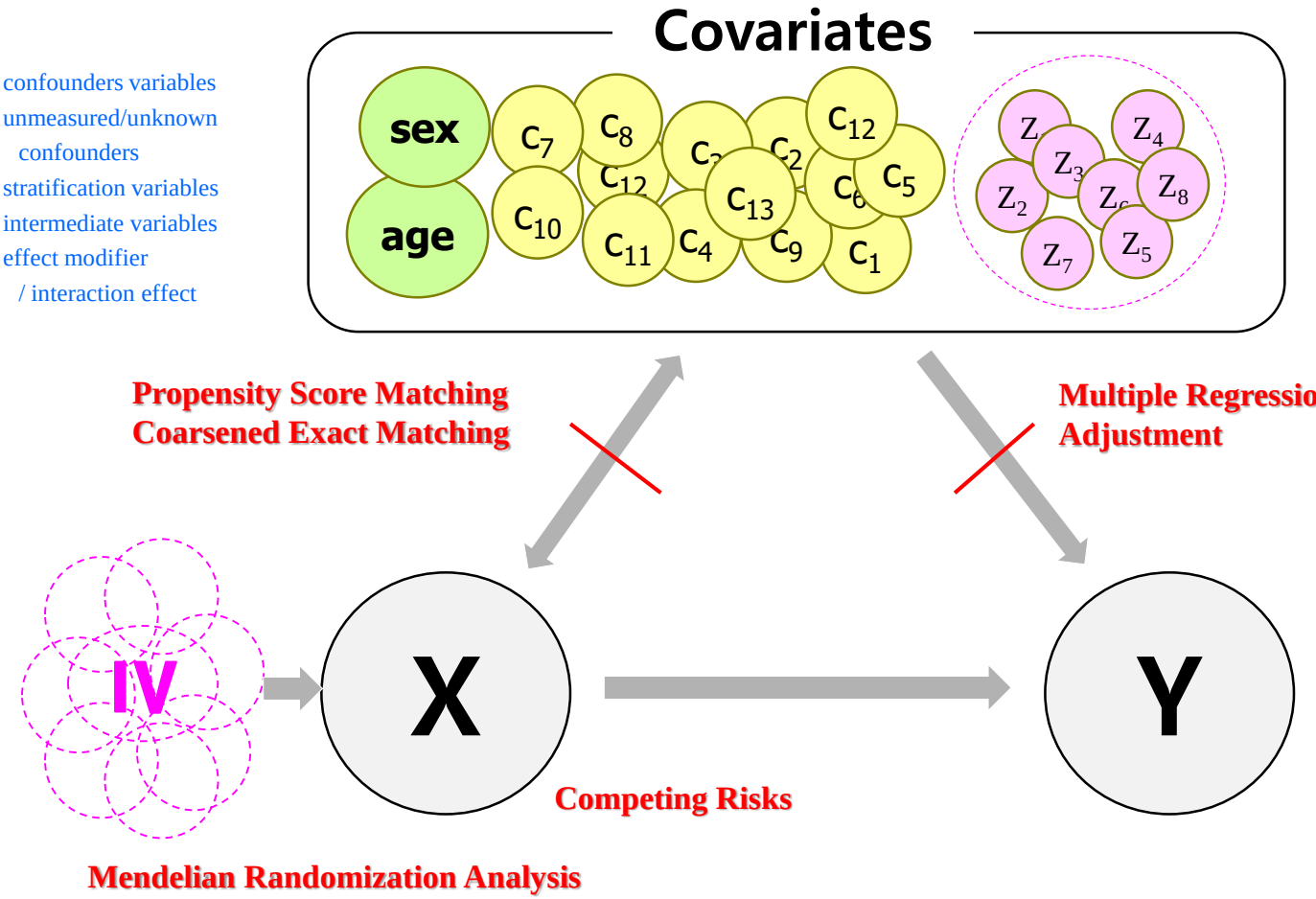
Franz H. Messerli, ***N Engl J Med*** 2012 Oct; 367(16):1562-4.

Chocolate Consumption — **Nobel Laureates**

Cognitive Function

Statistical Methods for Causal Inference

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



+ 'Time'을 어떻게 보정할 것인가?

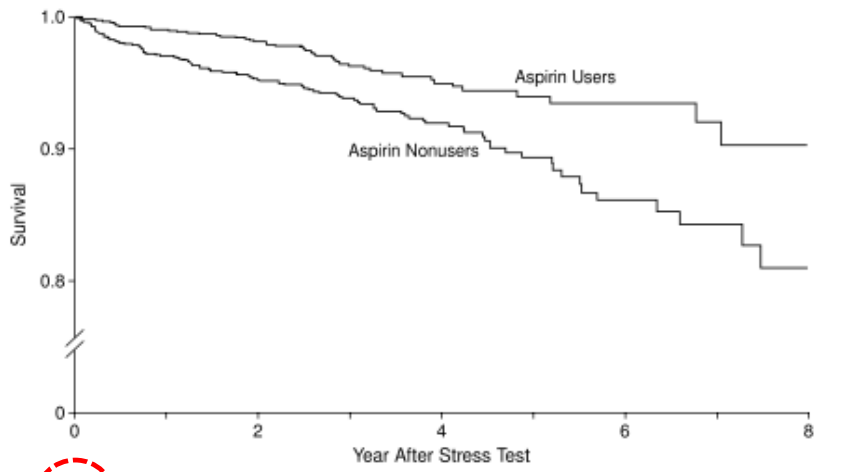
- Multiple Regression Analysis
- Logistic Regression Analysis
- Poisson Regression Analysis
- Cox's PHM
- Linear Mixed Model (**LMM**)
- Generalized Estimating Equation (**GEE**)

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

Variable	Aspirin (n = 1351) 2310	No Aspirin (n = 1351) 3864	*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/m²	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



No. at Risk									
Aspirin Use	1351	1338	1328	1314	1308	1305	1304	1303	1300
No Aspirin Use	1351	1312	1292	1280	1269	1260	1254	1252	1249

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

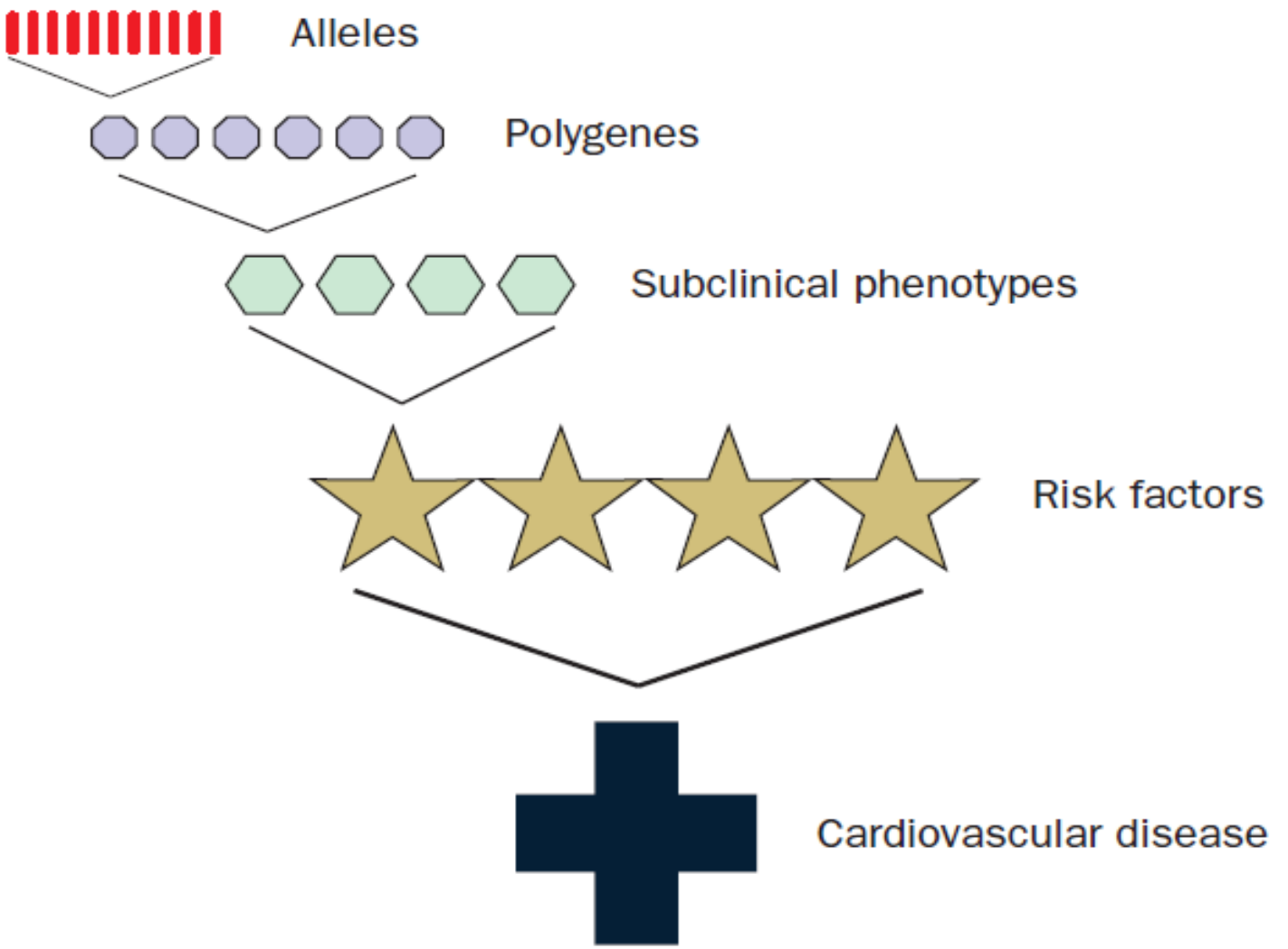
Daniel E. Ho
Stanford Law School

Kosuke Imai
Princeton University

Gary King
Harvard University

Elizabeth A. Stuart
Johns Hopkins University

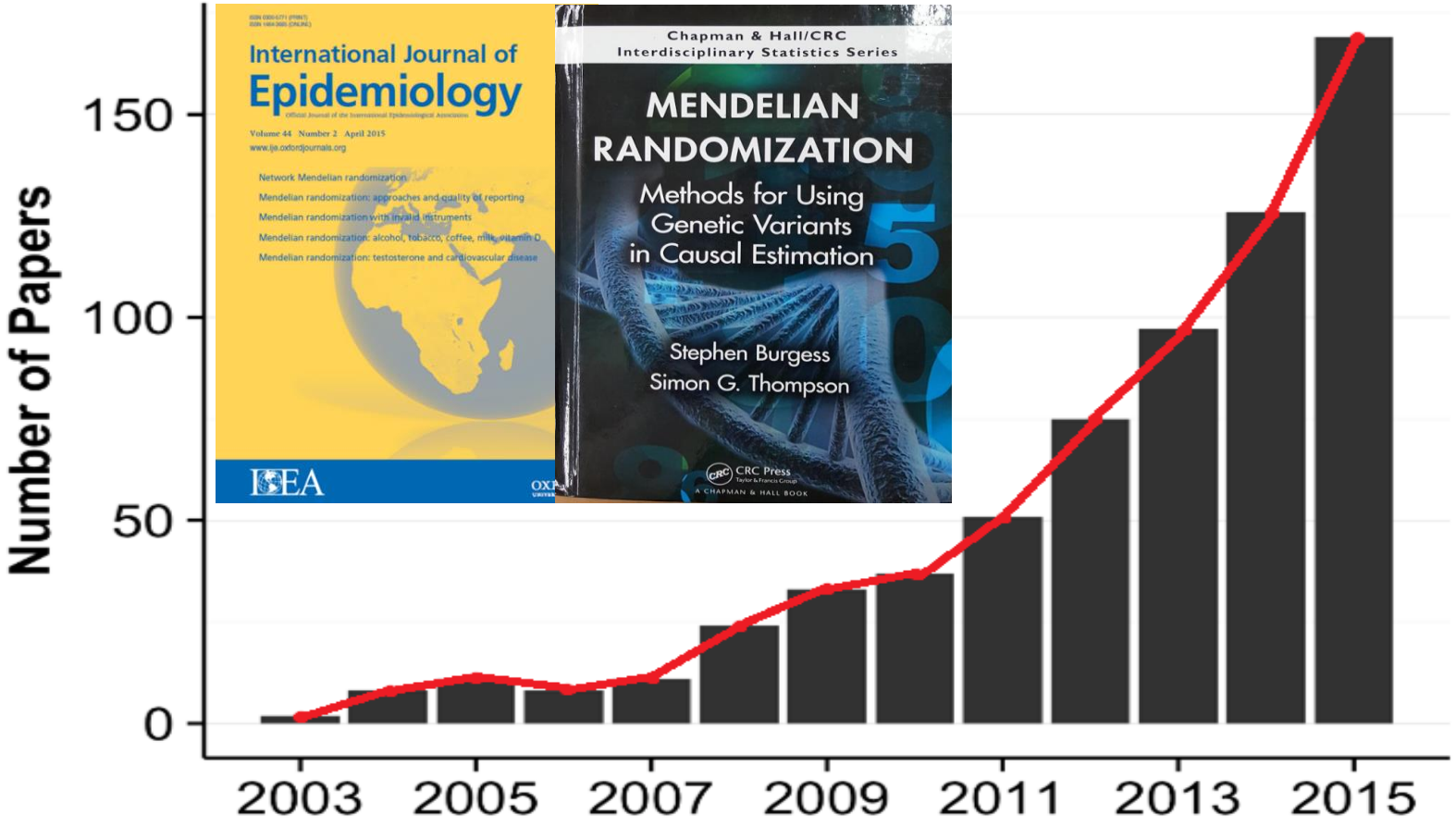
Journal of Statistical Software May 2011, Volume 42, Issue 8. http://www.jstatsoft.org/



Hierarchy in genetics of cardiovascular disease

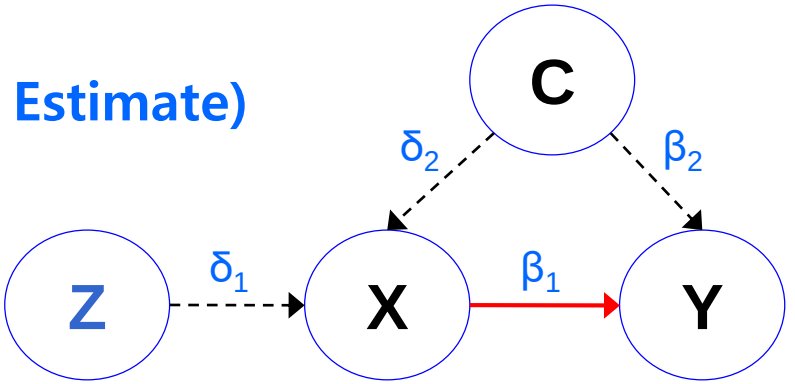
Source: Harrap et al., *Lancet* 2003;361:2149-51.

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)



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

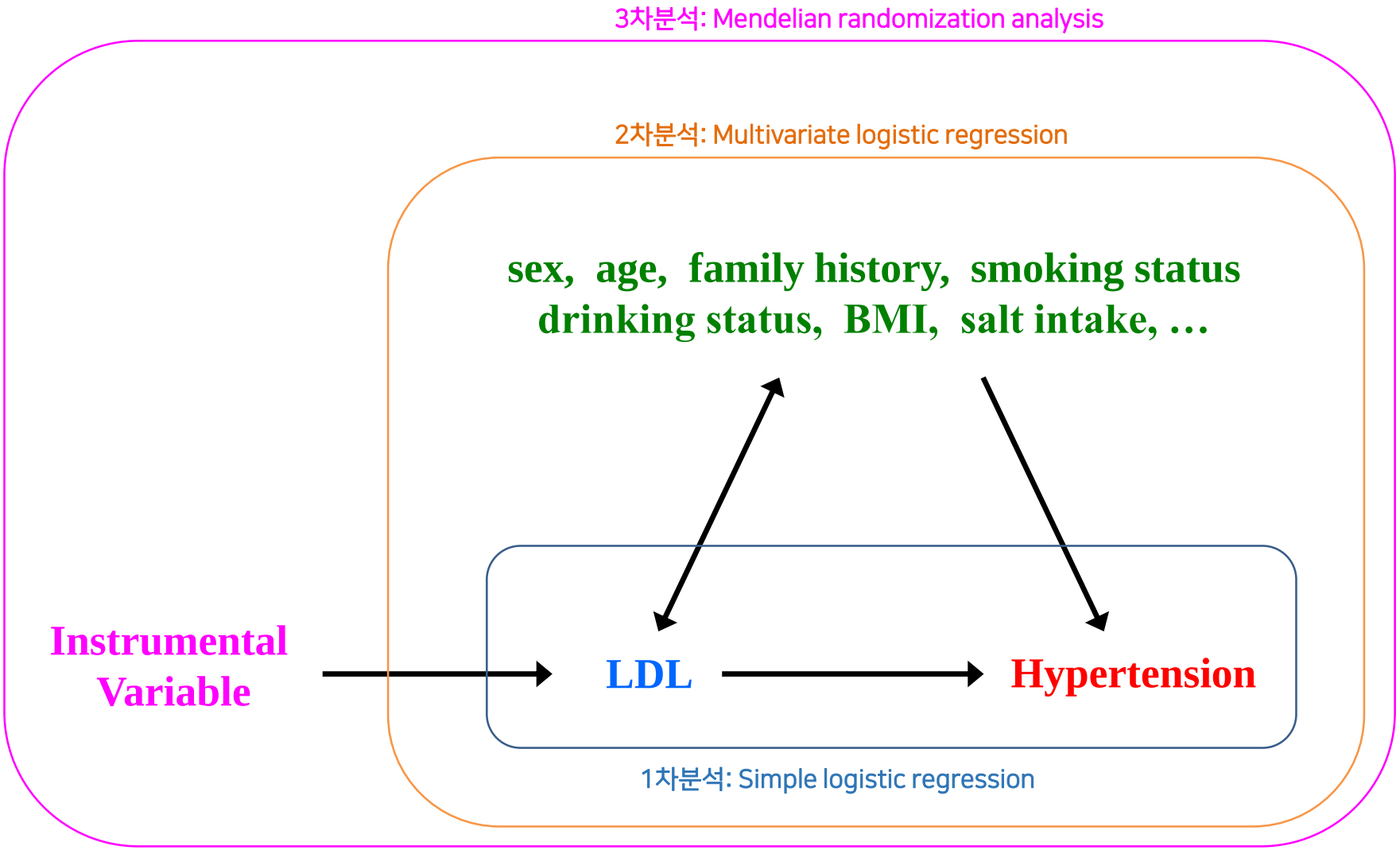
Stage 1 : regress of the X on the Z

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

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 */
```



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

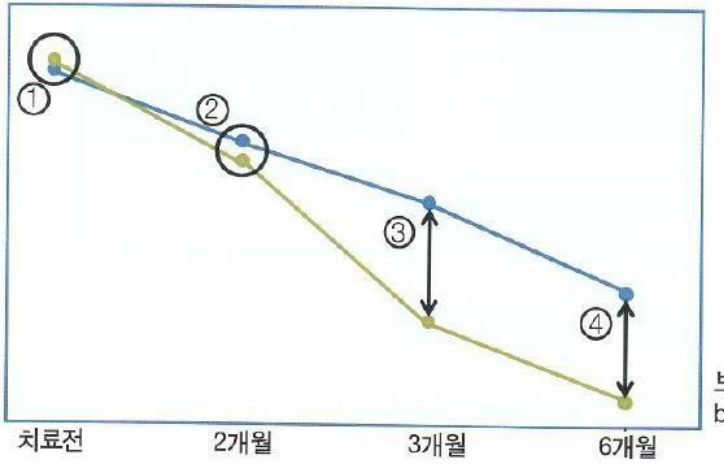
종속변수	독립변수	통계분석법
연속형	범주형(3개 이상)	ANOVA **Repeated Measures ANOVA**
연속형	연속형 + 범주형	회귀분석 General LM / **LMM** Linear Mixed Model
이분형	연속형 + 범주형	로지스틱 회귀분석 HGLM / **GEE** Generalized Estimating Equations
생존시간	연속형 + 범주형	**Cox PH 모형** Frailty 모형

반복측정 분산분석 에서의 세 가지 검정

군	두 군 간에 차이가 있는가 ?	개체간 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)**

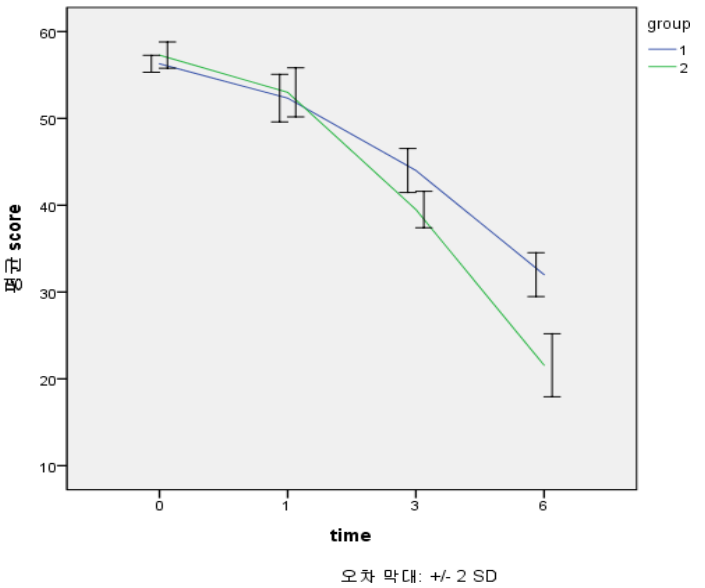
id	group	sex	baseline	month1	month3	month6
1	1	F	60	결측치	25	16
2	1	F	52	38	23	12
3	1	F	62	36	22	14
4	1	F	58	34	21	13
5	1	M	65	34	28	18
6	1	M	58	42	26	결측치
7	1	M	53	38	결측치	21
8	2	F	55	42	33	22
9	2	M	55	54	46	26
10	2	M	60	55	46	23
11	2	M	63	45	결측치	25
12	2	M	52	결측치	35	22
13	2	F	61	38	32	18
14	2	F	58	결측치	39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



- 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 곱해

일반적 통계분석 방법과 생존분석의 비교

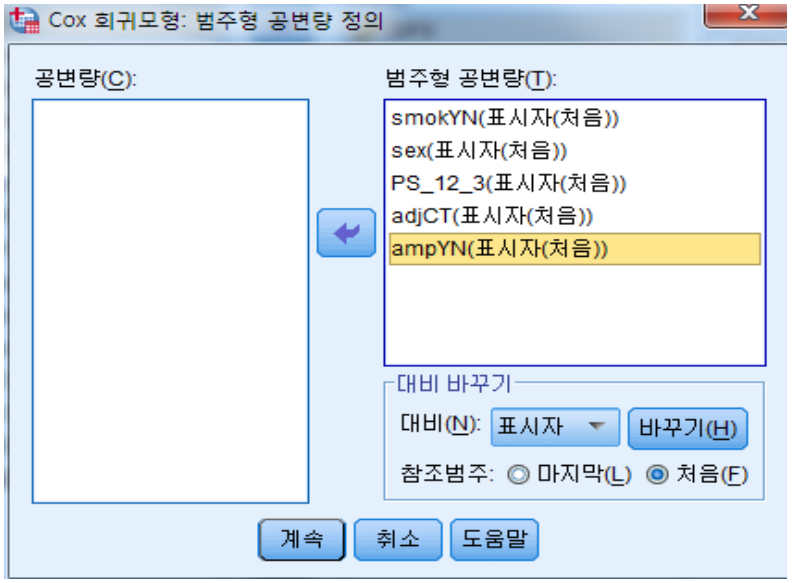
	일반적인 방법	생존분석
Demographic Graph	- Mean±SD (bar graph) - Median (min, max) or Median (Q1, Q3) - Box plot	- Kaplan-Meier method (1958)
1:1의 관계	- Independent two sample t-test, ANOVA - Mann-Whitney U test, Kruskal-Wallis test - Chi-square test (Fisher's exact test)	- Log-rank test (Mantel-Haenszel, 1959)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

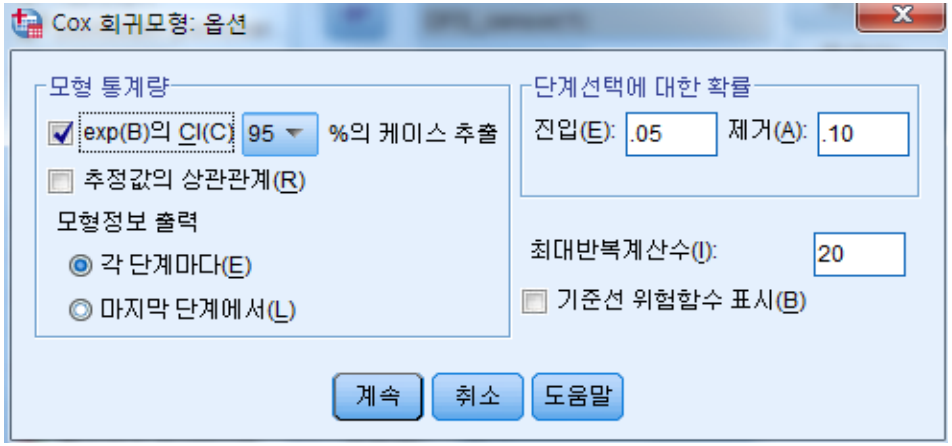
Cox Proportional Hazards Model (1972)

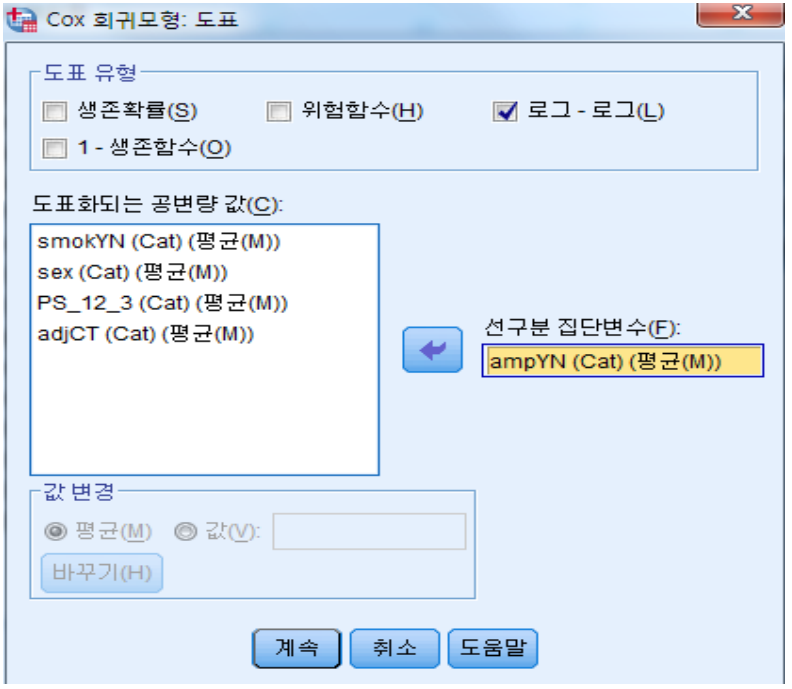
- 생존곡선에 영향을 주는 '위험요인'과의 관련성을 모형화하는 것이 목적
- 다른 변수들의 효과를 보정한 후 치료효과를 볼 수 있는 대표적인 통계모형

Baseline	$h_0(t)$: 모든 독립변수가 0일 때의 위험함수
Model	t시점에서 p개의 독립변수가 $x_1, x_{2,\ldots,} x_p$ 일 때의 위험함수
Semi-parametric model	$h(t) = h_0(t)\exp(\beta_1 x_1 + \ldots + \beta_k x_k)$ $\log h(t,x) = \log h_0(t) + \beta_1 x_1 + \ldots + \beta_k x_k$
Assumption	i번째 환자와 j번째 환자의 위험비가 시간과 무관하게 상수가 됨 $h_i(t)/h_j(t) = \exp(\beta_1(x_{i1} - x_{j1}) + \ldots + \beta_k(x_{ik} - x_{jk}))$
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

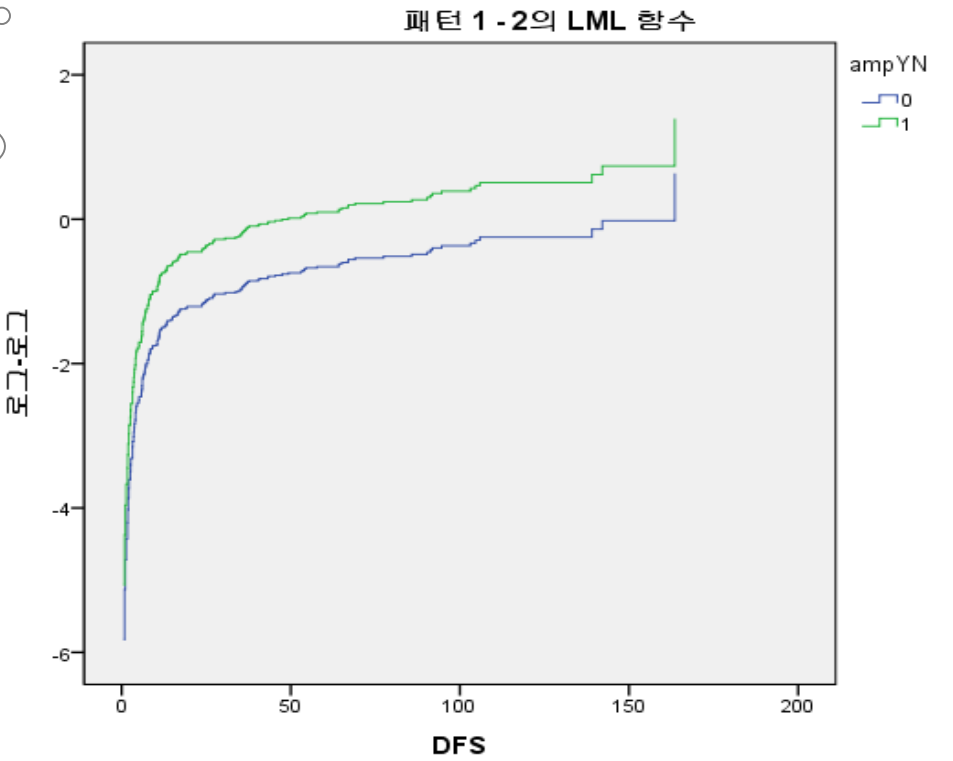

Cox 회귀모형: 옵션
모형 통계량
exp(B)의 CI(C)
95
%의 케이스 추출
추정값의 상관관계(R)
모형정보 출력
각 단계마다(E)
마지막 단계에서(L)
단계선택에 대한 확률
진입(E):
.05
제거(A):
.10
최대반복계산수(I):
20
기준선 위험함수 표시(B)
계속
취소
도움말


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).

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

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

- Hazard of the subdistribution

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

➔ 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)}

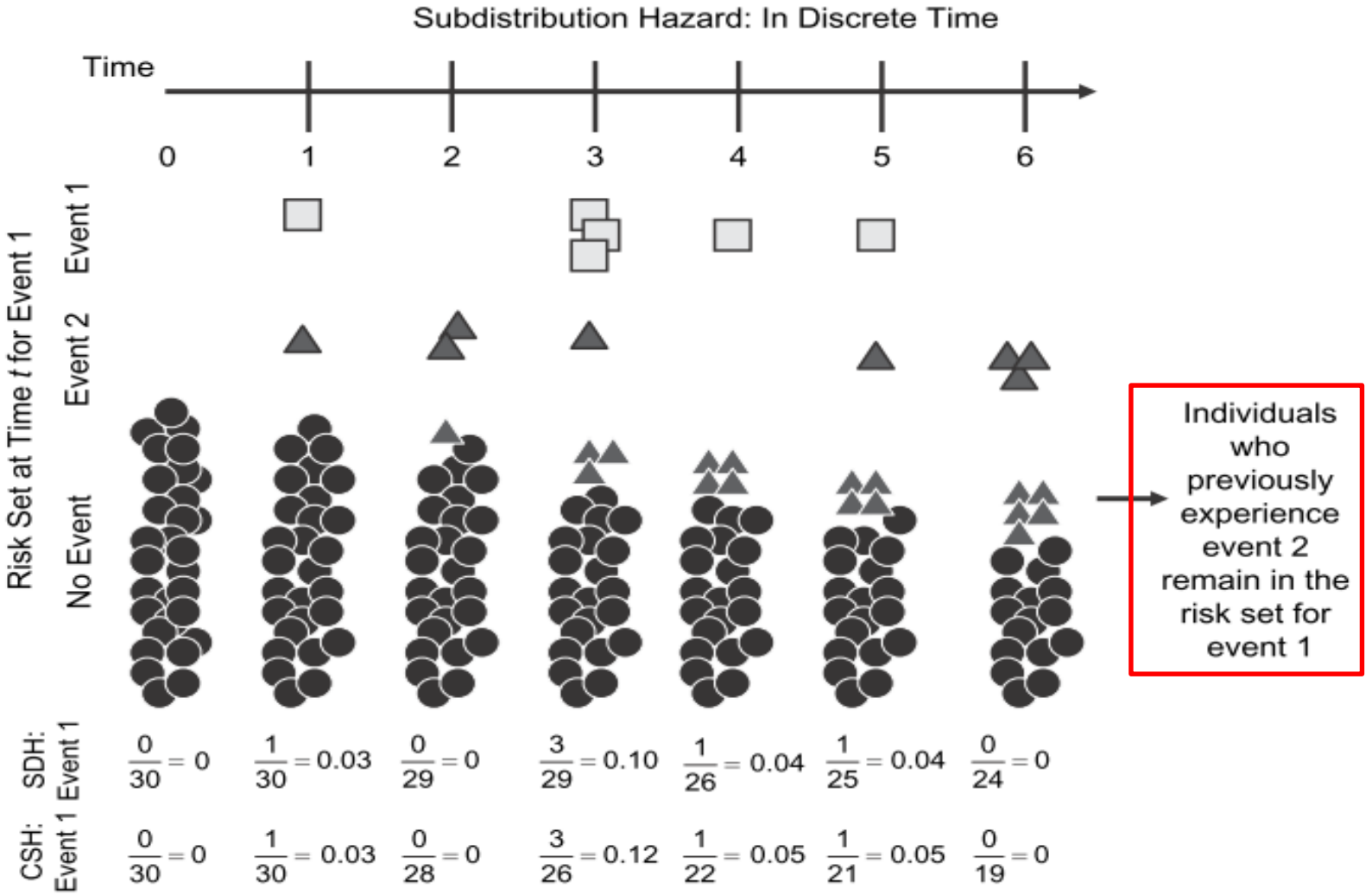


Figure 2. Subdistribution hazard(SDH) schematic. The risk set starts with **30** individuals (solid circles). Over time, individuals have either event 1 (square) or event 2 (triangle). As individuals have the competing event (event 2, triangle), they are maintained in the risk set as triangles. Thus, 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.

4 Major Bigdata in Bio-Healthcare

병원/개인 진료정보
EMR/EHR/PHR

GENOME

Multi-OMICS
KoGES

h·well 국민건강보험
National Health Insurance Service
건강보험심사평가원
질병관리본부
기상청
식품의약품안전처

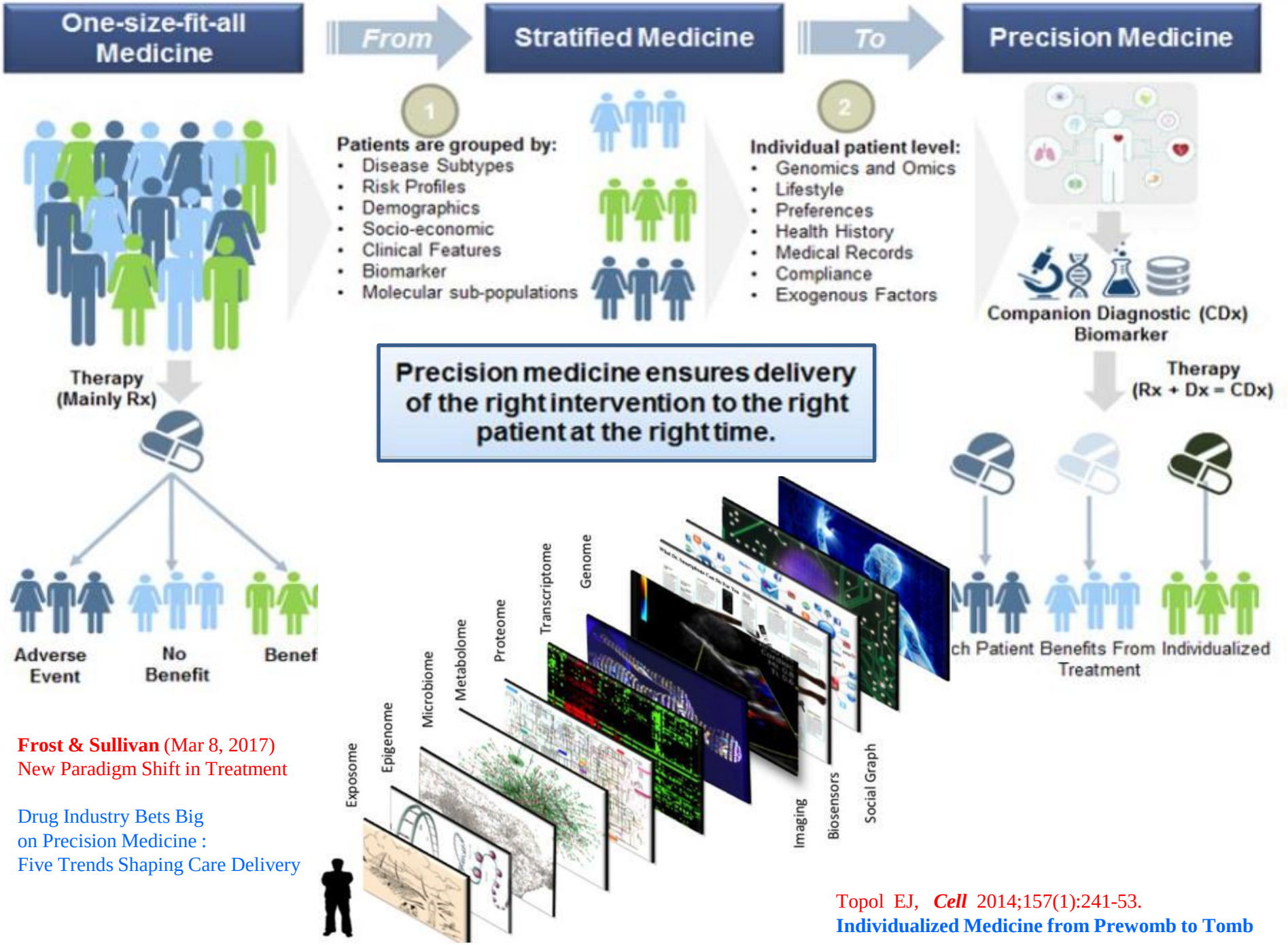
Lifelog Data
Wearable technology
Mobile devices
PGHD

PGHD
(Patient-Generated Health Data)란?
스마트폰, 웨어러블 디바이스 등의 발전으로 환자가 병원에 가지 않더라도 자신의 건강 데이터를 측정하고 저장할 수 있는 시대가 되면서, 환자가 자발적으로 생산한 건강데이터를 PGHD라 한다. 여기에는 건강에 대한 이력이나 생체 데이터, 생활습관 등이 모두 속한다. **PGHD가 중요한 이유는**

▲병원에 방문하지 않는 기간 환자의 상태를 파악이 가능하다는 점,
▲만성질환의 관리 혹은 예방에 가장 근접한 정보를 제공한다는 점,
▲복용중인 약물정보, 알러지 정보 등의 환자 안전성에 기여 가능하다는 점의 장점이 있기 때문이다.

Data

Information ➡ Knowledge ➡ Theory & Expertise



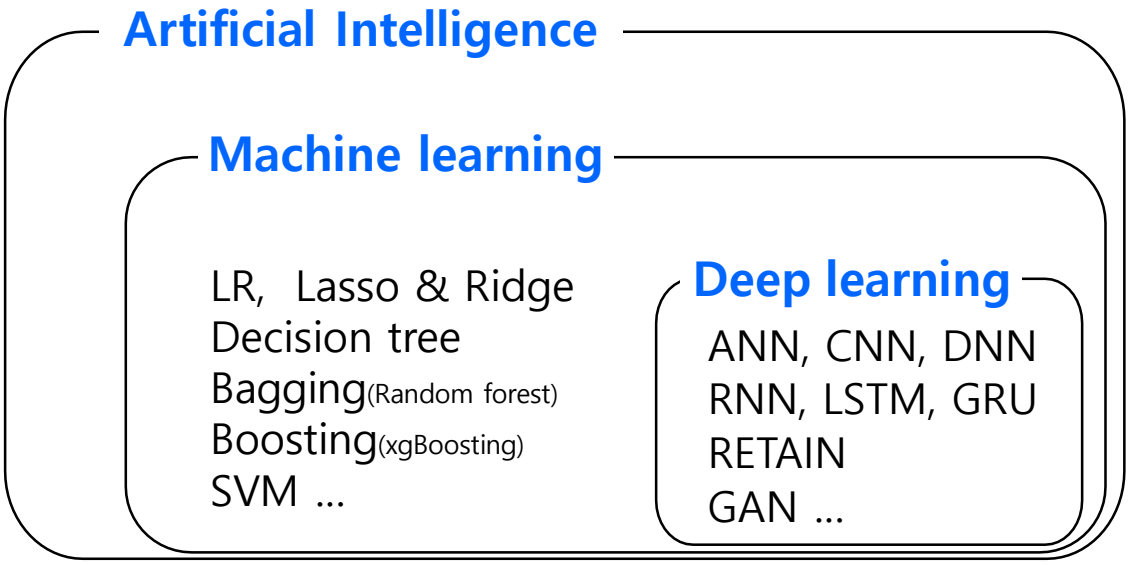
Frost & Sullivan (Mar 8, 2017)
New Paradigm Shift in Treatment

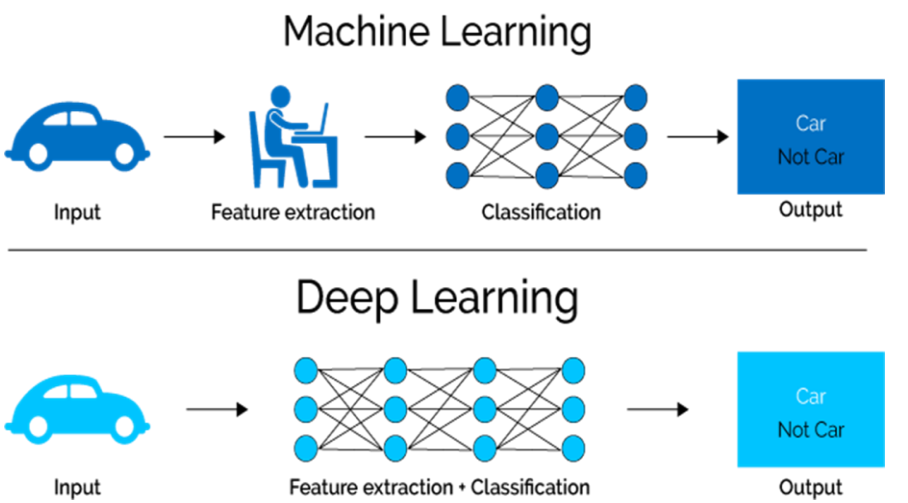
Drug Industry Bets Big
on Precision Medicine :
Five Trends Shaping Care Delivery

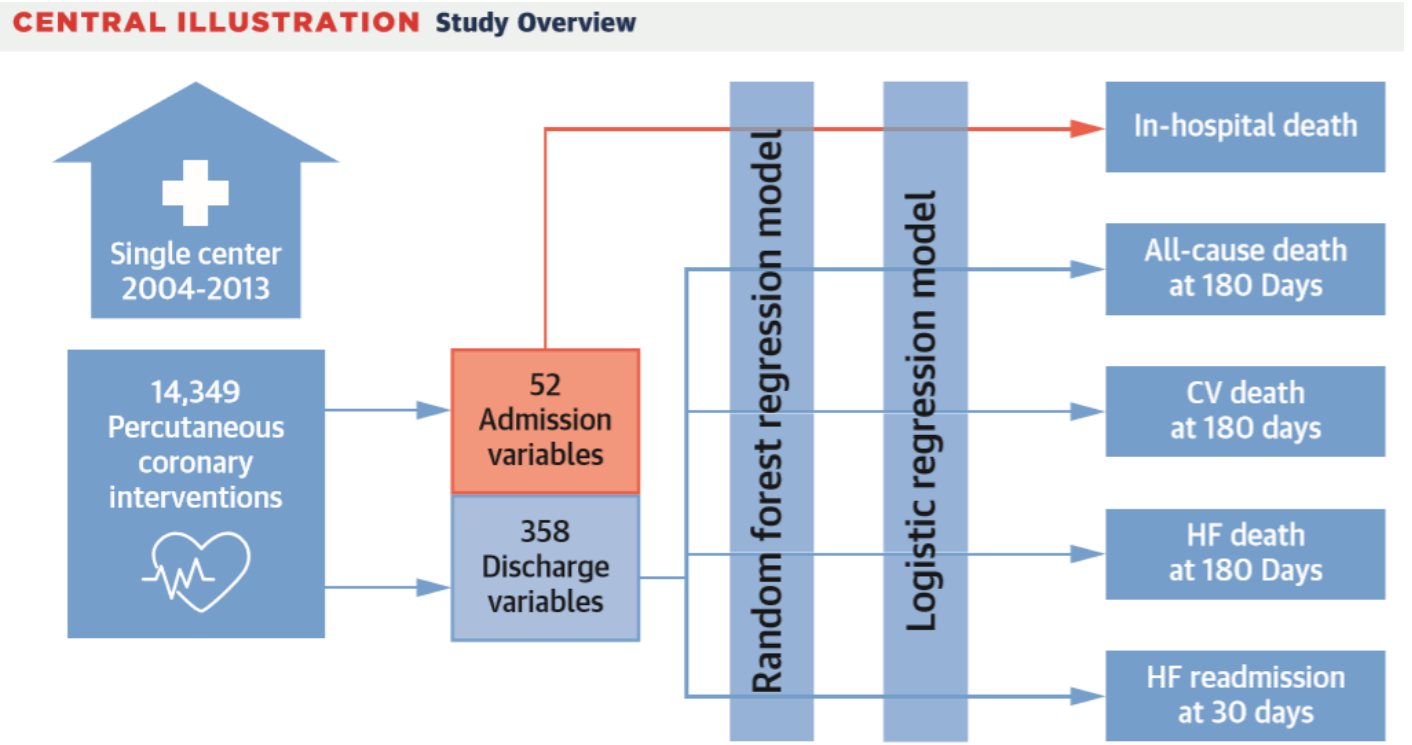
Topol EJ, ***Cell*** 2014;157(1):241-53.
Individualized Medicine from Prewomb to Tomb

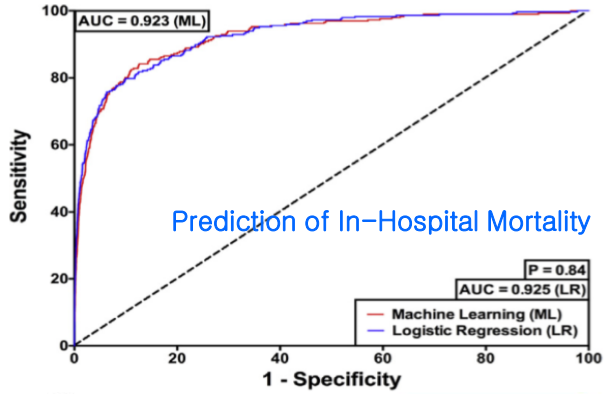
LR – ML – DL

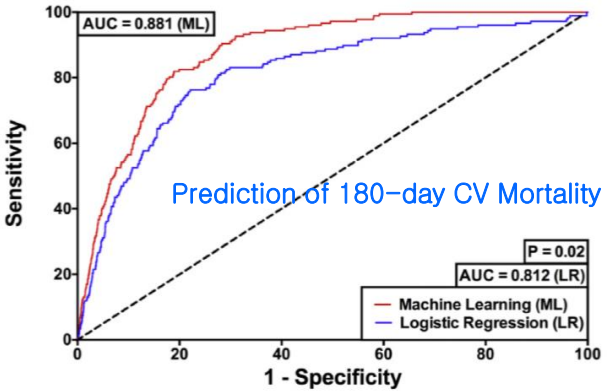
- There is no bright line between machine learning models and traditional statistical models
- **Deep learning** is well suited to learn from the complex and heterogeneous kinds of data that are generated from modern clinical care, such as medical notes entered by physician, medical images, continuous monitoring data from sensors, and genomic data to help make medically relevant predictions.

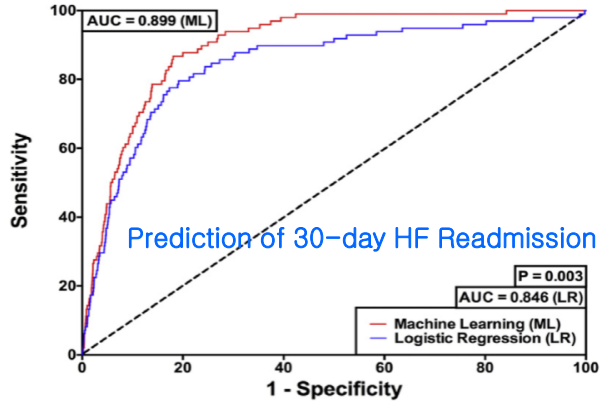












Leveraging Machine Learning Techniques to Forecast Patient Prognosis After Percutaneous Coronary Intervention (PCI)

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

경청해주셔서 감사합니다.