Virtual Congress

The Korean Society of Radiology

KSSR 2021

제8차 대한영상의학회
춘계종합심포지엄

The 8th Korean Spring Symposium of Radiology

2021.6.24 목 – 2021.6.25 금

Improving Collaboration, Quality, and Life

Clinical Research Methodology Course
-Intermediate Course

www.kssr.kr

제8차 대한영상의학회 춘계종합심포지엄

Clinical Research Methodology Course - Intermediate Course

(6월 24일 목요일 09:00~17:00)

Room 2

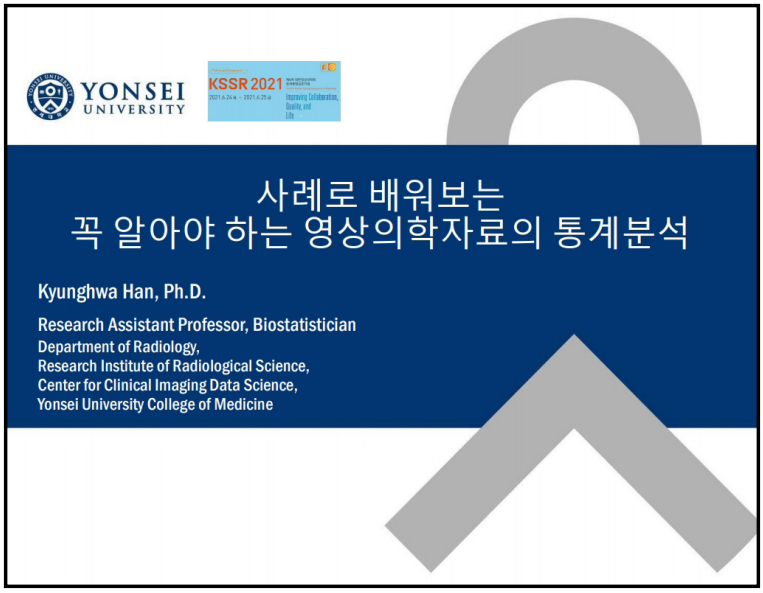
09:00-09:50	사례로 배워보는 꼭 알아야 하는 영상의학자료의 통계분석	한경화 (연세대학교)	1
09:50-10:00	Q&A; Break		
10:00-10:50	Statistical modeling for continuous outcome	송기준 (연세대학교)	32
10:50-11:00	Q&A; Break		
11:00-11:50	Statistical modeling for binary outcome	송기준 (연세대학교)	44
11:50-12:00	Q&A; Break		
12:00-13:10	점심식사		
13:10-14:20	Fundamentals of survival analysis	김선옥 (서울아산병원)	54
14:20-14:30	Q&A; Break		
14:30-15:20	How to construct a prediction model	한경화 (연세대학교)	98
15:20-15:50	How to validate and report a prediction model	한경화 (연세대학교)	122
15:50-16:00	Q&A; Break		
16:00-16:40	Noninferiority testing in radiology research	안소연 (분당서울대학교병원)	137
16:40-17:00	Q&A		

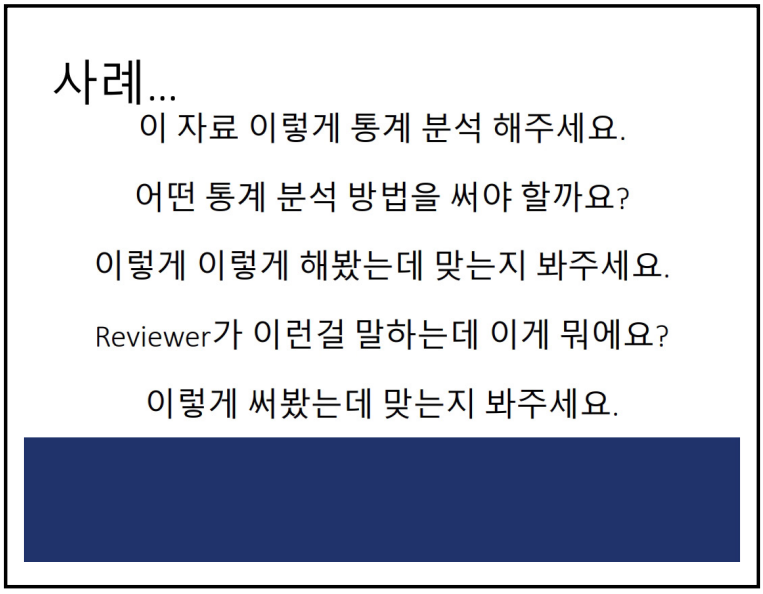
Clinical Research Methodology Course – Intermediate Course

09:00–09:50 Room 2

사례로 배워보는 꼭 알아야 하는 영상의학자료의 통계분석

한 경 화
연세대학교





YONSEI UNIVERSITY

Statistical Methods

KSSR 2021

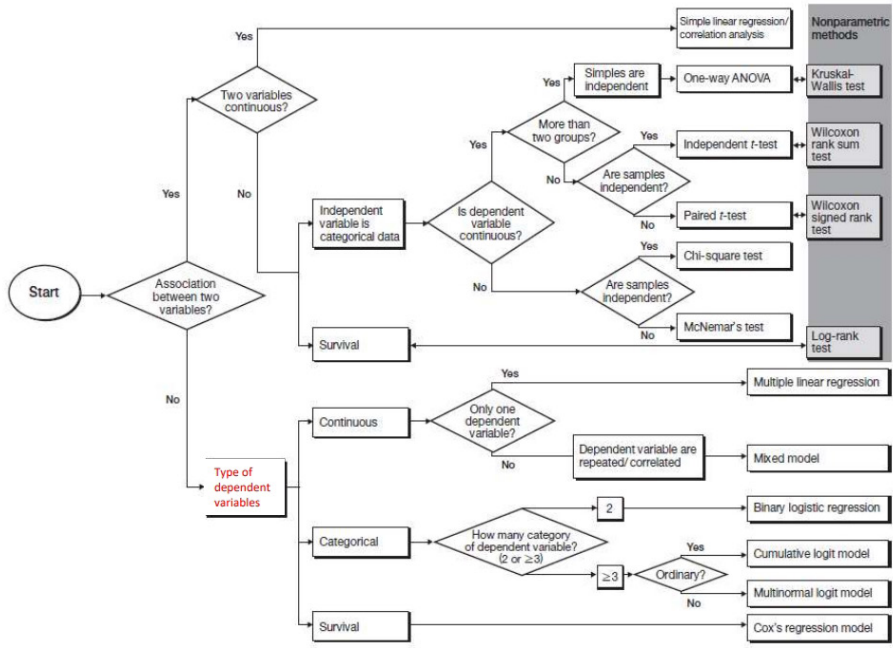


Figure 1. A flowchart of appropriate tests of statistical inference.

***JKMA* 2012 55(6): 573-581 (Korean)**

YONSEI UNIVERSITY

KSSR 2021

연구 흐름

- 연구 목적 (가설) 설정
- Eligibility criteria 설정
- 대상자 수 선정
- 조사할 변수 선정
- 자료 수집
- 통계 분석
- 논문 쓰기
- Submit and reject……Revision Revision…….
- Publish!!

4

YONSEI UNIVERSITY

(영상의학) 연구에서의 통계분석들

- Group 에 따른 연속형/범주형 변수 비교
- 측정치에 대한 평가자, 평가방법에 따른 일치도
- 진단 결과의 정확도
- 질병 발생에 대한 위험 인자 탐색
- (진단, 예후에 대한) 예측 모형
- Multiple radiologists
- Multiple lesions per patient

Technical validity
↓
Clinical validation
↓
Clinical utility

- 메타분석, 비용효과 분석 등

- Radiomics 연구
- 인공지능 성능 평가

5

YONSEI UNIVERSITY

무심코 지나치는.. 하지만 꼭 확인할 요소들

✓Sample size
✓Clustered data
: Multiple radiologists/Multiple lesions per patient
✓Multiple comparisons
✓Biases in Diagnostic test accuracy studies

EDITORIAL

Submissions to *Radiology*: Our Top 10 List of Statistical Errors[1]

Radiology

6

Sample size and Power
- Hypothesis testing -

YONSEI UNIVERSITY

KSSR 2021

Sample Size Estimation

- 왜 Sample size estimation이 필요한가?
  - (전향적 무작위 배정 연구에서) 제1종 오류 통제와 ***적절한 검정력***을 가지면서, 가설 검정을 수행하기 위해서!!
- 후향적 연구에서도 필요한가?
  - 반드시 필요한 것은 아니지만 연구에서 예상한 결과가 나오지 않을 때 sample이 너무 작아서 나오지 않는지 확인할 필요가 있고 때때로 reviewer가 이러한 이유로 power을 제시하라고 요구하기도 함.
  - 자료 구축에 들이는 시간 및 노력을 예상해볼 수 있음.
- 확증적 연구 vs. 탐색적 연구

8

YONSEI UNIVERSITY

Hematocrit Level and T1 Mapping Parameters in Pre- and Post-anemia Rat Models

Parameter	Preanemia Group (n = 13)	Postanemia Group (n = 13)	P Value
Hematocrit level (%)	59.0 ± 4.1	45.7 ± 5.2	<.01
Native T1 (msec)			
Myocardium	1186.7 ± 55.6	1174.8 ± 59.3	.47
Blood	1992.0 ± 214.9	2193.7 ± 334.4	<.01
Postcontrast T1 (msec)			
Myocardium	829.5 ± 80.7	794.3 ± 119.8	.08
Blood	690.2 ± 109.7	563.8 ± 155.7	<.01
ΔR1			
Myocardium	0.000374	0.000435	.12
Blood	0.000984	0.001465	<.01
Partition coefficient	38.2 ± 4.4	29.2 ± 3.5	<.01
ECV (%)	15.5 ± 2.0	16.0 ± 1.9	.24

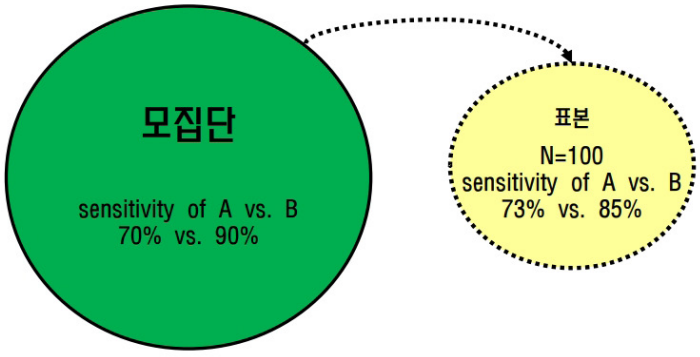
Note.—Data are mean ± standard deviation. ECV = extracellular volume fraction, R1 = 1/T1, $\Delta R1 = R1_{contrast\text{-}enhanced} - R1_{unenhanced}$.

Kim PK, et al. "Myocardial extracellular volume fraction and change in hematocrit level: MR evaluation by using T1 mapping in an experimental model of anemia." *Radiology* 288.1 (2018): 93-98.

9

YONSEI UNIVERSITY

Conceptual Summary of Hypothesis Testing



10

YONSEI UNIVERSITY KSSR 2021

Conceptual Summary of Hypothesis Testing

- Null hypothesis (H_0) vs. Alternative hypothesis (H_1)

Decision \ True	H_0 : True	H_1 : True
Fail to reject H_0	1-α	Type II error
Accept H_1	Type I error =α	power =1-β

11

YONSEI UNIVERSITY KSSR 2021

Conceptual Summary of Hypothesis Testing

- Sensitivity의 차이가 실제로 20%일 때에도 (H_0 true) 우연히 표본에서의 차이는 그보다 작거나 클 수 있고, 그 반대의 경우 (H_1 true) 도 가능하다.
- **P value**
  - Observed type I error
  - 귀무가설(H_0)이 맞다는 전제 하에, 실제로 주어진 자료로부터 계산된 검정 통계량 값보다 더욱 "극단적인 값"을 얻을 확률
  - 통계학적으로 정의되는 분포에 기반
    - Binomial, χ^2, t, F, Normal distribution,...

12

YONSEI UNIVERSITY

Sample size estimation

- 꼭 필요한 사항
  1. 일차 연구목적 (primary endpoint)에 따른 통계 분석 기법
     ex) 평균비교: t-test, 비율비교: Chi-square test
  2. 연구자가 밝히고자 하는 최소 유의한 차의 정도 및 분포
  3. Significance level: 일반적으로 0.05 적용
  4. Statistical power: 일반적으로 80% or 90% 로 가정

- 기타사항: Allocation ratio, Drop-out rate

$$N * = \frac{N}{1 - (\text{drop-out rate})\}}$$

13

YONSEI UNIVERSITY

Ex

4.2.1 Test for Equality

To test whether there is a difference between the mean response rates of the test drug and the reference drug, the following hypotheses are usually considered:

$$H_0 : \epsilon = 0 \quad \text{versus} \quad H_a : \epsilon \neq 0.$$

We reject the null hypothesis at the α level of significance if

type I error

$$\left| \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}} \right| > z_{\alpha/2}. \qquad (4.2.1)$$

Under the alternative hypothesis that $\epsilon \neq 0$, the power of the above test is approximately

$$\Phi\left(\frac{|\epsilon|}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}} - z_{\alpha/2} \right).$$

As a result, the sample size needed for achieving a power of $1 - \beta$ can be obtained by the following equation:

power

$$\frac{|\epsilon|}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}} - z_{\alpha/2} = z_\beta.$$

This leads to

sample size

$$n_1 = \kappa n_2$$
$$n_2 = \frac{(z_{\alpha/2} + z_\beta)^2}{\epsilon^2} \left[\frac{p_1(1-p_1)}{\kappa} + p_2(1-p_2) \right]. \qquad (4.2.2)$$

Chow, S. C., Wang, H., & Shao, J. (2007). *Sample size calculations in clinical research*. CRC press.

14

YONSEI UNIVERSITY KSSR 2021

Hematocrit Level and T1 Mapping Parameters in Pre- and Post-anemia Rat Models

Parameter	Preanemia Group (n = 13)	Postanemia Group (n = 13)	P Value
Hematocrit level (%)	59.0 ± 4.1	45.7 ± 5.2	<.01
Native T1 (msec)			
Myocardium	1186.7 ± 55.6	1174.8 ± 59.3	.47
Blood	1992.0 ± 214.9	2193.7 ± 334.4	<.01
Postcontrast T1 (msec)			
Myocardium	829.5 ± 80.7	794.3 ± 119.8	.08
Blood	690.2 ± 109.7	563.8 ± 155.7	<.01
ΔR1			
Myocardium	0.000374	0.000435	.12
Blood	0.000984	0.001465	<.01
Partition coefficient	38.2 ± 4.4	29.2 ± 3.5	<.01
ECV (%)	15.5 ± 2.0	16.0 ± 1.9	.24

Note.—Data are mean ± standard deviation. ECV = extracellular volume fraction, R1 = 1/T1, $\Delta R1 = R1_{contrast\text{-}enhanced} - R1_{unenhanced}$.

Kim PK, et al. "Myocardial extracellular volume fraction and change in hematocrit level: MR evaluation by using T1 mapping in an experimental model of anemia." *Radiology* 288.1 (2018): 93-98.

15

YONSEI UNIVERSITY KSSR 2021

Reviewer #4/Statistical Reviewer:

*1. As in any study with non-significant results for the primary hypothesis, a power analysis is required. What was the magnitude of the difference in ECV that would have been detectable with 80% power? The clinical significance of differences smaller than that must be considered.

- Paired t-test: $t = \frac{mean\ difference}{\frac{standard\ deviation\ of\ difference}{\sqrt{n}}}$

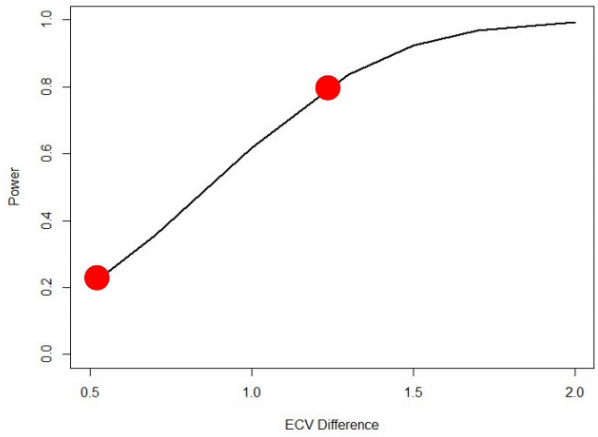
- **power analysis 결과)**
  - 논문 결과에 대한 statistical power = 20.44%
  - power 80%로 논문 결과를 보이기 위해 필요한 rat수 = 70 마리
  - 13마리로 power 80%를 얻을 수 있는 ECV 차이값 = 1.24 (※ 다만 이 결과는 ECV difference의 standard deviation은 현재 연구 자료와 같다는 가정하에 구한 결과입니다.)

Kim PK, et al. "Myocardial extracellular volume fraction and change in hematocrit level: MR evaluation by using T1 mapping in an experimental model of anemia." *Radiology* 288.1 (2018): 93-98.

16

YONSEI UNIVERSITY

N=13, alpha = 0.05)



17

YONSEI UNIVERSITY

Statistical Analyses

Continuous variables are expressed as mean ± standard deviation, and categorical variables are expressed as a frequency or percentage. The Shapiro-Wilk test was performed to evaluate the distribution of data. The significance of differences in hematocrit levels, LV functional parameters, and T1 values between the pre- and postanemia models was evaluated with the paired *t* test. A post hoc power analysis was performed to determine the difference between the pre- and postanemia models that was needed to achieve 80% power with our sample size. A linear mixed model with restricted maximum likelihood

Discussion

The ECV difference between the pre- and postanemia groups was insignificant. The difference between these two groups was 0.5, which was less than 1.24. This difference may be clinically insignificant, given that the width of a 95% CI for repeated measurement of ECV was 2.8% in the previous human studies (25,31).

There were a few limitations to this study. First, this study had a small sample size. To achieve a higher power, we would need to perform a study with a larger sample size. However, this was ultimately an exploratory study performed in animal subjects, and we were still able to obtain results with a small sample size. Second, we obtained measurements at only one time point

Kim PK, et al. "Myocardial extracellular volume fraction and change in hematocrit level: MR evaluation by using T1 mapping in an experimental model of anemia." *Radiology* 288.1 (2018): 93-98.

18

Because the association between kurtosis with SSF of 2 and outcome was significant not in the overall population but rather only in the non–triple-negative breast cancer population, we performed a post hoc power analysis to estimate the power to detect a significant association in the overall population. Considering the observed effect size (median difference, 0.6) standard deviation and sample sizes in the two outcome groups, the estimated post hoc power was 25%.

KSSR 2021

Table 2

Comparison of BFVs between Children and Young Adults

Variable	Preprandial BFV (L/min/m²) ≤18 Years Old (n = 19)	Preprandial BFV (L/min/m²) >18 Years Old (n = 20)	Preprandial BFV (L/min/m²) P Value*	Postprandial BFV (L/min/m²) ≤18 Years Old (n = 19)	Postprandial BFV (L/min/m²) >18 Years Old (n = 20)	Postprandial BFV (L/min/m²) P Value*
Cardiac output	3.92 (2.48–6.0)	3.1 (2.56–4.0)	.014 (0.93)†	4.04 (3.0–5.96)	3.46 (2.7–4.56)	.029 (0.90)†
Abdominal BFV	1.54 (1.18–2.39)	1.42 (1.19–1.89)	>.99 (0.45)	1.87 (1.37–2.54)	1.86 (1.27–2.44)	>.99 (0.16)
Cerebral blood flow	1.43 (1.03–2.30)	1.10 (0.70–1.69)	.039 (0.93)†	1.39 (0.91–2.29)	1.12 (0.88–1.59)	.008 (0.97)†
SMA	0.23 (0.17–0.30)	0.22 (0.13–0.48)	>.99 (0.10)	0.70 (0.46–1.15)	0.68 (0.31–0.94)	>.99 (0.22)
SMV	0.24 (0.13–0.36)	0.23 (0.15–0.38)	>.99 (0.06)	0.71 (0.34–1.25)	0.68 (0.37–1.04)	>.99 (0.07)
Portal vein	0.66 (0.36–0.86)	0.63 (0.45–0.85)	>.99 (0.04)	1.30 (0.85–1.74)	1.16 (0.96–1.67)	>.99 (0.08)
IVC						
Bifurcation	0.86 (0.33–1.61)	0.58 (0.39–1.17)	.14 (0.79)	0.75 (0.41–1.31)	0.58 (0.40–0.91)	.75 (0.64)
Suprarenal	1.49 (0.78–2.17)	1.23 (0.87–1.84)	.19 (0.75)	1.42 (1.0–2.41)	1.23 (0.97–1.75)	.57 (0.67)
Posthepatic	2.24 (1.63–4.22)	2.15 (1.6–2.65)	.70 (0.71)	2.66 (2.04–4.0)	2.38 (1.83–3.09)	>.99 (0.39)
Descending aorta						
Diaphragm	2.32 (1.45–3.7)	1.97 (1.60–2.42)	.075 (0.84)	2.65 (2.13–3.67)	2.36 (1.73–2.97)	.57 (0.66)
Bifurcation	0.86 (0.38–1.52)	0.56 (0.29–0.97)	.22 (0.79)	0.62 (0.42–1.27)	0.49 (0.32–1.02)	.13 (0.74)
Ascending aorta	3.61 (2.39–5.65)	3.02 (2.49–4.13)	.123 (0.81)	3.83 (2.8–5.72)	3.39 (2.72–4.93)	.123 (0.71)
Right common carotid artery	0.36 (0.23–0.51)	0.27 (0.20–0.32)	.002 (1.0)†	0.35 (0.24–0.55)	0.27 (0.19–0.35)	.009 (0.97)†
Left common carotid artery	0.37 (0.25–0.56)	0.28 (0.17–0.35)	.004 (0.99)†	0.39 (0.25–0.61)	0.28 (0.18–0.38)	.009 (0.97)†
Right internal jugular vein	0.32 (0.08–0.81)	0.28 (0.02–0.46)	>.99 (0.53)	0.40 (0.08–0.8)	0.28 (0.01–0.43)	.212 (0.77)
Left internal jugular vein	0.30 (0.01–0.76)	0.15 (0.05–0.4)	.212 (0.74)	0.26 (0.01–0.64)	0.15 (0.06–0.37)	.619 (0.62)
Hepatic vein	0.77 (0.5–1.85)	0.81 (0.45–1.1)	>.99 (0.20)	1.05 (0.83–1.34)	1.12 (0.64–1.51)	>.99 (0.06)
Hepatic artery	0.22 (−0.27 to 1.07)	0.20 (−0.20 to 0.47)	>.99 (0.20)	−0.14 (−0.51 to 0.47)	−0.11 (−0.49 to 0.36)	>.99 (0.11)
Renal	0.60 (0.38–1.27)	0.59 (0.36–0.98)	>.99 (0.06)	0.75 (0.45–1.19)	0.67 (0.45–1.17)	>.99 (0.23)

Note.—Unless otherwise indicated, data are medians, with the range in parentheses. Cardiac output is a sum of SVC flow and descending aorta at diaphragm. CBF is estimated from SVC flow, assuming negligible return from upper limbs at rest. Abdominal BFV here is a mean of methods 1 and 2.

* Data in parentheses are the power.

† Indicates a significant difference.

Despite our large sample size, our study was underpowered to show significant differences in FN rates owing to the infrequent occurrence of this adverse event. A post hoc power analysis based on the FN rate in our study sample calculated that in order to observe a proportional difference of 0.1 per 1000 screens at 80% power, the estimated sample size would need to be 2 278 662

- Chamming's, Foucauld, et al. "Features from computerized texture analysis of breast cancers at pretreatment MR imaging are associated with response to neoadjuvant chemotherapy." *Radiology* 286.2 (2018): 412-420.
- Muthusami, Prakash, et al. "Splanchnic, thoracoabdominal, and cerebral blood flow volumes in healthy children and young adults in fasting and postprandial states: determining reference ranges by using phase-contrast MR imaging." *Radiology* 285.1 (2017): 231-241.
- Durand, Melissa A., et al. "False-Negative Rates of Breast Cancer Screening with and without Digital Breast Tomosynthesis." *Radiology* (2020): 202858.

19

YONSEI UNIVERSITY

KSSR 2021

In reviewer's comment…

Please perform sample size calculation. This study is underpowered and will be impossible to make any inference

A Proposal to Mitigate the Consequences of Type 2 Error in Surgical Science

?

Yanik J. Bababekov, MD, MPH, Sahael M. Stapleton, MD, Jessica L. Mueller, BA, Zhi Ven Fong, MD, MPH, and David C. Chang, PhD, MPH, MBA

Don't Calculate Post-hoc Power Using Observed Estimate of Effect Size
Gelman, Andrew
Annals of Surgery. 269(1):e9-e10, January 2019.
☆ Favorites PDF © Get Content & Permissions

Letter to Editor: A Proposal to Mitigate the Consequences of Type 2 Error in Surgical Science
Helminen, Olli; Reito, Aleksi
Annals of Surgery. 269(1):e10-e11, January 2019.
☆ Favorites PDF © Get Content & Permissions

Post Hoc Power Calculation: Observing the Expected
Plate, Joost D. J.; Borggreve, Alicia S.; van Hillegersberg, Richard; More
Annals of Surgery. 269(1):e11, January 2019.
☆ Favorites PDF © Get Content & Permissions

Post Hoc Power: A Surgeon's First Assistant in Interpreting "Negative" Studies
Bababekov, Yanik J.; Chang, David C.
Annals of Surgery. 269(1):e11-e12, January 2019.

Comment on "Post-hoc Power: A Surgeons's First Assistant in Interpreting 'Negative' Studies" and "A Proposal to Mitigate the Consequences of Type 2 Error in Surgical Science"
Helminen, Olli; Reito, Aleksi
Annals of Surgery. 270(6):e77-e78, December 2019.
☆ Favorites PDF © Get Content & Permissions

Comment on "Post-hoc Power: If You Must, At Least Try to Understand"
Althouse, Andrew D.; Chow, Zad R.
Annals of Surgery. 270(6):e78-e79, December 2019.
☆ Favorites PDF © Get Content & Permissions

Response to Comment on "Misinterpretation of Results With P > 0.05 May Harm Quality and Patient Safety"
Bababekov, Yanik; Lee, Hang; Chang, David
Annals of Surgery. 270(6):e79-e80, December 2019.

20

YONSEI UNIVERSITY

Statistical vs. Clinical Significance

- N
- p value vs. 95% CI

21

P value
Uncertainty Metrics

YONSEI UNIVERSITY KSSR 2021

95% CI 꼭 적어야 하나요?

- Reviewer's comment

uncertainty measures are lacking, such as 95% CIs , for sensitivities, specificities, FOMs, etc.

- To quantify precision of the estimate

(e.g., Sn and Sp, odds ratio, hazard ratio,…)

23

YONSEI UNIVERSITY KSSR 2021

The advantage of CIs over significance tests (*P* values)

- the CIs shift the interpretation from a qualitative judgment about the role of chance to a quantitative estimation of the biologic measure of effect.
- Allow more reliable analysis, interpretation, and communication of clinical information among health care providers and between these providers and their patients.

Medina, L. Santiago, and David Zurakowski. "Measurement variability and confidence intervals in medicine: why should radiologists care?." *Radiology* 226.2 (2003): 297-301.

24

YONSEI UNIVERSITY

the P-value has nothing to do with the magnitude or the importance of an observed effect

- Two different results

	95% CI
OR = 0.59, p = 0.15	(0.2, 1.3)
OR = 0.83, p = 0.002	(0.7, 0.9)

- Due to a very large (small) sample size regardless of the effect size
- When there is a wide confidence interval that includes potentially important benefits or harms

Jung I. Some facts that you might be unaware of about the p-value. Arch Plast Surg 2017;44:93-4.

25

YONSEI UNIVERSITY

Uncertainty: SD, SE, 95% CI

- SD (Standard Deviation)
  각 관측치가 평균으로부터 떨어진 정도
- SE (Standard Error)
  표본으로부터 얻어진 통계량(평균, 비율, OR, HR 등등)이 모집단에서의 해당 통계량으로부터 떨어진 정도
- 95% CI (Confidence Interval)
  - CIs based on a 95% confidence level (=100 – type I error)
  - Sample size + Variability
  - 예: 평균의 95% CI
    - mean ± 1.96×SE = mean ± 1.96×SD/$\sqrt{N}$

26

YONSEI UNIVERSITY

KSSR 2021

95% CI for Odds Ratios

Imaging	Disease +	Disease -
+	a	b
-	c	d

- Odds Ratios: $OR = \frac{a/b}{c/d}$
- 95% CI for OR

$$\ln OR \pm 1.96 \times \sqrt{1/a + 1/b + 1/c + 1/d}$$

Q. 위 표에서 어느 한 경우의 sample size가 작다면 95% CI는?

A) 아주 넓은 CI가 산출됨 (ex, OR (95% CI): 2.5 (1.4, 225.399))

Q. 어느 한 경우는 자료에서 없는 경우 (0 cell) OR와 95% CI는?

A) CI가 <.001이나 >999.999 로 표시됨

Firth's correction, re-categorization,…

27

YONSEI UNIVERSITY

KSSR 2021

Table 4

Results of ROC Analysis for 2D Mammography versus Single-View Tomosynthesis for Average Diagnostic Accuracy

Variable	2D Mammography vs Single-View Tomosynthesis ForAll Readers	2D Mammography vs Single-View Tomosynthesis For Masses	2D Mammography vs Single-View Tomosynthesis For Calcification
2D mammography*	0.774	0.781	0.775
Single-view tomosynthesis*	0.775	0.788	0.742
2D mammography vs single-view tomosynthesis*	−0.001	−0.007	0.032
95% CI	−0.066, 0.064	+0.85, 0.070	−0.080, 0.145
P value	0.79	.85	.57

* Data are figures of merit.

✓95% CI for the difference

✓Multiple radiologist

Table 5

Results of ROC Analy… eader Experience

Variable	2D Mammography vs Two-View Tomosynthesis			2D Mammography vs Two-View Tomosynthesis For Masses			2D Mammography vs Two-View Tomosynthesis For Calcification		
	≥10*	<10*	<10*†	≥10*	<10*	<10*†	≥10*	<10*	<10*†
2D mammography‡	0.811	0.734	0.783	0.812	0.712	0.750	0.818	0.765	0.829
Two-view tomosynthesis‡	0.858	0.844	0.859	0.854	0.834	0.846	0.871	0.870	0.883
2D mammography vs two-view tomosynthesis‡	−0.047	−0.110	−0.076	−0.042	−0.122	−0.096	−0.053	−0.105	−0.054
95% CI	−0.135, 0.040	−0.204, −0.015	−0.121, −0.032	−0.142, 0.058	−0.201, −0.042	−0.161, −0.031	−0.155, 0.049	−0.259, 0.048	−0.142, 0.034
P value	0.25	0.03	0.002	0.38	0.008	0.007	0.29	0.14	

* Reader experience in years.
† The outlier was removed.
‡ Data are figures of merit.

28

Wallis MG., et al. "Two-view and single-view tomosynthesis versus full-field digital mammography: high-resolution X-ray imaging observer study." *Radiology* 262.3 (2012): 788-796.

YONSEI UNIVERSITY

95% CI for diagnostic accuracy

- 95% CI for accuracy (*p*: proportion)

$$p \pm 1.96 \times \sqrt{\frac{p(1-p)}{N}}$$

- CIs are needed to help one to be more certain about the clinical value of any screening or diagnostic test and to decide to what degree one can rely on the results.
- Sample size can be estimated to achieve a desired **CI width.**

29

YONSEI UNIVERSITY

Borderline p value

Hematocrit Level and T1 Mapping Parameters in Pre- and Post-anemia Rat Models

Parameter	Preanemia Group (n = 13)	Postanemia Group (n = 13)	P Value
Hematocrit level (%)	59.0 ± 4.1	45.7 ± 5.2	<.01
Native T1 (msec)			
Myocardium	1186.7 ± 55.6	1174.8 ± 59.3	.47
Blood	1992.0 ± 214.9	2193.7 ± 334.4	<.01
Postcontrast T1 (msec)			
Myocardium	829.5 ± 80.7	794.3 ± 119.8	.08
Blood	690.2 ± 109.7	563.8 ± 155.7	<.01
ΔR1			
Myocardium	0.000374	0.000435	.12
Blood	0.000984	0.001465	<.01
Partition coefficient	38.2 ± 4.4	29.2 ± 3.5	<.01
ECV (%)	15.5 ± 2.0	16.0 ± 1.9	.24

Note.—Data are mean ± standard deviation. ECV = extracellular volume fraction, R1 = 1/T1, $\Delta R1 = R1_{contrast\text{-}enhanced} - R1_{unenhanced}$.

Kim PK, et al. "Myocardial extracellular volume fraction and change in hematocrit level: MR evaluation by using T1 mapping in an experimental model of anemia." *Radiology* 288.1 (2018): 93-98.

30

YONSEI UNIVERSITY

KSSR 2021

- Statistically NS results
  - Negative?
  - Inconclusive?
  - Comparable?

  cf. noninferiority, equivalence

To make it seem more interesting.

better trends of improvement (p=0.056)
bordered on a statistically significant value (p=0.06)
bordered on being significant (p>0.07)
bordered on being statistically significant (p=0.0502)
bordered on but was not less than the accepted level of significance (p>0.05)
bordered on significant (p=0.09)
borderline conventional significance (p=0.051)
borderline level of statistical significance (p=0.053)
borderline significant (p=0.05)
borderline significant trends (p=0.099)
close to a marginally significant level (p=0.06)
close to being significant (p=0.06)
close to being statistically significant (p=0.055)
close to borderline significance (p=0.072)
close to the boundary of significance (p=0.06)
close to the level of significance (p=0.07)
close to the limit of significance (p=0.17)
close to the margin of significance (p=0.055)
close to the margin of statistical significance (p=0.075)
closely approaches the brink of significance (p=0.07)
closely approaches the statistical significance (p=0.0669)
closely approximating significance (p>0.05)
closely not significant (p=0.06)
closely significant (p=0.058)
close-to-significant (p=0.09)
did not achieve conventional threshold levels of statistical significance (p=0.08)
did not exceed the conventional level of statistical significance (p<0.08)
did not quite achieve acceptable levels of statistical significance (p=0.054)

https://mchankins.wordpress.com/2013/04/21/still-not-significant-2/

31

YONSEI UNIVERSITY

KSSR 2021

THE AMERICAN STATISTICIAN
2016, VOL. 70, NO. 2, 129–133
http://dx.doi.org/10.1080/00031305.2016.1154108

EDITORIAL

The ASA's Statement on *p*-Values: Context, Process, and Purpose

Q: Why do so many colleges and grad schools teach $p = 0.05$?
A: Because that's still what the scientific community and journal editors use.
Q: Why do so many people still use $p = 0.05$?
A: Because that's what they were taught in college or grad school.

32

Multiple Comparison

YONSEI UNIVERSITY

Multiple testing problem

- Multiple comparison (btw group)
- Multiplicity (multiple endpoints)
- 어떤 보정 방법?
- 어떤 변수 (또는 비교) 에 대해서 할까?
- 확증적 연구 vs. 탐색적 연구

34

YONSEI UNIVERSITY

KSSR 2021

Statistical Analyses

with cardiac MR imaging being the reference standard. Bonferroni correction was performed, and a P value of less than .0167 indicated a significant difference. For quantitative analysis of MDE, interobserver agreements were assessed

Table 1

Pairwise Comparison of Subjective Image Quality Score and Contrast-to-Noise Ratio between Conventional and Monochromatic CT

CT Examination	Image Quality	*P* Value*	CNR	*P* Value*
Conventional	3.15 ± 0.43	...	3.93 ± 1.33	...
60-keV	3.05 ± 0.39	.0455	3.61 ± 1.07	.1172
70-keV	3.38 ± 0.54	.0067	4.26 ± 1.38	.0047
80-keV	3.45 ± 0.55	.0005	4.10 ± 1.41	.0190

Note.—Unless otherwise indicated, data are mean ± standard deviation.

* Indicates pairwise comparison between conventional and monochromatic CT examinations. $P < .0167$ indicated a significant difference, with Bonferroni correction.

Chang S., Utility of Dual-Energy CT-based Monochromatic Imaging in the Assessment of Myocardial Delayed Enhancement in Patients with Cardiomyopathy. *Radiology* 287.2 (2018): 442-451

YONSEI UNIVERSITY

KSSR 2021

Statistical Analysis

Statistical analyses were performed with softwar Windows, version 20.0, IBM, Armonk, NY; an 3.5.1, R Foundation for Statistical Computing, V tria). The baseline characteristics and clinical outc three groups were compared. For the global diffe the three groups, all continuous variables (which w mally distributed) were analyzed with the Kruskal and categorical variables were analyzed with the Pearson χ^2 test or Fisher exact test. Treatment and outcomes were compared between groups, and group 1 was used as the reference. Multiplicity adjustments were not performed because our study was exploratory and did not have a confirmatory primary hypothesis for multiple end points. Multivariable logistic regression was performed to evaluate the statistical significance of the group factor for clinical outcome at 90 days with adjustment for onset to recanalization time. P values less than .05 were considered to indicate a statistically significant difference.

Table 2: Comparison of Treatment and Outcomes among Basilar Artery Occlusion Subtypes

Characteristic	Patients (n = 82)	Embolism without VA Steno-occlusion (group 1, n = 34)	Embolism from Tandem VA Steno-occlusion (group 2, n = 28)	In Situ Atherosclerotic Thrombosis (group 3, n = 20)	*P* Value (group 1 vs group 2)	*P* Value (group 1 vs group 3)
Onset to puncture time (min)*	218 (151–298)	218 (138–281)	225 (165–371)	240 (150–264)	.44	.96
Procedure time (min)*	60 (35–101)	49 (31–84)	66 (55–121)	63 (47–117)	.01	.06
Onset to recanalization time (min)*	286 (220–375)	260 (201–367)	296 (235–472)	290 (240–341)	.1	.54
First-line endovascular treatment	...	...	...	...	.14	.22
Stent retriever	53 (65)	26 (76)	15 (54)	12 (60)	...	...
Aspiration	22 (27)	6 (18)	10 (36)	6 (30)	...	...
Angioplasty	1 (1)	0 (0)	0 (0)	1 (5)†	...	...
Rescue treatment	12 (15)	5 (15)	5 (18)	2 (10)	.74	>.99
Adjuvant treatment angioplasty	15 (18)	0 (0)	8 (29)	7 (35)	.03	.04
Intra-arterial tirofiban	12 (15)	2 (6)	4 (14)	6 (30)	...	...
Intra-arterial urokinase	22 (27)	8 (24)	7 (25)	7 (35)	...	...
mTICI grade 2b or 3	64 (78)	29 (85)	24 (86)	11 (55)	>.99	.01
mRS score at 90 d*	4 (1–5)	3 (0–5)	5 (2–5)	5 (2–5)	.03	.05
mRS score 0–2 at 90 d	30 (37)	18 (53)	8 (29)	4 (20)	.05	.02
Mortality at 90 d	17 (21)	6 (18)	7 (25)	4 (20)	.48	>.99
Any hemorrhagic complication	13 (16)	5 (15)	5 (18)	3 (15)	.74	>.99
Symptomatic intracerebral hemorrhage	5 (6)	2 (6)	1 (4)	2 (10)	>.99	.62
Subarachnoid hemorrhage	2 (2)	1 (3)	0 (0)	1 (5)	>.99	>.99
Parenchymal hemorrhage type 1	2 (2)	1 (3)	1 (4)	0 (0)	>.99	>.99
Parenchymal hemorrhage type 2	1 (1)	0 (0)	0 (0)	1 (5)	...	.37

Note—Unless otherwise indicated, data are numbers of patients, with percentages in parentheses. VA = vertebral artery, mRS = modified Rankin Scale, mTICI = modified Thrombolysis in Cerebral Infarction scale.

* Data are medians, with interquartile ranges in parentheses.

† Primary angioplasty was performed in a patient without mechanical thrombectomy attempts.

Baik SH et al. Mechanical Thrombectomy in Subtypes of Basilar Artery Occlusion: Relationship to Recanalization Rate and Clinical Outcome. *Radiology* 291.3 (2019): 730-737.

YONSEI UNIVERSITY

Table 3 Univariate analysis of the impact of MRI radiomic features on overall survival

Feature Category	Feature Name	HR (95% CI)	*P*-value
Haralick	Entropy	0.33 (0.17, 0.64)	**0.001***
LoG	Mean	0.68 (0.54, 0.85)	**0.001***
LoG	Standard deviation	0.76 (0.65, 0.89)	**0.001***
Gabor	Mean	0.41 (0.24, 0.68)	**0.001***
Sobel	Mean	0.89 (0.83, 0.96)	**0.002***
Sobel	Standard deviation	0.86 (0.78, 0.95)	**0.004***
First order	Mean	0.93 (0.88, 0.98)	**0.005***
Gabor	Kurtosis	0.8 (0.68, 0.94)	**0.007***
Haralick	Contrast	0.92 (0.87, 0.99)	**0.016***
First order	Standard deviation	0.97 (0.95, 1.00)	**0.019***
Gabor	Standard deviation	0.95 (0.9, 0.99)	**0.026***
First order	Skewness	0.99 (0.98, 1.00)	0.0288
LoG	Kurtosis	0.68 (0.48, 0.96)	0.0301
LoG	Skewness	0.93 (0.86, 1.0)	0.036
Haralick	Energy	0.91 (0.83, 1.0)	0.04
Sobel	Kurtosis	0.87 (0.75, 1.0)	0.054
Sobel	Skewness	0.64 (0.39, 1.04)	0.073
Haralick	Correlation	0.9 (0.79, 1.03)	0.13
Haralick	Homogeneity	0.94 (0.87, 1.02)	0.139
First order	Kurtosis	1.02 (0.98, 1.06)	0.378
Gabor	Skewness	1.0 (0.96, 1.06)	0.862

*Significant after FDR correction.

Bhatia et al. MRI radiomic features are associated with survival in melanoma brain metastases treated with immune checkpoint inhibitors. *Neuro-Oncology* 2020

37

YONSEI UNIVERSITY

In high-dimensional data

- 무수히 많은 검정 수행
- Bonferroni correction? 하고 나면 논문을 못..
- Alternative correction method
  - Holm-Bonferroni's step-down
  - Hochberg's step-up
  - Step-wise correction (Fixed-sequence, fallback)
  - FDR correction
  - ...

38

DIAGNOSTIC PERFORMANCE

YONSEI UNIVERSITY

KSSR 2021

On 2 X 2contingency table

test / standard ref.	D^+	D^-	Total
T^+	a TP	b FP	a+b
T^-	c FN	d TN	c+d
Total	a+c	b+d	N

40

YONSEI UNIVERSITY

Comparison with the benchmark

- If a reference standard is available: Sn, Sp
- If a reference standard is available, but impractical: bias corrected Sn, Sp
- If a reference standard is not available or unacceptable for your particular intended use and/or intended use population: consider whether one can be constructed.
- If a reference standard is not available and cannot be constructed: NOT accuracy, agreement

41

FDA. Statistical guidance on reporting results from studies evaluating diagnostic tests. 2011

Biases in Diagnostic Accuracy Studies

YONSEI UNIVERSITY

KSSR 2021

Example

- CT에서 평가한 소견으로 두 질환을 구분하고자 하는데 한쪽이 상대적으로 드뭅니다.
- A질환 90명, B질환 10명입니다.
- 즉 CT를 안보고 전부 A질환이라고 해도 90/100=90%의 PPV for A가 예상됩니다.
- Matching을 해서 연구 해야 할까요?
- 실제 prevalence를 반영하는 게 더 좋지 않을까요?

43

YONSEI UNIVERSITY

KSSR 2021

Predictive Values (PPV, NPV)

- Population/prevalence dependent value

TABLE 3 Disease Prevalence 50%

Index Test	Reference Text		Row Total
	Positive	Negative	
Positive	90	10	100
Negative	10	90	100
Column total	100	100	200

Note.—Diagnostic test with 90% sensitivity, specificity, and positive and negative predictive values. Prevalence of disease is a relatively high 50%.

- Sn, Sp = 90%
- PPV = 90%
- NPV = 90%

TABLE 4 Disease Prevalence 1%

Index Test	Reference Text		Row Total
	Positive	Negative	
Positive	9	99	108
Negative	1	891	892
Column total	10	990	1000

Note.—Decreasing the disease prevalence to 1% leaves the sensitivity and specificity at 90%; however, the positive predictive value has decreased to 8% and the negative predictive value has increased to 99.9%.

- Sn, Sp = 90%
- PPV = 8%
- NPV = 99.9%

44

YONSEI UNIVERSITY

Predictive Values (PPV, NPV)

- PPV나 NPV의 오용
  - 환자 대조군 연구

Population: $PPV = \frac{a}{a+b}$

진단 / 질병	D^+	D^-
T^+	a	b
T^-	c	d

Sample: $PPV = \frac{f_1 a}{f_1 a + f_2 b}$

진단 / 질병	D^+	D^-
T^+	f_1a	f_2b
T^-	f_1c	f_2d

- f_1: 환자군에서의 sampling rate
- f_2: 대조군에서의 sampling rate
- 대부분 $f_1 \neq f_2$ 이므로 population과 sample의 PPV가 다름

45

YONSEI UNIVERSITY

Corrected Predictive Values (PPV, NPV)

$$\widehat{PPV} = \frac{p \times \mathrm{Sn}}{p \times \mathrm{Sn} + (1-p) \times (1-\mathrm{Sp})}$$

$$\widehat{NPV} = \frac{(1-p) \times \mathrm{Sp}}{(1-p) \times \mathrm{Sp} + p \times (1-\mathrm{Sn})}$$

Where, p: prevalence or pretest probability

Pretest probability: based on the patient's previous medical history, previous and recent exposures, current signs and symptoms, and results of other screening and diagnostic tests performed.

Weinstein S et al. Clinical Evaluation of Diagnostic Tests. *AJR* 2005;184:14–19
Halpern EF, Gazelle GS. Probability in Radiology. *Radiology 2003;*226(1):12–15

46

YONSEI UNIVERSITY

KSSR 2021

Table 2

Sensitivity, Specificity, PPV, and NPV for Detecting Knee Instability at Arthrometric Testing with Various MR Imaging Signs of Knee Instability

MR Imaging Sign and Observer	Sensitivity (%)	Specificity (%)	PPV (%)*	NPV (%)
Buckling of PCL				
Observer 1	50 (19, 81)	49 (32, 65)	20 (7, 41)	79 (58, 93)
Observer 2	50 (19, 81)	46 (30, 63)	19 (7, 39)	78 (56, 93)
PCL line				
Observer 1	10 (0, 45)	82 (66, 92)	12 (0, 53)	78 (62, 89)
Observer 2	10 (0, 45)	80 (64, 91)	11 (0, 48)	78 (62, 89)
Posterior femoral line				
Observer 1	0 (0, 31)	100 (91, 100)	NA	80 (66, 90)
Observer 2	0 (0, 31)	100 (91, 100)	NA	80 (66, 90)
PCL angle <100°				
Observer 1	10 (0, 45)	87 (73, 96)	17 (0, 64)	79 (64, 90)
Observer 2	10 (0, 45)	87 (73, 96)	17 (0, 64)	79 (64, 90)
PCL angle <107°				
Observer 1	30 (7, 65)	72 (55, 85)	21 (5, 51)	80 (63, 92)
Observer 2	50 (19, 81)	62 (45, 77)	25 (9, 49)	83 (64, 94)
PCL curvature ratio				
Observer 1	10 (0, 45)	87 (73, 96)	17 (0, 64)	79 (64, 90)
Observer 2	10 (0, 45)	82 (66, 92)	12 (0, 53)	78 (62, 89)
Uncovering of the posterior horn of the lateral meniscus				
Observer 1	0 (0, 31)	97 (87, 100)	0.0 (0, 98)	79 (65, 90)
Observer 2	0 (0, 31)	100 (91, 100)	NA	90 (66, 90)

Note.—Numbers in parentheses are 95% confidence intervals.

* NA = not applicable; the PPV is undefined because the denominator (true positive + false positive findings) is zero.

Discussion

However, the low prevalence will certainly affect the PPV of these MR imaging signs.

Conclusion: Although MR imaging signs of knee laxity in the presence of an intact ACL graft have a high specificity, the low PPV means that MR imaging is of little value in predicting anterior knee laxity as demonstrated with mechanical testing.

Naraghi AM et al. Anterior cruciate ligament reconstruction: MR imaging signs of anterior knee laxity in the presence of an intact graft. *Radiology* 2012;263(3):802–810.

47

YONSEI UNIVERSITY

KSSR 2021

Comment on the Correct Use of Predictive Values for Evaluating Diagnostic Tests

From
Fang-yao Chen, MS,* Christine P. Yang, PhD,† and Ping-yan Chen, MD*
Department of Biostatistics, School of Public Health and Tropical Medicine, Southern Medical University, Guangzhou, Guangdong, China 510515*

Therefore, it is plausible to pay special attention to this problem. We suggest that *(a)* when a test is used for screening the general population, neither PPV nor NPV can be estimated if the prevalence is not known, and *(b)* when a test is applied for diagnostic purposes in a specific population of human beings, such as patients in a clinical setting, both PPV and NPV should be calculated on the basis of the known prevalence even if it is very difficult to obtain (to obtain a precise prevalence, the subjects should be sampled randomly from the target population). The authors should explain how to select the subjects and define the prevalence, and the conclusion is therefore only eligible for the target population. In addition, if predictive values generated from different diagnostic tests are compared, these tests should have the same prevalence. In conclusion, predictive values for evaluating the accuracy of diagnostic tests should be applied cautiously to avoid any misleading reports.

Editors' Note

From
Deborah Levine, MD, Senior Deputy Editor
Elk F. Halpern, PhD, Consultant to the Editor
Herbert Y. Kressel, MD, Editor

We thank Dr Chen and colleagues for raising this important point. The PPV was originally defined as the value of a single positive diagnostic test in an unselected population (1). It applies the sensitivity and specificity of a test to the general population in which the test was performed. Both PPVs and NPVs of a diagnostic test depend on the disease prevalence. Many of the studies reported in the imaging literature have selection or inclusion bias in the study populations and/or have artificially enriched populations, and, as Dr Chen and colleagues note, the PPV generated in these types of studies cannot be compared with those of other studies with differing disease prevalence. In such

Radiology 266.1 (2013): 364-366.

48

YONSEI UNIVERSITY

KSSR 2021

Table 4. Diagnostic Performance of CT for Detection of Mitral Paravalvular Leakage Using Surgical Findings as the Standard Reference: Comparison With TTE and TEE

	TP	TN	FP	FN	Sensitivity, %	Specificity, %	PPV, %	NPV, %	Accuracy, %
CT	31	45	1	1	96.9 (31/32)	97.8 (45/46)	96.9 (31/32)	97.8 (45/46)	97.4 (76/78)
TTE	26	43	2	6	81.3 (26/32)	95.6 (43/45)	92.9 (26/28)	87.8 (43/49)	89.6 (69/77)
TEE	25	23	1	1	96.2 (25/26)	95.8 (23/24)	96.2 (25/26)	95.8 (23/24)	96.0 (48/50)
P value (CT and TTE)					0.086	0.558	0.479	0.089	0.073
P value (CT and TEE)					0.884	0.647	0.879	0.637	0.658
P value (TTE and TEE)					0.065	0.929	0.362	0.207	0.110

CT indicates computed tomography; FN, false-negative; FP, false-positive; NPV, negative predictive value; PPV, positive predictive value; TEE, transesophageal echocardiography; TN; true-negative; TP, true-positive; and TTE, transthoracic echocardiography.

Suh YJ et al. Assessment of Mitral Paravalvular Leakage After Mitral Valve Replacement Using Cardiac Computed Tomography: Comparison With Surgical Findings. *Circulation: Cardiovascular Imaging*, 2016 *9*(6), e004153.

49

YONSEI UNIVERSITY

KSSR 2021

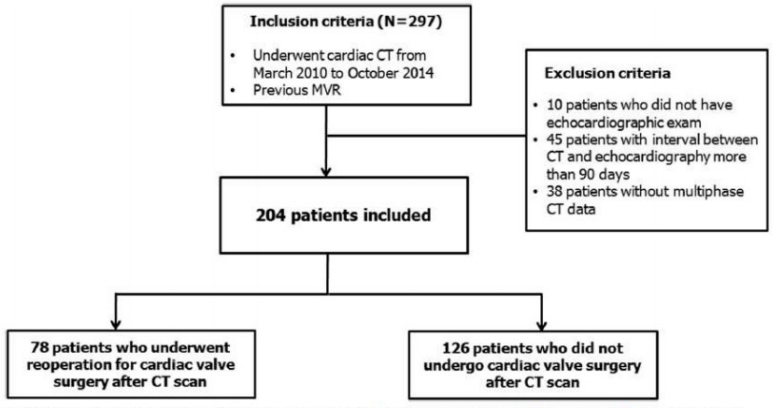


Figure 1. Flow chart of the study population. CT indicates computed tomography; and MVR, mitral valve replacement.

- Reviewer's comment:
  we would be interested in considering a new manuscript if you are able to expand the study cohort and include the patients who did not undergo re-do surgery.

Suh YJ et al. Assessment of Mitral Paravalvular Leakage After Mitral Valve Replacement Using Cardiac Computed Tomography: Comparison With Surgical Findings. *Circulation: Cardiovascular Imaging*, 2016 *9*(6), e004153.

50

YONSEI UNIVERSITY

KSSR 2021

Verification Bias:

An Underrecognized Source of Error in Assessing the Efficacy of Medical Imaging

Jonelle M. Petscavage, MD, MPH, Michael L. Richardson, MD, Robert B. Carr, MD

Acad Radiol 2011; 18:343–346

TABLE 1. Frequency Table of Data Collected from Original Research Articles in Four Journals, November 2006–October 2009

	American Journal of Roentgenology	Academic Radiology	Radiology	European Journal of Radiology	All
Original research articles	1,004	422	1,043	500	2,969
Sensitivity and specificity listed as study end point (%)	24.7	19.2	24.9	37.4	26.1
Potential verification bias (%)	36.4	23.4	29.5	13.4	27.2
Bias acknowledged in discussion (%)	4.4	26.3	28.9	20	17.1

51

YONSEI UNIVERSITY

KSSR 2021

Verification bias

- Work up bias
- Referral bias
- Sometimes it is not feasible to obtain disease status verification for all study subjects-costly or invasive reference test.
- High risk subjects may be more likely to have disease status assessed.
- Analysis of only those with disease ascertainment can result in biased estimates of the diagnostic test accuracy.
- Missing data exist in reference test

52

YONSEI UNIVERSITY

Bias correction methods (1)

<Disease Verification is obtained for everyone>

V	D	$T=1$	$T=0$
1	1	80	20
1	0	90	810
0	Missing	0	0
Total:		170	830

<Disease Verification is obtained for all Test (+) but only 10% of Test (-)>

V	D	$T=1$	$T=0$
1	1	80	2
1	0	90	81
0	Missing	0	747
Total:		170	830

	Sensitivity	Specificity	PPV	NPV
Full data	80/100=80%	810/900=90%	same	same
Verification biased data	80/82=98% Overestimate	81/171=47% Underestimate		

53

YONSEI UNIVERSITY

Discussion

Our study has several limitations. First, our study was a retrospective study from a single institution. However, to avoid bias in patient selection, CT images were blindly analyzed without clinical information of the prosthetic valve, echocardiographic data, and surgical findings. In addition, because only 38.2% of our study population (78 of 204 patients) underwent redo-surgery, the remaining 126 patients were excluded from the analysis of diagnostic performance. Therefore, the verification bias may be present and can result in overestimation or underestimation of diagnostic performance of imaging studies. Second, some patients may have had poor quality CT images, which can affect the diagnos

Suh YJ et al. Assessment of Mitral Paravalvular Leakage After Mitral Valve Replacement Using Cardiac Computed Tomography: Comparison With Surgical Findings. *Circulation: Cardiovascular Imaging*, 2016 9(6), e004153.

54

YONSEI UNIVERSITY

KSSR 2021

Imperfect Reference Standard Bias

- MRA for Carotid Artery Stenosis
  - reference standard: catheter angiogram
  - Let, catheter angio has sensitivity 90%, specifcity 70%

catheter/ actual D	D^+	D^-
+	**90** (90%)	30
-	10	**70** (70%)
합계	100	100

  - Let, MRA's true Se = 80%, Sp = 60%

	catheter			
MRA	+	-	+	-
+	72	8	12	28
-	18	2	18	42
	90	10	30	**70**

	catheter	
MRA	+	-
+	84 (70%)	36
-	36	44 (55%)
합계	120	80

Underestimate

55

YONSEI UNIVERSITY

KSSR 2021

Bias in Research Studies[1]

Radiology 2006; 238:780–789

TYPES OF BIAS DISCUSSED IN THIS REVIEW

SELECTION BIAS	INFORMATION BIAS
Sample	Recall
Loss to Follow-up	Interviewer
Disease Spectrum	Verification or Work-up
Referral	Follow-up or Surveillance
Participation	Response
Image-based Selection	Reviewer
Study Exam	Diagnostic Review
Self-selection	Test Review
	Incorporation
	Imperfect Standard
	Reader Order
	Measurement
	Clustering or Repeated Measurement
	Context
	Publication

Types of bias discussed in this review.

56

YONSEI UNIVERSITY

KSSR 2021

EDITORIAL

Submissions to *Radiology*: Our Top 10 List of Statistical Errors[1]

Radiology

Top 10 list of statistical errors

- **Radiology 2009;** 253:288–290.

1. Consult a statistician during the study design phase to review study size, the data to be collected, and the type of analysis that will be performed on the data obtained.
2. Make sure that the size of the study group is sufficient to justify the conclusions you are reporting. Account for the statistical power (or lack thereof) in your study.
3. Analyze all of the data from each step in the methods.
4. In a diagnostic performance study, be sure to account for true-negative cases in your population.
5. Use confidence intervals to assess the extent of difference.

57

YONSEI UNIVERSITY

KSSR 2021

Top 10 list of statistical errors

6. Use a statistical test that considers clustering effects when a study subject has more than one lesion.
7. Use a statistical test that corrects for multiple comparisons, when a large number of variables are being analyzed.
8. Understand the interpretation of a *P* value.
9. Understand the difference between correlation and accuracy.
10. Report on variability in readers.

P=0.03
Does not mean that there is a 3% probability for the difference is not significance

58

YONSEI UNIVERSITY

KSSR 2021

Statistics in Radiology journal

1789

RadioGraphics 2015; 35:1789–1801

Statistics 101 for Radiologists[1]

Arash Anvari, MD
Elkan F. Halpern, PhD
Anthony E. Samir, MD, MPH

Abbreviations: ANOVA = analysis of variance, ROC = receiver operating characteristic

RadioGraphics 2015; 35:1789–1801

Published online 10.1148/rg.2015150112

Content Codes: RS SO

[1]From the Department of Radiology, Ultrasound Division, Massachusetts General Hospital, Harvard Medical School, 55 Fruit St, White 270, Boston, MA 02114. Received April 13, 2015; revision requested May 19 and received June 30; accepted July 1. For this journal-based SA-CME activity, the authors, editor, and reviewers have disclosed no relevant relationships. **Address correspondence to** A.A. (e-mail: *Anvari.arash@mgh.harvard.edu*).

Diagnostic tests have wide clinical applications, including screening, diagnosis, measuring treatment effect, and determining prognosis. Interpreting diagnostic test results requires an understanding of key statistical concepts used to evaluate test efficacy. This review explains descriptive statistics and discusses probability, including mutually exclusive and independent events and conditional probability. In the inferential statistics section, a statistical perspective on study design is provided, together with an explanation of how to select appropriate statistical tests. Key concepts in recruiting study samples are discussed, including representativeness and random sampling. Variable types are defined, including predictor, outcome, and covariate variables, and the relationship of these variables to one another. In the hypothesis testing section, we explain how to determine if observed differences between groups are likely to be due to chance. We explain type I and II errors, statistical significance, and study power, followed by an explanation of effect sizes and how confidence intervals can be used to generalize observed effect sizes to the larger population. Statistical tests are explained in

59

YONSEI UNIVERSITY

KSSR 2021

Radiology Statistical Concepts Series (November 2002 – March 2004)

Contents *(Page numbers in PDF file)*

60

YONSEI UNIVERSITY

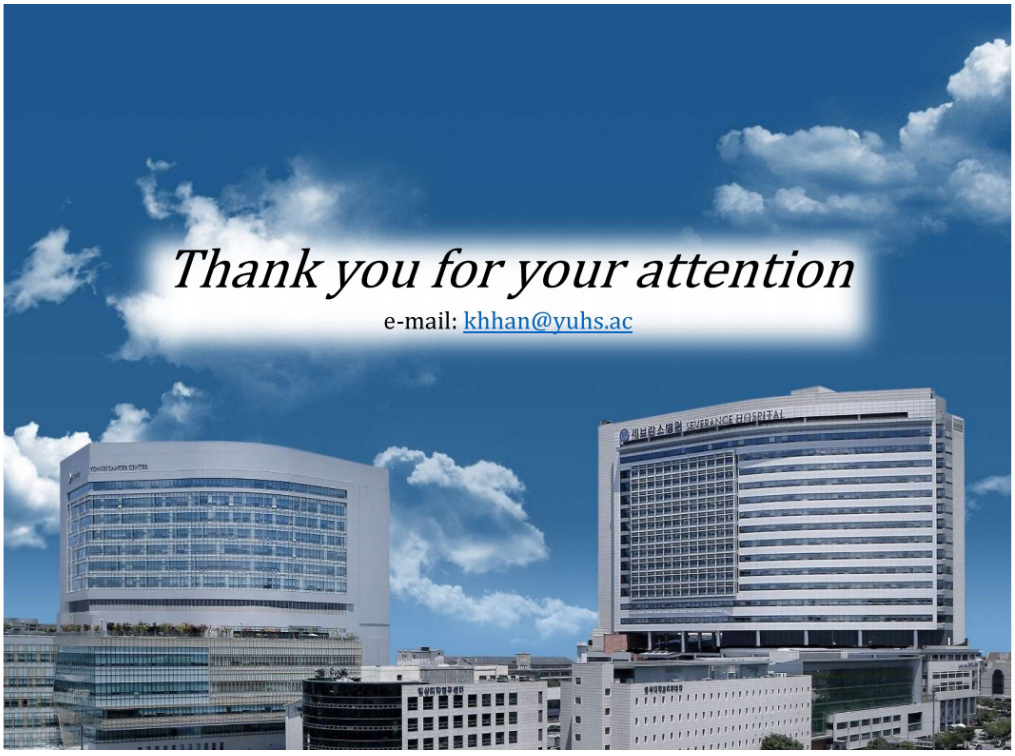
KSSR 2021

Fundamentals of Clinical Research for Radiologists

AJR series

1. Introduction, which appeared in February 2001
2. The Research Framework, April 2001
3. Protocol, June 2001
4. Data Collection, October 2001
5. Population and Sample, November 2001
6. Statistically Engineering the Study for Success, July 2002
7. Screening for Preclinical Disease: Test and Disease Characteristics, October 2002
8. Exploring and Summarizing Radiologic Data, January 2003
9. Visualizing Radiologic Data, March 2003
10. Introduction to Probability Theory and Sampling Distributions, April 2003
11. Observational Studies in Radiology, November 2004
12. Randomized Controlled Trials, December 2004
13. Clinical Evaluation of Diagnostic Tests, January 2005
14. ROC Analysis, February 2005
15. Statistical Inference for Continuous Variables, April 2005
16. Statistical Inference for Proportions, April 2005
17. Reader Agreement Studies, May 2005
18. Correlation and Regression, July 2005
19. Survival Analysis, July 2005
20. Multivariate Statistical Methods, August 2005
21. Decision Analysis and Simulation Modeling for Evaluating Diagnostic Tests on the Basis of Patient Outcomes, September 2005
22. Radiology Cost and Outcomes Studies: Standard Practice and Emerging Methods, October 2005

61



Clinical Research Methodology Course – Intermediate Course

10:00–10:50 Room 2

Statistical modeling for continuous outcome

송 기 준
연세대학교

Statistical modeling for *continuous* outcome

- Linear regression analysis

선형회귀분석(linear regression analysis): example

TABLE 2
Total Procedure Time and Dose of CT Fluoroscopy–guided Procedures, by Means of the Quick-Check Method

Subject No.	*x* Data: Log Time (ln[min])	*y* Data: Log Dose (ln[rad])
1	3.61	1.48
2	3.87	1.24
3	3.95	2.08
4	4.04	1.70
5	4.06	2.08
6	4.11	2.94
7	4.19	2.24
8	4.20	1.85
9	4.32	2.84
10	4.32	3.93
11	4.42	3.03
12	4.42	3.23
13	4.45	3.87
14	4.50	3.55
15	4.52	2.81
16	4.57	4.07
17	4.58	4.44
18	4.61	3.16
19	4.74	4.19

선형회귀분석(linear regression analysis): example

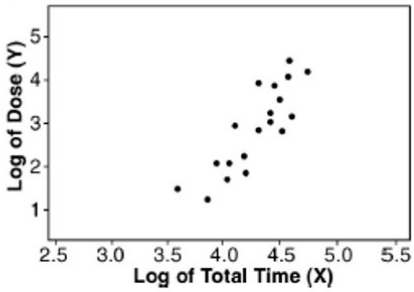


Figure 3. Scatterplot of the log of dose (y axis) versus the log of total time (x axis). Each point in the scatterplot represents the values of two variables for a given observation.

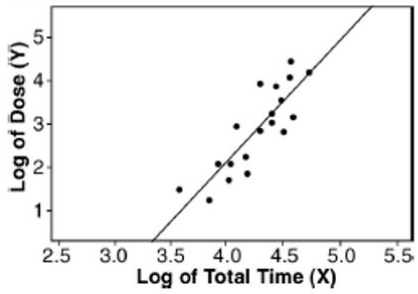


Figure 4. Scatterplot of the log of dose (y axis) versus the log of total time (x axis). The regression line has the intercept $a = -9.28$ and slope $b = 2.83$. We conclude that there is a possible association between the radiation dose and the total time of the procedure.

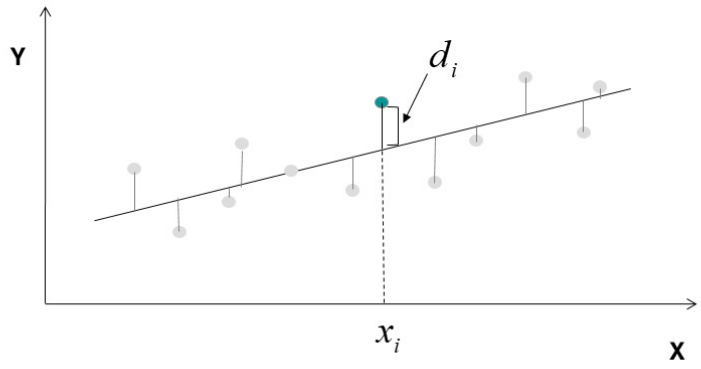
선형회귀분석

- 회귀분석
  - 회귀모형을 이용하여 독립변수(들)와 종속변수간의 선형적 (인과)관계를 알아보기 위한 분석 방법
- n개의 대상에 대하여 독립변수 X, 종속변수 Y의 각 관찰치를 $x_1, x_2, \cdots, x_n$과 $y_1, y_2, \cdots, y_n$이라 할 때 단순선형회귀모형은 다음과 같음.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, 2, \cdots, n$$

- 회귀계수(regression coefficient; β_0, β_1)
  - β_0 : intercept(y 절편), X=0일 때 종속변수 Y의 값
  - β_1 : slope(기울기), X 값이 한 단위 증가할 때, Y값의 변화량
- 기본가정
  - 오차항 $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ 은 서로 독립이며, 평균은 0, 분산은 σ^2인 정규분포를 따름. 즉, 오차항의 독립성(independency), 등분산성(constancy), 정규성(normality)을 만족함.

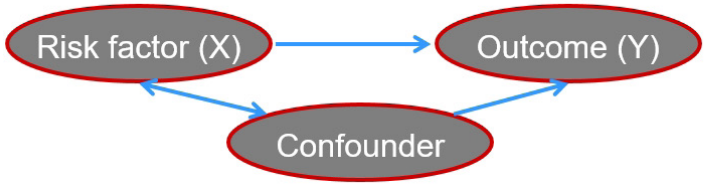
단순(simple) 선형회귀분석: 회귀계수(β_0, β_1)의 추정



- 각 관찰치에서 회귀직선까지 거리(d_i)의 제곱의 합을 최소화하는 회귀계수를 추정 !!!

➔ 최소 제곱 추정법(LSE; Least Square Estimation)

혼란변수 (Confounders)



- 종류 : positive confounder, negative confounder
- Positive confounder(PC)

  : 위험요인과 질병에 모두 같은 방향으로 영향을 미치는 경우
- Negative confounder(NC)

  : 위험요인과 질병에 미치는 영향이 서로 다른 방향인 경우

- 만약 PC를 통제하지 못하면? 관련성 크기는 과대추정
- 만약 NC를 통제하지 못하면? 관련성 크기는 과소추정

혼란변수의 영향을 통제하는 방법

연구설계를 통한 방법

- 연구대상자 선정 범위 제한 (inclusion, exclusion criteria 등)
- 짝짓기 (matching): 자료를 모을 때 부터
- 완전확률화를 통한 방법 (randomization) : best of best !!!

통계학적 모형을 이용하는 방법

- 각종 회귀모형을 이용하여 분석하는 방법 :
  선형회귀모형, 로지스틱 회귀모형, Cox의 비례위험 회귀모형 등

※ 연구설계 단계를 통해서는 혼란변수의 영향을 제거(elimination) 혹은 배제(exclusion) 할 수 있지만, 통계학적 모형을 이용하는 것은 보정(adjustment) 혹은 통제(control)하는 수준에 지나지 않음.

Multiple linear regression analysis : Example

Primer on Multiple Regression Models for Diagnostic Imaging Research[1]

This article provides an introduction to multiple regression analysis and its application in diagnostic imaging research. We begin by examining why multiple regression models are needed in the evaluation of diagnostic imaging technologies. We then examine the broad categories of available models, notably multiple linear regression models for continuous outcomes and logistic regression models for binary outcomes. The purpose of this article is to elucidate the scientific logic, meaning, and interpretation of multiple regression models by using examples from the diagnostic imaging literature.

WHY ARE MULTIPLE REGRESSION MODELS USED IN DIAGNOSTIC IMAGING?

Multiple Factors of Interest

Adjustment for Potential Confounding

Prediction

TABLE 1
Results of Multiple Linear Regression Analysis to Examine the Number of Annual Procedures per FTE Radiologist in Diagnostic Radiology Groups

Variable	Coefficient (β)	Standard Error*	*P* Value
Intercept (β_0)	10,403	2,154	.001
Academic status (X_1)	−2,238	1,123	.05
Annual hours per FTE (X_2)	0.43	1.11	.70
Group size (FTE) (X_3)	−59.7	32.5	.07
Proportion of high productivity procedures (X_4)†	−4,782	11,975	.69

Note.—Adapted and reprinted, with permission, from reference 1.
* Standard error of the estimated coefficient.
† High-productivity procedures included computed tomography (CT) and magnetic resonance (MR) imaging, and interventional or angiographic procedures that required more mental effort, stress, physical effort, and training than did other types of procedures.

다중선형회귀분석(multiple linear regression analysis)

- 선형회귀모형을 이용하여 두 개 이상의 독립변수들이 연속형 종속변수에 **선형적으로** 영향을 미치는지 파악하기 위한 방법

- 자료의 형태

대상	종속변수	독립변수 1	독립변수 2	…	독립변수 k
1	y_1	x_{11}	x_{21}	…	x_{k1}
2	y_2	x_{12}	x_{22}	…	x_{k2}
⋮	⋮	⋮	⋮	⋮	⋮
n	y_n	x_{1n}	x_{2n}	…	x_{kn}

다중선형회귀모형의 설정

- 통계학적 모형
  - 종속변수를 y로 k개의 독립변수를 $X_1, X_2, \cdots, X_k$ 로 나타내면 다중선형회귀모형은 다음과 같이 표현할 수 있음.

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i ,$$

$$i = 1, \cdots, n$$

- 기본 가정
  - 오차항 $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$은 서로 독립이며, 평균은 0, 분산은 σ^2인 정규분포를 따름. (독립성, 등분산성, 정규성)

※ **독립변수들 간에는 선형적 관련성이 존재하지 않아야 함.**

회귀모형의 유의성 검정

- 검정절차
  - 가설 설정

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \quad vs. \quad H_1 : not\ H_0$$

  - 분산분석 table 및 검정통계량

	제곱합(SS)	자유도(df)	평균 제곱합(MS)	F^*
회귀(Regression)	SSR	k	MSR=SSR/k	MSR/MSE
잔차(Residual)	SSE	n-k-1	MSE=SSE/(n-k-1)	
합	SST	n-1		

$$F^* \sim F_{k,\, n-k-1}$$

  - 의사결정 원칙(귀무가설 기각)

F^* 에 해당하는 p-value < 0.05

개별 회귀계수의 유의성 검정

- 검정절차
  - 가설 설정

$$H_0 : \beta_k = 0 \quad vs. \quad H_1 : \beta_k \neq 0$$

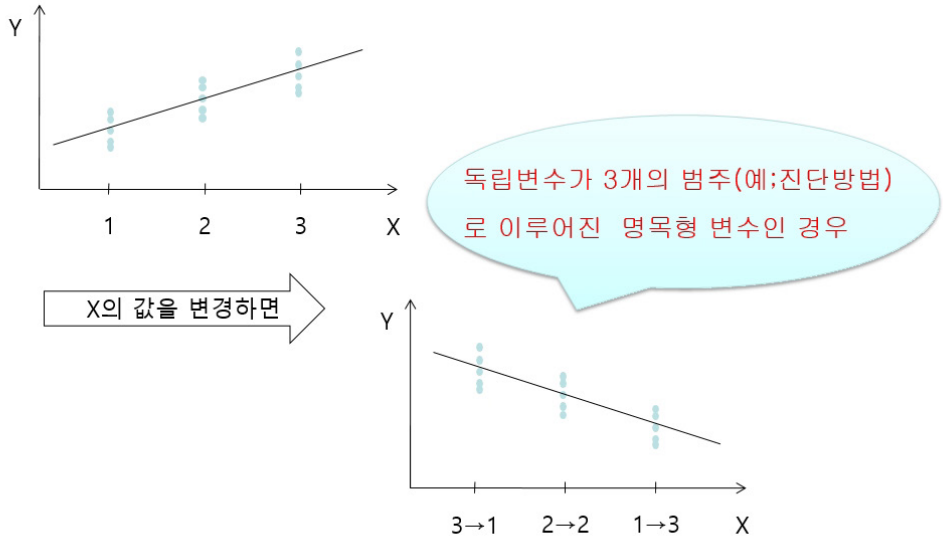
  - 검정통계량

$$t^* = \frac{\hat{\beta}_k}{SE(\hat{\beta}_k)} \sim t_{n-k-1}$$

  - 의사결정 원칙(귀무가설 기각)

$|t^*|$ 에 해당하는 p-value < 0.05

가변수(dummy variable)를 사용한 회귀분석



가변수(dummy variable)를 사용한 회귀분석

- 독립변수들 중에서 연속형 변수(양적인 변수) 외에 명목형 변수(질적인 변수)가 있을 경우 이에 대한 가변수 생성
- 참조범주(reference category)를 제외한 (범주의 수-1)개의 가변수 생성

  예) 범주의 수가 3개인 경우

X	X_1	X_2
1	1	0
2	0	1
3	0	0

- 가변수에 대한 회귀계수의 해석
  - 참조범주에 비해 다른 범주들의 종속변수가 얼마나 차이가 나는지 보여주는 지표, 일반적인 기울기의 의미가 아님.

가변수에 대한 회귀계수의 해석

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

$$\Rightarrow \quad X = 1;\ \hat{y}_{X=1} = \hat{\beta}_0 + \hat{\beta}_1 + 0$$

$$X = 2;\ \hat{y}_{X=2} = \hat{\beta}_0 + 0 + \hat{\beta}_2$$

$$X = 3;\ \hat{y}_{X=3} = \hat{\beta}_0 + 0 + 0$$

	X		
	1	2	3
X_1	1	0	0
X_2	0	1	0

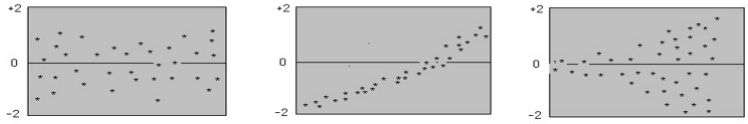
$\hat{\beta}_1$: $X = 3$인 집단에 비해 $X = 1$ 인 집단의 Y의 평균적인 차이

$\hat{\beta}_2$: $X = 3$인 집단에 비해 $X = 2$ 인 집단의 Y의 평균적인 차이

다중회귀분석 주의점: 기본가정 검토

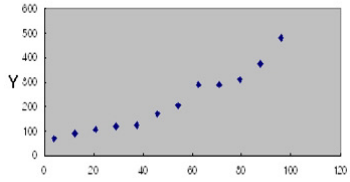
- 기본가정(독립성, 등분산성, 정규성)에 대한 검토 : 잔차분석을 이용
  - 잔차(오차의 추정치; residual): $y_i - \hat{y}_i$
  - 잔차도(residual plot)를 이용하여 선형회귀모형의 가정이 잘 맞는지 검토

1) 잔차의 분포에 따른 가정 검토 : 독립성, 등분산성



(a) (b) (c)

2) 잔차의 분포에 따른 가정 검토 : 정규성



Normal probability plot (Q-Q plot)

다중회귀분석 주의점: 다중공선성 (multicollinearity)

- 독립변수들간에 선형적 상관성이 존재하는 경우
- 특정한 독립변수(X_k)가 다른 독립변수들에 의해 많이 설명된다면?

$$VIF_k = 1/(1-R_k^2)$$

- VIF(Variance Inflation Factor; 분산확대인자)가 10 이상인 독립변수는 다중공선성의 문제가 있다고 판단

※ 다중공선성의 해결 방안

- 다중공선성이 있는 변수를 회귀모형에서 제외하여 분석
- 다중공선성이 있는 변수를 "centering $(X-\bar{X})$" 시켜 분석에 포함 시킴.

다른 변수들을 보정한 후 유의한 factor를 찾는 연구

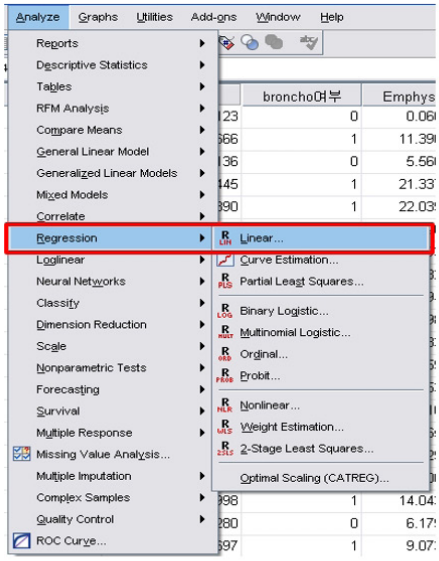
- Outcome과 관련 있다고 여겨지는 후보 독립변수(들)와 보정할 변수들 조사
- 각 변수에 대해 outcome과 univariate analysis 수행
- 다중회귀분석에 포함시킬 변수 선택
  - 적절한 독립 변수의 개수(결측치가 없는 전체자료 수의 1/10 혹은 1/15) 범위 내에서 분석에 포함시킬 변수 선택
  - 기존에 outcome과 관계가 익히 알려진 변수(반드시 보정해야 할 변수)들 선택(통계학적 유의성에 상관없이)
  - 일변량 분석 결과, 통계학적으로 유의한 변수 선택
  - 선택된 후보 변수들간에 상관성 파악(다중공선성 확인)
- 최종적으로 선택된 변수들을 이용하여 다중회귀분석 수행 후 결과 해석

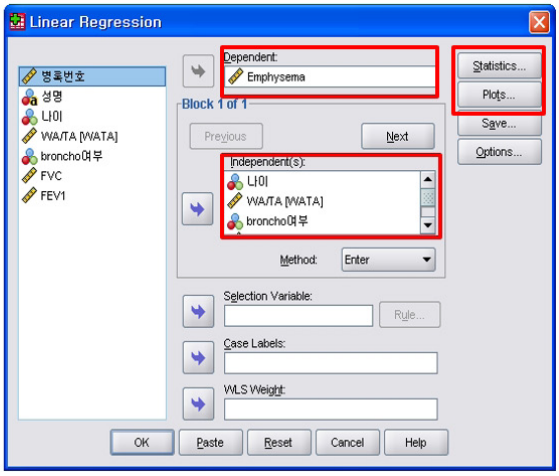
다중회귀분석 : SPSS example

- 84명의 만성 폐쇄성 폐질환 환자에 있어서 Emphysema Scores를 측정한 자료. 이 때 나이, WA/TA, broncho여부, FVC, FEV1이 Emphysema score와 관련이 있는지 파악하고자 함.

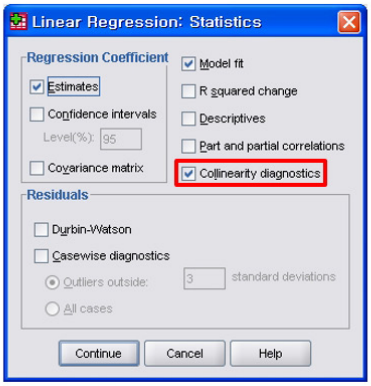
나이	WA/TA	broncho여부	FVC	FEV1	Emphysema score
59	0.756509	0	75	70	0.07
76	0.7743406	1	32	35	11.39
67	0.734191	0	62	47	5.56
59	0.718514	1	35	52	21.34
54	0.6979343	1	41	50	22.04
71	0.8081131	0	20	25	4.92
67	0.7683278	1	42	35	27.09
60	0.7820918	0	40	36	18.33
…	…	…	…	…	…
83	0.7414517	1	52	38	27.25
84	0.7541633	0	35	42	28.30

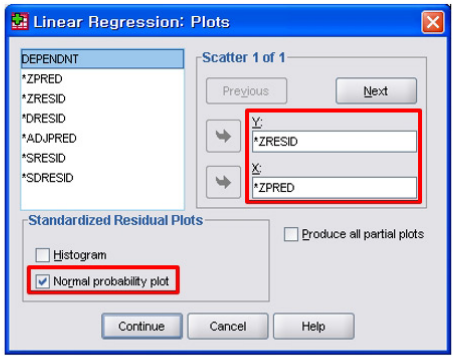
실제구현 : 다중회귀분석





실제구현 : 다중회귀분석





결과 : 다중회귀분석

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.762[a]	.581	.506	8.451969297 E0%

a. Predictors: (Constant), FEV1, broncho여부, WA/TA, 나이, FVC

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2773.528	5	554.706	7.765	.000[a]
	Residual	2000.202	28	71.436		
	Total	4773.730	33			

a. Predictors: (Constant), FEV1, broncho여부, WA/TA, 나이, FVC

b. Dependent Variable: Emphysema

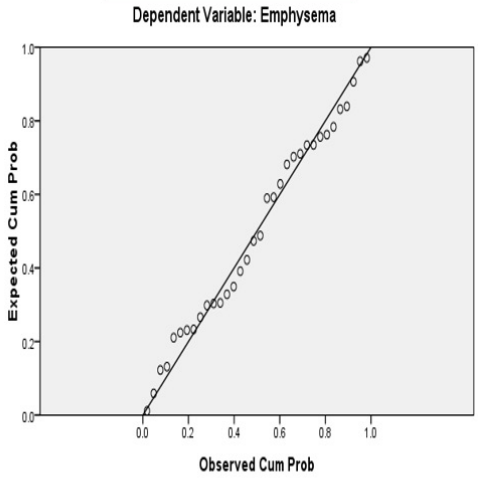
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	160.368	52.837		3.035	.005		
	나이	-.056	.233	-.034	-.239	.813	.732	1.366
	WA/TA	-182.953	68.592	-.360	-2.667	.013	.823	1.216
	broncho여부	6.251	3.196	.247	1.956	.061	.940	1.064
	FVC	7.837	2.488	.574	3.150	.004	.450	2.221
	FEV1	-17.386	3.258	-1.059	-5.337	.000	.380	2.631

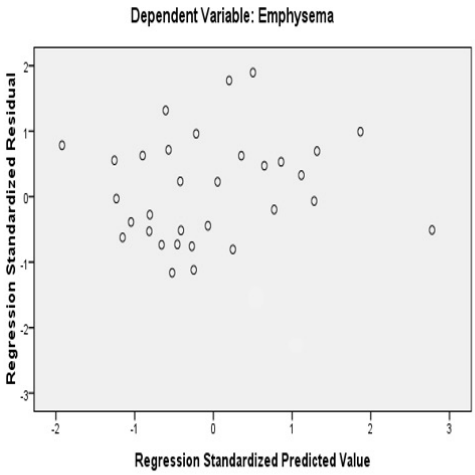
a. Dependent Variable: Emphysema

결과 : 다중회귀분석

Normal probability plot



Residual plot



Clinical Research Methodology Course – Intermediate Course

11:00–11:50 Room 2

Statistical modeling for binary outcome

송 기 준
연세대학교

Statistical modeling for *binary* outcome

- Logistic regression analysis

Logistic regression analysis : Example

Pulmonary Embolism at CT Angiography: Implications for Appropriateness, Cost, and Radiation Exposure in 2003 Patients[1]

Radiology: Volume 256: Number 2—August 2010

Purpose: To determine whether thromboembolic risk factor assessment could accurately indicate the pretest probability for pulmonary embolism (PE), and if so, computed tomographic (CT) angiography might be targeted more appropriately than in current usage, resulting in decreased costs and radiation exposure.

Materials and Methods: Institutional review board approval was obtained. Electronic medical records of 2003 patients who underwent CT angiography for possible PE during 1½ years (July 2004 to February 2006) were reviewed retrospectively for thromboembolic risk factors. Risk factors that were assessed included immobilization, malignancy, hypercoagulable state, excess estrogen state, a history of venous thromboembolism, age, and sex. Logistic regressions were conducted to test the significance of each risk factor.

Results: Overall, CT angiograms were negative for PE in 1806 (90.16%) of 2003 patients. CT angiograms were positive for PE in 197 (9.84%) of 2003 patients; 6.36% were Emergency Department patients, and 13.46% were inpatients. Of the 197 patients with CT angiograms positive for PE, 192 (97.46%) had one or more risk factors, of which age of 65 years or older (69.04%) was the most common. Of the 1806 patients with CT angiograms negative for PE, 520 (28.79%) had no risk factors. The sensitivity and negative predictive value of risk factor assessment in all patients were 97.46% and 99.05%, respectively. All risk factors, except sex, were significant in the multivariate logistic regression ($P < .031$).

Conclusion: In the setting of no risk factors, it is extraordinarily unlikely (0.95% chance) to have a CT angiogram positive for PE. This selectivity and triage step should help reduce current costs and radiation exposure to patients.

Logistic regression analysis : Example

Table 4

Frequency of Specific Risk Factors in Patients with CT Angiograms Positive for PE

Risk Factor	No. of Patients	Percentage
Age 65 y or older	136	69.04
Immobilization	106	53.81
Male sex	100	50.76
Malignancy	73	37.06
Prior PE and/or DVT	28	14.21
Hypercoagulable state	16	8.12
Excess estrogen state	12	6.09

Table 6

Multivariate Logistic Regression Analyses and PE Odds Ratios

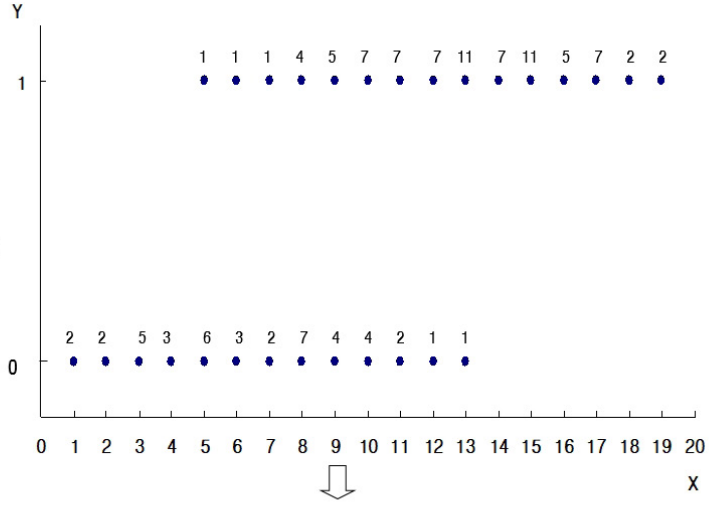
Risk Factor	Odds Ratio: Lower 95% Confidence Limit	Odds Ratio: Point Estimate*	Odds Ratio: Upper 95% Confidence Limit	*P* Value†
Immobilization	4.08	5.95	8.68	<.001
Hypercoagulable state	5.94	8.42	11.93	<.001
Malignancy	1.05	1.75	2.90	<.031
Prior PE and/or DVT	3.56	7.56	15.93	<.001
Excess estrogen state	1.77	3.63	7.45	<.001
Age‡	1.16	1.67	2.41	<.006
Sex§	0.95	1.33	1.86	<.1

* The point estimate is the odds ratio that determines how much more likely a PE is to occur.
† A difference with $P < .05$ was considered significant.
‡ Coded with score 1 for age 65 years or older and score 0 for age younger than 65 years.
§ Coded with score 1 for male sex and score 0 for female sex.

종속변수가(Y)가 이분형(예; 0,1) 변수인 경우

X	Y
1	0
1	0
2	0
2	0
3	0
⋮	⋮
17	1
18	1
18	1
19	1
19	1

< Y vs. X 산점도 (n=120) >



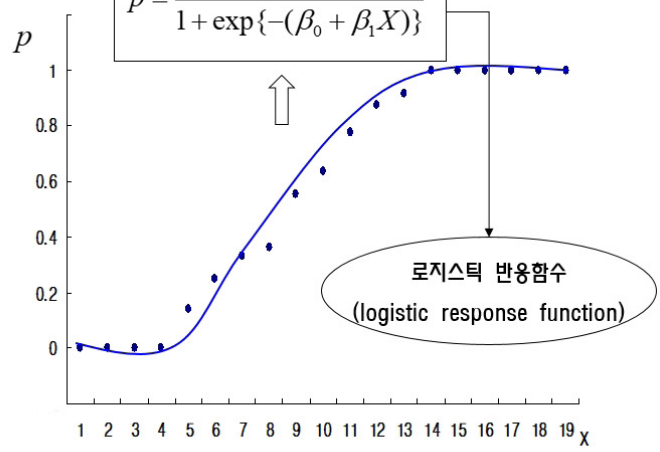
X와 Y의 관계를 선형적으로 표현하기 어려움

P(Y=1) vs. X 산점도

$$\exp = exponetial = e = \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n = 2.718281\cdots\cdots$$

X	Y=1	Y=0	*P*(Y=1)
1	0	2	0
2	0	2	0
3	0	5	0
4	0	3	0
5	1	6	0.14
⋮	⋮	⋮	⋮
9	5	4	0.56
10	7	4	0.64
11	7	2	0.78
⋮	⋮	⋮	⋮
18	2	0	1
19	2	0	1

$$p = \frac{1}{1+\exp\{-(\beta_0+\beta_1 X)\}}$$

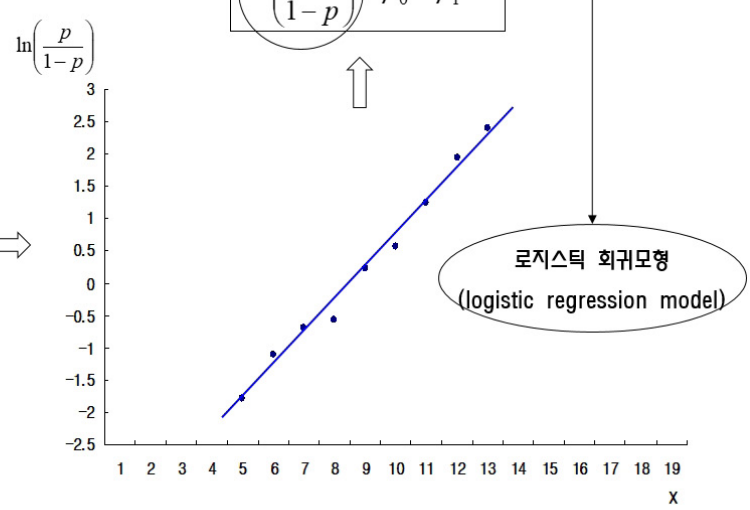


$\ln\left(\frac{p}{1-p}\right) = \ln(odds)$ vs. X 산점도

X	P(Y=1)	$\ln\left(\frac{p}{1-p}\right)$
1	0	
2	0	
3	0	
4	0	
5	0.14	−1.82
⋮	⋮	⋮
9	0.56	0.24
10	0.64	0.58
11	0.78	1.27
⋮	⋮	⋮
18	1	
19	1	

Logit link

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$



로지스틱 회귀분석(Logistic regression analysis)

- 종속변수가 두 개의 범주(이분형; binary)로 측정되는 경우
- 로지스틱 반응함수: $P = P(Y=1) = \frac{1}{1+\exp\{-(\beta_0+\beta_1 X)\}}$
- 로지스틱 회귀모형

$$\text{Odds} = \frac{P}{1-P} = \frac{\frac{1}{1+\exp\{-(\beta_0+\beta_1 X)\}}}{1-\frac{1}{1+\exp\{-(\beta_0+\beta_1 X)\}}} = \exp(\beta_0+\beta_1 X)$$

$$\Rightarrow \ln(\text{odds}) = \ln\left(\frac{P}{1-P}\right) = \text{logit}(P) = \underline{\ln\{\exp(\beta_0+\beta_1 X)\} = \beta_0+\beta_1 X}$$

$\boxed{log_e e^a = a}$

β_0 : x가 0일 때 사건이 일어날 $\ln(odds)$

β_1 : x가 한 단위 증가할 때 사건이 일어날 $\ln(odds)$의 증 / 감분

로지스틱 회귀분석 : 회귀계수(β) 추정

- 최대우도추정법(MLE; Maximum Likelihood Estimation) 이용
  - 우도(Likelihood): 회귀모형을 이용하여 종속변수를 예측할 확률(가능성)
  - 로그-우도함수: $\ln L(\beta_0,\beta_1) = \sum_{i=1}^{n}[Y_i \ln(p) + (1-Y_i)\ln(1-p)]$
  - 정규 방정식: $\sum_{i=1}^{n}(Y_i - p) = 0$, $\sum_{i=1}^{n} X_i(Y_i - p) = 0$

    ⇒ 회귀계수 추정치가 단일한(하나의) 값으로 얻어지지 않음.

    ⇒ Fisher-scoring method 또는 Newton-Raphson method를 이용하여 수많은 반복을 통해 <u>로그-우도함수 값을 최대가 되게 하는</u> 회귀계수 β 추정

회귀계수(β)와 Odds Ratio(OR)의 관계

- Odds Ratio의 정의
  - Odds

$$Odds_0 = \frac{p_0}{1-p_0} = \frac{X=0\text{일 때 } Y=1\text{이 될 확률}}{X=0\text{일 때 } Y=0\text{이 될 확률}}$$

$$Odds_1 = \frac{p_1}{1-p_1} = \frac{X=1\text{일 때 } Y=1\text{이 될 확률}}{X=1\text{일 때 } Y=0\text{이 될 확률}}$$

  - Odds ratio

$$\boxed{OR = \frac{Odds_1}{Odds_0} = \frac{p_1/(1-p_1)}{p_0/(1-p_0)} = \frac{\exp(\beta_0+\beta_1)}{\exp(\beta_0)} = \exp(\beta_1)}$$

- Odds Ratio에 대한 95% 신뢰구간

$$\exp\{\hat{\beta}_1 \pm 1.96 \times SE(\hat{\beta}_1)\}$$

다중 로지스틱 회귀분석

- 다중 로지스틱 회귀모형

$$\log\left(\frac{p}{1-p}\right) = \text{logit}(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

- 회귀모형의 유의성 검정
  - 가설 : $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \quad vs. \quad H_1 : not\ H_0$
  - 검정통계량 : 우도비 검정 통계량(Likelihood ratio test statistic)

$$G^2 = -2(L_0 - L_1) \sim \chi^2_k$$

  - 의사결정 원칙(귀무가설 기각)

G^2에 해당하는 p-value < 0.05

다중 로지스틱 회귀분석

- 개별 회귀계수의 검정
  - 가설 : $H_0 : \beta_k = 0 \quad vs. \quad H_0 : \beta_k \neq 0$
  - 검정통계량 : 왈드 검정통계량(Wald chi-square test statistic)

$$W = \left(\frac{\hat{\beta}_k}{SE(\hat{\beta}_k)} \right)^2 \sim \chi_1^2$$

  - 의사결정 원칙(귀무가설 기각)

    W 에 해당하는 p-value < 0.05

로지스틱 회귀분석 적용에서 주의사항

- 독립변수들끼리 상관성(multi-collinearity) 검토
- 적절한 표본 크기 및 독립변수의 개수
  - 종속변수의 관심 있는 사건의 발생 비율이 10% 이상인 경우가 적절
  - "The rule of 10 events per parameter" by Peduzzi et al(1996):

    독립변수의 개수 ≤ {min(no. of events, no. of non-events)}/10

로지스틱 회귀분석 적용에서 주의사항

- 이분형 독립변수와 종속변수 간의 2X2 table을 만들었을 때, 한 개의 cell이라도 0인 경우 odds ratio 값이 무한대(굉장히 큰 숫자)혹은 0으로 추정됨.

	1	0
1	a	b
0	**0**	d

$$OR = \frac{a \times d}{b \times 0}$$

	1	0
1	**0**	b
0	c	d

$$OR = \frac{0 \times d}{b \times c}$$

- 연속형 독립변수의 완전 분리(complete separation) 문제
  - 연속형 독립변수의 특정한 값을 중심으로 종속변수의 범주가 확연히 구분되는 경우(예를 들어, 연령이 60세 이상이면 모두 질병이 있고 60세 미만이면 모두 질병이 없는 자료), 회귀계수의 최대우도 추정치가 존재하지 않음.

	1	0
≥60	a	0
<60	0	d

• 이분형 독립변수와 종속변수 간의 2X2 table에서 odds ratio의 분모가 0인 경우

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	69.315[a]	.350	.519

a. Estimation terminated at iteration number 20 because maximum iterations has been reached. Final solution cannot be found.

Classification Table[a]

Observed		Predicted Y 0	Predicted Y 1	Percentage Correct
Step 1 Y	0	50	25	66.7
	1	0	25	100.0
Overall Percentage				75.0

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	X	21.203	5684.147	.000	1	.997	1615474523	.000	.
	Constant	-21.203	5684.147	.000	1	.997	.000		

a. Variable(s) entered on step 1: X.

- 연속형 독립변수에서 완전 분리(complete separation)가 발생하는 경우

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	.000[a]	.745	1.000

a. Estimation terminated at iteration number 20 because maximum iterations has been reached. Final solution cannot be found.

Classification Table[a]

Observed			Predicted Y 0	Predicted Y 1	Percentage Correct
Step 1	Y	0	40	0	100.0
		1	0	30	100.0
	Overall Percentage				100.0

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	95% C.I.for EXP(B) Upper
Step 1[a]	X	24.411	632.334	.001	1	.969	3.997E+10	.000	.
	Constant	-1452.475	37624.971	.001	1	.969	.000		

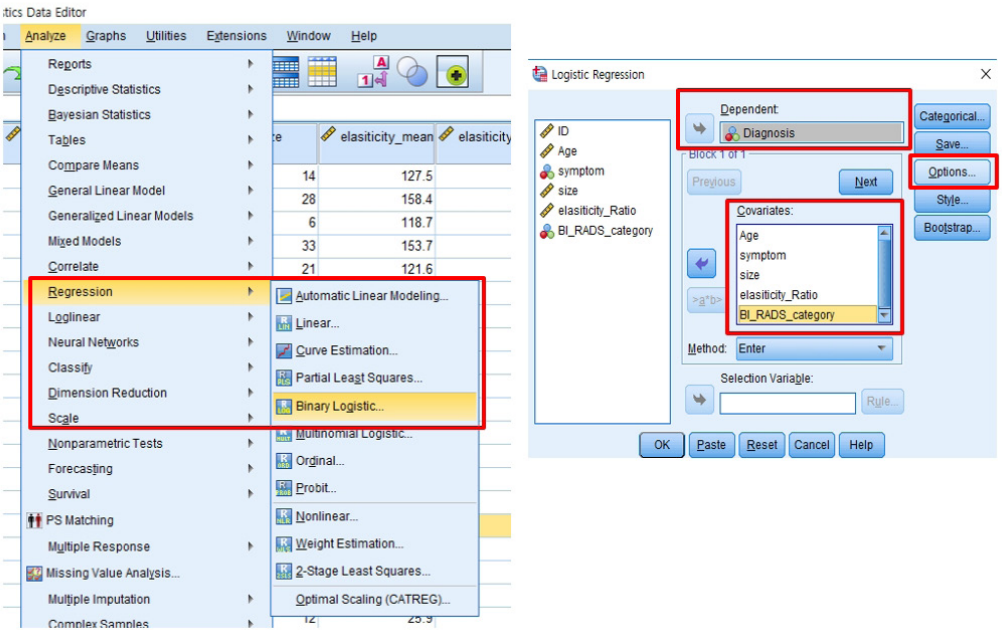
a. Variable(s) entered on step 1: X.

로지스틱 회귀분석 : SPSS example

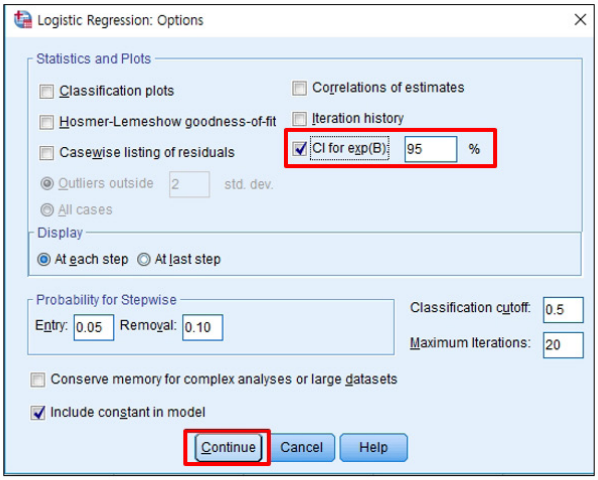
- 유방 종괴에 대한 전단파 탄성 초음파를 실시한 330명의 여성에서 최종적인 유방암 진단(diagnosis)과 관련된 인자를 확인하고자 함.

ID	Age	symptom	size	Elasticity Ratio	BI-RADS category	Diagnosis
1	36	1	14	8.93	1	1
2	36	1	28	12.98	0	1
3	52	0	6	13.15	1	1
4	53	1	33	14.82	0	1
5	70	1	21	10.96	1	1
6	28	1	13	10.95	1	1
7	45	1	9	9.15	1	1
8	45	0	11	7.17	1	1
9	45	1	20	11.07	1	1
10	52	1	5	2.96	0	2
…	…	…	…	…	…	…
324	43	1	5	1.63	1	2
325	53	0	8	1.43	1	2
326	53	1	7	1.70	0	2
327	44	1	11	2.88	1	2
328	40	0	13	3.70	1	2
329	38	1	16	4.12	0	2
330	47	1	17	2.53	0	2

실제구현 : 로지스틱 회귀분석



실제구현 : 로지스틱 회귀분석



결과 : 로지스틱 회귀분석

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	416.242	5	.000
	Block	416.242	5	.000
	Model	416.242	5	.000

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Age	-.123	.081	2.280	1	.131	.884	.754	1.037
	symptom	.915	1.925	.226	1	.634	2.498	.057	108.603
	size	-.020	.188	.011	1	.915	.980	.679	1.416
	elasiticity_Ratio	-1.965	.493	15.880	1	.000	.140	.053	.368
	BI_RADS_category	4.042	1.632	6.136	1	.013	56.959	2.326	1394.994
	Constant	11.923	5.876	4.117	1	.042	150648.995		

a. Variable(s) entered on step 1: Age, symptom, size, elasiticity_Ratio, BI_RADS_category.

Clinical Research Methodology Course – Intermediate Course

13:10–14:20 Room 2

Fundamentals of survival analysis

김 선 옥
서울아산병원

Survival (time to event) Analysis

서울아산병원 의학통계학과

김 선 옥

Fundamentals of survival analysis 1/88

통계적 모형

Outcome	Model
Continuous	Linear regression $Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$
	Generalized additive model $Y = \alpha + f_1(x_1) + f_2(x_2) + \varepsilon$
Binary	Logistic regression $log\left\{\frac{P(Y=1)}{1-P(Y=1)}\right\} = \alpha + \beta_1 x_1 + \beta_2 x_2$
Survival	Cox PH model $h(t;X) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2)$

Fundamentals of survival analysis 2/88

목차

- 생존분석의 개념과 자료형태
- 생존함수의 가정과 생존함수의 정의
- 생존자료의 요약
  - Life-table methods (생명표법)
  - **Kaplan-Meier analysis (product-limit method)**
- 그룹 간 생존율 및 평균 생존기간 비교
  - **Log-rank test (로그순위 검정)**
- 생존율에 영향을 미치는 요인분석
  - **Cox proportional hazards model with time independent and time dependent covariates**

생존분석(Survival analysis)이란?

- **사건**의 발생과 그러한 사건이 일어날 **시점**을 예측하고 설명하는 통계학의 방법론 중 하나
  - 사건 : 질병으로 인한 사망, 질병의 발생, 재발 (relapse from remission), 회복(recovery)
  - 사건의 표현 : 0 (발생 X) 또는 1 (발생 O)로 코딩
- 일정한 조건을 갖춘 연구대상을 추적-관찰하면서 질병의 발생이나 재발 혹은 생명현상의 종결인 사망의 **확률**을 **시간의 함수**로 분석하는 방법

생존자료의 형태 및 특성

- 결과변수 (Y_1) : 이분형 변수
  - 일반적인 범주형분석과 차이점: 사건의 발생과 비발생으로 나누어지긴 하지만, 비발생에 불확실한 정보에 대한 관찰이 포함됨
  - 사건과 탈락에 따른 관찰 중단 여부
    - ✓ 사건 발생 (event) – 완전 절단 (uncensoring)
    - ✓ 사건 비발생 (no event) – 중도 절단 (censoring)
- 결과변수 (Y_2) : 시간 (사건이 일어날 때 까지 관찰 기간)
  - Uncensored- 사건 발생까지 관찰기간
  - Censored- 실제 관찰기간
- 설명 변수 (X)
  - 사건 발생과 관련될 것으로 기대되는 여러 요인(특성)

생존자료의 이해 (1)

- 완전절단(un-censoring, complete observation) : 관찰시작 이후 사건이 발생한 경우
- 중도절단(censoring) : 어떤 표본이 관심 사건의 발생 여부와 관계없이 어떤 이유로든 그 실제 값을 알 수 없게 되는 경우
  - 연구대상자가 연구 종료 전 우리 병원에서 다른 병원으로 옮긴다거나(FU loss), 참여를 거부하거나 조건이 맞지 않아 중도 탈락(drop-out) 하는 경우
  - 관찰기간 동안 사건을 경험하지 못하게 되는 경우 (예. 연구종료 시점까지 사건 일어나지 않음 or 다른 질병으로 사망)

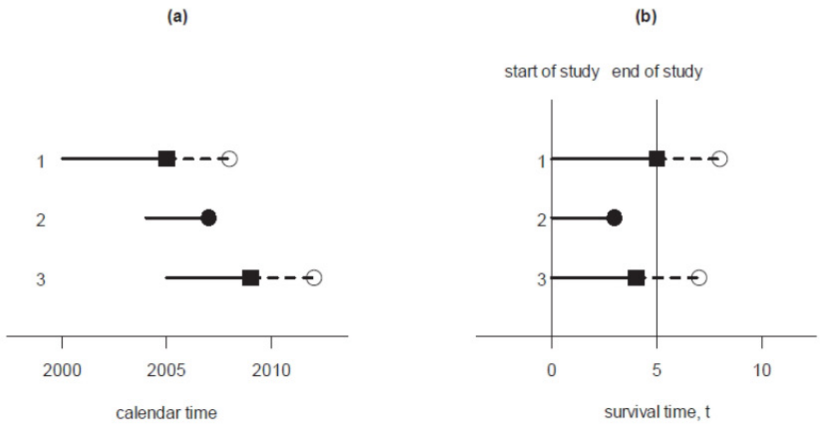
생존자료의 이해 (2)

- 생존기간 : 말 그대로 환자가 관심 사건을 경험하기 전까지 생존한 기간
  - 연속적 시간 자료 또는 비연속적 시간 자료

- f/u 이 드물게 반복되는 경우 비연속적 시간 자료에 해당하는데, 이 때는 중간에 사건이 발생하는 경우에 interval censoring 이라 한다 (규칙적인 f/u의 필요성)

생존자료의 이해 (3)

- **Calendar time (a) vs. survival time (b)**



John Fox. Introduction to survival analysis.2014
Available from: http://socserv.mcmaster.ca/jfox/Courses/soc761/survival-analysis.pdf

2014년 1월 1일 부터 A와 B 치료법을 무작위로 배정하여 투여하고 추적관찰을 시작하여 2015년 12월 31일에 연구종료하여 아래와 같은 자료를 얻었다. 생존여부와 생존기간을 확인해 보시오.

환자번호	성별	연령	치료법	치료개시일	2015.12.31 생존 여부	사망연월일	최종생사여부확인일
1	M	62	A	2014-04-03	사망	2014-06-15	
2	F	57	A	2014-04-16	생존		
3	M	49	B	2014-05-12	사망	2014-10-18	
4	M	72	B	2014-06-16	사망	2014-08-04	
5	F	63	A	2014-06-18	소식불명		2014-06-25
6	M	51	A	2014-07-26	생존		
7	F	68	B	16-Aug-14	사망	2014-11-02	
8	M	40	B	23-Aug-14	소식불명		2014-12-20
9	F	38	A	27-Sep-14	생존		
10	F	67	B	2014-10-16	생존		
11	M	81	B	2014-10-26	사망	2015-11-19	
12	M	54	A	2014-11-11	사망	2015-12-20	
13	M	57	A	2015-01-14	생존		
14	F	63	A	2015-01-20	생존		
15	M	48	B	2015-02-05	생존		
16	F	35	B	2015-03-07	사망	2015-12-03	
17	M	62	B	2015-04-14	생존		
18	M	59	A	2015-04-17	소식불명		15年 8月 5日
19	F	75	A	2015-04-26	사망	15年 12月 15日	
20	M	71	B	2015-05-09	사망	15年 9月 3日	
21	M	60	A	2015-06-24	생존		
22	M	77	B	2015-06-27	사망	2015-08-30	
23	F	42	A	7월 23일	생존		
24	M	54	B	8월 5일	생존		
25	F	63	A	2015-08-19	사망	2015-10-25	
26	M	59	B	2015-09-23	생존		
27	M	72	B	2015-09-26	사망	2015-11-05	
28	F	44	A	2015-09-26	생존		
29	M	67	A	2015-10-04	생존		
30	M	50	A	Nov-15	생존		

환자번호	성별	연령	치료법	치료개시일	2015.12.31 생존 여부	사망연월일	최종생사여부확인일	추적관찰 종료일	생존일수	생존개월수	
1	M	62	A	2014-04-03	사망	2014-06-15		2014-06-15	73	2	
2	F	57	A	2014-04-16	생존			2015-12-31	624	21	+
3	M	49	B	2014-05-12	사망	2014-10-18		2014-10-18	159	5	
4	M	72	B	2014-06-16	사망	2014-08-04		2014-08-04	49	2	
5	F	63	A	2014-06-18	소식불명		2014-06-25	2014-06-25	7	0	+
6	M	51	A	2014-07-26	생존			2015-12-31	523	17	+
7	F	68	B	16-Aug-14	사망	2014-11-02		2014-11-02	78	3	
8	M	40	B	23-Aug-14	소식불명		2014-12-20	2014-12-20	119	4	+
9	F	38	A	27-Sep-14	생존			2015-12-31	460	15	+
10	F	67	B	2014-10-16	생존			2015-12-31	441	15	+
11	M	81	B	2014-10-26	사망	2015-11-19		2015-11-19	389	13	
12	M	54	A	2014-11-11	사망	2015-12-20		2015-12-20	404	13	
13	M	57	A	2015-01-14	생존			2015-12-31	351	12	+
14	F	63	A	2015-01-20	생존			2015-12-31	345	11	+
15	M	48	B	2015-02-05	생존			2015-12-31	329	11	+
16	F	35	B	2015-03-07	사망	2015-12-03		2015-12-03	271	9	
17	M	62	B	2015-04-14	생존			2015-12-31	261	9	+
18	M	59	A	2015-04-17	소식불명		15年 8月 5日	2015-08-05	110	4	+
19	F	75	A	2015-04-26	사망	15年 12月 15日		2015-12-15	233	8	
20	M	71	B	2015-05-09	사망	15年 9月 3日		2015-09-03	117	4	
21	M	60	A	2015-06-24	생존			2015-12-31	190	6	+
22	M	77	B	2015-06-27	사망	2015-08-30		2015-08-30	64	2	
23	F	42	A	7월 23일	생존			2015-12-31	161	5	+
24	M	54	B	8월 5일	생존			2015-12-31	148	5	+
25	F	63	A	2015-08-19	사망	2015-10-25		2015-10-25	67	2	
26	M	59	B	2015-09-23	생존			2015-12-31	99	3	+
27	M	72	B	2015-09-26	사망	2015-11-05		2015-11-05	40	1	
28	F	44	A	2015-09-26	생존			2015-12-31	96	3	+
29	M	67	A	2015-10-04	생존			2015-12-31	88	3	+
30	M	50	A	Nov-15	생존			2015-12-31	48	2	

① 날짜변수 연,월,일 순서 자리수 등 확인
② 추적관찰 종료일 : 사망-사망일, 생존-관찰종료일, 소식불명-최종생사여부확인일
③ 생존일수 : 추적관찰종료일-치료개시일

생존분석의 기본 가정

- **생존/사망 - 완전한 무작위(random) 현상**

  → 관찰의 시작/종료 - 생존여부와 무관

  – 생존/사망에 영향을 주지 않음

- **중도절단(Censoring) – 완전한 무작위(random) 현상**

  "non-informative" censoring

  → Censoring은 시간에 따른 생존여부와 무관

  Censoring은 생존여부에 대해 독립적

생존함수

- T: **생존시간을 나타내는 확률변수**
- t: **특정 시간**
- **생존함수**: 환자가 t 시간 이상 생존할 확률,
- t시점까지의 구간 생존확률의 누적값
- $S(t) = P(T > t)$
- $S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t)$
- $F(t) = P(T \leq t) = 1 - S(t)$
- 확률밀도함수
- $f(t) = \frac{dF(t)}{dt} = \frac{d(1-S(t))}{dt} = -\frac{dS(t)}{dt}$

생존함수와 위험함수

- **위험함수** : t 시점까지는 생존했다고 가정하고 바로 직후 순간적으로 사망할 조건부확률
- $h(t) = \lim_{\delta t \to 0} \left\{ \frac{P(t<T\le t+\delta t \mid T>t)}{\delta t} \right\} = \frac{f(t)}{S(t)}$
- $h(t) = -\frac{dS(t)/dt}{S(t)} = -\frac{d}{dt}\{logS(t)\}$
- $S(t) = exp\left\{-\int_0^t h(x)dx\right\} = \exp\{-H(t)\}$ **누적위험함수**

Fundamentals of survival analysis 13/88

생존자료의 요약 - 누적생존율 추정

- 모수적 (parametric)
  - 생존시간의 분포형태가 알려져 있을 경우
  - 지수분포, 와이블분포, 로그-정규분포, 로그-로지스틱분포 등
  - 공업제품에 대한 실험결과를 분석하는데 흔히 적용됨
- 비모수적(non-parametric)
  - 사람을 대상으로 하는 연구에서 주로 쓰임
  - 생명표법 (life-table method, actuarial method, Cutler-Ederer method)
  - 누적한계 추정법 (product-limit method, Kaplan-Meier method)

Fundamentals of survival analysis 14/88

생존자료 요약: Kaplan-Meier 방법

- 어떤 특정 시점(t_j)보다 더 생존할 확률 계산
- $S(T= t_j) = Pr(T > t_j)$

중도절단 (Censored)

4 | 4.5+ | 10 | 14 | 20+ | 24

Event 발생
(4/6=66.7%)

Time	Event
4	1
4.5	0
10	1
14	1
20	0
24	1

분석자료

Kaplan-Meier 방법

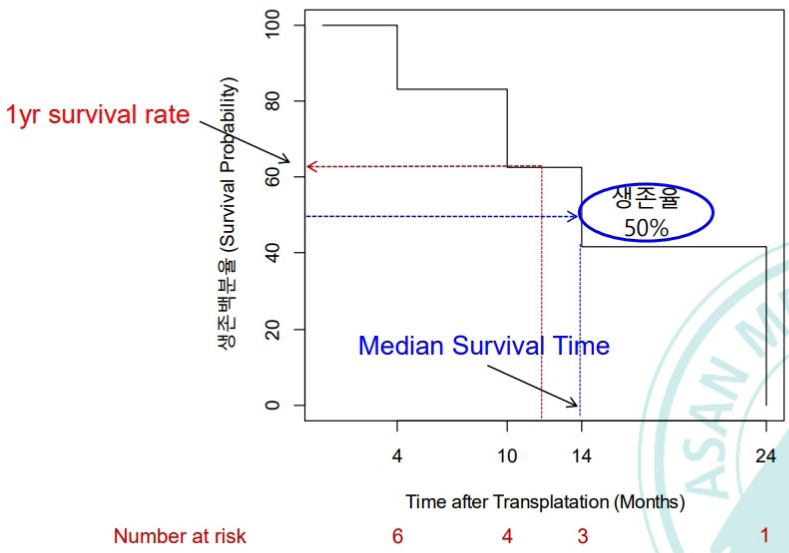
- 간이식후 사망까지 시간

4 4.5+ 10 14 20+ 24

시간 (개월)	초기 생존자	사망자	각 시간의 사망자 분율	각 시간의 생존자 분율	누적생존율
	①	②	③=②/①	④=1-③	⑤
4	6	1	0.167	0.833	0. 833
10	4	1	0.250	0.750	0.625
14	3	1	0.333	0.667	0.417
24	1	1	1	0	0.000

위험 노출 수는 그 시점에서 초기 생존자 수

Kaplan-Meier 곡선



두 집단 이상에서의 생존율 비교

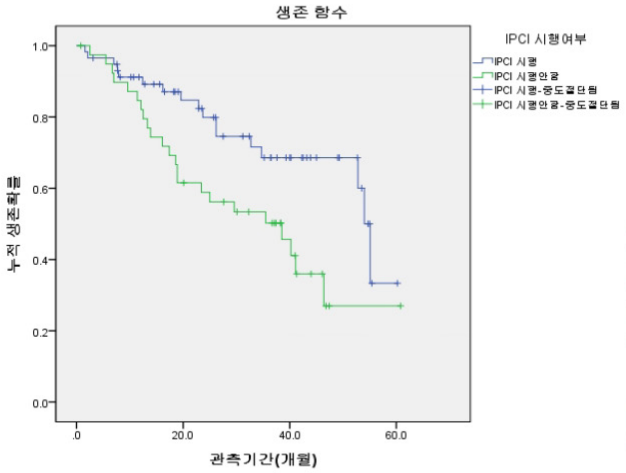
- 모수적 (parametric)
  - Likelihood ratio test
- 비모수적(non-parametric)
  - Log-rank method (generalized Mantel-Haenszel method)
  - Gehan's generalized Wilcoxon method
  - Tarone-Ware test
  - Peto-Peto test

생존곡선의 평가

- 두 개 이상의 생존 곡선을 통계적으로 비교

→ Log-rank test

(로그순위 검정)



로그-순위 검정 (Log-rank test)

- 두 개 이상의 생존 곡선을 통계적으로 비교
  - 두 그룹 비교일 때
  - 귀무가설 (H_0) : 생존곡선들은 차이가 없다.

    ($S_1(t) = S_2(t)$ for all t, 모든 시점)
  - 대립가설 (H_1): 적어도 두 개의 생존곡선의 차이가 있다.

    ($S_1(t) \neq S_2(t)$ for some t, 시점이 존재한다.)
- **관찰된 모든 시점에서** 평균적인 생존율의 차이 평가
- 전 구간에 걸쳐 일정한 차이가 있을 때 가장 검정력 높음

  (생존곡선의 교차가 있을 경우에는 부적절)

- 두 그룹을 섞은 후 $t_1 < t_2 < \ldots t_k$ 으로 정리한 후 모든 t_i에서

	사망	생존	계
그룹 1	D_{1i}	$N_{1i}-D_{1i}$	N_{1i}
그룹 2	D_{2i}	$N_{2i}-D_{2i}$	N_{2i}
계	D_i	N_i-D_i	N_i

- N_{1i}, N_{2i}, D_i가 고정되어 있다고 가정하면
  D_{1i}~ 초기하분포(hypergeometric distribution)
- 평균 $E(D_{1i})=E_{1i}=N_{1i}\,D_i/N_i$
- 분산 $V(D_{1i})=V_{1i}=\frac{N_{1i}N_{2i}}{N_i-1}\frac{D_i(N_i-D_i)}{N_i^2}$

$$T=\frac{\left\{\sum_{i=1}^{k}(D_{1i}-E_{1i})\right\}^2}{\sum_{i=1}^{k}V_{1i}}\sim\chi^2_{g-1}$$

- $T>x^2$ 임계치, 귀무가설 기각

- 로그-순위 검정법은 각 시점에서 같은 가중치를 준다 $(W_i=1)$

$$T=\frac{\{\sum_{i=1}^{k}w_i(D_{1i}-E_{1i})\}^2}{\sum_{i=1}^{k}w_i^2V_{1i}}$$

- 대안
  - 연구의 초기 차이에 보다 많은 비중을 둔 것
  - Number at risk에 비례
  - Gehan (Wilcoxon)의 방법

$$w_i=N_i/(N+1)$$

  - Tarone-Ware방법

$$w_i=\sqrt{N_i/(N+1)}$$

생존자료 분석 예제

- 예제: 관상동맥 질환(coronary artery disease)를 앓고 있는 환자에서 drug-eluting stent (DES) 삽입 시술 환자와 관상동맥우회술 (coronary artery bypass grafting, CABG) 환자의 예후 비교 연구*(American Journal of Cardiology, 2012;109:1548-1557)*
  - 환자의 예후를 시술시점부터 adverse event (①사망, ②composite outcome : death, MI or stroke, ③혈관재생수술) 를 경험할 때까지 시간으로 정의 (특정 intervention에 대한 비교연구이므로 시술시점을 on-set time 으로 고려)

생존자료 분석 예제(SPSS)

- Coronary artery disease 환자의 생존자료

중도절단여부
사망=1
중도절단=0

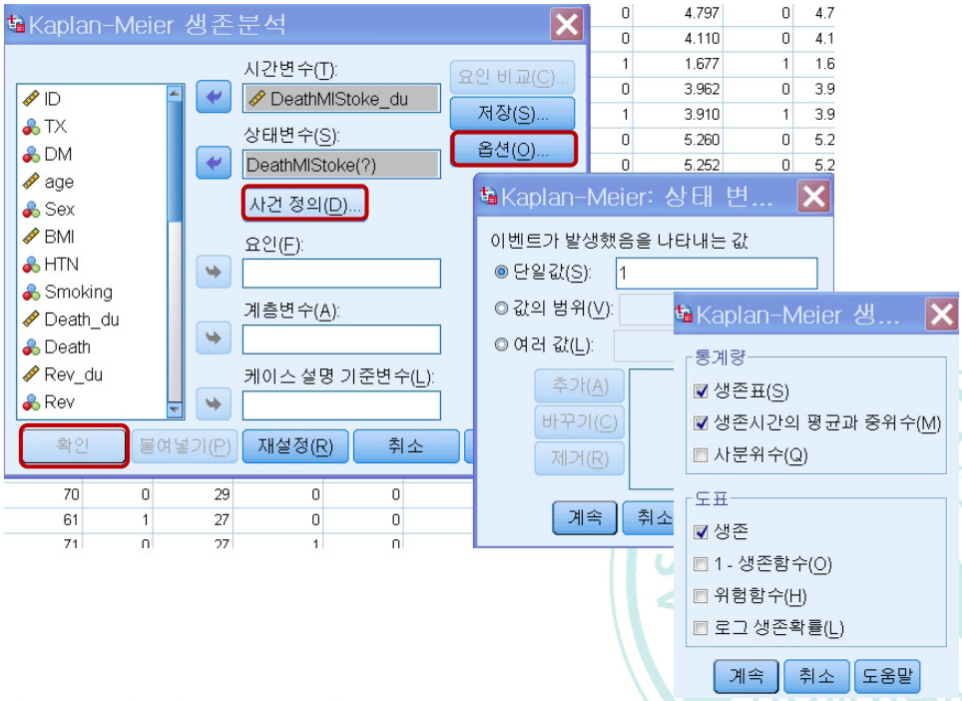
생존관련 예후요인

생존시간(년)

ID	TX	DM	age	Sex	BMI	HTN	Smoking	DeathMIStoke_du	DeathMIStoke
1	1	2	45	1	22	0	0	4.797	0
2	1	2	51	1	25	0	1	4.110	0
3	1	1	71	1	25	1	0	1.677	1
4	1	2	88	0	23	0	0	3.962	0
5	0	2	83	1	22	1	0	2.049	1
6	1	1	44	1	25	1	1	5.260	0
7	1	2	54	1	23	0	1	5.252	0
8	1	1	63	0	24	1	1	5.137	0
9	1	2	63	0	25	0	1	3.751	1
10	0	2	56	1	23	1	1	6.447	0
11	1	2	51	1	27	0	1	5.164	0

생존확률 계산: K-M 방법(SPSS)

- 분석 → 생존확률 → Kaplan-Meier 생존분석



Kaplan-Meier 분석 결과(SPSS)

케이스 처리 요약

합계 N	사건 수	중도절단	
		N	퍼센트
300	43	257	85.7%

생존표

	시간	상태	시간에 누적 생존 비율		누적 사건 수	남아 있는 케이스 수
			추정값	표준 오차		
1	.000	1	.997	.003	1	299
2	.008	1	.993	.005	2	298
3	.055	1	.990	.006	3	297
4	.066	1	.987	.007	4	296
5	.071	1	.983	.007	5	295
6	.074	0	.	.	5	294
7	.079	1	.980	.008		
8	.148	1	.977	.009		
9	.216	1	.973	.009		
10	.255	0	.	.		
11	.285	0	.	.		
12	.474	0	.	.		
13	.501	0	.	.		
14	.578	1	.970	.010		

- ✓ 0.074년 때 Number at risk 295명
- ✓ 사망 5명, 중도절단 1명
- ✓ 누적생존확률: 0.983
- ✓ 사망 발생마다 생존율 추정

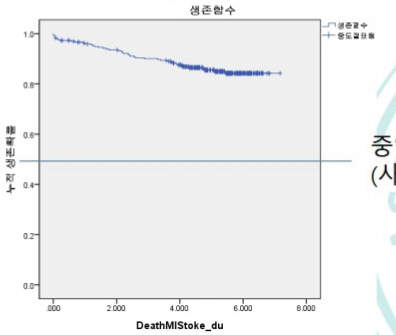
생존시간의 평균과 중위수 추정(SPSS)

생존 시간에 대한 평균 및 중위수

평균[a]				중위수			
		95% 신뢰구간				95% 신뢰구간	
추정값	표준 오차	하한	상한	추정값	표준 오차	하한	상한
6.422	.109	6.208	6.637	.	.	.	.

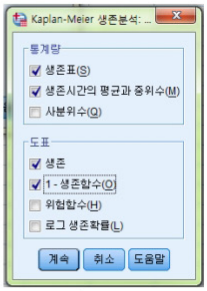
a. 중도절단된 경우 추정값은 가장 큰 생존시간으로 제한됩니다.

- 평균 생존시간(95% CI)은 6.422(6.208 – 6.637)로 추정
- 사건이 50% 미만으로 발생하였기 때문에 median survival time 추정 불가

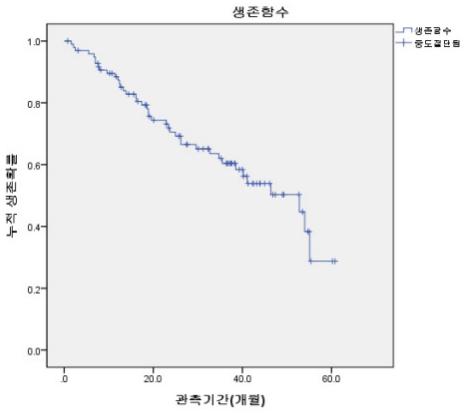


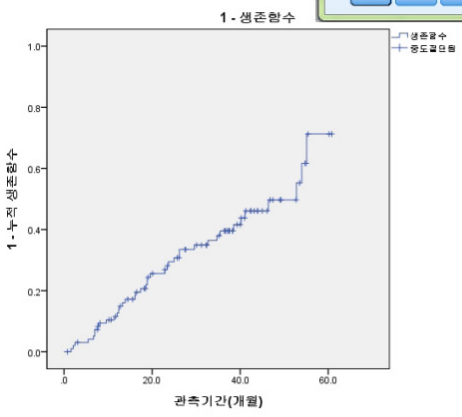
중앙 생존 시간 추정 불가
(사건 발생 50% 시점)

Kaplan-Meier 곡선(SPSS)



- 변수해석에 따라 (1–생존)곡선을 그리기도 함



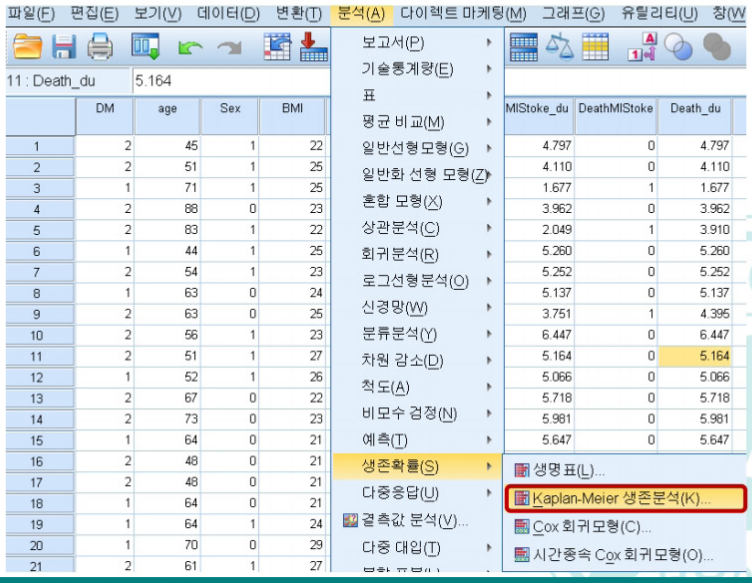


Log-rank test 예제(SPSS)

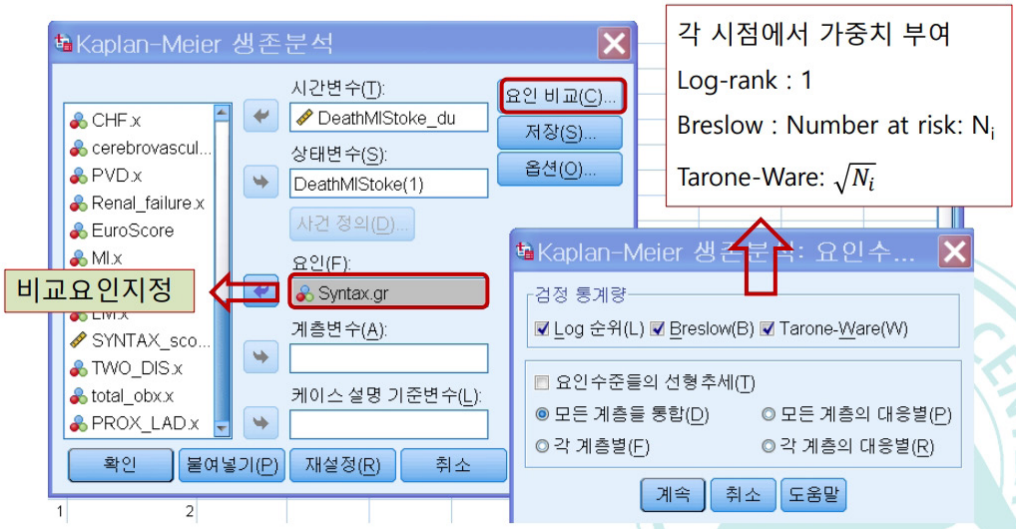
- 세 군 비교: SYNTAX score에 따른 adverse event rate 비교 (low: <=22, intermediate: 23-32, high: >=33)
  - SYNTAX score 정도에 따라 AE rate의 차이가 있는가 ?

Log-rank test 예제(SPSS)

- 분석 → 생존확률 → Kaplan-Meier 생존분석



- **SYNTAX score 정도에 따라 AE rate의 차이가 있는가 ?**



각 시점에서 가중치 부여

Log-rank : 1

Breslow : Number at risk: N_i

Tarone-Ware: $\sqrt{N_i}$

Kaplan-Meier 분석 결과 (SPSS)

케이스 처리 요약

syntax.gr	합계 N	사건 수	중도절단 N	중도절단 퍼센트
low (22이하)	178	18	160	89.9%
intermediate (23-32)	86	21	65	75.6%
high (33이상)	36	4	32	88.9%
전체	300	43	257	85.7%

생존표

생존율 추정
(syntax.gr: low)

syntax.gr		시간	상태	시간에 누적 생존 비율 추정값	시간에 누적 생존 비율 표준 오차	누적 사건 수	남아 있는 케이스 수
low (22이하)	1	.066	발생	.994	.006	1	177
	2	.285	미발생	.	.	1	176
	3	.474	미발생	.	.	1	175
	4	.501	미발생	.	.	1	174
	5	.578	발생	.989	.008	2	173
	6	.638	미발생	.	.	2	172
	7	.770	미발생	.	.	2	171
	8	.822	미발생	.	.	2	170
	9	.932	발생	.983	.010	3	169
	10	.975	발생	.977	.011	4	168

생존율 추정
(syntax.gr: intermediate)

intermediate (23-32)	1	.008	발생	.988	.012	1	85
	2	.055	발생	.977	.016	2	84
	3	.071	발생	.965	.020	3	83
	4	.074	미발생	.	.	3	82
	5	.148	발생	.953	.023	4	81
	6	.216	발생	.942	.025	5	80
	7	.255	미발생	.	.	5	79
	8	.671	발생	.930	.028	6	78
	9	1.249	발생	.918	.030	7	77
	10	1.764	발생	.906	.032	8	76

생존율 추정
(syntax.gr: high)

	85	6.751	미발생	.	.	21	1
	86	6.797	미발생	.	.	21	0
high (33이상)	1	.000	발생	.972	.027	1	35
	2	.079	발생	.944	.038	2	34
	3	.970	미발생	.	.	2	33
	4	2.181	발생	.916	.047	3	32
	5	3.701	미발생	.	.	3	31
	6	4.038	발생	.886	.054	4	30
	7	4.449	미발생	.	.	4	29
	8	4.814	미발생	.	.	4	28
		4.953	미발생	.	.	4	27
		5.104	미발생	.	.	4	26

마지막 추적 관찰 기간까지
50% 이상 생존 :
중앙생존 시간 추정 불가

생존 시간에 대한 평균 및 중위수

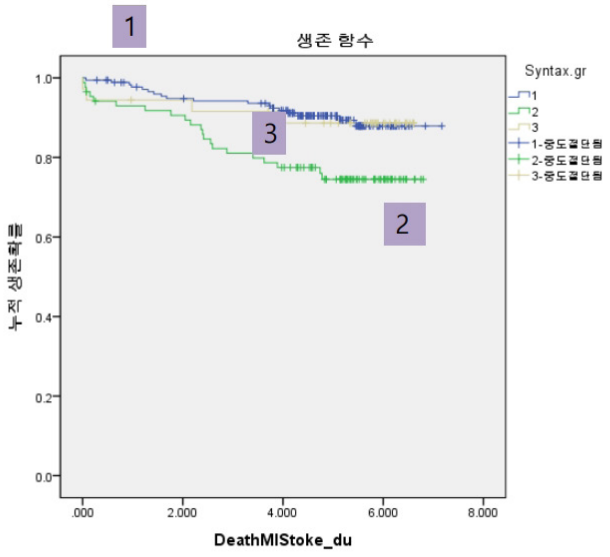
Syntax.gr	평균[a]				중위수			
			95% 신뢰구간				95% 신뢰구간	
	추정값	표준 오차	하한	상한	추정값	표준 오차	하한	상한
1	6.652	.119	6.418	6.886	.	.	.	.
2	5.616	.235	5.155	6.077	.	.	.	.
3	6.043	.282	5.490	6.596	.	.	.	.
전체	6.422	.109	6.208	6.637	.	.	.	.

- 평균 AE 경험시간은 syntax grading=2일 때 가장 빠르다. (2<3<1)
- 중앙생존시간은 사건이 50%미만으로 발생하여 추정 불가

전체 비교

	카이제곱검정	자유도	유의확률
Log Rank (Mantel-Cox)	9.777	2	.008
Breslow (Generalized Wilcoxon)	10.351	2	.006
Tarone-Ware	10.162	2	.006

SYNTAX grading은 AE rate과 관련성이 있음 (P=0.008)



SYNTAX grading이 AE rate과 관련성이 있다면 어느 군에서 유의한지 사후검정으로 확인가능!

- **SYNTAX grading에 따른 사후 검정**

대응별 비교

	Syntax.gr	1		2		3	
		카이제곱검정	유의확률	카이제곱검정	유의확률	카이제곱검정	유의확률
Log Rank (Mantel-Cox)	1			8.979	.003	.004	.951
	2	8.979	.003			2.624	.105
	3	.004	.951	2.624	.105		
Breslow (Generalized Wilcoxon)	1			9.907	.002	.080	.777
	2	9.907	.002			2.367	.124
	3	.080	.777	2.367	.124		
Tarone-Ware	1			9.578	.002	.034	.853
	2	9.578	.002			2.499	.114
	3	.034	.853	2.499	.114		

- Grade 1 과 2의 AE율의 차이 유의 (P=0.003)
- Grade 1 과 3의 AE율의 차이는 유의하지 않음 (P=0.951)
- Grade 2 과 3의 AE율의 차이는 유의하지 않음 (P=0.105)

- 다중비교에 의한 1종 오류(type I error) 증가를 막기 위해 유의수준 0.05/3=0.016 사용(Bonferroni correction)
- 유의수준 0.016에서 grade 1 vs 2 사이의 AE율의 유의한 차이가 있었고 grade 1 vs 3 또는 grade 2 vs 3 은 유의한 차이가 있다고 할 수 없음

생존자료 분석 예

Purpose To investigate differences in VDT (Volume Doubling Times) between the predominant histologic subtypes of primary lung adenocarcinomas and to assess the correlation between VDT and prognosis.

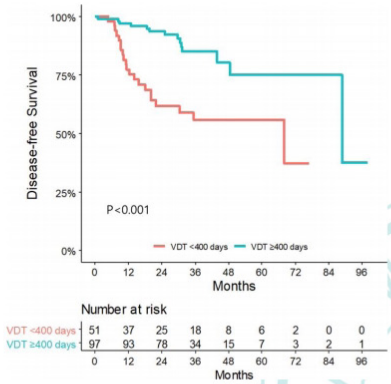


Figure 4b: Kaplan-Meier curves for disease-free survival. **(a)** Kaplan-Meier curves for prognosis-based subtype groups and **(b)** for v olume doubling time (VDT) class (<400 days and ≥400 days) are plotted for the survival analysis of 148 patients. *P* values were obta ined by using the log-rank test.

Radiology. 2020 Jun;295(3):703-712

생존자료 분석 예

Purpose To develop and validate a preoperative risk scoring system using clinical and CT variables to predict recurrence-free survival (RFS) after upfront surgery in patients with resectable PDAC.

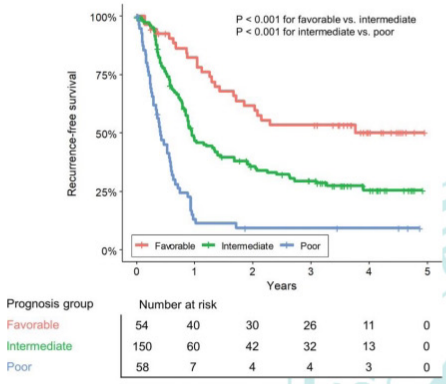


Figure 3a: Graphs show recurrence-free survival curves of three prognosis groups based on risk score in **(a)** development set and test set according to CT interpretations of **(b)** reader 1 and **(c)** reader 2.

Radiology. 2020 Sep;296(3):541-551

생존자료에 대한 회귀분석

- 모수적(parametric) 방법
  - 생존 시간(T)에 대해 특정 분포를 미리 가정
  - 각 환자의 생존시간, 생존여부를 종속변수로 하여 특정 모형 적합
- 반모수적(semiparametric) 방법
  - 생존 시간에 대한 특정 분포를 미리 가정하지 않는다. 단지 위험 요인(X)들의 결합에 대해서 특정 형태를 가정
- 콕스 비례위험 모형(Cox proportional hazard(PH) model)

$$h(t;X) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)$$

Linear in the X's

Baseline hazard function (기저 위험함수)
모든 X값이 0일 때의 위험함수

콕스 비례 위험 모형 (Cox Proportional Hazards Model)

- 위험함수와 위험요인 (risk factors) 사이의 관련성을 모형화하는 것이 목적
  - 환자의 예후에 영향을 주는 위험요인을 찾아낼 수 있음
  - 주요한 예후 인자를 이용하여 예측모형 개발
  - 교란변수의 영향을 보정하여, 치료법, 수술법 등의 효과를 추정

➢ $h_0(t)$에 대한 특정 분포를 가정하지 않더라도 β에 대한 추정 가능

➢ 위험의 추론은 hazard ratio(HR)로 이루어 지고 β를 통해 계산

Hazard Ratio 계산

- 두 개 hazard rates의 비(ratio): the hazard for one individual divided by the hazard for a different individual

$$\widehat{HR} = \frac{\hat{h}(t, X^*)}{\hat{h}(t, X)} = \frac{\hat{h}_0(t) e^{\sum_{i=1}^{p} \hat{\beta}_i X_i^*}}{\hat{h}_0(t) e^{\sum_{i=1}^{p} \hat{\beta}_i X_i}} = e^{\sum_{i=1}^{p} \hat{\beta}_i (X_i^* - X_i)}$$

여기서 ***X**** 와 ***X*** 는 각 환자의 관측된 위험요인

예, ***X**** = (X^*_1, X^*_2,⋯,X^*_p), where X^*_1 =1 for DES group

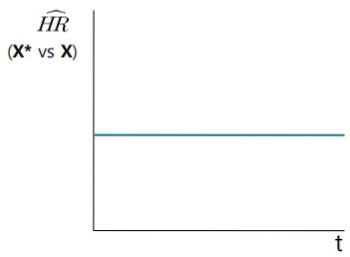
그리고 ***X*** = (X_1, X_2,⋯,X_p), where X_1 =0 for CABG group

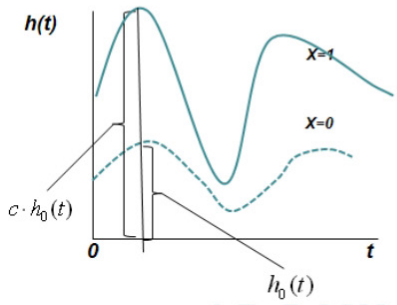
$$\boxed{\widehat{HR} = \exp[\hat{\beta}_1 (X_1^* - X_1)] = \exp[\hat{\beta}_1 (1-0)] = \exp(\hat{\beta}_1)}$$

Hazard Ratio 해석

- $\widehat{HR}$ >1, DES환자(X_1=1)위험이 CABG 환자(X_1=0)의 위험보다 높다.
- $\widehat{HR}$ <1, DES환자(X_1=1)위험이 CABG 환자(X_1=0)의 위험보다 낮다.
- $\widehat{HR}$ =1, DES환자(X_1=1)위험은 CABG 환자(X_1=0)의 위험과 같다.

비례위험(Proportional Hazards)





- 예측된 HR는 constant (not dependent on time)
- Hazard function for one individual is proportional to the hazard function for another individual, where the proportionality constant(c), which does not dependent on time.

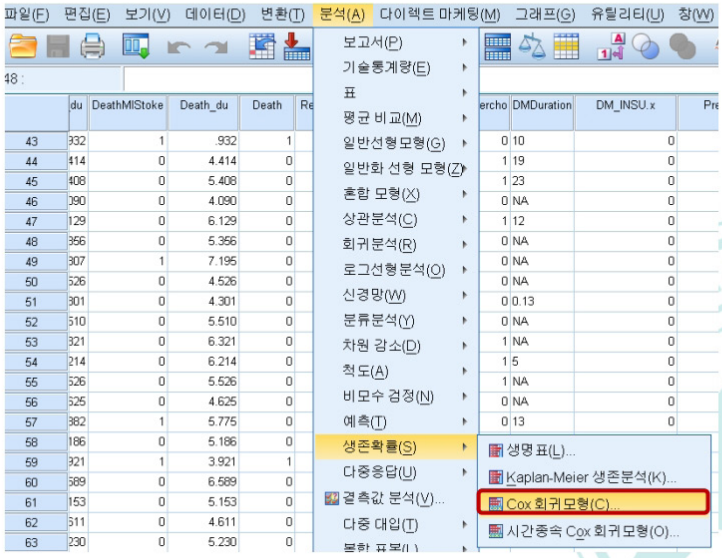
$$\hat{h}(t, X^*) = c \times \hat{h}(t, X)$$

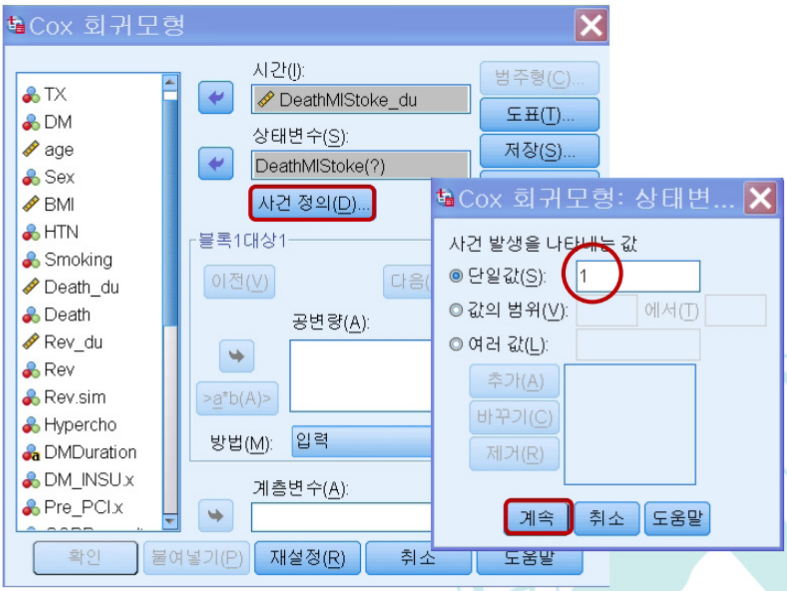
Cox PH 회귀모형 예제

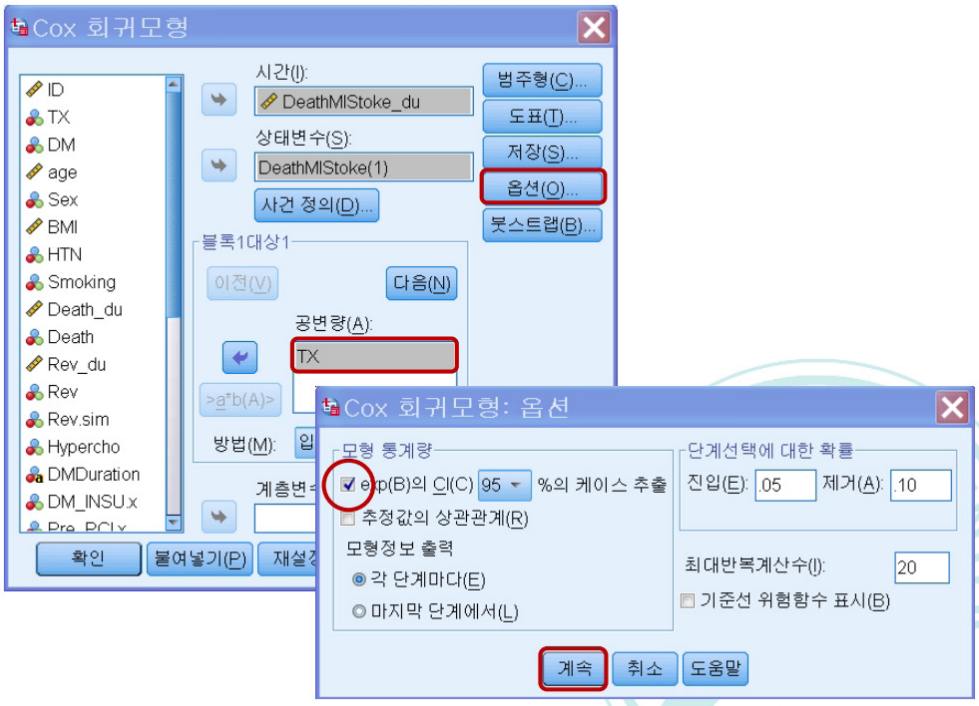
- 단변량 분석 (Univariate analysis)
  - 시술법에 따른 hazard rate 비교
  - SYNTAX grade에 따른 hazard rate 비교
- 다변량 분석 (Multivariable analysis)
  - 다른 유의미한 covariate으로 보정했을 때 시술법에 따른 hazard rate 비교

Cox PH 회귀모형(Uni.) 예제 (SPSS)

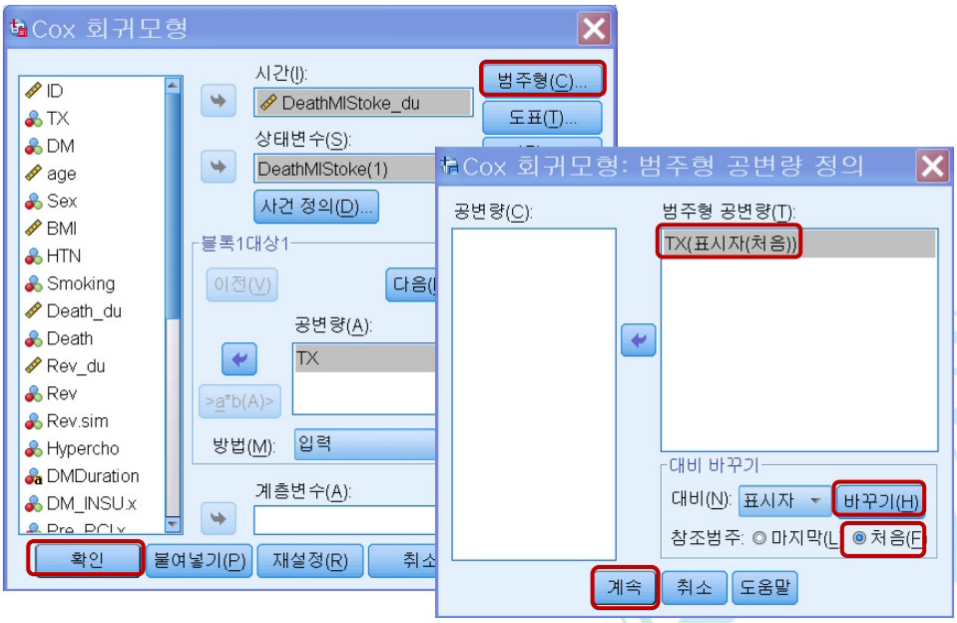
- 분석 → 생존확률 → Cox 회귀모형







- 가변수 설정: TX=0(CABG) 참조범주 설정



Cox PH 회귀모형 결과 (SPSS)

케이스 처리 요약

		N	퍼센트
분석가능한 케이스	사건[a]	43	14.3%
	중도절단	257	85.7%
	전체	300	100.0%
삭제 케이스	결측 케이스	0	0.0%
	음의 시간을 갖는 케이스	0	0.0%
	계층에서 가장 최근 사건 이전까지의 중도절단 케이스	0	0.0%
	전체	0	0.0%
전체		300	100.0%

a. 종속변수 DeathMIStoke_du: DeathMIStoke_du

범주형 변수 코딩[a]

		빈도	(1)
TX[b]	0	91	0
	1	209	1

a. 범주변수: TX

b. 표시형 파라미터 코딩

- TX는 두 개의 수준이므로 reference coding=0

방정식의 변수

	B	표준오차	Wald	자유도	유의확률	Exp(B)	Exp(B)에 대한 95.0% CI	
							하한	상한
TX	-.623	.307	4.102	1	.043	.537	.294	.980

→ TX=1(DES)일 때 TX=0(CAGB)에 비해 hazard ratio는 0.537배, 즉 DES는 위험율을 46% 정도 낮춤. (P=0.043, 95% CI for HR = (0.294 to 0.980))

- **만약 명목형 변수가 세 수준(이상)이라면..**

범주형 변수 코딩[a]

		빈도	(1)	(2)
Syntax.gr[b]	1	178	0	0
	2	86	1	0
	3	36	0	1

a. 범주변수: Syntax.gr

b. 표시형 파라미터 코딩

참조범주(grade=1) 대비 grade=2일 때 1 증가

참조범주(grade=1) 대비 grade=3일 때 1 증가

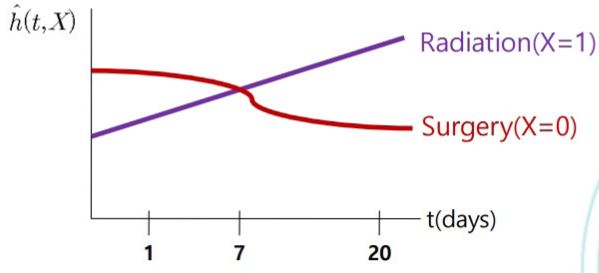
방정식의 변수

	B	표준오차	Wald	자유도	유의확률	Exp(B)	Exp(B)에 대한 95.0% CI 하한	상한
Syntax.gr			9.114	2	.010			
변수 이름 Syntax.gr(1)	.930	.321	8.379	1	.004	2.535	1.350	4.758
변수 이름 Syntax.gr(2)	.051	.553	.009	1	.926	1.052	.356	3.113

- SYNTAX grade는 overall 하게 유의함 (P = 0.010)
- Grade 1 과 2는 유의하게 차이 (P=0.004);
  (Grade 2 의 hazard rate는 Grade 1에 비해 2.535배 높다)
- Grade 1 과 3는 유의한 차이 없음 (P=0.926)

If Hazard functions cross…

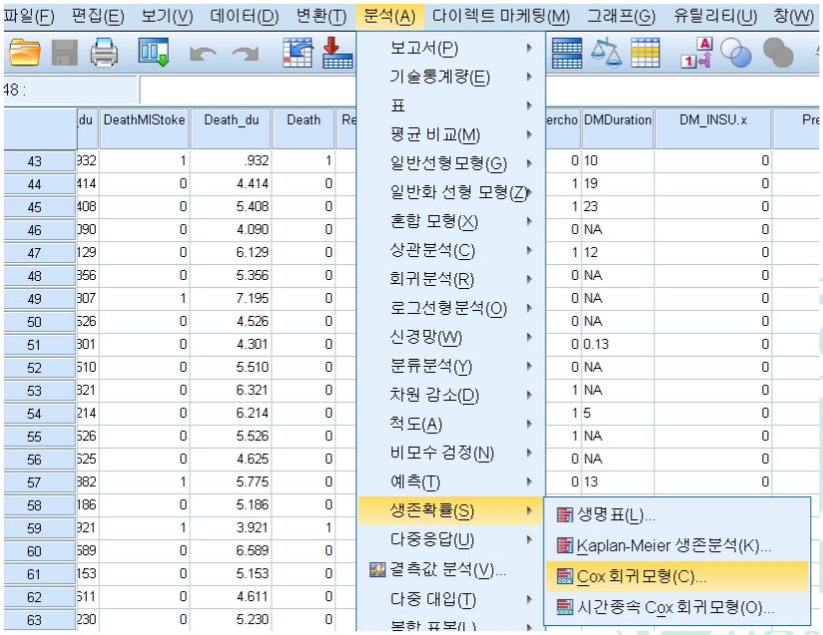
- 예) Cancer 환자에 대해 수술법과 방사선요법 비교
  - 만약 수술로 종양을 제거한 후 early time 에 합병증으로 high risk 존재하지만 일단 early critical period를 지나면 surgery 의 benefit 이 훨씬 크다면…

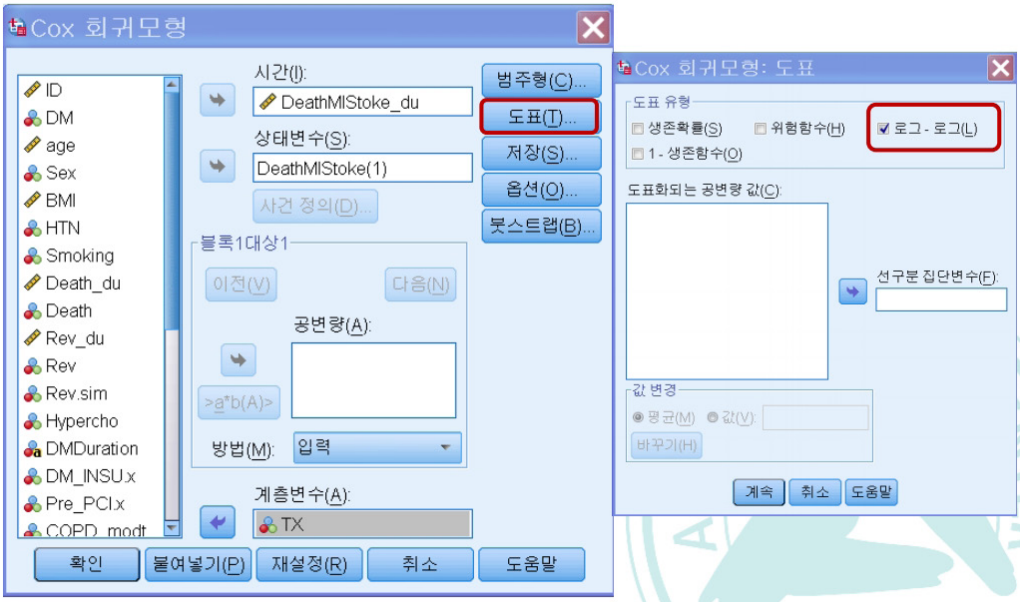


- 1 days:
  $$\frac{\hat{h}(t=1,X=1)}{\hat{h}(t=1,X=0)} < 1$$
  but
- 20 days:
  $$\frac{\hat{h}(t=20,X=1)}{\hat{h}(t=20,X=0)} > 1$$

→ PH model is not appropriate

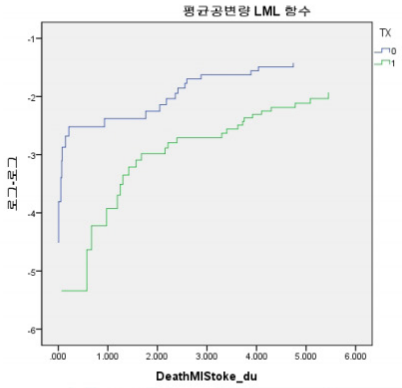
비례위험 가정 확인 : LML 도표





비례 위험 가정 확인(1)

- 로그-로그 Plot 사용 (LML plot)
  - 각 시점에서 생존율을 log[-log(S)] 변환
  - Time과 변환시킨 log[-log(S)] 값으로 그래프를 나타냄
  - 모든 시점에서 두 군의 생존율 차이가 일정하면, 그 요인은 PH 가정을 만족하는 것임



- 두 군의 생존율 차이가 일정
  → PH 가정 만족
  → 독립변수 효과는 시간에 관계없이 일정

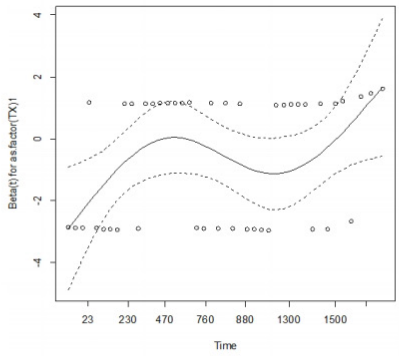
비례 위험 가정 확인 (2)

- Schoenfeld residuals을 이용한 검정
- Schoenfeld residuals defined for
  - every subject who has event
  - each covariate in model

  → 각 환자로부터 각 변수에 대해 Schoenfeld residual 계산
- Schoenfeld residual이 time과 correlation이 없다면 PH가정 만족
- 편잔차(P) in SPSS

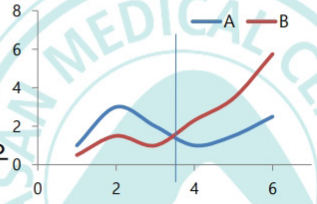
비례 위험 가정: Schoenfeld residuals R

```
> schfit=cox.zph(fit, transform='km')
> schfit
                  rho chisq      p
as.factor(TX)1 0.276  3.31 0.0688
> plot(schfit)
```



When PH assumption not satisfied

- Use time-dependent variables
  - Defined to analyze a time-independent predictor not satisfying the PH assumption
- Stratified Cox model
  - PH가정을 만족하지 않는 변수를 층화변수로 이용
- Partition the time axis
  - 짧은 기간 내에서 PH가정이 만족하는 경우
- Accelerated failure time or additive hazards model



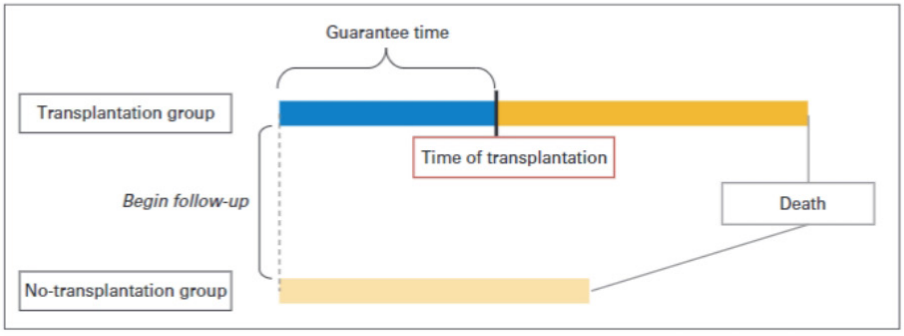
Time independent vs dependent

- Time independent variables
  - $h(t, X) = h_0(t)exp(\sum_{i=1}^{p1} \beta_i X_i)$
    - ✓ 시간에 따라 변하지 않는 변수 (Baseline characteristics 등), 또는 시점이 정해진 변수(수술 전, 진단 전 변수) 등
- Time dependent variables
  - $h(t, X(t)) = h_0(t)exp(\sum_{j=1}^{p2} \beta_j X_j(t))$
- Extended Cox Model
  - $h(t, X(t)) = h_0(t)exp(\sum_{i=1}^{p1} \beta_i X_i + \sum_{j=1}^{p2} \beta_j X_j(t))$

Time-varying covariates

- 연구의 관찰기간 동안 group의 분류가 변화하는 경우
- transplantation
- seroconversion
- the occurrence of objective disease response
- use of drug
- onset of toxicity

Immortal time bias



- immortal time bias, guarantee time bias, survivor bias, and survivor treatment selection bias

Time-varying covariates

- time-varying X(t)
  - Ex) Heart transplant status at time t_0

    HT(t)=1 if received transplant at some time $t_0 \leq t$

    HT(t)=0 if did not receive transplant by time t

    ✓ Transplant H(t): 0000...011111 (t_0)

    ✓ No transplant H(t): 0000...00000
- $h(t, X(t)) = h_0(t) exp(\delta \cdot HT(t))$
- δ represents the overall effect of HT(t)
  - But, PH is not satisfied
  - HR(t) is time-dependent because HT(t) is time-dependent

Research Article

EASL | JOURNAL OF HEPATOLOGY

Incidence of hepatocellular carcinoma after HBsAg seroclearance in chronic hepatitis B patients: A need for surveillance

Gi-Ae Kim[1], Han Chu Lee[1,*], Min-Ju Kim[2], Yeonjung Ha[1], Eui Ju Park[1], Jihyun An[1], Danbi Lee[1], Ju Hyun Shim[1], Kang Mo Kim[1], Young-Suk Lim[1]

[1]*Department of Gastroenterology, Asan Liver Center, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea;*
[2]*Department of Clinical Epidemiology and Biostatistics, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea*

Background & Aims: Little is known about whether surveillance for hepatocellular carcinoma (HCC) is worthwhile in chronic hepatitis B virus (HBV)-infected patients who have achieved HBsAg seroclearance.
Methods: A retrospective analysis of 829 patients (mean age: 52.3 years; 575 males; 98 with cirrhosis) achieving HBsAg seroclearance was performed at a tertiary hospital in Korea between 1997 and 2012. We evaluated incidence rates of HCC, and validated CU-HCC score based on data at the time of HBsAg seroclearance.
Results: During a follow-up of 3464 patient-years, 19 patients developed HCC (annual rate: 0.55%). Liver cirrhosis (hazard ratio [HR]: 10.80; 95% confidence interval [CI]: 4.25–27.43), male gender (HR: 8.96; 95% CI: 1.17–68.80), and age ⩾50 years at the time of HBsAg seroclearance (HR: 12.14; 95% CI: 1.61–91.68) were independently associated with HCC. The estimated annual incidence of HCC was 2.85% and 0.29% in patients with and without cirrhosis, respectively. Among the non-cirrhotic patients, the annual rate of HCC was higher in the male patients than in the females (0.40% *vs.* 0%, respectively), and all the HCCs developed after age 50. The time-dependent area under the receiver operating characteristic curves for the CU-HCC score for 5 year and 10 year HCC prediction were 0.85 and 0.74, respectively.
Conclusions: HCC surveillance should be considered for cirrhotic patients and non-cirrhotic male patients over age 50, even after HBsAg seroclearance, especially those infected with HBV genotype C. HBsAg seroclearance at age ⩾50 years was also an independent predictor for HCC.

Introduction

Hepatitis B surface antigen (HBsAg) seroclearance is considered to be the most important end point of chronic hepatitis B virus (HBV) infection [1–4] because both spontaneous and therapy-induced-HBsAg seroclearance are associated with histological improvement, a reduced risk of hepatocellular carcinoma (HCC), and prolonged survival [5–11]. However, several reports have shown that clinical complications, such as hepatic decompensation or HCC, may occur even after HBsAg seroclearance, particularly in patients superinfected with other viruses or in those with liver cirrhosis [5,7,11–14].

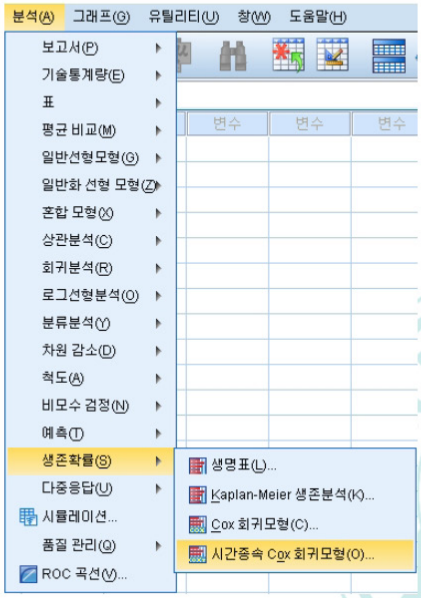
Surveillance for HCC is cost-effective when the annual risk of HCC exceeds 0.2% in non-cirrhotic hepatitis B patients and 1.5% in cirrhotic patients [15,16]. However, little is known about whether surveillance for HCC is worthwhile in chronic HBV-infected patients who have achieved HBsAg seroclearance. Moreover, the reported rates of HCC after HBsAg seroclearance

예제-1

- hcc: hcc 발생여부 no 0, yes 1
- hcc_yr: hcc 발생까지 기간
- **HBsAg 혈전전환 여부** (time-dependent 변수)
  - sab_yr1: HBsAg 혈전전환까지 기간, 혈전전환이 생기지 않은 경우는 999로 코딩

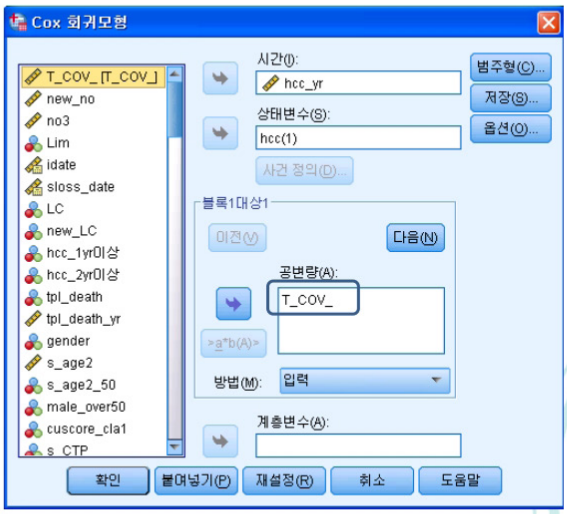
	new_no	no3	Lim	idate	sloss_date	LC	new_LC	hcc	hcc_yr	sab_yr1
1	10	7	0	2002/07/03	2005/04/13	1	1	1	6.40	6.19
2	729	59	0	2000/07/22	2012/01/12	1	1	0	1.68	.62
3	693	91	0	2003/01/03	2011/10/04	0	0	0	1.71	1.71
4	683	99	0	1999/03/16	2011/08/12	0	0	0	1.89	1.89
5	715	72	0	2000/04/26	2011/12/21	0	0	0	2.09	1.59
6	497	270	0	1999/05/03	2009/06/12	1	1	0	4.39	2.95
7	379	378	0	2005/11/11	2007/05/18	0	0	0	4.94	.56
8	420	342	1	1997/10/28	2008/02/21	0	0	0	5.90	2.70
9	347	407	0	2001/05/14	2006/11/08	0	0	0	7.11	1.60
10	280	471	0	1997/03/04	2005/12/09	0	0	0	7.92	2.90
11	1	18	0	2003/07/22	2005/08/24	0	1	1	8.40	999.00
12	811	731	1	2000/08/07	2013/01/10	0	0	0	.50	999.00
13	699	776	1	1999/07/22	2011/10/31	0	0	0	.56	999.00
14	814	728	0	1999/12/27	2013/01/15	0	0	0	.64	999.00

SPSS 실습:
분석->생존확률->시간종속 Cox 회귀모형



HBsAg 혈전전환까지 걸린 시간이 HCC발생까지 시간보다 작으면 1,
혈전전환까지 시간이 결측이거나 크면 0

hcc발생까지 시간



생성된 Time dependent covariate를 공변량에 넘겨줌

SPSS output

- Time dependent covariate로 고려한 경우

방정식의 변수

	B	표준오차	Wald	자유도	유의확률	Exp(B)	Exp(B)에 대한 95.0% CI	
							하한	상한
T_COV_	-.802	.539	2.213	1	.137	.449	.156	1.290

- 관찰기간 내 혈전전환이 있는지(Time independent covariate)로 구분한 경우

방정식의 변수

	B	표준오차	Wald	자유도	유의확률	Exp(B)	Exp(B)에 대한 95.0% CI	
							하한	상한
sab_01	-1.619	.529	9.363	1	.002	.198	.070	.559

When PH assumption not satisfied

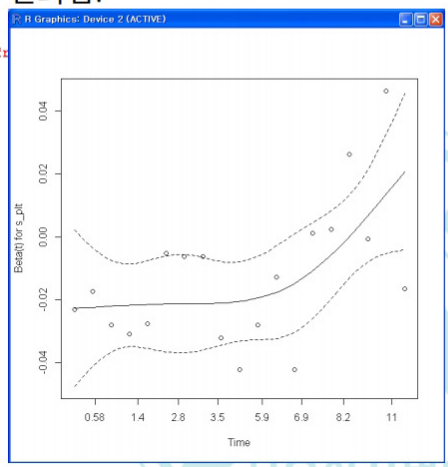
- Use time-dependent coefficients
  - Defined to analyze a time-independent predictor not satisfying the PH assumption
  - $h(t) = h_0(t)\exp(\beta X + \delta(X \times g(t)))$
    - ✓ Check PH assumption for X
    - ✓ $\hat{HR}(t) = \exp(\hat{\beta} + \hat{\delta}t)$

      $\hat{\delta} > 0 \Rightarrow \hat{HR}(t) \uparrow \text{ as } t \uparrow$

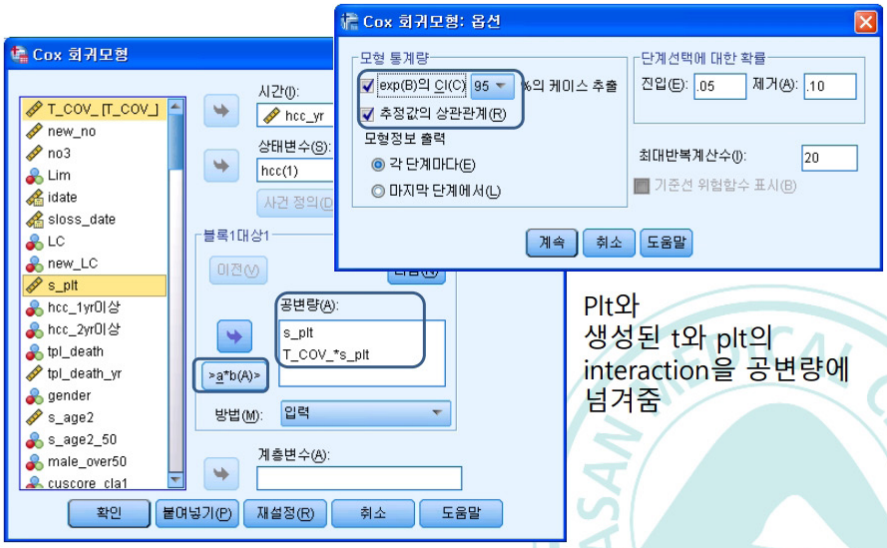
예제-2

- PH assumption에 위배되는 경우
- Platelets(baseline 때): 시간에 따라 변화하지는 않지만, 시간에 따라 platelet이 미치는 영향력(beta)이 변화함.

```
> fit<-coxph(Surv(hcc_yr,hcc)~s_plt,data=c,method="efr
>
>
> schfit<-cox.zph(fit,transform='rank')
> schfit
        rho chisq       p
s_plt 0.512  7.56 0.00598
> plot(schfit)
>
```



SPSS분석:
분석->생존확률->시간종속 Cox 회귀모형



Plt와
생성된 t와 plt의
interaction을 공변량에
넘겨줌

SPSS output

방정식의 변수

	B	표준오차	Wald	자유도	유의확률	Exp(B)	Exp(B)에 대한 95.0% CI	
							하한	상한
s_plt	-.029	.008	14.737	1	.000	.972	.957	.986
T_COV_*s_plt	.003	.001	6.969	1	.008	1.003	1.001	1.005

$$h(t) = h_0(t)\exp(\beta \cdot plt + \delta(plt \times t))$$

PH assumption에 위배

- $\hat{\beta} = -0.029, \hat{\delta} = 0.003$
- $\hat{HR}$ depends on $\hat{\beta}$ and $\hat{\delta}$
  - Time=1, HR=exp(β+δ*t)=exp(-0.029+0.003*1)=0.974
  - Time=2, HR=exp(-0.029+0.003*2)=0.977
  - Time=5, HR=exp(-0.029+0.003*5)=0.985

방정식의 변수

	B	표준오차	Wald	자유도	유의확률	Exp(B)	Exp(B)에 대한 95.0% CI	
							하한	상한
s_plt	-.029	.008	14.737	1	.000	.972	.957	.986
T_COV_*s_plt	.003	.001	6.969	1	.008	1.003	1.001	1.005

회귀계수의 상관행렬

	s_plt
T_COV_*s_plt	-.844

$$\mathrm{cov}(\hat{\beta},\hat{\delta}) = \rho_{\hat{\beta},\hat{\delta}} \cdot s_{\hat{\beta}} \cdot s_{\hat{\delta}} = -0.844 \times 0.008 \times 0.001$$

- HR의 95% CI

$$\exp\left[\left(\hat{\beta}+\hat{\delta}t\right) \pm 1.96 \cdot \sqrt{\mathrm{Var}\left(\hat{\beta}+\hat{\delta}t\right)}\right],$$

$$\mathrm{Var}\left(\hat{\beta}+\hat{\delta}t\right) = s_{\hat{\beta}}^2 + t^2 s_{\hat{\delta}}^2 + 2t\,\mathrm{cov}(\hat{\beta},\hat{\delta})$$

$$= (0.008)^2 + t^2(0.001)^2 + 2t(-0.0000068)$$

Cox PH 회귀모형: 다변량 분석

- $h(t;X) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)$
- 여러 독립변수들을 하나의 모형에 포함
  - 다중공선성을 고려하여 모형에 포함될 후보변수 선택
  - 충분한 event가 필요
    - ✓ Rule of thumb(Peduzzi et al.(1995)):
      at least **10** events per **1** covariate
  - 비례위험 가정(PH assumption) 확인

Fundamentals of survival analysis 77/88

변수선택

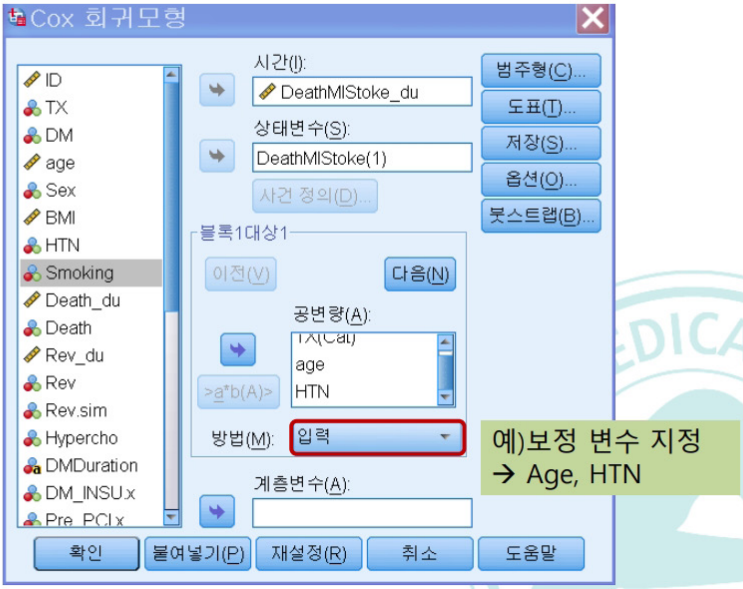
- 연구목적에 따라서 변수 선택 strategy 결정
  - Risk factor 분석 : 주로 통계적으로 의미 없는 변수 제외
    - ✓ Care must be exercised (False positive 문제)
  - Causal 분석 : single factor is under investigation
    - ✓ RCT(Causal 분석)연구에서 종종 보정변수(adjustment factor)를 미리 protocol에 지정 (False positive 문제)
  - Prognostic 모델 : calibration, discrimination
- Common choices : semi-automated
  - Stepwise, Backward and Forward 등
  - 통계적인 유의성에만 근거한 모델은 임상적으로 의미가 없을 수 있음 (Henderson and Velleman, 1981)

Fundamentals of survival analysis 78/88

Cox PH 회귀모형 예제

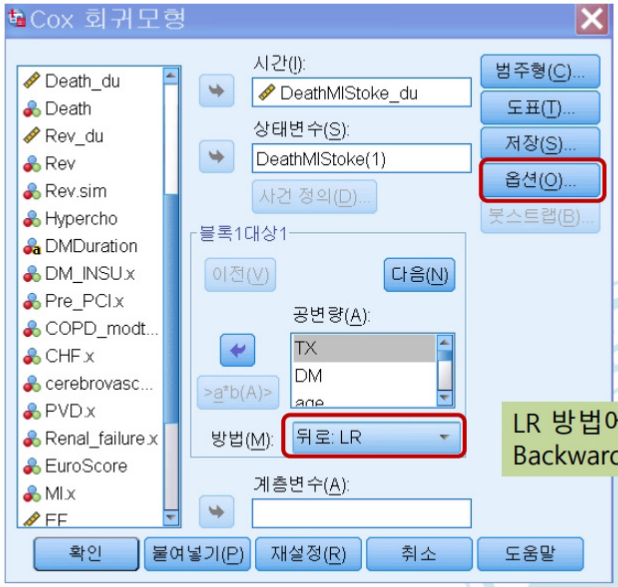
- 단변량 분석 (Univariate analysis)
  - 시술법에 따른 hazard rate 비교
  - SYNTAX grade에 따른 hazard rate 비교
- 다변량 분석 (Multivariable analysis)
  - 다른 유의미한 covariate으로 보정했을 때 시술법에 따른 hazard rate 비교

Cox PH 회귀모형(Multi.) 예제 (SPSS)

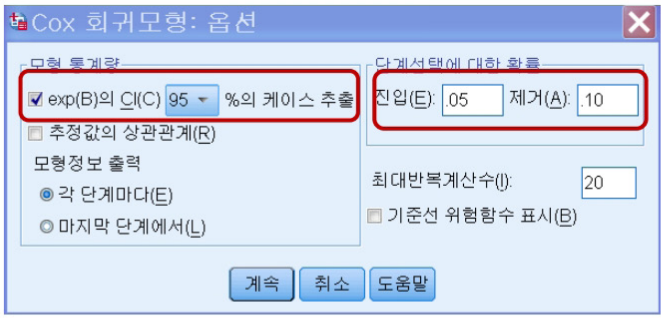


예)보정 변수 지정
→ Age, HTN

Cox PH 회귀모형(Multi.) 예제 (SPSS)



LR 방법에 근거한
Backward elimination법



[Stepwise 변수 선택 시]
진입: 0.05 → AE 에 관련하여 가장 유의한 변수부터 차례로 선택하되
유의수준 0.05에서 선택
제거: 0.1 → 새로 선택된 변수를 보정한 후 기존에 선택된 변수들이 유의수준
0.1에서 유의하지 않을 때 제거

[Backward 변수 선택 시]
모든 변수를 모형에 놓고 시작 → 차례로 유의수준 0.1 기준으로 단계적 제거

[Forward 변수 선택 시]
가장 유의한 변수부터 모형에 첨가 → under-fit 가능성

단계별 변수 제거 후 마지막 까지 남은 변수

방정식의 변수

		B	표준오차	Wald	자유도	유의확률	Exp(B)	Exp(B)에 대한 95.0% CI 하한	상한
단계 1	TX	-.625	.561	1.238	1	.266	.536	.178	1.609
	age	.081	.031	6.988	1	.008	1.085	1.021	1.152
	HTN	1.042	.416	6.268	1	.012	2.835	1.254	6.408
	DM	.147	.360	.167	1	.683	1.158	.572	2.346
	Sex	-.129	.428	.091	1	.762	.879	.380	2.033
	BMI	.040	.054	.545	1	.460	1.041	.936	1.158
	Smoking	.526	.428	1.516	1	.218	1.693	.732	3.914
	Hypercho	.373	.374	.995	1	.319	1.453	.697	3.026
	Pre_PCI.x	.692	.391	3.140	1	.076	1.998	.929	4.297

17 단계에서 TX 제거됨

단계 18	age	.065	.019	12.236	1	.000	1.068	1.029	1.107
	HTN	1.022	.378	7.314	1	.007	2.780	1.325	5.833
	Pre_PCI.x	.604	.348	3.017	1	.082	1.829	.925	3.615
	Renal_failure.x	1.336	.401	11.089	1	.001	3.805	1.733	8.354
	EF	-.042	.014	9.525	1	.002	.959	.934	.985

TX 첨가하여 모형 re-fit

방정식의 변수

	B	표준오차	Wald	자유도	유의확률	Exp(B)	Exp(B)에 대한 95.0% CI 하한	상한
TX	-.509	.310	2.696	1	.101	.601	.327	1.104
age	.065	.019	11.672	1	.001	1.067	1.028	1.108
HTN	.981	.378	6.729	1	.009	2.667	1.271	5.598
Renal_failure.x	1.270	.391	10.536	1	.001	3.561	1.654	7.668
EF	-.040	.013	9.319	1	.002	.961	.936	.986

→Age, HTN, Renal failure, EF 보정 후 시술법(TX)은 AE rate에 대해 유의하지 않음 (P=0.101)

통계분석 방법 기술

- Statistical analysis

Treatment-related differences in long-term outcomes between the 2 procedures were analyzed separately in patients with and without medically treated DM. Prevalence rates of risk factors and characteristics of the patients in the 2 treatment groups were compared using *t* test or Wilcoxon rank-sum test for continuous variables and with chi-square statistics or Fisher's exact test for categorical variables. Survival curves were constructed using the Kaplan–Meier method and compared using log-rank test.

Differences in risk-adjusted long-term rates of study outcomes between patients in the DES and CABG groups were assessed using multivariable Cox proportional hazards regression. Adjusted covariates included patient age and gender, presence or absence of different clinical and coexisting conditions, left ventricular function, and number and extent of diseased vessels. The proportional hazards assumption was confirmed by examination of log(–log [survival]) curves and by testing of partial (Schoenfeld) residuals, and

- 두 군에서의 환자 특징 비교 : t-test/Wilcoxon rank sum test or chi-square test
- 생존곡선 비교 : K-M method
- Adjusted 위험률의 차이 비교 : Cox 모형
- PH 가정 평가 : LML plot, Schoenfeld residuals.

- American Journal of Cardiology, 2012;109:1548-1557

분석결과 제시방법1

Table 3
Hazard ratios for clinical adverse outcomes after drug-eluting stents compared to coronary artery bypass grafting according to diabetic status*

Outcomes	Total Number of Events/ Number of Patients		Unadjusted		Multivariable Adjusted†		
	DES	CABG	HR (95% CI)	*p* Value	HR (95% CI)	*p* Value	Interaction *p* Value for Diabetic Status
Death							
Diabetic patients	57/489	60/402	0.82 (0.57–1.17)	0.27	1.37 (0.86–2.17)	0.19	0.32
Nondiabetic patients	72/1,058	115/1,093	0.68 (0.51–0.91)	0.01	0.85 (0.63–1.15)	0.30	
Composite outcome (death, myocardial infarction, or stroke)							0.12
Diabetic patients	72/489	76/402	0.80 (0.58–1.10)	0.16	1.38 (0.92–2.08)	0.12	
Nondiabetic patients	99/1,058	158/1,093	0.67 (0.52–0.86)	0.002	0.79 (0.61–1.02)	0.07	
Repeat revascularization							0.46
Diabetic patients	91/489	22/402	3.88 (2.43–6.20)	<0.001	3.61 (2.25–5.77)	<0.001	
Nondiabetic patients	168/1,058	65/1,093	3.12 (2.33–4.16)	<0.001	3.12 (2.34–4.17)	<0.001	

* Hazard ratios are for the drug-eluting stent compared to the coronary artery bypass grafting group.
† Hazard ratios were adjusted for age; gender; diabetes; duration of diabetes; presence or absence of congestive heart failure; chronic obstructive pul peripheral arterial disease, and renal failure; European System for Cardiac Operative Risk Evaluation; history or no history of myocardial infarction before presence or absence of involvement of the proximal left anterior descending or left main coronary artery; total obstruction; and SYNTAX score.
HR = hazard ratio; IPTW = inverse probability-of-treatment weighting. Other abbreviation as in Table 2.

- American Journal of Cardiology, 2012;109:1548-1557

분석결과 제시방법2

Table 3: Univariable and Multivariable Cox Proportional Hazard Analyses of Postoperative Recurrence-Free Survival in Development Set

	Univariable Cox Proportional Hazard Analysis			Multivariable Cox Proportional Hazard Analysis		
Parameter	Regression Coefficient	Hazard Ratio	*P* Value	Regression Coefficient	Hazard Ratio	*P* Value
Age	0.01	1.01 (0.99, 1.03)	.30	…	…	…
Male sex	0.25	1.29 (0.94, 1.76)	.12	…	…	…
Body mass index (kg/m²)	−0.04	0.96 (0.91, 1.01)	.11	…	…	…
Tumor size (cm)	0.37	1.44 (1.26, 1.66)	<.001	0.21	1.23 (1.05, 1.44)	.009
Dominant location			.72			…
Head	1	1 [reference]	…	…	…	…
Body	−0.13	0.88 (0.57, 1.35)	.55	…	…	…
Tail	0.09	1.09 (0.72, 1.67)	.67	…	…	…
Tumor density in AP			.02			…
Isodense or hyperdense	1	1 [reference]	…	…	…	…
Hypodense	0.73	2.07 (1.12, 3.82)	…	…	…	…
Tumor density in PVP			<.001			.04
Isodense or hyperdense	1	1 [reference]	…	…	…	…
Hypodense	0.92	2.51 (1.55, 4.04)	…	0.51	1.66 (1.01, 2.73)	…
Tumor conspicuity in AP			.008			…
Poor	1	1 [reference]	…	…	…	…
Moderate	0.70	2.01 (1.17, 3.46)	.01	…	…	…
Well	0.92	2.50 (1.41, 4.44)	.002	…	…	…
Tumor conspicuity in PVP			.003			…
Poor	1	1 [reference]	…	…	…	…
Moderate	0.65	1.91 (1.12, 3.28)	.02	…	…	…
Well	0.93	2.52 (1.47, 4.35)	.001	…	…	…
Tumor necrosis	1.07	2.91 (2.00, 4.25)	<.001	0.714	2.04 (1.38, 3.03)	<.001
Peripancreatic tumor infiltration	0.69	1.99 (1.44, 2.75)	<.001	0.406	1.50 (1.07, 2.11)	.02
Contact to SMV or PV	0.10	1.11 (0.79, 1.55)	.55	…	…	…
Adjacent organ invasion	0.37	1.45 (1.06, 1.98)	.02	…	…	…
Suspicious metastatic lymph nodes	0.76	2.15 (1.53, 3.01)	<.001	0.662	1.94 (1.38, 2.72)	<.001
Cancer antigen 19–9	0.00	1 (1–1)	.06	…	…	…
Bilirubin	0.04	1.04 (0.97–1.11)	.30	…	…	…
Albumin	−0.30	0.74 (0.54–1.02)	.07	…	…	…
Lymphocytes	0.00	1 (1–1)	.34	…	…	…

Note.—Data in parentheses are 95% confidence intervals. AP = arterial phase, PV = portal vein, PVP = portal venous phase, SMV = superior mesenteric vein.

Radiology. 2020 Sep;296(3):541-551

감사합니다.

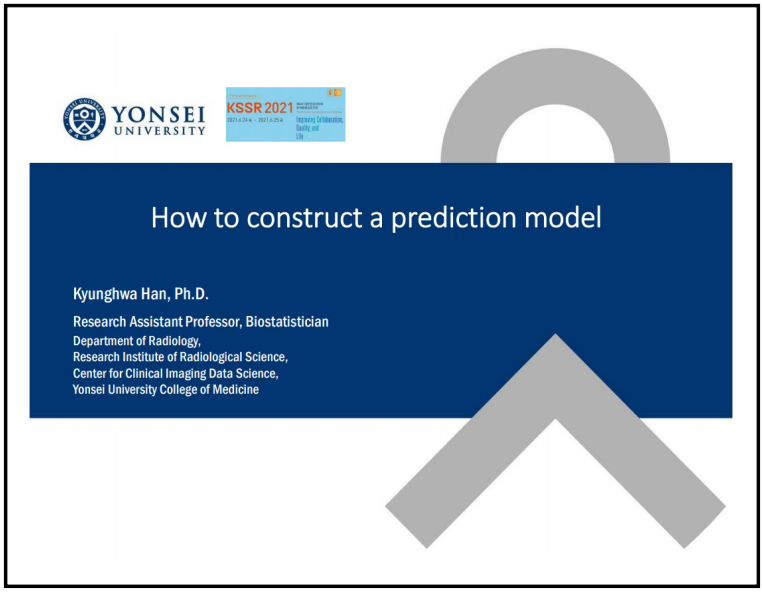
seonok@amc.seoul.kr

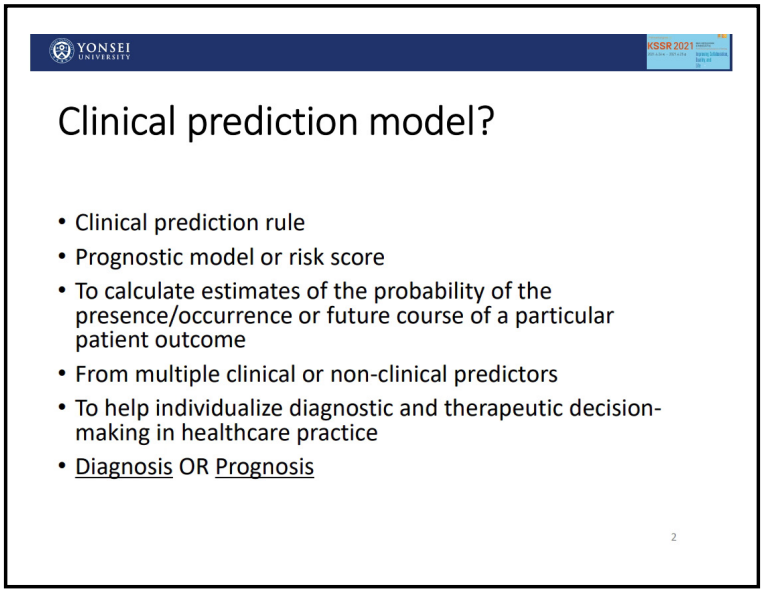
Clinical Research Methodology Course – Intermediate Course

14:30–15:20 Room 2

How to construct a prediction model

한 경 화
연세대학교





YONSEI UNIVERSITY

한국인을 위한 뇌졸중 발생 예측모형 개발

고려대학교 의과대학 의학통계학교실, 을지대학교 을지병원 신경과[a], 서울의료원 신경과[b], 순천향대학교병원 신경과[c], 을지대학교 을지대학병원 신경과[d], 인제대학교 일산백병원 신경과[e], 서울대학교 의과대학 분당서울대학교병원 신경과[f]

이지성 박종무[a] 박태환[b] 이경복[c] 이수주[d] 조용진[e] 한문구[f] 배희준[f] 이준영

Background: Assessing an individual's risk of stroke can be a starting point for stroke prevention. The aim of this study was to develop a stroke prediction model that can be applied to the Korean population, using the best available current knowledge.

Methods: A sex- and age-specific stroke prediction model that is applicable specifically to Koreans was developed using Gail's breast cancer prediction model, which is based on competing risk theory.

Results: The relative risks for major stroke risk factors, including hypertension, diabetes, hypercholesterolemia, atrial fibrillation, ischemic heart disease, previous stroke, obesity, and smoking status, were obtained from a recent systematic review of stroke risk factors among Koreans. The results were incorporated into the concept of a proportional hazard regression model. For baseline age- and sex-specific hazard rates for stroke, we employed Jee's 10-year stroke-risk prediction model with its reference categories for predictor variables. Death-certificate data from the Korea National Statistical Office were used to calculate competing risks of stroke in our model.

Conclusions: Our prediction model for stroke incidence may be useful for predicting an individual's risk of stroke based on his/her age, sex, and risk factors. This model will contribute to the development of individualized risk-specific guidelines for the prevention of stroke.

J Korean Neurol Assoc 28(1):13-21, 2010

3

Effect of Microvascular Invasion Risk on Early Recurrence of Hepatocellular Carcinoma After Surgery and Radiofrequency Ablation

Sunyoung Lee, MD,††† Tae Wook Kang, MD, Kyoung Doo Song, MD,* Min Woo Lee, MD,* Hyunchul Rhim, MD,* Hyo Keun Lim, MD,*‡ So Yeon Kim, MD,† Dong Hyun Sinn, MD,§ Jong Man Kim, MD,¶ Kyunga Kim, PhD,‡|| and Sang Yun Ha, MD***

TABLE 3. Multivariable Analysis of Predictors of Microvascular Invasion and Creation of the Microvascular Invasion Risk Score

Variable	Multivariable Analysis OR (95% CI)	P	β Coefficient	MVI Risk Points
α-FP ≥15, ng/mL [α-FP<15]	3.46 (1.62–7.39)	0.001	1.242	1.0
PIVKA-II ≥48, mAU/mL [PIVKA-II<48]	3.41 (1.54–7.55)	0.003	1.225	1.0
Arterial peritumoral enhancement [absence]	5.07 (2.36–10.87)	<0.001	1.622	1.5
Peritumoral hypointensity on HBP [absence]	15.98 (6.73–37.97)	<0.001	2.771	2.5

The reference category for each categorical variable is in the square brackets in first column. Multivariable logistic regression model was performed using stepwise backward variable selection. The scaled coefficients were simplified by rounding them to nearest half. The MVI risk score is obtained by adding the total number of points scored in each of the 4 variables.

MELD indicates Model for End-Stage Liver Disease.

Point value for each variable

Serum α-fetoprotein level (ng/mL)	Points
<15	0
≥15	1

Serum PIVKA-II level (mAU/mL)	Points
<48	0
≥48	1

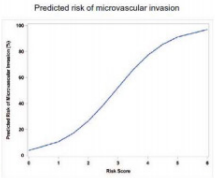
Peritumoral arterial enhancement	Points
(-)	0
(+)	1.5

Peritumoral hypointensity on HBP	Points
(-)	0
(+)	2.5

A

Summation of Points:	
Serum α-fetoprotein level	+
Serum PIVKA-II level	+
Peritumoral arterial enhancement	+
Peritumoral hypointensity on HBP	+
Total Risk Score =	

Predicted risk of microvascular invasion



B

Total Risk Score	Predicted Risk of MVI of HCC	Percentage of Patients in Risk Category, %	
		Derivation Cohort (n = 276)	External Validation Cohort (n = 101)
0	5.9	31.2	22.8
1	10.7	25.7	22.8
1.5	13.2	4.7	10.9
2	18.6	7.6	7.9
2.5	23.9	9.4	9.9
3.5	37.4	5.8	6.9
4	45.2	2.5	5.9
4.5	53.2	0.4	1.0
5	61.0	7.6	7.9
6	74.8	5.1	4.0

FIGURE 2. Four-variable risk index for microvascular invasion in patients with a small (≤3 cm) hepatocellular carcinoma. This model was able to stratify MVI risk ranging from less than 5.9% in those with a risk score of 0 to higher than 74.8% in those with a risk score of 6 in the external validation cohort.

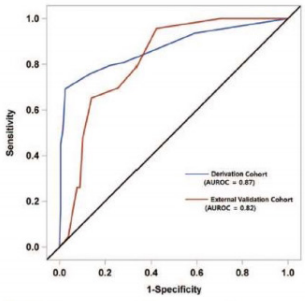


FIGURE 3. Receiver operating characteristic curves for the prediction model for microvascular invasion in the derivation and external validation cohorts. The area under the receiver operating characteristic curve for the MVI prediction model was 0.87 (95% confidence interval: 0.82–0.92) and 0.82 (95% confidence interval: 0.74–0.90) in the derivation and external validation cohorts, respectively.

Ann Surg 2021;273:564–571

YONSEI UNIVERSITY

KSSR 2021

When do we need to develop a prediction model? in radiologic research

- To show **improvement** in the predictive ability

  by adding imaging features
  - Conventional imaging findings
  - Radiomics
  - Artificial Intelligence
- Need to compare between…
  - Clinical only vs. Clinical + Imaging
  - Imaging only vs. Clinical + Imaging
  - OOO + Imaging #1 vs. OOO + Imaging #2

Table 2: Risk Factors for Lymph Node Metastasis in Biliary Tract Cancer

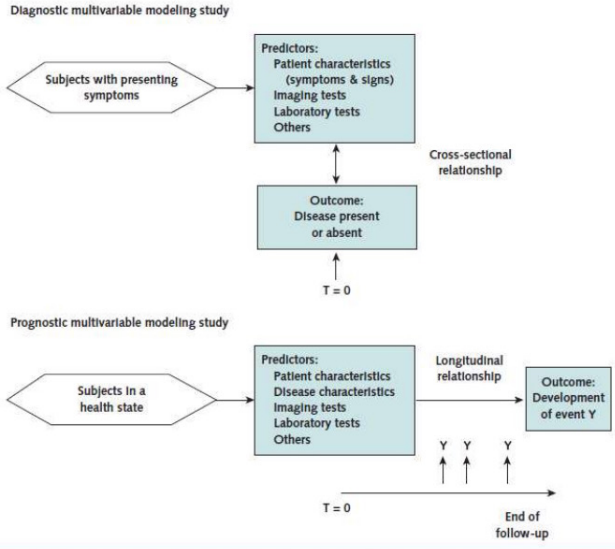
	Radiomics Model		Clinical Model	
Variable	Odds Ratio	*P* Value	Odds Ratio	*P* Value
CA 19–9 level	2.10 (0.85, 5.22)	.11	1.82 (0.78, 4.25)	.17
CT-reported tumor size	0.46 (0.17, 1.28)	.14	2.83 (1.44, 5.55)	.003
CT-reported vascular invasion	1.54 (0.68, 3.47)	.30	1.84 (0.87, 3.88)	.11
CT-reported LN status	2.81 (1.21, 6.55)	.02	3.03 (1.39, 6.57)	.005
Radiomics signature	6.24 (2.91, 13.40)	<.001	NA	NA

Note.—Data are results of the multivariable regression analysis. Data in parentheses are 95% confidence intervals. The clinical model was built on the basis of independent predictors of nodal metastasis without the addition of radiomics signature. CA 19–9 = carbohydrate antigen 19–9, LN = lymph node, NA = not available.

GW Ji, et al., Biliary Tract Cancer at CT: A Radiomics-based Model to Predict Lymph Node Metastasis and Survival Outcomes. Radiol 2019; 290: 90-98

5

Box A. Schematic representation of diagnostic and prognostic prediction modeling studies.



The nature of the prediction in diagnosis is estimating the probability that a specific outcome or disease is present (or absent) within an individual, at this point in time–that is, the moment of prediction (T = 0). In prognosis, the prediction is about whether an individual will experience a specific event or outcome within a certain time period. In other words, in diagnostic prediction the interest is in principle a cross-sectional relationship, whereas prognostic prediction involves a longitudinal relationship. Nevertheless, in diagnostic modeling studies, for logistical reasons, a time window between predictor (index test) measurement and the reference standard is often necessary. Ideally, this interval should be as short as possible without starting any treatment within this period.

6

Ann Intern Med. 2015;162:55-63. doi:10.7326/M14-0697

YONSEI UNIVERSITY

Prediction Model

Define the outcome
- Type of outcome : continuous, binary, time to event

Developing New Model
OR
Improving Existing Model

Development of the Prediction Model
- Variables (Predictors) Selection
- Determining of Statistical Modeling
- Evaluation of Model Performance (Comparison with Existing Model)

Validation of the Prediction Model
- Internal Validation: Performance and Optimism
- External Validation: Generalizability, Usability in Clinical Practice

Updating
- Evaluation of Model Performance

7

YONSEI UNIVERSITY

Source of data and Sample size

- Prospective longitudinal cohort study vs. RCT?
- Individual participant data from multiple studies or large existing data sets
- Clustered Data ⇒ a weighted approach
- An "adequate" sample size is unclear
- A rule of thumb for sample size

  - at least 10 events are required per candidate predictor
- Readily available large cohort or registry

8

YONSEI UNIVERSITY

KSSR 2021

BMJ 2020;368:m441 doi: 10.1136/bmj.m441 (Published 18 March 2020) Page 1 of 12

RESEARCH METHODS & REPORTING

Check for updates

Calculating the sample size required for developing a clinical prediction model

Clinical prediction models aim to predict outcomes in individuals, to inform diagnosis or prognosis in healthcare. Hundreds of prediction models are published in the medical literature each year, yet many are developed using a dataset that is too small for the total number of participants or outcome events. This leads to inaccurate predictions and consequently incorrect healthcare decisions for some individuals. In this article, the authors provide guidance on how to calculate the sample size required to develop a clinical prediction model.

Richard D Riley *professor of biostatistics*[1], Joie Ensor *lecturer in biostatistics*[1], Kym I E Snell *lecturer in biostatistics*[1], Frank E Harrell Jr *professor of biostatistics*[2], Glen P Martin *lecturer in health data sciences*[3], Johannes B Reitsma *associate professor*[4], Karel G M Moons *professor of clinical epidemiology*[4], Gary Collins *professor of medical statistics*[5], Maarten van Smeden *assistant professor*[4 5 6]

9

YONSEI UNIVERSITY

KSSR 2021

Type of Model to estimate p or y

- Continuous Outcome
  - Linear regression: $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$
  - for predicting outcome values
- Binary Outcome
  - Logistic regression: $\text{logit}(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$
  - for predicting short-term events

$\exp(\beta) = \text{Odds ratio}$

$$\Rightarrow \quad \hat{p} = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)}$$

- Time to event (Survival) Outcome
  - Cox proportional hazard regression:
    $\log h(t, x) = \log h_0(t) + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$
  - for predicting long-term prognostic outcomes

10

YONSEI UNIVERSITY

Considerations in Selecting Predictors

- Before / During modeling
- Clinical reasoning + statistical significance
- Categorization for continuous variables
- Automated predictor selection strategies
- Multicollinearity
- Missing values

11

YONSEI UNIVERSITY

Variable selection

- Group comparison (univariable analysis)
  - Inequality test for mean or proportion
  - Multiple testing problem
- Subset selection
  - p개의 independent variables 중 특정 k개만을 최종 모형에 포함시키도록 하면서 prediction accuracy가 가장 높아지는 subset of variables를 선정하는 방법
- Automatic selection

12

YONSEI UNIVERSITY

KSSR 2021

Assumptions for Model Development

- Linear regression
  - linear relationship, normality, homoscedasticity
- Logistic regression
  - linear relationship
- Cox proportional hazard regression
  - linear relationship, proportionality of hazard
- When the assumptions were not met,
  - transformation, nonlinear modeling, stratified analysis

13

YONSEI UNIVERSITY

KSSR 2021

Annals of Internal Medicine RESEARCH AND REPORTING METHODS

Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement

Gary S. Collins, PhD; Johannes B. Reitsma, MD, PhD; Douglas G. Altman, DSc; and Karel G.M. Moons, PhD

Annals of Internal Medicine RESEARCH AND REPORTING METHODS

Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration

Karel G.M. Moons, PhD; Douglas G. Altman, DSc; Johannes B. Reitsma, MD, PhD; John P.A. Ioannidis, MD, DSc; Petra Macaskill, PhD; Ewout W. Steyerberg, PhD; Andrew J. Vickers, PhD; David F. Ransohoff, MD; and Gary S. Collins, PhD

Ann Intern Med. 2015;162:55-63.
Ann Intern Med. 2015;162:W1-W73.

14

TRIPOD Checklist: Prediction Model Development and Validation

Section/Topic	Item		Checklist Item	Page
Title and abstract				
Title	1	D;V	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	
Introduction				
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model or both.	
Methods				
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	
Participants	5a	D;V	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	
	5b	D;V	Describe eligibility criteria for participants.	
	5c	D;V	Give details of treatments received, if relevant.	
Outcome	6a	D;V	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	
	6b	D;V	Report any actions to blind assessment of the outcome to be predicted.	
Predictors	7a	D;V	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	
	7b	D;V	Report any actions to blind assessment of predictors for the outcome and other predictors.	
Sample size	8	D;V	Explain how the study size was arrived at.	
Missing data	9	D;V	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses.	
	10b	D	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	
	10c	V	For validation, describe how the predictions were calculated.	
	10d	D;V	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	
	10e	V	Describe any model updating (e.g., recalibration) arising from the validation, if done.	
Risk groups	11	D;V	Provide details on how risk groups were created, if done.	
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	

15

Results				
Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	
	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	
	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	
Model development	14a	D	Specify the number of participants and outcome events in each analysis.	
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome.	
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	
	15b	D	Explain how to the use the prediction model.	
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model.	
Model-updating	17	V	If done, report the results from any model updating (i.e., model specification, model performance).	
Discussion				
Limitations	18	D;V	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	
Interpretation	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data.	
	19b	D;V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	
Implications	20	D;V	Discuss the potential clinical use of the model and implications for future research.	
Other information				
Supplementary information	21	D;V	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	
Funding	22	D;V	Give the source of funding and the role of the funders for the present study.	

16

YONSEI UNIVERSITY

KSSR 2021

Assessing the model performance

TABLE 1. Characteristics of Some Traditional and Novel Performance Measures

Aspect	Measure	Visualization	Characteristics
Overall performance	R^2, Brier	Validation graph	Better with lower distance between Y and $\hat{Y}$. Captures calibration and discrimination aspects
Discrimination	c statistic	ROC curve	Rank order statistic; interpretation for a pair of subjects with and without the outcome
	Discrimination slope	Box plot	Difference in mean of predictions between outcomes; easy visualization
Calibration	Calibration-in-the-large	Calibration or validation graph	Compare mean (y) versus mean ($\hat{y}$); essential aspect for external validation
	Calibration slope		Regression slope of linear predictor; essential aspect for internal and external validation; related to "shrinkage" of regression coefficients
	Hosmer-Lemeshow test		Compares observed to predicted by decile of predicted probability
Reclassification	Reclassification table	Cross-table or scatter plot	Compare classifications from 2 models (one with, one without a marker) for changes
	Reclassification statistic		Compare observed outcomes to predicted risks within cross-classified categories
	Net reclassification index (NRI)		Compare classifications from 2 models for changes by outcome for a net calculation of changes in the right direction
	Integrated discrimination index (IDI)	Box plots for 2 models (one with, one without a marker)	Integrates the NRI over all possible cut-offs; equivalent to difference in discrimination slopes
Clinical usefulness	Net benefit (NB) Decision curve analysis (DCA)	Cross-table Decision curve	Net number of true positives gained by using a model compared to no model at a single threshold (NB) or over a range of thresholds (DCA)

EW Steyerberg et al. et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010;21:128-38.

Statistical Analysis

Subsequent analysis was performed using R v3.4.0. Patients were randomly allocated to a discovery and validation set (2:1 ratio with $n = 120$ patients in the discovery set and $n = 61$ patients in the validation set) with the distribution of MGMT promoter methylation kept balanced between both sets (stratified random split). Distribution of epidemiological, clinical, and molecular characteristics between the discovery and validation sets was compared with the chi-square test for categorical parameters and the Wilcoxon test for continuous parameters.

A total of 386 out of the 1043 extracted radiomic features (37.0%) were identified as stable and reproducible based on a separate prospective test-retest study and selected for further analysis (methodology and results of this preceding analysis are outlined in Supplementary Table S6).

A Cox regression model via penalized maximum likelihood (lasso) was fitted on the discovery set to identify a subset of radiomic features and construct a radiomic signature from the high-dimensional radiomic dataset associated with outcome (as measured by OS; using the glmnet package[34,35]). The tuning parameter λ, which is the global regularization parameter, was identified via 10-fold cross-validation. The performance of the identified radiomic signature for stratifying PFS and OS in the discovery and validation sets was assessed by comparing models that included (i) molecular features alone (MGMT promoter methylation status and global DNA methylation subgroups), (ii) clinical features alone (including patient's age, KPS at diagnosis, extent of resection [EOR; gross total resection (GTR) vs subtotal resection (STR) or biopsy] and adjuvant treatment [radiotherapy plus concomitant and adjuvant TMZ (RT+TMZ) vs RT or TMZ only]), (iii) standard imaging features alone (tumor volumes from contrast enhancement, necrosis, and edema), (iv) radiomic signature alone, and (v) different combinations of the above stated models to assess the incremental value of combining parameters from different layers (ie, molecular, clinical, standard imaging, radiomics).

For each model, we assessed the overall performance with prediction error curves (PECs) over time and the integrated Brier score (IBS) (using the pec function of the pec library[36,37]). The IBS can range from 0 for a perfect model to 0.25 for a non-informative model with a 50% incidence of the outcome. Specifically, the discovery set was supplied to the traindata argument of the pec function, whereas the validation set was used for estimating the PECs and IBS (data argument of the pec function). Furthermore, ANOVA was used to determine whether additional predictors significantly increase the model fit (ie, reduction in the log-likelihood). Multivariate Cox regression models were used

Kickingereder P, et al. Radiomic subtyping improves disease stratification beyond key molecular, clinical, and standard imaging characteristics in patients with glioblastoma. *Neuro-oncology* 20.6 (2017): 848-857.

Table 1. (a) Analysis of deviance for different Cox regression models (ANOVA) was used to determine whether the radiomic signature or the tumor volume increased the model fit beyond key molecular and clinical characteristics. (b) Performance metrics of the different Cox regression models based on prediction error curves over time with the integrated Brier score (lower values indicate better performance)

(a) Analysis of Deviance for Different Cox Regression Models (ANOVA)

Model 1	Model 2	Discovery Set				Validation Set			
		OS		PFS		OS		PFS	
		P	chi²	*P*	chi²	*P*	chi²	*P*	chi²
Molecular[1] + Clinical[2]	Molecular[1] + Clinical[2] + Radiomic signature	<0.01	34.3	0.01	6.2	<0.01	10.4	<0.01	8.0
Molecular[1] + Clinical[2]	Molecular[1] + Clinical[2] + Tumor volumes[3]	0.79	1.0	0.19	4.7	0.21	4.6	0.14	5.4

(b) Performance Metrics of the Different Cox Regression Models

Model		Integrated Brier Score (IBS) (percent reduction of IBS compared with the null model[4])			
		OS		PFS	
Single layer	Molecular[1]	0.149	−9%	0.121	−12%
	Clinical[2]	0.133	−18%	0.126	−9%
	Tumor volumes[3]	0.160	−2%	0.135	−2%
	Radiomic signature	0.137	−16%	0.125	−9%
Two layers	Molecular[1] + Clinical[2]	0.119	−27%	0.098	−29%
	Clinical[2] + Radiomic signature	0.116	−29%	0.117	−15%
	Molecular[1] + Radiomic signature	0.122	−25%	0.109	−21%
Three layers	Molecular[1] + Clinical[2] + Radiomic signature	0.103	−37%	0.089	−36%

Annotation: 1 = including MGMT promoter methylation status and global DNA methylation glioblastoma subtypes; 2 = including patient's age, KPS, EOR, and adjuvant treatment; 3 = including tumor volumes from contrast enhancement, necrosis, and edema; 4 = IBSs for the null (reference) models were 0.163 for OS and 0.138 for PFS.

19

YONSEI UNIVERSITY

Analysis of deviance for different Cox regression models : Likelihood ratio test (LRT)

- Conventional ANOVA
  - to compare the means between groups

ANOVA

Time

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	91.467	2	45.733	4.467	.021
Within Groups	276.400	27	10.237		
Total	367.867	29			

- ANOVA to determine the model fit [in R]

```
anova(fm1, fm2)

    Model df      AIC      BIC    logLik   Test  L.Ratio p-value
fm1     1  5 239.4856 251.6397 -114.7428
fm2     2  6 238.9662 253.5511 -113.4831 1 vs 2 2.519406  0.1125
```

20

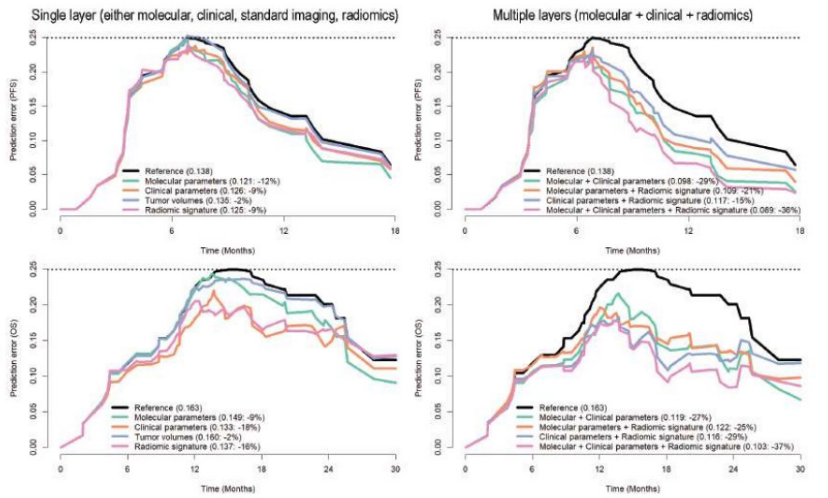


Fig. 2 Prediction error curves for stratifying PFS (upper row) and OS (lower row) based on a single layer (left column)—ie, either molecular (including MGMT promoter methylation status + global DNA methylation glioblastoma subtypes) or clinical (patient's age + KPS, EOR, adjuvant treatment) information or standard imaging parameters (tumor volumes from contrast enhancement, necrosis, and edema) or the radiomic signature—or (right column) combining the information from multiple layers. Prediction error rates are given in brackets (including the percentage reduction compared with the null model with no explanatory value). Combining the information from multiple layers (right column) allowed reduction of the prediction error beyond every single layer model (left column). The identified radiomic signature reduced the prediction error beyond molecular and clinical features and combining molecular + clinical information and the radiomic signature yielded the highest accuracy, with a reduction of the prediction error by 36% for PFS and 37% for OS (compared with 29% and 27% for a model without the radiomic signature that includes only molecular and clinical information).

Prediction error curve
Integrated Brier score

- Prediction error: time-dependent expected Brier score
- Integrated Brier score
  - Weighted average of Brier score
  - 0, perfect model
  - 0.25, non-informative model with a 50% incidence of the outcome

* Brier score: the squared difference between observed survival status and a model based prediction of surviving time t.

22

YONSEI UNIVERSITY

Assessing the model performance

TABLE 1. Characteristics of Some Traditional and Novel Performance Measures

Aspect	Measure	Visualization	Characteristics
Overall performance	R^2, Brier	Validation graph	Better with lower distance between Y and $\hat{Y}$. Captures calibration and discrimination aspects
Discrimination	c statistic	ROC curve	Rank order statistic; interpretation for a pair of subjects with and without the outcome
	Discrimination slope	Box plot	Difference in mean of predictions between outcomes; easy visualization
Calibration	Calibration-in-the-large	Calibration or validation graph	Compare mean (y) versus mean ($\hat{y}$); essential aspect for external validation
	Calibration slope		Regression slope of linear predictor; essential aspect for internal and external validation; related to "shrinkage" of regression coefficients
	Hosmer-Lemeshow test		Compares observed to predicted by decile of predicted probability
Reclassification	Reclassification table	Cross-table or scatter plot	Compare classifications from 2 models (one with, one without a marker) for changes
	Reclassification statistic		Compare observed outcomes to predicted risks within cross-classified categories
	Net reclassification index (NRI)		Compare classifications from 2 models for changes by outcome for a net calculation of changes in the right direction
	Integrated discrimination index (IDI)	Box plots for 2 models (one with, one without a marker)	Integrates the NRI over all possible cut-offs; equivalent to difference in discrimination slopes
Clinical usefulness	Net benefit (NB)	Cross-table	Net number of true positives gained by using a model compared to no model at a single threshold (NB) or over a range of thresholds (DCA)
	Decision curve analysis (DCA)	Decision curve	

EW Steyerberg et al. et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010;21:128-38.

YONSEI UNIVERSITY

Evaluation of prediction model

- Calibration

  : The ability to distinguish between yes/no, 0/1 on the dependent variable
  - Hosmer-Lemeshow test, Calibration plot
- Discrimination

  : The ability to generate predicted probabilities that reflect the true probability of a 0 or 1

  : Not dependent on arbitrary threshold choices
  - ROC(Receive-Operating Characteristic) curve
  - C-statistics

24

YONSEI UNIVERSITY

KSSR 2021

Discrimination ability for binary outcome

- AUC (Area under the ROC curve)

  = Averaged sensitivity for all possible values of specificity

  = Probability that abnormal case rated higher than normal case

  = ROC curve for predicted probability based on logistic reg.

  = c-statistic (concordance index) for binary outcome

$$c = \frac{\text{Number of concordant pairs} + 0.5(\text{number of tied pairs})}{\text{Number of all informative pairs}}$$

where, taking all possible pairs of subjects consisting of one subject who experienced the event of interest and one subject who did not experience the event of interest

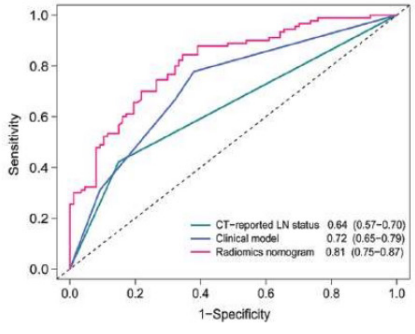
- Concordant pair?

  : the subject who experienced the event had a higher predicted probability of experiencing the event than the subject who did not experience the event

25

YONSEI UNIVERSITY

KSSR 2021



GW Ji, et al., Biliary Tract Cancer at CT: A Radiomics-based Model to Predict Lymph Node Metastasis and Survival Outcomes. Radiol 2019; 290: 90-98

26

YONSEI UNIVERSITY

The TRIPOD Statement: Explanation and Elaboration

1. Traditional Measures

Discrimination refers to the ability of a prediction model to differentiate between those who do or do not experience the outcome event. A model has perfect discrimination if the predicted risks for all individuals who have (diagnostic) or develop (prognosis) the outcome are higher than those for all individuals who do not experience the outcome. Discrimination is commonly estimated by the so-called concordance index (*c*-index). The *c*-index reflects the probability that for any randomly selected pair of individuals, one with and one without the outcome, the model assigns a higher probability to the individual with the outcome (526). The *c*-index is identical to the area under the receiver-operating characteristic curve for models with binary endpoints, and can be generalized for time-to-event (survival) models accounting for censoring. For survival models, a number of different *c*-indices have been proposed (527); authors should state clearly which measure is used, including an appropriate reference. More recently, extensions to the *c*-index for models with more than 2 outcome categories (528), competing risks (529), and clustering have been proposed (170, 171).

27

YONSEI UNIVERSITY

Discrimination ability for survival outcome

- C-statistic (Harrell's)
  - Harrell et al., Evaluating the yield of medical tests. JAMA 1982; 247(18): 2543–2546.
  - Harrell et al., Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15:361–87
- Modified c-statistic
  - Uno's c-index, Gonen and Heller's c-index
  - Censoring pattern 고려 (skewed or heavy censoring)
- Time-dependent ROC curve
  - Heagerty PJ, Lumley T and Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics 2000; 56(2): 337–344.,
  - 비교적 덜 보수적
- iAUC with bootstrapping
  - Integrated AUC
  - Heagerty PJ and Zheng Y. Survival model predictive accuracy and ROC curves. Biometrics 2005; 61(1): 92–105.
  - 시간의 흐름에 따라 비교 가능

28

YONSEI UNIVERSITY

KSSR 2021

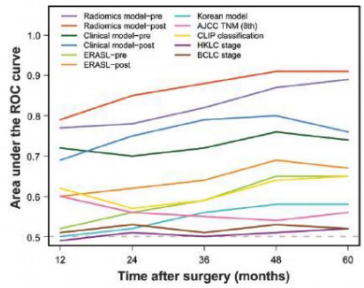
Table 3: Prognostic Performance of Radiomics-based Models Compared with Other Models and Staging Systems

	Development Cohort				Test Cohort			
Model	C Index	Time-Dependent AUC	IBS	*P* Value	C Index	Time-Dependent AUC	IBS	*P* Value
Preoperative radiomics model	0.77 (0.72, 0.82)	0.82	0.14	Reference	0.78 (0.72, 0.84)	0.84	0.14	Reference
Postoperative radiomics model	0.81 (0.76, 0.85)	0.88	0.12	Reference	0.82 (0.76, 0.87)	0.88	0.13	Reference
Preoperative clinical model	0.68 (0.62, 0.74)	0.72	0.18	<.001*	0.64 (0.57, 0.71)	0.71	0.18	<.001*
Postoperative clinical model	0.73 (0.67, 0.78)	0.76	0.17	<.001†	0.63 (0.55, 0.71)	0.67	0.18	<.001†
Preoperative ERASL model	0.57 (0.51, 0.63)	0.59	0.21	<.001*	0.57 (0.49, 0.64)	0.59	0.20	<.001*
Postoperative ERASL model	0.61 (0.54, 0.68)	0.64	0.19	<.001†	0.59 (0.51, 0.67)	0.63	0.19	<.001†
Korean model	0.53 (0.46, 0.59)	0.56	0.21	<.001†	0.56 (0.48, 0.64)	0.58	0.20	<.001†
BCLC stage	0.54 (0.37, 0.71)	0.52	0.21	.005*	0.54 (0.33, 0.74)	0.51	0.19	.01*
HKLC stage	0.51 (0.34, 0.69)	0.51	0.21	.004*	0.51 (0.34, 0.67)	0.50	0.19	.001*
CLIP classification	0.66 (0.57, 0.76)	0.62	0.20	.01*	0.68 (0.56, 0.80)	0.63	0.19	.04*
AJCC TNM (eighth edition)	0.65 (0.52, 0.79)	0.56	0.21	.009†	0.69 (0.55, 0.84)	0.60	0.20	.03†

Note.—Numbers in parentheses are the 95% confidence interval. The time-dependent AUC represents the median AUC at various time points. All *P* values were obtained from analyses comparing the C indexes of various models by using the "survcomp" package in R software. AJCC = American Joint Committee on Cancer, AUC = area under the receiver operating characteristic curve, BCLC = Barcelona Clinic Liver Cancer, CLIP = Cancer of the Liver Italian Program, ERASL = Early Recurrence After Surgery for Liver Tumor, HKLC = Hong Kong Liver Cancer, IBS = integrated Brier score.

* *P* value versus preoperative radiomics model.

† *P* value versus postoperative radiomics model.



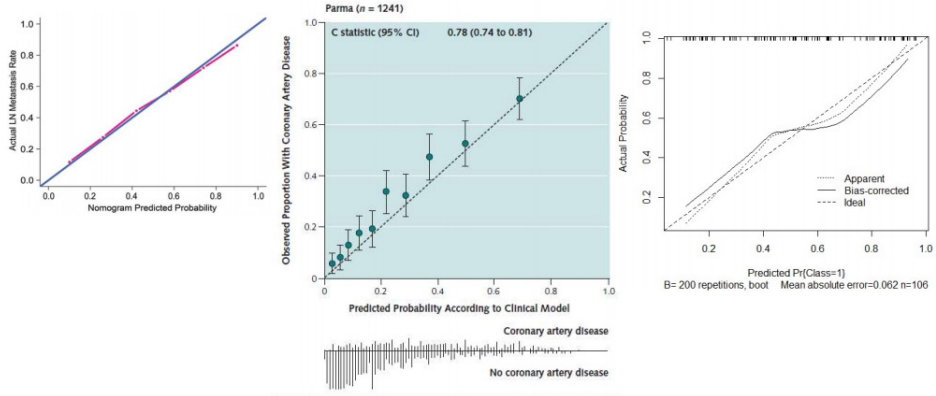
Ji GW., et al. (2020). Radiomic Features at Contrast-enhanced CT Predict Recurrence in Early Stage Hepatocellular Carcinoma: A Multi-Institutional Study. *Radiology, 294(3), 568-579.*

YONSEI UNIVERSITY

KSSR 2021

Calibration plot

- Observed probability vs. Predicted probability



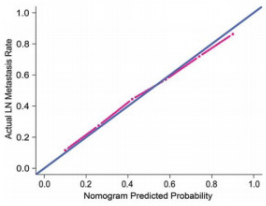
- GW Ji, et al., Biliary Tract Cancer at CT: A Radiomics-based Model to Predict Lymph Node Metastasis and Survival Outcomes. Radiol 2019; 290: 90-98
- TRIPOD statement: explanation and elaboration
- 'rms' R package

30

YONSEI UNIVERSITY

Calibration-in-the-large

- In calibration plot,
  - Intercept: the extent that predictions are systematically too low or too high
  - Slope: should be 1
- At validation, calibration-in-the-large problems are common.
  - Slope < 1: overfitting



- GW Ji, et al., Biliary Tract Cancer at CT: A Radiomics-based Model to Predict Lymph Node Metastasis and Survival Outcomes. Radiol 2019; 290: 90-98

31

YONSEI UNIVERSITY

Test for Calibration

: Hosmer-Lemeshow test

- P > 0.05 then the model fits well.
- have limited statistical power to evaluate poor calibration.
- sensitive to the grouping and sample size.
- often nonsignificant for small *N* and nearly always significant for large *N*.
- no indication of magnitude or direction of any miscalibration

∴ Prefer to give calibration plots.

32

YONSEI UNIVERSITY

KSSR 2021

Assessing the model performance

TABLE 1. Characteristics of Some Traditional and Novel Performance Measures

Aspect	Measure	Visualization	Characteristics
Overall performance	R^2, Brier	Validation graph	Better with lower distance between Y and $\hat{Y}$. Captures calibration and discrimination aspects
Discrimination	c statistic	ROC curve	Rank order statistic; interpretation for a pair of subjects with and without the outcome
	Discrimination slope	Box plot	Difference in mean of predictions between outcomes; easy visualization
Calibration	Calibration-in-the-large	Calibration or validation graph	Compare mean (y) versus mean ($\hat{y}$); essential aspect for external validation
	Calibration slope		Regression slope of linear predictor; essential aspect for internal and external validation; related to "shrinkage" of regression coefficients
	Hosmer-Lemeshow test		Compares observed to predicted by decile of predicted probability
Reclassification	Reclassification table	Cross-table or scatter plot	Compare classifications from 2 models (one with, one without a marker) for changes
	Reclassification statistic		Compare observed outcomes to predicted risks within cross-classified categories
	Net reclassification index (NRI)		Compare classifications from 2 models for changes by outcome for a net calculation of changes in the right direction
	Integrated discrimination index (IDI)	Box plots for 2 models (one with, one without a marker)	Integrates the NRI over all possible cut-offs; equivalent to difference in discrimination slopes
Clinical usefulness	Net benefit (NB) Decision curve analysis (DCA)	Cross-table Decision curve	Net number of true positives gained by using a model compared to no model at a single threshold (NB) or over a range of thresholds (DCA)

EW Steyerberg et al. et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010;21:128-38.

YONSEI UNIVERSITY

KSSR 2021

The TRIPOD Statement: Explanation and Elaboration

2. Quantifying the Incremental Value of an Additional Predictor

The advantage of multivariable analysis in contrast to single-marker or test research is that it generates direct evidence whether a test or marker has incremental value. However, quantifying the incremental value of adding a certain, often new, predictor to established predictors or even to an existing prediction model, by using the increase or improvement in the general, traditional performance measures (such as calibration, discrimination, or R^2), is difficult to interpret clinically (339, 340). Furthermore, there are concerns that such performance measures as the c-index are insensitive for assessing incremental value (341, 342), although its role as a descriptive measure still remains useful (343). Finally, statistical significance tests can be misleading, because statistically significant associations of new but weak predictors are easily found in a large sample.

34

YONSEI UNIVERSITY

Comparison of Prediction Models

- Assess the improvement in discrimination
  - ✓ Clinical only vs. Clinical + Imaging
  - ✓ Imaging only vs. Clinical + Imaging
  - ✓ OOO + Imaging #1 vs. OOO + Imaging #2
- The difference of two AUCs hardly significant.
- Need to quantify the improvement.

Pencina M, D'Agostino R, D'Agostino R, Vasan R (2008) Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med* 27(2):157–172
Pencina MJ, D'Agostino RB, Steyerberg EW (2011) Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat Med* 30(1):11–21

35

YONSEI UNIVERSITY

Alternative index for comparison

- NRI (Net Reclassification Improvement)

  = Event NRI + Non-event NRI

  = [P(up|D =1) - P(down|D =1)] + [P(down|D =0) - P(up|D =0)]

  ⇒ Report separately.

- Category-free NRI
  - Continuous NRI
  - considers any change in predicted risk for each individual

Table 3. Reclassification Tables Korean J Radiol 2016;17(3):339-350

Model without CCTA Finding	Model with CCTA Finding		
	< 10%	10–20%	≥ 20%
Death (n = 92)			
< 10%	17 (18.5)	13 (14.1)	0 (0.0)
≥ 10% and < 20%	5 (5.4)	4 (4.4)	19 (20.7)
≥ 20%	0 (0.0)	5 (5.4)	29 (31.5)
Survivor (n = 868)			
< 10%	525 (60.5)	70 (8.1)	0 (0.0)
≥ 10% and < 20%	104 (12.0)	25 (2.9)	58 (6.7)
≥ 20%	12 (1.4)	35 (4.0)	39 (4.5)

Values are numbers (percentages). Event NRI = (13 + 19 + 0) / 92 - (5 + 5 + 0) / 92 = (14.1% + 20.7%) - (5.4% + 5.4%) = 24.0%, Non-event NRI = (104 + 35 + 12) / 868 - (70 + 58 + 0) / 868 = (12.0% + 4.0% + 1.4%) - (8.1% + 6.7% + 0.0%) = 2.6%, Category-based NRI = 0.240 + 0.026 = 0.266 (95% CI, 0.131–0.400), Category-free NRI = 0.840 (95% CI, 0.654–1.025).
CCTA = coronary computed tomographic angiography, CI = confidence interval, NRI = net reclassification improvement

YONSEI UNIVERSITY

KSSR 2021

Alternative index for comparison

- IDI (Integrated Discrimination Improvement)

  = $(\bar{\hat{p}}_{new,events} - \bar{\hat{p}}_{old,events}) - (\bar{\hat{p}}_{new,nonevents} - \bar{\hat{p}}_{old,nonevents})$
- the difference in mean predicted probability between the two groups
- Example)

	Subject	Pr_new model	Pr_old model
Event	1	0.6998	0.8498
	...	...	...
	a	0.8556	0.3465
Non-event	a+1	0.8309	0.6493
	...	...	...
	b	0.4062	0.1433

IDI = 0.057 (= 0.051 - [-0.005])

37

YONSEI UNIVERSITY

KSSR 2021

The TRIPOD Statement: Explanation and Elaboration

3. Utility Measures

Explanation

Both discrimination and calibration are statistical properties characterizing the performance of a prediction model, but neither captures the clinical consequences of a particular level of discrimination or degree of miscalibration (359, 360). New approaches, such as decision curve analysis (361-363) and relative utility (364-366), offer insight to the clinical consequences or net benefits of using a prediction model at specific thresholds (349). They can also be used to compare the clinical usefulness of different models: for example, a basic and extended model fitted on the same data set, or even 2 different models (developed from 2 different data sets) validated on the same independent data set (367).

38

Research Article
Statistics in Medicine
Received 21 September 2012, Accepted 16 March 2014 Published online 9 April 2014 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/sim.6165

Assessing the incremental predictive performance of novel biomarkers over standard predictors

Table II. Comparison among the four methods—simulation results (scenario 2).

	Mean	Standard deviation	Median
Method 1 (add W to X_1, X_2, X_3, X_4)			
C before adding W	0.841	0.069	0.851
C after adding W	0.873	0.044	0.874
Difference in C-statistic	**0.031**	0.028	0.022
NRI	0.530	0.129	0.531
IDI	0.057	0.034	0.051
Method 2 (add W to risk score from current study)			
C before adding W	0.841	0.069	0.851
C before adding W	0.730	0.108	0.738
C after adding W	0.815	0.047	0.806
Difference in C-statistic	**0.085**	0.113	0.075
NRI	0.621	0.120	0.628
IDI	0.093	0.043	0.089
Method 4 (add W to a model with only intercept, null model)			
C before adding W	0.500	N/A	0.500
C after adding W	0.761	0.008	0.761
Difference in C-statistic	**0.260**	N/A	N/A
NRI	0.767	0.033	0.767
IDI	0.153	0.022	0.155

Poorer performance in the new sample

- **Overfitting**
- **Difference in populations**

39

YONSEI UNIVERSITY

Assessing the model performance

TABLE 1. Characteristics of Some Traditional and Novel Performance Measures

Aspect	Measure	Visualization	Characteristics
Overall performance	R^2, Brier	Validation graph	Better with lower distance between Y and $\hat{Y}$. Captures calibration and discrimination aspects
Discrimination	c statistic	ROC curve	Rank order statistic; interpretation for a pair of subjects with and without the outcome
	Discrimination slope	Box plot	Difference in mean of predictions between outcomes; easy visualization
Calibration	Calibration-in-the-large	Calibration or validation graph	Compare mean (y) versus mean ($\hat{y}$); essential aspect for external validation
	Calibration slope		Regression slope of linear predictor; essential aspect for internal and external validation; related to "shrinkage" of regression coefficients
	Hosmer-Lemeshow test		Compares observed to predicted by decile of predicted probability
Reclassification	Reclassification table	Cross-table or scatter plot	Compare classifications from 2 models (one with, one without a marker) for changes
	Reclassification statistic		Compare observed outcomes to predicted risks within cross-classified categories
	Net reclassification index (NRI)		Compare classifications from 2 models for changes by outcome for a net calculation of changes in the right direction
	Integrated discrimination index (IDI)	Box plots for 2 models (one with, one without a marker)	Integrates the NRI over all possible cut-offs; equivalent to difference in discrimination slopes
Clinical usefulness	Net benefit (NB) Decision curve analysis (DCA)	Cross-table Decision curve	Net number of true positives gained by using a model compared to no model at a single threshold (NB) or over a range of thresholds (DCA)

EW Steyerberg et al. et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010;21:128-38.

YONSEI UNIVERSITY

KSSR 2021

Decision Curve Analysis: A Novel Method for Evaluating Prediction Models

Andrew J. Vickers, PhD, Elena B. Elkin, PhD

Med Decis Making 2006;26:565–574

Harm of missed treatment

= ?

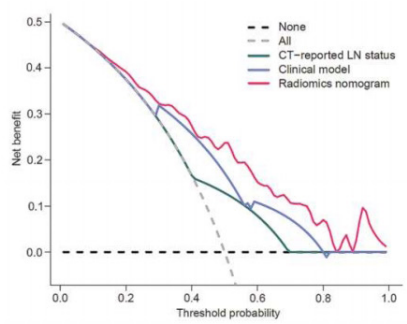
Harm of unnecessary treatment

41

YONSEI UNIVERSITY

KSSR 2021

Decision curve



- All: the assumption that all patients have LN metastases.
- None: the assumption that no patients have LN metastases.
- (y-axis) Net Benefit: summing the benefits (TP) and subtracting the harms (FP), weighting the latter by **the relative harm** of forgoing treatment compared with the negative consequences (harm) of an unnecessary treatment.
- Relative harm: $\frac{p_t}{1-p_t}$,
  p_t : threshold probability; where the expected benefit of treatment is equal to the expected benefit of avoiding treatment

GW Ji, et al., Biliary Tract Cancer at CT: A Radiomics-based Model to Predict Lymph Node Metastasis and Survival Outcomes. Radiol 2019; 290: 90-98

42

YONSEI UNIVERSITY

- p_t, threshold probability

test	+	-
+	a	b
-	c	d

$$p_t a + (1-p_t) b = p_t c + (1-p_t) d.$$

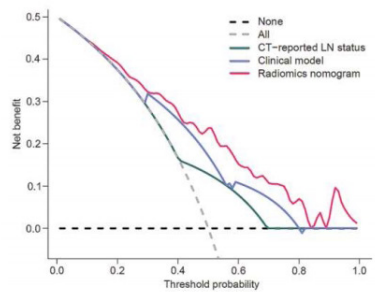
By some simple algebra:

$$p_t a - p_t c = (1-p_t) d - (1-p_t) b$$
$$\Rightarrow p_t (a-c) = (1-p_t)(d-b)$$
$$\Rightarrow \frac{a-c}{d-b} = \frac{1-p_t}{p_t}.$$

- a-c: harm associated with a FN
- d-b: harm associated with a FP

43

YONSEI UNIVERSITY



- **Interpretation**
  - If the threshold probability is over 10%, the application of radiomics model to predict lymph-node (LN) metastasis adds more benefit than treating all or none of the patients, clinical prediction model, and CT reported LN status.
  - The net benefit was comparable in lower threshold probability, on the basis of the radiomics nomogram and the clinical model.
  - If the test were harmful, the net benefit ≈ of the "ALL".

GW Ji, et al., Biliary Tract Cancer at CT: A Radiomics-based Model to Predict Lymph Node Metastasis and Survival Outcomes. Radiol 2019; 290: 90-98

44

YONSEI UNIVERSITY

KSSR 2021

How to make a decision curve?

1. Select a p_t
2. Positive test defined as $\hat{p} \geq p_t$
3. Calculate "Clinical Net Benefit" as:

$$\frac{TruePositiveCount}{n} - \frac{FalsePositiveCount}{n}\left(\frac{p_t}{1-p_t}\right)$$

4. Vary p_t over an appropriate range

- Extension
  - Net Benefit – *test harm*

"holistic" estimate of the negative consequence of having to take the test (cost, inconvenience, medical harms, etc.) in the units of a true-positive result.
Ex. FN is 50 times worse than having to undergo testing,
⇒ test harm=0.02
= If the test was perfect, we would probably perform no more than 50 tests to find a cancer

45

YONSEI UNIVERSITY

KSSR 2021

Available software

Outcome	Measures	SPSS (Menu)	MedcalC (Menu)	R (Packages or Functions)
Binary	Calibration Test	[Analyze] – [Regression] – [Binary Logistic]	[Regression] – [Logistic regression]	PredictABEL ResourceSelection rms
	c-index	[Analyze] – [ROC Curve]	[Statistics] – [ROC curves]	PredictABEL pROC
	NRI, IDI	Not Available	Not Available	PredictABEL Hmisc
Survival (Time-to-event)	Calibration Test	Not Available	Not Available	Rms pec
	c-index	Not Available	Not Available	Survival pec
	NRI, IDI	Not Available	Not Available	Hmisc survIDINRI

46

YONSEI UNIVERSITY

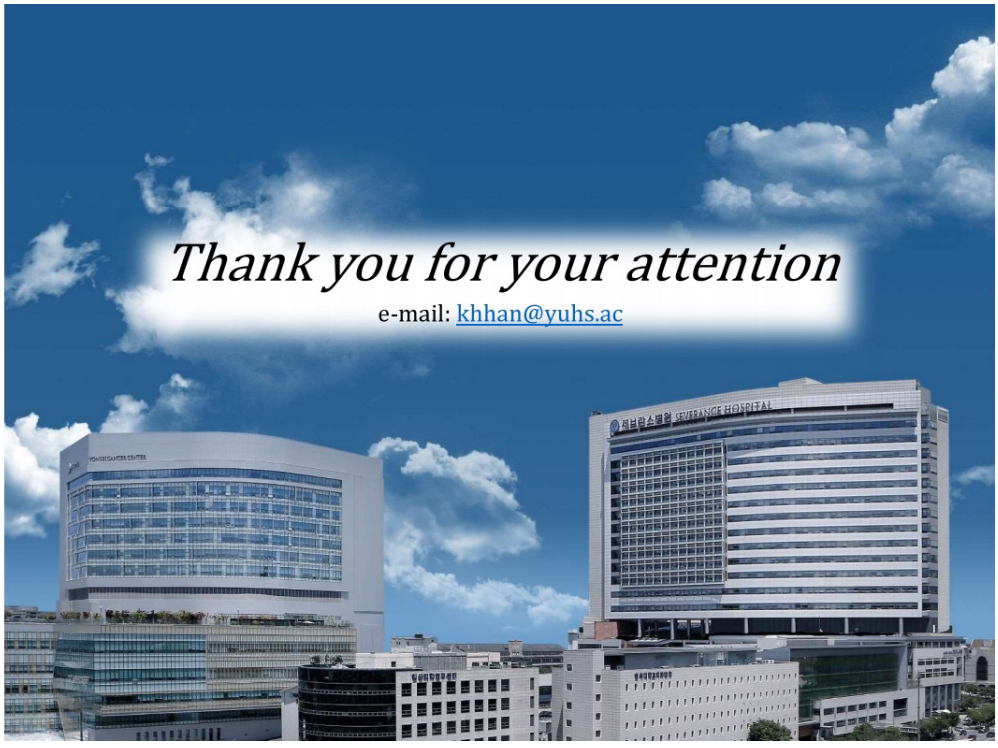
Assessing the model performance

TABLE 1. Characteristics of Some Traditional and Novel Performance Measures

Aspect	Measure	Visualization	Characteristics
Overall performance	R^2, Brier	Validation graph	Better with lower distance between Y and $\hat{Y}$. Captures calibration and discrimination aspects
Discrimination	c statistic	ROC curve	Rank order statistic; interpretation for a pair of subjects with and without the outcome
	Discrimination slope	Box plot	Difference in mean of predictions between outcomes; easy visualization
Calibration	Calibration-in-the-large	Calibration or validation graph	Compare mean (y) versus mean ($\hat{y}$); essential aspect for external validation
	Calibration slope		Regression slope of linear predictor; essential aspect for internal and external validation; related to "shrinkage" of regression coefficients
	Hosmer-Lemeshow test		Compares observed to predicted by decile of predicted probability
Reclassification	Reclassification table	Cross-table or scatter plot	Compare classifications from 2 models (one with, one without a marker) for changes
	Reclassification statistic		Compare observed outcomes to predicted risks within cross-classified categories
	Net reclassification index (NRI)		Compare classifications from 2 models for changes by outcome for a net calculation of changes in the right direction
	Integrated discrimination index (IDI)	Box plots for 2 models (one with, one without a marker)	Integrates the NRI over all possible cut-offs; equivalent to difference in discrimination slopes
Clinical usefulness	Net benefit (NB) Decision curve analysis (DCA)	Cross-table Decision curve	Net number of true positives gained by using a model compared to no model at a single threshold (NB) or over a range of thresholds (DCA)

- ✓ Internal validation
- ✓ External validation
- ✓ Model Updating

EW Steyerberg et al. et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010;21:128-38.

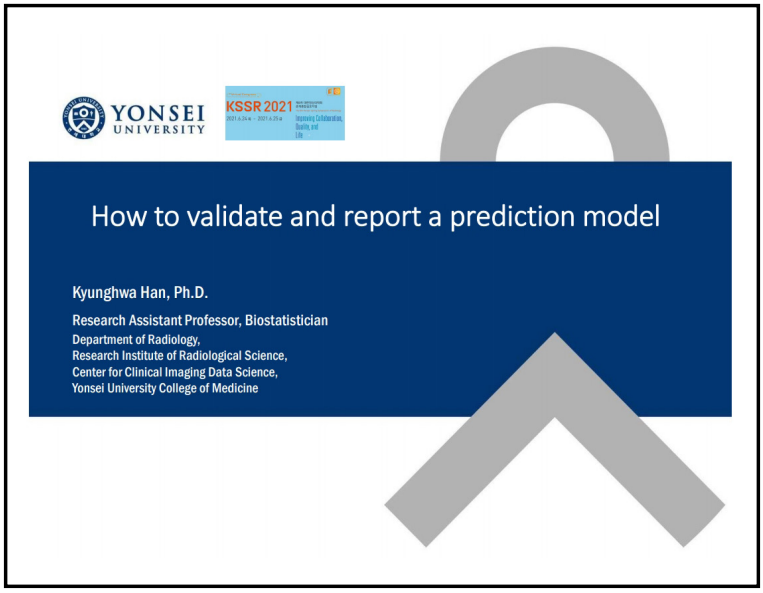


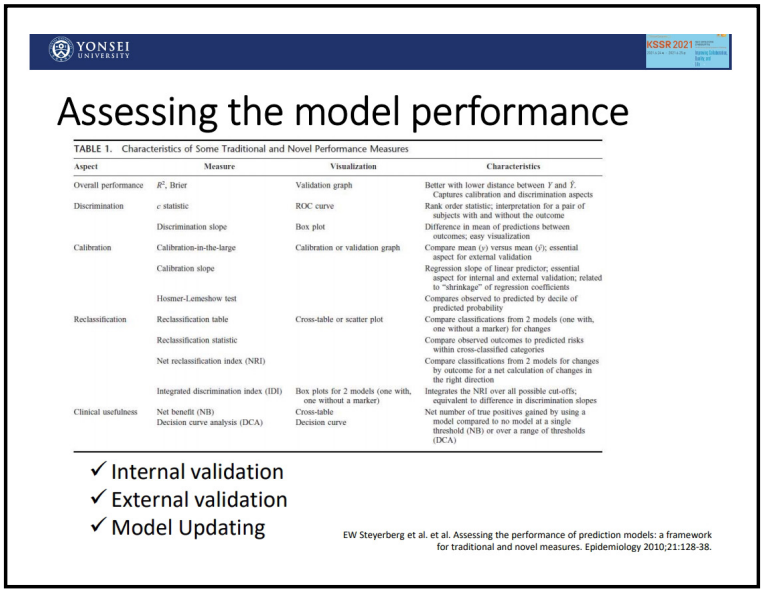
Clinical Research Methodology Course – Intermediate Course

15:20–15:50 Room 2

How to validate and report a prediction model

한 경 화
연세대학교





YONSEI UNIVERSITY

Validation of Prediction Model

	Internal validation	External validation
Purpose	· Reproducibility · Preventing against over-interpretation of current data	· Generalizability · External applicability
Method	· Split-sample validation · Cross-validation · Bootstrap validation	· Temporal validation · Fully independent validation

3

YONSEI UNIVERSITY

Split-sample validation

Table 1: Demographic and Clinical Characteristics of Included Patients

Variable	Training Cohort (n = 183)	Test Cohort (n = 133)
Median age (y)*	64.5 (59–71)	63 (58–71)
Median PSA (ng/mL)*	6.6 (4.9–9.5)	7.5 (5.4–11)
Median PSA density*	0.16 (0.10–0.26)	0.16 (0.11–0.23)
No. of patients without MRI-detected lesions	26	12
No. of patients with MRI-detected lesions†	157 (100)	121 (100)
1 lesion	86 (55)	47 (39)
2 lesions	58 (37)	53 (44)
3 lesions	11 (7)	18 (15)
4 lesions	2 (1)	3 (2)
No. of patients with specified maximum Gleason score†		
No prostate cancer	76 (42)	50 (38)
6 (3+3)	35 (19)	34 (25)
7a (3+4)	49 (27)	31 (23)
7b (4+3)	8 (4)	7 (5)
8 (4+4)	4 (2)	8 (6)
9a (4+5)	7 (4)	2 (2)
9b (5+4)	4 (2)	1 (1)
No. of patients with specified MRI index lesion†		
No lesion	26 (14)	12 (9)
PI-RADS 2	11 (6)	1 (1)
PI-RADS 3	42 (23)	30 (23)
PI-RADS 4	60 (33)	54 (40)
PI-RADS 5	44 (24)	36 (27)
No. of MRI-detected lesions negative for sPC†	163 (67)	159 (73)
No. of MRI-detected lesions positive for sPC†	80 (33)	60 (27)
Peripheral zone	54 (22)	37 (17)
Transition zone	26 (11)	23 (10)
No. of lesions with specified MRI assessment†		
Total	243 (100)	219 (100)
PI-RADS 2	21 (9)	4 (2)
PI-RADS 3	80 (33)	82 (37)
PI-RADS 4	91 (37)	88 (40)
PI-RADS 5	51 (21)	45 (21)
No. of MRI-detected lesions with specified zone distribution†		
Peripheral zone	144 (59)	131 (60)
Transition zone	75 (31)	79 (36)
Anterior fibromuscular stroma	17 (7)	9 (4)
Central zone	7 (3)	0 (0)

Note.—PSA = prostate-specific antigen, PI-RADS = Prostate Imaging Reporting and Data System, sPC = clinically significant prostate cancer.
* Data in parentheses are the interquartile range.
† Data in parentheses are percentages.

- Training
  - 2015.5-2016.1
- Test
  - 2016.1-2016.9

OR

- Random sampling

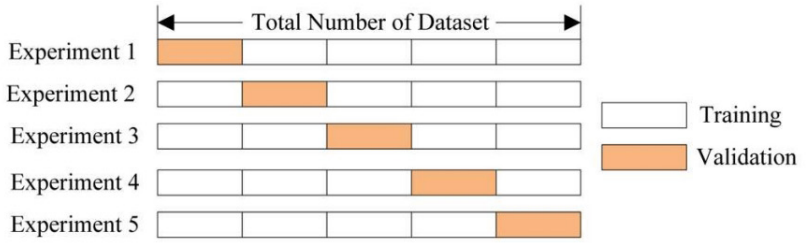
Bonekamp D., et al. (2018). Radiomic Machine Learning for Characterization of Prostate Lesions with MRI: Comparison to ADC Values. *Radiology, 289(1), 128-137*

4

YONSEI UNIVERSITY

KSSR 2021

5-fold cross-validation



5

YONSEI UNIVERSITY

KSSR 2021

nested cross-validation

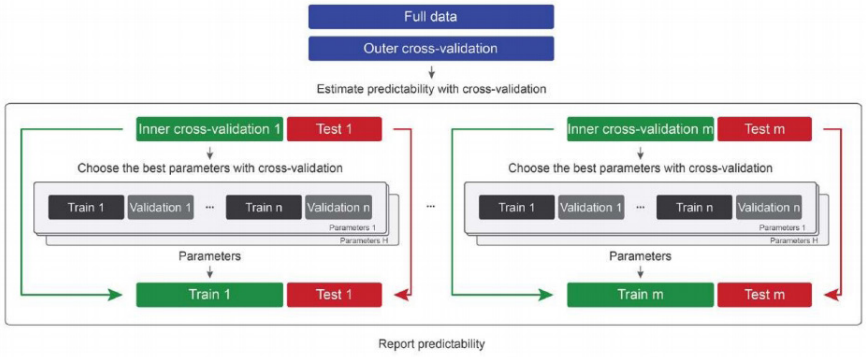


Fig 2. In the "Nested cross-validation" approach, first (outer) cross-validation is performed to estimate predictability of the data. In each iteration, data are divided into training and test sets. Before training, another (inner) cross-validation loop is used to optimize parameters. As model weights (fitted models) and parameters are different at every partition, it is not possible to report accuracy or statistical significance about a particular set of parameters or model weights.

doi:10.1371/journal.pone.0161788.g002

6

An Efficient Data Partitioning to Improve Classification Performance While Keeping Parameters Interpretable. PLoS ONE 2016

YONSEI UNIVERSITY

ness-of-fit test.[12,13] The primary models were internally validated using bootstrap resampling for 200 iterations.[14] For each resampling, the stepwise selection process was rerun, and the discrimination of the bootstrap model was assessed in the bootstrap sample and the full data set. The mean difference between these bootstrap model values was defined as the "optimism," and was subtracted from the final reported discrimination of the models.[15]

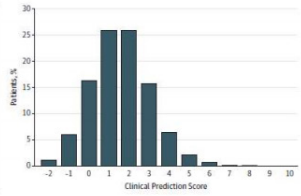
the model. The ischemic model had moderate discrimination (*c* statistic, 0.70 [95% CI, 0.68-0.73]) and was well calibrated (goodness-of-fit $P = .81$).

Increasing age was a significant independent predictor of bleeding, but not of ischemic events (Table 2). No tested interactions between covariates and randomized treatment for bleeding were retained in the model. The bleeding model showed similar discrimination to the ischemia model (*c* statistic, 0.68 [95% CI, 0.65-0.72]) and was well calibrated (goodness-of-fit $P = .34$). After bootstrap internal validation, optimism-corrected *c* statistics for both the ischemia (0.68 [95% CI, 0.65-0.70]) and bleeding models (0.66 [95% CI, 0.62-0.70]) were similar.

Figure 2. Elements of Clinical Prediction Score and Distribution of Score Among Randomized DAPT Study Patients (Derivation Cohort, 11 648 Patients)

Clinical Prediction Score

Variable	Points
Age, y	
≥75	−2
65–<75	−1
<65	0
Cigarette smoking	1
Diabetes mellitus	1
MI at presentation	1
Prior PCI or prior MI	1
Paclitaxel-eluting stent	1
Stent diameter <3 mm	1
CHF or LVEF <30%	2
Vein graft stent	2
Total score range: −2 to 10	



Yeh et al., Development and Validation of a Prediction Rule for Benefit and Harm of Dual Antiplatelet Therapy Beyond 1 Year After Percutaneous Coronary Intervention. *JAMA.* 2016;315(16):1735-1749

7

YONSEI UNIVERSITY

Bootstrap validation

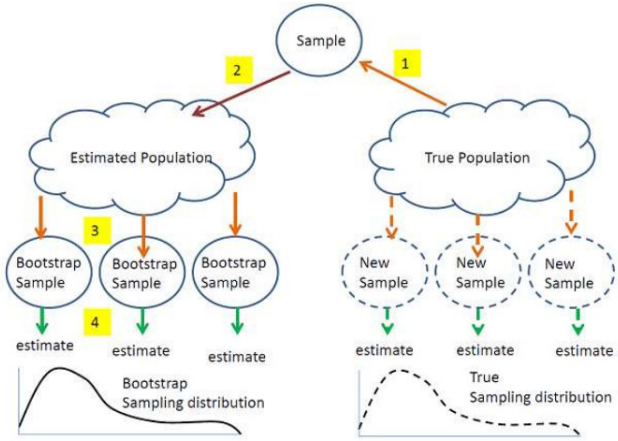
- Internal validated performance (추정 과정)
  - B개의 bootstrap sample을 추출
  - 각 sample에서 model을 만듬: bootstrap model
  - bootstrap model 을 이용하여 해당 sample에 대한 AUC 계산: AUC_{bi}
  - bootstrap model 을 이용하여 원래 자료에 대한 AUC 계산: AUC_{oi}
    ⇒이 과정을 B개의 bootstrap sample에 대해 시행
  - Bootstrap-validated estimate of the AUC
    {원래 자료의 AUC} - {B개의 차이들의 평균(mean(AUC_{bi} – AUC_{oi}))}
    Optimism correction

8

YONSEI UNIVERSITY

Bootstrap validation

- Simulation-based



https://newonlinecourses.science.psu.edu/stat555/node/119/

9

YONSEI UNIVERSITY

Bootstrap validation

- Bootstrap with 500 resampling (when c=0.737)

Replicate	c (bootstrap sample)	c (original data)	optimism	c (optimism corrected)
1	0.7531	0.7411	0.0120	0.7250
2	0.7356	0.6914	0.0441	0.6929
…	…	…	…	…

- Bootstrap-validated estimate of the AUC

  {원래 자료의 AUC} - {B개의 차이들의 평균(mean(AUC_{bi} – AUC_{oi}))}

  = 0.737 – 0.034 = 0.703 (± 0.03), (95% CI: 0.634, 0.741)

10

YONSEI UNIVERSITY

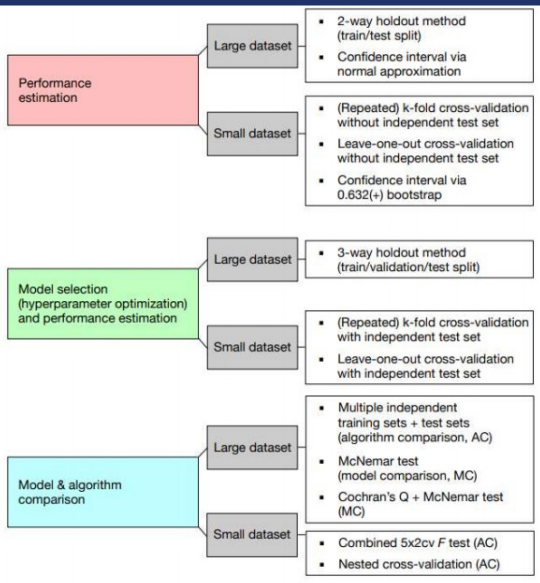


Figure 23: A recommended subset of techniques to be used to address different aspects of model evaluation in the context of small and large datasets. The abbreviation "MC" stands for "Model Comparison," and "AC" stands for "Algorithm Comparison," to distinguish these two tasks.

Sebastian Raschka. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. arXiv:1811.12808 [cs.LG]

11

YONSEI UNIVERSITY

ELSEVIER

Check for updates

Journal of Clinical Epidemiology 103 (2018) 131–133

Journal of Clinical Epidemiology

COMMENTARY

Validation in prediction research: the waste by data splitting

Ewout W. Steyerberg[a,b,*]

[a]Professor of Clinical Biostatistics and Medical Decision Making, Chair, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands

[b]Professor of Medical Decision Making, Department of Public Health, Erasmus MC, Rotterdam, The Netherlands

Accepted 24 July 2018; Published online 29 July 2018

- Split data or internal validation?
  - random data splitting should be abolished for validation of prediction models
  - In small samples, cross-validation and bootstrapping are more efficient approaches.

12

YONSEI UNIVERSITY

KSSR 2021

ELSEVIER

CrossMark

Journal of Clinical Epidemiology 79 (2016) 76–85

Journal of Clinical Epidemiology

Geographic and temporal validity of prediction models: different approaches were useful to examine model performance

Peter C. Austin[a,b,c,*], David van Klaveren[d,e], Yvonne Vergouwe[d], Daan Nieboer[d], Douglas S. Lee[a,b,f], Ewout W. Steyerberg[d]

- Geographical or temporal?
  - Validation studies of clinical prediction models should carefully describe whether overall validity of a model is reported, or that transportability is addressed by assessment of geographical or temporal variability in performance.

13

YONSEI UNIVERSITY

KSSR 2021

ELSEVIER

CrossMark

Journal of Clinical Epidemiology 68 (2015) 279–289

Journal of Clinical Epidemiology

ORIGINAL ARTICLES

A new framework to enhance the interpretation of external validation studies of clinical prediction models

Thomas P.A. Debray[a,*], Yvonne Vergouwe[b], Hendrik Koffijberg[a], Daan Nieboer[b], Ewout W. Steyerberg[b,1], Karel G.M. Moons[a,1]

- Internal validation studies assess model reproducibility.
- External validation studies do not necessarily assess model transportability (to a large extent).
- When externally validating a prediction model, researchers **should evaluate and quantify the relatedness between the population of the development and validation samples**
- otherwise, inferences on the actual clinical value or transportability of a prediction model may be misleading and cause prediction models to be implemented in incompatible populations.

14

YONSEI UNIVERSITY

Sample size calculation for external validation

ELSEVIER
Journal of Clinical Epidemiology 135 (2021) 79–89
Journal of Clinical Epidemiology

ORIGINAL ARTICLE

External validation of clinical prediction models: simulation-based sample size calculations were more reliable than rules-of-thumb

Kym I.E. Snell[a,*], Lucinda Archer[a], Joie Ensor[a], Laura J. Bonnett[b], Thomas P.A. Debray[c], Bob Phillips[d], Gary S. Collins[e,f], Richard D. Riley[a]

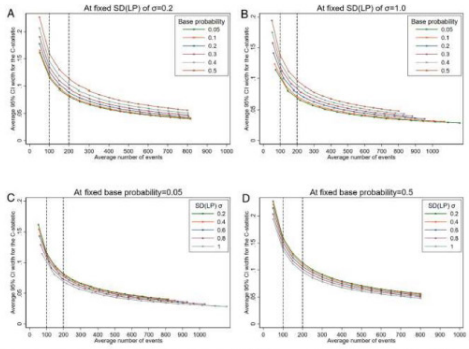


Fig. 2. Average 95% confidence interval width for the C-statistic at different effective sample sizes (based on average number of events in the simulation scenario) comparing by base probabilities at fixed SD(*LP*) (panels A and B), or comparing by SD(*LP*) at fixed base probabilities (panels C and D).

Table 3. Sample size and number of events required to target precise performance measures in an external validation study of a DVT prediction model, with an assumed linear predictor that follows a Normal(-1.75, 1.47^2) distribution and assuming the model is well calibrated ($\gamma = 0$ and $S = 1$ in Eq. 2).

Performance Measure	Targeted 95% CI width	Sample size (events) required to achieve CI width
C-statistic	0.1	385 (85)
Calibration slope	0.2	2430 (531)
Ln(observed/expected)	0.2	1379 (302)

15

YONSEI UNIVERSITY

Sample size calculation for external validation

Minimum sample size for external validation of a clinical prediction model with a binary outcome

Richard D. Riley[1] | Thomas P. A. Debray[2] | Gary S. Collins[3,4] | Lucinda Archer[1] | Joie Ensor[1] | Maarten van Smeden[2] | Kym I. E. Snell[1]

to precisely estimate

- calibration (Observed/Expected and calibration slope)
- discrimination (C-statistic)
- clinical utility (net benefit)

Statistics in Medicine. 2021 in press

16

Reporting guidelines

YONSEI UNIVERSITY

KSSR 2021

http://www.equator-network.org/

equator network

Enhancing the QUAlity and Transparency Of health Research

EQUATOR resources in German | Portuguese | Spanish

Home | About us | Library | Toolkits | Courses & events | News | Blog | Librarian Network | Contact

Your one-stop-shop for writing and publishing high-impact health research

find reporting guidelines | improve your writing | join our courses | run your own training course | enhance your peer review | implement guidelines

Library for health research reporting

The Library contains a comprehensive searchable database of reporting guidelines and also links to other resources relevant to research reporting.

- Search for reporting guidelines
- Not sure which reporting guideline to use?
- Reporting guidelines under development
- Visit the library for more resources

Reporting guidelines for main study types

Randomised trials	CONSORT	Extensions
Observational studies	STROBE	Extensions
Systematic reviews	PRISMA	Extensions
Study protocols	SPIRIT	PRISMA-P
Diagnostic/prognostic studies	STARD	TRIPOD
Case reports	CARE	Extensions
Clinical practice guidelines	AGREE	RIGHT
Qualitative research	SRQR	COREQ
Animal pre-clinical studies	ARRIVE	
Quality improvement studies	SQUIRE	Extensions
Economic evaluations	CHEERS	

See all 463 reporting guidelines

Developing a new reporting guideline?

LET THE WORLD KNOW!

Register with us

EQUATOR Network

18

STARD 2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies[1]

Radiology

Table 1

The STARD 2015 List

Section and Topic	No.	Item
TITLE OR ABSTRACT		
	1	Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity, specificity, predictive values, or AUC)
ABSTRACT		
	2	Structured summary of study design, methods, results, and conclusions (for specific guidance, see STARD for Abstracts)
INTRODUCTION		
	3	Scientific and clinical background, including the intended use and clinical role of the index test
	4	Study objectives and hypotheses
METHODS		
Study design	5	Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study)
Participants	6	Eligibility criteria
	7	On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, inclusion in registry)
	8	Where and when potentially eligible participants were identified (setting, location and dates)
	9	Whether participants formed a consecutive, random or convenience series

19

YONSEI UNIVERSITY

Test methods	10a	Index test, in sufficient detail to allow replication
	10b	Reference standard, in sufficient detail to allow replication
	11	Rationale for choosing the reference standard (if alternatives exist)
	12a	Definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory
	12b	Definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing pre-specified from exploratory
	13a	Whether clinical information and reference standard results were available to the performers/readers of the index test
	13b	Whether clinical information and index test results were available to the assessors of the reference standard
Analysis	14	Methods for estimating or comparing measures of diagnostic accuracy
	15	How indeterminate index test or reference standard results were handled
	16	How missing data on the index test and reference standard were handled
	17	Any analyses of variability in diagnostic accuracy, distinguishing pre-specified from exploratory
	18	Intended sample size and how it was determined
RESULTS		
Participants	19	Flow of participants, using a diagram
	20	Baseline demographic and clinical characteristics of participants
	21a	Distribution of severity of disease in those with the target condition
	21b	Distribution of alternative diagnoses in those without the target condition
	22	Time interval and any clinical interventions between index test and reference standard
Test results	23	Cross tabulation of the index test results (or their distribution) by the results of the reference standard
	24	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)
	25	Any adverse events from performing the index test or the reference standard
DISCUSSION		
	26	Study limitations, including sources of potential bias, statistical uncertainty, and generalisability
	27	Implications for practice, including the intended use and clinical role of the index test
OTHER INFORMATION		
	28	Registration number and name of registry
	29	Where the full study protocol can be accessed
	30	Sources of funding and other support; role of funders

20

YONSEI UNIVERSITY

KSSR 2021

Annals of Internal Medicine RESEARCH AND REPORTING METHODS

Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement

Gary S. Collins, PhD; Johannes B. Reitsma, MD, PhD; Douglas G. Altman, DSc; and Karel G.M. Moons, PhD

Annals of Internal Medicine RESEARCH AND REPORTING METHODS

Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration

Karel G.M. Moons, PhD; Douglas G. Altman, DSc; Johannes B. Reitsma, MD, PhD; John P.A. Ioannidis, MD, DSc; Petra Macaskill, PhD; Ewout W. Steyerberg, PhD; Andrew J. Vickers, PhD; David F. Ransohoff, MD; and Gary S. Collins, PhD

Ann Intern Med. 2015;162:55-63.
Ann Intern Med. 2015;162:W1-W73.

21

TRIPOD

TRIPOD Checklist: Prediction Model Development and Validation

Section/Topic	Item		Checklist Item	Page
Title and abstract				
Title	1	D;V	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	
Introduction				
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model or both.	
Methods				
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	
Participants	5a	D;V	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	
	5b	D;V	Describe eligibility criteria for participants.	
	5c	D;V	Give details of treatments received, if relevant.	
Outcome	6a	D;V	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	
	6b	D;V	Report any actions to blind assessment of the outcome to be predicted.	
Predictors	7a	D;V	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	
	7b	D;V	Report any actions to blind assessment of predictors for the outcome and other predictors.	
Sample size	8	D;V	Explain how the study size was arrived at.	
Missing data	9	D;V	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses.	
	10b	D	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	
	10c	V	For validation, describe how the predictions were calculated.	
	10d	D;V	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	
	10e	V	Describe any model updating (e.g., recalibration) arising from the validation, if done.	
Risk groups	11	D;V	Provide details on how risk groups were created, if done.	
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	

22

YONSEI UNIVERSITY

Results				
Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	
	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	
	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	
Model development	14a	D	Specify the number of participants and outcome events in each analysis.	
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome.	
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	
	15b	D	Explain how to the use the prediction model.	
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model.	
Model-updating	17	V	If done, report the results from any model updating (i.e., model specification, model performance).	
Discussion				
Limitations	18	D;V	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	
Interpretation	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data.	
	19b	D;V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	
Implications	20	D;V	Discuss the potential clinical use of the model and implications for future research.	
Other information				
Supplementary information	21	D;V	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	
Funding	22	D;V	Give the source of funding and the role of the funders for the present study.	

23

YONSEI UNIVERSITY

How to present Prediction Model?

- Regression formula
- Scoring system
- Nomogram
- etc…

24

YONSEI UNIVERSITY

KSSR 2021

Table 2. Multivariable Logistic Regression Analysis in Derivative Cohort

	Old Model			New Model		
	Adjusted OR	95% CI	*P*	Adjusted OR	95% CI	*P*
Age, years	1.073	1.021–1.128	0.005	1.059	1.005–1.115	0.031
Sex, male	3.899	2.381–6.385	< 0.001	3.311	1.996–5.492	< 0.001
Hypertension	1.458	0.861–2.468	0.161	1.282	0.745–2.206	0.369
Diabetes	2.755	1.750–4.338	< 0.001	2.407	1.504–3.852	< 0.001
Hyperlipidemia	0.838	0.457–1.538	0.569	0.754	0.403–1.413	0.379
Significant CAD at CCTA				4.669	2.789–7.816	< 0.001

CAD = coronary artery disease, CCTA = coronary computed tomographic angiography, CI = confidence interval, OR = odds ratio

- The predicted probability for a patient to death

$$p = \frac{\exp(-8.527 + 0.057\ \text{age} + 1.197\ \text{male} + 0.249\ \text{hypertension} + 0.878\ \text{diabetes} - 0.282\ \text{hyperlipidemia} + 1.541\ \text{significant CAD})}{1 + \exp(-8.527 + 0.057\ \text{age} + 1.197\ \text{male} + 0.249\ \text{hypertension} + 0.878\ \text{diabetes} - 0.282\ \text{hyperlipidemia} + 1.541\ \text{significant CAD})}$$

- EX) the predicted probability for a 77-year-old man with both hypertension and diabetes and significant CAD on CCTA

$$p = \frac{\exp(-8.527 + 0.057 \times 77 + 1.197 + 0.249 + 0.878 + 1.541)}{1 + \exp(-8.527 + 0.057 \times 77 + 1.197 + 0.249 + 0.878 + 1.541)} = 43.22\%.$$

Korean J Radiol 2016;17(3):339-350

YONSEI UNIVERSITY

KSSR 2021

Table 4. Scoring System to Calculate Point Values for Risk Score

Variables	β (1)	Categories (2)	Reference Value (W) (2)	β (W - W_{REF}) (3)	Points$_i$ = β (W - W_{REF}) / B (4, 5)
Age	0.057	70–74*	72 (W_{REF})	0	0
		75–79	77	0.285	1
		80–84	82	0.570	2
		85–92	88.5	0.941	3
Sex	1.197	Female*	0 (W_{REF})	0	0
		Male	1	1.197	4
Hypertension	0.249	No*	0 (W_{REF})	0	0
		Yes	1	0.249	1
Diabetes	0.878	No*	0 (W_{REF})	0	0
		Yes	1	0.878	3
Hyperlipidemia	-0.282	No*	0 (W_{REF})	0	0
		Yes	1	-0.282	-1
Significant CAD	1.541	No*	0 (W_{REF})	0	0
		Yes	1	1.541	5

*Reference category
1) Estimate the regression coefficients (β) of the multivariable model
2) Organize the risk factors into categories, determine the reference category, and reference values for each variable
3) Determine how far each category is from the reference category in regression units
4) Set the base constant (constant B)
5) Determine the number of points for each of the categories of each variable
CAD = coronary artery disease

Korean J Radiol 2016;17(3):339-350

Effect of Microvascular Invasion Risk on Early Recurrence of Hepatocellular Carcinoma After Surgery and Radiofrequency Ablation

Sunyoung Lee, MD,††† Tae Wook Kang, MD, Kyoung Doo Song, MD,* Min Woo Lee, MD,* Hyunchul Rhim, MD,* Hyo Keun Lim, MD,*‡ So Yeon Kim, MD,† Dong Hyun Sinn, MD,§ Jong Man Kim, MD,¶ Kyunga Kim, PhD,‡|| and Sang Yun Ha, MD***

TABLE 3. Multivariable Analysis of Predictors of Microvascular Invasion and Creation of the Microvascular Invasion Risk Score

Variable	Multivariable Analysis OR (95% CI)	P	β Coefficient	MVI Risk Points
α-FP ≥15, ng/mL [α-FP<15]	3.46 (1.62–7.39)	0.001	1.242	1.0
PIVKA-II ≥48, mAU/mL [PIVKA-II<48]	3.41 (1.54–7.55)	0.003	1.225	1.0
Arterial peritumoral enhancement [absence]	5.07 (2.36–10.87)	<0.001	1.622	1.5
Peritumoral hypointensity on HBP [absence]	15.98 (6.73–37.97)	<0.001	2.771	2.5

The reference category for each categorical variable is in the square brackets in first column. Multivariable logistic regression model was performed using stepwise backward variable selection. The scaled coefficients were simplified by rounding them to nearest half. The MVI risk score is obtained by adding the total number of points scored in each of the 4 variables.
MELD indicates Model for End-Stage Liver Disease.

Point value for each variable

Serum α-fetoprotein level (ng/mL)	Points
<15	0
≥15	1

Serum PIVKA-II level (mAU/mL)	Points
<48	0
≥48	1

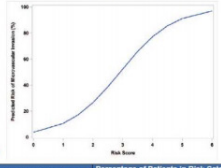
Peritumoral arterial enhancement	Points
(-)	0
(+)	1.5

Peritumoral hypointensity on HBP	Points
(-)	0
(+)	2.5

A

Summation of Points:	
Serum α-fetoprotein level	+
Serum PIVKA-II level	+
Peritumoral arterial enhancement	+
Peritumoral hypointensity on HBP	+
Total Risk Score =	

Predicted risk of microvascular invasion



B

Total Risk Score	Predicted Risk of MVI of HCC	Percentage of Patients in Risk Category, %	
		Derivation Cohort (n = 276)	External Validation Cohort (n = 101)
0	5.9	31.2	22.8
1	10.7	25.7	22.8
1.5	13.2	4.7	10.9
2	18.6	7.6	7.9
2.5	23.9	9.4	9.9
3.5	37.4	5.8	6.9
4	45.2	2.5	5.9
4.5	53.2	0.4	1.0
5	61.0	7.6	7.9
6	74.8	5.1	4.0

FIGURE 2. Four-variable risk index for microvascular invasion in patients with a small (≤3 cm) hepatocellular carcinoma. This model was able to stratify MVI risk ranging from less than 5.9% in those with a risk score of 0 to higher than 74.8% in those with a risk score of 6 in the external validation cohort.

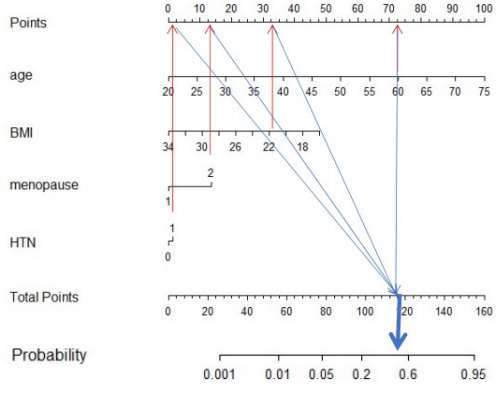
Ann Surg 2021;273:564–571

27

YONSEI UNIVERSITY

Nomogram

- Greek νόμος *nomos*, "law" and γραμμή *grammē*, "line"
- Parallel coordinate system ⇒ individualized predictions



28

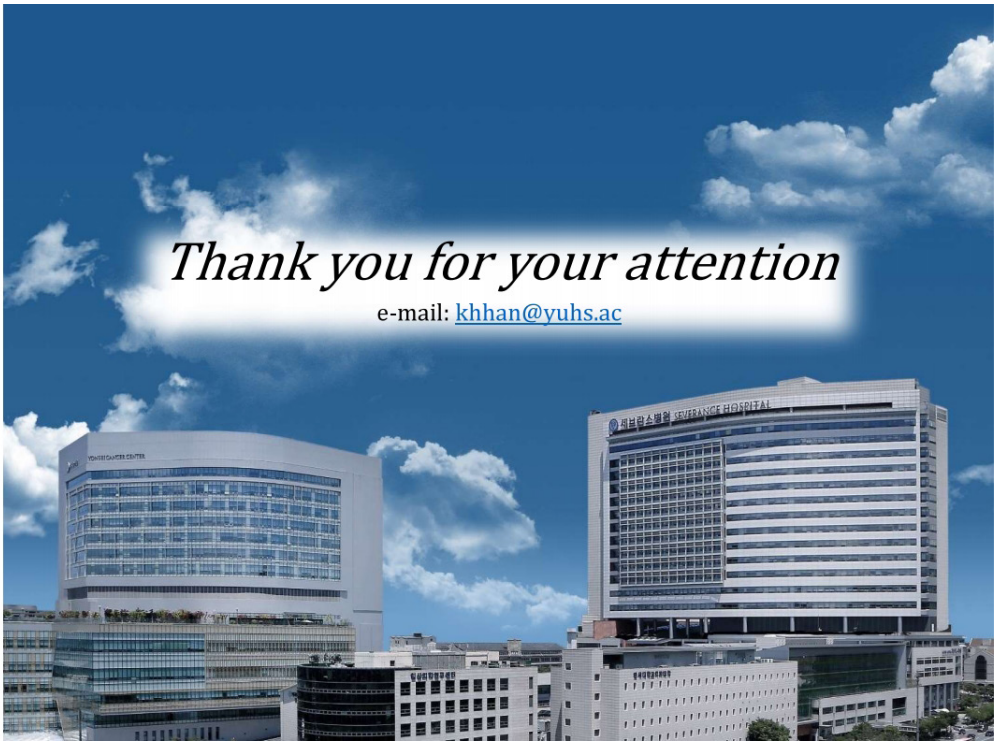
YONSEI UNIVERSITY

KSSR 2021

Summary

- Multivariable regression modeling
- Machine Learning classifier
- Predictors should be selected using both clinical knowledge and statistical reasoning.
- The model performance should be evaluated in terms of both calibration and discrimination.
- The validation, especially external validation, is an important aspect of establishing a predictive model.
- Performance of different predictive models can be compared using c-index, NRI, and IDI.
- Presentation of a predictive model

29

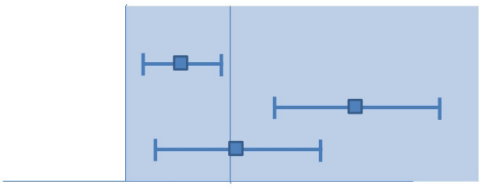


Clinical Research Methodology Course – Intermediate Course

16:00–16:40 Room 2

Noninferiority testing in radiology research

안 소 연
분당서울대학교병원



Noninferiority testing in radiology research

2021-06-24

제8차 대한영상의학회 춘계종합심포지엄(Korean Spring Symposium of Radiology; KSSR)
분당서울대학교병원
안소연

- **Introduction: rationale, examples**
- Statistical Concept of Equivalence/Noninferiority
  - hypothesis
- Noninferiority Margin
  - General Principles
  - Radiologic Perspective
- Examples

2

nomenclature

- Active control, standard treatment/modality (AC)
- Test, new treatment/modality (T)
- Placebo, sham control (P)

3

Noninferiority in Radiology research

- Superiority
  - radiology has been a highly technology driven field
- Why noninferiority
  - diagnostically saturated
  - safer, more convenient, and less costly

4

Example 1

전향적
RCT

- Hausleiter et al. 2010

- T: Low-dose CT (100-kVp tube voltage scan)
- AC: Standard-dose CT (120 kVp tube voltage scan)

- Primary outcome/margin:
  - image quality (continuous) / -0.2

JACC: CARDIOVASCULAR IMAGING
© 2010 BY THE AMERICAN COLLEGE OF CARDIOLOGY FOUNDATION
PUBLISHED BY ELSEVIER INC.
VOL. 3, NO. 11, 2010
ISSN 1936-878X/$36.00
DOI:10.1016/j.jcmg.2010.08.016

Image Quality and Radiation Exposure With a Low Tube Voltage Protocol for Coronary CT Angiography

Results of the PROTECTION II Trial

Jorg Hausleiter, MD,* Stefan Martinoff, MD,† Martin Hadamitzky, MD,* Eugenio Martuscelli, MD,‡ Iris Pschierer, MD,§ Gudrun M. Feuchtner, MD,‖ Paz Catalán-Sanz, MD,¶# Benedikt Czermak, MD,** Tanja S. Meyer, MD,* Franziska Hein, MD,* Bernhard Bischoff, MD,* Miriam Kuse,* Albert Schomig, MD,* Stephan Achenbach, MD††

Munich, Landshut, and Erlangen, Germany; Rome, Italy; Innsbruck, Austria; Oviedo and Barcelona, Spain

OBJECTIVES [illegible] image quality and radiation dose using a 100 kVp tube voltage scan protocol compared with standard 120 kVp for coronary computed [illegible]

BACKGROUND Concerns have been raised about radiation exposure during coronary CTA. The use of a 100 kVp tube voltage scan protocol effectively lowers coronary CTA radiation dose compared with standard 120 kVp, but it is unknown whether image quality is maintained.

METHODS We enrolled 400 nonobese patients who underwent coronary CTA: 202 patients were randomly assigned to a 100 kVp protocol and 198 patients to a 120 kVp protocol. The primary end point was to demonstrate noninferiority in image quality with the 100 kVp protocol, which was assessed by [illegible] (1 = nondiagnostic, 4 = excellent image quality). For the noninferiority analysis, [illegible] quality score points for the difference between both scan protocols was pre defined. Secondary end points included radiation dose and need for additional diagnostic tests during follow up.

RESULTS The mean image quality scores in patients scanned with 100 kVp and 120 kVp were 3.30 ± 0.67 and 3.28 ± 0.68, respectively ($p = 0.742$); image quality of the 100 kVp protocol was not inferior, as demonstrated by the 97.5% confidence interval of the difference, which did not cross the pre defined noninferiority margin of −0.2. The 100 kVp protocol was associated with a 31% relative reduction in radiation exposure (dose length product: 868 ± 317 mGy × cm with 120 kVp vs. 599 ± 255 mGy × cm with 100 kVp; $p < 0.0001$). At 30 day follow up, the need for additional diagnostic studies did not differ (13.4% vs. 19.2% for 100 kVp vs. 120 kVp, respectively; $p = 0.114$).

CONCLUSIONS A coronary CTA protocol using 100 kVp tube voltage maintained image quality, but reduced radiation exposure by 31% as compared with the standard 120 kVp protocol. Thus, 100 kVp scan protocols should be considered for nonobese patients to keep radiation exposure as low as reasonably achievable. (Prospective Randomized Trial on Radiation Dose Estimates of Cardiac CT Angiography in Patients Scanned With a 100 kVp Protocol [PROTECTION II]; NCT00611780) (J Am Coll Cardiol Img 2010;3:1113–23) © 2010 by the American College of Cardiology Foundation

5

Example 2

전향적
RCT

- Kim et al. 2012

- T: Low-dose CT
- AC: Standard-dose CT

- Primary outcome/margin:
  - negative appendectomy rate (diagnostic false positive rate)
  - absolute difference / 5.5%

The NEW ENGLAND JOURNAL of MEDICINE
ORIGINAL ARTICLE

Low-Dose Abdominal CT for Evaluating Suspected Appendicitis

Kyuseok Kim, M.D., Young Hoon Kim, M.D., So Yeon Kim, M.D., Suyoung Kim, R.N., Yoon Jin Lee, M.D., Kwang Pyo Kim, Ph.D., Hye Seung Lee, M.D., Soyeon Ahn, Ph.D., Taesun Kim, M.D., Seung-sik Hwang, M.D., Ki Jun Song, Ph.D., Sung Bum Kang, M.D., Duck Woo Kim, M.D., Seong Ho Park, M.D., and Kyoung Ho Lee, M.D.

From the Departments of Emergency Medicine (K.K., T.K.), Radiology (Y.H.K., S.K., Y.J.L., K.H.L.), Pathology (H.S.L.), and Surgery (S.-B.K., D.-W.K.), Seoul National University College of Medicine, and the Medical Research Collaborating Center (S.A.), Seoul National University Bundang Hospital, and the Department of Nuclear Engineering, Kyung Hee University (K.P.K.) — both in Gyeonggi-do; the Department of Radiology and Research Institute of Radiology, University of Ulsan College of Medicine, Asan Medical Center (S.Y.K., S.H.P.), and the Department of Biostatics, Yonsei University College of Medicine, Yonsei University Hospital (K.J.S.) — both in Seoul; and the Department of Social and Preventive Medicine, Inha University School of Medicine, Incheon (S.M.) — all in South Korea. [illegible] Department of Surgery, Seoul National University College of Medicine, 300 Gumi-dong, Bundang-gu, Seongnam-si, Gyeonggi-do 463-707, South Korea; [illegible]

N Engl J Med 2012;366:1596-605
Copyright © 2012 Massachusetts Medical Society

ABSTRACT

BACKGROUND
Computed tomography (CT) has become the predominant test for diagnosing acute appendicitis in adults. In children and young adults, exposure to CT radiation is of particular concern. We evaluated the rate of negative (unnecessary) appendectomy after low-dose versus standard-dose abdominal CT in young adults with suspected appendicitis.

METHODS
In this single-institution, single-blind, noninferiority trial, we randomly assigned 891 patients with suspected appendicitis to either low-dose CT (444 patients) or standard-dose CT (447 patients). The median radiation dose in terms of dose–length product was 116 mGy·cm in the low-dose group and 521 mGy·cm in the standard-dose group. The primary end point was the percentage of negative appendectomies among all nonincidental appendectomies, with a noninferiority margin of 5.5 percentage points. Secondary end points included the appendiceal perforation rate and the proportion of patients with suspected appendicitis who required additional imaging.

RESULTS
The negative appendectomy rate was 3.5% (6 of 172 patients) in the low-dose CT group and 3.2% (6 of 186 patients) in the standard-dose CT group (difference, 0.3 percentage points; 95% confidence interval, −3.8 to 4.6). The two groups did not differ significantly in terms of the appendiceal perforation rate (26.5% with low-dose CT and 23.3% with standard-dose CT, $P = 0.46$) or the proportion of patients who needed additional imaging tests (3.2% and 1.6%, respectively; $P = 0.09$).

CONCLUSIONS
Low-dose CT was noninferior to standard-dose CT with respect to negative appendectomy rates in young adults with suspected appendicitis. (Funded by GE Healthcare Medical Diagnostics and others; ClinicalTrials.gov number, NCT00913380.)

6

Example 3

전향적
RCT

- Kasivisvanathan et al. 2018
- T: Multiparametric magnetic resonance imaging (MRI), with or without targeted biopsy
- AC: Standard transrectal ultrasonography–guided biopsy
- Primary outcome/margin:
  - clinically significant cancer
  - absolute difference / -5% + *

The NEW ENGLAND JOURNAL of MEDICINE
ESTABLISHED IN 1812 MAY 10, 2018 VOL. 378 NO. 19

MRI-Targeted or Standard Biopsy for Prostate-Cancer Diagnosis

[illegible] Kasivisvanathan, A.S. Rannikko, M. Borghi, V. Panebianco, L.A. Mynderse, M.H. Vaarala, A. Briganti, L. Budäus, G. Hellawell, R.G. Hindley, M.J. Roobol, S. Eggener, M. Ghei, A. Villers, F. Bladou, G.M. Villeirs, J. Virdi, S. Boxler, G. Robert, P.B. Singh, W. Venderink, B.A. Hadaschik, A. Ruffion, J.C. Hu, D. Margolis, S. Crouzet, L. Klotz, S.S. Taneja, P. Pinto, F. Gill, C. Allen, F. Giganti, A. Freeman, S. Morris, S. Punwani, N.R. Williams, C. Brew-Graves, J. Deeks, Y. Takwoingi, M. Emberton, and C.M. Moore, for the PRECISION Study Group Collaborators[illegible]

ABSTRACT

BACKGROUND
Multiparametric magnetic resonance imaging (MRI), with or without targeted biopsy, [illegible] not undergone biopsy. However, comparative evidence is limited.

[illegible] randomized, noninferiority trial, we assigned men with a clinical suspicion of prostate cancer who had not undergone biopsy previously to undergo MRI [illegible] biopsy, followed by standard transrectal ultrasonography–guided biopsy (with 12 standard biopsy cores) if MRI suggested the presence of prostate cancer. Men whose MRI results were not suggestive of prostate cancer were not offered biopsy. Standard biopsy was a 10-to-12-core, transrectal ultrasonography–guided biopsy. The primary outcome was the proportion of men who received a diagnosis of clinically significant cancer. Secondary outcomes included the proportion of men who received a diagnosis of clinically insignificant cancer.

RESULTS
A total of 500 men underwent randomization. In the MRI-targeted biopsy group, 71 of 252 men (28%) had MRI results that were not suggestive of prostate cancer, so they did not undergo biopsy. Clinically significant cancer was detected in 95 men (38%) in the MRI-targeted biopsy group, as compared with 64 of 248 (26%) in the standard-biopsy group (adjusted difference, 12 percentage points; 95% confidence [illegible] to 20; P=0.005). MRI, with or without targeted biopsy, was noninferior to standard biopsy, and the 95% confidence interval indicated the superiority of this strategy over standard biopsy. Fewer men in the MRI-targeted biopsy group [illegible] standard-biopsy group received a diagnosis of clinically insignificant cancer (adjusted difference, −13 percentage points; 95% CI, −19 to −7; P<0.001).

CONCLUSIONS
The use of risk assessment with MRI before biopsy and MRI-targeted biopsy was superior to standard transrectal ultrasonography–guided biopsy in men at clinical risk for prostate cancer who had not undergone biopsy previously. (Funded by the National Institute for Health Research and the European Association of Urology Research Foundation; PRECISION ClinicalTrials.gov number, NCT02380027.)

The complete list of members of the PRECISION Study Group is provided in the Supplementary Appendix, available at NEJM.org.

This article was published on March 19, 2018, at NEJM.org.
N Engl J Med 2018;378:1767-77
DOI: 10.1056/NEJMoa1801993

7

AI-based Strategies to Reduce Workload in Breast Cancer Screening with Mammography and Tomosynthesis: A Retrospective Evaluation

후향적
paired 연구

José Luis Raya-Povedano, MD • Sara Romero-Martín, PhD, MD • Esperanza Elías-Cabot, MD • Albert Gubern-Mérida, PhD • Alejandro Rodríguez-Ruiz, PhD • Marina Álvarez-Benito, PhD, MD

From the Breast Cancer Unit, Department of Radiology, Hospital Universitario Reina Sofía, Av Menéndez Pidal s/n, Córdoba 14004, Spain (J.L.R.P., S.R.M., E.E.C., M.Á.B.); Maimonides Institute for Biomedical Research of Córdoba, Córdoba, Spain (J.L.R.P., S.R.M., E.E.C., M.Á.B.); and Department of Clinical Science, ScreenPoint Medical, Nijmegen, the Netherlands (A.G.M., A.R.R.). Received August 31, 2020; revision requested October 23; revision received January 5, 2021; accepted January 14.
Address correspondence to J.L.R.P. (e-mail: *joseluisrayapovedano@gmail.com*).

The study was funded by the Hospital Universitario Reina Sofía in Córdoba, Spain.

Conflicts of interest are listed at the end of this article.

Radiology 2021; 000:1–9 • https://doi.org/10.1148/radiol.2021203555 • Content codes: BR AI

Background: The workflow of breast cancer screening programs could be improved given the high workload and the high number of false-positive and false-negative assessments.

Purpose: To evaluate if using an artificial intelligence (AI) system could reduce workload without reducing cancer detection in breast cancer screening with digital mammography (DM) or digital breast tomosynthesis (DBT).

Materials and Methods: Consecutive screening-paired and independently read DM and DBT images acquired from January 2015 to December 2016 were retrospectively collected from the Córdoba Tomosynthesis Screening Trial. The original reading settings were single or double reading of DM or DBT images. An AI system computed a cancer risk score for DM and DBT examinations independently. Each original setting was compared with a simulated autonomous AI triaging strategy (the least suspicious examinations for AI are not human-read; the rest are read in the same setting as the original, and examinations not recalled by radiologists but graded as very suspicious by AI are recalled) in terms of workload, sensitivity, and recall rate. The McNemar test with Bonferroni correction was used for statistical analysis.

Results: A total of 15 987 DM and DBT examinations (which included 98 screening-detected and 15 interval cancers) from 15 986 women (mean age ± standard deviation, 58 years ± 6) were evaluated. In comparison with double reading of DBT images (568 hours needed, 92 of 113 cancers detected, 706 recalls in 15 987 examinations), AI with DBT would result in 72.5% less workload ($P < .001$, 156 hours needed), noninferior sensitivity (95 of 113 cancers detected, $P = .38$), and 16.7% lower recall rate ($P < .001$, 588 recalls in 15 987 examinations). Similar results were obtained for AI with DM. In comparison with the original double reading of DM images (222 hours needed, 76 of 113 cancers detected, 807 recalls in 15 987 examinations), AI with DBT would result in 29.7% less workload ($P < .001$), 25.0% higher sensitivity ($P < .001$), and 27.1% lower recall rate ($P < .001$).

Conclusion: Digital mammography and digital breast tomosynthesis screening strategies based on artificial intelligence systems could reduce workload up to 70%.

T: AI
AC: no AI

유방암 스크리닝 시
워크 로드를 줄이면서
cancer detection을 유지 하는 지

8

Original Article | Thyroid
eISSN 2005-8330
https://doi.org/10.3348/kjr.2019.0581
Korean J Radiol 2020;21(3):369-376
Check for updates
Korean Journal of Radiology KJR

Computer-Aided Diagnosis System for the Evaluation of Thyroid Nodules on Ultrasonography: Prospective Non-Inferiority Study according to the Experience Level of Radiologists

Sae Rom Chung, MD[1], Jung Hwan Baek, MD, PhD[1], Min Kyoung Lee, MD[1], Yura Ahn, MD[1], Young Jun Choi, MD, PhD[1], Tae-Yon Sung, MD, PhD[2], Dong Eun Song, MD, PhD[3], Tae Yong Kim, MD, PhD[4], Jeong Hyun Lee, MD, PhD[1]

Departments of [1]Radiology and Research Institute of Radiology, [2]Surgery, [3]Pathology, and [4]Endocrinology and Metabolism, University of Ulsan College of Medicine, Asan Medical Center, Seoul, Korea

Objective: To determine whether a computer-aided diagnosis (CAD) system for the evaluation of thyroid nodules is non-inferior to radiologists with different levels of experience.
Materials and Methods: Patients with thyroid nodules with a decisive diagnosis of benign or malignant nodule were consecutively enrolled from November 2017 to September 2018. Three radiologists with different levels of experience (1 month, 4 years, and 7 years) in thyroid ultrasound (US) reviewed the thyroid US with and without using the CAD system. Statistical analyses included non-inferiority testing of the diagnostic accuracy for malignant thyroid nodules between the CAD system and the three radiologists with a non-inferiority margin of 10%, comparison of the diagnostic performance, and the added value of the CAD system to the radiologists.
Results: Altogether, 197 patients were included in the study cohort. The diagnostic accuracy of the CAD system (88.5%, 95% confidence interval [CI] = 82.7–92.5) was non-inferior to that of the radiologists with less experience (1 month and 4 year) of thyroid US (83.0%, 95% CI = 76.5–88.0; $p < 0.001$), whereas it was inferior to that of the experienced radiologist (7 years) (95.8%, 95% CI = 91.4–98.0; $p = 0.138$). The sensitivity and negative predictive value of the CAD system were significantly higher than those of the less-experienced radiologists were, whereas no significant difference was found with those of the experienced radiologist. A combination of US and the CAD system significantly improved sensitivity and negative predictive value, although the specificity and positive predictive value deteriorated for the less-experienced radiologists.
Conclusion: The CAD system may offer support for decision-making in the diagnosis of malignant thyroid nodules for operators who have less experience with thyroid US.
Keywords: *Computer-aided diagnosis; Thyroid nodule; Thyroid cancer; Ultrasonography*

전향적
paired 연구

CAD가

radiologist 비해

노듈 evaluation
(accuracy) 측면에서

9

Assessment of an AI Aid in Detection of Adult Appendicular Skeletal Fractures by Emergency Physicians and Radiologists: A Multicenter Cross-sectional Diagnostic Study

Loïc Duron, MD, MSc • Alexis Ducarouge, MSc • André Gillibert, MD, MSc • Julia Lainé, MD, MSc • Christian Allouche • Nicolas Cherel, MSc • Zekun Zhang, MSc • Nicolas Nitche, MSc • Elise Lacave, MSc • Aloïs Pourchot, MSc • Adrien Felter, MD • Louis Lassalle, MD, MSc • Nor-Eddine Regnard, MD, MSc • Antoine Feydy, MD, PhD

From the Department of Radiology, Hôpital Fondation A. de Rothschild, 25 rue Manin, 75019 Paris, France (L.D.); Faculty of Medicine, Université de Paris, Paris, France (L.D., A. Feydy); Gleamer, Paris, France (A.D., C.A., N.C., Z.Z., N.N., E.L., A.P., N.E.R.); Department of Biostatistics, CHU Rouen, Rouen, France (A.G.); Department of Radiology, Hôpital Hôtel-Dieu, Assistance Publique-Hôpitaux de Paris, Paris, France (J.L.); Department of Radiology, Hôpital Ambroise-Paré, Assistance Publique-Hôpitaux de Paris, Boulogne-Billancourt, France (A. Felter); Department of Radiology, Hôpital Raymond-Poincaré, Assistance Publique-Hôpitaux de Paris, Garches, France (A. Felter); and Department of Radiology B, Hôpital Cochin, Assistance Publique-Hôpitaux de Paris, Paris, France (L.L., N.E.R., A. Feydy). Received September 30, 2020; revision requested December 23; revision received January 26, 2021; accepted March 4. **Address correspondence to** L.D. (e-mail: *lduron@for.paris*).

This study was funded by Gleamer.

Conflicts of interest are listed at the end of this article.

Radiology 2021; 000:1–10 • https://doi.org/10.1148/radiol.2021203886 • Content codes: MK AI

Background: The interpretation of radiographs suffers from an ever-increasing workload in emergency and radiology departments, while missed fractures represent up to 80% of diagnostic errors in the emergency department.

Purpose: To assess the performance of an artificial intelligence (AI) system designed to aid radiologists and emergency physicians in the detection and localization of appendicular skeletal fractures.

Materials and Methods: The AI system was previously trained on 60 170 radiographs obtained in patients with trauma. The radiographs were randomly split into 70% training, 10% validation, and 20% test sets. Between 2016 and 2018, 600 adult patients in whom multiview radiographs had been obtained after a recent trauma, with or without one or more fractures of shoulder, arm, hand, pelvis, leg, and foot, were retrospectively included from 17 French medical centers. Radiographs with quality precluding human interpretation or containing only obvious fractures were excluded. Six radiologists and six emergency physicians were asked to detect and localize fractures with ($n = 300$) and fractures without ($n = 300$) the aid of software highlighting boxes around AI-detected fractures. Aided and unaided sensitivity, specificity, and reading times were compared by means of paired Student t tests after averaging of performances of each reader.

Results: A total of 600 patients (mean age ± standard deviation, 57 years ± 22; 358 women) were included. The AI aid improved the sensitivity of physicians by 8.7% (95% CI: 3.1, 14.2; $P = .003$ for superiority) and the specificity by 4.1% (95% CI: 0.5, 7.7; $P < .001$ for noninferiority) and reduced the average number of false-positive fractures per patient by 41.9% (95% CI: 12.8, 61.3; $P = .02$) in patients without fractures and the mean reading time by 15.0% (95% CI: −30.4, 3.8; $P = .12$). Finally, stand-alone performance of a newer release of the AI system was greater than that of all unaided readers, including skeletal expert radiologists, with an area under the receiver operating characteristic curve of 0.94 (95% CI: 0.92, 0.96).

Conclusion: The artificial intelligence aid provided a gain of sensitivity (8.7% increase) and specificity (4.1% increase) without loss of reading speed.

후향적
~~cross-sectional~~
paired

AI가
detection/localization of
appendicular skeletal fracture
측면에서

radiologists
/emergency physicians
보조하는 지

10

MR Enterography for the Evaluation of Small-Bowel Inflammation in Crohn Disease by Using Diffusion-weighted Imaging without Intravenous Contrast Material: A Prospective Noninferiority Study[1]

Yeun Seo, MD
Seong Ho Park, MD
Kyung-Jo Kim, MD
Ji-Kyeong Kang, MD
Kedaun Lee, MD
Suk-Kyun Yang, MD
Byong Duk Ye, MD
Sang Hyoung Park, MD
So Yeon Kim, MD
Seunghee Baek, PhD
Kyunghwa Han, PhD
Hyun Kwon Ha, MD

[1] From the Department of Radiology and Research Institute of Radiology (N.S., Seong H. Park, S.Y.K., H.K.H.), Department of Gastroenterology (K.J.K., S.K.Y., B.D.Y., Sang H. Park), and Department of Clinical Epidemiology and Biostatistics (S.B.), University of Ulsan College of Medicine, Asan Medical Center, 88 Olympic-ro 43-gil, Songpa-gu, Seoul, South Korea; Department of Radiology, Hanyang University Medical Center, Seoul, South Korea (B.K.K.); Department of Radiology, Inje University College of Medicine, Busan, South Korea (Y.L.); and Biostatistics Collaboration Unit, Gangnam Medical Research Center, Yonsei University College of Medicine, Seoul, South Korea (K.H.). Received April 4, 2015; revision requested May 18; revision received June 3; accepted June 29; final version accepted June 29. Supported by a grant from Dongkook Pharmaceutical, Seoul, South Korea. The funder had no role in the study design; the collection, analysis, and interpretation of data; the writing of the report; or the decision to submit the manuscript for publication. The corresponding author had full access to all of the study data and had final responsibility for the decision to submit the manuscript for publication. **Address correspondence to** Seong Ho Park (e-mail: parksh.radiology@gmail.com).

Purpose: To determine whether magnetic resonance (MR) enterography performed with diffusion-weighted imaging (DWI) without intravenous contrast material is noninferior to contrast material–enhanced (CE) MR enterography for the evaluation of small-bowel inflammation in Crohn disease.

Materials and Methods: Institutional review board approval and informed consent were obtained for this prospective noninferiority study. Fifty consecutive adults suspected of having Crohn disease underwent clinical assessment, MR enterography, and ileocolonoscopy within 1 week. MR enterography included conventional imaging and DWI ($b = 900$ sec/mm^2). In 44 patients with Crohn disease, 171 small-bowel segments that were generally well distended and showed a wide range of findings, from normalcy to severe inflammation (34 men, 10 women; mean age ± standard deviation, 26.9 years ± 6.1), were selected for analysis. Image sets consisting of *(a)* T2-weighted sequences with DWI and *(b)* T2-weighted sequences with CE T1-weighted sequences were reviewed by using a crossover design with blinding and randomization. Statistical analyses included noninferiority testing regarding proportional agreement between DWI and CE MR enterography for the identification of bowel inflammation with a noninferiority margin of 80%, correlation between DWI and CE MR enterography scores of bowel inflammation severity, and comparison of accuracy between DWI and CE MR enterography for the diagnosis of terminal ileal inflammation by using endoscopic findings as the reference standard.

Results: The agreement between DWI and CE MR enterography for the identification of bowel inflammation was 91.8% (157 of 171 segments; one-sided 95% confidence interval: ≥88.4%). The correlation coefficient between DWI and CE MR enterography scores was 0.937 ($P < .001$). DWI and CE MR enterography did not differ significantly regarding the sensitivity and specificity for the diagnosis of terminal ileal inflammation ($P > .999$). DWI and CE MR enterography concurred in the diagnosis of penetrating complications in five of eight segments.

Conclusion: DWI MR enterography was noninferior to CE MR enterography for the evaluation of inflammation in Crohn disease in generally well-distended small bowel, except for the diagnosis of penetration.

전향적
cross-over RCT

MR DWI without 경정맥 조영제가

with 경정맥 조영제에 비해

크론병 small-bowel inflammation evaluation 측면에서

11

The NEW ENGLAND JOURNAL of MEDICINE
ESTABLISHED IN 1812 MAY 10, 2018 VOL. 378 NO. 19

MRI-Targeted or Standard Biopsy for Prostate-Cancer Diagnosis

V. Kasivisvanathan, A.S. Rannikko, M. Borghi, V. Panebianco, L.A. Mynderse, M.H. Vaarala, A. Briganti, L. Budäus, G. Hellawell, R.G. Hindley, M.J. Roobol, S. Eggener, M. Ghei, A. Villers, F. Bladou, G.M. Villeirs, J. Virdi, S. Boxler, G. Robert, P.B. Singh, W. Venderink, B.A. Hadaschik, A. Ruffion, J.C. Hu, D. Margolis, S. Crouzet, L. Klotz, S.S. Taneja, P. Pinto, I. Gill, C. Allen, F. Giganti, A. Freeman, S. Morris, S. Punwani, N.R. Williams, C. Brew-Graves, J. Deeks, Y. Takwoingi, M. Emberton, and C.M. Moore, for the PRECISION Study Group Collaborators*

ABSTRACT

BACKGROUND
Multiparametric magnetic resonance imaging (MRI), with or without targeted biopsy, is an alternative to standard transrectal ultrasonography–guided biopsy for prostate-cancer detection in men with a raised prostate-specific antigen level who have not undergone biopsy. However, comparative evidence is limited.

METHODS
In a multicenter, randomized, noninferiority trial, we assigned men with a clinical suspicion of prostate cancer who had not undergone biopsy previously to undergo MRI, with or without targeted biopsy, or standard transrectal ultrasonography–guided biopsy. Men in the MRI-targeted biopsy group underwent a targeted biopsy (without standard biopsy cores) if the MRI was suggestive of prostate cancer; men whose MRI results were not suggestive of prostate cancer were not offered biopsy. Standard biopsy was a 10-to-12-core, transrectal ultrasonography–guided biopsy. The primary outcome was the proportion of men who received a diagnosis of clinically significant cancer. Secondary outcomes included the proportion of men who received a diagnosis of clinically insignificant cancer.

RESULTS
A total of 500 men underwent randomization. In the MRI-targeted biopsy group, 71 of 252 men (28%) had MRI results that were not suggestive of prostate cancer, so they did not undergo biopsy. Clinically significant cancer was detected in 95 men (38%) in the MRI-targeted biopsy group, as compared with 64 of 248 (26%) in the standard-biopsy group (adjusted difference, 12 percentage points; 95% confidence interval [CI], 4 to 20; P = 0.005). MRI, with or without targeted biopsy, was noninferior to standard biopsy, and the 95% confidence interval indicated the superiority of this strategy over standard biopsy. Fewer men in the MRI-targeted biopsy group than in the standard-biopsy group received a diagnosis of clinically insignificant cancer (adjusted difference, −13 percentage points; 95% CI, −19 to −7; P<0.001).

CONCLUSIONS
The use of risk assessment with MRI before biopsy and MRI-targeted biopsy was superior to standard transrectal ultrasonography–guided biopsy in men at clinical risk for prostate cancer who had not undergone biopsy previously. (Funded by the National Institute for Health Research and the European Association of Urology Research Foundation; PRECISION ClinicalTrials.gov number, NCT02380027.)

The authors' full names, academic degrees, and affiliations are listed in the Appendix. Address reprint requests to Dr. Kasivisvanathan at the Division of Surgery and Interventional Science, UCL, 3rd Fl., Charles Bell House, 43-45 Foley St., London W1W 7TS, United Kingdom, or at veeru.kas@ucl.ac.uk.

*A complete list of members of the PRECISION Study Group is provided in the Supplementary Appendix, available at NEJM.org.

This article was published on March 19, 2018, at NEJM.org.

N Engl J Med 2018;378:1767-77.
DOI: 10.1056/NEJMoa1801993

전향적
RCT

MR-targeted biopsy가

standard biopsy 에 비해

detection 측면에서

12

the rationale

Placebo-controlled trial is unethical: a clinical equipoise.

(1) no standard treatment (usual care, for non-pharmacological) exists
(2) standard treatment is not better than placebo
(3) standard treatment is a placebo (or no treatment)
(4) new evidence has shown uncertainty of risk-benefit profile of the standard treatment
(5) effect treatment is not readily available due to cost or supply issues

Non-inferiority trials are unethical

(1) they disregard patients' interests
(2) no relevant clinical questions
(3) commercial aims, not patients' interests: it is enough to show that they are similar
(4) no limits to the non-inferiority limit
(5) enrolling patients in non-inferiority trials betrays their trust

13

Lancet 2007; 370: 1875–77

the rationale: EMA

demonstrate the efficacy

bioequivalence studies are not possible
the use of a placebo arm is not possible

risk-benefit assessment

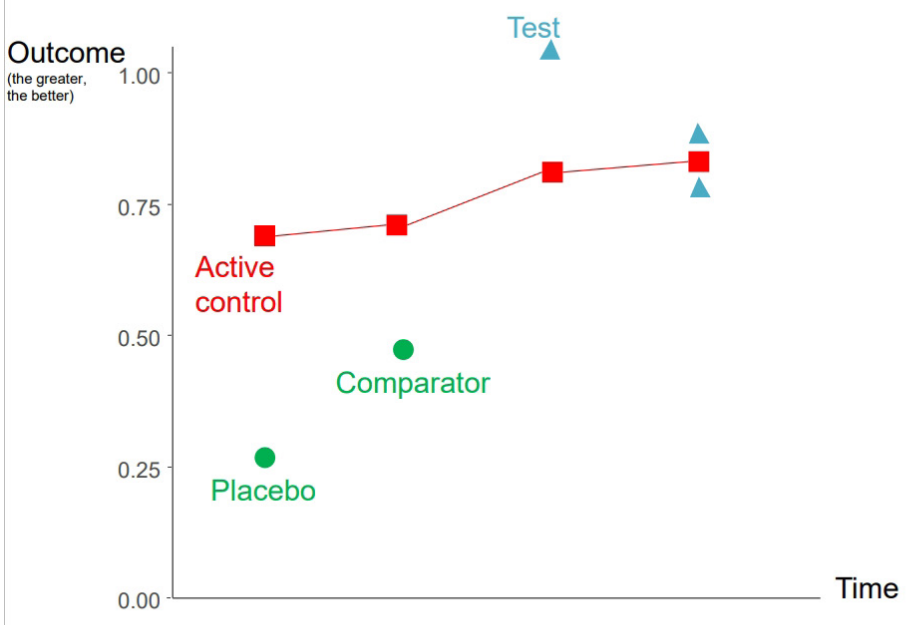
no important loss of efficacy
a direct comparison: risk/benefit
a potential safety advantage, an efficacy comparison

14

the rationale: JAMA

- Available efficacious active treatments can make use of **placebo controls unethical**
- New treatment offers **important advantages** over reference treatments.
  - greater availability
  - reduced cost
  - less invasiveness
  - fewer adverse effects (harms)
  - greater ease of administration

15



16

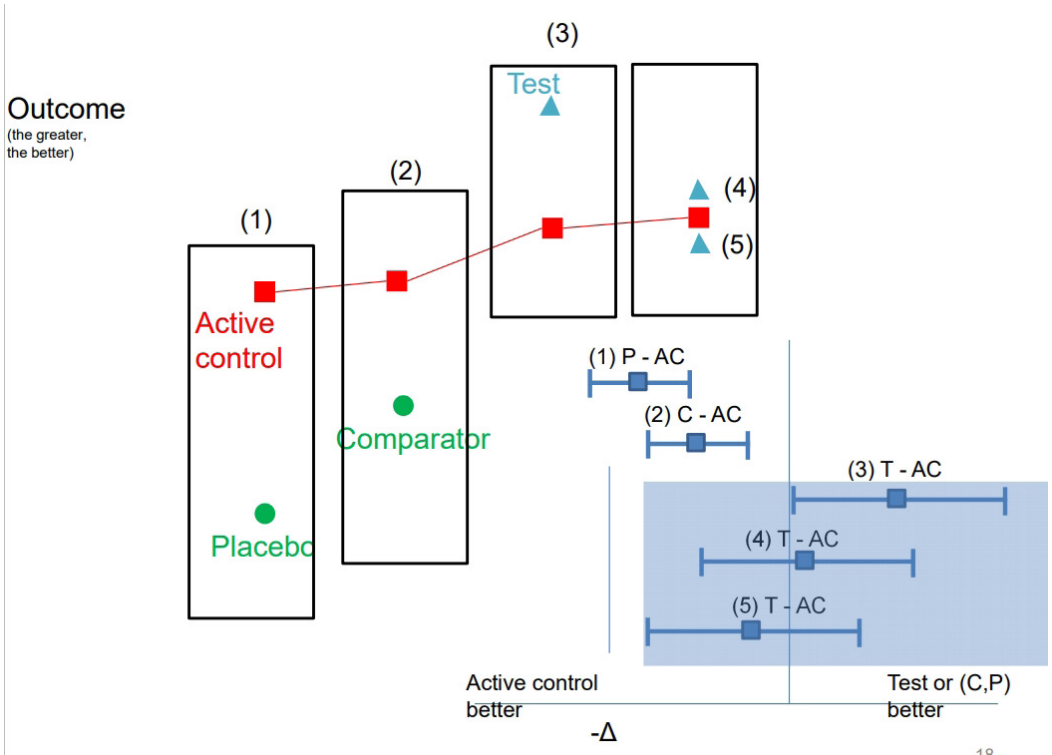
NI in general

- Superior Efficacy: P < AC (P < T; systematic review)

- In general, AC > T
- Additional advantages: AC' < T'
- Risk-benefit: AC ≒ T

- Efficacy : AC < T + Δ
- -Δ < T - AC

17



18

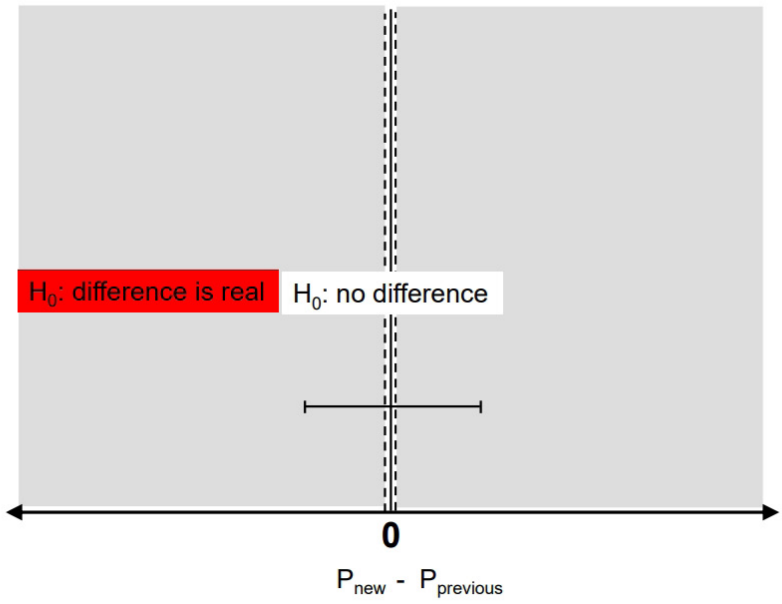
- Introduction: rationale, examples
- Statistical Concept of Equivalence/Noninferiority
  - **hypothesis**
- Noninferiority Margin
  - General Principles
  - Radiologic Perspective
- Examples

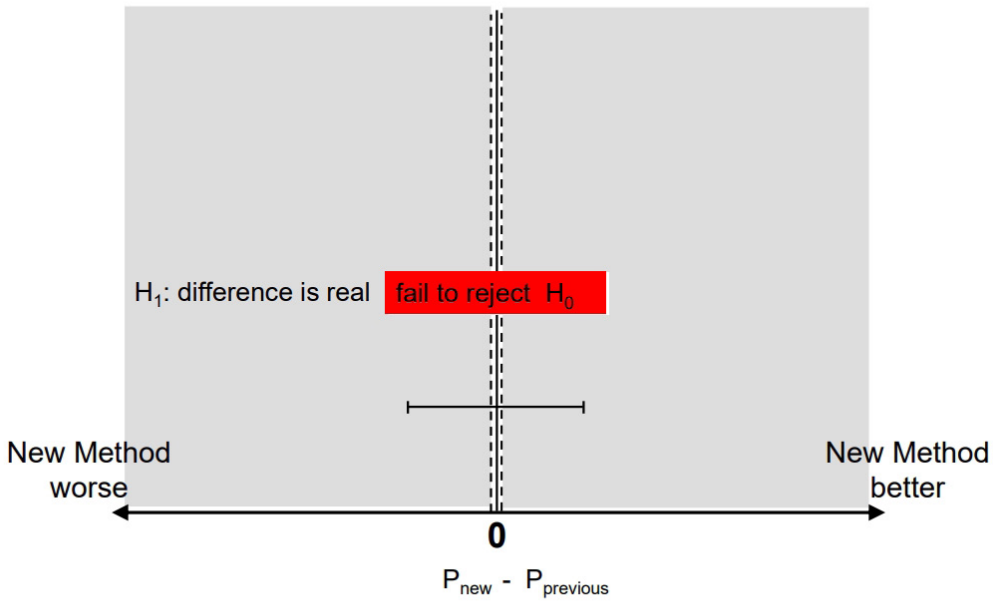
19

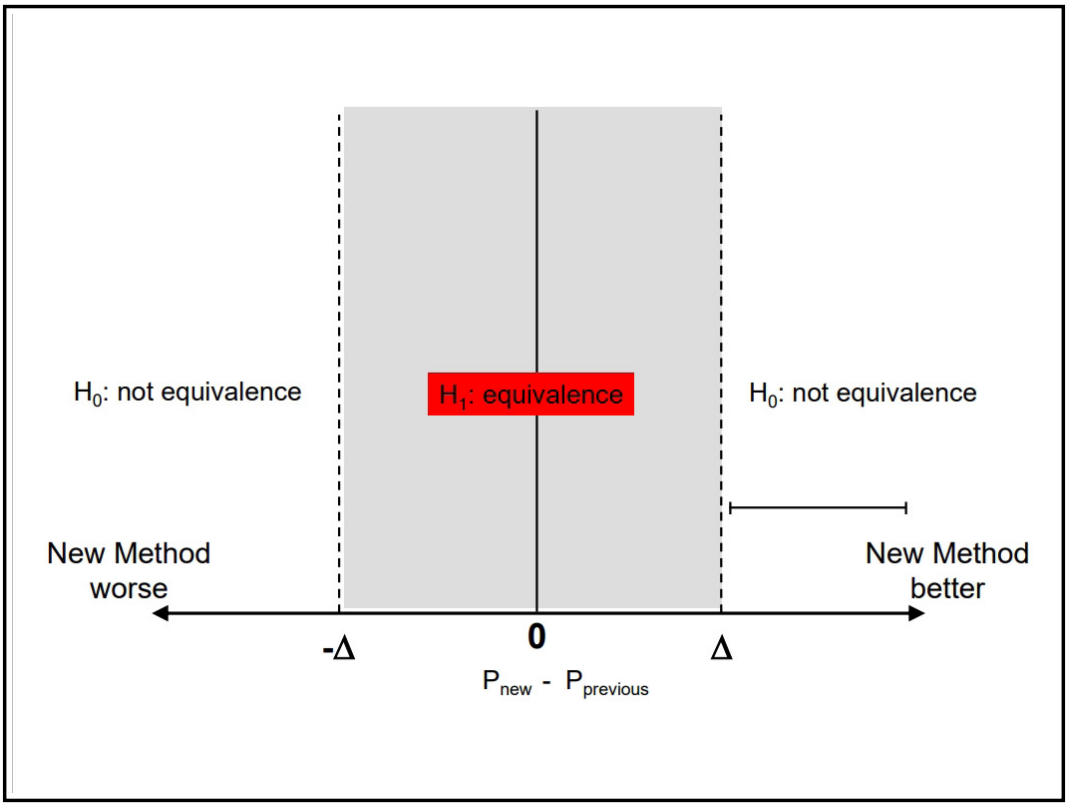
absence of evidence is not evidence of absence
no significant difference≠ the same

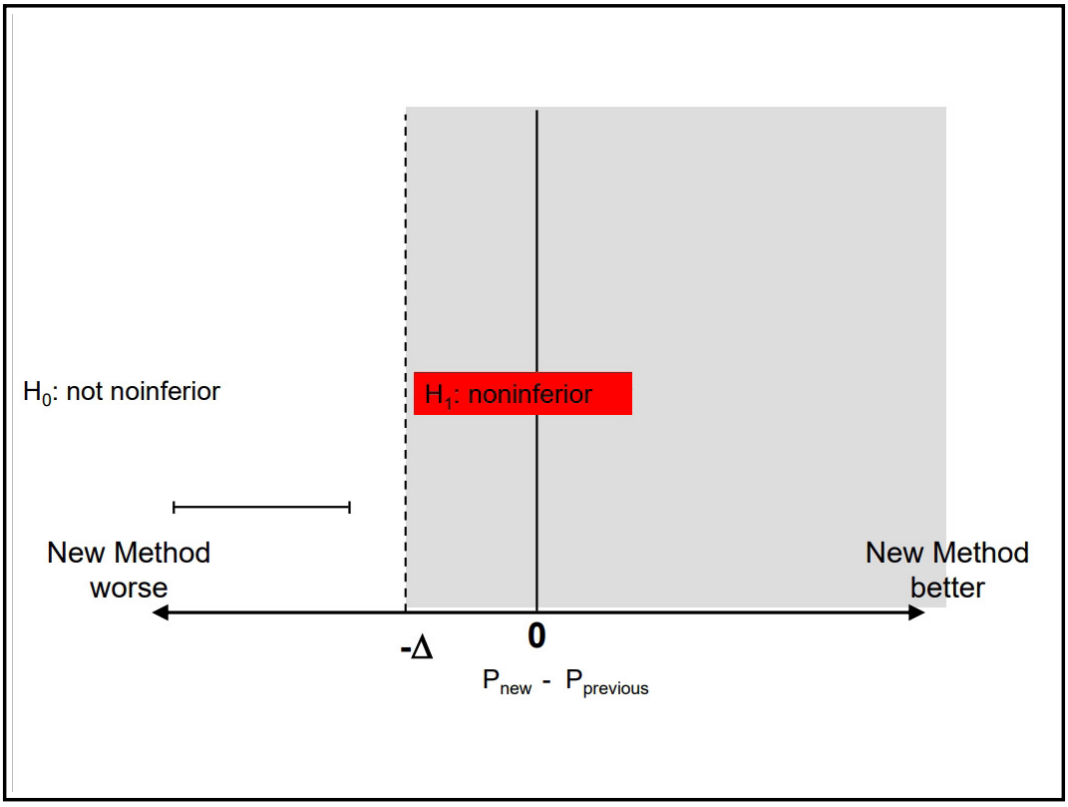
- H_0: no difference
- H_1: the difference is real
  - $P < 0.05$
    - the difference is real
  - $P > 0.05$
    - ~~no difference~~
    - there is insufficient evidence to make a conclusion
    - fail to reject H_0

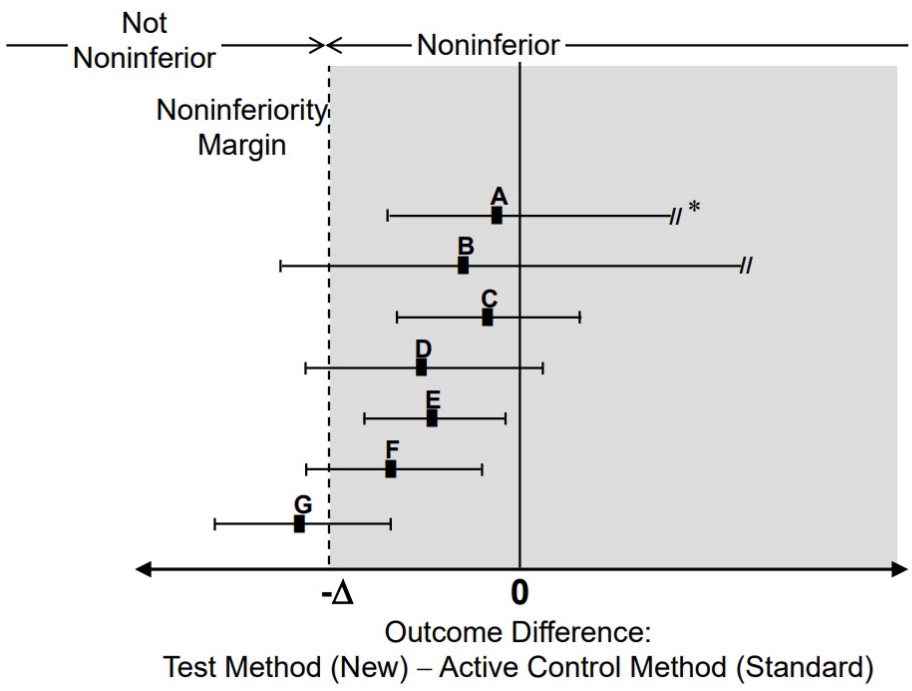
20

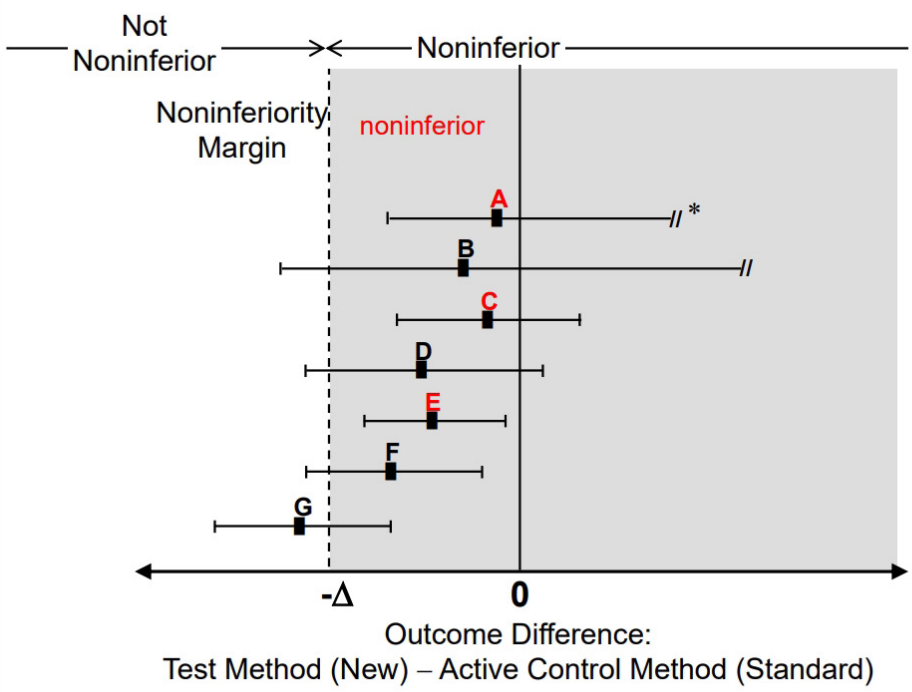

H_0: difference is real
H_0: no difference
0
P_{new} - $P_{previous}$


H_1: difference is real
fail to reject H_0
New Method worse
New Method better
0
P_{new} - $P_{previous}$


H0: not equivalence
H1: equivalence
H0: not equivalence
New Method worse
New Method better
-Δ
0
Δ
Pnew - Pprevious


H0: not noinferior
H1: noninferior
New Method worse
New Method better
-Δ
0
Pnew - Pprevious


Not
Noninferior
Noninferior
Noninferiority
Margin
A
B
C
D
E
F
G
*
-Δ
0
Outcome Difference:
Test Method (New) – Active Control Method (Standard)


Not
Noninferior
Noninferior
Noninferiority
Margin
noninferior
A
B
C
D
E
F
G
*
-Δ
0
Outcome Difference:
Test Method (New) – Active Control Method (Standard)

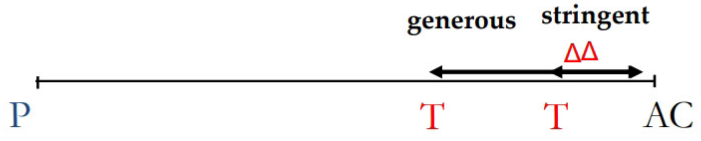
- Introduction: rationale, examples
- Statistical Concept of Equivalence/Noninferiority
  - hypothesis
- **Noninferiority Margin**
  - **General Principles**
  - Radiologic Perspective
- Examples

27

Noninferiority Trial

- Active control (AC): standard test
- Test (T): new test
- Placebo (P): placebo

- New treatment (T) is not worse than Standard treatment (AC) by amount of Δ
- Margin: Generous / **Stringent**
- Outcome: Absolute difference / **Relative difference**



28

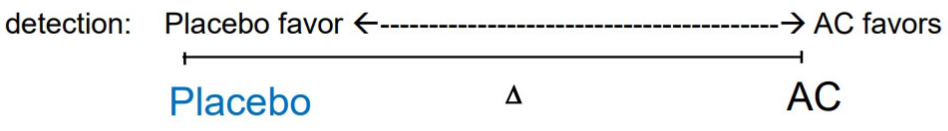
The ABC of non-inferiority margin

- Assay sensitivity
- Bias
- Constancy assumption

29

Noninferiority margin: fixed-margin

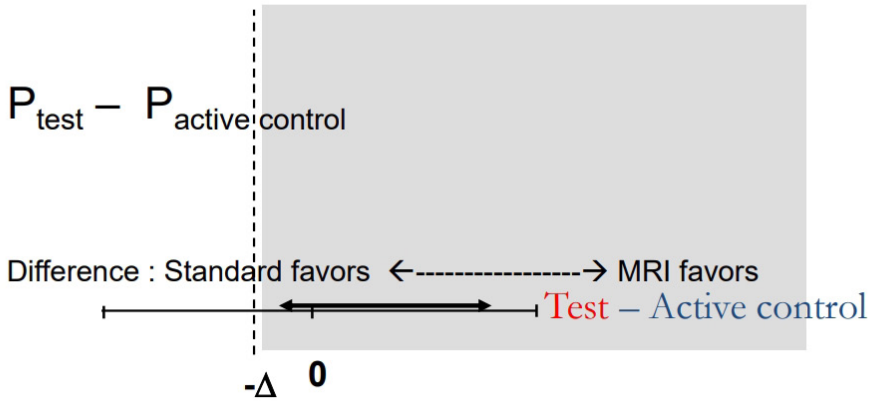
- (1) (AC-P) effect (95% CI)



30

Noninferiority margin: fixed-margin

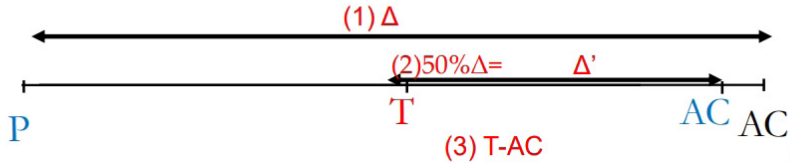
- (2) T-AC effect (95% CI)



31

Noninferiority margin: fixed-margin

- (1) 95% CI of (AC – P)
  - not available
  - (if exists, take a lower limit)
- (2) retention
  - 50%
- (3) 95% CI of (T – AC)



32

	Parallel	Paired
Binary	$N = 4\frac{(Z_{crit} + Z_{pwr})^2 P(1-P)}{\Delta^2}$ $N \approx \frac{42P(1-P)}{\Delta^2}$ 2.5% one-sided type I error 90% power	
Continuous	$N = 4\frac{(Z_{crit} + Z_{pwr})^2 \sigma^2}{\Delta^2}$ $N \approx \frac{42\sigma^2}{\Delta^2}$ 2.5% one-sided type I error 90% power	$N = 4\frac{(Z_{crit} + Z_{pwr})^2 \sigma_d^{\ 2}}{\Delta_d^{\ 2}}$ $N \approx \frac{42\sigma_d^{\ 2}}{\Delta_d^{\ 2}}$ 2.5% one-sided type I error 90% power

33

N : total, Zcrit = 1.96 (one-sided 2.5% = two-sided 5%), Zpwr =1.28 (90% power)

95% CI

	Parallel	Paired
Binary	$p_1 - p_2 \pm 1.96\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$	
Continuous	$m_1 - m_2 \pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	$m_1 - m_2 \pm 1.96\sqrt{\frac{s^2}{n}}$

34

N : total, Zcrit = 1.96 (one-sided 2.5% = two-sided 5%), Zpwr =1.28 (90% power)

- Introduction: rationale, examples
- Statistical Concept of Equivalence/Noninferiority
  - hypothesis
- Noninferiority Margin
  - General Principles
  - Radiologic Perspective
- **Examples**

35

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 MAY 10, 2018 VOL. 378 NO. 19

MRI-Targeted or Standard Biopsy for Prostate-Cancer Diagnosis

V. Kasivisvanathan, A.S. Rannikko, M. Borghi, V. Panebianco, L.A. Mynderse, M.H. Vaarala, A. Briganti, L. Budäus, G. Hellawell, R.G. Hindley, M.J. Roobol, S. Eggener, M. Ghei, A. Villers, F. Bladou, G.M. Villeirs, J. Virdi, S. Boxler, G. Robert, P.B. Singh, W. Venderink, B.A. Hadaschik, A. Ruffion, J.C. Hu, D. Margolis, S. Crouzet, L. Klotz, S.S. Taneja, P. Pinto, I. Gill, C. Allen, F. Giganti, A. Freeman, S. Morris, S. Punwani, N.R. Williams, C. Brew-Graves, J. Deeks, Y. Takwoingi, M. Emberton, and C.M. Moore, for the PRECISION Study Group Collaborators*

ABSTRACT

BACKGROUND

Multiparametric magnetic resonance imaging (MRI), with or without targeted biopsy, is an alternative to standard transrectal ultrasonography–guided biopsy for prostate-cancer detection in men with a raised prostate-specific antigen level who have not undergone biopsy. However, comparative evidence is limited.

METHODS

In a multicenter, randomized, noninferiority trial, we assigned men with a clinical suspicion of prostate cancer who had not undergone biopsy previously to undergo MRI, with or without targeted biopsy, or standard transrectal ultrasonography–guided biopsy. Men in the MRI-targeted biopsy group underwent a targeted biopsy (without standard biopsy cores) if the MRI was suggestive of prostate cancer; men whose MRI results were not suggestive of prostate cancer were not offered biopsy. Standard biopsy was a 10-to-12-core, transrectal ultrasonography–guided biopsy. The primary outcome was the proportion of men who received a diagnosis of clinically significant cancer. Secondary outcomes included the proportion of men who received a diagnosis of clinically insignificant cancer.

RESULTS

A total of 500 men underwent randomization. In the MRI-targeted biopsy group, 71 of 252 men (28%) had MRI results that were not suggestive of prostate cancer, so they did not undergo biopsy. Clinically significant cancer was detected in 95 men (38%) in the MRI-targeted biopsy group, as compared with 64 of 248 (26%) in the standard-biopsy group (adjusted difference, 12 percentage points; 95% confidence interval [CI], 4 to 20; P=0.005). MRI, with or without targeted biopsy, was noninferior to standard biopsy, and the 95% confidence interval indicated the superiority of this strategy over standard biopsy. Fewer men in the MRI-targeted biopsy group than in the standard-biopsy group received a diagnosis of clinically insignificant cancer (adjusted difference, −13 percentage points; 95% CI, −19 to −7; P<0.001).

CONCLUSIONS

The use of risk assessment with MRI before biopsy and MRI-targeted biopsy was superior to standard transrectal ultrasonography–guided biopsy in men at clinical risk for prostate cancer who had not undergone biopsy previously. (Funded by the National Institute for Health Research and the European Association of Urology Research Foundation; PRECISION ClinicalTrials.gov number, NCT02380027.)

The authors' full names, academic degrees, and affiliations are listed in the Appendix. Address reprint requests to Dr. Kasivisvanathan at the Division of Surgery and Interventional Science, UCL, 3rd Fl., Charles Bell House, 43-45 Foley St., London W1W 7TS, United Kingdom, or at veeru.kasi@ucl.ac.uk.

*A complete list of members of the PRECISION Study Group is provided in the Supplementary Appendix, available at NEJM.org.

This article was published on March 19, 2018, at NEJM.org.

N Engl J Med 2018;378:1767-77.
DOI: 10.1056/NEJMoa1801993

36

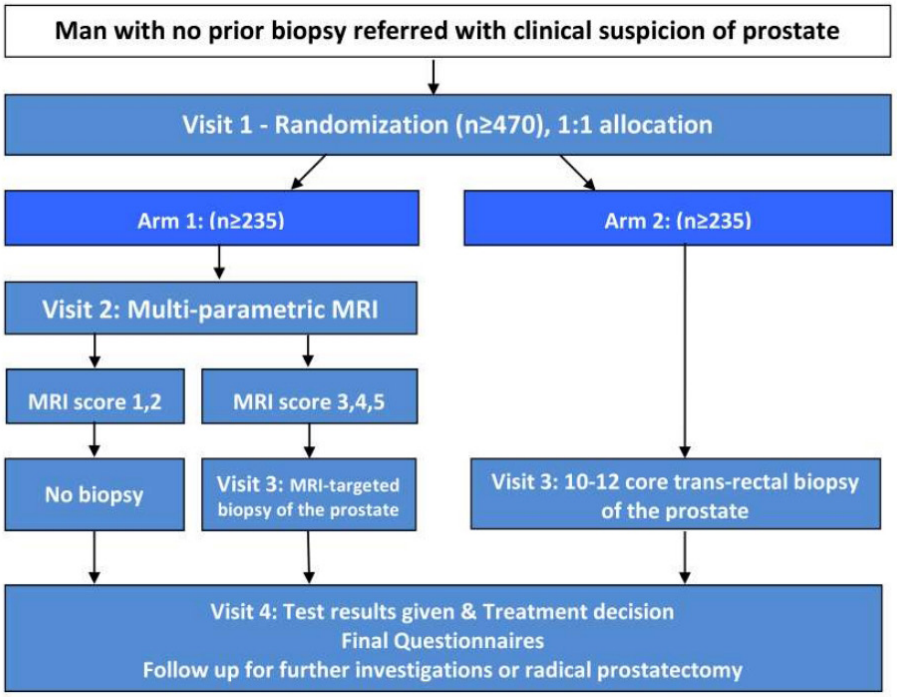


Table 1. Features of Noninferiority Studies.

Consideration	Explanation	Challenges
Active control	Select active control on the basis of a previous randomized superiority trial comparing active control with placebo; active control represents current standard of care	Placebo-controlled trials may not have been performed
End-point selection	Is the end point clinically relevant, and are there historical data comparing the active control with placebo for the selected end point?	Composite end points may be difficult to interpret; the relevance of end points may change in the course of follow-up
Choice of noninferiority margin	Is the margin less than the treatment effect of the active control versus placebo? Is there consensus about the margin of reduced effectiveness that is still acceptable in light of potential benefits (e.g., improved safety, lower cost, lower risk of side effects)?	It is important not to accept new therapies that are less effective over time than previous therapies (known as "biocreep"*); historical data are not always available to determine the difference between placebo and control (e.g., in the case of antiinfective agents)
Assay sensitivity	If the active control were compared with placebo, would superiority be evident?	A "positive control" usually cannot be assessed in the study, since placebo is not feasible or ethical
Constancy and metrics	Have the conditions changed between the trial establishing superiority of the active control over placebo and the noninferiority trial? What type of metric (between-group difference in absolute risk or relative risk) is more likely to be constant between studies and therefore a reliable metric for comparison and margin definition?	Characteristics of the study population or concomitant therapies may have changed since the effect of active therapy was established, making a determination of noninferiority unreliable; constancy is not always present for absolute effects; a lower-than-expected event rate may make a risk-difference margin clinically inappropriate if viewed from a relative-risk perspective; a higher-than-expected event rate may result in lower-than-expected power
Execution	Are the assigned treatments administered adequately? Is ascertainment of the end point accurate and complete?	Lack of attention to execution in the control group or misclassification or missing data on the end point may bias the study toward a conclusion of noninferiority
Analysis	If treatment crossover or nonadherence occurs, what is the appropriate analysis (intention-to-treat or per-protocol)?	Treatment crossover may bias an intention-to-treat analysis toward a conclusion of noninferiority, but a per-protocol analysis may also introduce bias, since baseline characteristics are no longer balanced between study groups

* Biocreep was defined in a 1992 "Points to Consider" Food and Drug Administration briefing document.[4]

Mauri L, D'Agostino RB. Challenges in the Design and Interpretation of Noninferiority Trials. N Engl J Med. 2017

Active control

- **(Test) MRI-Targeted Biopsy**
- **(Active control) Standard Biopsy**

- Select active control on the basis of a previous randomized superiority trial comparing active control with placebo; active control represents current standard of care

- Challenges: Placebo-controlled trials may not have been performed

39

End-point selection

- **clinically significant prostate cancer rate (Gleason grade 3+4 disease or greater)**

- Is the end point clinically relevant, and are there historical data comparing the active control with placebo for the selected end point?

- Challenges: Composite end points may be difficult to interpret; the relevance of end points may change in the course of follow-up

40

Choice of noninferiority margin

- **The choice of 5% as the margin of non-inferiority represents a difference that would be considered clinically unimportant in the detection rates.**
- **+ 10 % difference**
- Is the margin less than the treatment effect of the active control versus placebo? Is there consensus about the margin of reduced effectiveness that is still acceptable in light of potential benefits (e.g., improved safety, lower cost, lower risk of side effects)?
- It is important not to accept new therapies that are less effective over time than previous therapies (known as "biocreep"*); historical data are not always available to determine the difference between placebo and control (e.g., in the case of antiinfective agents)

41

- For the non-inferiority hypothesis, using 90% power and 2.5% one-sided α, using an estimate for detection rate of clinically significant cancer for targeted biopsy of 40% and an estimate of detection rate for TRUS biopsy of 30% and using a margin of clinical unimportance of 5%, 211 men per arm will be required.

42

Assay sensitivity

- ?
- If the active control were compared with placebo, would superiority be evident?
- A "positive control" usually cannot be assessed in the study, since placebo is not feasible or ethical

43

Constancy and metrics

Outcome	MRI-Targeted Biopsy Group (N=252)	Standard-Biopsy Group (N=248)	Difference†	P Value
Clinically significant cancer¶				
Intention-to-treat analysis — no. (%)	95 (38)	64 (26)	12 (4 to 20)	0.005

- Rates of clinically significant cancer detection from targeted-alone biopsy in a population with no prior biopsy have been shown to be 50%.
- Assuming 20% of men avoid biopsy in the MRI arm of PRECISION, this would correspond to a 50% detection rate in 80% of the participants in this arm =40% overall detection rate of clinically significant cancer in the MRI arm.
- Rates of clinically significant cancer detection from one of the largest studies of TRUS biopsy in men without prior biopsy are shown to be 27%.
- Have the conditions changed between the trial establishing superiority of the active control over placebo and the noninferiority trial? What type of metric (between-group difference in absolute risk or relative risk) is more likely to be constant between studies and therefore a reliable metric for comparison and margin definition?
- Characteristics of the study population or concomitant therapies may have changed since the effect of active therapy was established, making a determination of noninferiority unreliable; constancy is not always present for absolute effects; a lower-than-expected event rate may make a risk-difference margin clinically inappropriate if viewed from a relative-risk perspective; a higher-than-expected event rate may result in lower- than-expected power

44

Execution

- Of the 71 men with negative results on MRI and no biopsy, 3 (4%) were discharged, 62 (87%) were referred for monitoring of the PSA level, 3 (4%) underwent further prostate biopsy (all had negative results), 1 (1%) underwent an additional multiparametric MRI, and 2 (3%) had missing information.
- Are the assigned treatments administered adequately? Is ascertainment of the end point accurate and complete?
- Lack of attention to execution in the control group or misclassification or missing data on the end point may bias the study toward a conclusion of noninferiority

45

Analysis

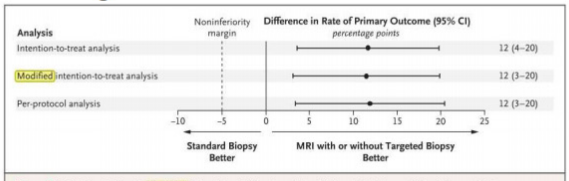


Figure 2. Intention-to-Treat, Modified Intention-to-Treat, and Per-Protocol Analyses of the Primary Outcome for the Detection of Clinically Significant Prostate Cancer.
Shown are the absolute differences between the MRI-targeted biopsy group and the standard-biopsy group in the rates of detection of clinically significant cancer. The intention-to-treat analysis included all the participants who underwent randomization, the modified intention-to-treat analysis excluded participants who did not complete a diagnostic test strategy, and the per-protocol analysis included only participants who underwent the randomly assigned testing procedure as specified in the protocol. If the lower boundary of the two-sided 95% confidence interval for the difference (MRI-targeted biopsy group minus standard-biopsy group) was greater than −5 percentage points (dashed line), then MRI, with or without targeted biopsy, would be deemed to be noninferior. If the lower boundary was greater than zero (solid line), superiority would be claimed.

- If treatment crossover or nonadherence occurs, what is the appropriate analysis (intention-to-treat or per-protocol)?
- Treatment crossover may bias an intention-to-treat analysis toward a conclusion of noninferiority, but a per- protocol analysis may also introduce bias, since baseline characteristics are no longer balanced between study groups

46

Sample size calculation and two-sided 95% CI

$$N = 4\frac{(Z_{crit} + Z_{pwr})^2 P(1-P)}{\Delta^2}$$

targeted biopsy of 40% and an estimate of detection rate for TRUS biopsy of 30% and using a margin of clinical unimportance of 5%

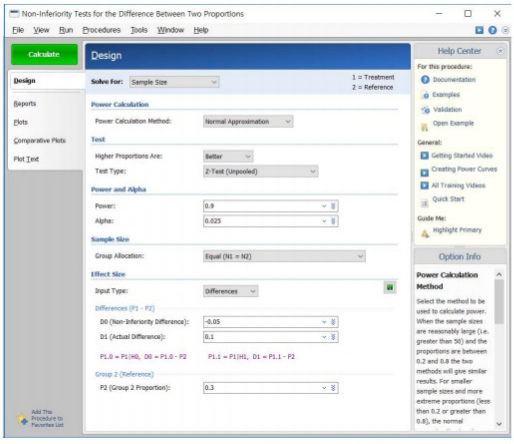
```
> 4*(1.96+1.282)^2*0.3*(1-0.3)/(0.05+0.1)^2
[1] 393
```

$$p_1 - p_2 \pm 1.96\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Outcome	MRI-Targeted Biopsy Group (N=252)	Standard-Biopsy Group (N=248)	Difference	P Value
Clinically significant cancer				
Intention-to-treat analysis — no. (%)	95 (38)	64 (26)	12 (4 to 20)	0.005

```
> 0.38-0.26 - 1.96*sqrt(0.38*(1-0.38)/252 + 0.26*(1-0.26)/248)
[1] 0.04
> 0.38-0.26 + 1.96*sqrt(0.38*(1-0.38)/252 + 0.26*(1-0.26)/248)
[1] 0.20
```

47



Non-Inferiority Tests for the Difference Between Two Proportions

Numeric Results for Non-Inferiority Tests for the Difference Between Two Proportions
Test Statistic: Z-Test with Unpooled Variance
H0: P1 - P2 ≤ D0 vs. H1: P1 - P2 = D1 > D0.

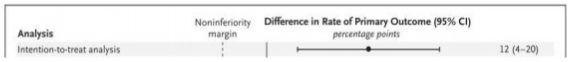
Target Power	Actual Power*	N1	N2	N	Ref. P2	P1|H0 P1.0	P1|H1 P1.1	NI Diff D0	Diff D1	Alpha
0.90	0.90115	211	211	422	0.3000	0.2500	0.4000	-0.0500	0.1000	0.025

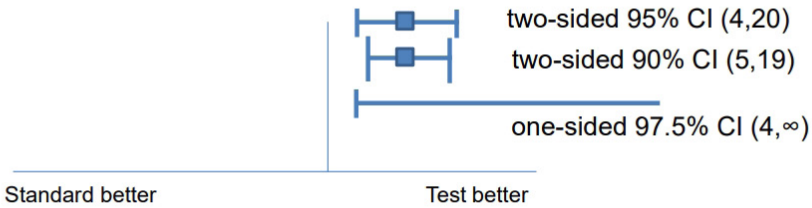
* Power was computed using the normal approximation method.

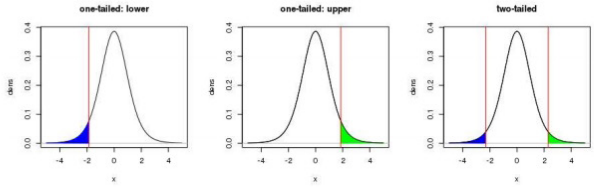
48

Q1. two-sided 95% and one-sided 97.5% CI

양측 95% 신뢰구간 = 단측 97.5% 신뢰구간







49

Q2. retrospective, prospective / paired, parallel

전향적, 후향적 / paired, 평행

- Paired design
  - McNemar's test
  - Generalized Estimating Equation
  - Bootstrapping
- Parallel design
  - independent two-sample test

50

Q3. noninferiority and superiority?
비열등성과 우위성 동시 설계

- In some cases, a study planned as an NI study may show superiority to the active control. Recommendations in International Conference on Harmonisation guidance E9: Statistical Principles for Clinical Trials (ICH E9) and FDA policy have been that this superiority finding arising in an NI study can be interpreted without adjustment for multiplicity.
- (but pre-planned)

51

Conclusions

- Introduction: rationale, examples
  - Active control, standard treatment/modality(AC)
  - Test, new treatment/modality (T)
- Statistical Concept of Equivalence/Noninferiority
  - hypothesis: margin
- Noninferiority Margin
  - General Principles: 95-95 fixed margin
  - Radiologist's Perspective: previous experiences
- Study design
  - General Principle: prospective, parallel
  - Radiologist's Perspective: retrospective/prospective, paired
- **Endpoint**
  - General Principle: Clinically relevant
  - Radiologist's Perspective: Evaluation of diagnostic performance

52