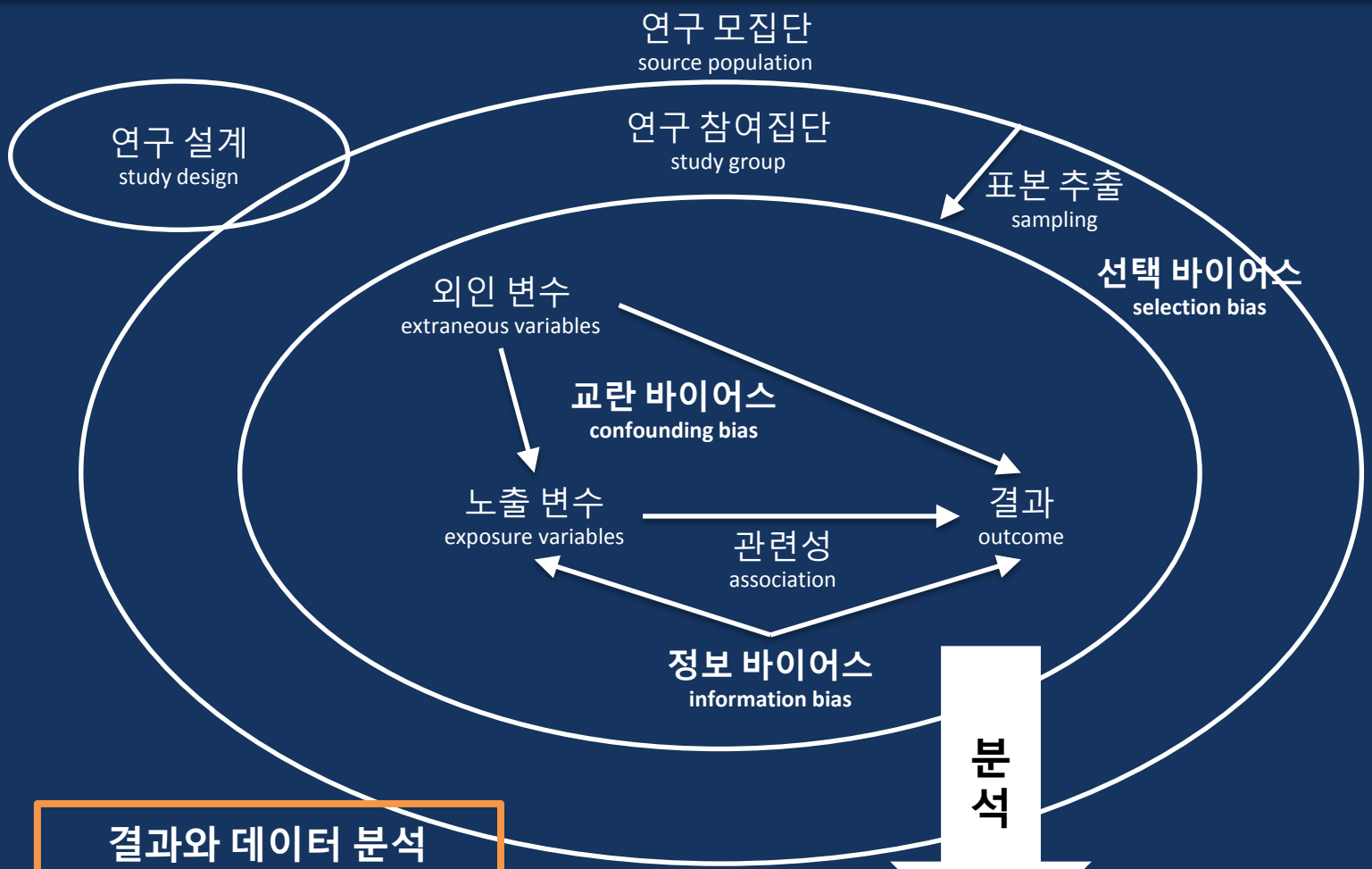
춘계 심혈관 통합학술대회

Assessment of Intra- and Inter-Observer Variability

Jun-Bean Park

Seoul National University Hospital

역학 연구의 핵심 요소



Dohoo I, Martin W, Stryhn H. (2012) Methods in Epidemiologic Research, VER Inc.

Accuracy of Echocardiography Versus Electrocardiography in

Journal of the American College of Cardiology
© 1999 by the American College of Cardiology
Published by Elsevier Science Inc.

Vol. 34, No. 5, 1999
ISSN 0735-1097/99/$20.00
PII S0735-1097(99)00396-4

Echocardiography

Reliability of Echocardiographic Assessment of Left Ventricular Structure and Function

Journal of the American College of Cardiology
© 2004 by the American College of Cardiology Foundation
Published by Elsevier Inc.

Vol. 44, No. 4, 2004
ISSN 0735-1097/04/$30.00
doi:10.1016/j.jacc.2004.05.050

Echocardiography

Reproducibility and Accuracy of Echocardiographic Measurements of Left Ventricular Parameters Using Real-Time Three-Dimensional Echocardiography

Carly Jenkins, BS, Kristen Bricknell, BS, Lizelle Hanekom, MD, Thomas H. Marwick, MD, PhD, FACC
Brisbane, Australia

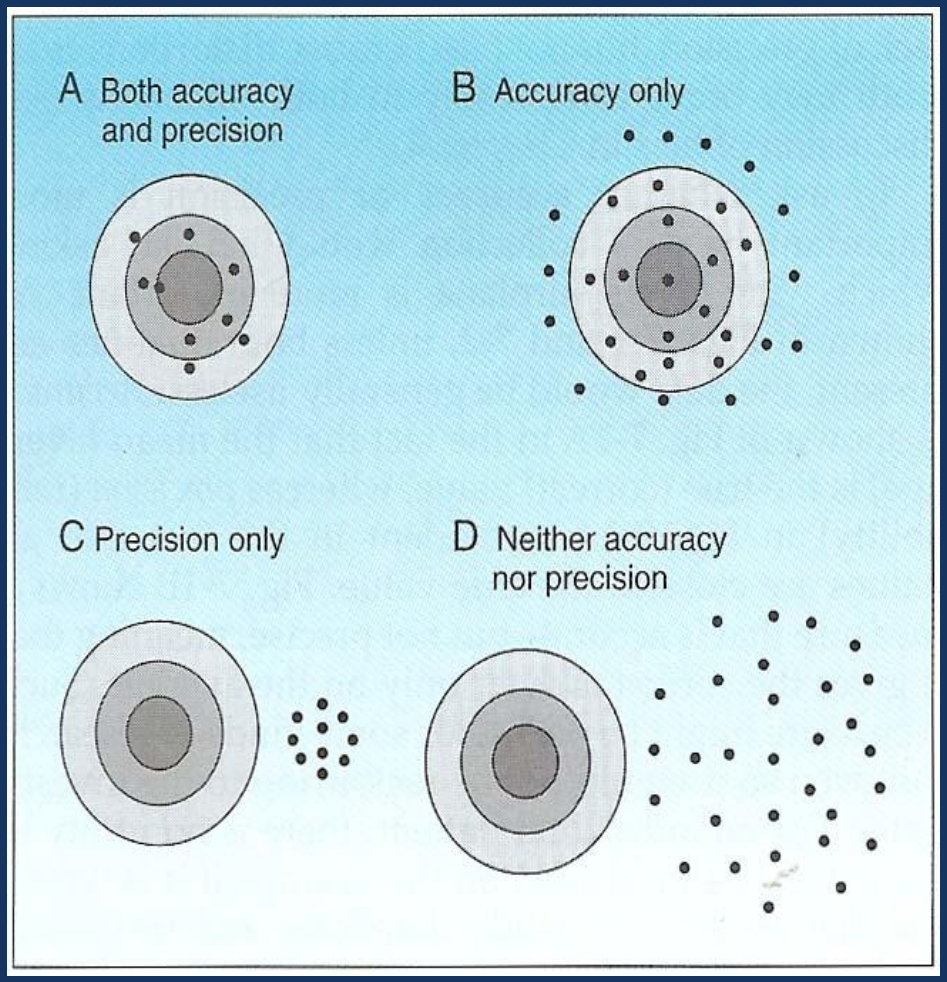
검사법의 타당도와 신뢰도

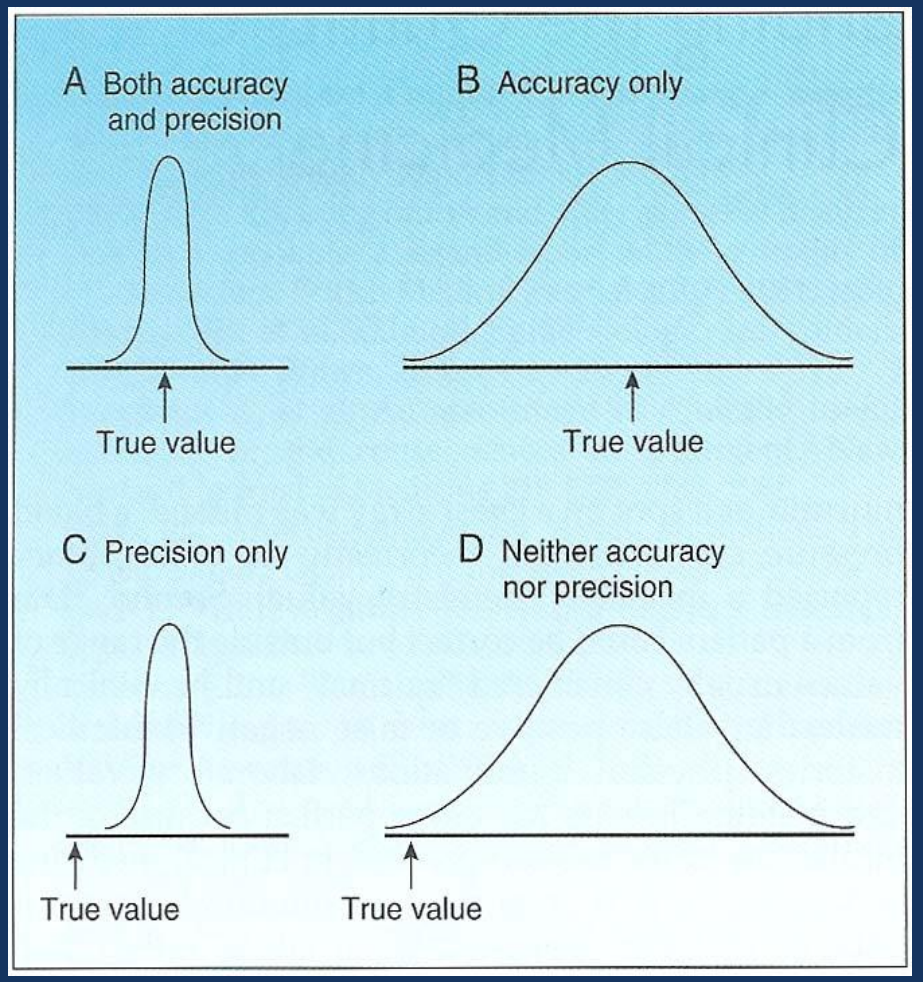
정의

- **타당도 (validity)**
  - 검사법이 진단하고자 하는 질병의 유무를 얼마나 정확하게 판정하는가에 대한 능력
  - 정확도 (accuracy)

- **신뢰도 (reliability)**
  - 측정 조건에 따라 검사 결과가 얼마나 일관되게 나타나는지에 대한 능력
  - 타당도의 전제 조건, 검사자 간 변이 중요
  - 정밀도 (precision), 재현성 (reproducibility), 반복성 (repeatability)
  - Concordance = test–retest reliability

타당도 vs. 신뢰도


A Both accuracy and precision
B Accuracy only
C Precision only
D Neither accuracy nor precision


A Both accuracy and precision
B Accuracy only
True value
True value
C Precision only
D Neither accuracy nor precision
True value
True value

검사법의 타당도 기준

검사 결과	질병 상태		전체
	있음	없음	
양성	a	b	a+b
음성	c	d	c+d
전체	a+c	b+d	a+b+c+d

- **민감도 (감수성, sensitivity) = a / (a + c)**
- **특이도 (특이성, specificity) = d / (b + d)**
- 위양성 (거짓양성, false-positive rate) = b / (b + d)
- 위음성 (거짓음성, false-negative rate) = c / (a + c)
- 유병률 (prevalence) = (a + c) / (a + b + c + d)
- **양성 예측도 (positive predictive value) = a / (a + b)**
- **음성 예측도(negative predictive value) = d / (c + d)**
- 양성 가능도 비 (likelihood ratio positive, LR+) = [a / (a + c)] / [b / (b + d)]
- 음성 가능도 비 (likelihood ratio negative, LR-) = [c / (a + c)] / [d / (b + d)]

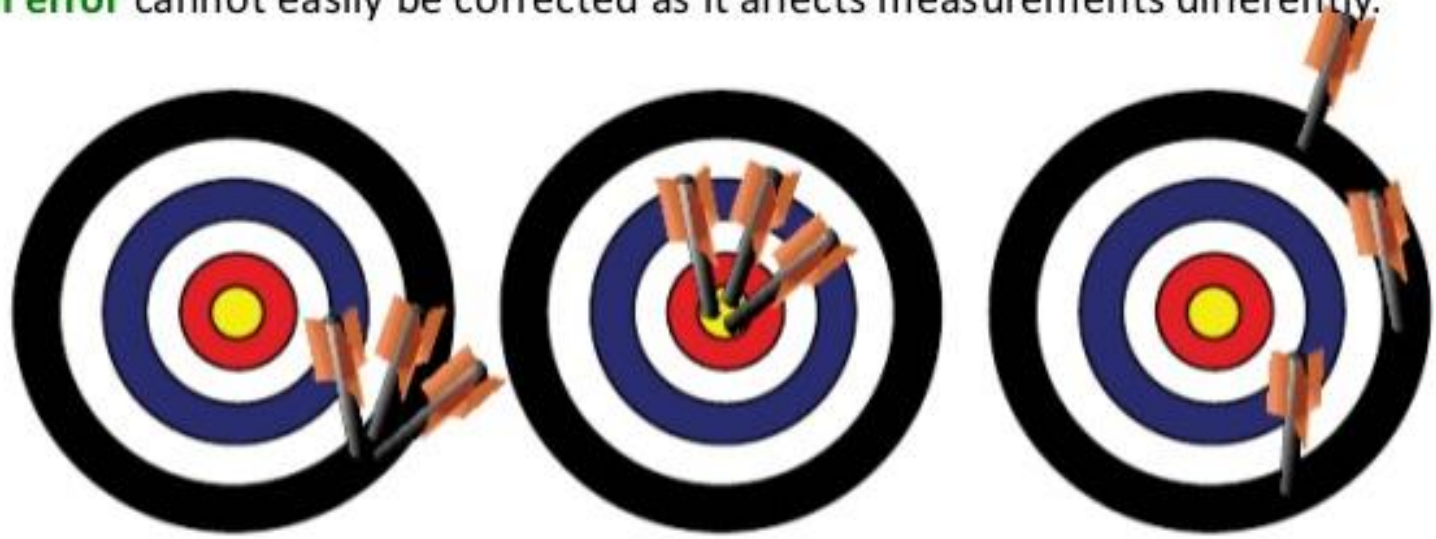
신뢰도 (정밀도, 재현성, 반복성)

- Reliability, precision, reproducibility, repeatability

- 한 실험자가 반복 검사 또는 여러 실험자가 동일한 검사 수행 시 얼마나 일치하는가?
  - **관찰자 내 변이 (intra-observer variation)**와 **관찰자 간 변이 (inter-observer variation)**
  - 무작위 오차 (random error)가 높으면 신뢰도 낮음.
  - 계통 오차 (systematic error)만 있는 경우 신뢰도 높을 수 있음.

Systematic Error vs Random Error

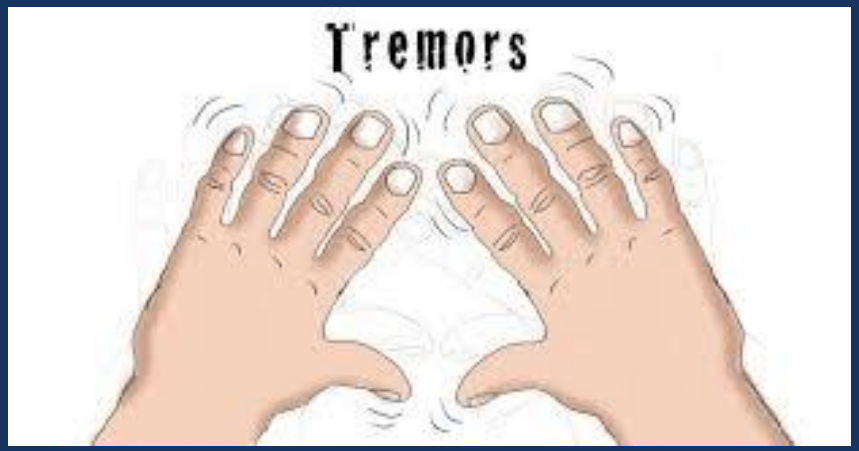
- **Systematic errors** are repeated in the same way throughout an investigation, such as using a balance incorrectly in the same way for each measurement. This can be corrected. Precision describes how repeatable they are.
- **Random error** cannot easily be corrected as it affects measurements differently.

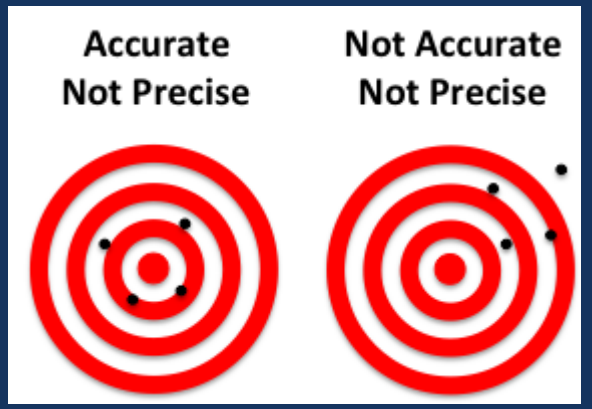


	Results A	Results B	Results C
Systematic error		None	No
Random Error	No	None	


Accurate
Precise


Not Accurate
Precise

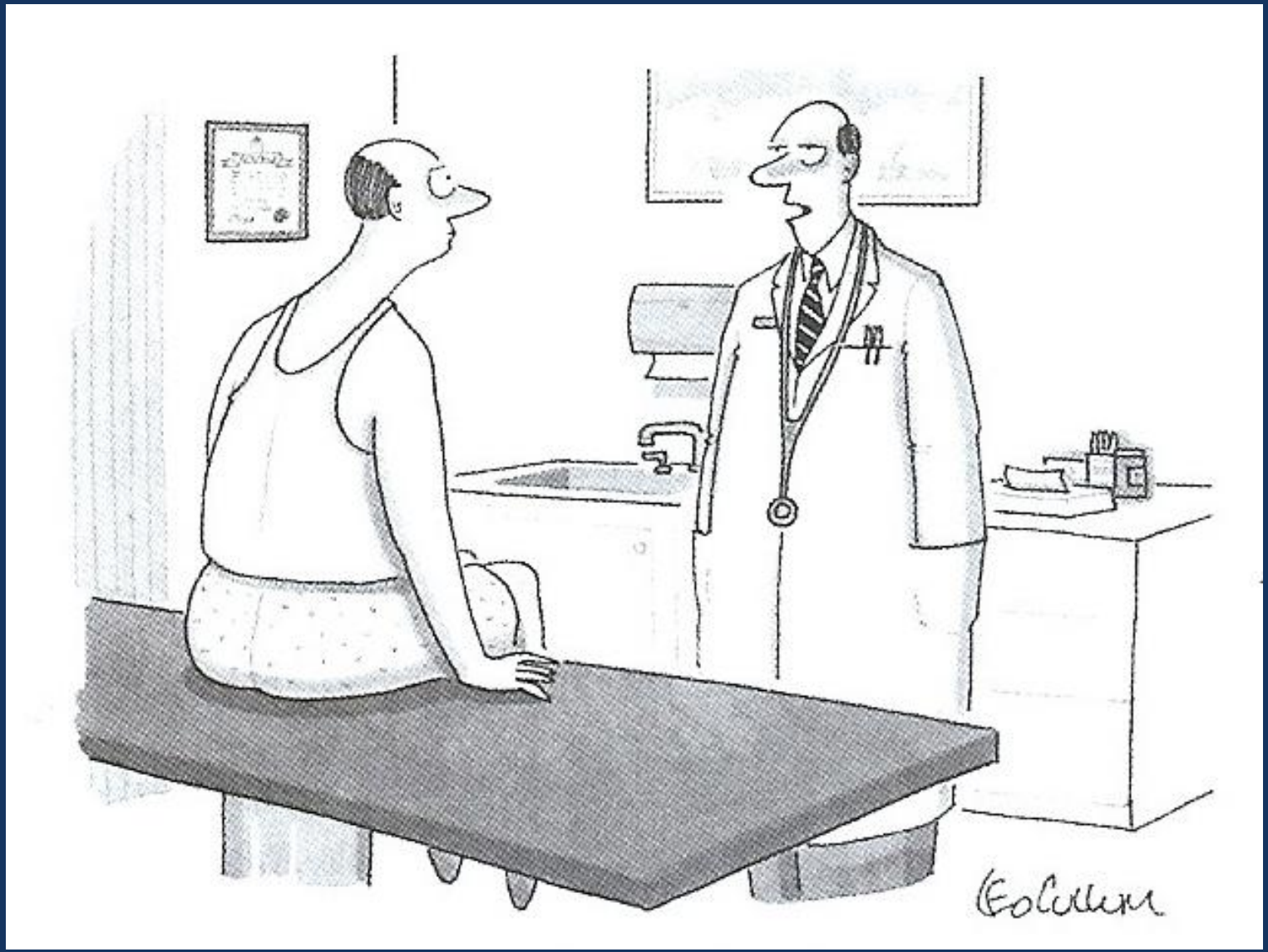

Tremors


Accurate
Not Precise
Not Accurate
Not Precise

신뢰도에 영향을 미치는 변이

1. **관찰자 내 변이 (intra-observer variation)**
   - 같은 임상의가 같은 사람에 대해 혈압이나 신장을 연속적으로 측정하거나,
   - 같은 엑스선 사진을 몇 번에 걸쳐서 판독할 때 생기는 측정과 해석의 차이.

2. **관찰자 간 변이 (inter-observer variation)**
   - 두 명의 다른 임상의가 같은 사람의 혈압을 측정하거나
   - 같은 엑스선 사진을 각각 판독할 때 생기는 차이



“지금 소견은 좀 다른데요. 처음에는 환자분이 다른 질병이 있다고 생각했지만...”

신뢰도 측정 방법

- 순위 척도의 경우
  - **일치율 (agreement percent)**
  - **카파 통계량 (kappa statistics, kappa value)**
  - 스피어맨 순위 상관계수 (Spearman's rank correlation coefficient)

- 연속 척도의 경우
  - **급 내 상관계수 (intra-class correlation coefficient, ICC)**
  - **블랜드-앨트먼 도표 (Bland-Altman plot)**

*일치도 측정 시 피어슨 (Pearson's) 상관계수는 권장하지 않음.

	관찰자 1		
관찰자 2	양성	음성	전체
양성	30 (a)	7 (b)	37
음성	3 (c)	60 (d)	63
전체	33	67	100

1. 관찰 일치도 = 30 + 60 = 90
2. 최대 가능 일치도 = 30 + 7 + 3 + 60 = 100
3. **일치율 = (30 + 60) / (30 + 7 + 3 + 60) = 90 / 100 = 90%**
4. 칸 a 기대 일치도 = [(30 + 7)(30 + 3)] / 100 = [(37)(33)] / 100 = 12.2
5. 칸 d 기대 일치도 = [(3 + 60)(7 + 60)] / 100 = [(63)(67)] / 100 = 42.2
6. 전체 기대 일치도=12.2+42.2=54.4
7. **카파 통계량**

=[(관찰된 일치율) - (우연에 의해 기대된 일치율)] /
[100% - (우연에 의해 기대된 일치율)]
=(90 - 54.4) / (100 - 54.4) = 0.78

Kappa statistics

Value of κ	Strength of agreement
<0.20	Poor
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	**Good**
0.81-1.00	Very good

Altman DG. (1991) Practical Statistics for Medical Research, p 404.

Intra-class correlation coefficient (ICC)

- ICC = $\dfrac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2}$

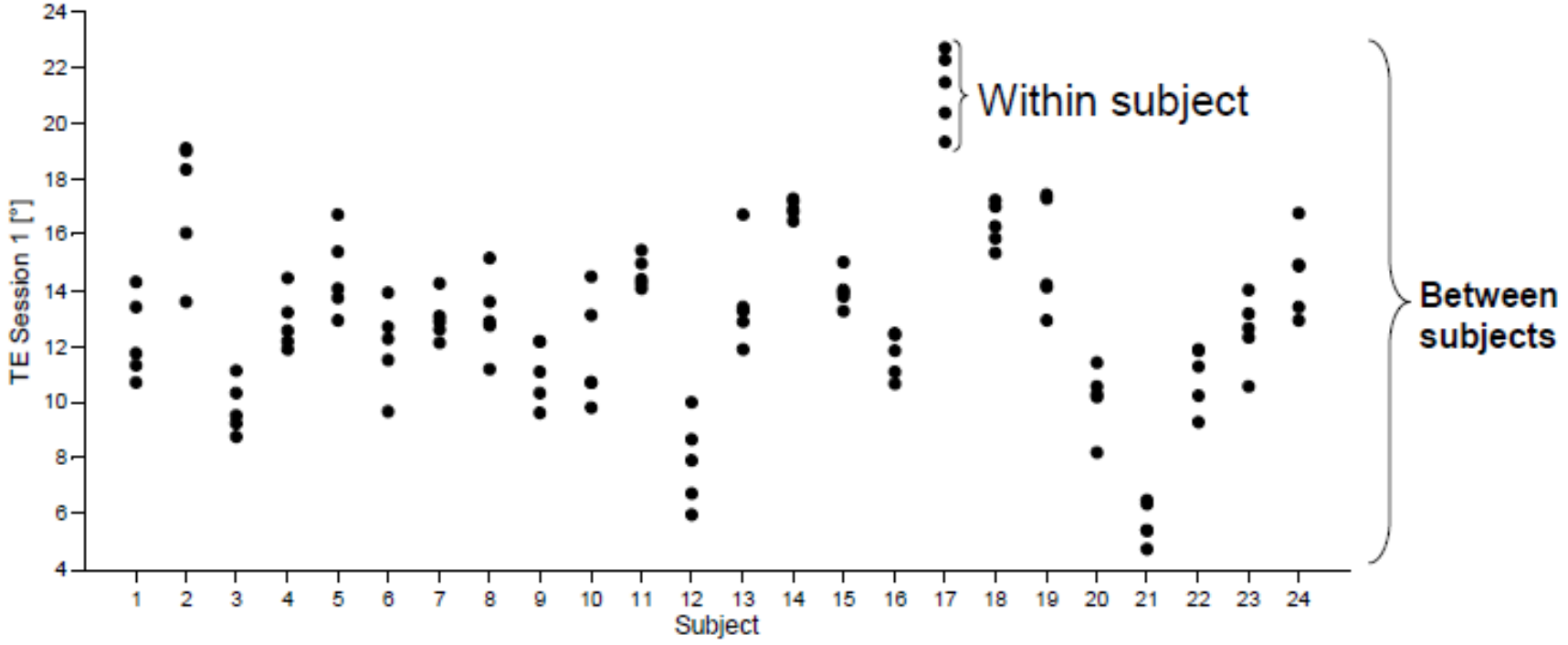
  σ_u^2 = Between-subject variance

  σ_e^2 = Within-subject (measurement error) variance

- 결과 해석은 kappa value와 동일

Between vs. within subject variance

Intraclass Correlation Coefficient



Bland-Altman plot

STATISTICAL METHODS FOR ASSESSING AGREEMENT BETWEEN TWO METHODS OF CLINICAL MEASUREMENT

J. Martin Bland, Douglas G. Altman

Department of Clinical Epidemiology and Social Medicine, St. George's Hospital Medical School, London SW17 ORE; and Division of Medical Statistics, MRC Clinical Research Centre, Northwick Park Hospital, Harrow, Middlesex

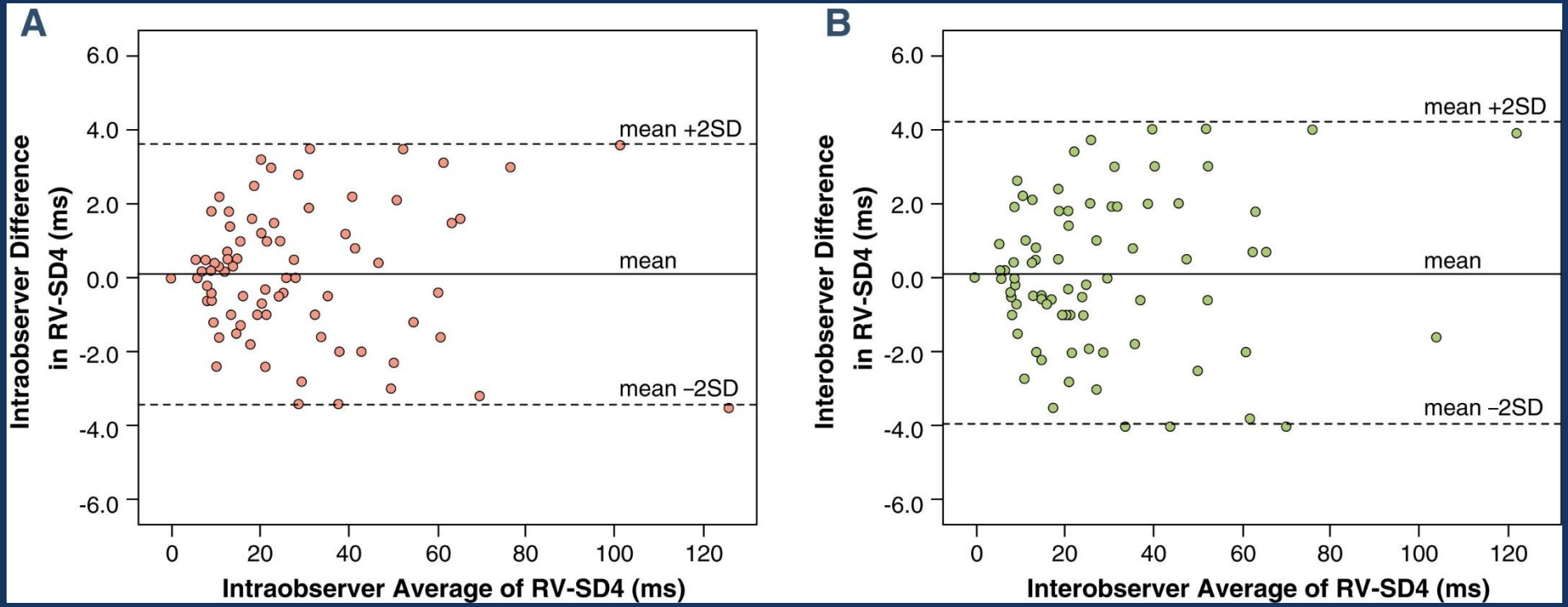
SUMMARY

In clinical measurement comparison of a new measurement technique with an established one is often needed to see whether they agree sufficiently for the new to replace the old. Such investigations are often analysed inappropriately, notably by using correlation coefficients. The use of correlation is misleading. An alternative approach, based on graphical techniques and simple calculations, is described, together with the relation between this analysis and the assessment of repeatability.

(*Lancet*, 1986; **i:** 307-310)

Bland-Altman plot

Intra-observer and Inter-observer variability assessed by the Bland-Altman method



Badagliacca, R, et al. (2015) JACC: CV Imaging

Bland-Altman plot

- 두 단계로 살펴 봐야 함.
  - **바이어스 (bias)와 변이 (variation)**

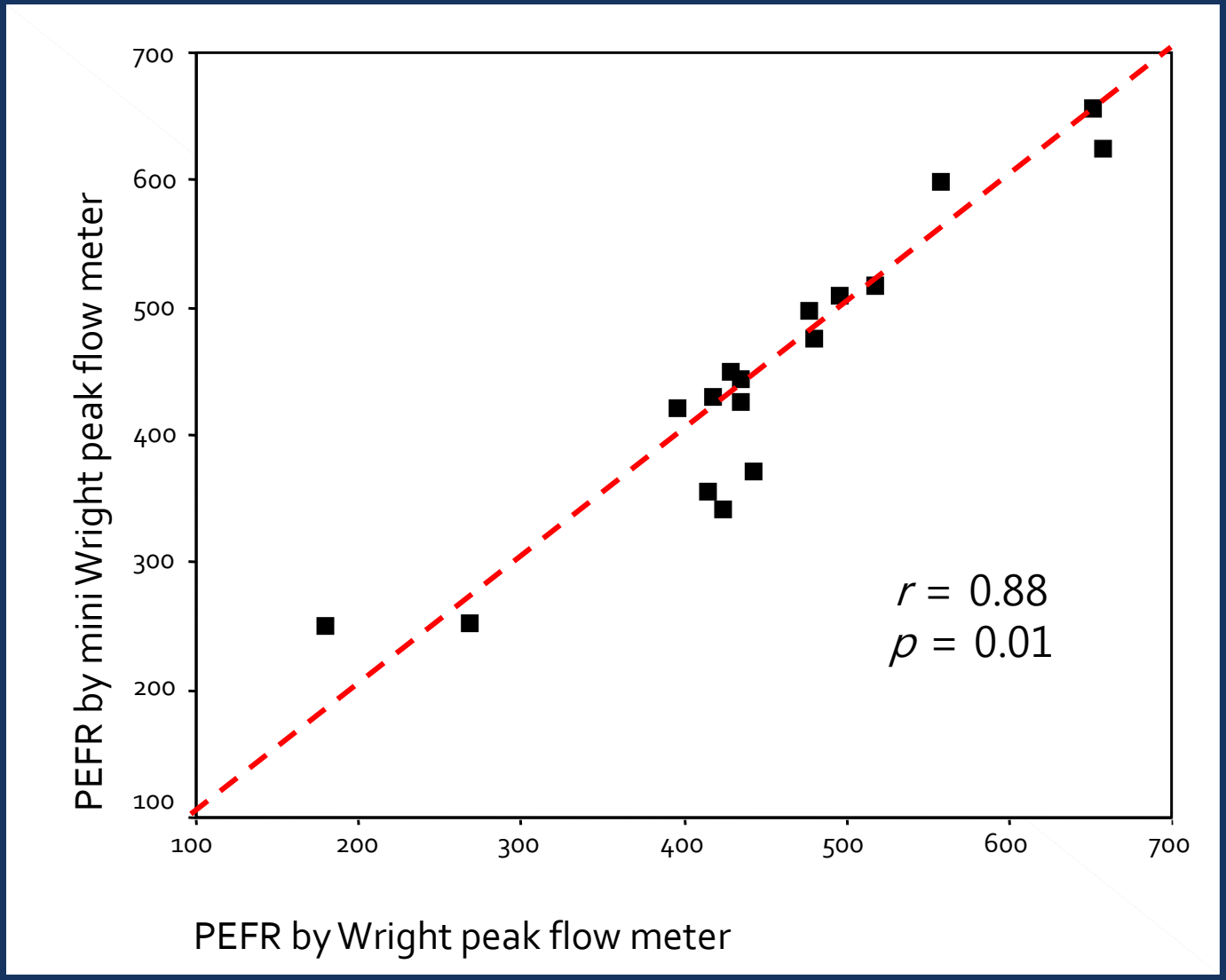
Bias & variation

- **바이어스 (bias)**
  - **측정법이 평균적으로 (on average) 일치하는가?, 또는 한 측정법이 다른 측정법보다 높은 /낮은 값을 읽는 경향이 있는가?**
  - 두 측정값 차이의 평균을 이용
- **변이 (variation)**
  - **측정법이 개별적으로 (for an individual) 일치하는가?**
  - 두 측정값 차이의 표준 편차를 이용

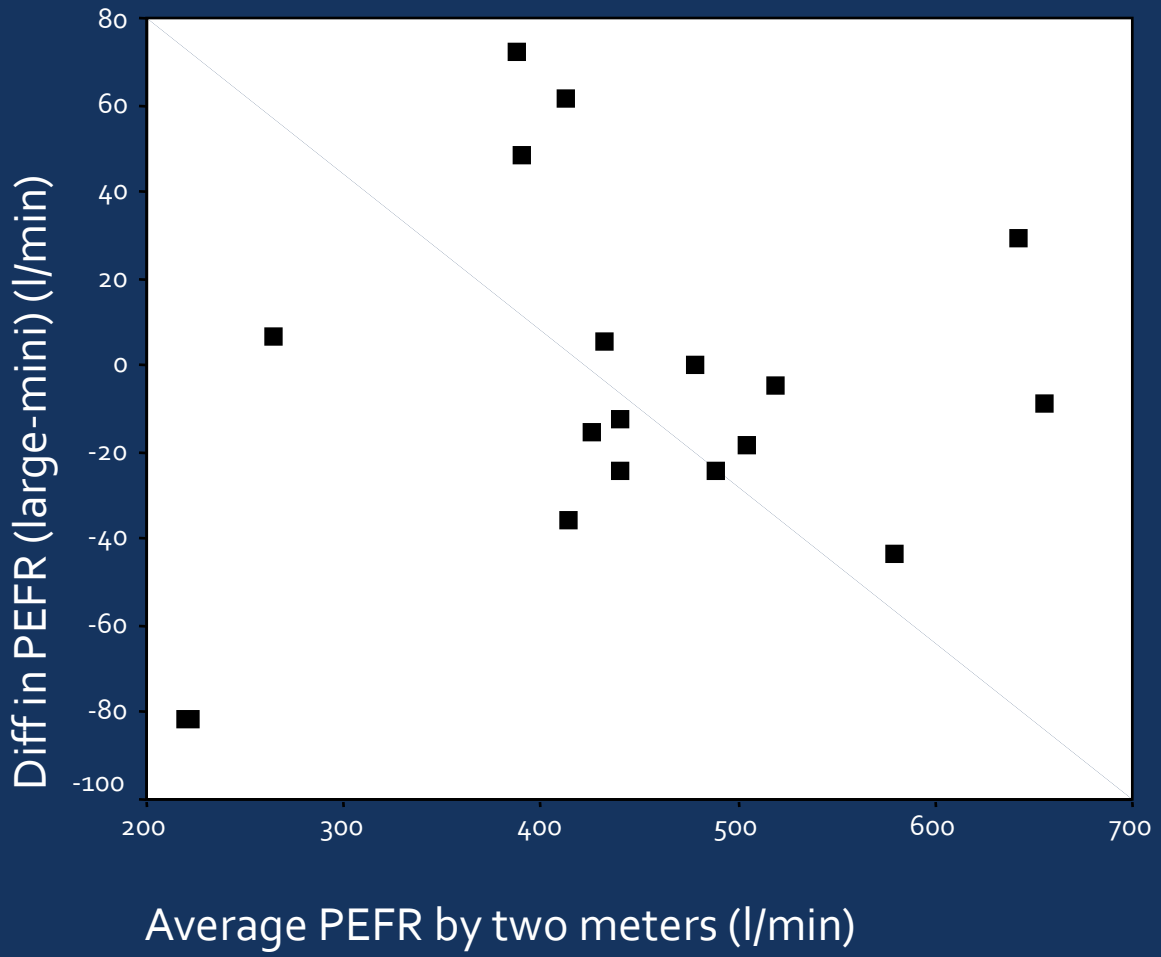
예시

- 최대 호기 유속 (Peak Expiratory Flow Rate (PEFR), l/min) 을 측정하는 서로 다른 두 측정법을 평가
  - 두 측정법은 같은 연속형 변수

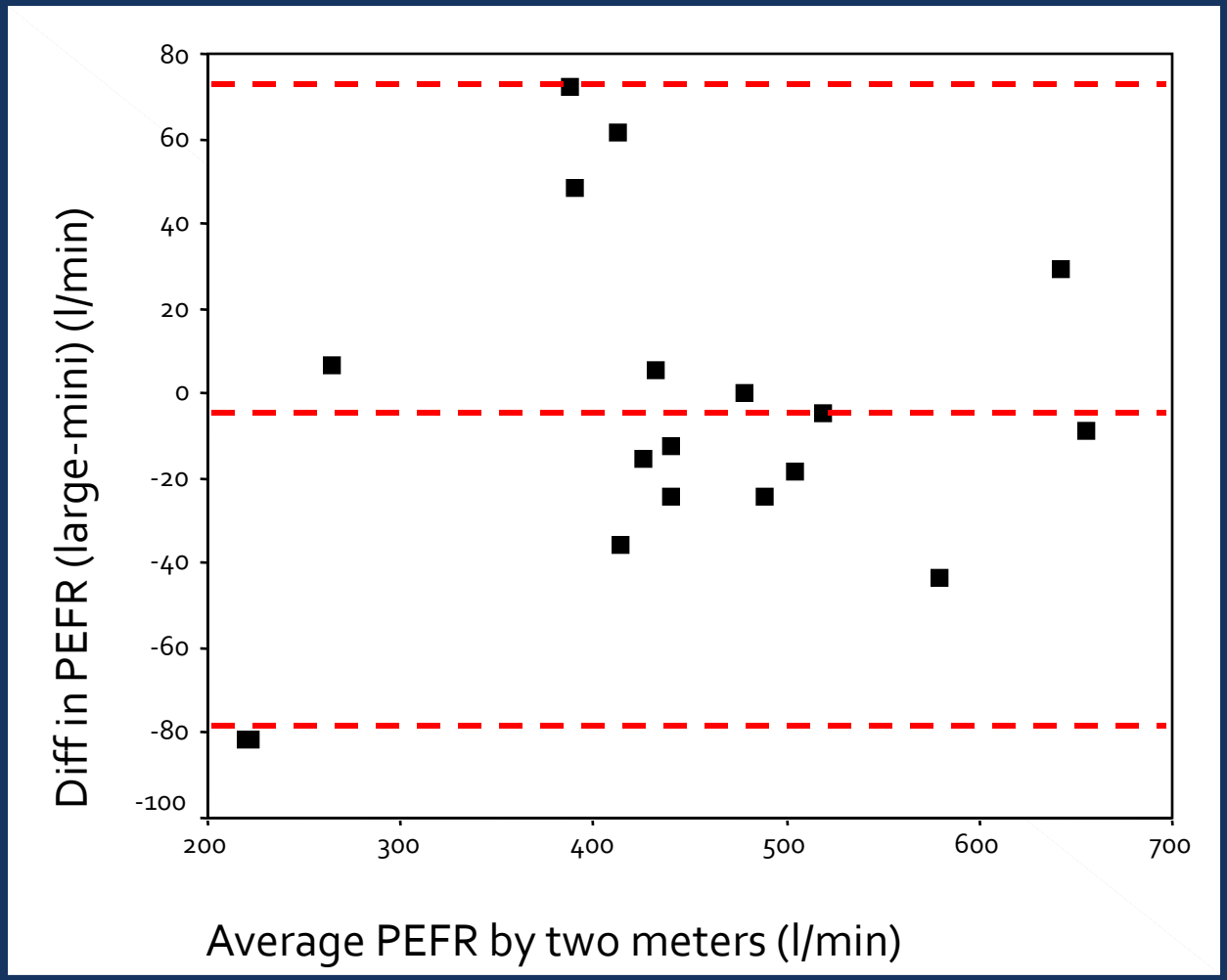
1단계: 일치선 위에 두 측정값 산점도 그리기



2단계: 두 측정값의 평균을 x축으로, 차이를 y축으로 산점도 그리기



3단계: 차이의 평균 (bias)과 일치도 한계 (variation)를 계산



4단계: 해석

- 바이어스 = 차이의 평균 = -2.1 l/min
- 일치 한계:

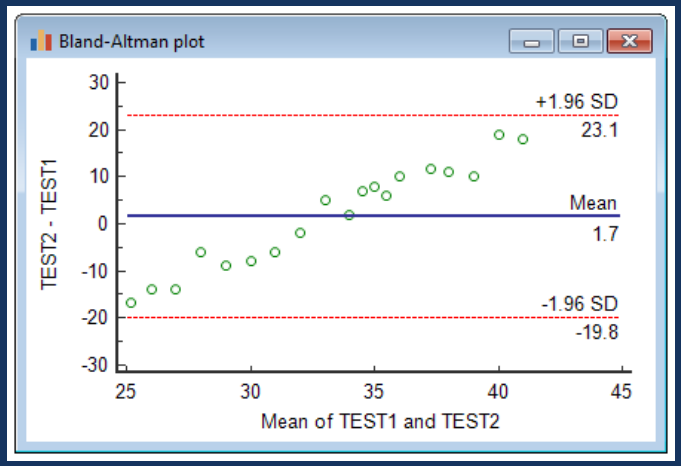
  mean difference ± 2 standard deviations

  = -2.1 - (2 x 38.8), -2.1 + (2 x 38.8)

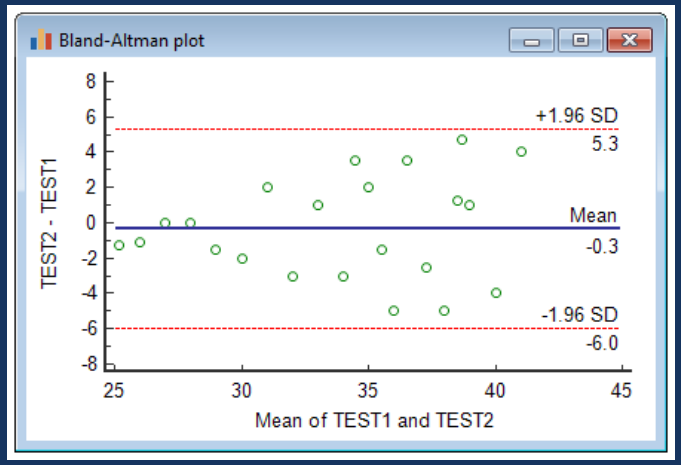
  = -79.7, 75.5 l/min
- The mini meter may be 80 l/min below or 76 l/min above the large meter. This is not acceptable for clinical purposes, but not immediately apparent from the scatterplot (nor from the correlation coefficient).

Interpretation of Bland-Altman plot

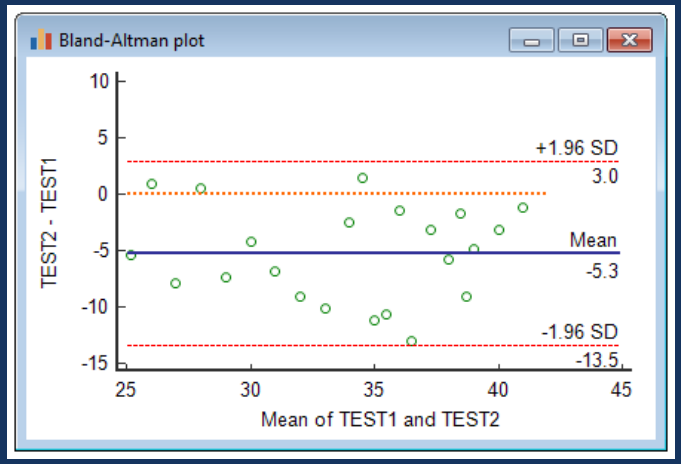
- Bland-Altman plot은 일반적으로 비공식적으로 해석. 아래 세 질문을 확인.
  - **두 측정법 사이에 평균적 불일치 (bias)가 얼마나 큰가?**
    - 임상적으로 해석. 불일치가 중요할 정도로 충분히 큰가? 이는 임상적 질문이지, 통계적 질문이 아님.
  - **경향이 있는가?**
    - 측정법 사이의 차이가 평균이 증가함에 따라 커지는 (또는 작아지는) 경향이 있는가?
  - **변동성 (variability)이 그래프 전체적으로 일정한가?**
    - Bias 선 주위로 흩어진 정도가 평균이 커짐에 따라 더 커지는가?



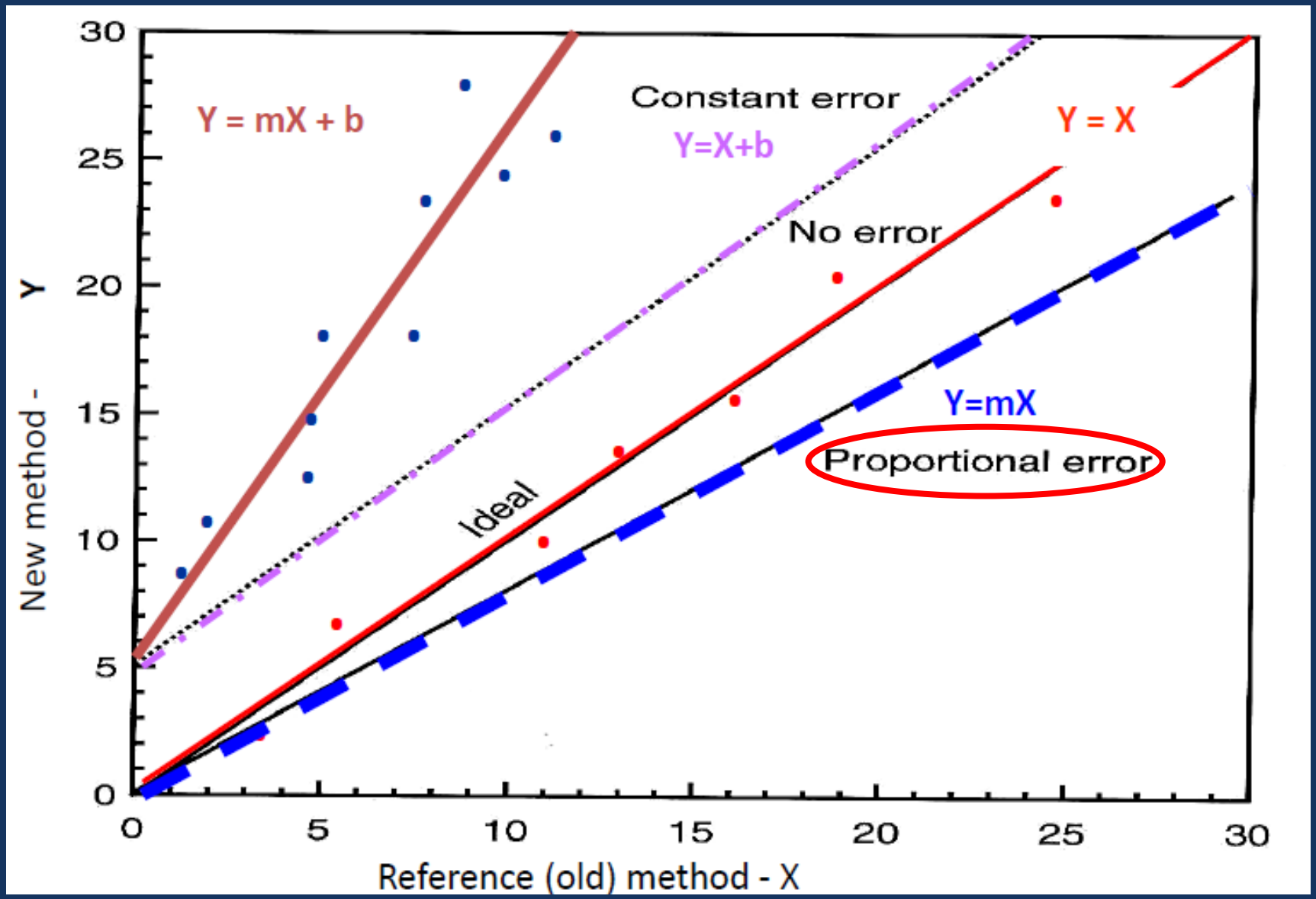
→ Example 1: Case of a proportional error.



→ Example 2: Case where the variation of at least one method depends strongly on the magnitude of measurements.

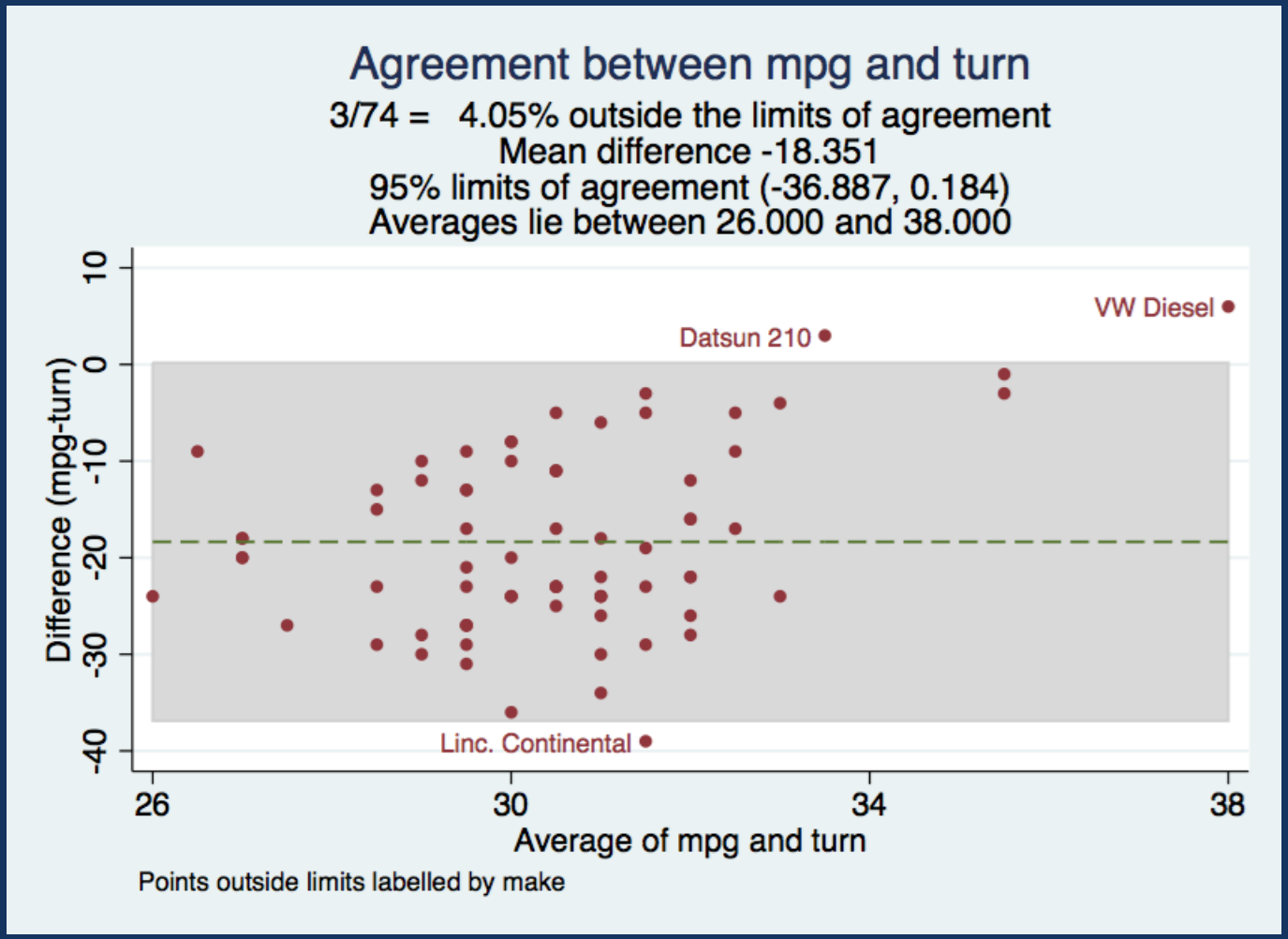


→ Example 3: Case of an absolute systematic error


Y = mX + b
Constant error
Y=X+b
Y = X
No error
Ideal
Y=mX
Proportional error
New method - Y
Reference (old) method - X
0
5
10
15
20
25
30

```
. sysuse auto, clear
(1978 Automobile Data)

. batplot mpg turn, title(Agreement between mpg and turn) info valabel(make) notrend xlab(26(4)3
> ) moptions(mlabp(9))
Mean difference      = -18.35135135135135
Limits of agreement = (-36.88690235522478,.1841996525220715)
Averages lie between  26.000 and 38.000
```

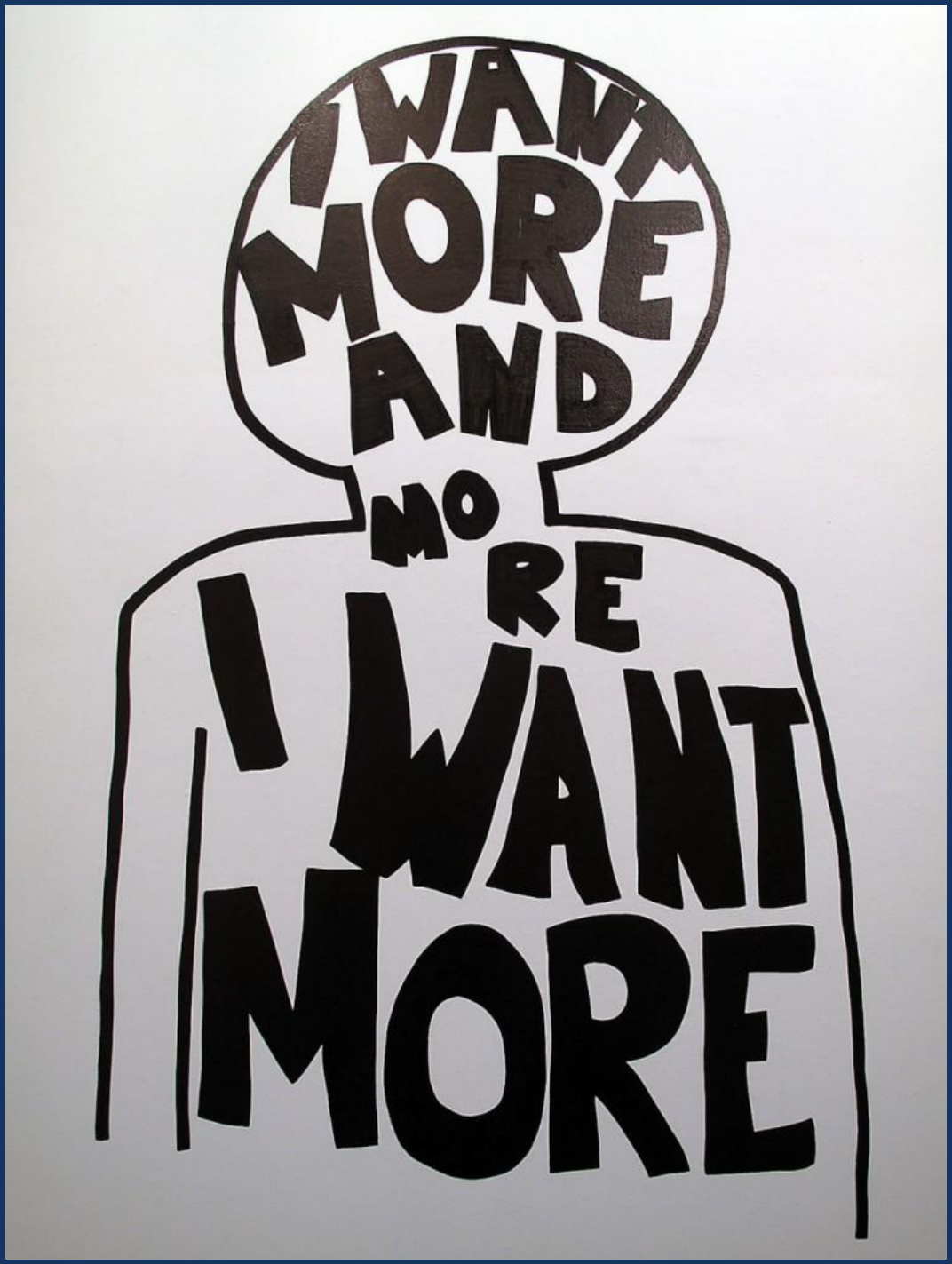


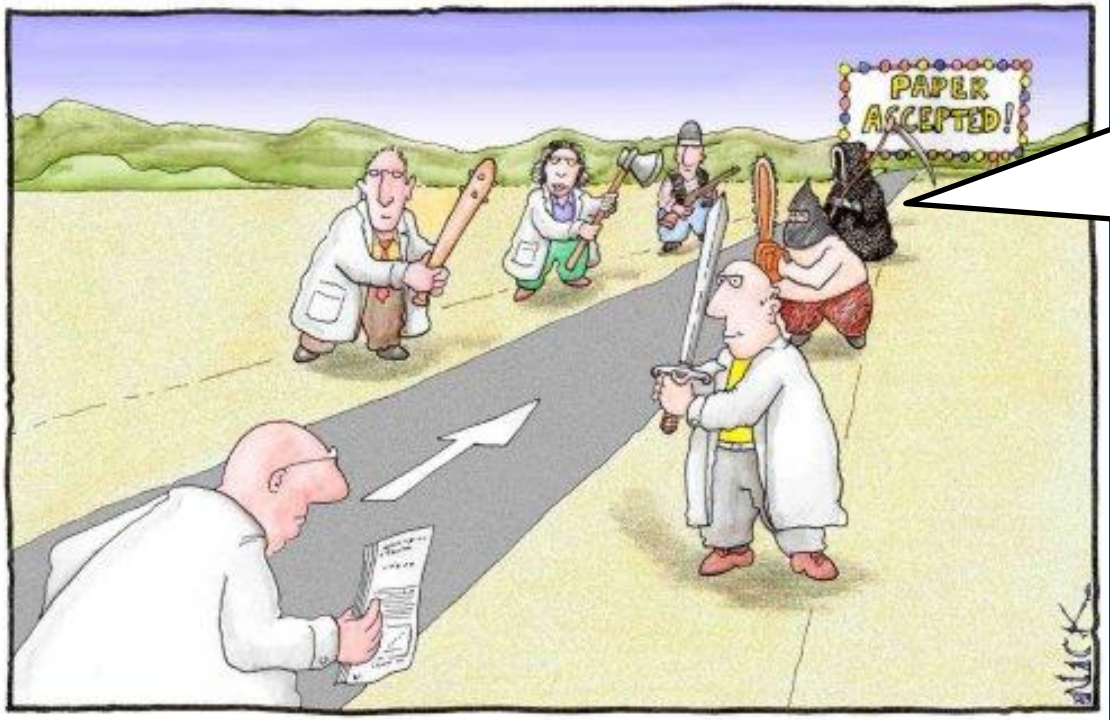
Which approach to be used?

Summary of Indices or Graphic Approaches Most Frequently Used for the Assessment of Validity and Reliability

		Mostly Used to Assess…	
Type of Variable	*Index or Technique*	*Validity*	*Reliability*
Categorical	**Sensitivity / Specificity**	++	
	Youden's *J* statistic	++	+
	Percent agreement	+	++
	Percent positive agreement	+	++
	Ordinal correlation coefficient (Spearman)	+	+
	Kappa statistic	+	++
Continuous	Scatter plot (correlation plot)	+	++
	Linear correlation coefficient (Pearson)	+	+
	Intra-class correlation coefficient	+	++
	Mean within-pair difference	+	++
	Coefficient of variation	+	++
	Bland-Altman plot	++	++

Modified from Szklo M, Nieto FJ. *Epidemiology Beyond the Basics*, 2nd Ed. (2007)


I WANT
MORE
AND
MO
RE
I WANT
MORE



Most scientists regarded the new streamlined peer-review process as 'quite an improvement.'

귀하의 논문에서는 새로운 검사법의 정상 / 비정상 구분을 cutoff value XX를 사용하여 분석하였습니다. Intra-observer 혹은 inter-observer variability가 이 cutoff value 보다 클 수 있을 것 같은데요?

Measures of reliability

Popular measures of **relative reliability**	Popular measures of **absolute reliability**
Intra-class correlation coefficient (ICC)	Limits of Agreement (LoA): **Bland-Altman plot** → 2 measurements
Pearson's r correlation coefficient	**Root mean square error (RMSE)** → 2 or more measurements
	Coefficient of Variation (CV)

RMSE (root mean square error)

$$Bias = \frac{1}{N}\sum_{i=1}^{n}(M_i - O_i)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(M_i - O_i)^2}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{n}|M_i - O_i|$$

M_i: 모의치
O_i: 실측치
N: 샘플링 개수

RMSE (root mean square error)

The error between simulation results and **field data**, which is calculated by using the repeated measure ANOVA.

This is also known as **the within-subject standard deviation**, which represents the **within-subject variation from test to test**, averaged over all subjects, **reflecting absolute reliability**.

Interpretation of RMSE

The difference **between a subject's measurement** and **the true value** would be expected to be **less than 1.96 x RMSE** for 95% of observations.

Another useful way of presenting measurement error is sometimes called the ***repeatability***, which is **√2 x 1.96 x RMSE**. **The difference between two measurements for the same subject** is expected to be less than **√2 x 1.96 x RMSE** for 95% of observations.

[Bland & Altman, 1996]

예시

실시간 3차원 심장초음파를 이용하여 RV-EF를 측정 시, intra-observer variability와 inter-observer variability를 평가

변수명:

1) id: 환자 고유번호
2) analyzer: 관찰자 구분번호 (ex. 관찰자 1, 관찰자 2)
3) time: 관찰자 내 분석 순서 (ex. 첫 번째 분석, 두 번째 분석)
4) EFp: RV-EF

예시

```
. anova EFp analyzer / id|analyzer time analyzer#time, repeated(time)

                           Number of obs =        60     R-squared     =  0.9718
                           Root MSE      =   2.23959     Adj R-squared =  0.9123
```

Source	Partial SS	df	MS	F	Prob > F
Model	3279.67075	40	81.9917688	16.35	0.0000
analyzer	10.8993571	1	10.8993571	0.13	0.7238
id|analyzer	3267.02896	38	85.9744462		
time	.257599598	1	.257599598	0.05	0.8231
analyzer#time	0	0			
Residual	95.2993423	19	5.01575486		
Total	3374.97009	59	57.2028829		

예시

- RMSE = 2.24%
- 1.96 x RMSE = 4.39%
- Repeatability = √2 x 1.96 x RMSE = 6.21%

이번 연구에서 RV-EF가 5% 호전을 보이면 치료 효과가 있는 것으로 정의하였는데, 2번 측정치의 차이가 6% 정도 될 수 있구나.

Take home message

1. 검사법 관련 연구: 타당도 / 신뢰도

2. 신뢰도 영향 주는 변이: 관찰자 내 / 간 변이

3. 신뢰도 측정 방법 (1): 순위 / 연속 척도

4. 신뢰도 측정 방법 (2): 상대적 / 절대적 분석

Thank you for your attention!