Reliability:
Why does it matter?

Robert B.M. Landewé, MD
Professor in rheumatology

Amsterdam Rheumatology & immunology Center (ARC-amC)
Consultant in rheumatology
Zuyderland medical center, Heerlen

The Netherlands

Measuring in medicine

- Blood pressure (mmHg)
- Body temperature (°C)
- Number of swollen joints (n)
- Response to treatment (yes vs.no)
- Erosions present vs absent on an X-ray (yes vs.no)
- mNY status on a pelvic radiograph (0,1,2,3,4)
- Counting a SLE criterion (present vs. absent)

Measuring is omnipresent

Studies

- “Systematic joint counts in studies”
- “X-rays scores for a study”
- “Biopsy specimens scores of test animals”
- “ELISA-test results for studies”
- “FACS-analysis on cell-samples”

Measuring is omnipresent

Clinical practice

- “History taking”
- “Doing physical examination and interpret the results”
- “Performing US”
- “Judging X-rays”
- “Interpreting the results of lab tests”
- “Making medical decisions”

Why do we measure patients?

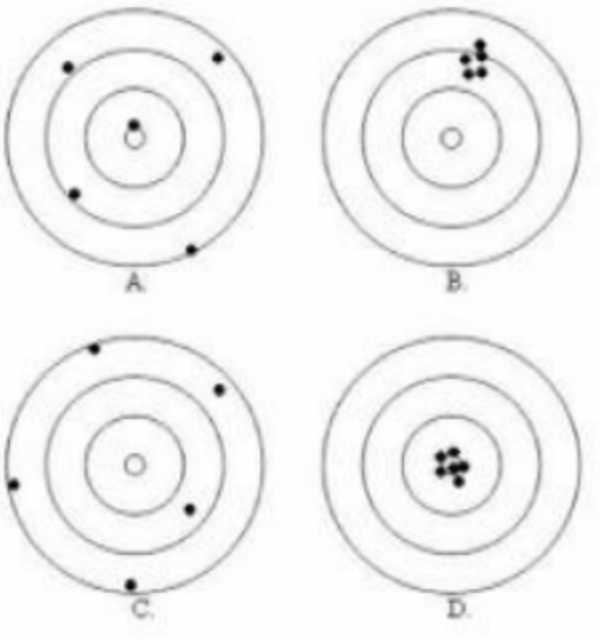
- We want to know *'the truth'*
- *Problem I:* By definition the truth is not known
- We can only approximate the truth by measuring the processes that we think are relevant
- The more often we measure the closer we approximate the truth
- *Problem II:* Only rarely we can measure more than once in medicine

Tests should be accurate

Accuracy vs. Precision

Figure 1

Accuracy and Precision



Reliability

How to find out?

1. Tests for use in clinical practice

We cannot measure reliability; we have to rely on others

"The test should be spot-on, since it guides a medical decision"

2. Tests for use in studies:

The extent to which a particular measure is insensitive to varying conditions:

- Retesting under unchanged conditions
- Retesting by other observers

"Some lenience can be allowed, since group-inference will help a lot, and consequences are rarely immediate"

Does reliability really matter?

One example

2003:

- Ankylosing spondylitis and mNY-status
- TNFi-treatment was thought to be promising
- Inclusion in RCTs was based on ‘local readings’ of pelvic X-rays
- RCTs showed clear separation between TNFi and placebo
- TNFi was smoothly approved for the indication of AS

Does reliability really matter?

One example

10 years later:

- Non-radiographic axSpA trials
- Enrolment based on local reads
- Reads were re-read by a panel of 3 expert readers (consensus)
- Agreement beyond chance was *moderate* at best
- Many ‘AS-patients’ have been treated with TNFi on unjust grounds
- Potentially dangerous overtreatment (~~€€€€~~)
- Pelvic X-ray reading for mNY status is *unreliable* for the clinic

Reliability implies variability

Total agreement among observers implies *no variability* in measurement results

- Total agreement in biology / biological systems is rare (if not absent)
- *Ergo:* Variability is inherent to studies with cells, tissues, animals, humans etc.

Testing the reliability of instruments aims at checking the level of agreement (and improve it) in order to constrain the variability of measurement results

Two levels of variation

- The instrument or test itself
- The measurer or observer or test-interpreter

Test the instrument I

Reproducibility

- Same thermometer used twice by the same observer in the same patient
- Same questionnaire completed twice by the same patients in unchanged conditions
- Same X-ray scored twice by the same observers

Test the instrument II

Agreement between different observers

If the *truth* is not known:

- The mean of two (or more) observations is closer to *the truth*
- High precision (& accuracy) provides increased statistical power in studies
- High precision (& accuracy) provides credit to a test to be used in clinical practice

Interobserver agreement

1. Reliability assessment embedded in the study (post-hoc):

Pro: convenient
Contra: prevalence imbalance, entire scale range not necessarily covered

2. Separate experimental setting (a-priori):

Pro: custom-fit selection of patients and observers
Contra: time-consuming & expensive

Three types of measurement results

- Dichotomous results/decisions (yes vs. no)
- Multinomial (categorical) results
- Continuous values/scores

Example:

Occiput-to-wall distance in ankylosing spondylitis

	Assessor 1		Assessor 2	
Patient	First test	Second test	First test	Second test
1	10	11	5	9
2	4	6	4	5
3	12	13	10	8
4	6	5	4	7
5	0	0	2	3
6	0	2	3	2
7	3	4	3	3
8	5	8	7	6
9	3	3	5	4
10	0	0	2	2
11	1	1	1	0
12	1	3	3	3
13	0	2	2	1
14	7	9	9	10
15	13	16	11	11
16	15	15	14	12
17	0	1	2	5
18	20	17	19	20
19	6	8	9	8
20	9	11	7	6
mean	5,75	6,75	6,3	6,25

Example:

Occiput-to-wall distance in ankylosing spondylitis

	Assessor 1		Assessor 2	
Patient	**First test**	**Second test**	**First test**	**Second test**
1	10	11	5	9
2	4	6	4	5
3	12	13	10	8
4	6	5	4	7
5	0	0	2	3
6	0	2	3	2
7	3	4	3	3
8	5	8	7	6
9	3	3	5	4
10	0	0	2	2
11	1	1	1	0
12	1	3	3	3
13	0	2	2	1
14	7	9	9	10
15	13	16	11	11
16	15	15	14	12
17	0	1	2	5
18	20	17	19	20
19	6	8	9	8
20	9	11	7	6
mean	5,75	6,75	6,3	6,25

- Grand mean: 6,26
- Assessor 1: 6,25
- Assessor 2: 6,27

Intraclass-correlation coefficients (ICC)

"That proportion of variance in a reliability experiment that is attributable to (true) variance among patients (and not to measurement error)"

In a formula:

$$ICC\ (0,1) = \frac{VAR_{pt}}{VAR_{pt}+VAR_{obs}+VAR_{rand}}$$

ICC = *Relative* agreement

Define carefully what you are interested in

Reliability of mTSS scores in a RA trial:

- "At baseline inter-reader reliability is excellent" (ICC = 0.98)
- "Reliability of change scores was moderate" (ICC = 0.65)
- VAR_{pt} depends on a range between 0 and 150 units for absolute scores
- VAR_{pt} depends on a range between -2,5 and +2,5 for change scores

If you are interested in *change* score as an outcome, look at the reliability of measuring *change* scores !!

Example:

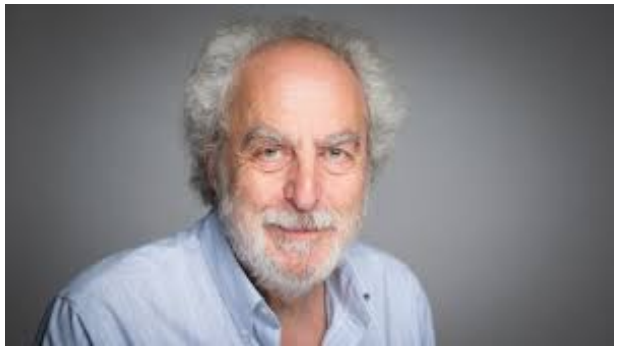
Occiput-to-wall distance in ankylosing spondylitis

	Assessor 1			Assessor 2		
Patient	First test	Second test	Delta	First test	Second test	Delta
1	10	11	1	5	9	4
2	4	6	2	4	5	1
3	12	13	1	10	8	-2
4	6	5	-1	4	7	3
5	0	0	0	2	3	1
6	0	2	2	3	2	-1
7	3	4	1	3	3	0
8	5	8	3	7	6	-1
9	3	3	0	5	4	-1
10	0	0	0	2	2	0
11	1	1	0	1	0	-1
12	1	3	2	3	3	0
13	0	2	2	2	1	-1
14	7	9	2	9	10	1
15	13	16	3	11	11	0
16	15	15	0	14	12	-1
17	0	1	1	2	5	3
18	20	17	-3	19	20	1
19	6	8	2	9	8	-1
20	9	11	2	7	6	-1
mean	5,75	6,75	1	6,3	6,25	-0.05

Ass 1: Systematic error+ random error

Ass 2: Random error

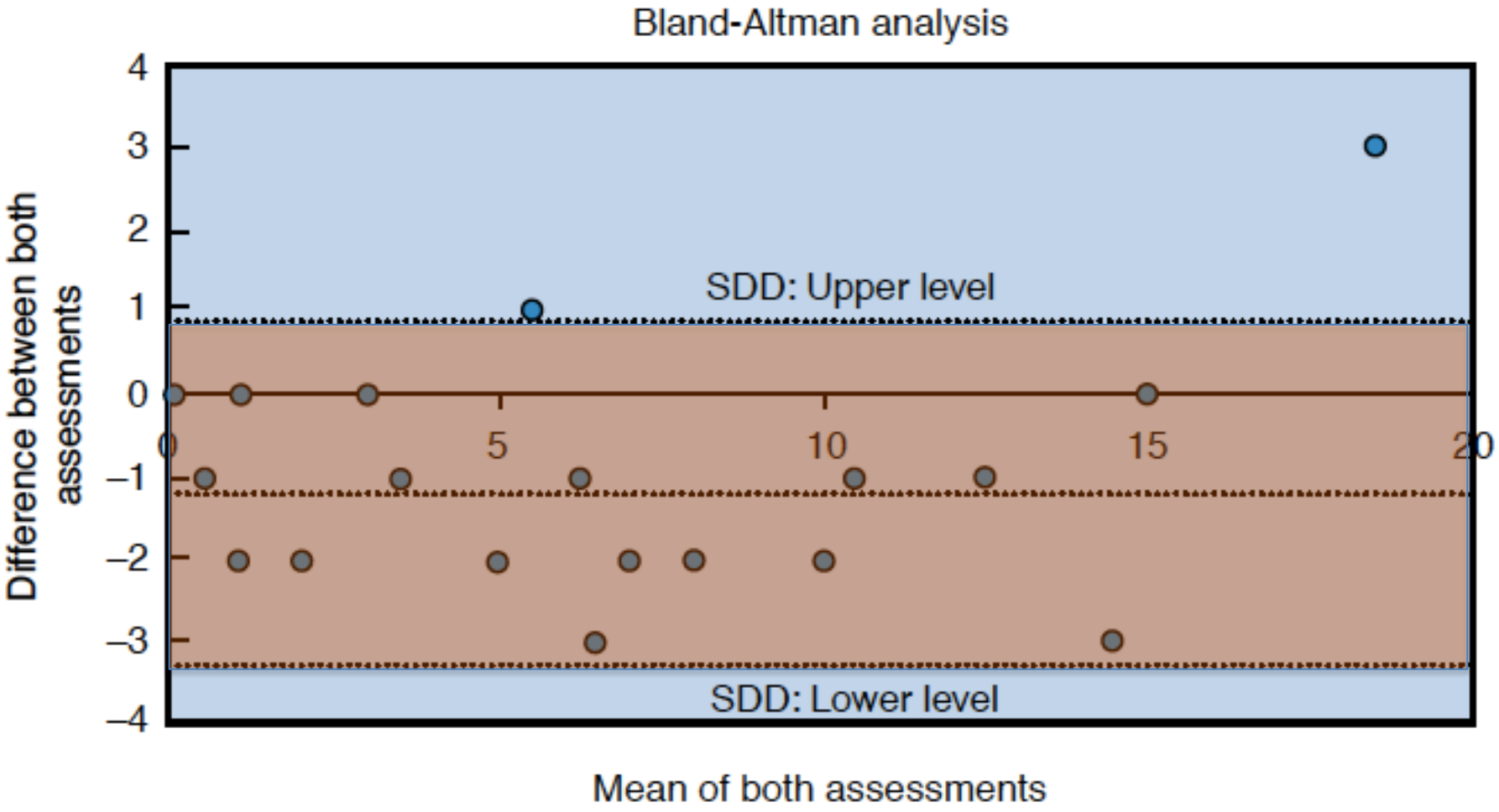
Bland&Altman plot



Bland JM, Altman DG.
Statistical methods for assessing agreement between 2 methods of clinical measurement.
Lancet i, (1986) 307-310

Absolute agreement

Assessor 1



95% Limits of agreement

How to calculate?

$$95\%\text{LoA} = \text{mean difference score} \pm \frac{1{,}96 * \text{SD (difference scores)}}{\sqrt{2}}$$

Bland & Altman plot

What can you learn from it?

- Is there systematic error?
- How *wide* is the 95% band of agreement?
- Is the deviation from optimal ‘bandwise’ (*homoscedastic*) or dependent on the scale range (*heteroscedastic*)?
- Have you included *all potential values* in your reliability experiment?

Smallest detectable *difference* (SDD) or Smallest detectable *change* (SDC)?

- The word ‘difference’ in SDD is a ‘misnomer’
- SDD includes 95% of all variation within the bands of the *limits of agreement*: It is simply distribution-theory
- Upper/lower LoA: mean difference ± $1{,}96 * SD_{\text{interobs differences}} / \sqrt{n_{\text{observers}}}$
- Smallest detectable change (SDC)-concept:
  “*Suppose you have a patient with a baseline value and a follow up value, and you want to decide if that patient has improved/worsened: What is the cut-off level beyond which you may safely assume that this patient has changed beyond measurement error?”*
- SDC = SDD / $\sqrt{n}$, (in which n=number of measurements).

Intermezzo: SPSS in reliability studies

What to choose?

Model specification:

Type: Absolute agreement or consistency definition?

- (2,4), (4,6), (6,8) are in perfect agreement using a 'consistency definition' (ICC=1) but not an absolute agreement definition (ICC=0.67)

Model: 'Two-way mixed or two-way random'?

- All patients are measured by all readers, and these readers are the only readers of interest (Model 3 or 'two way-mixed')

Output:

Which ICC: Single-measures or average-measures?

- Always ask yourself the question how you want to utilize the score!
  In case of doubt, report the most conservative one (single-measures).

Kappa or ICC?

A kappa is an ICC for binomial (nominal) test results

	Positive	Negative	Total
Positive	50	50	100
Negative	50	350	400
Total	100	400	500

Agreement in (50+350)=400/500 patients (80%)

Chance agreement is already substantial, given the distribution

Kappa is the proportion agreement beyond chance

Cohen J: Psychological Measurement 1960; 20:37-46
Viera A and Garrett J: Fam Med 2005;37:360-3

	Positive	Negative	Total
Positive	50 (a) 20	50 (b)	100
Negative	50 (c)	350 (d) 320	400
Total	100	400	500

$$\text{Kappa} = \frac{\text{Observed agreement} - \text{Expected agreement}}{(1\text{-Expected agreement})}$$

Observed agreement = (50+350) / 500 = 0.80

Expected agreement = (Exp cell frequency (a) + Exp cell frequency (d)) / 500 = 0.68

Exp cell frequency (a) = (50+50)(50+50) / 500 = 20

Exp cell frequency (d) = (50+350)(50+350) / 500 = 320

Kappa = (0.80 – 0.68) / (1-0.68) = 0.375

How to interpret kappa?

Interpretation of Kappa

	Poor	Slight	Fair	Moderate	Substantial	Almost perfect
Kappa	0.0	.20	.40	.60	.80	1.0

Kappa	Agreement
< 0	Less than chance agreement
0.01–0.20	Slight agreement
0.21– 0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–0.99	Almost perfect agreement

Weighted kappa

	Normal	Abnormal but benign	Abnormal malignant
RADIOLOGIST 1	x		
RADIOLOGIST 2		x	
RADIOLOGIST 3			x

Partial agreement vs. No agreement

Cohen J: Psychological Bulletin 1968;70:213-20

The paradox: High agreement low kappa

		Central read		
		Abnormal	Normal	Total
Local read	Abnormal	1	6	7
	Normal	9	84	93
	Total	10	90	100

Kappa: 0.04
Observed agreement: 0.85

"*Symmetrically unbalanced marginal totals*"

Cichetti D, Feinstein A: J Clin Epidemiol 1990; 43: 551-8

The paradox: High agreement low kappa

		Central read		
		Abnormal	Normal	Total
Local read	Abnormal	1	6	7
	Normal	9	84	93
	Total	10	90	100

Replace cells *a* and *d* by their average
Replace cells *b* and *c* by their average

Prevalence-adjusted-bias-adjusted kappa (PABAK)

Byrt T et al: J Clin Epidemiol 1993, 46;423-9

Prevalence-adjusted Bias adjusted Kappa (PABAK)

		Central read		
		Abnormal	Normal	Total
Local read	Abnormal	1	6	7
	Normal	9	84	93
	Total	10	90	100

Observed Kappa=0.04

		Central read		
		Abnormal	Normal	Total
Local read	Abnormal	*43*	*7*	50
	Normal	*8*	*42*	50
	Total	51	49	100

Adjusted PABAK=0.90

Summary

- A test and/or measurement instrument should be reliable
- It ideally measures the *truth* (no systematic or random error)
- It should be precise (reproducible, high agreement)
- Ideally reliability experiments are designed and performed befóre the actual study takes place
- Calculating measures of reliability (ICC, Kappa) requires a computer and a software program
- Interpreting measures of reliability (ICC, Kappa) needs some skills and training