11 Nov 2020

Advanced Topics in Biostatistics: Measuring Agreement

Prof. Dr. Annette Kopp-Schneider
Biostatistics – C060
kopp@dkfz.de

DEUTSCHES KREBSFORSCHUNGSZENTRUM IN DER HELMHOLTZ-GEMEINSCHAFT

Outline

- What is meant by 'Agreement'

- Measuring agreement for categorical data:
  - Cohen's Kappa
  - weighted Kappa
  - Fleiss' Kappa

- Measuring agreement for quantitative data
  - Bland-Altman Plot
  - Bland-Altman Plot using SigmaPlot
  - Shapes in Bland-Altman Plot
  - Alternative intervals: Prediction interval, Tolerance interval, Total Deviation Index, Coverage Probability
  - Scaled indices: Concordance Correlation Coefficient, Intraclass Correlation Coefficient

Agreement: The set-up

n items, e.g. patients

R raters or measurement methods (to assess the n items)

Ratings/measurements can be categorical or continuous.

Assumption:

There is no gold standard (true value) known.

Aim:

Explore degree of agreement between the raters/measurement methods on the items („interrater/intermethod agreement“).

Note: agreement ≠ correlation

Situations

Categorical outcome:

Raters categorize items into 2 or more categories.

Quantitative outcome:

Raters measure a continuous value.

Situations

Categorical outcome:
Raters categorize items into 2 or more categories.

Quantitative outcome:
Raters measure a continuous value.

Example

Doctor A and doctor B both independently see 100 patients and rate whether they have a disease:

	Doctor A		
Doctor B	yes	no	total
yes	50	15	65
no	5	30	35
total	55	45	100

How well do Doctors A and B agree in their diagnosis?

Problem setup

- n patients
- Two raters, A and B.
- Each rater classifies each patient as belonging to one of two categories such as „yes“/“no“; e.g. „tumor present“/ “tumor absent“
- Assumption: The raters perform their ratings without knowledge of each other

	Rater A		
Rater B	yes	no	total
yes	a	b	a+b
no	c	d	c+d
total	a+c	b+d	n=a+b+c+d

Naive approach

Question: What is the agreement between the two raters?

Naive approach: consider the proportion of observed agreement

$$p_O = \frac{a+d}{n}$$

as a measure of agreement between the two raters

Rater A			
Rater B	yes	no	total
yes	a	b	a+b
no	c	d	c+d
total	a+c	b+d	n=a+b+c+d

Problem with naive approach

Rater A			
Rater B	yes	no	total
yes	50	15	65
no	5	30	35
total	55	45	100

Rater C			
Rater D	yes	no	total
yes	0	20	20
no	0	80	80
total	0	100	100

Problem with naive approach

How much agreement do we expect to occur by chance alone?

	Rater A		
Rater B	yes	no	total
yes	50	15	65
no	5	30	35
total	55	45	100

	Rater C		
Rater D	yes	no	total
yes	0	20	20
no	0	80	80
total	0	100	100

Agreement expected by chance

$$p_{A,yes} = \frac{a+c}{n}$$

$$p_{A,no} = \frac{b+d}{n}$$

$$p_{B,yes} = \frac{a+b}{n}$$

$$p_{B,no} = \frac{c+d}{n}$$

	Rater A		
Rater B	yes	no	total
yes	a	b	a+b
no	c	d	c+d
total	a+c	b+d	n=a+b+c+d

Agreement expected by chance :

$$p_E = p_{A,yes} \cdot p_{B,yes} + p_{A,no} \cdot p_{B,no}$$

Cohen's kappa

Cohen's kappa $$\kappa = \frac{p_O - p_E}{1 - p_E}$$

p_O : observed agreement

p_E : agreement expected by chance

1: maximum possible agreement

Interpretation: κ is the agreement adjusted for agreement expected by chance

$\kappa \leq p_O$

If $p_O = 1$ then $\kappa = 1$

$\kappa = 0 \leftrightarrow p_O = p_E$: $\kappa = 0$ means no agreement

Cohen's kappa in example

For A, B:

$p_O = 0.8$

$$p_E = \frac{55}{100} \cdot \frac{65}{100} + \frac{45}{100} \cdot \frac{35}{100} = 0.515$$

$$\kappa = \frac{0.8 - 0.515}{1 - 0.515} = 0.588$$

	Rater A		
Rater B	yes	no	total
yes	50	15	65
no	5	30	35
total	55	45	100

For C, D:

$p_O = 0.8$

$$p_E = \frac{0}{100} \cdot \frac{20}{100} + \frac{100}{100} \cdot \frac{80}{100} = 0.8$$

$$\kappa = \frac{0.8 - 0.8}{1 - 0.8} = 0$$

	Rater C		
Rater D	yes	no	total
yes	0	20	20
no	0	80	80
total	0	100	100

Assessing the value of Cohen’s kappa

Interpretation of Cohen’s kappa according to Landis and Koch (1977):

Value of kappa	strength of agreement
<0.00	poor
0.00-0.20	slight
0.21-0.40	fair
0.41-0.60	moderate
0.61-0.80	substantial
0.81-1.00	almost perfect

Confidence interval for Cohen's kappa

The standard error for κ is given by

$$se(\kappa) = \sqrt{\frac{p_O(1-p_O)}{n(1-p_E)^2}}$$

The 100·(1-α)% Confidence Interval is given by

$$\left[\kappa - z_{1-\alpha/2}se(\kappa), \kappa + z_{1-\alpha/2}se(\kappa)\right]$$

where $z_{1-\alpha/2}$ is the (1-α/2)-quantile of the Standard Normal Distribution

e.g. for α = 0.05: $z_{1-\alpha/2} = z_{0.975} = 1.96$

→ test of kappa, e.g.

H_0: κ = 0 vs. H_1: κ ≠ 0 is significant ($p < 0.05$) if 95%-CI does not contain 0

H_0: κ = c vs. H_1: κ ≠ c is significant ($p < 0.05$) if 95%-CI does not contain c

Confidence interval for kappa in examples

$p_E = 0.515$, $p_O = 0.8$, $\kappa = 0.588$

$$se(\kappa) = \sqrt{\frac{p_O(1-p_O)}{n(1-p_E)^2}} = \sqrt{\frac{0.8(1-0.8)}{100(1-0.515)^2}} = 0.082$$

	Rater A		
Rater B	yes	no	total
yes	50	15	65
no	5	30	35
total	55	45	100

95% Confidence Interval is given by

$[\kappa - 1.96 \cdot 0.082, \kappa + 1.96 \cdot 0.082] = [0.588 - 1.96 \cdot 0.082, 0.588 + 1.96 \cdot 0.082]$

$= [0.427, 0.749]$

$p_E = 0.8$, $p_O = 0.8$, $\kappa = 0$

$se(\kappa) = 0.2$

95% Confidence Interval is given by

$[\kappa - 1.96 \cdot 0.2, \kappa + 1.96 \cdot 0.2] = [-0.392, 0.392]$

	Rater C		
Rater D	yes	no	total
yes	0	20	20
no	0	80	80
total	0	100	100

Assessing the value of Cohen's kappa

Caveat: kappa depends on prevalence

	Rater A		
Rater B	yes	no	total
yes	1	1	2
no	1	97	98
total	2	98	100

kappa=0.49

	Rater A		
Rater B	yes	no	total
yes	0	1	1
no	1	98	99
total	1	99	100

kappa=-0.01

	Rater A		
Rater B	yes	no	total
yes	1	1	2
no	0	98	98
total	1	99	100

kappa=0.67

	Rater A		
Rater B	yes	no	total
yes	1	0	1
no	0	99	99
total	1	99	100

kappa=1

Assessing the value of Cohen’s kappa

Caveat: kappa depends on prevalence

	Rater A		
Rater B	yes	no	total
yes	40	9	49
no	6	45	51
total	46	54	100

kappa=0.70

	Rater A		
Rater B	yes	no	total
yes	80	10	90
no	5	5	10
total	85	15	100

kappa=0.32

	Rater A		
Rater B	yes	no	total
yes	45	15	60
no	25	15	40
total	70	30	100

kappa=0.13

	Rater A		
Rater B	yes	no	total
yes	25	35	60
no	5	35	40
total	30	70	100

kappa=0.26

Cohen's kappa for more than two categories

- 50 cancer patients
- Two raters, A and B
- For each patient both raters estimate the degree of spread to regional lymph nodes (N0,N1,N2,N3)

	Rater A				
Rater B	N0	N1	N2	N3	total
N0	**3**	2	3	2	10
N1	3	**3**	3	3	12
N2	1	4	**6**	6	17
N3	3	1	3	**4**	11
Total	10	10	15	15	50

$$p_O = \frac{3+3+6+4}{50} = 0.32$$

$$p_E = \frac{10}{50}\cdot\frac{10}{50} + \frac{10}{50}\cdot\frac{12}{50} + \frac{15}{50}\cdot\frac{17}{50} + \frac{15}{50}\cdot\frac{11}{50} = 0.256$$

Cohen's kappa for more than two categories

$$p_O = 0.32$$

$$p_E = 0.256$$

$$\kappa = \frac{0.32 - 0.256}{1 - 0.256} = 0.086$$

$$se(\kappa) = \sqrt{\frac{0.32(1-0.32)}{50(1-0.256)^2}} = 0.089$$

	Rater A				
Rater B	N0	N1	N2	N3	total
N0	**3**	2	3	2	10
N1	3	**3**	3	3	12
N2	1	4	**6**	6	17
N3	3	1	3	**4**	11
Total	10	10	15	15	50

95%-CI for κ : [0.086 – 1.96·0.089, 0.086 + 1.96·0.089] = [-0.088,0.260]

Extensions of Cohen‘s Kappa

For more than two categories:

- κ does not take degree of disagreement into account

  → If appropriate, for ordinal measurement:

  Use weighted kappa κ_w with weights, e.g. quadratic weights

Rater B	Rater A: N0	N1	N2	N3
N0	1	0.89	0.56	0
N1	0.89	1	0.89	0.56
N2	0.56	0.89	1	0.89
N3	0	0.56	0.89	1

- Calculate p_O and p_E using weights

- If weights are 1 for agreement and 0 for disagreement: $\kappa_w = \kappa$

- Decision about which weights to use before data collection!

For more than two raters:

- Use Fleiss‘ kappa

Situations

Categorical outcome:

Raters categorize items into 2 or more categories.

Quantitative outcome:

Raters measure a continuous value.

Example

Quantitative outcome

Evaluation of diameter measurements for thoracic endovascular aortic repair

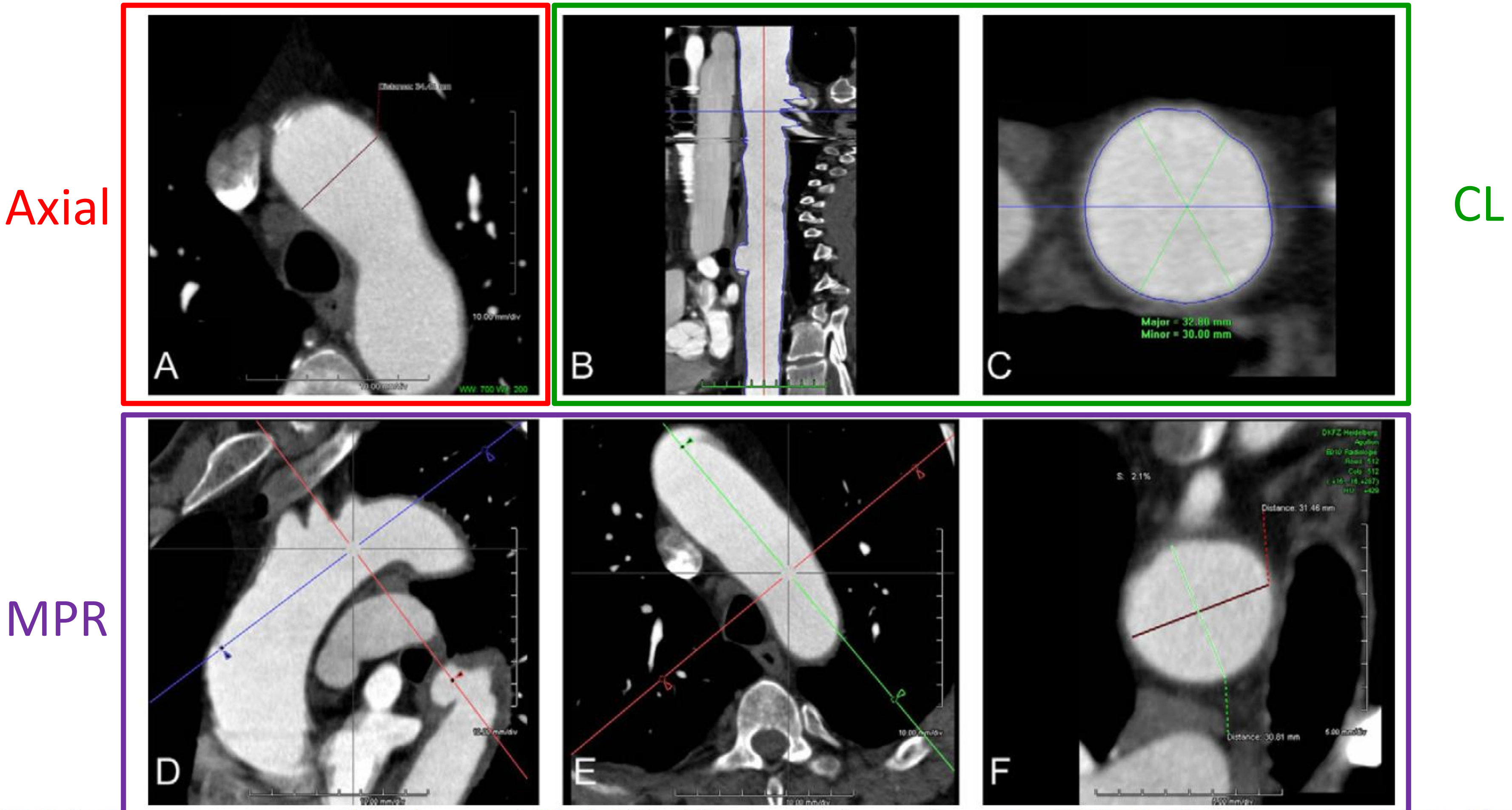


Figure 1. Diameter assessment based on axial (A), double oblique multiplanar reformation (MPR; D—F) and centerline (CL, B—C) techniques. Measuring on axial images, the course of the aorta can only be assessed visually (A), whereas MPR and centerline analysis allow for measurements in a plane perpendicular to the vessel course (C, F).

Müller-Eschner et al. 2013

Quantitative outcome

Example

Data:

- 30 patients
- 3 evaluation methods:
  - Axial
  - Manual double-oblique multiplanar reformations (MPRs)
  - Semiautomatic centerline analysis (CL)
- Measured at several aortic positions: P1,..,P4
- Two experts

Question:

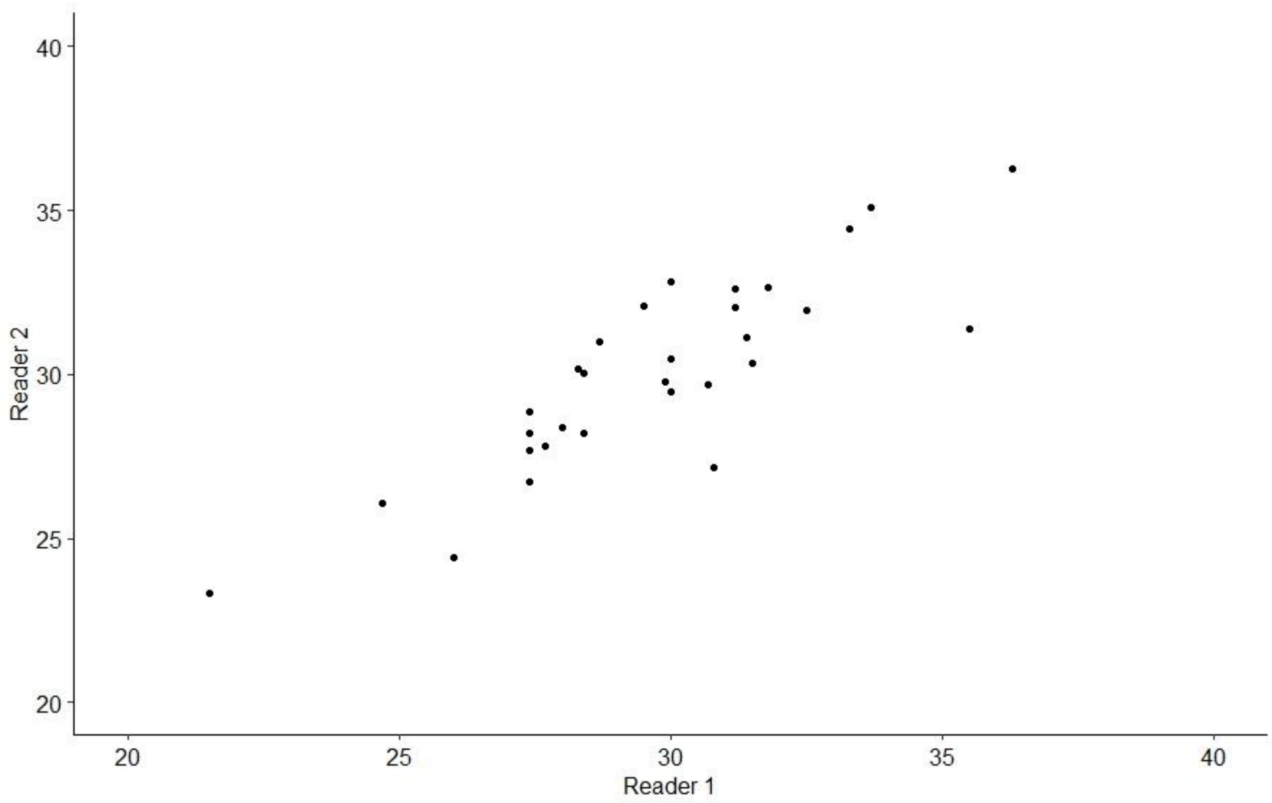
How well do measurements agree:

- from different methods for a given expert?
- from different experts for a given method?

Scatterplots

Quantitative outcome

Axial, P1



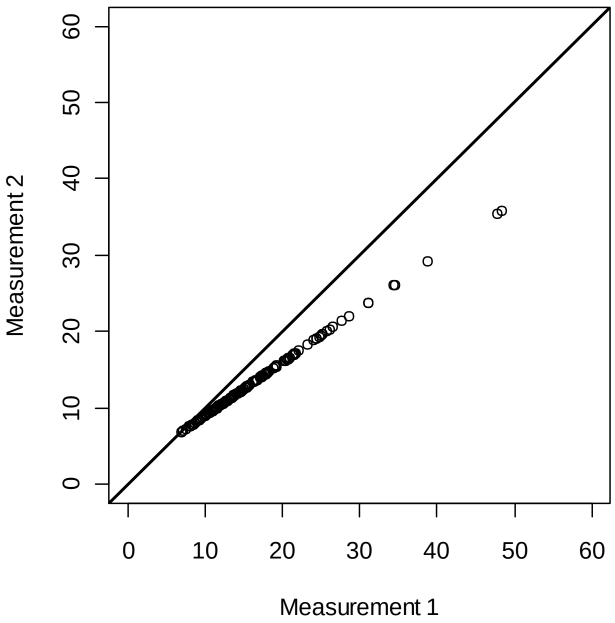
r=0.81
95%-CI: 0.72 to 0.93
p<<0.001

→ **Interpretation?**

Correlation coefficient?

Correlation ≠ Agreement:

Correlation r=1, p<0.0001



Evaluating the difference

Focus of interest: difference between measurements $Y_1 - Y_2$

For Axial, P1, Reader1 and Reader2:

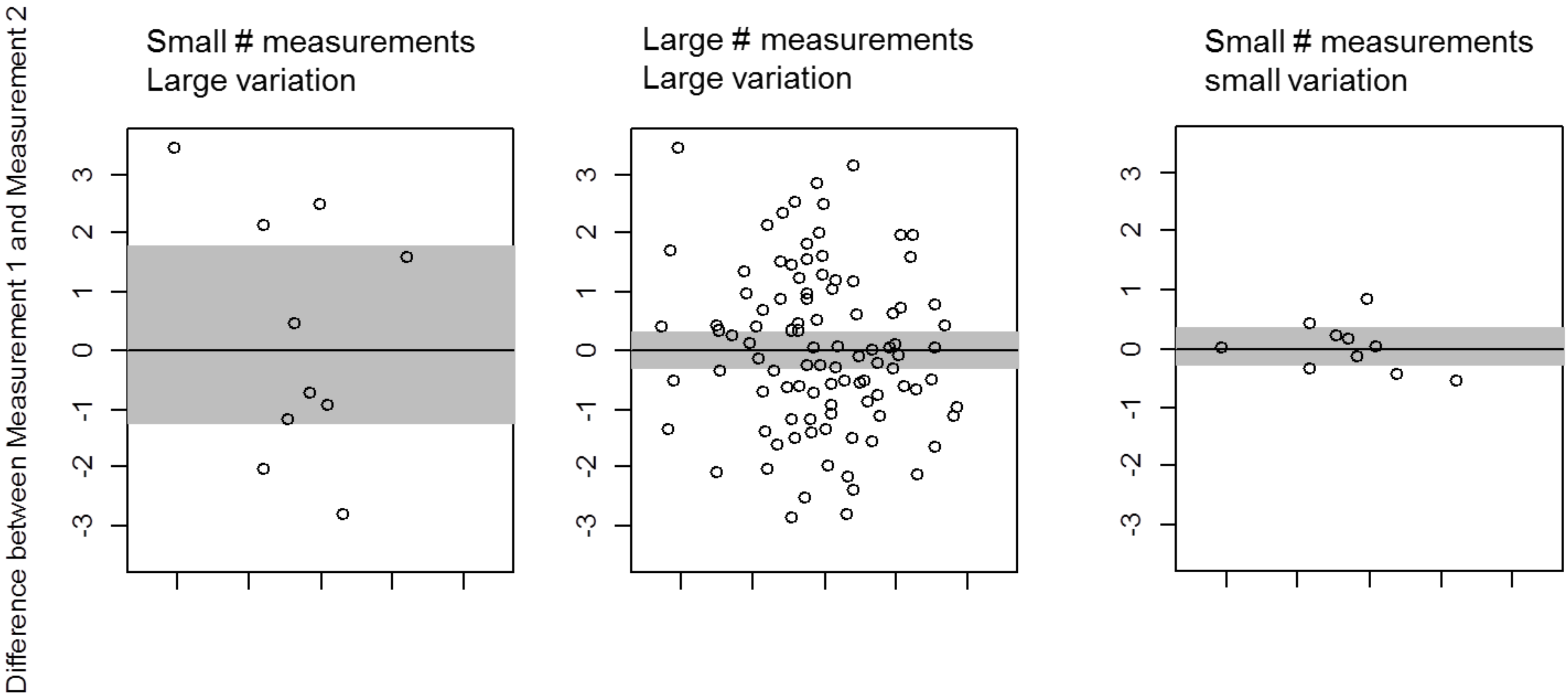
$\bar{d}$ = mean($Y_1 - Y_2$) = - 0.3 mm, s = sd($Y_1 - Y_2$) = 1.6 mm,

95%-CI for difference [-0.9,0.3] indicates no systematic bias between measurements.

Evaluating the difference

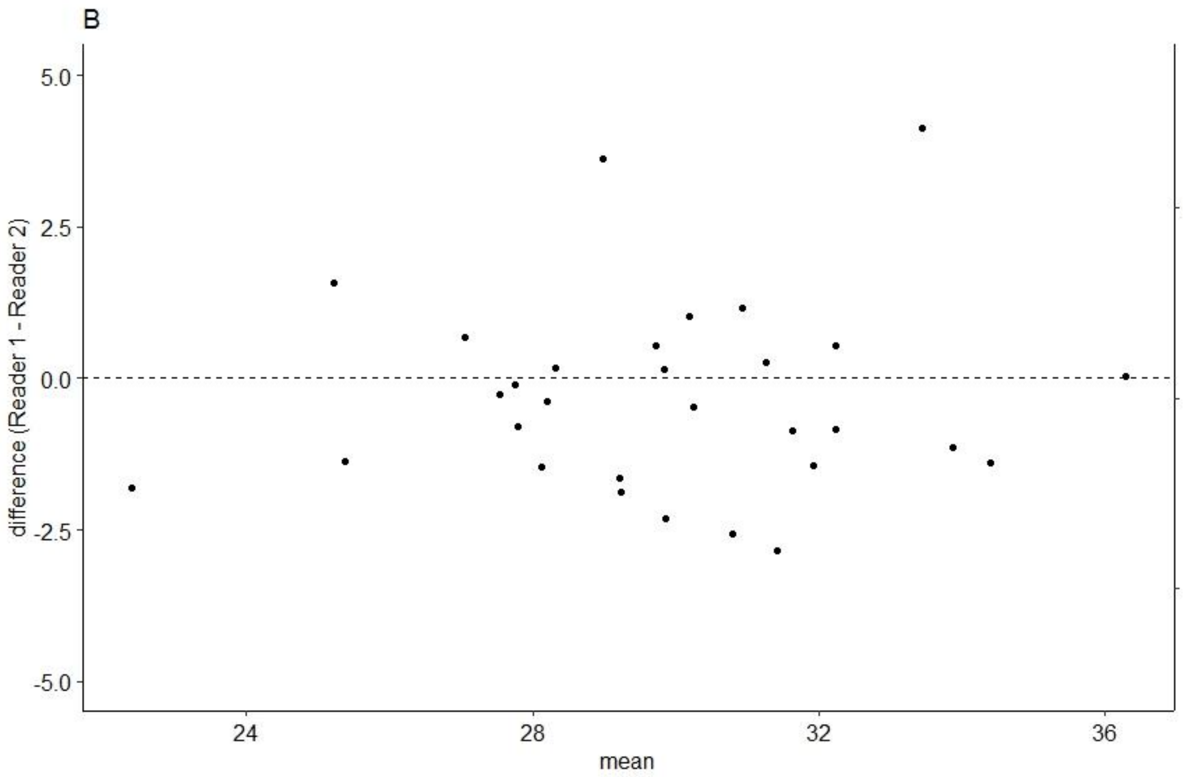
How to interpret Confidence interval?

Artificial examples:



Plotting the difference versus average: Bland-Altman Plot

Idea: Plot Difference of measurements vs. Average of measurements

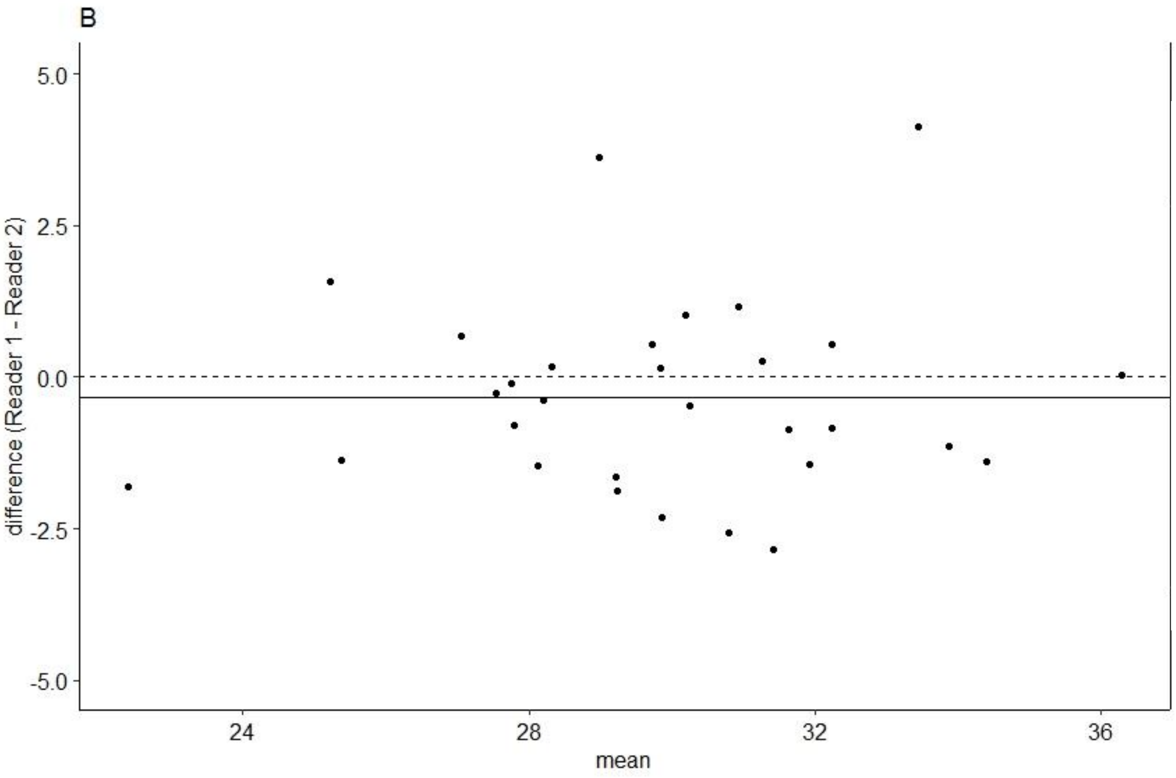


Aim: Identify outliers, identify trends

Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **8476**, 307-10.

Plotting the difference versus average: Bland-Altman Plot

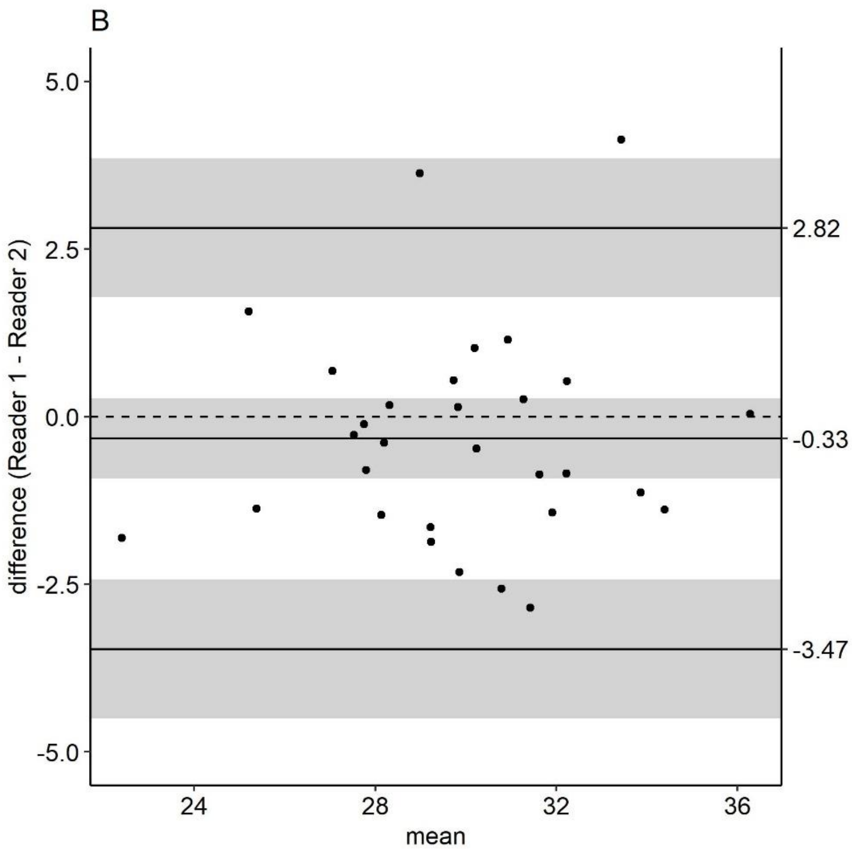
Idea: Plot Difference of measurements vs. Average of measurements, indicate mean difference



Mean difference $\bar{d}$

Plotting the difference versus average: Bland-Altman Plot

Idea: Plot Difference of measurements vs. Average of measurements
Indicate mean difference
Show upper and lower limits of agreement (LoA). Grey zones: 95%-CIs.

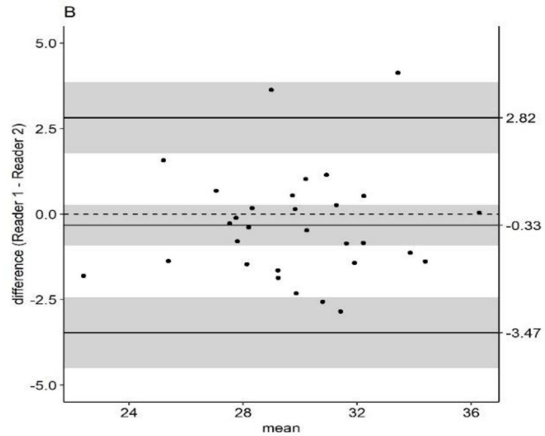


Upper limit of agreement $\bar{d} + 1.96\,s$

Mean difference $\bar{d}$

Lower limit of agreement $\bar{d} + 1.96\,s$

Plotting the difference versus average: Bland-Altman Plot



Upper limit of agreement $\bar{d} + 1.96\ s$

Mean difference $\bar{d}$

Lower limit of agreement $\bar{d} + 1.96\ s$

Interpretation of LoA:

Limits define a range within which most differences between measurements will lie.

Justification:

For normally distributed measurements from N(μ, σ^2):

95% measurements are expected in the range $\mu \pm 1.96\ \sigma$

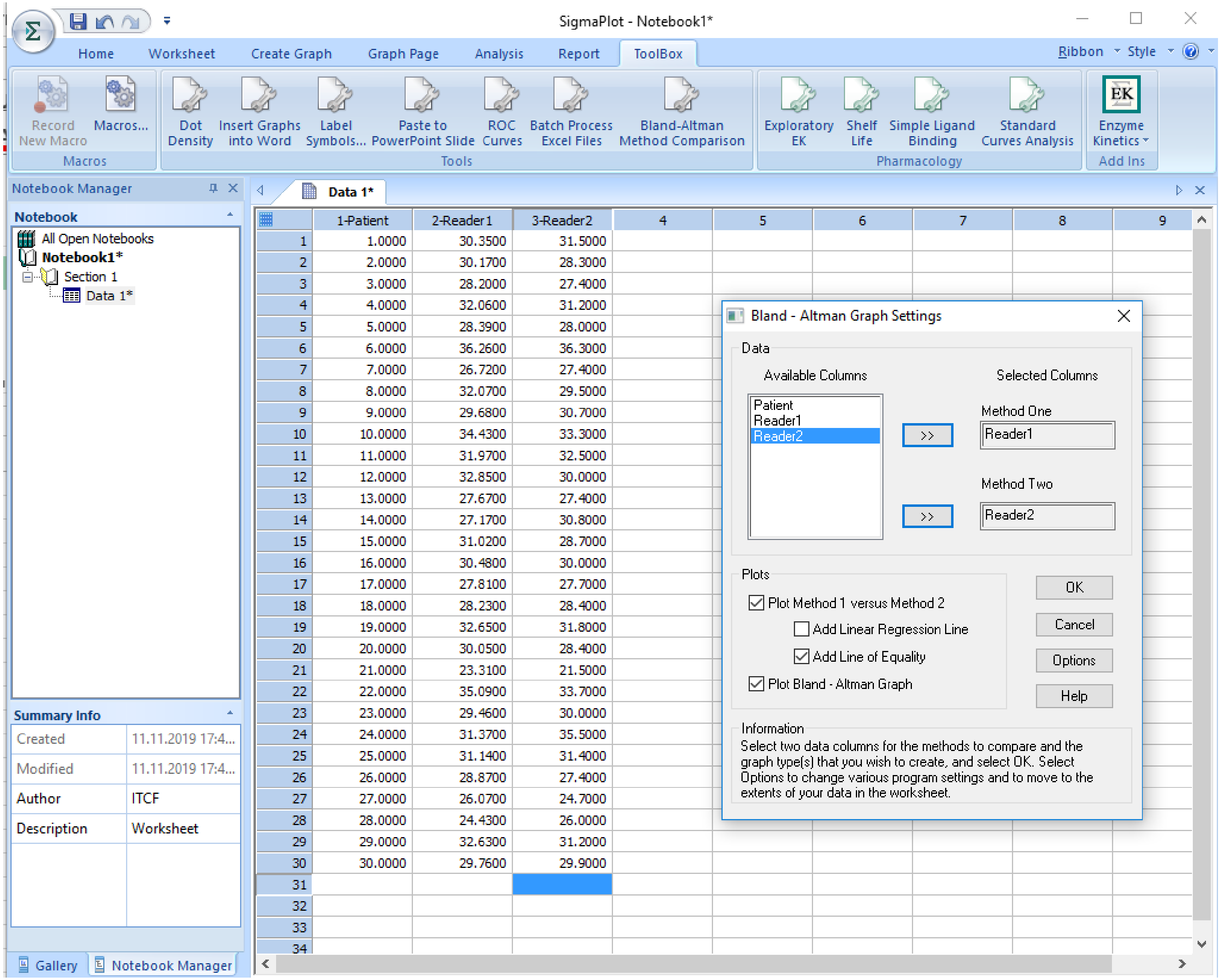
Note:

The decision whether the LoA are acceptable is left to the investigator, it is not a decision for the statistician!

If LoA are clinically acceptable, the measurement methods can be used interchangeably.

Bland-Altman plot with SigmaPlot

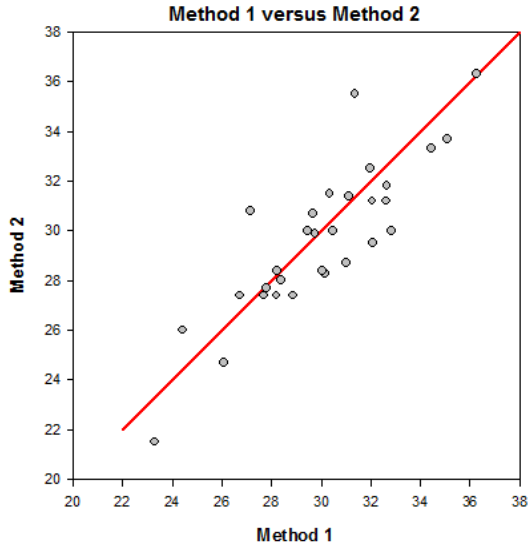
Quantitative outcome

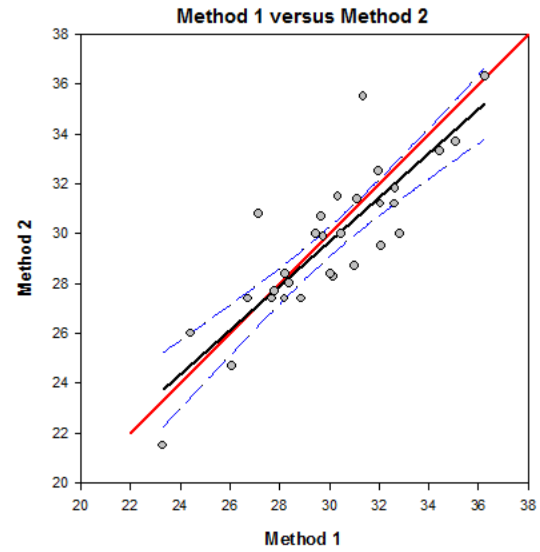


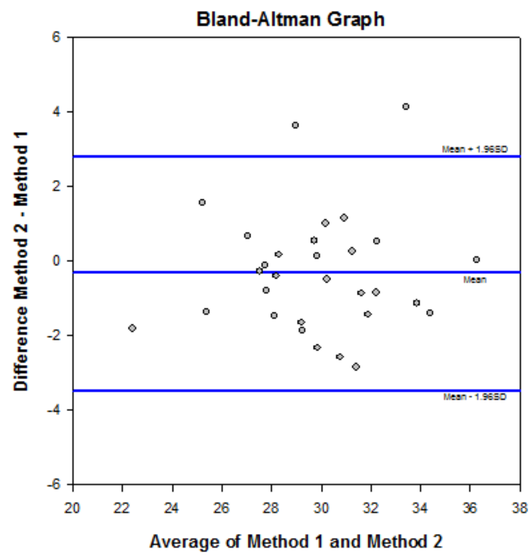
dkfz.

Bland-Altman plot with SigmaPlot

Quantitative outcome







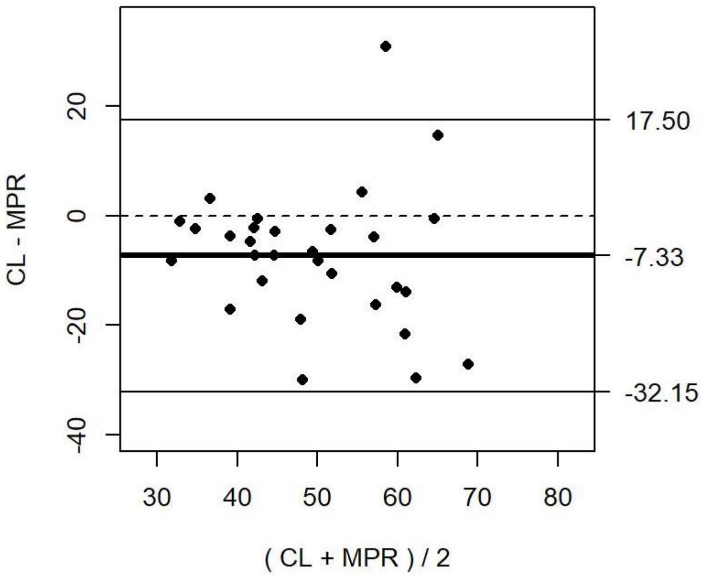
Bias = -.3253
Std Dev = 1.6029
Limits of Agreement = -3.4669, 2.8163
Bias CI
95% CI = -0.9247 To 0.274
Lower Limit of Agreement CI
95% CI = -4.505 to -2.4289
Upper Limit of Agreement CI
95% CI = 1.7782 to 3.8543

dkfz.

Bland-Altman Plot

What if size of difference changes with size of average?

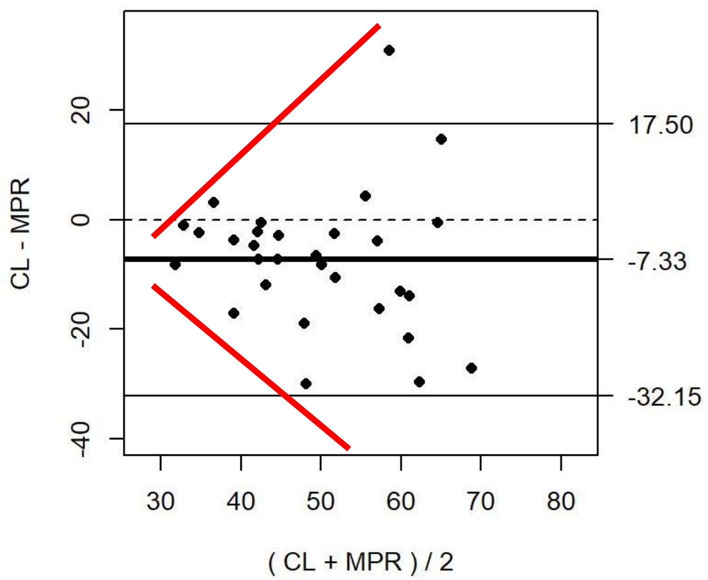
Reader 1, CL and MPR at P4



Bland-Altman Plot

What if size of difference changes with size of average?

Reader 1, CL and MPR at P4

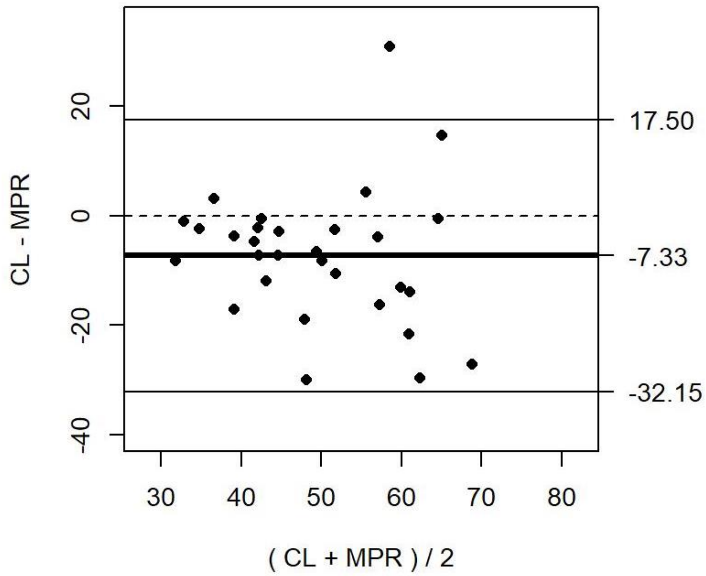


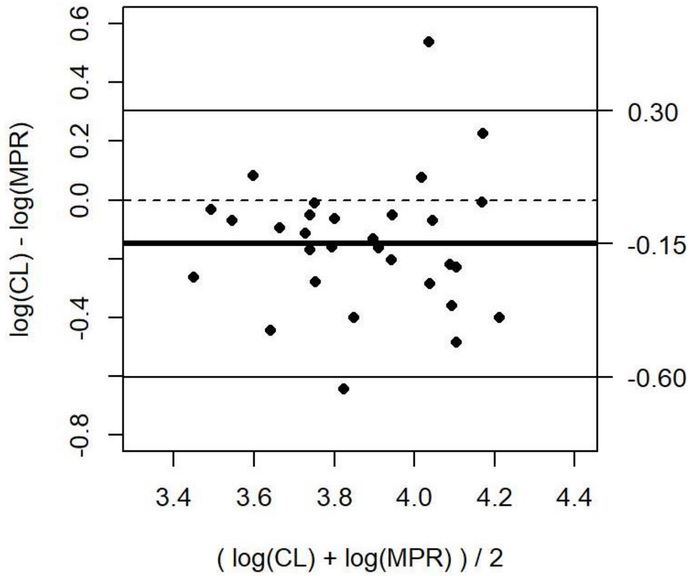
Bland-Altman Plot

What if size of difference changes with size of average?

Reader 1, CL and MPR at P4

→ log-transformation of measurements

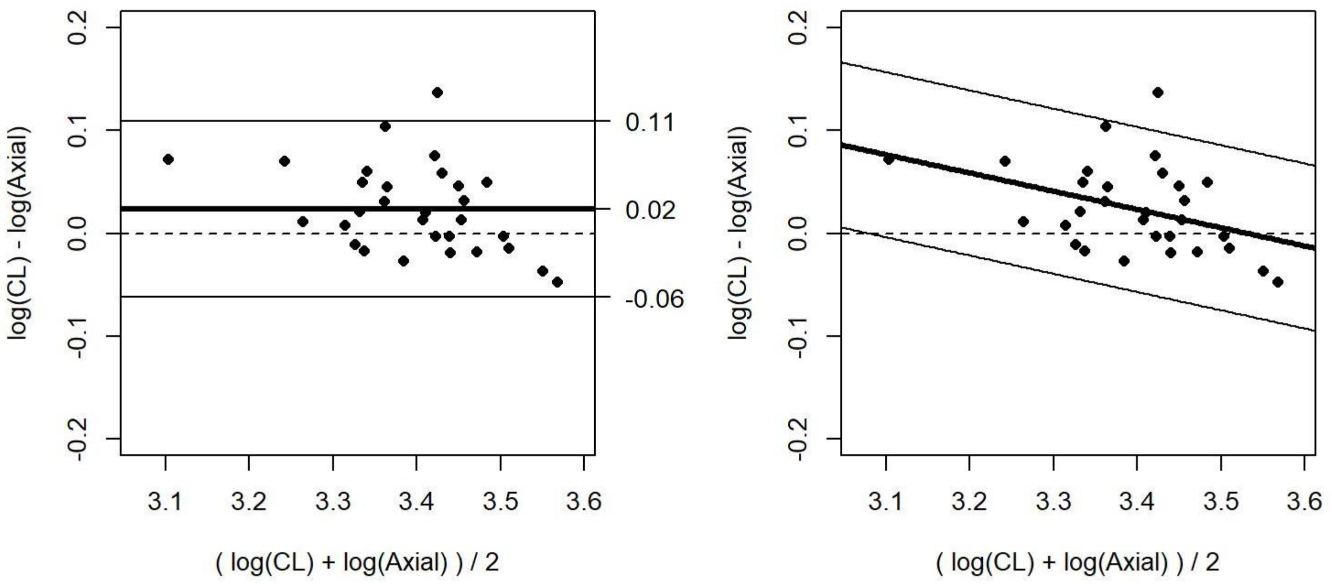




Bland-Altman Plot

What if size of difference changes with size of average?

Reader 1, Axial and CL at P1



Trend indicates that Axial has more variability than CL.

Bland-Altman Plot: Extensions

- Until now: one measurement per subject and method/rater.
- Sometimes multiple measurements per subject:
  - equal number of replicates
  - different number of replicates
  - paired replicates
- Bland-Altman Plots can be drawn for multiple measurements per subject.
- Limits of Agreement can be derived.

Alternative Intervals

- LoA are not suited as interval to predict where a future observation will lie, this would be a **prediction interval**, limits of this:

$$\bar{d} \pm t_{n-1,0.975}\sqrt{(n+1)/n}\ s$$

- **Tolerance interval**: Interval in which 95% of the future observations will lie with 95% probability.

Intervals symmetric around 0:

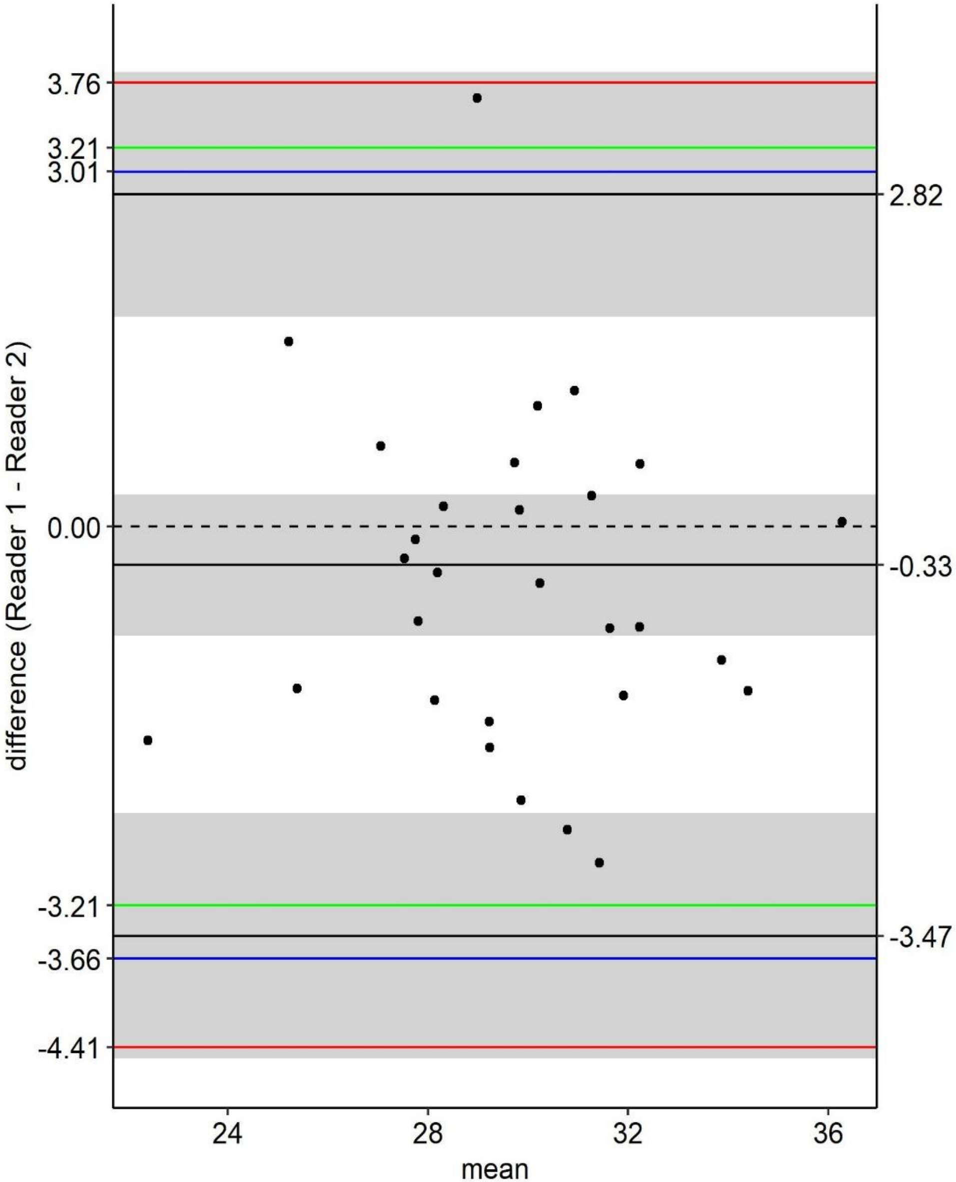
- **Total Deviation Index**, e.g. $TDI_{0.95}$:

$$P(|Y_1 - Y_2| < TDI_\pi) = \pi$$

- **Coverage Probability**:

$$P(|Y_1 - Y_2| < d) = CP_d$$

e.g. $CP_3 = 0.92$

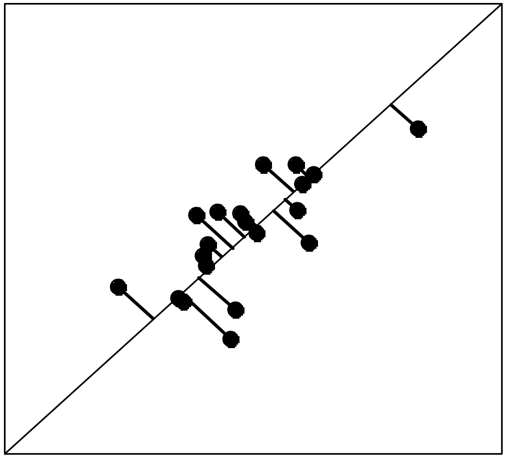


Francq and Govaerts (2016), Lin (2000)

Scaled indices:
Concordance Correlation Coefficient (CCC)

Summarize agreement in one number:

$$CCC = 1 - \frac{\text{Expected squared perpendicular deviation from } 45^\circ \text{ line}}{\text{Expected square perpendicular deviation from } 45^\circ \text{ line if uncorrelated}}$$



CCC=1: complete agreement, CCC=0: no agreement

Lin LI (1989)

Scaled indices:
Intraclass Correlation Coefficient (ICC)

Summarize agreement in one number:

- ICC based on ANOVA model
- assumes identical precision of the methods
- Different ICCs exist: identify the one that reflects the research question
- If interest is in the interchangeability of measurements

  → use ICC measuring the so-called absolute agreement rather than just consistency, which ignores systematic shifts between raters.

Comparison of several approaches

Dependence on variability of subjects

	Full data set (n=30)	**Restricted data set (n=18, ‚middle‘ 60%)**
Standard deviation of differences d	1.60	1.58
95%-Limits of Agreement (bias)	-3.47;2.82 (-0.33)	-3.56;2.62 (-0.47)
95%-CIs of LoAs	LL: -4.50;-2.43 UL:1.78;3.85	LL:-4.92;-2.20 UL:1.26;3.97
95%-Prediction interval	-3.66;3.01	-3.89;2.94
95%-Tolerance interval (with 95% confidence)	-4.41;3.76	-4.91;3.97
TDI with π = 95% (95%-CI)	3.21 (2.30;4.05)	3.23 (2.14;4.19)
CP at d_{max}=3	92%	91%
Correlation r (95%-CI)	0.86 (0.72;0.93)	0.49 (0.03;0.78)
CCC (95%-CI)	0.85 (0.72;0.93)	0.45 (0.04;0.74)
ICC (95%-CI)	0.86 (0.72;0.93)	0.47 (0.04;0.76)

Summary
Categorical outcome

- Explore degree of agreement between raters/measurement methods on the items
- Cohen‘s kappa
  - Derive κ and 95%-CI
  - Assessment of Kappa value on a heuristic scale
- for ordinal ratings: weighted Kappa
- for more raters: Fleiss‘ Kappa

Summary
Quantitative outcome

- Always plot the data!
- Bland-Altman Plot
  - Identify extreme differences and shapes/ trends
  - Identify bias by 95%-CI for mean difference
  - LoA to assess whether agreement is acceptable
  - Extensions to multiple measurements per subject
- Various other intervals exist:
  - Prediction interval
  - Tolerance interval
  - Symmetric around 0: Total Deviation Index, Coverage Probability
- Scaled indices:
  - Concordance Correlation Coefficient (CCC)
  - Intraclass Correlation Coefficients (ICC)
  - careful: CCC and ICC values depend on variability of subjects

    → comparison between different data sets is difficult!

Software

Kappa:

- GraphPad
- http://vassarstats.net/kappa.html
- R packages psych and irr

Bland-Altman Plot + some further analyses

- SigmaPlot
- GraphPad
- MedCalc
- R package MethComp
- R package biostatUZH

References

- Bland JM, Altman DG (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* Feb 8;1:307-10.
- Francq BG, Govaerts B (2016). How to regress and predict in a Bland–Altman plot? Review and contribution based on tolerance intervals and correlated-errors-in-variables models. Statistics in Medicine 35: 2328-2358.
- Kopp-Schneider A, Hielscher T (2019). How to evaluate agreement between quantitative measurements. *Radiotherapy and Oncology* 141:321-326.
- Kwiecien R, Kopp-Schneider A, Blettner M(2011). Concordance analysis: part 16 of a series on evaluation of scientific publications. *Dtsch Arztebl Int.* Jul; 108(30):515-21.
- Lin LI (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 45; 255-268.
- Lin L (2000). Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. *Statistics in Medicine* 19: 255-270.

Next lextures

Change in schedule

Thomas Hielscher

18 November: Survival analysis: Kaplan-Meier, logrank

25 November: Survival analysis: Cox PH regression

Dr. Diana Tichy:

2 December: Diagnostic tests