Clinical Agreement

Dr Wan Nor Arifin
Biostatistics and Research Methodology Unit, Universiti Sains Malaysia.
wnarifin@usm.my

Updated: 27 Mar 2023

Agreement

- Applicable in following 3 conditions (Abd Aziz, 2013):

    1. Interrater agreement
    2. Intrarater agreement
    3. Test-retest agreement

Interrater agreement

- To what extend **different raters/assessors** agree on **measurement values** of a **stable phenomenon** of **same subject**, using **same tool** at **one particular time**.
- A subject should have same value on a particular stable phenomenon (e.g. weight, height, disease status etc.), although measured by different raters within relatively short period of time.
- Abu with true SBP of 120mmHg → Staff nurse A, B, C should report same blood pressure when checking Abu's blood pressure, otherwise they are not in agreement to each other.
- One X-Ray film with bone fracture → Radiologist A and B should report that there is bone fracture, otherwise they contradict each other.
- It is particularly very important in clinical, as we want different doctors, staff nurses and medical personnel to agree on something of clinical interest (e.g. fracture/no fracture, blood pressure value etc.), otherwise clinical practice could be jeopardized. *It would be horrible to think of doctors arguing whether you have broken your bone or not while you writhe in pain on bed.
- **Reliability** is "the extent to which repeated measurements of a stable phenomenon – by different people and instruments, at different times and places – get similar result" (Flether, Flether and Wagner, 1996).
- As we are assessing *agreement*, we want to determine their *reliability*.
- *Interrater agreement → Interrater reliability.*

Intrarater agreement

- To what extend **same rater** agrees (consistent) on **repeated measurement values** of a **stable phenomenon** of **same subject**, using **same tool** at **different time**.
- A subject should have same value on a particular stable phenomenon (e.g. weight, height, disease status etc.), although measured by same rater repeatedly at different times.
- Ali that actually weight 80kgs, should be consistently being recorded as weighing 80kgs by same staff nurse when assessed, let say four times in a day.
- Similarly, a doctor should report an X-ray film with similar finding although being asked a number of times for confirmation.
- Recall the component in our definition of reliability, "the extent to which repeated measurements of a stable phenomenon … at different times ... get similar result" (Flether, Flether and Wagner, 1996).
- *Intrarater agreement → Intrarater reliability*.

Test-retest agreement

- It is concerned with the tool itself, e.g. questionnaire.
- To what extend **same tool** agrees (consistent) on **repeated measurement values** of a **stable phenomenon** of **same subject**, at **different time** (usually 7 days to 14 days, depending on stability of phenomenon of interest).
- It is justified that if the questionnaire is reliable, the answers/scores should be similar consistent from time-to-time, if the responses are expected to tap into stable phenomena, e.g. gender, ethnicity, personality etc. *As a matter of fact, the researcher would be surprised if someone had a change of gender within 1 – 2 weeks time.
- As an example, someone who scores 65% on a personality test should obtain the same score after 1 week gap.
- Recall the component in our definition of reliability, "the extent to which repeated measurements of a stable phenomenon … at different times ... get similar result" (Flether, Flether and Wagner, 1996).
- *Test-retest agreement → Test-retest reliability*.

True Score Theory

- Observed reading/score is thought to be made of true reading and error.

  Observed reading = True reading + Error

  $X = T + e_x$

  in another way:

  Variance of observed reading = Variance of true reading + Variance of error

  $VAR(X) = VAR(T) + VAR(E_x)$

Theory of Reliability

- Going back to our *true score theory*, reliability is defined as:

$$\text{Reliability}\left(\rho_{XX}\right)=\frac{\text{T}}{\text{X}}=1-\frac{e_X}{\text{X}}$$

  or in term of variability:

$$\rho_{XX}=\frac{\text{VAR(T)}}{\text{VAR(X)}}=1-\frac{\text{VAR}(\text{E}_x)}{\text{VAR(X)}}$$

- Ranging from 0 (totally unreliable) – 1 (perfect reliability)

Agreement analysis

- Agreement on:
  - → Numerical data
    - → **Intraclass correlation**
    - → *Bland-Altman plot*
    - → Pearson's correlation
  - → Categorical data
    - → **Kappa** (**Unweighted**, *weighted, Fleiss's kappa*)
    - → Intraclass correlation (ordinal data)

Definition

- For numerical data
- Intra – within; class – same metric/measurement scale.
- Was developed by Fisher to describe repeated measurements on *same* variable (Streiner & Norman, 2008), called as *intra*class correlation vs *inter*class correlation (Pearson's correlation).
- **Height** ↔ **Height** vs **Height** ↔ *Weight*
- Basic of ICC → ANOVA!

Cases of ICCs **not crime cases.*

- Cases of ICCs (McGraw & Wong, 1996) are determined by combinations of the following factors, which also determine formula used, interpretation and application of the ICCs:

  Model
  - → One-way: One factor – Row effect.
  - → Two-way: Two factors – Row and column effects.

  Effect
  - → Random: Subject/Row – Random, Rater/Column – Random.
  - → Mixed: Subject/Row – Random, Rater/Column – Fixed.

  Measurement
  - → Single: Reliability of a measurement from any rater is of concern.
  - → Average: Reliability of average of all ratings from raters is of concern.

  Type
  - → Consistency: Consistency in giving rating. As long as the ratings by different raters are in similar direction (positive, negative) then the reliability would be high, although the ratings given are totally different.
  - → Absolute agreement: Absolute match/agreement between ratings is of concern.

Case 1 ICC – One-way random model

- The simplest case of all, but could be confusing at times. The following factors are applicable to this case:

  Effect
  → Subject (usually termed as row effect) is random. Rater is not applicable here, hence one-way as the is only one factor of concern, which is the subject.

  Measurement
  → Single
  → Average

  Type
  → Absolute agreement
- The data would look this way:

Subject	Rating 1	Rating 2	...	Rating k
1				
2				
...				
n				

* For the subsequent formulas, the following notations are used,

MS_R = mean square for rows
MS_W = mean square for residual sources of variance (within rows)
MS_E = mean square error
MS_C = mean square for columns
σ^2_r = row variance
σ^2_w = within variance
σ^2_e = error variance
σ^2_c or θ^2_c= column variance

Case 1 ICC(1) – One-way random model, single measure:

- Scenario: On entry interview to medical school, applicants performance are rated by 10 groups of lecturers, 3 lecturers each. The applicants may be rated by any of the group, so that no applicants are rated by all of the lecturers, and no lecturers rated all of the applicants. How reliable is performance rating for one applicant?

  * at least two rating per subject i.e repeated measures.
  * regardless of raters.

- Formula: $\frac{\sigma_r^2}{\sigma_r^2+\sigma_w^2}=\frac{MS_R-MS_W}{MS_R+(k-1)MS_W}$

- Context: **Single** performance rating for an applicant is reliable.

Case 1 ICC(k) – One-way random model, average measure:

- Scenario: How reliable is average performance rating for one applicant?

- Formula: $\frac{\sigma_r^2}{\sigma_r^2+\sigma_w^2/k}=\frac{MS_R-MS_W}{MS_R}$

- Context: **Average** performance rating for an applicant is reliable.

Case 2 – Two-way random model

- The following factors are applicable to this case:

  Effect
  → Both subject and rater are random.
  Measurement
  → Single
  → Average
  Type
  → Consistency
  → Absolute agreement

- Data would look this way:

Subject	Rater 1	Rater 2	...	Rater k (random)
1				
2				
...				
n				

Case 2 ICC(C, 1) – Two-way random model, consistency, single measure:

- Scenario: Scenario: On entry interview to medical school, applicants performance are rated by 1 group of 5 lecturers, which are representative sample of all lecturers in the university. All 5 lecturers rated all applicants. How consistent is the rating given by a lecturer from that university?

- Formula: $\frac{\sigma_r^2}{\sigma_r^2+\sigma_e^2}=\frac{MS_R-MS_E}{MS_R+(k-1)MS_E}$

- Context: Rating by **a lecturer** from the **university** is reliable. Any one of lecturer's rating is reliable and can be trusted on its own in rating an applicant.

Case 2 ICC(C, k) – Two-way random model, consistency, average measure:

- Scenario: On entry interview to medical school, applicants performance are rated by 1 group of 5 lecturers, which are representative sample of all lecturers in the university. All 5 lecturers rated all applicants. How consistent is the average rating given by a group of 5 consisting of lecturers from that university?

- Formula: $\frac{\sigma_r^2}{\sigma_r^2+\sigma_e^2/k}=\frac{MS_R-MS_E}{MS_R}$

- Context: **Average** rating by the **group of 5 lecturers** from the **university** is reliable.

Case 2 ICC(A, 1) – Two-way random model, absolute agreement, single measure:

- Scenario: On entry interview to medical school, applicants performance are rated by 1 group of 5 lecturers, which are representative sample of all lecturers in the university. All 5 lecturers rated all applicants. To what extend is the rating given by a lecturer from that university agrees with each other?

- Formula: $$\frac{\sigma_r^2}{\sigma_r^2+\sigma_c^2+\sigma_e^2}=\frac{MS_R-MS_E}{MS_R+(k-1)MS_E+\frac{k}{n}(MS_C-MS_E)}$$

- Context: Rating given by **a lecturer** from the **university** is reliable and in agreement with others. So any lecturer can be chosen and expected to give similar rating for a given applicant.

Case 2 ICC(A, k) – Two-way random model, absolute agreement, average measure:

- Scenario: On entry interview to medical school, applicants performance are rated by 1 group of 5 lecturers, which are representative sample of all lecturers in the university. All 5 lecturers rated all applicants. How reliable the average rating given a group of 5 lecturers from that university in absolute agreement term.

- Formula: $$\frac{\sigma_r^2}{\sigma_r^2+(\sigma_c^2+\sigma_e^2)/k}=\frac{MS_R-MS_E}{MS_R+\frac{MS_C-MS_E}{n}}$$

- Context: **Average** rating given by a **group of 5 lecturers** from the **university** is reliable and in agreement with others. So group of 5 consisting of lecturers from the university can be chosen and expected to give similar rating for a given applicant.

Case 3 – Two-way mixed model

- Most suitable to many clinical agreement situation. The following factors are applicable to this case:

  Effect
  → Subject is random.
  → Rater is fixed, which means the reliability would be applicable to the same set of raters only, not generalizable to other pool of raters.
  Measurement
  → Single
  → Average
  Type
  → Consistency
  → Absolute agreement
- The data would look this way:

Subject	Rater 1	Rating 2	...	Rater k (fixed)
1				
2				
...				
n				

Case 3 ICC(C, 1) – Two-way mixed model, consistency, single measure:

- Scenario: On entry interview to medical school, applicants performance are rated by 1 group of 5 lecturers. All 5 lecturers rated all applicants. How consistent is the rating given by a lecturer from the group?

- Formula: $$\frac{\sigma_r^2}{\sigma_r^2+\sigma_e^2}=\frac{MS_R-MS_E}{MS_R+(k-1)MS_E}$$

- Context: Rating by **a lecturer** of the **group** is reliable. Any of lecturer's rating is reliable and can be trusted on its own.
- This is the model to be used for test-retest situation (Weir, 2005), which is equivalent to ICC(3, 1) of Shrout and Fleiss (1979).

Case 3 ICC(C, k) – Two-way mixed model, consistency, average measure:

- Scenario: On entry interview to medical school, applicants performance are rated by 1 group of 5 lecturers. All 5 lecturers rated all applicants. How consistent is the average rating given by the group?

- Formula: $$\frac{\sigma_r^2}{\sigma_r^2+\sigma_e^2/k}=\frac{MS_R-MS_E}{MS_R}$$

- Context: **Average** rating by the **group of 5 lecturers** is reliable.

Case 3 ICC(A, 1) – Two-way mixed model, absolute agreement, single measure:

- Scenario: On entry interview to medical school, applicants performance are rated by 1 group of 5 lecturers. All 5 lecturers rated all applicants. To what extend the rating given by a lecturer from the group agrees to each other?

- Formula: $$\frac{\sigma_r^2}{\sigma_r^2+\theta_c^2+\sigma_e^2}=\frac{MS_R-MS_E}{MS_R+(k-1)MS_E+\frac{k}{n}(MS_C-MS_E)}$$

- Context: Rating given by **a lecturer** from the **group** is reliable and in agreement with others in that group. So any lecturer from the group can be chosen and expected to give similar rating for a given applicant.

Case 3 ICC(A, k) – Two-way mixed model, absolute agreement, average measure:

- Scenario: On entry interview to medical school, applicants performance are rated by 1 group of 5 lecturers. All 5 lecturers rated all applicants. How reliable the average rating given by the group in absolute agreement term.

- Formula: $$\frac{\sigma_r^2}{\sigma_r^2+(\theta_c^2+\sigma_e^2)/k}=\frac{MS_R-MS_E}{MS_R+\frac{MS_C-MS_E}{n}}$$

- Context: **Average** rating given by that **group of 5** is reliable as they are in agreement with each others. The same group of 5 lecturers is expected to give reliable for a given applicant. *However if all raters totally agree with each other, column variance θ_c^2 is zero, thus Case 1 ICC(k) should be used (McGraw & Wong, 1996).

  *Cases of ICC with interaction are skipped in this lecture for ease of understanding. Refer to McGraw & Wong (1996) for details.

Standard error and confidence interval for ICC

- As the SE and CI are specific for each of the ICC cases, student is encouraged to refer to McGraw and Wong (1996) paper on ICC. *Hint: I would not ask you to calculate manually SE and CI for ICC.

Interpretation

- The values of ICC ranges from -1 to 1, interpreted similarly to any reliability coefficient.
- It is helpful to interpret the values according to Cichetti (1994) as follows:

ICC value	Strength of agreement
< 0.40	Poor
0.40 – 0.59	Fair
0.60 – 0.74	Good
0.75 – 1.00	Excellent

Result presentation

- Case of ICC used must be stated.
- The effect, measurement and type must also be stated clearly as it affects its interpretation.
- e.g "*Rating by a single lecturer from the university is reliable with Case 2 ICC(C,1) of 0.85 (95% CI: 0.810, 0.890)*".

Hands-on

- Dataset 1: **ICC_BP Lecture.sav**. Consists of 11 subjects and 5 raters. SBP and DBP was measured. Apply all cases of ICCs to the dataset and compare the results.
- R script: **icc & kappa.R**
- R packages: **irr, psych**

Definition

- For categorical data – nominal and ordinal.
- Interrater agreement between two raters/two methods on categorical rating.
- Assessing ability of the raters/methods to classify subjects into different groups (Altman, 1991)
- Also for intrarater and test-retest.
- **Fracture status (Doctor 1) ↔ Fracture status (Doctor 2)**
- **Cancer staging (Pathologist 1) ↔ Cancer staging (Pathologist 2)**
- **HIV status (Rapid test) ↔ HIV status (ELISA)**

Proportion of agreement

Simple index of agreement → Just proportion of exact agreements between the raters.

Table 2(a). Assessment of fracture status from x-ray films by two doctors.

		Doctor 2		
		Fracture	No fracture	Total
Doctor 1	Fracture	**30**	5	35
	No Fracture	15	**30**	45
	Total	45	35	**80**

$$\text{Proportion of agreement} = \frac{\text{sum of observed agreement}}{\text{total}} = \frac{\sum f_{ii}}{n} = \frac{(30+30)}{80} = .75$$

Cohen's Kappa

- For two raters/methods only. Commonly referred only as kappa.
- Percentage of agreement method does not take into account possibility of agreement that can happen by chance. It is possible that if the raters just guess the categories, there could be still some degree of agreement among them.
- Kappa takes into account this chance agreement. Chance agreements are discarded in the calculation.

$$\text{Kappa, } \kappa = \frac{\text{observed proportion of agreement} - \text{expected proportion of agreeement by chance}}{1 - \text{expected proportion of agreeement by chance}}$$

$$= \frac{p_o - p_e}{1 - p_e}$$

Table 2(b). Assessment of fracture status from x-ray films by two doctors with expected frequencies.

		Doctor 2		
		Fracture	No fracture	Total
Doctor 1	Fracture	**30 (19.7)**	5	35
	No Fracture	15	**30 (19.7)**	45
	Total	45	35	**80**

$$p_o = \frac{\text{sum of observed agreement}}{\text{total}} = \frac{\sum f_{ii}}{n} = .75$$

$$p_e = \frac{\text{sum of expected agreement by chance}}{\text{total}} = = \frac{\sum r_i c_i / n}{n}$$

$$= \frac{(35 \times 45)/80 + (45 \times 35)/80}{80} = \frac{19.7 + 19.7}{80} = \frac{39.4}{80} = .49$$

$$\kappa = \frac{p_o - p_e}{1 - p_e} = \frac{.75 - .49}{1 - .49} = \frac{.26}{.51} = .51 \quad \rightarrow \text{Look how misleading proportion of agreement is.}$$

Standard error and confidence interval for κ

- Standard error of κ is given by

$$SE(\kappa) = \sqrt{\frac{p_o(1-p_o)}{n(1-p_e)^2}}$$

- Confidence interval of κ is given by

$$\kappa \pm z_{(1-\alpha/2)} \times SE(\kappa)$$

Interpretation

- The following guidelines by Landis and Koch (1977) is helpful for interpretation of κ:

κ value	**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

Result presentation

- Give κ value together with 95% CI.
- State the interpretation.
- e.g. "*The agreement between Rapid Test Novo with ELISA was k = 0.92 (95% CI: 0.880, 0.960)*".

Hands-on

- Dataset 2: Enter Table 2(a) data into R. Compare your result with hand-calculated result.
- Dataset 3: Enter Table 3 below into R.

Table 3. Assessment of lung infection severity from x-ray films by two doctors.

		Doctor 2			
		mild	moderate	severe	Total
Doctor 1	mild	**44**	4	0	48
	moderate	5	**38**	5	48
	severe	1	2	**21**	24
	Total	50	44	26	**120**

Additional reading

- Do self-study on the following topics:

    1. *Bland-Altman plot for assessment of numerical agreement*
    2. *Ordinal data – Weighted kappa*
    3. *More than 2 raters – Fleiss' kappa*

- You understanding will be assessed via assignment on these topics.

Compulsory reading

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, *8*(1), 23.

References

Abd Aziz, A. (2013). Reliability agreement 1. *2nd Questionnaire validation workshop 2013*. October 7 – 9, MIP Laboratory, Universiti Sains Malaysia, Kubang Kerian, Kelantan, Malaysia.

Altman, D. G. (1991). Practical statistics for medical research. London: Chapman and Hall.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological assessment*, *6*(4), 284-290.

Fletcher, R. H., Fletcher, S. W., & Wagner, E. H. (1996). Clinical epidemiology: the essentials (3rd ed.). Maryland: Williams & Wilkins.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159-174.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1): 30-46.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, *86*(2), 420.

Streiner, D. L. & Norman, G. R. (2008). Health measurement scales: a practical guide to their development and use. New York: Oxford University Press.

Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *The Journal of Strength & Conditioning Research*, *19*(1), 231-240.