South Eur J Orthod Dentofac Res

SEJODR

StaTips Part IV: Selection, interpretation and reporting of the intraclass correlation coefficient

Perinetti, Giuseppe *

* *Private practice, Nocciano (PE), Italy*

ABSTRACT

The intraclass correlation coefficient (ICC) is an index or repeatability that reflects both the degree of correlation and agreement between measurements. The ICC is widely used in orthodontic research for any continuous data set that satisfies assumptions for using the parametric methods. However, the ICC comprises a total of ten different variants not always recognized by researchers, which may give different outcomes. Here, a practical guide to choose the corrected variant of the ICC based on study design, and three different aspects (referred to as 'model', 'type' and 'definition') is provided. Finally, a full example of correct data interpretation and reporting is included.

Perinetti G. StaTips Part IV: Selection, interpretation and reporting of the intraclass correlation coefficient. South Eur J Orthod Dentofac Res. 2018;5(1):3-5.

FRAMING OF THE PROBLEM

In StaTips Part II [1] stress has been given to the repeatability analysis as a fundamental part of clinical research. Similarly, the distinction between systematic and random errors associated with data recording has been given earlier. [1] A very common index of repeatability not yet reported in StaTips is the intraclass correlation coefficient (ICC). [2] Briefly, the ICC is an index that reflects both the degree of correlation and agreement between measurements of continuous data.

The ICC is widely used in orthodontic research for any continuous data set that satisfies assumptions for using parametric methods (see StaTips Part I [3]). For instance, a rater repeating a subset of cephalometric recordings over time represents a case of intra-rater repeatability. On the contrary, when 2 or more raters perform cephalometric analyses on the same subset of subjects under investigation represents a case of inter-rater repeatability. In many of the orthodontic studies, both intra-rater and inter-rater repeatability is reported.

The ICC cannot be used for ordinal data, such as the skeletal maturation stages, or for assessing the interchangeability of different measurement methods/parameters. [4] On the contrary, the ICC may be used in some circumstances for dichotomous data as previously reported [5] (not dealt herein). The ICC exists in ten different variants and it thus constitutes an important piece of the repeatability analysis in the orthodontic field deserving a dedicated article.

Corresponding Author:
Perinetti Giuseppe
Via San Lorenzo 69/1,
65010 Nocciano (PE), Italy.
e-mail: G.Perinetti@yahoo.com

CHOICE OF THE CORRECT INTRACLASS CORRELATION COEFFICIENT

Although the wording ICC may be though to refer to a specific statistical entity, the ICC comprises a total of ten different variants (six described decades ago [2,6] and other four subsequently added [7]) which are not always recognized by researchers, and that may give different outcomes when applied to the very same data set. The reason behind such high number of variants resides in the concept that the ICC has to be flexible to be applied to many different circumstances (for full information, see [6-8]). Therefore, proper use of the ICC begins with the correct choice of the variant for that specific study, which is not a complex procedure.

When selecting the proper ICC, the study design is the first step to consider. For a repeatability issue, this includes three possibilities as follows: 1) test-retest repeatability; 2) intra-rater repeatability; and 3) inter-rater repeatability. As a rule of thumb, it has been reported that preferred repeatability studies should involve at least 30 heterogeneous samples (i.e. subjects, cephalometric recordings and so on) recorded by at least 3 raters (when dealing with inter-rater repeatability). [8] A brief explanation of each of these repeatability studies is

reported in Table 1. Thereafter, it is important to deal with three different aspects strictly related to the way the ICC is calculated.

Table 1. *Definition of the different studies of repeatability.*

Study	Brief explanation
Test-retest repeatability	It refers to the repeatability between consecutive sessions taken by the same instrument under the same conditions (longitudinal recordings). In such a case, raters are not primarily involved.
Intra-rater repeatability	It refers to the repeatability of the same (single) rater recording a given parameter in consecutive sessions under the same conditions (longitudinal recordings).
Inter-rater repeatability	It refers to the repeatability between 2 or more raters recording a given parameter on the same group of subjects under the same conditions (cross-sectional recordings).

These are as follows: 1) model (including one-way random-effect, two-way random-effect and two-way mixed-effect); 2) type (including single rater/measurement and mean of multiple raters/measurements); and 3) definition (including absolute agreement and consistency). A brief explanation of each of these model, type and definition is reported in Table 2. According to the different combinations of the study designs, models, types and definitions, specific variants of the ICC are obtained. In particular, the six ICC variants classified by Shrout and Fleiss [6] and the subsequent four variants reported by McGraw and Wong [7] are reported in Table 3.

Table 2. *Definition of the different criteria used to select the intraclass correlation coefficient.*

Criterium	Brief explanation	Further notes
Model		
One-way random-effect	Each recording made by a different set of raters who were randomly chosen from a population (different groups of subjects involved).	Results extended to the whole populations from which raters are selected (rarely used).
Two-way random-effect	Each recording made by the same set of raters who were randomly chosen from a population (a unique group of subjects involved).	Results extended to the whole population from which raters are selected (most used).
Two-way mixed-effect	Each recording made by the same set of raters who were not randomly chosen from a population (a unique group of subjects involved).	Results limited to the specific set of raters involved (rarely used).
Type		
Single rater/ measurement	Recordings from a single rater (or an instrument, in case of test-retest study) are the basis of the analysis.	-
Mean of multiple raters/ measurements	The mean recording from a set of raters (or a set of measurements from an instrument, in case of test-retest study) is the basis of the analysis.	-
Definition		
Absolute agreement	Repeatability based on the exact same scores among recordings. It takes into account systematic error among raters/recordings.	Recommended when systematic error among raters/recordings is expected to be relevant (more frequent).
Consistency	Repeatability based on the correlation among recordings. It does not take into account systematic error among raters/ recordings.	Recommended when systematic error among raters/recordings is not expected to be relevant (less frequent).

Table 3. *The ten different variants of the intraclass correlation coefficient according to the model, type, and definition criteria.*

Model	Type	Definition	
		Absolute agreement	Consistency
One-way random-effect	Single measurement/rater	A	--
	Mean of multiple measurements/raters	A	--
Two-way random-effect	Single measurement/rater	A	B
	Mean of multiple measurements/raters	A	B
Two-way mixed-effect	Single measurement/rater	B	A
	Mean of multiple measurements/raters	B	A

A, ICC classified by Shrout and Fleiss [6] classification; B, ICC added by McGraw and Wong [7]. --, not existing.

A simplified diagram illustrating the steps in choosing the proper ICC variant is shown in Figure 1. According to this diagram, the selection of the ICC would be very easy in case of test-retest and intra-rater repeatability studies. A slightly more complex situation regards the case of inter-rater repeatability study, while rare cases (such as those using the one-way random-effect model) have been partially omitted to simplify the picture. Finally, the ICC may be interpreted and numbers converted into repeatability outcomes (as fair, moderate, good and excellent) according to pre-determined ranges as reported in Table 4. An average statistical software carries functions to calculate all the ICC variants.

Table 4. *Common guidelines for the interpretation of the intraclass correlation coefficient.*

Repeatability outcome	Intervals	Further notes
	Ko and Li [8]	Cicchetti and Sparrow [9]
Poor	<0.50	<0.40
Fair	0.50-0.75	0.40-0.60
Good	0.75-0.90	0.60-0.75
Excellent	0.90-1	0.75-1

REPORTING OF THE INTRACLASS CORRELATION COEFFICIENT

Considering the existence of different ICC variants, a correct reporting of the study of repeatability should include details regarding the choice of the ICC. Specifically, model, type and definitions considered should be reported (eventually for each of the different ICC used). The 95% confidence intervals are also of importance and should be reported along with the corresponding ICC estimations.

The intervals used to interpret the ICC have to be reported in the methods and confidence intervals have also to be taken into account when interpreting the ICC (Table 4) and accordingly, a range of outcomes, i.e. good to excellent, will be reported. A typical example of how to report the ICC in a scientific paper is provided:

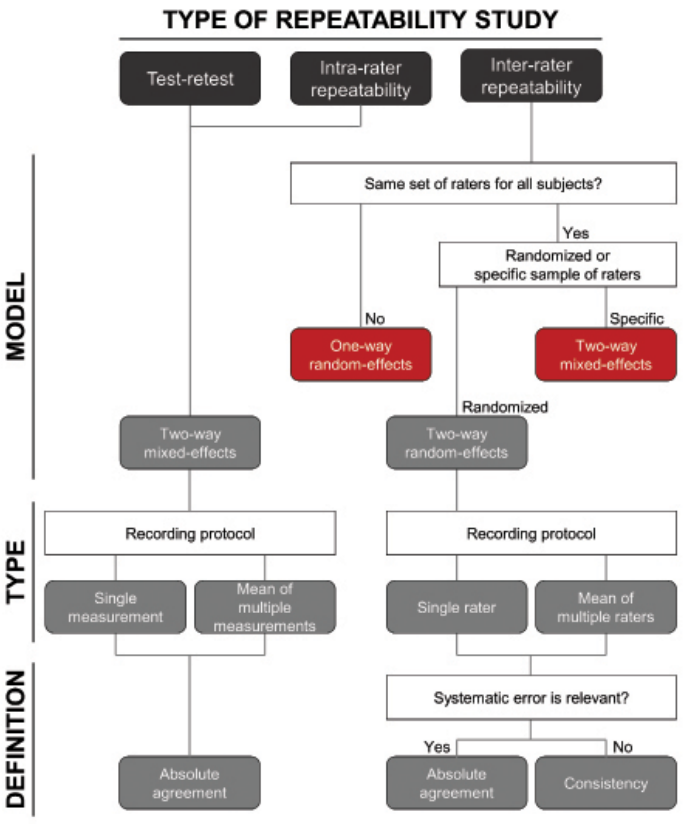
Statistical analysis

Repeatability analysis was performed on a subset of 30 samples randomly chosen and assessed at two different time points by two raters. After having tested the existence of a normal distribution of the data sets, the intraclass correlation coefficient (ICC) was used for the analysis. A two-way random-effect model based on single ratings and absolute agreement assessed the inter-rater repeatability, and a two-way mixed-effect model based on single rating assessed the intra-rater repeatability for either rater. Mean estimations along with 95% confidence intervals (CI) were reported for each ICC. Interpretation was as follows: <0.50, poor; between 0.50 and 0.75, fair, between 0.75 and 0.90 good; above 0.90, excellent.

Results

The ICC for inter-rater reliability was between fair and excellent being 0.85 (0.66-0.94), the ICC's for the intra-rater repeatability was between good and excellent being 0.92 (0.89-0.98) and 0.88 (0.72-0.95) for raters 1 and 2, respectively.

Figure 1. *A simplified diagram illustrating the selection process of the intraclass correlation coefficient for the most frequent cases in clinical research.*



In red, models rarely used (without further information provided on the diagram). Adapted from Bartko[2] and Koo and Li[8].

CONFLICT OF INTEREST

The author declares no conflict of interest.

REFERENCES

1. Perinetti G. StaTips Part II: Assessment of the repeatability of measurements for continuous data. South Eur J Orthod Dentofac Res. 2016;3(2):33-34.
2. Bartko JJ. The intraclass correlation coefficient as a measure of reliability. Psychol Rep. 1966;19(1):3-11.
3. Perinetti G. StaTips Part I: Choosing statistical test when dealing with differences. South Eur J Orthod Dentofac Res. 2016;3(1):4-5.
4. Bland JM, Altman DG. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. Comput Biol Med. 1990;20(5):337-40.
5. Fleiss JL. Measuring agreement between two judges on the presence or absence of a trait. Biometrics. 1975;31(3):651-9.
6. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979;86(2):420-8.
7. McGraw KO, Wong SP. Forming Inferences about some intraclass correlation coefficients. Psychological Methods. 1996;1(1):30-46.
8. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15(2):155-63.
9. Cicchetti DV, Sparrow SA. Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. Am J Ment Defic. 1981;86(2):127-37.