CURRENT CONCEPTS REVIEW

Methodology for Online Reliability Studies: A Primer for Orthopedic Surgeons

Femke M.A.P Claessen, MD PhD[1]; Ruben Zwiers, MD[2]; Thierry G. Guitton, MD, PhD[3]; Job N. Doornberg, MD, PhD[4]

Research performed at the Orthotrauma Research Center in Amsterdam, Amsterdam, the Netherlands

Received: 22 November 2019 *Accepted: 17 October 2022*

Abstract

In orthopedic surgery, there is an increasing number of papers about online studies on the reliability of classification systems. Useful classification systems need to be reliable and valid. Measurement of validity can be variable and is prone to observer bias.

These online collaboratives derive adequate power to study reliability by having a large group of trained surgeons review a small number of cases instead of the "classic" reliability studies in which a small number of observers evaluate many cases. Large online interobserver studies have advantages (i.e., less than 15 minutes to complete the survey, the ability to randomize, and the ability to study factors associated with reliability, accuracy, or decision-making). This 'handbook' paper gives an overview of current methods for online reliability studies. We discuss the study design, sample size calculation, statistical analyses of results, pitfalls, and limitations of the study design.

Level of evidence: V

Keywords: Interobserver studies, Methods, Reliability

Introduction

Classification systems can help surgeons characterize the nature of clinical problems, indicate a potential prognosis, provide guidance for optimal treatment decision-making, establish an expected outcome for a natural history of a condition or injury, and offer a standardized approach to reporting, documenting, and comparing results from clinical and epidemiological data. [1,2]

Useful classification systems need to be reliable and valid (accurate). [1] Measurement of validity can be variable and prone to observer bias. Thus, studies often tend to focus on establishing reliability (precision) at the outset and as a minimum requirement – often focusing on intra-observer (agreement between repeated observations by one observer) and inter-observer (agreement between different observers) reliability. Classification systems should have substantial to a good interobserver agreement to be useful in clinical practice. Since the level of agreement can be highly variable, it is important to understand the approach and sources of bias and disagreement.

This 'handbook' paper gives an overview of current methods for online reliability studies. We discuss the study design, sample size calculation, statistical analyses of results, pitfalls, and limitations of the study design. This 'handbook' paper is a systematic overview of the methodological framework for online reliability studies [Figure 1]. These studies are performed mainly by (surgeon) researchers.

Study Design

In "classic" reliability studies, a small number of observers (around six or so) evaluate many cases (usually more than 50). In contrast, online collaboratives derive adequate power to study reliability (e.g., COAST collaborative and the Science of Variation Group

Corresponding Author: Femke M.A.P. Claessen, Haaglanden Medisch Centrum, Leidschendam, Orthotrauma Research Center Amsterdam, Amsterdam, the Netherlands
Email: femke__claessen@hotmail.com

THE ONLINE VERSION OF THIS ARTICLE ABJS.MUMS.AC.IR

(SOVG)) by having a large group of trained surgeons (more than 60) review a small number of images or patient scenarios. [3] The advantages of these large online interobserver studies are that the survey can be completed within 15 minutes and the ability to randomize and study factors associated with reliability, accuracy, or decision-making.

To have a large database of observers, the SOVG recruited trained surgeons from all over the world. Furthermore, new independent members through societies of different specialties –i.e., Orthopedic Trauma (www.traumaplatform.org), Shoulder and Elbow platform (www.shoulderelbowplatform.org) and Foot & Ankle surgeons (www.ankleplatform.com) were included.

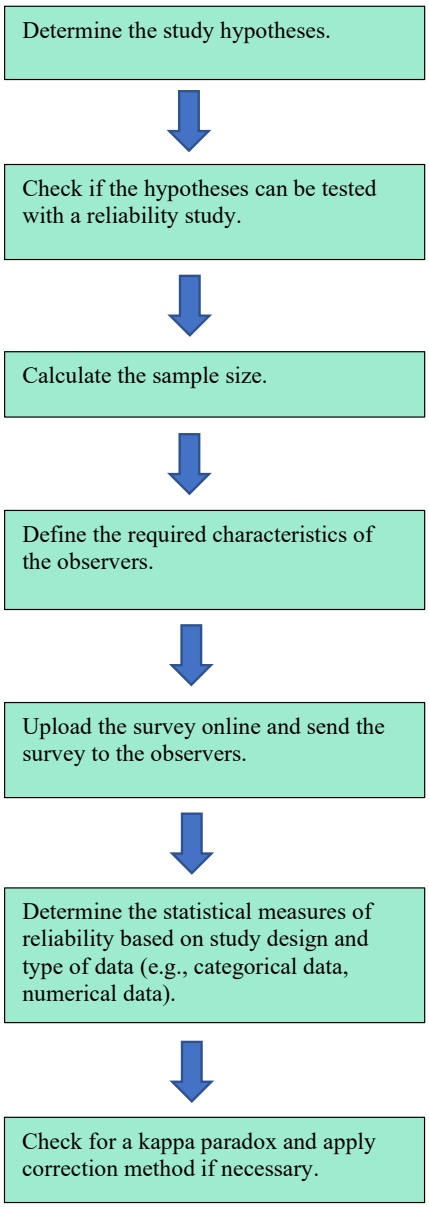

Figure 1. Flowchart of the methodological process for reliability studies

With the introduction of online reliability studies with many observers, new designs for interobserver studies were developed. Depending on the number of observers and cases, an appropriate design has to be chosen. In most studies, all measurements or images will be rated by all observers, a factorial or fully crossed design. In studies with many cases or multiple measurements or images per patient and/or where the rating is time-consuming, selecting a subset of observers that rates only a subset of measurements/images can be more practical and efficient. When this latter option is chosen, it must be decided whether the measurements or images of a single patient will be rated by the same set of observers or whether the assessments are completely random. These split-plot designs are frequently used in clinical imaging studies. [4]

Example

To identify the optimal projection angle to visualize the posterior aspect of the talus, radiographs of 13 different projection angles of 40 ankles were used. Because of the large number of images, a split-plot design was chosen. The images were distributed using a web-based rating application following an eight-block balanced split-plot design. Each block consisted of radiographs of 13 different projection angles of five different patients. Images were randomly assigned to the eight blocks, and each observer was randomly assigned to one of the blocks. Thus, each observer assessed 65 (13 x 5) images.

Online Presentation of Cases

Previously the SOVG used youtube.com and surveymonkey.com, which comprised movies of MRIs and 2-dimensional (2D)/ 3-dimensional (3D) CTs to present the cases. New features, such as a web-based viewer to facilitate online window leveling and contrast enhancement to improve the visualization of radiographs, were included via www.dicomlibrary.com. In addition, a digital imaging and communications in medicine (DICOM) viewer facilitates scrolling through MRI and CT images.

After logging in to the website, observers had to provide demographic and professional information: (1) sex, (2) location of practice, (3) years in independent practice, (4) supervising trainees in the operating room, (5) number of fractures/injuries treated per year, and (6) subspecialty. [5] Usually, observers are unaware of the hypothesis to reduce the chance of observer bias.

Primary and Secondary Hypotheses

Online interobserver studies can test hypotheses regarding classification systems, fracture/injury characteristics, radiographic methods, and surgical techniques. The null hypothesis should test no difference because it can be rejected if the P-value is smaller than

0.05. If the number of observers is high enough, it is possible to test a secondary hypothesis.

Example

We addressed the following study question: is the Minami, Berndt and Harty, Ferkel and Sgaglione, or Anderson Classification system more reliable for classification of osteochondritis dissecans of the humeral capitellum.[6-9] Our null hypothesis was that there is no difference in reliability of the Minami-, Berndt and Harty-, Ferkel and Sgaglione- and Anderson classification systems for the classification of osteochondritis dissecans of the humeral capitellum. [10]

If the number of observers is high enough, it is possible to test a secondary hypothesis. Subgroup analysis can be performed. For example, suppose the primary null hypothesis assesses no agreement between surgeons on which implants will loosen or break after surgery for a distal humerus fracture in the subgroup analysis. In that case, no difference in interobserver agreement according to subgroups (e.g., experience surgeons, location of practice) can be tested. [11]

Recommended Sample size

Although reliability studies are common, the literature on sample size estimation for interobserver reliability studies is limited. [12-17] In general, statistical power is based on the total number of observations. To sustain enough power, a higher number of observers is needed if there are fewer cases. [18] But, a larger number of cases better represents case distribution in day-to-day practice. Depending on the availability of observers and the purpose of the study, the most appropriate ratio of observers and patients should be selected.

Two ways of calculating the sample size are described based on the hypothesis testing approach and confidence interval approach. [19,20] With the hypothesis testing approach, the hypothesis that the reliability coefficient is above a predefined level is tested. The confidence interval approach allows one to obtain a prespecified level of precision around an estimated reliability coefficient.

Usually, it is relevant to compare the reliability coefficients of different classifications or diagnostic methods. Different complex calculations are developed to compare these correlated reliability coefficients. [21] But, if the number of observations is large enough, the coefficient distribution is approaching a normal distribution. So, two-sample independent z-tests can be used to compare two different reliability coefficients. In addition, sample size calculation could be based on this test. [18]

Number of Observers versus Number of Cases
Example

After defining the null hypothesis, the number of cases and the minimal number of observers that will result in a power of at least 80% should be calculated. The power is based on the number of observers and cases. For example, if a higher number of cases is used, a lower number of observers is needed for significant power.

"Classic" study: five observers – 23 cases

Doornberg et al. assessed whether three-dimensional reconstructed CT scans have greater intraobserver and interobserver reliability and improved accuracy compared with two-dimensional images on the characterization, classification, and treatment choice in the evaluation of fractures of the distal aspect of the humerus.[14]

Power analysis revealed that a minimum sample of 23 fractures would provide 80% power (a = 0.05, ß = 0.20) to detect significant intraobserver and interobserver agreement using the kappa coefficient.

"Contemporary" study: 107 observers – 15 cases

Bruinsma et al. tested the null hypothesis that interobserver reliability of the Arbeitsgemeinschaft für Osteosynthesefragen classification of proximal humeral fractures, the preferred treatment, and fracture characteristics is the same for 2-D and 3-D CT. [22]

It was calculated that the 107 observers would have yielded 80% power (a = 0.05, ß = 0.20) to detect a difference of 0.15 in the kappa value for the fracture classification.

Pre- and Post-hoc Power Analysis

In a pre-hoc analysis, the power and number of observers are calculated before the experiment, contrary to a posthoc analysis, where the power is calculated after the experiment. In practice, post-hoc analyses are usually concerned with finding associations between subgroups that would otherwise be undetected and if the pre-hoc analysis was not possible. [1] It is also important to notice that more observers are needed to detect a smaller effect size to result in statistical power.

Example
Six observers – xx cases
Pre-hoc analysis

Doornberg et al. assessed whether 3D CT reconstructions can improve the reliability of complex tibial plateau fracture characterization and classification. [23]

A power analysis revealed that a sample size of 45 fractures would provide 95% power (a = 0.05; b = 0.95) to detect a significant difference in intraobserver and interobserver agreement with the use of the kappa coefficient.

xx observers – 15 cases
Post-hoc analysis

Claessen et al. tested the primary null hypothesis that the Minami-, Berndt and Harty-, Ferkel and Sgaglione- and Anderson classification systems are equally reliable for the classification of osteochondritis dissecans of the humeral capitellum. [10]

It was calculated that a minimum sample of 22 cases evaluated by a minimum of 32 observers would provide 80% power to detect a clinically significant difference of one categorical rating of k = 0.20.

Pre-hoc analysis						
Number of observers	**Alpha**	**Beta**	**Effect size**	**Sd1**	**Sd2**	**Power**
62	0.05	0.20	0.03	0.05	0.03	80%
14	0.05	0.20	0.07	0.05	0.03	80%
6	0.05	0.20	0.15	0.05	0.03	80%
876	0.05	0.20	0.03	0.2	0.1	80%
164	0.05	0.20	0.07	0.2	0.1	80%
38	0.05	0.20	0.15	0.2	0.1	80%
4364	0.05	0.20	0.03	0.3	0.4	80%
804	0.05	0.20	0.07	0.3	0.4	80%
178	0.05	0.20	0.15	0.3	0.4	80%
Post-hoc analysis						
100	0.05	0.20	0.03	0.05	0.03	95%
100	0.05	0.20	0.07	0.05	0.03	100%
100	0.05	0.20	0.15	0.05	0.03	100%
100	0.05	0.20	0.03	0.2	0.1	16%
100	0.05	0.20	0.07	0.2	0.1	60%
100	0.05	0.20	0.15	0.2	0.1	98%
100	0.05	0.20	0.03	0.3	0.4	7%
100	0.05	0.20	0.07	0.3	0.4	17%
100	0.05	0.20	0.15	0.3	0.4	56%

Sd= Standard deviation

Hypothetical study

Diagnostic Performance Characteristics

For diagnostic scenarios with a reliable reference standard, diagnostic performance characteristics can be calculated. This is uncommon in interobserver studies.

Sensitivity and Specificity

The sensitivity of a test is the proportion of patients with the disease with a positive test. The specificity of a test is the proportion of patients without the disease with a negative test. These parameters are usually less relevant in the clinical setting because the clinician typically does not know whether the patient has the disease or not. [1]

Negative- and Positive Predictive Value

More relevant in clinics are the probability of the patient having the disease when the test result is positive (positive predictive value) and the probability of the patient not having the disease when the test result is negative (negative predictive value). [1] The calculation of the diagnostic performance depends on the study design. In studies whereby all observers have rated all cases, the average performance of all observers can be used to calculate the diagnostic performance characteristics. In more complex designs, like split-plot studies, the calculation of diagnostic performance is more challenging and requires advanced statistics [Table 1].

Absence of Reference Standard

Without a "gold standard," it is difficult to calculate diagnostic performance characteristics. In this case, latent class analysis is a statistical method that can be used. [24,25] The latent class analysis evaluates groups of test results representing disease probability levels. The latent classes cannot be assessed directly (e.g., a fracture), but the resultant (e.g., no bone bridging) can be observed. [24]

Example

Buijze et al. [24] compared diagnostic performance characteristics of CT, MRI, bone scintigraphy, and physical examination to identify true fractures among suspected scaphoid fractures.

Percentage of Agreement

Interobserver variability is usually reported with kappa values or intra-class correlation (ICC) instead of percentages of agreement. [1] When reporting only percentages of the agreement, no adjustment for agreement due to chance alone is tested. [26]

Reliability

Many different statistical reliability measures have been described based on study design and type of data (e.g., categorical data, numerical data).

Table 1. Different designs and statistical analysis

	Diagnostic performance	Interobserver reliability	
	Sensitivity/Specificity	*Categorical data*	*Nummerical*
Factorial Design (Fully-crossed)			
All cases assessed by all observers All observers asses all cases	Average sensitivity/ specificity	Lights' kappa Davies and Fleiss Krippendorf's alpha	Two-Way mixed or fixed ICC Krippendorf's alpha
Split-plot design			
Each subset of cases assessed by a subset of observers Each observers rates all images/measurements of a patient	More advanced statistics like general linear models	Lights' kappa Davies and Fleiss Krippendorf's alpha	Two-Way mixed or fixed ICC Krippendorf's alpha
Each subset of cases assessed by a subset of random observers Each observer assess a random group of cases	More advanced statistics like general linear models	Fleis generalized kappa Krippendorf's alpha	One-way ICC Krippendorf's alpha

Categorical data

Observer reliability is commonly reported in kappa values. Kappa is a chance-corrected measure of agreement that compares the observed measure of agreement with the level of agreement expected by chance alone for categorical data. [26-28] Zero represents no agreement. The value of -1.00 means total disagreement, and +1.00 indicates perfect agreement. [27,29]

The original kappa described by Cohen is only suitable when there are two observers. [30] In online interobserver studies, there are usually more than two observers; in that case, standard kappa statistics are unsuitable. There are some variants for situations with more than two observers. Fleiss described a kappa-like formula in which, for each case, a constant number of observers is sampled randomly from a large population. [31] It is assumed that for each case, a different sample of observers is selected, and thus Fleiss kappa coefficient is not suitable for studies with fully crossed designs.

For studies in which all cases are rated by the same group of observers, Light proposed calculating a kappa statistic for each pair of observers and using the arithmetic mean of the kappa coefficients. [32] A similar approach was described by Davies and Fleiss. [33]

In addition to the kappa statistics, Krippendorf proposed an alternative reliability measure, Krippendorff's alpha.[34] This alpha is less well-known but more flexible and can be used for all data types. Besides, it allows missing data, which makes it suitable for incomplete designs. It embraces a large number of reliability coefficients. It calculates disagreements instead of correcting the percentage of agreement, resulting in fewer limitations. Alpha is presented as a scale from 0.000, indicating the absence of reliability to 1.000 (perfect reliability). [35]

Numerical data

ICC is a commonly used measure to assess the reliability of numerical data and is obtained from the analysis of variance models for quantitative measurements (ANOVA). [29] ICC is a measure of agreement between observers, adjusted for agreement due to chance alone, taking into account the measure of disagreement. Potential ICC values ranging from -1 to 1 (perfect agreement), with negative values indicating systematic disagreement and 0 only random agreement. It can be used for two or more observers and is suitable for different study designs.

There are several variants of the ICC. Depending on the type of data and study purpose, the correct type of ICC has to be chosen. The one-way ICC is required if a random observer is selected for every case. Two-way is chosen if there is a sample of observers for the different cases. Then, it has to be decided whether the study aims to generalize the results of the observers in the study to a larger population. In other words, are the observers a random sample of a larger population (random or mixed method), or are you only interested in these specific observers (fixed method). For online interobserver studies, generally, the two-way mixed method is appropriate. Furthermore, the ICC can be based on absolute values or based on consistency. When the average score of many ratings is used, an average measure ICC should be calculated. [35,36]

Interpretation of reliability coefficients

Although some arbitrary classifications exist, there is no standard for interpreting reliability coefficients. [36] Landis and Koch proposed a commonly used classification for kappa values: 0.01 to 0.20 indicating slight agreement; 0.21 to 0.40, fair agreement; 0.41 to 0.60, moderate agreement; 0.61 to 0.80, substantial agreement; and 0.81 or more, almost perfect agreement. [29] However, Krippendorff suggested a more conservative classification: no conclusions can be made for values less than 0.67, conclusions can tentatively be made for values between 0.67 and 0.80, and definite conclusions can be made for values above 0.80. [34]

Discussion

Treatment variation and differences in prognosis are unwanted in medical practice.

Classification systems help indicate a potential prognosis and guide optimal treatment decision-making. Classification systems need to be reliable and valid, but measurement of validity can be variable and is prone to observer bias. Thus, there is a growing number of papers about online studies on the reliability of classification systems. This 'handbook' paper is a systematic overview of the methodological framework for online reliability studies. It can be helpful for researchers who want to conduct a reliability study.

The strengths of interobserver studies include a large number of observers, which allows randomization and subgroup analysis to increase the generalizability of the results. [22] A large number of experienced and internationally distributed observers, and many cases improving both statistical power and external validity. [37]

Most studies are limited to the interobserver agreement only because it is less relevant to clinical practice, as surgeons mostly agree with themselves and not so much with each other. In addition, it is less time efficient if an observer has to complete the survey several times.

Another limitation is the possible observer bias that can arise with this study type. To reduce the chance of observer bias, the observers are usually unaware of the hypothesis.

Many investigators are not aware of an important disadvantage of kappa. Similar to positive and negative predictive values, the calculation of the kappa is affected by prevalence. If the prevalence of an outcome is low, it generates a lower kappa than one might expect. [22,27,28,38]

The first kappa paradox –a low value but high values of the observed proportion of agreement– will occur only if the marginal totals are highly symmetrically unbalanced. The second paradox –producing higher kappa values for asymmetrical than for symmetrical imbalance– can be associated with high values of the observed proportion of agreement if the imbalances have different degrees of high symmetry. [27]

Example

Bruinsma et al. [22] found poor to fair (k= 0.09-0.35) interobserver agreement after testing the null hypothesis that interobserver reliability of the Arbeitsgemeinschaft fur Osteosynthesefragen and Neer classifications, preferred treatment, and fracture characteristics are the same for two-dimensional CT and three-dimensional CT scans, while the percentage of agreement was higher (87-97%).

Data should be tested for the kappa paradoxes, and if this is the case, it is useful to describe the percentage of agreement as well. There are some formulas to correct for the kappa paradox.

Siegel and Castellan proposed an alternative formula that corrects the prevalence and imbalance problems. [39]

For ordinal data, Cohen described a weighted kappa variant in which disagreements are penalized based on the amount of disagreement. [30] There is a linear and a quadratic variant. The quadratic variant of the weighted kappa is similar to the two-way mixed, consistency intra-class correlation coefficient (ICC).

In conclusion, a high agreement between observers is important to guide treatment and predict clinical outcomes. Thus, reliable and valid classification systems should be developed, and several existing classification systems should be simplified. In future studies, observer error and bias may be reduced by clear consensus definitions, measurement techniques, and diagnostic criteria.

Femke M.A.P Claessen MD PhD[1]
Ruben Zwiers MD[2]
Thierry G. Guitton MD PhD[3]
Job N. Doornberg MD PhD[4]

1 Haaglanden Medisch Centrum, Leidschendam, Orthotrauma Research Center Amsterdam, Amsterdam, the Netherlands
2 Orthotrauma Research Center Amsterdam, Amsterdam, the Netherlands
3 Department of Plastic, Reconstructive, Hand en Burn Surgery, Martini Hospital, Burn Center Groningen, the Netherlands
4 Flinders Medical Center, Royal Adelaide Hospital, Adelaide, Australia

References

1. Kocher MS, Zurakowski D. Clinical epidemiology and biostatistics: a primer for orthopaedic surgeons. J Bone Joint Surg Am. 2004;86-A(3):607-20.
2. Hegedus EJ, Stern B. Beyond SpPIN and SnNOUT: Considerations with Dichotomous Tests During Assessment of Diagnostic Accuracy. J Man Manip Ther. 2009;17(1):E1-5. doi: 10.1179/jmt.2009.17.1.1E.
3. Doornberg J, Lindenhovius A, Kloen P, van Dijk CN, Zurakowski D, Ring D. Two and three-dimensional computed tomography for the classification and management of distal humeral fractures. Evaluation of reliability and diagnostic accuracy. J Bone Joint Surg Am. 2006;88(8):1795-801. doi: 10.2106/JBJS.E.00944.
4. Obuchowski NA, Gallas BD, Hillis SL. Multi-reader ROC studies with split-plot designs: a comparison of statistical methods. Acad Radiol. 2012;19(12):1508-17. doi:10.1016/j.acra.2012.09.012

5. Lindenhovius A, Karanicolas PJ, Bhandari M, van Dijk N, Ring D, Collaboration for Outcome Assessment in Surgical T. Interobserver reliability of coronoid fracture classification: two-dimensional versus three-dimensional computed tomography. J Hand Surg Am. 2009;34(9):1640-6. doi: 10.1016/j.jhsa.2009.07.009.
6. Minami M, Nakashita K, Ishii S, et al. Twenty-five cases of osteochondritis dissecans of the elbow. Rinsho Seikei Geka. 1979;14:805-810.
7. Berndt AL, Harty M. Transchondral fractures (osteochondritis dissecans) of the talus. J Bone Joint Surg Am. 1959;41-A:988-1020.
8. Ferkel RD, Zanotti RM, Komenda GA, et al. Arthroscopic treatment of chronic osteochondral lesions of the talus: long-term results. Am J Sports Med. 2008;36(9):1750-62. doi:10.1177/0363546508316773
9. Anderson IF, Crichton KJ, Grattan-Smith T, Cooper RA, Brazier D. Osteochondral fractures of the dome of the talus. J Bone Joint Surg Am.1989;71(8):1143-52.
10. Claessen FM, van den Ende KI, Doornberg JN, et al. Osteochondritis dissecans of the humeral capitellum: reliability of four classification systems using radiographs and computed tomography. J Shoulder Elbow Surg. 2015;24(10):1613-8.doi:10.1016/j.jse.2015.03.029
11. Claessen FM, Stoop N, Doornberg JN, et al. Interpretation of Post-operative Distal Humerus Radiographs After Internal Fixation: Prediction of Later Loss of Fixation. J Hand Surg Am. 2016;41(10):e337-e341. doi:10.1016/j.jhsa.2016.07.094
12. Altaye M, Donner A, Klar N. Inference procedures for assessing interobserver agreement among multiple raters. Biometrics. 2001;57(2):584-8. doi: 10.1111/j.0006-341x.2001.00584.x.
13. Cantor AB. Power calculation for the log rank test using historical data. Control Clin Trials. 1996;17(2):111-6. doi: 10.1016/s0197-2456(96)80002-x.
14. Donner A, Eliasziw M. Sample size requirements for reliability studies. Stat Med. 1987;6(4):441-8. doi: 10.1002/sim.4780060404.
15. Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. Stat Med. 1998;17(1):101-10. doi: 10.1002/(sici)1097-0258(19980115)17:1<101::aid-sim727>3.0.co;2-e.
16. Cichetti DV. Methodological Commentary The Precision of Reliability and Validity Estimates Re-Visited: Distinguishing Between Clinical and Statistical Significance of Sample Size Requirements. J Clin Exp Neuropsychol. 2010;23(5):695-700. doi: 10.1076/jcen.23.5.695.1249.
17. Shoukri MM, Asyali MH, Donner A. Sample size requirements for the design of reliability study: review and new results. Statistical methods in medical research. 2004;13(4):251-71.
18. Guitton TG, Ring D, Science of Variation G. Interobserver reliability of radial head fracture classification: two-dimensional compared with three-dimensional CT. J Bone Joint Surg Am. 2011;93(21):2015-21. doi:10.2106/JBJS.J.00711
19. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther. 2005;85(3):257-68.
20. Rotondi MA, Donner A, Koval JJ. Evidence-based sample size estimation based upon an updated meta-regression analysis. Res Synth Methods. 2012;3(4):269-84. doi:10.1002/jrsm.1055
21. Vanbelle G. In the beginning was a relationship. Revue belge de medecine dentaire. 2008;63(2):77-80.
22. Bruinsma WE, Guitton TG, Warner JJ, Ring D, Science of Variation G. Interobserver reliability of classification and characterization of proximal humeral fractures: a comparison of two and three-dimensional CT. J Bone Joint Surg Am. 2013;95(17):1600-4. doi:10.2106/JBJS.L.00586
23. Doornberg JN, Rademakers MV, Van Den Bekerom MP, et al. Two-dimensional and three-dimensional computed tomography for the classification and characterisation of tibial plateau fractures. Injury. 2011;42(12):1416-25. doi: 10.1016/j.injury.2011.03.025.
24. Buijze GA, Mallee WH, Beeres FJ, Hanson TE, Johnson WO, Ring D. Diagnostic performance tests for suspected scaphoid fractures differ with conventional and latent class analysis. Clin Orthop Relat Res. 2011;469(12):3400-7. doi:10.1007/s11999-011-2074-9
25. Mallee W, Doornberg JN, Ring D, van Dijk CN, Maas M, Goslings JC. Comparison of CT and MRI for diagnosis of suspected scaphoid fractures. J Bone Joint Surg Am. 2011;93(1):20-8. doi: 10.2106/JBJS.I.01523.
26. Cole RJ, Bindra RR, Evanoff BA, Gilula LA, Yamaguchi K, Gelberman RH. Radiographic evaluation of osseous displacement following intra-articular fractures of the distal radius: reliability of plain radiography versus computed tomography. J Hand Surg Am. 1997;22(5):792-800. doi: 10.1016/s0363-5023(97)80071-8.
27. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol. 1990;43(6):543-9. doi: 10.1016/0895-4356(90)90158-l.
28. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Family medicine. 2005;37(5):360-3.
29. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-74.
30. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull. 1968;70(4):213-20. doi: 10.1037/h0026256.
31. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull. 1971;76(5):378.
32. Light RJ. Measures of response agreement for qualitative data: some generalizations and alternatives. Psychol Bull. 1971;76:365-377.
33. Davies M, Fleiss JL. Measuring Agreement for Multinominal Data. Biometrics. 1982;38(4):1047-1051.
34. Krippendorff K. Validity in content analysis. Computerstrategien fur die Komminikationsanalyse. 1980:69-112.

35. Hallgren KA. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. Tutor Quant Methods Psychol. 2012;8(1):23-34. doi: 10.20982/tqmp.08.1.p023.
36. Streiner DLNGR. Health Measurement Scales. 2008; https://academic.oup.com/book/6813
37. van Kollenburg JA, Vrahas MS, Smith RM, Guitton TG, Ring D, Science of Variation G. Diagnosis of union of distal tibia fractures: accuracy and interobserver reliability. Injury. 2013;44(8):1073-5. doi: 10.1016/j.injury.2012.10.034.
38. Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol. 1990;43(6):551-8. doi: 10.1016/0895-4356(90)90159-m.
39. Sidney S. Nonparametric statistics for the behavioral sciences. The Journal of Nervous and Mental Disease. 1957;125(3):497.