**Ophthalmic Statistics Note 14: Method agreement studies in ophthalmology – the Intraclass Correlation Coefficient?**

Catey Bunce * [1,2,3], Gabriela Czanner[4], Joanna Moschandreas[5], Irene Stratton[6], Caroline J Doré[7], Nicholas Freemantle[7]

*Corresponding author

1 School of Population Health and Environmental Sciences, King's College London
2 NIHR BRC for Ophthalmology at Moorfields and the UCL Institute of Ophthalmology
3 London School of Hygiene & Tropical Medicine, London UK
4 Department of Applied Mathematics, Liverpool John Moores University, Liverpool, UK
5 Nuffield Department of Clinical Neurosciences, Oxford University, Oxford, UK
6 Gloucestershire Retinal Research Group, Cheltenham General Hospital, Sandford Road, Cheltenham, UK
7 Institute of Clinical Trials and Methodology, University College London, UK

**Keywords:** applied medical statistics, correlation, method agreement, measurement

**Word count: 1904**

**Correspondence to Dr Catey Bunce, School of Population Health and Environmental Sciences, King's College London, London SE1 1UL, catey.bunce@kcl.ac.uk**

**Introduction**

Our previous note outlined why method agreement studies are so important in ophthalmology. [1] Technology moves at a relentless pace and clinicians are keen to adopt innovative techniques that may offer benefits to their patients, such as shorter or less invasive testing, in addition to creating richer datasets that may increase the research potential of the data captured. Researchers and clinicians must, however, use caution to ensure that any differences observed between measurements made on a patient with different methods of measurement are truly due to changes in pathology rather than the method of measurement, the observer making the measurement or other variables that might influence the measurement. Even if two machines appear to report the same characteristic it is possible that one machine is measuring a different anatomical feature than another machine but used the same name. An example is in studies of keratometry and topography where the term Kmax is used to describe both the steepest meridian of the cornea in the central 3 mm (also called K2) and the power of the steepest point of the cornea. [2, 3] This is of particular importance in trials investigating treatments for keratoconus where Kmax may be the primary outcome measure or used to determine subject eligibility. [4]

Our previous note outlined that the regulation of devices is quite different to that of medicines. There is a tension between innovation and safety and whilst measurement

reproducibility may not immediately be seen as relevant to harm, measurements are used to make decisions about diagnosis, progression and treatment. For example, a large change in intraocular pressure between two visits in a child with glaucoma may indicate the need for examination under anaesthetic, whilst a large change in K2 readings in an individual with keratoconus might indicate a discussion about crosslinking. So there is a clear need to ensure that a change is real and not the result of a difference in the manner of measurement. [4, 5]

When a study is designed to compare a new method for measuring a continuous characteristic such as intraocular pressure or retinal nerve fibre layer thickness with a standard method, it is our view that limits of agreement (LOA) is the best approach. [1] We have described the need to consider agreement within individuals (precision or random error) and agreement on average (bias or systematic error). It is important, however, to be aware that if a method does not agree well with itself it is clearly unlikely to agree well with another method. Studies examining whether a method agrees with itself are assessing the **test-retest reliability** of that method. [6] Another term in common use for this, however, is the repeatability of the test. [6] Even more confusingly within method (measuring instrument) comparison studies is the common use of a further term **reproducibility**. [6] **Repeatability and reliability** are terms used when repeat measurements are made on the same subject under identical conditions (the same observer, the same patient, measurements made sufficiently close in time that their condition can be assumed not to have changed). [6] Repeatability and reliability thus require consideration of the same device being used and not consideration of different devices. An alternative definition of reliability is the ratio of variability between subjects to total variability. [7, 8] This definition whilst less commonly used is provided here since it is a literal translation of the arithmetic formula that is often used to compute an intraclass correlation coefficient (ICC) which is a statistic often used in reliability studies.

**Reproducibility** is the term used when measurements are made on subjects under changing conditions (different observers, different methods of measurement). (6) Reproducibility is thus the term used to describe method agreement studies since these will always involve different methods of measurement. [6]

The terms reliability and agreement are unfortunately often used interchangeably despite being conceptually distinct. [9]

**The Bland-Altman approach** can be used for both types of study (method agreement studies – which compare one device with another and reliability studies – which look at repeated use of the same instrument). If the Bland-Altman approach is used for repeatability /reliability there may be no need to test for bias between measurements since the measurements have been made under identical conditions. Note, however, that even if measurements are made under identical conditions only the first measurement will correspond to the first time a patient is measured by a particular machine and the measurement may differ the second time a patient is measured by the machine simply because of measurement error or indeed due to patient characteristics. If a patient is anxious their heart rate might increase and so it is possible that there is a difference which

is why bias is often still assessed! Such a difference might be termed a white-coat effect or a learning curve effect and is still worth considering should a difference be detected. It should be noted that there are real challenges in assessing reliability of methods of measurement that impact on the patient such there is a change induced by that method. An example would be that of applanation tonometry which indents the cornea as part of its assessment and this itself may impact upon the measurement obtained.

A frequently used statistic in repeatability / reliability studies is the intraclass correlation coefficient (ICC) and a question often asked is whether the ICC needs reporting instead of, or in conjunction with limits of agreement in method agreement studies. This is despite a clear steer given by Bland and Altman to the contrary in relation to method agreement. [10] One needs to understand that use of ICC and of Bland-Atman approaches depend on the goal of analysis as well as on the design of the collected data.

Scenario 1

I have submitted a paper for publication and I have used the limits of agreement approach to assess the reliability and reproducibility of the Pentacam in people with keratoconus. A reviewer has advised that I must include the ICC since it is commonly seen in such reports.

I don't recall any mention of an ICC when taught statistics at medical school so search the internet for references that might assist.

I learn from a paper by Bland and Altman that the ICC was introduced to assess the relationship between variables in classes. [10] For example to assess the relationship between measurements made on subjects who are paired in some way where the order is arbitrary or interchangeable – for example looking e looking at the binocular best corrected visual acuity of identical twins, because in twins saying which twin is number 1 or 2 is arbitrary.

Suppose we have measures on ten sets of identical twins so we have binocular visual acuity from two twins, albeit different individuals. To calculate the usual correlation coefficient we would designate one twin as X and the other twin as Y. The assignment of these labels to the twins would be completely arbitrary and if I were to repeat this process many times, then in one instance Peter might be twin X and Paul be twin Y but in another case Peter is called twin Y and Paul is called twin X. Different assignments of measurements of X and Y in the calculation of the correlation coefficient (r), would produce different values of r, which of course does not make sense, as there should be only one value of r. To allow for this, the ICC is used, as a ratio of variability between subjects over the total variability, where the total variability is the variability between subjects plus the variability within twins plus measurement error. The ICC can be interpreted as the mean correlation across all possible orderings. [11]

Consider a different situation when we have paired measures of IOP from the same patient and on the same day from two machines. The measures are paired in that they have been

measured on the same patient. The measures however for a reproducibility study will come from different machines - for example Goldman tonometer and Tonosafe – providing pairs of data. Here one measurement is clearly from Goldman and one from Tonosafe, which is something that will be ignored by the ICC. The ICC would assign the labels X and Y to each pairing but this would then mean that X in one patient would mean tonometer and in another patient would mean Tonosafe. Because the assignment of labels X and Y to the pairs of measures is no longer arbitrary (since we would wish it to define the type of machine that is used) the ICC is clearly not appropriate for reproducibility studies. Furthermore, as the name suggests, the ICC is a "correlation" coefficient and as such it will tell us if the two measurements are correlated or whether there is concordance of items in genetics (the area whether the ICC was first developed). [12] The objective of our study however is not to assess correlation but to assess agreement.

What, however, about repeatability studies – where the same machine has been used on more than one occasion in the same subject?

I search the internet further and learn that there are actually a number of different ICC statistics in use, each for a different designed study, and much debate about which ICC statistic is appropriate to assess repeatability. [13] Some of the ICCs are parametric (ie their computation assumes
underlying statistical distributions) and others are non parametric (considering only the ranking of the data). Some of the ICCs treat the methods being compared as a random sample from all possible methods whereas in a method comparison study there are two (or more) specific methods being studied. [13] I learn too that different ICCs may yield different answers for the same data set and that the ICC is influenced by the range of data used for its calculation. [12] Since the ICC is the ratio of variability between subjects divided by the total variability, then if the variance between subjects is high (e.g. in terms of IOP), the value of ICC will be higher than if the ICC is calculated on a group of patients who are similar to each other. [10] Including a more heterogeneous group will therefore yield larger values of the ICC which may be interpreted by some to indicate better agreement. Whilst any study must consider very carefully the representativeness of the study subjects, it is important to understand that the ICC might be inflated simply by ensuring inclusion of more variable subjects. This is not a problem with Bland-Altman analyses since it does not involve a ratio of variability between subjects to variability in total.

I have identified sufficient concerns in relation to the ICC to decide that it provides no additional value beyond the limits of agreement in method agreement studies. My senior collaborator is persuaded by my comprehensive review of the subject and we agree to omit the ICC and respond to the reviewer.

**Lessons learned**

**There are different types of ICC and these can yield different answers for the same data set. The study design will determine which type of ICC is best used.**

**The ICC can be increased simply by including more variable subjects but this does not mean that the measure is more reliable.**

**Measurement error can impact upon patient care which is why method agreement studies are so important.**

**A paper reporting LOA when examining method agreement does not need to include also the ICC.**

Contributors: CB drafted the paper. IS, GC and JM contributed to the second version of the paper. CJD and NF critically reviewed and revised the paper.

Funding: CB is partly funded/supported by the National Institute for Health Research (NIHR) Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust and King's College London. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

Competing interests: The authors declare no competing interests.

References

1. Bunce C, Stratton IM, Elders A, Czanner G, Grzeda MT, Doré CJ, Freemantle N. Ophthalmic statistics note 13: method agreement studies in ophthalmology—please don't carry on correlating. British Journal of Ophthalmology 2019;**103**:1201-1203. doi: 10.1136/bjophthalmol-2018-313759.

2. Wittig-Silva C, Chan E, Islam FM, Wu T, Whiting M, Snibson GR. A randomized, controlled trial of corneal collagen cross-linking in progressive keratoconus: three-year results. Ophthalmology 2014;**121**(4):812-21. doi: 10.1016/j.ophtha.2013.10.028.

3. Brunner M, Czanner G, Vinciguerra R, Romano V, Ahmad S, Batterbury M, Britten C, Willoughby CE, Kaye SB. Improving precision for detecting change in the shape of the cornea in patients with keratoconus. Sci Rep 2018;**8**(1):12345. doi: 10.1038/s41598-018-30173-7.

4. Chowdhury K, Doré C, Burr JM, Bunce C, Raynor M, Edwards M, Larkin DFP. A randomised, controlled, observer-masked trial of corneal cross-linking for progressive keratoconus in children: the KERALINK protocol. BMJ Open 2019; **9**:e028761. doi:10.1136/bmjopen-2018-028761.

5. Dahlmann-Noor AH, Puertas R, Tabasa-Lim S, El-Karmouty A, Kadhim M, Wride NK, Lewis A, Grosvenor D, Rai P, Papadopoulos M, Brookes J, Bunce C, Khaw PT. Comparison of handheld rebound tonometry with Goldmann applanation tonometry in children with glaucoma: a cohort study. BMJ Open. 2013 Apr 2;3(4). pii: e001788. doi: 10.1136/bmjopen-2012-001788.

6. Kottner J, Audigé L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, Roberts C, Shoukri M, Streiner DL. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. J Clin Epidemiol 2011;**64**(1):96-106. doi: 10.1016/j.jclinepi.2010.03.002.

7. Dunn G. Statistical Evaluation of Measurement Errors: Design and Analysis of Reliability Studies, 2nd ed 2004 Arnold, London.

8. Streiner DL, Norman GR. Health Measurement Scales: A Practical Guide to Their Development and Use. 4th ed 2008 Oxford University Press, Oxford.

9. de Vet HC, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. J Clin Epidemiol 2006; **59**(10):1033-9.

10. 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.

11. Fay MP. Random marginal agreement coefficients: rethinking the adjustment for chance when measuring agreement. Biostatistics 2005;**6**:171-180.

12. Müller R, Büttner P. A critical discussion of intraclass correlation coefficients. Stat Med 1994; **13**(23-24):2465-76.

13. Zaki R, Bulgiba A, Ismail R, Ismail NA. Statistical methods used to test for agreement of medical instruments measuring continuous variables in method comparison studies: a systematic review. PLoS ONE 2012; 7(5): e37908. doi.org/10.1371/journal.pone.0037908