

Agreement between methods of measurement



Gabriela Czanner PhD CStat
Department of Biostatistics
Department of Eye and Vision Science
University of Liverpool
and
Clinical Eye Research Center
Royal Liverpool University Hospital

Email: czanner@liverpool.ac.uk

Web: <http://pcwww.liv.ac.uk/~czanner/>

5 October 2016 (4.15 and 4.45)

MERSEY POSTGRADUATE TRAINING PROGRAMME

This is part of the workshop series:

Basic Statistics for Eye Researchers and Clinicians

Temperature measurement in children to diagnose fever

- How many cases of fever are missed?
- How many afebrile children are incorrectly classified as having fever?
- How well do measurements taken by two raters agree using a mercury thermometer?
- [Similar examples in ophthalmology?](#)



“Hey - 104.5? Your kid’s temperature happens to be my favorite radio station!”

Learning objectives

- What is the proper statistical method of agreement
 - ... when we do **measurements on continuous scale**
- Two components of agreement
 - **bias** and **precision**
- What is Bland-Altman plot
 - how to calculate limits of agreement

Method Comparison Studies in Ophthalmology

- ❖ Method comparison studies are studies that compare two or more ways of measurement (e.g. visual acuity, intraocular pressure, extent of retinal vein occlusion)
- ❖ Technology moves at an advanced rate
- ❖ New ways for measuring ocular characteristics seem to be emerging constantly
- ❖ New devices can be introduced once it has been shown that they are safe
- ❖ Focus of this talk, is on **continuous** data (different methods for categorical)

Method Comparison Studies in Ophthalmology

Intraocular pressure (pressure in the eye)

- ❖ Goldmann applanation tonometer
- ❖ Tono-Pen XL
- ❖ Perkins applanation tonometer
- ❖ Draeger tonometer
- ❖ iCare rebound tonometer



A comparison of four methods of tonometry: method agreement and interobserver variability.
Tonnu PA, Ho T, Sharma K, White E, **Bunce C**, Garway-Heath D.
Br J Ophthalmol. 2005 Jul;89(7):847-50

Method Comparison Studies in Ophthalmology

Retinal thickness (retina = layer at the back of the eye)

- ❖ Confocal laser ophthalmoscope
- ❖ Scanning laser polarimeter
- ❖ Optical coherence tomography

Optical coherence tomography analysis of the macula after scleral buckle surgery for retinal detachment.

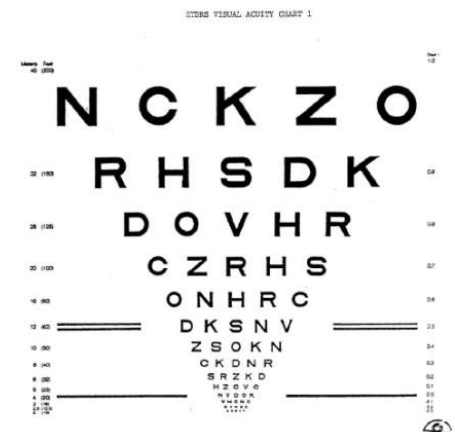
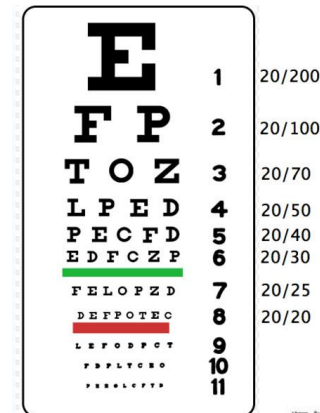
Benson SE, Schlottmann PG, **Bunce C**, Xing W, Charteris DG. Ophthalmology. 2007 Jan;114(1):108-12. Epub 2006 Nov 7



Method Comparison Studies in Ophthalmology

Visual Acuity (letters seen on a chart)

- ❖ Snellen Chart
- ❖ ETDRS (LogMAR)
- ❖ Frieberg visual acuity test
- ❖ Smart-phone based visual acuity apps



The Eye Phone Study: reliability and accuracy of assessing Snellen visual acuity using smartphone technology.

Perera C, Chakrabarti R, Islam FM, Crowston J.

Eye (Lond). 2015 May 1. doi: 10.1038/eye.2015.60. [Epub ahead of print]

Method Comparison Studies in Ophthalmology

Keratometry (steepness of the eye)

- ❖ Manual Javal Kereatometer
- ❖ Automated Topcon kerato-refractometer
- ❖ IOLMaster



Keratometry with five different techniques: a study of device repeatability and inter-device agreement.
Mehravaran S, Asgari S, Bigdeli S, Shahnazi A, Hashemi H.
Int Ophthalmol. 2014 Aug;34(4):869-75. doi: 10.1007/s10792-013-9895-3. Epub 2014 Feb 23.

Method Comparison Studies in Ophthalmology: goal

❖ Goal:

❖ **Do the different methods for measuring the same ocular parameter agree well enough to be “interchanged”?**

❖ May be comparing conventional equipment with newer, faster, cheaper method

❖ May simply be comparing two methods where neither can be said to be the truth

Method Comparison Studies in Ophthalmology

Common Steps in Analysis

- ❖ **Step 1:** plot the values obtained by each method as a scatter plot
- ❖ **Step 2:** (judging from many publications) is to calculate correlation coefficient...

Question: Is it correct to use correlation coefficient?

Answer: NO

Example 1: Intraocular pressure

GOAL: comparison of **iCare** rebound tonometer with Goldmann applanation tonometry (**GAT**) when measuring Intraocular pressure

- ❖ 100 children with glaucoma
- ❖ 1st tested with iCare, then tested with GAT
- ❖ Ideally tests in random order
- ❖ Ideally also, both tests should have been used twice

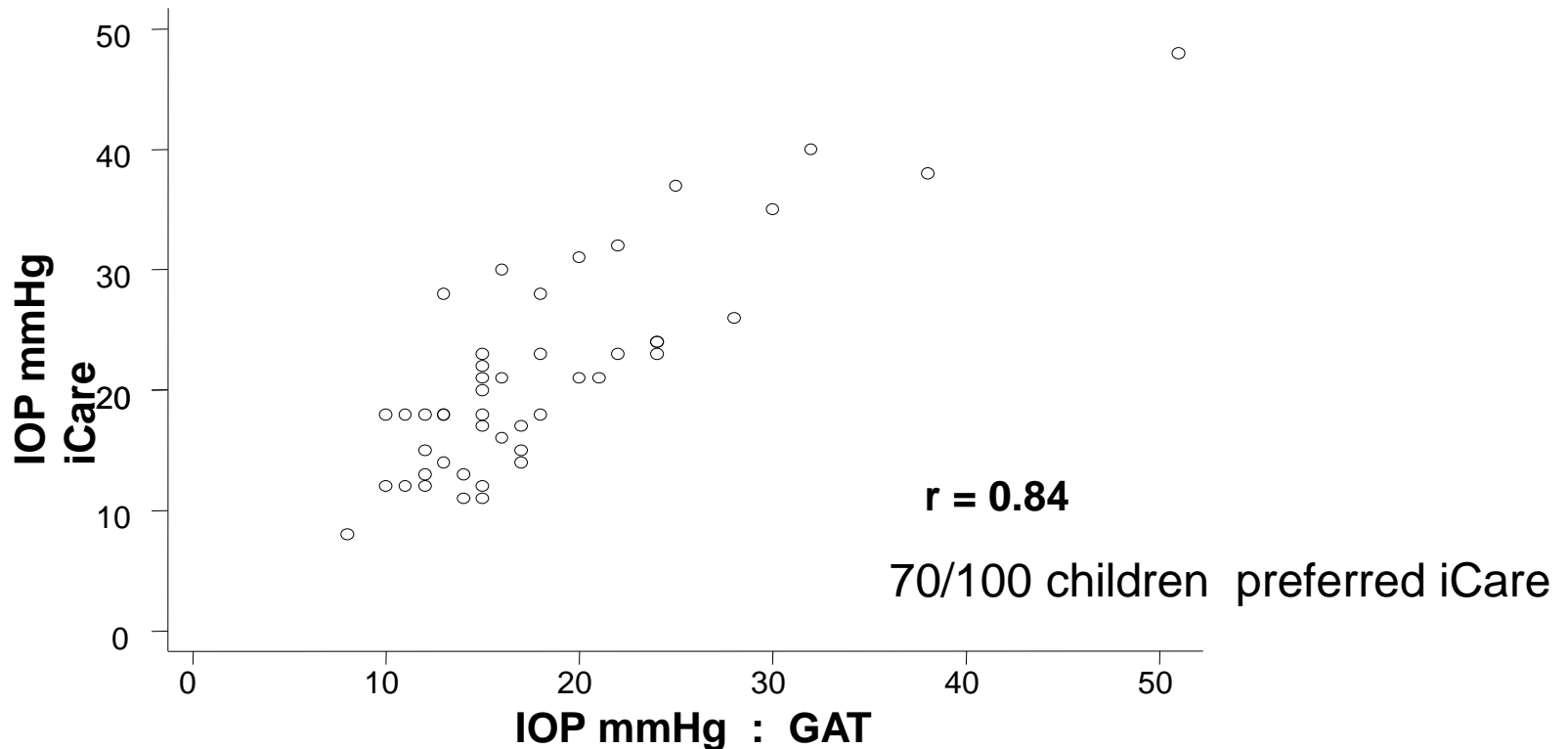
Comparison of handheld rebound tonometry with Goldmann applanation tonometry in children with glaucoma: a cohort study.

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.

BMJ Open. 2013 Apr 2;3(4). pii: e001788. doi: 10.1136/bmjopen-2012-001788. Print 2013

Example 1: Intraocular pressure

Next the researcher Dr Confused plots the data and calculates the correlation. He hopes to use that to find out if iCare and GAT agree enough to be used interchanged.

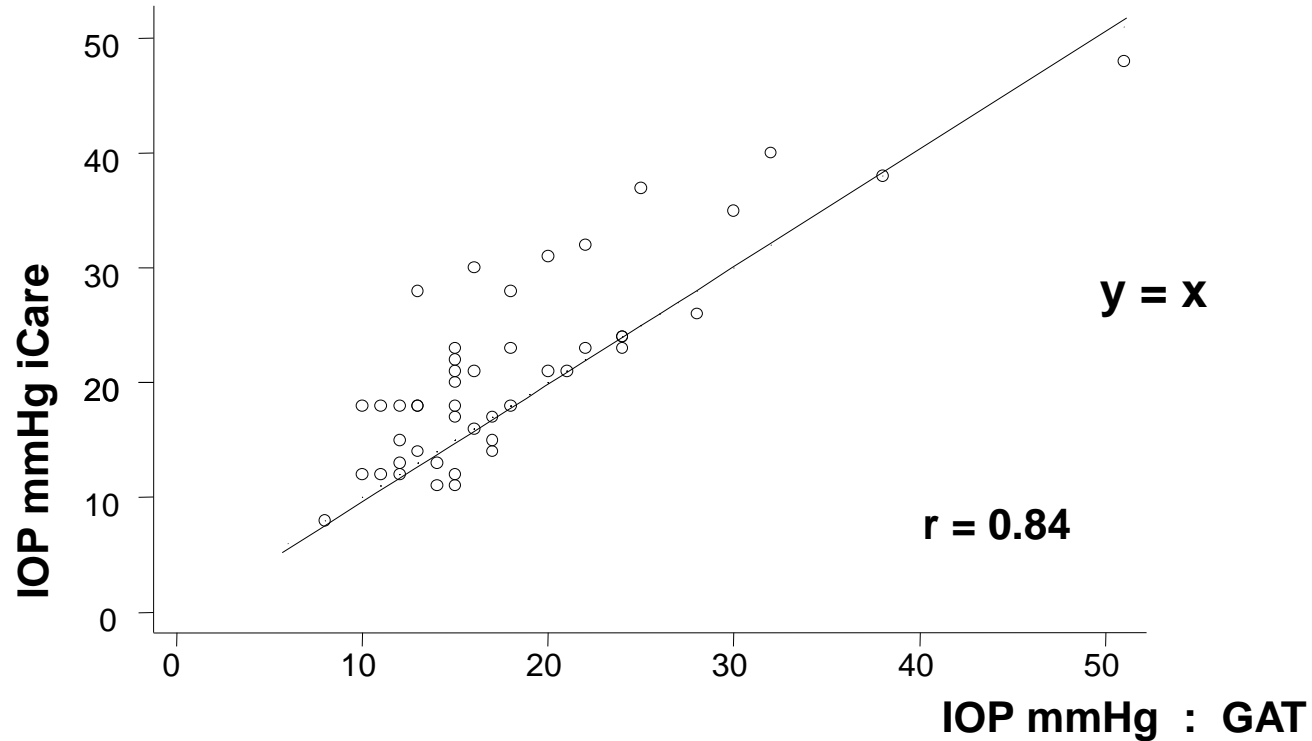


Question 1: Can iCare and GAT be used interchangeably?

Question 2: What does the correlation tells us? Does is help us to answer the question 1?

Example 1: Intraocular pressure

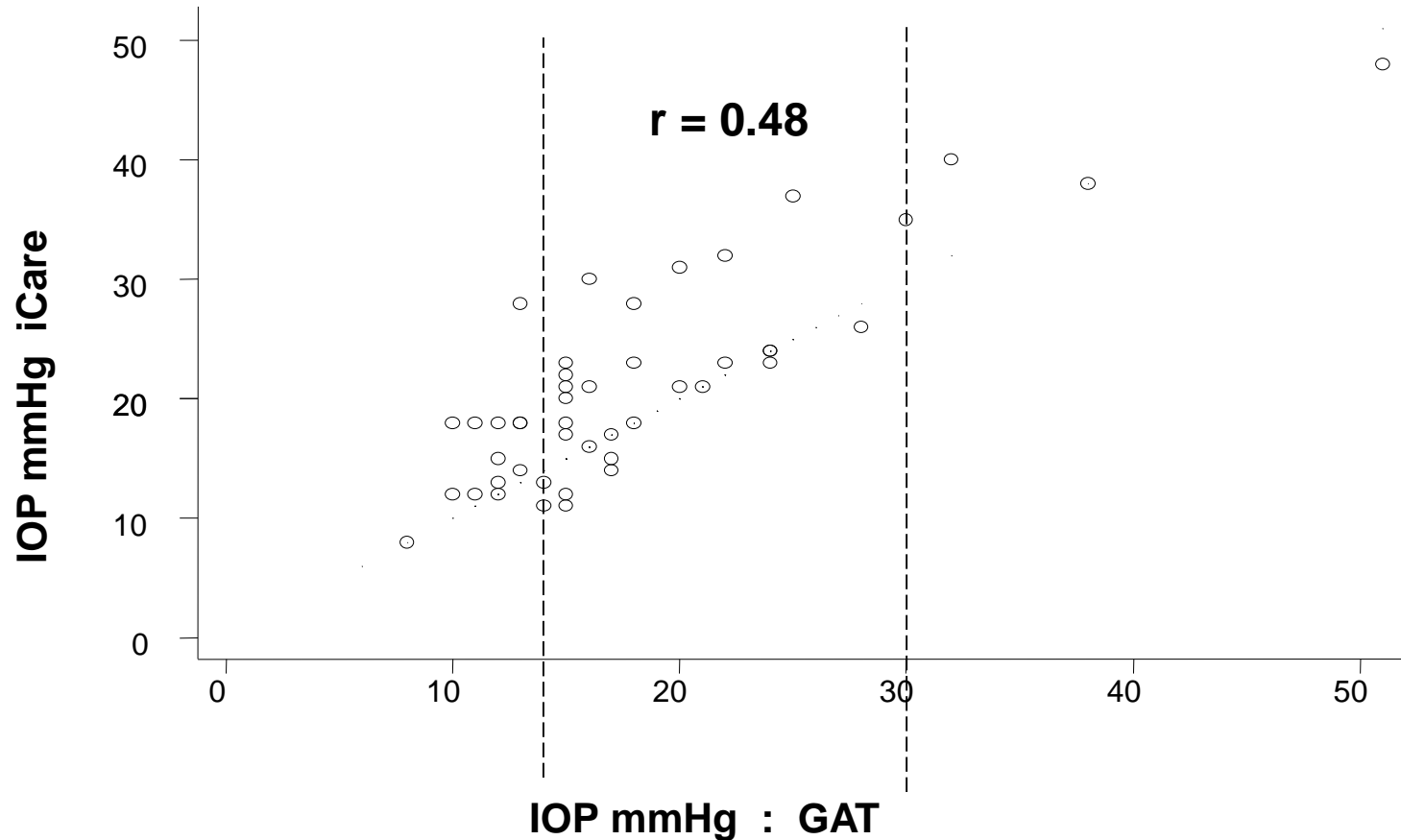
Next, we assume that Dr. Confused plots the $y=x$ line.



Question 1: What do you think now about the agreement between iCare and GAT?

Example 1: Intraocular pressure

Next we assume that Dr Confused is curious about the agreement in the middle range.



http://onlinestatbook.com/stat_sim/restricted_range/index.html

Questions: Why is the correlation smaller now? Is it useful for our goal?

Learning point: This is one reason to not use correlation in agreement studies.

Method Comparison Studies in Ophthalmology: Should we use correlation?

- ❖ Use of correlation coefficient for this analysis (i.e. the analysis of agreement) is misguided
- ❖ Stems from common failure to appreciate what the correlation coefficient gives
- ❖ Correlation measures linear association. In these studies we are assessing agreement.

Correlation \neq Agreement

Method Comparison Studies in Ophthalmology: bit of history

- ❖ First highlighted in 1983 by Doug Altman and Martin Bland
- ❖ How well do the methods (e.g. iCare vs GAT) agree on average?
- ❖ How well do the methods agree for individuals?

Altman DG, Bland JM. (1983). Measurement in medicine: the analysis of method comparison studies. *The Statistician* 32, 307-317

Method Comparison Studies in Ophthalmology: correct analysis

- ❖ To do comparison study correctly we need to look into two components
 - ❖ Bias
 - ❖ Precision
 - ❖ ...and visualise on Bland-Altman plot

Method Comparison Studies in Ophthalmology: First component is bias

- ❖ **Bias: measures average agreement**
 - ❖ For each person calculate the difference between iCare and Goldmann
 - ❖ Mean of these differences = estimate of bias
 - ❖ Test of significance (e.g. *t-test*) against null hypothesis that there is no systematic bias

Example 1. Intraocular pressure: Bias

- ❖ Measures average agreement
- ❖ Bias (95% CI) estimate: 3.15 (1.73, 4.57) mmHg
- ❖ Test of significance, (*t*-test) $t_{(45)} = 4.46, P < 0.001$

Question: Is there an evidence of systematic bias between instruments?

Method Comparison Studies in Ophthalmology: Second component is precision

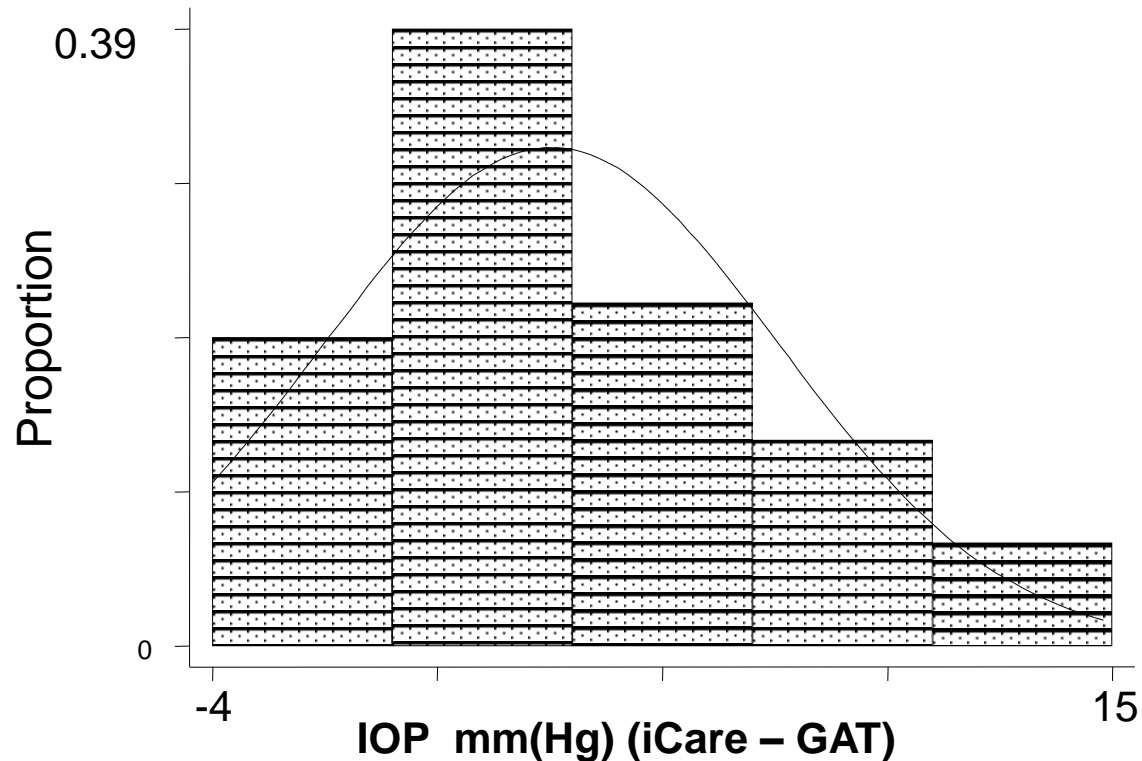
- ❖ **Precision:** *SD* of the differences i.e. how well do the methods agree for individuals
- ❖ For symmetric distributions, 95% of the observations will lie within the range

Mean difference \pm 1.96 SD differences (*)

- ❖ 95% limits of agreement (LoA)
- ❖ But, valid use of these limits depends on adherence to two assumptions !
(see next slides)

Assumption 1 for calculation of limits of agreement*

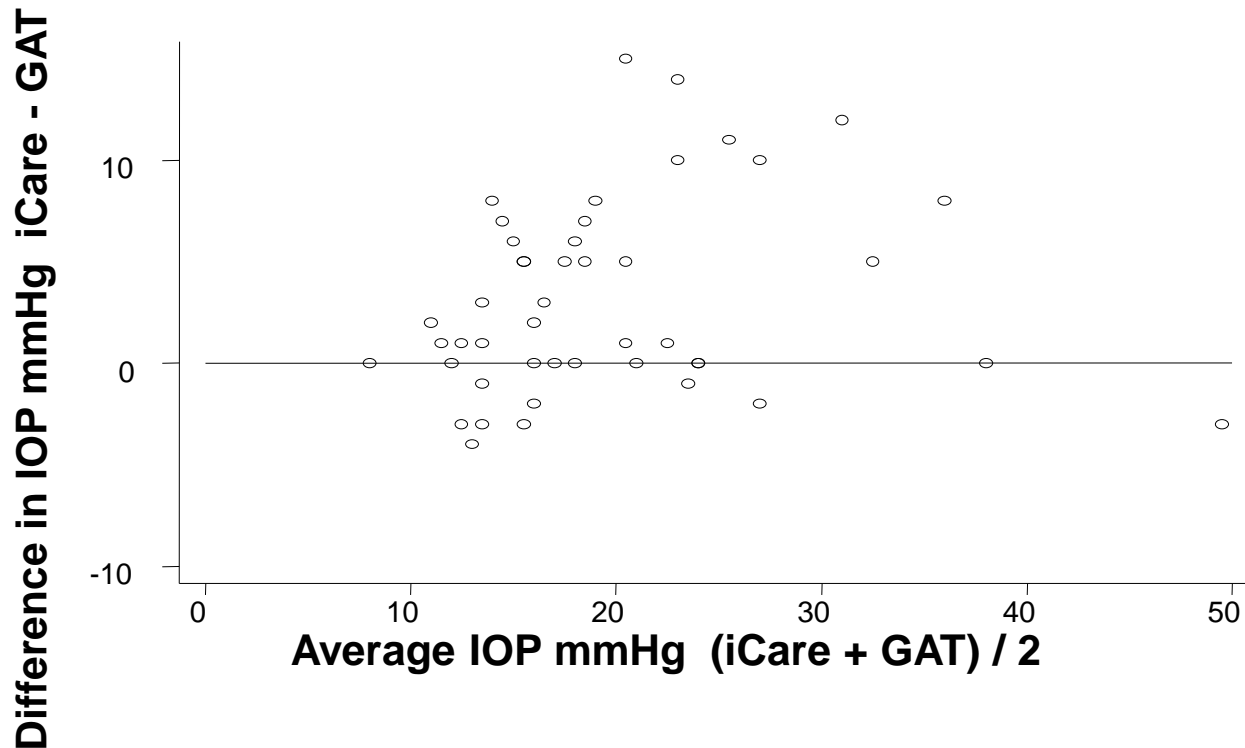
Assumption 1: The differences are approximately normally distributed



Question: Is the assumption 1 satisfied here?

Assumption 2 for calculation of limits of agreement*

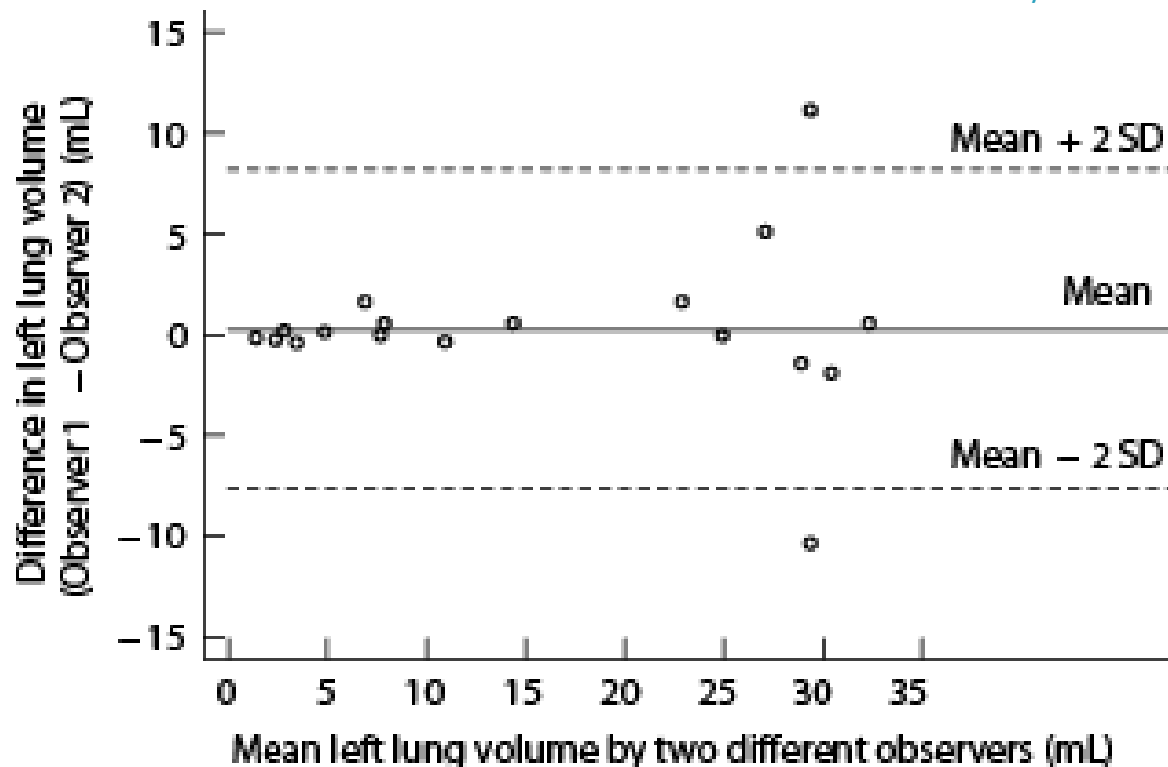
Assumption 2: There is no relationship between the difference and the magnitude of the characteristic being assessed.



Question: Is the assumption 2 satisfied here?

Example 2: Agreement between two observers, in measuring the lung volume

Applying the right statistics: analyses of measurement studies.
Bland JM, Altman DG.
Ultrasound Obstet Gynecol. 2003 Jul;22(1):85-93. Review.



Question 1: Is the assumption of normality of differences satisfied?

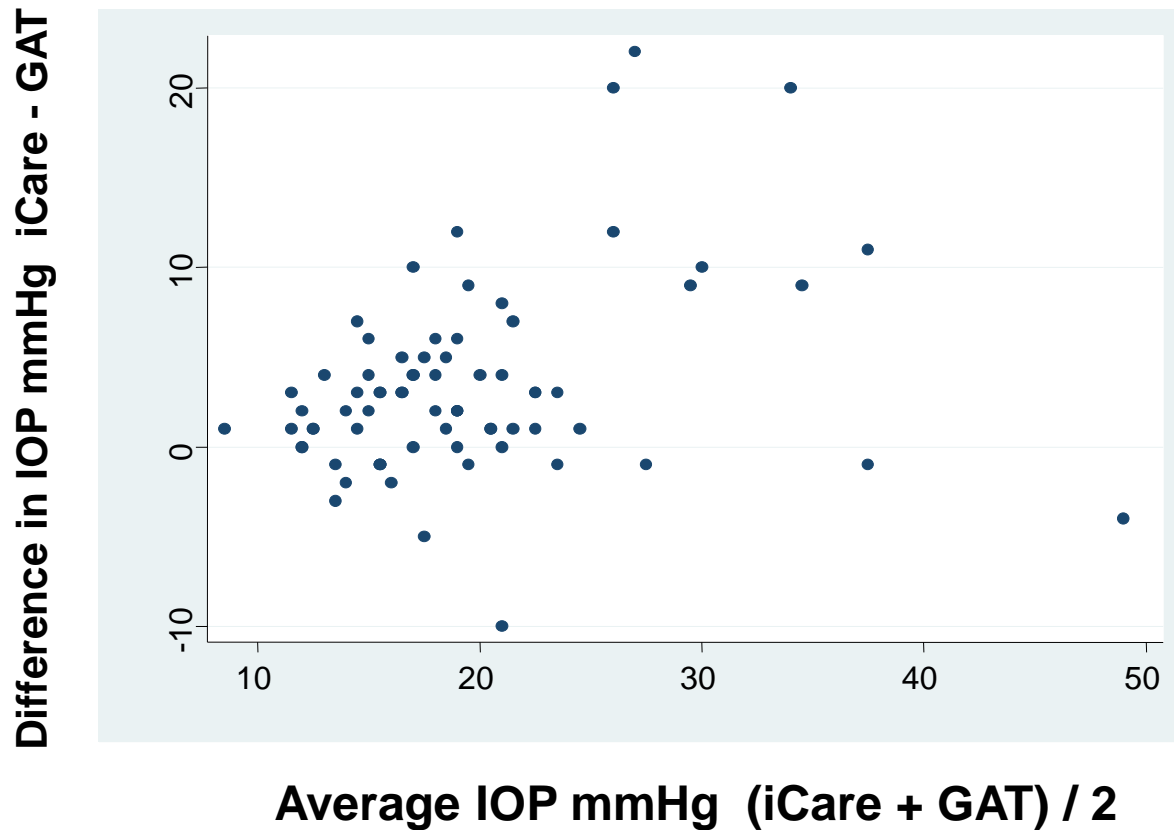
Question 2: Is the assumption of no relationship between difference and mean satisfied?

Example 1. Intraocular pressure: Precision

- ❖ 95% limits of agreement (-6.24 mmHg, 12 mmHg)
- ❖ For 95% of individuals a measurement made by the iCare tonometer will lie between 6.24 mmHg less and 12 mmHg more than a measurement by the Goldmann tonometer
- ❖ Ideally, before you start the study you will define how narrow the range should be in order for you to decide that methods agree. This is a clinical rather than a statistical decision.

Example 1. Intraocular pressure

Bland-Altman Plot using “reliable” iCare data

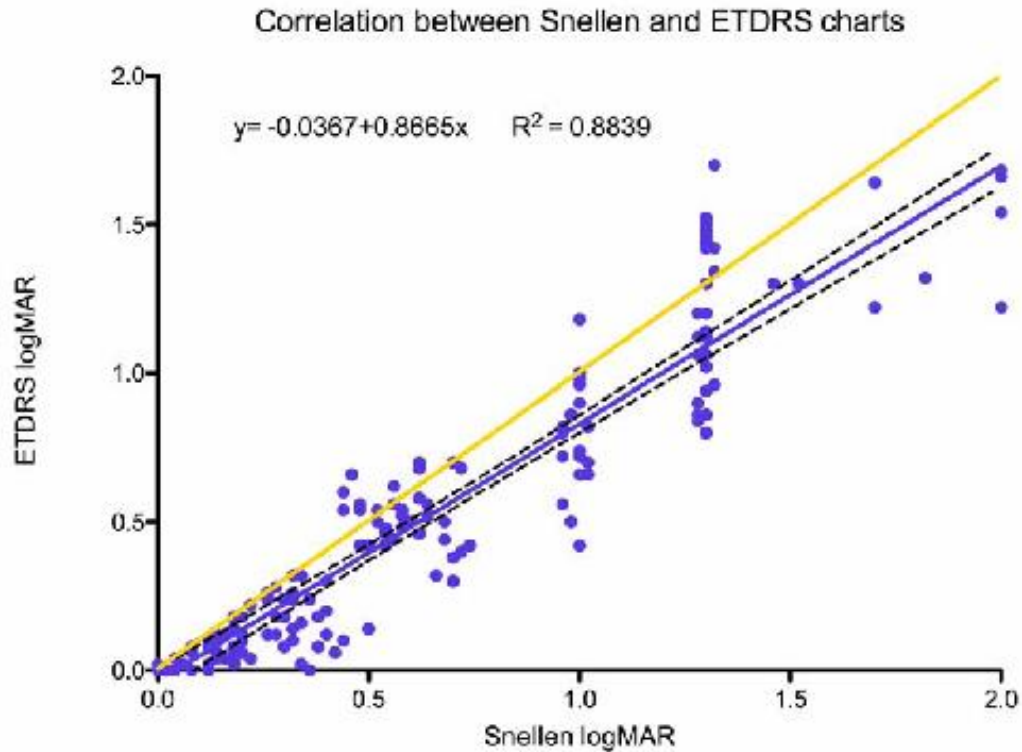


Example 3: Visual acuity – ETDRS vs Snellen

- ❖ ETDRS vs. Snellen
- ❖ Measurements made on both charts can be expressed in LogMAR but, is an ETDRS visual acuity expressed in LogMAR the same as a Snellen visual acuity in LogMAR?
- ❖ 163 eyes from 163 patients. Both charts used.

Prospective evaluation of visual acuity assessment: a comparison of Snellen versus ETDRS charts in clinical practice (An AOS Thesis). Kaiser PK. Trans Am Ophthalmol Soc. 2009 Dec;107:311-24

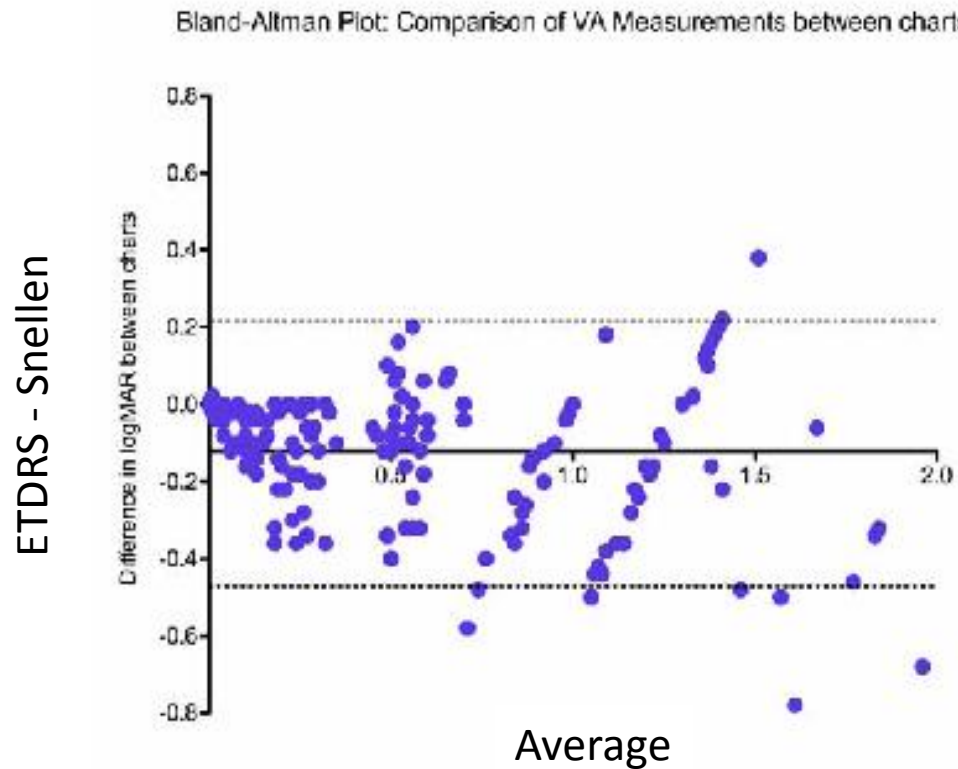
Example 3: Visual acuity – ETDRS vs Snellen



- Yellow line = equivalence line ($y=x$ line)
- Blue line = regression line
- - - Dotted lines = 95 % CI of regression line

Question: Which measure gives better vision score on logMAR?

Example 3: Visual acuity – ETDRS vs Snellen



Solid line = mean difference in LogMAR acuity (-0.13)
Dotted lines = 95 % limits of agreement.

Example 3: Visual acuity – ETDRS vs Snellen

Limits of agreement (LoA) calculated for three categories of vision:

- ❖ “Poor” vision ($< 6/60$) (n = 56)
 - ❖ Average disagreement 10 ETDRS letters
 - ❖ **LoA (-14.5, 34.5) letters**

- ❖ “Intermediate” vision (n = 38)
 - ❖ Average disagreement 5 ETDRS letters
 - ❖ **LoA (-9.7, 19.7) letters**

- ❖ “Good” vision ($> 6/18$) (n = 69)
 - ❖ Average disagreement 4 ETDRS letters
 - ❖ **LoA (-5.8, 13.8) letters**

Example 3: Visual acuity – ETDRS vs Snellen

Conclusions from the study:

- ❖ As visual acuity worsens, disagreement between charts increases
- ❖ This was also shown in a study using data from 104 participants, 80 of whom had some degree of AMD. They concluded that there was “poor” agreement between Snellen and ETDRS charts which was more pronounced in “poor vision”.

Comparison of visual acuity in macular degeneration patients measured with snellen and early treatment diabetic retinopathy study charts. Falkenstein IA, Cochran DE, Azen SP, Dustin L, Tammewar AM, Kozak I, Freeman WR. *Ophthalmology*. 2008 Feb;115(2):319-23.

Example 3: Visual acuity – ETDRS vs Snellen

Why does this matter?

- ❖ It means that a **patient with poor vision**, measured with the ETDRS chart will have a better LogMAR score than they would have had they been assessed on the Snellen chart.
- ❖ Current guidelines for registrations as sight impaired (which often acts as a trigger for social support) relate to measurements made on Snellen charts – patients who would have been eligible for CVIs (certificate for vision impairment) may no longer meet the criteria

What we learned

- The proper statistical method of agreement for **measurements on continuous scale is**
 - Bland-Altman plot together with limits of agreement
 - We should not use correlation coefficient to describe the agreement

Challenge – Correlation coefficient still (wrongly) persists as main method of evaluating the agreement in published studies

- ❖ Correlation coefficient still prevails and limits of agreement are not in use or are clearly misunderstood
- ❖ This message was highlighted in 1983
- ❖ The message has been repeated – in general medical journals and in ophthalmic journals

Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; i: 307-310

Patton N, Aslam T, Murraray G. Statistical strategies to assess reliability in ophthalmology. *Eye* 2005 1-6

Bunce C. Correlation, agreement and Bland Altman analysis – statistical analysis of method Comparison studies. *Am J Ophthalmol.* 2009 Jul;148(1):4-6

WHY?

Challenge – Correlation coefficient still (wrongly) persists as main method of evaluating the agreement in published studies

WHY?

1. Resistance to change?
2. Desire to conform, follow what others have done?
3. Correlation is a word in common usage. The statistical definition is more specific than the non-statistical definition. When a researcher speaks about correlation they are not necessarily speaking about linear association and do not therefore comprehend that this is what a correlation coefficient is doing
4. The need to think about your results: you don't get a yes or no to "do they agree", you get a range and have to think, is this reasonable
5. Methods of Bland and Altman – simple if assumptions adhered to – not so straight forward if not
6. No P-value?
7. Other suggestions? (from audience)

Event: NIHR Statistics Group Ophthalmology Research Section Meeting on “Methods of agreement studies”

- Time: 3 November 2016
- Place: London, Moorfields
- Invited talks:
 - Professor Garway Heath – consultant ophthalmologist
 - Professor Martin Bland – statistician – his view of correlation persistence

Method Comparison Studies in Ophthalmology

Additional References

- Dewitte K, Fierens C, Stockl D, Thienpont LM. Application of the Bland–Altman Plot for Interpretation of Method -Comparison Studies: A Critical Investigation of Its Practice. *Clin. Chem.*, May 2002; 48: 799 - 801.
- Mantha S, Roizen MF, Fleisher LA, Thisted R, Foss J. Comparing Methods of Clinical Measurement: Reporting Standards for Bland and Altman Analysis. *Anesth Analg.* 2000 Mar;90(3):593-602.
- Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res.* 1999 Jun;8(2):135-60. Review. PubMed PMID: 10501650.
- Donner A, Eliasziw M. Sample size requirements for reliability studies. *Stat Med.* 1987 Jun;6(4):441-8. PubMed PMID: 3629046.
- 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 Jan;64(1):96-106. doi:10.1016/j.jclinepi.2010.03.002. PubMed PMID: 21130355.
- Shoukri MM, Asyali MH, Donner A. Sample size requirements for the design of reliability study: review and new results. *Stat Methods Med Res.* 2004;13:251–271.
- Bland JM, Altman DG. Comparing methods of measurement: why plotting difference against standard method is misleading. *The Lancet.*

THANK
YOU!

