DOI: 10.7860/JCDR/2017/23897.9314

View Point

Education Section

Method Comparison (Agreement) Studies: Myths and Rationale

AJAY G PHATAK[1], SOMASHEKHAR M NIMBALKAR[2]

ABSTRACT

Unprecedented technological growth in the last quarter of twentieth century has resulted in improved health care and opened new domains of health care research. This technological leap also facilitated the paradigm shift from hospital care to home care through development of 'point of care' devices. As early diagnoses and timely referral is a key to health management, these devices play an important role in improving health. Validation of the new technology in different settings is necessary before adopting it to practice. For a binary result like pregnancy test, it is trivial to use statistical tools like sensitivity, specificity etc. For a continuous variable like blood glucose level the analysis is not straightforward. Many of us misinterpret 'association' as 'agreement'. This misinterpretation is reflected in studies which have compared two different technologies. The findings of well conducted studies do not contribute to the evidence base just because of wrong analysis strategy. We delineate on finer nuances of analysis and interpretation of method comparison studies.

Keywords: Agreement, Association, Bland-Altman, Correlation, Method Comparison

INTRODUCTION

The last quarter of twentieth century witnessed unprecedented growth in technology. Advent of faster and miniature computers had cascading effect leading to exponential growth of technology in almost all fields. This technological leap has not only resulted in better health care but also opened new avenues in health care research. While CT scan, MRI, Linear accelerator with multi-leaf collimators etc., facilitated better outcome and improved Quality of Life, Microarray experiments is a ray of hope in identifying genes responsible for particular diseases so that gene therapy could be developed to manage such diseases effectively.

A paradigm shift in health management from hospital care to home care that has started in developed countries will eventually find its way into the developing world with some time lag. This paradigm shift is supported partially through development of 'point of care' devices in various fields of health care especially diagnostic tools. Most of us in India have been using such devices such as digital thermometers, blood pressure monitors and even pregnancy test kits with varying levels of confidence. Even in laboratory setting, manual methods are steadily replaced by Auto Analysers. As greater focus is placed on early diagnosis and timely referral from remote areas, it is likely that more and more such point of care devices will find their way in health care management. So, now one can check his/her glycaemic control or determine whether she is pregnant or not at home.

While, it is quite tempting to use the 'point of care' devices because of its portability, simplicity and many times non-invasiveness, it is always important to validate the new technology before adopting it. For a binary result (like in pregnancy test) it is trivial to use statistical tools like sensitivity, specificity etc. The theoretical dilemma creeps in when a researcher wants to validate the result of a continuous variable like blood pressure, blood sugar level etc. In this scenario it is incumbent on the academic research community to provide appropriate ways and means to validate various such devices.

The technological advancement provided new challenges to the discipline of statistics. In statistics circles, a common banter is "the half-life of statistical knowledge is more than the life span of an elephant". It appears that statistics has acquired some pace. Contemporary techniques like Structural Equation Modeling, path analysis etc., were developed as per the need to deal with complex situations in research. For example, while dealing with microarray experiments; the signal and noise are entangled so much that statistics had to come up with new techniques to separate them. Statistics as a discipline appears to be very much successful in addressing the new challenges in health care research.

The Right Question: Before adopting a new method; we need to ensure that it consistently provides accurate results. In other words, we want to know whether the new method can replace the standard method. The theoretical dilemma starts with the statisticians' habit of plotting the data to get the feel of the data. As the scatterplot is constructed; rightly so for a bivariate data, correlation/regression analysis is done intuitively. Regrettably many of us fail to understand that correlation/regression is a measure of association and NOT of agreement. Excellent association can be achieved when the data is clustered around any straight line but the agreement is good only when it is clustered around line of equity (Y=X). Moreover, because both the methods are measuring same quantity, we expect good association theoretically.

This monograph delineates on finer nuances of analysis and interpretation of method comparison studies. It also reiterates the conceptual framework of "Method Comparison Studies" and highlight common mistakes done by researchers in analysis and reporting of the findings.

Quantitation of HIV-1 RNA level is important serological and molecular diagnosis of HIV infection. Few studies compared the levels in Dried Blood Spot (DBS) and Dried Plasma Spot (DPS) specimens using the same assay while few studies compared the plasma samples using two (or more) different assays. The astute reader will quickly realize that without loss of generality, the core question is same: whether we can replace one method with the other. Choi JY et al., compared HIV-1 RNA level in plasma using two different assays [1], while Reigadas S et al., compared the HIV-1 RNA level in plasma and dried blood spot using the same assay [2]. Both presented the correlation coefficient/ regression analysis to conclude satisfactory agreement. Uslu S et al., compared different methods of temperature measurement in sick

newborns [3]. He also reported correlation coefficient/ regression analysis and concluded satisfactory agreement. The authors and the journal editors both missed the difference between association and agreement and the egregious analysis that followed.

To elaborate the difference, imagine two bathroom scales: one digital and one analogue. Suppose that both the scales provide consistently accurate results but due to some electronic circuit problems, the digital scale shows exactly two times the actual weight. The correlation coefficient in this case is exactly +1 but any scientist will not allow the use of digital scale (may be except bariatric surgeons).

Other methods of assessing associations like t-test/Analysis of Variance (ANOVA) are also inappropriate to analyse agreement studies for the same reasons. Duran R et al., used ANOVA for comparing 3 methods of measuring temperature of low birth weight pre-term infants [4]. He concluded good agreement because there was no statistically significant difference between mean temperatures of mid-forehead and axillary temperatures.

Bhaskar Shahbabu et al., reported multiple t-tests and analysis based on dichotomizing a scale variable for comparing accuracy of aneroid and digital sphygmomanometers in reference to mercury sphygmomanometers [5]. The authors concluded superiority of aneroid over digital sphygmomanometers.

Imagine two series of measurements viz., 1,2,3,4,5,6,7 and 7,6,5,4,3,2,1. The mean and SD of both the series are same and hence the t-test will yield no statistical difference between them. However, this indicates a very poor agreement as smallest reading of first series corresponds to highest reading of second and vice versa. The moral is “Good association does not necessarily mean good agreement”.

Recognizing the pitfall, Bland and Altman developed a method to measure agreement between two methods which involves plotting the difference against the average of the two methods to get 95% confidence limits for clinical consideration [6]. The article was published in The Lancet in 1986 and despite the overwhelming response, the technique developed by Bland and Altman has not been sufficiently adopted and we encounter papers reporting correlation/regression/ANOVA in method comparison studies. The serious issue is that the findings of well conducted studies are invalid just because of wrong analysis strategy.

Three (common) mistakes in analysis and interpretation of method comparison studies

Although these are rampant in the literature, I will restrict to just one example of each due to space constraint:

1. Completely wrong analysis: This is the most common scenario. Gupta PP and Dipti Agarwal compared Peak exploratory flow by two methods [7]. They reported the comparison using t-test and Pearson’s correlation coefficient. The authors and the journal editors both missed this egregious analysis.

2. Mentioning Bland Altman method without using it: Bland Altman method is probably misused most by not using it at all. Jaramillo et al., compared MR imaging with conventional Arthrography [8]. In the methods section Bland Altman method was cited but the results section reports correlations and rank correlations with p-values without any mention of 95% confidence limits which is an integral part of Bland Altman method. This is unacceptable when the correct analysis strategy is known to the researcher.

3. Doing the analysis properly but insufficient reporting: The quality of a published paper lies in its reproducibility and not just in generalizability. Priya M et al., compared capillary whole blood glucose with venous plasma glucose for screening in diabetes [9]. They did the analysis appropriately but the abstract section reports correlation coefficient with other relevant details without mentioning 95% confidence limits of agreement. While accepting the fact that it is readers' responsibility to read the full paper, many times readers rely on the abstract due to lack of resources. An astute reader may just skip the paper considering the authors have not done the analysis properly if the abstract does not contain Bland-Altman limits. On the contrary, a naive researcher may do similar wrong analysis in absence of full text. Further if Bland-Altman analysis is performed, no secondary inappropriate analysis is required and such results should be avoided even in main manuscript. The authors strongly urge including 95% limits of agreement in the abstract also for maintaining completeness of the abstract.

It appears that authors confuse Bland-Altman Plot with Bland-Altman Analysis. Simply plotting the Bland-Altman plot does not mean appropriate analysis. Kane CT et al., reported 94% of the results were within 2SD limits and concluded good agreement without discussing the width of the 95% confidence limits [10]. The confusion continued and David S et al., claimed good agreement between DBS and Plasma samples for the same reason [11]. If the distribution is fairly normal, about 95% observations fall between mean±2SD. This is a statistical fact and should not be used for claiming good agreement.

Mridula Madiyal et al., compared Assay Performance of ELISA and Chemiluminescence Immunoassay in Detecting Antibodies to Hepatitis B Surface Antigen [12]. The authors mentioned Bland-Alman analysis in methods section without mentioning the results of Bland-Altman Analysis in the abstract. Further, the authors correctly identified a trend in Bland-Altman plot without calculating the modified confidence limits. In conclusion section the authors presented conflicting statements. First statement approved the interchangeability of both methods based on categorization of titer values whereas second statement indicated that the methods are not interchangeable based on absolute titer values.

Other considerations: It is necessary to set 'a priory' criterion for the acceptable limits. The expected accuracy of the method as well as clinical implications should be considered while setting this criterion.

As this is a kind of 'estimation problem', the sample size should be sufficiently large. Bland recommended a minimum of 100 observations [13].

It is to be noted that there is an appropriate way to use regression that Bland himself used before developing this technique but it is quite complicated. Alanen E proposed a different technique based on Structural Equation Modeling (SEM) [14]. It is criticized that Bland Altman method considers reliability as distinct from method comparison. Bland and Altman clearly mentioned about reliability in their original article published in Lancet [6] in 1986 and advocated repeated measure design in their paper published in 1999 [15]. An excellent step by step illustration of analysis and interpretation of method comparison studies is provided by Hanneman SK [16]. Gwet KL elaborated few shortcomings of Bland-Altman method under specific situations but he also endorsed the simplicity of the technique especially in exploratory stage [17].

Another commandment in method comparison is that it is insufficient to validate the technology alone. The study setting, study population etc. may influence the 95% confidence limits to some extent. The instrument, its calibration, least count and workmanship also influence the agreement and these along with validity of the technology should be considered before adopting it in practice. This may appear farfetched but unfortunately quite common in science. Chiappini E et al., found good agreement between infrared thermometry and axillary thermometry under ideal setting and recommended its use in neonatal setting [18]. The authors found unacceptable agreement (and wider 95% limits of agreement) when replicated the study with another brand and in different setting [19]. Fortuna EL et al., and Sener S et al.,

reported similar findings [20,21]. Unfortunately, few well conducted studies could not contribute to the evidence base just because of inappropriate analysis [3,4].

CONCLUSION

There is a difference between "Association and Agreement". When the outcome is scale (continuous) variable, correlation/regression/ANOVA should not be used to report agreement between two methods of measurements. Bland-Altman method can be considered as a first step of a continuous process of validation and quality improvement. Popularity of Bland Altman method stems from the essence of its simplicity in conceptual understanding and practical application.

REFERENCES

[1] Choi JY, Kim EJ, Rho HJ, Kim JY, Kwon OK, Lee JH, et al. Evaluation of the NucliSens EasyQ HIV-1 v1.1 and RealTime HIV-1 kits for quantitation of HIV-1 RNA in plasma. J Virol Methods. 2009;161(1):7-11.

[2] Reigadas S, Schrive MH, Aurillac-Lavignolle V, Fleury HJ. Quantitation of HIV-1 RNA in dried blood and plasma spots. J Virol Methods. 2009;161(1): 177-80.

[3] Uslu S, Ozdemir H, Bulbul A, Comert S, Bolat F, Can E, et al. A comparison of different methods of temperature measurements in sick newborns. J Trop Pediatr. 2011;57(6):418-23.

[4] Duran R, Vatansever U, Acunaş B, Süt N. Comparison of temporal artery, mid-forehead skin and axillary temperature recordings in preterm infants <1500 g of birthweight. J Paediatr Child Health. 2009;45(7-8):444-47.

[5] Shahbabu B, Dasgupta A, Sarkar K, Sahoo SK. Which is More Accurate in Measuring the Blood Pressure? A Digital or an Aneroid Sphygmomanometer. Journal of Clinical and Diagnostic Research. 2016;10(3):LC11-LC14.

[6] Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;1(8476):307-10.

[7] Agarwal D, Gupta PP. A comparison of peak expiratory flow measured from forced vital capacity and peak flow meter manoeuvres in healthy volunteers. Ann Thorac Med. 2007;2(3):103-06.

[8] Jaramillo D, Galen TA, Winalski CS, DiCanzio J, Zurakowski D, Mulkern RV, et al. Legg-Calvé-Perthes disease: MR imaging evaluation during manual positioning of the hip--comparison with conventional arthrography. Radiology. 1999;212(2):519-25.

[9] Priya M, Mohan Anjana R, Pradeepa R, Jayashri R, Deepa M, Bhansali A, et al. Comparison of capillary whole blood versus venous plasma glucose estimations in screening for diabetes mellitus in epidemiological studies in developing countries. Diabetes Technol Ther. 2011;13(5):586-91.

[10] Kane CT, Ndiaye HD, Diallo S, Ndaiye I, Wade AS, Diaw PA, et al. Quantitation of HIV-1 RNA in dried blood spots by the real time Nuclisens EasyQ HIV-1 assay in Senegal. J Virol Methods. 2008;148:291-95.

[11] David S, Sachithanandham J, Jerobin J, Parasuram S, Kannangai R. Comparison of HIV-1 RNA level estimated with plasma and DBS samples: a pilot study from India (South). Indian J Med Microbiol. 2012;30(4):403-06.

[12] Madiyal M, Sagar S, Vishwanath S, Banerjee B, Eshwara VK, Chawla K. Comparing Assay Performance of ELISA and Chemiluminescence Immunoassay in Detecting Antibodies to Hepatitis B Surface Antigen. Journal of Clinical and Diagnostic Research. 2016;10(11):DC22-DC25.

[13] Available at http://www-users.york.ac.uk/~mb55/meas/sizemeth.htm assessed on 3rd September, 2016.

[14] Alanen E. Everything all right in method comparison studies? Stat Methods Med Res. 2012;21(4):297-309.

[15] Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res. 1999;8(2):135-60.

[16] Hanneman SK. Design, analysis, and interpretation of method-comparison studies. AACN Adv Crit Care. 2008;19(2):223-34.

[17] Gwet KL. Handbook of interrater reliability: The definitive guide to measuring the extent of agreement among multiple raters, Fourth ed. Gaithersburg, MD: Advanced Analytics LLC. 2014.

[18] Chiappini E, Sollai S, Longhi R, Morandini L, Laghi A, Osio CE, et al. Performance of non-contact infrared thermometer for detecting febrile children in hospital and ambulatory settings. J Clin Nurs. 2011;20(9-10):1311-18.

[19] Sethi A, Patel D, Nimbalkar A, Phatak A, Nimbalkar S. Validation of Forehead Infrared Thermometry in Neonates. Indian Pediatr. 2013;50(12):1153-54.

[20] Fortuna EL, Carney MM, Macy M, Stanley RM, Younger JG, Bradin SA, et al. Accuracy of non-contact infrared thermometry versus rectal thermometry in young children evaluated in the emergency department for fever. J Emerg Nurs. 2010;36(2):101-04.

[21] Sener S, Karcioglu O, Eken C, Yaylaci S, Ozsarac M. Agreement between axillary, tympanic, and mid-forehead body temperature measurements in adult emergency department patients. Eur J Emerg Med. 2012;19(4):252-56.

PARTICULARS OF CONTRIBUTORS:

1. Manager, Central Research Services, Charutar Arogya Mandal, Karamsad, Anand, Gujarat, India.
2. Head, Central Research Services, Charutar Arogya Mandal, Karamsad, Anand, Gujarat, India.

NAME, ADDRESS, E-MAIL ID OF THE CORRESPONDING AUTHOR:

Mr. Ajay G Phatak,
Central Research Services, Academic centre Charutar Arogya Mandal Gokal Nagar, Karamsad, Anand, Gujarat, India.
E-mail: ajaygp@gmail.com

FINANCIAL OR OTHER COMPETING INTERESTS: None.

Date of Submission: **Sep 04, 2016**
Date of Peer Review: **Oct 12, 2016**
Date of Acceptance: **Nov 16, 2016**
Date of Publishing: **Jan 01, 2017**