ORIGINAL ARTICLE

Intraclass correlation in practice: Assessment of reliability of manual blood pressure measurement among assessors in a clinical trial

Wan Nor Arifin[1], Wan Arfah Nadiah[1], Muhammad Irfan[1], Chen Xin Wee[3], Nani Draman[2], Nyi Nyi Naing[1]

[1]Biostatistics and Research Methodology Unit, [2]Department of Family Medicine, [3]medical student, School of Medical Sciences, Universiti Sains Malaysia, Kelantan, Malaysia

Abstract

Background: To ensure the reliability of manual blood pressure (BP) readings in a clinical trial, sources of error due to measurement must be reduced as much as possible. Apart from following standard procedure for BP measurement and ensuring good equipments, the measurement errors that come from the assessors themselves should be assessed. **Objective**: To demonstrate the use of two-way random effects, interactions absent, absolute agreement (Type A), single measures (Type 1) intraclass correlation coefficient (ICC) in the assessment of reliability of manual BP readings among assessors involved in a clinical trial using manual BP measurement, by using an interrater reliability study conducted by the authors as an example study. **Methods**: The steps involved in obtaining ICC in the study were discussed. Sample size given the number of assessors in the study was calculated. BP was measured using regularly maintained mercury sphygnomanometers, following recommendations by Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure (JNC 7) for BP measurement in office setting. The outcomes were systolic and diastolic BP readings. A type of ICC, two-way random effects, interactions absent, absolute agreement (Type A), single measures (Type 1) ICC was chosen for the analysis and specifically discussed. Pre-requisite assumptions for ICC were meticulously checked and described. The interrater reliability for systolic and diastolic BP readings as expressed by ICC (single measure) were presented with confidence interval (CI). The ICCs obtained in the example study were discussed and concluded. The flaws of the study were also criticised. **Results**: The interrater reliability for systolic and diastolic BP measurements as expressed by ICC (single measure) were 0.87 (95% confidence interval [CI] 0.722, 0.956) and 0.77 (95% CI 0.560, 0.918) respectively. **Conclusion**: We demonstrated the steps required to obtain ICC. Since the use of manual BP measurement using mercury sphygmomanometer is still considered as gold standard of BP measurement, it is important that studies in which the BP outcome is measured using such method conduct interrater reliability studies properly.

Keywords: interrater reliability, intraclass correlation coefficient, manual blood pressure, mercury sphygmomanometer, blood pressure measurement standardization

Introduction

Blood pressure measurement using mercury sphygmomanometer remains a gold standard for BP measurement [1, 2]. As such, it depends very much on the reliability of human operator, apart from other factors that may affect the validity and reliability of of a BP reading. Thus, it is important to establish reliability of an assessor or assessors of BP measurement should a researcher want to conduct a clinical trial in which the outcome is BP reading. The specific reliability that is of concern is interrater reliability.

In general, to obtain a valid and reliable blood pressure (BP) reading, it is important to address and reduce the possible errors in measurement. Errors generally divided into two categories: systematic and random errors [3]. In BP measurement, the systematic errors may be due to poorly maintained equipments, non-standard procedures, and assessors' digit preference [2-4]. The random errors may come from use of unreliable equipments and also unreliable assessors such as due to poor concentration, erroneous recording, and poor technique [2, 3, 4, 5]. The systematic errors affect the validity or accuracy of a BP reading, while the random errors affect its reliability [3]. Both facets should be addressed in a reliability study, in which the systematic errors should be controlled with well planned methodology, while the random errors should be established objectively with statistical analysis.

Reliability word is synonymous to the words repeatability, reproducibility, consistency, and precision [3, 6, 7]. To be more specific, it is “the extent to which repeated measurements of a stable phenomenon – by different people and instruments, at different times and places – get similar result” [3].

Speaking of interrater reliability, intraclass correlation coefficient (ICC) is a common way to assess such reliability. ICC is a statistical method to measure reliability of two or more raters or assessors when data measurement is on interval scale [8]. It measures the relationship among variables of a common class which share common metric and variance, hence the term intraclass [9]. In line with our discussion, the manual BP measurement mainly consists of two components that make it suitable for this type of analysis; BP reading is in interval scale, and it involves multiple assessors. The BP readings of the assessors would be assessed for their consistency between themselves and then expressed in term of ICC. With this, we obtain an objective way to look at the reliability.

We did a literature search using Google Scholar with search terms: "blood pressure" ("mercury sphygmomanometer" OR "mercury manometer") ("clinical trial" OR "controlled trial"), restricted to articles published between 2006 to 2010, in medicine, pharmacology, and veterinary science., which resulted in 1210 studies. Once we added "intraclass correlation" to the search terms, only 28 matches were found, or 2.3 percent. We also conducted a literature search using similar strategies on PubMed, without restricting the date of publication and field of science, and found three studies with "intraclass correlation" term added out of 113 studies with the term removed, or 2.7 percent. Thus, we feel that intraclass correlation use in clinical trial involving manual blood pressure measurement using mercury sphygmomanometer is lacking. However it should be noted that the literature search was conducted to check for the availability of such studies, not for the purpose of appraising each of the studies in detail.

Unlike studies using questionnaires as measurement tools, which most of the time reported or made reference to relevant validation studies, it seems that interrater reliability is neglected. In our opinion, interrater reliability is not something to be neglected by researchers, particularly when human factor contributes to the measurement error, as such in manual BP measurement.

In this paper, it is our aim to demonstrate the use of two-way random effects, interactions absent, absolute agreement (Type A), single measures (Type 1) intraclass correlation (ICC) in the assessment of reliability of manual BP readings among assessors in a clinical trial. We use our previous study on blood pressure reduction among hypertensive patient using "One Minute Exercise" qigong technique as an example to illustrate the methods involved. The validity of the measurement was controlled at the methodology of the study.

Methodology

The study

Interrater reliability of assessors in clinical trial on blood pressure reduction among hypertensive patients using "One Minute Exercise" qigong technique was conducted.

Study Subjects

The study was approved by School of Medical Sciences and Human Ethical Committee, Universiti Sains Malaysia. Participants for this reliability study were recruited among staffs in Biostatistics and Research Methodology Units, and Department of Medical Education under the School of Medical Sciences, which consisted of medical lecturers and administrative staffs. Such sampling frame was chosen as the staffs were easily approachable, logistic reason, time constraint and budget constraint. Every staffs in respective departments were approached personally for voluntary involvement in the study, and explanations were given with regard to the study. Participants were adults aged 18 years and above, and were able to understand study protocol and simple instructions. Participants without serious medical illnesses were included in the study. Participants who had serious heart disease, respiratory disease, renal failure, liver failure, or diagnosed with malignancy were excluded from this study. Participants were selected on the basis of voluntary participation and fulfillment of inclusion and exclusion criteria. All participants (n = 11) approached for the study fulfilled the criteria and agreed to participate in the study.

Study Design

A reliability study on BP measurement among assessors was conducted on May 13 2010. This study was conducted to establish reliability of BP readings among assessors participating in a subsequent clinical trial (Wee et al., unpublished manuscript), which requires regular BP measurement. The study was conducted between 10am to 1pm, which took 3 hours to be completed. There were five participating assessors.

Pre-calculated sample size or number of subjects, k was 10.3 using Walter, Eliasziw and Donner [10] table for determining sample size for reliability study based on ICC (acceptable reliability, ρ_0 = 0.7; desired reliability, ρ_1; number of observations per subject, n = 5 (five assessors hence five repeated observations); significant level, α = 0.05; power of study, $1\text{-}\beta$ = 0.8). n notation used by Walter, Eliasziw and Donner [10] is not to be confused with commonly used n notation to denote sample size or study subjects, in which k is used in place to denote number of

subjects. The sample size of 11 subjects for this study was adequate.

Outcome Measurement

The outcomes for this study were BP readings, which were measured using mercury sphygmomanometers. To maintain the standard in BP measurement, the following recommendations by Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure (JNC 7) on BP measurement in office settings [11] were adapted:

1. Trained operator.
2. Auscultatory method of BP measurement.
3. Patient will be seated comfortably in a chair for at least 5 minutes, with his/her feet on the floor and arm positioned and supported at heart level.
4. Caffeine, smoking and exercise 30 minutes prior to measurement is contraindicated.
5. Appropriate sized cuff, with cuff bladder encircling at least 80 percent of arm.
6. At least 2 readings made and average recoded as BP reading.
7. Palpated radial pulse obliteration pressure is used to estimate systolic BP.
8. Cuff then inflated to 20-30mmHg above the estimated systolic BP based on palpation.
9. Systolic BP is taken as a point when the first two or more Korotkoff sounds are heard (onset of phase 1).
10. Diastolic BP is taken as a point of disappearance of Korotkoff sound (onset of phase 5).

Procedure

Five mercury sphygmomanometers and stethoscopes were obtained from Multidisciplinary Laboratory of Universiti Sains Malaysia, where the instruments were regularly checked and maintained. All five assessors had a briefing session on JNC 7 [11] standard in BP measurement to ensure similar steps and techniques were used during BP measurement session. The assessors were well versed with BP measurement, of which four are medical doctors and one is a trained nurse. Participants were seated comfortably in a chair for at least five minutes on prior to BP measurement with feet on the floor. Screening for caffeine, smoking or exercise 30 minutes prior to BP measurement was not done. The sphygmomanometer cuffs used were of suitable and appropriate size for the participants. The arm on which the cuff was applied was supported at heart level on a table next to the chair. Steps 7 to 10 of JNC 7 [11] as listed before were properly followed by every assessors. Each assessor measured two BP readings for each participant and the readings were written in a table with two columns per assessor per subject (Table 1). The average was taken as BP reading for the participant, which was written in one column below the two BP readings columns (Table 1). The assessors were unaware of other assessors BP readings until they wrote down their readings in the table. The participants were informed of their BP readings after the session.

Data entry and statistical analysis

Data entry and analyses were done using PASW version 18. The systolic and diastolic BP readings for each participant were keyed both in wide and long format. Demographic characteristics of participants were described in frequencies and percentages for categorical characteristics (gender, occupation and hypertensive status). Continuous

characteristic (age) was described in term of mean and standard deviation after checking for normality.

To check for underlying assumptions of ICC two-way model, residuals of systolic and diastolic BP readings were inspected for equality of variances across observers, normality and outliers. Visual exploration of the residuals was done by using residual plot against fitted values for equality of variances across observers; histograms with normality curve, box plots and normal probability plots (Q-Q and P-P plots) for normality check; while visual inspection for outliers was facilitated by using stem-and-leaf, box plots and residual plots against fitted values. Statistically, normality of the residuals was checked with Shapiro-Wilk test of normality, while Brown-Forsythe test (using median) was used to test for equality of variances. Non-multiplicative interaction (additivity) of the model was checked with Turkey's test of non-additivity.

Interrater reliability for BP readings was assessed with ICC, separately done for systolic and diastolic readings. Taking into account the variability due to subjects and also due to assessors, the ICC for each systolic and diastolic BP was done with two-way random model [12]. Absolute agreement type [9] was chosen for ICC analysis as we were interested to know whether the BP readings were similar among assessors and as systematic variability due to assessors was important. Single measures SPSS output of the ICC analysis would be taken as ICC values for this study indicative of reliability of BP readings from a single assessor [8]. Case A (absent interaction) model was chosen as it was assumed that there was no interaction between assessor factor and subject factor. In short, ICC two-way random effects, interactions absent, absolute agreement (Type A), single measures (Type 1) was chosen, or denoted as ICC Case 2A (A,1) [9].

Results

All 11 (100%) participants remained throughout the study. Demographically, six (54.5%) of them were females and five (45.5%) were males. Six (54.5%) participants were medical lecturers and five (45.5%) participants were administrative staffs of different positions. Two (18.2%) of them were known to be hypertensive on medical treatments, while the remaining nine (81.8%) participants had no known medical illness. The mean age was 39.6 (14.09).

The residuals of systolic BP readings plots against observers showed equality of variances for all observers except observer 4 which had a larger extent of scatter of the residuals around zero. The residuals of diastolic BP readings against observers did not show any marked difference in the extent of scatter of the residuals around zero for all observers, indicative of equality of variances. Brown-Forsythe test (using median) showed that the residuals of systolic and diastolic BP across observers had equal variances, with p=0.104 for residuals of systolic BP and p=0.783 for residuals of diastolic BP. For normality check, the residuals of systolic BP readings across observers were combined in consideration for small factor level sample size and equality of variances [13]. The combination was also done to residuals of diastolic BP readings. Histograms with normality curve, box plots, P-P and Q-Q plots for the residuals of systolic and diastolic BP showed normal distribution. There were no extreme outliers noted on inspection of stem-and-leaf, box plots and residual plots against observers. Shapiro-Wilk test of normality resulted in p=0.217 for residuals of systolic BP and p=0.121 for residuals of diastolic BP, indicative of normality for the residuals. Tukey's tests of non-additivity resulted in p=0.28 for systolic BP readings and p=0.20 for diastolic BP readings, with insignificant results

indicative of additive (non-multiplicative) interactions between subjects and assessors in the ICC model.

The interrater reliability for systolic and diastolic BP readings as expressed by ICC (single measure) were 0.87 (95% confidence interval [CI] 0.722, 0.956) and 0.77 (95% CI 0.560, 0.918) respectively.

Discussion

The reliability of BP readings for the assessors was established in this study. In reference to Altman [14] guidelines on interpretation of interrater agreement (Kappa) value in relation to its strength of agreement, the reliability for systolic BP readings was very good (0.81 - 1.0) with ICC (single measure) of 0.87, while the reliability for diastolic readings was considered good (0.61 – 0.8) with ICC (single measure) of 0.77. Even though the guidelines by Altman [14] were laid out for Kappa, which was intended for measure of agreement for categorical data, it could be interpreted similarly for ICC [8].

To recapitulate, for the analysis of the reliability, ICC two-way random effects, interactions absent, absolute agreement (Type A), single measures (Type 1) or ICC Case 2A (A,1) was chosen. Two-way model was chosen as row factor (subjects or targets) and column factor (assessors or number of readings) were taken into consideration. Both factors were considered as random factors. Basically it is a randomized block design of ANOVA.

However, in this study, the subjects were not selected by random sampling, and the number of assessors was fixed. How can both of the factors considered as random? For the subjects, it was ideal to choose them on random sampling as opposed to convenient sampling used. Thus it would threaten the generalizably of the results to subjects other than those involved in the study as one of the basic assumption of ANOVA was violated. This is the strict use of the term random, which is impractical as it is not always possible.

Factor can be considered random even when the sampling is not random as pointed out by Jackson and Brashers [15] and they outlined three ways to decide whether a factor is random or fixed. Firstly, replaceability test in which if the subjects or levels can be replaced with other subjects without affecting the research question, thus the factor is random. Secondly, it depends whether we want to restrict conclusion to the selected subjects (levels) only, or we want to generalize to other subjects (levels). Thirdly, if making conclusion on the subjects (levels) of a factor is meaningful enough, then the factor is fixed, otherwise it is random. We found out that our subjects were replaceable; it did not matter who were chosen as long as there were subjects for our study. We also wanted to conclude the study to other subjects, who were the subjects of a following clinical trial, thus we were unwilling to restrict to the selected subjects only, and it was meaningless make conclusion to fixed number of subjects only. Thus, we decided that the subjects (row factor) were indeed random. However, despite the criteria, the same authors recommended that the sample should be representative of target population as closely as possible. In our sample, the subjects were adults subjects with age ranging from 20 to 63 years old, with mean of 39.6 years of age. We could not verify whether our sample was a representative one.

As for the assessors (column factor), we also decided that the factor was random based on replaceability criterion as the assessors were replaceable with other assessors if available. However, we wanted to restrict to our selected number of assessors (levels) only and doing so was enough for this study, therefore

did not completely fulfil the criteria mentioned above. To explain the choice of random column factor, Shrout and Fleiss [12] explained in their landmark paper on ICC that when results from different assessors are pooled together for analysis (in substantive study), the assessors contribute to the variability, thus considered as random factor. In our case, this reliability study preceded our following clinical trial (substantive study) in which the results from different assessors would be combined and analysed, thus we decided for random column factor. Thus, row factor (subjects) and column factor (assessors) both contributed to the variability for ICC calculation, and considered as random effects.

In this reliability study, we carefully selected the most suitable type of ICC and checked for the assumptions of ICC. The sample size exceeded the pre-calculated sample size. Necessary steps as explained in the methodology were taken to reduce systematic errors. Though careful consideration was given to comply with JNC 7 [11] steps in BP measurement, the step to exclude smoking, caffeine or exercise was not properly ensured. However, we deemed the step unnecessary in this study because the objective was to measure BP of participants, not to establish whether they were hypertensive or not. Hence this particular step in JNC 7 [11] was skipped.

Nevertheless, there were several weaknesses of the study. The sampling method was of non-probabilistic sampling, which might result in our sample being unrepresentative of the target population. Further, some assessors read out loud the BP readings, which may affect other assessors' readings as they were aware of the other assessors' readings for particular participants. Ideally assessors should be in separate rooms to reduce assessor bias. Terminal digit preference among the assessors was also noted. The assessors reported the readings with "5" and "0" terminal digits. The recommended cuff deflation rate is 2mmHg per second [11], to allow a reading to closest 2mmHg [16]. However, the assessors noted that in practice, rounding to closest 5mmHg is clinically more practical. In a clinical trial, the observations should be as precise as possible when statistical analysis is concerned. The assessors should have been noted on this aspect. It was not briefed to the assessors as it was not stated explicitly in JNC 7 [11] steps.

The demonstration of the use of ICC 2A (A, 1) in this study to assess the reliability of manual BP measurement using mercury sphygmomanometer among assessors is of value to other reliability study of similar nature, particularly for reliability study preceding a clinical trial involving manual BP measurement. We emphasized on selection of suitable ICC type for statistical analysis and assumptions checking, apart from ensuring compliance to BP measurement protocol and working equipments. The weaknesses in this study can be used as check list points for a better design of reliability study on BP measurement.

Conclusion

Since the use of manual BP measurement using mercury sphygmomanometer is still pertinent given its status as a gold standard of BP measurement, studies in which the BP outcome is measured using such method have to ensure the interrater reliability of their assessors. Such assurance is provided with the use of ICC as an objective way of expressing the reliability.

Acknowledgement

We are grateful to the School of Medical Sciences, Universiti of Science Malaysia for approving and funding this study. We are also thankful to the subjects who volunteered to participate in the study, consisted of administrative staffs and lecturers from Biostatistics and Research Methodology Unit and Medical Education Department, for the high level of cooperation given throughout the study period, however short it was. The authors have no conflict of interest to declare.

Reference

1. Beevers DG, Lip GYH, O'Brien E. ABC of Hypertension. 5th ed. Singapore: Blackwell Publishing; 2007.
2. Jones DW, Appel LJ, Sheps SG, Roccella EJ, Lenfant C. Measuring blood pressure accurately. JAMA: The Journal of the American Medical Association. 2003;289(8):1027. http://dx.doi.org/10.1001/jama.289.8.1027
3. Fletcher RH, Fletcher SW, Wagner EH. Clinical epidemiology: the essentials. 3rd ed. Maryland: Williams & Wilkins; 1996.
4. Neufeld PD, Johnson DL. Observer error in blood pressure measurement. CMAJ: Canadian Medical Association Journal. 1986;135(6):633.
5. O'Brien ET, O'Malley K. ABC of blood pressure measurement. Reconciling the controversies: a comment on" the literature". British Medical Journal. 1979;2(6199):1201.
6. Gordis L. Epidemiology. 4th ed. Philadelphia: Saunders; 2009.
7. Trochim WMK. Research methods knowledge base [Internet]. 2006 [updated July 12, 2010]. Available from: http://www.socialresearchmethods.net.
8. Garson GD. Statnotes: Topics in multivariate analysis [Internet]. 2010 [cited 2010 July 12]. Available from: http://faculty.chass.ncsu.edu/garson/pa765/statnote.htm.
9. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychological Methods. 1996;1(1):30-46. http://dx.doi.org/10.1037/1082-989X.1.1.30
10. Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. Statistics in Medicine. 1998;17(1):101-10. http://dx.doi.org/10.1002/(SICI)1097-0258(19980115)17:1<101::AID-SIM727>3.3.CO;2-5
11. Complete report: Seventh report of the Joint National Committee on prevention, detection, evaluation, and treatment of high blood pressure: U.S. Department of Health and Human Services; 2004. Available from: http://www.nhlbi.nih.gov/guidelines/hypertension/jnc7full.pdf.
12. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin. 1979;86(2):420-8. http://dx.doi.org/10.1037/0033-2909.86.2.420
13. Kutner MH, Nachtsheim CJ, Neter J, Li W. Applied linear statistical models. 5th ed. Singapore: McGraw-Hill; 2005.
14. Altman DG. Practical statistics for medical research. London: Chapman and Hall; 1991.
15. Jackson S, Brashers DE. Random Factor in ANOVA. California: Sage Publications; 1994.
16. Williams B, Poulter N, Brown M, Davis M, McInnes G, Potter J, et al. Guidelines for management of hypertension: report of the fourth working party of the British Hypertension Society, 2004 - BHS IV.

Journal of Human Hypertension. 2004;18(3):139-85. http://dx.doi.org/10.1038/sj.jhh.1001683

Table 1: Blood pressure readings form for data collection

Subject ID	Blood Pressure (Systolic/Diastolic mmHg)									
	Assessor ID									
	1		2		3		4		5	
1										
2										
3										
4										
5										

Corresponding Author: Dr Wan Nor Arifin, Unit of Biostatistics and Research Methodology, School of Medical Sciences, Universiti Sains Malaysia, 16150 Kubang Kerian, Kelantan, Malaysia. **Email**: wnarifin@yahoo.com

Accepted: November 2011 Published: December 2011