Investigating the Reliability of Contemporary Chinese Pulse Diagnosis

Karen Bilton*[1] BAppSc
Sean Walsh[1] PhD
Narelle Smith[1] PhD
Leon Hammer[2] MD

1. University of Technology, Sydney, Australia
2. Dragon Rises College of Oriental Medicine, Gainesville, Florida, USA

ABSTRACT

There have been few studies that evaluate the reliability of the clinical use of pulse diagnosis despite it being a fundamental part of Oriental medicine diagnostics. The objective of this study was to determine the levels of intra-rater and inter-rater reliability of practitioners using an operationally defined method, Contemporary Chinese Pulse Diagnosis (CCPD), to evaluate the radial pulse of volunteer subjects. The study utilised a real-life design to investigate CCPD in a clinical setting. Fifteen volunteer subjects and six testers skilled in the CCPD method were recruited. Two episodes of data collection were conducted 28 days apart as a practical test and retest. For each subject, 30 pulse categories defined by the CCPD system were assessed and reassessed by the same four testers during both phases of testing. All assessments were conducted according to the CCPD method. Intra-rater reliability was measured by comparing individual tester results on day one with day two, while inter-rater agreement and reliability were determined by comparing all testers across both days. The data were analysed using Cohen's kappa coefficient. Kappa values were interpreted according to recommendations from previous clinical studies and parameters considered acceptable when using a tool such as CCPD to assist in clinical diagnosis. Results for intra-rater reliability showed excellent agreement in 43.2%, moderate to good agreement in 42.5% and poor agreement in 14.3% of the raw kappa calculations. Inter-rater agreement demonstrated excellent agreement in 23.5%, moderate to good agreement in 46% and poor agreement in 30.5% of the raw kappa calculations. In conclusion, when the system of pulse diagnosis is operationally defined, acceptable levels of reliability can be achieved. Disagreement was either intrinsic to the subject or indicative of ambiguity within the CCPD system. Accordingly, review of the terminology of the appropriate pulse categories and their clinical reliability is recommended.

KEYWORDS pulse diagnosis, intra-rater reliability, inter-rater reliability, Chinese medicine, acupuncture, Cohen's kappa coefficient, Contemporary Chinese Pulse Diagnosis.

Introduction

PULSE DIAGNOSIS AND CLINICAL PRACTICE

The practice of pulse diagnosis is obfuscated in the modern context due to a range of commonly accepted assumptions that have little basis in clinical fact,[1] the most notable being the 'correctness' of the historical pulse literature as a reliable means for the diagnostic interpretation of pulse findings within clinical practice.[2-5] The classics have been shown under a range of experimental conditions[6-10] to be inadequate for this task therefore discounting this long held supposition. Accordingly, Ramholz[5] describes the classics as 'the starting point for study and research, not the accumulation or final arbiter of what can be known.'

[*] Correspondent author; e-mail: karenbilton@mac.com

Aust J Acupunct Chin Med 2010;5(1):3–13.

PULSE DIAGNOSIS – THE RESEARCH

There have been too few studies undertaken to assess the reliability of pulse diagnosis, which is surprising, given there remains a questionable validity to the results obtained from any study on acupuncture or Chinese medicine using untested methods of assessment. Using pulse diagnosis as an example, without demonstration of the clinical dependability of the pulse-taking procedure, it is not possible to have confidence in any assertions concerning information gained from the pulse.[11]

Of the studies undertaken in English, Cole[7] was the first to report upon pulse diagnosis reliability. Studying four practitioners feeling the pulse of several subjects, Cole reported that low levels of agreement found between pulse assessors reflected the conflicting and confusing nature of information available in the literature. Krass[8] and Craddock[9] have also reported similar findings. Separately they each concluded that inter-rater reliability decreases with increasing levels of complexity of pulse qualities being measured. Together, the findings of these studies support the assertion that variation of definitions and pulse terms in the literature has limited the reliability of manual pulse measurement.[6,11,12] In response, King and Walsh[10,11] undertook a series of studies assessing the reliability of manual pulse diagnosis. Using a standardised evaluation procedure and concrete operational definitions[10,11] they demonstrated that high levels of inter-rater reliability can be achieved under these conditions.

CONTEMPORARY PRACTICE

Practitioners depend upon the literature to assist and guide their classification of a pulse as healthy or not, thus contributing to their process of diagnosis.[7-10] Given the current state of the literature and the subjective nature of pulse diagnosis overall, it is not surprising to find reports in the literature, anecdotal and otherwise, of practitioners' reduced confidence in pulse assessment to contribute meaningful information to diagnosis.[11,13] For this reason, there has been a resurgence in systems of pulse diagnosis based on traditional texts and theoretical knowledge that have been further developed to clarify the problems of ambiguity contained in the classics, while still remaining clinically relevant to current methods of practice.[1,5,14] One such system is *Contemporary Chinese Pulse Diagnosis (CCPD)*.

CONTEMPORARY CHINESE PULSE DIAGNOSIS

CCPD is a trademarked method introduced by Dr John HF Shen,[14] a prominent modern practitioner who trained with the Ding family physicians, one of three influential lineages in Menghe medicine.[15] It is believed the vast body of Chinese medical knowledge possessed by the traditional family lineages influenced his development of this pulse system. CCPD is ostensibly standardised or operationally defined in the text *Chinese Pulse Diagnosis, A Contemporary Approach*.[16] Although using different definitions than those fixed by King and colleagues,[11] feasibly, a rigorous standard of testing to measure agreement levels between practitioners using this system can be applied. In this context, high levels of reliability should be achievable if the definitions of pulse characteristics, and how these are assessed, are being replicated every time.

CCPD incorporates six *principal* positions, 22 *complementary* positions, 80 pulse qualities, and eight depths.[16] The radial arteries are palpated bilaterally with differing amounts of pressure to assess pulse qualities at three main depths (termed the qi, blood, and organ depth).[16] Pulse qualities are described by sensation and defined using fixed terms in an attempt to eliminate the metaphoric ambiguity of the classical literature. In addition, each pulse term is given a specific diagnostic interpretation. Individual pulse positions are described using standard anatomical terms and the procedure for evaluating and identifying pulse qualities is clearly explained.[16] As the methods, terminology and interpretation are allegedly definitive and consistent within the system,[16] theoretically it is possible to undertake testing of intra-rater and inter-rater reliability.

Within CCPD the pulse is thought to represent the state of the organs, substances, pathogens and metabolic activity, or, the health of the person.[16] If health remains unchanged then the founding principles of this method indicate the fundamental characteristics of the pulse readings should remain accordingly stable, thus allowing appropriate inquiry. Accordingly, a study was designed to investigate the reliability of practitioners assessing pulses with standard pulse definitions and procedures using the CCPD system.

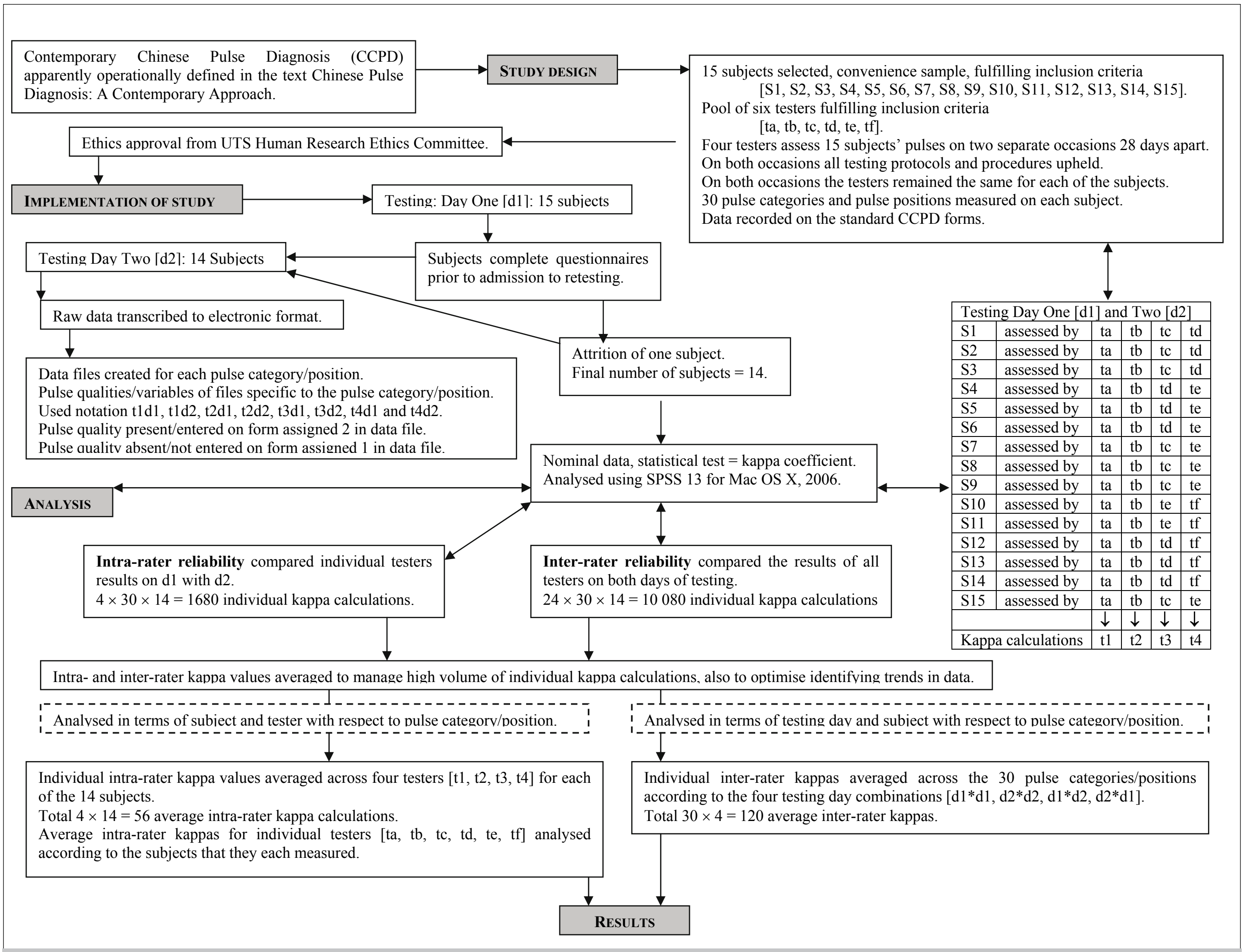
The Study: Materials and Methods

AIM AND OBJECTIVES

This study aimed to determine the reliability of practitioners using CCPD to assess the radial pulse of patients, including the examination of (1) intra-rater reliability by measuring agreement between single testers' assessments within the same subject made on different occasions, and (2) inter-rater reliability by measuring agreement between different testers assessing a single subject on the same as well as different occasions.

DESIGN

The study incorporated a real-life design, where testing conditions reflected clinical practice, utilising the same procedure, positioning and documentation of findings (Figure 1). Two separate episodes of data collection were conducted as a practical test and retest 28 days apart (to replicate female subjects' menstrual cycles) at the same time of day (to control for diurnal pulse variance). The recorded data were



Contemporary Chinese Pulse Diagnosis (CCPD) apparently operationally defined in the text Chinese Pulse Diagnosis: A Contemporary Approach.

STUDY DESIGN

15 subjects selected, convenience sample, fulfilling inclusion criteria [S1, S2, S3, S4, S5, S6, S7, S8, S9, S10, S11, S12, S13, S14, S15].
Pool of six testers fulfilling inclusion criteria [ta, tb, tc, td, te, tf].
Four testers assess 15 subjects' pulses on two separate occasions 28 days apart.
On both occasions all testing protocols and procedures upheld.
On both occasions the testers remained the same for each of the subjects.
30 pulse categories and pulse positions measured on each subject.
Data recorded on the standard CCPD forms.

Ethics approval from UTS Human Research Ethics Committee.

IMPLEMENTATION OF STUDY

Testing: Day One [d1]: 15 subjects

Subjects complete questionnaires prior to admission to retesting.

Testing Day Two [d2]: 14 Subjects

Raw data transcribed to electronic format.

Attrition of one subject.
Final number of subjects = 14.

Data files created for each pulse category/position.
Pulse qualities/variables of files specific to the pulse category/position.
Used notation t1d1, t1d2, t2d1, t2d2, t3d1, t3d2, t4d1 and t4d2.
Pulse quality present/entered on form assigned 2 in data file.
Pulse quality absent/not entered on form assigned 1 in data file.

Testing Day One [d1] and Two [d2]					
S1	assessed by	ta	tb	tc	td
S2	assessed by	ta	tb	tc	td
S3	assessed by	ta	tb	tc	td
S4	assessed by	ta	tb	td	te
S5	assessed by	ta	tb	td	te
S6	assessed by	ta	tb	td	te
S7	assessed by	ta	tb	tc	te
S8	assessed by	ta	tb	tc	te
S9	assessed by	ta	tb	tc	te
S10	assessed by	ta	tb	te	tf
S11	assessed by	ta	tb	te	tf
S12	assessed by	ta	tb	td	tf
S13	assessed by	ta	tb	td	tf
S14	assessed by	ta	tb	td	tf
S15	assessed by	ta	tb	tc	te
		↓	↓	↓	↓
Kappa calculations		t1	t2	t3	t4

ANALYSIS

Nominal data, statistical test = kappa coefficient.
Analysed using SPSS 13 for Mac OS X, 2006.

Intra-rater reliability compared individual testers results on d1 with d2.
4 × 30 × 14 = 1680 individual kappa calculations.

Inter-rater reliability compared the results of all testers on both days of testing.
24 × 30 × 14 = 10 080 individual kappa calculations

Intra- and inter-rater kappa values averaged to manage high volume of individual kappa calculations, also to optimise identifying trends in data.

Analysed in terms of subject and tester with respect to pulse category/position.

Analysed in terms of testing day and subject with respect to pulse category/position.

Individual intra-rater kappa values averaged across four testers [t1, t2, t3, t4] for each of the 14 subjects.
Total 4 × 14 = 56 average intra-rater kappa calculations.
Average intra-rater kappas for individual testers [ta, tb, tc, td, te, tf] analysed according to the subjects that they each measured.

Individual inter-rater kappas averaged across the 30 pulse categories/positions according to the four testing day combinations [d1*d1, d2*d2, d1*d2, d2*d1].
Total 30 × 4 = 120 average inter-rater kappas.

RESULTS

FIGURE 1 Study design

compared to determine the levels of agreement and reliability of testers using this method. All testing was conducted at Dragon Rises College of Oriental Medicine (DRCOM) in Gainesville, Florida, USA.

PARTICIPANTS

Fifteen participants were recruited as a volunteer convenience sample from the Gainesville region. Eight of the 15 were students at DRCOM, the remaining were recruited from the local community by advertising. Inclusion criteria specified that the subjects were at least 18 years of age, had not previously completed a pulse assessment with the testers and, if receiving any form of medical intervention, it remain consistent for the duration of the testing period. Exclusion criteria included signs or symptoms of short-term acute illness (e.g., colds, respiratory and/or intestinal flu), febrile conditions, radical lifestyle changes and modification to medical management or daily supplements. These parameters aimed to reduce influences that would alter the 'normal' pulse for the subjects and unduly affect the testing procedure.

Prior to the second stage of testing, all subjects completed a questionnaire detailing their health status at the time. No exclusion criteria were reported so all were admitted to re-testing. Of the 15, one failed to return for re-assessment, reducing the final number of participants to 14.

ETHICAL CONSIDERATIONS

Prior to data collection, ethics approval was obtained from the Human Research Ethics Committee of the University of Technology, Sydney, Australia.

PULSE EXAMINERS

The study incorporated a pool of six testers. Inclusion criteria were that all had detailed knowledge of CCPD, had been using it clinically for more than seven years and were actively involved in maintaining familiarity with the application of the system. To fulfil these requirements, testers were recruited from Dragon Rises Seminars (DRS) instructors.

One of the testers had documented the CCPD system while the remaining had received equivalent training from that tester. Their clinical experience using the current standard CCPD definitions and procedures ranged from seven to 15 years and all had been DRS instructors for more than three years. In addition to match tester skills, each attended six-monthly workshops to ensure consistency in palpation. Four of the six testers were lecturers at DRCOM.

METHOD

Throughout testing, four testers evaluated three subjects daily for five consecutive days. Working in pairs, testers were designated one hour to assess each subject. Testers were allowed ten minutes individual bilateral wrist palpation to assess pulse categories that constitute the *large segments* of the pulse, followed by 20 minutes for each side to assess the *small segments*, or individual pulse positions. In total four testers required two hours to evaluate each of the participants who were allowed a five-minute break between each pair of testers.

During test and retest, the combination of four testers remained the same for each subject (Figure 1). Testers employed standard CCPD operational evaluation procedures and all data were recorded on the standardised pulse forms. During testing, talking was not allowed and the testers were blinded to each other's findings. The subjects were allowed to drink water if requested.

As comparing the combination of pulse qualities entered for each category (reliability) was the exclusive interest of the study, and not diagnostic interpretation of these qualities (validity), the subjects remained in view of the testers. Immediately after the first round of testing all pulse forms were collected and secured independently to prevent untoward comparisons of findings.

DATA MANAGEMENT

Thirty pulse positions or pulse categories were measured and recorded for each subject. Table 1 lists these along with the number of possible pulse qualities or variables for each. Table 2 presents all variables grouped according to their pulse classification.

The testers' responses were transcribed to electronic format. For each pulse position/category, a master file was constructed that defined the range of possible qualities. The results of all testers on both days were entered into these electronic data files. If a pulse quality was present, that is, entered on the pulse form by the tester, number 2 was assigned. If a pulse quality was absent, that is, not entered on the pulse form, number 1 was assigned. For each file, the data were organised according to tester and day of testing using the notation t1d1, t1d2, t2d1, t2d2, t3d1, t3d2, t4d1 and t4d2 where t=tester, and d=day. In total, 30 separate files were created for each of the 14 subjects, an example of which is presented in Table 3.

DATA ANALYSIS

The data were analysed using Cohen's kappa coefficient[17] and SPSS 13 for Mac OS X, 2006. Kappa, the preferred measure of

TABLE 1 List of pulse categories or positions and associated number of variables

Pulse category or pulse position	Number of variables
Large segment of pulse (categories assessed by simultaneous bilateral wrist palpation)	
1. First impressions	41
2. First impressions (left side)	51
3. First impressions (right side)	51
4. Rhythm	4
5. Above qi depth	6
6. Qi depth	29
7. Blood depth	34
8. Organ – qi depth	31
9. Organ – blood depth	31
10. Organ – organ depth	31
11. Waveform	6
Small segment of pulse (positions assessed by unilateral wrist palpation)	
Principal positions (found on the main artery)	
12. Left distal position	39
13. Right distal position	39
14. Left middle position	43
15. Right middle position	43
16. Left proximal position	43
17. Right proximal position	43
Complementary positions (related to a principle position)	
18. Left neuropsychological position	28
19. Right neuropsychological position	28
20. Mitral valve	22
21. Left special lung position	37
22. Right special lung position	37
23. Diaphragm position	13
24. Gall bladder position	32
25. Stomach pylorus extension position	32
26. Large intestine position	33
27. Small intestine position	33
28. Left pelvic lower body position	32
29. Right pelvic lower body position	32
30. Combined complementary position	11

rater reliability for nominal data,[18] measures the reliability of agreement between two or more independent raters[17] using a rating scheme with mutually exclusive categories.[17-19] Kappa is an extension of simple percent agreement[28,29] and corrects this for the proportion of agreement that would be expected due to chance alone.[18-23] Kappa values lie between -1.00 and 1.00. Those approaching 1.00 represent perfect agreement, 0.00 represents agreement due to chance alone[18] and negative values indicate agreement less than what is expected by chance.[24,25] Definitive kappa interpretations have been proposed.[20,22,26-29] However, for most purposes values ≤0.40 represent poor agreement, values between 0.40 and 0.75 represent moderate to good agreement and values ≥0.75 indicate excellent agreement.[29] Values <0.00 are a rare outcome as rater training usually results in a kappa value >0.00.[24,25]

TABLE 2 List of variables or pulse qualities included in the study

Pulse quality grouping	Variable – pulse quality	Variable number
Qi wild	Empty	1
	Change in quality	2
	Change in intensity	3
	Unstable	4
	Scattered	5
	Minute	6
	Leather	7
	Intensity change side to side	8
	Qualities shifting side to side	9
Volume (robust)	Hollow full-overflowing	10
	Robust pounding	11
	Flooding excess	12
	Inflated	13
Volume (reduced)	Yielding qi depth	14
	Diminished qi depth	15
	Feeble at qi depth	16
	Spreading	17
	Reduced substance	18
	Reduced pounding	19
	Diffuse	20
	Deep	21
	Feeble – absent	22
	Flat	23
	Suppressed pounding	24
	Muffled	25
	Dead	26
Depth	Floating tight	27
	Floating tense	28
	Floating yielding	29
	Floating smooth vibration	30
	Floating slippery	31
	Cotton	32
	Hollow	33
Width (narrow)	Thin	34
Length	Short	35
	Restricted	36
	Long	37
Shape (fluid)	Slippery	38
Shape (non-fluid–hard even)	Taut	39
	Tense [tense-tight]	40
	Tight [tight-tense]	41
	Wiry	42
	Ropy	43
Shape (non-fluid–hard uneven)	Choppy	44
	Smooth vibration	45
	Rough vibration	46

TABLE 2 continued List of variables or pulse qualities included in the study

Pulse quality grouping	Variable – pulse quality	Variable number
Shape (miscellaneous)	Biting	47
	Doughy	48
	Amorphous	49
	Hard-leather	50
	Electrical	51
	Bean 'spinning'	52
	Split vessel	53
Modifiers	Transient	54
	Separating	55
	Rough	56
Anomalous	Fan Quan Mai/ San Yin Mai	57
	Ganglion	58
	Local trauma	59
Wave	Normal wave	60
	Flooding deficient	61
	Hesitant	62
	Suppressed	63
	{Hollow full-overflowing}	
	{Flooding excess}	
Rhythm	Change in rate at rest	64
	Intermittent	65
	Interrupted	66
	Normal rhythm	67
Width (wide)	Blood unclear	68
	Blood heat	69
	Blood thick	70
Sides (amplitude–intensity)	Sides equal	71
	Left > right	72
	Right > left	73
Diaphragm	Inflation equal bilateral	74
	Inflation left > right	75
	Inflation right > left	76
Variables specific to the combined complimentary positions		
Associated principle position	Variable – complementary position	Variable number
Heart	Pericardium	77
	Large vessel	78
	Heart enlarged	79
Lung	Pleura	80
Liver	Distal liver engorged	81
	Radial liver engorged	82
	Ulna liver engorged	83
Stomach – Spleen	Oesophagus	84
	Spleen special	85
	Pancreas – peritoneal cavity	86
	Duodenum	87

TABLE 3 First Impressions: Subject 1 (2 = present, 1 = absent)

Possible Qualities First Impressions		t1 d1	t1 d2	t2 d1	t2 d2	t3 d1	t3 d2	t4 d1	t4 d2
Qi Wild									
Empty	1	1	1	1	1	1	1	1	1
Change in quality	2	1	1	1	1	1	1	1	1
Change in intensity	3	2	2	2	2	1	2	2	2
Scattered	5	1	1	1	1	1	1	1	1
Minute	6	1	1	1	1	1	1	1	1
Leather	7	1	1	1	1	1	1	1	1
Volume – Robust									
Hollow full-overflowing	10	1	1	1	1	1	1	1	1
Robust pounding	11	2	2	1	2	2	2	2	2
Flooding excess	12	1	1	1	1	1	1	1	1
Volume – Reduced									
Yielding qi depth	14	1	1	1	1	1	1	1	1
Diminished qi depth	15	1	1	1	1	1	1	1	1
Feeble at qi depth	16	1	1	1	1	1	1	1	1
Spreading	17	1	1	1	1	1	1	1	1
Reduced substance	18	2	2	2	2	1	2	1	2
Reduced pounding	19	1	1	1	1	1	1	1	1
Diffuse	20	2	2	1	1	1	1	1	1
Deep	21	1	1	1	1	1	1	1	1
Feeble – absent	22	1	1	1	1	1	1	1	1
Muffled	25	2	1	2	1	2	2	1	1
Depth									
Hollow	33	1	1	1	1	1	1	1	1
Width									
Thin	34	2	2	2	2	2	2	2	2
Length									
Short	35	1	1	1	1	1	1	1	1
Long	37	1	1	1	1	1	1	1	1
Shape – Fluid									
Slippery	38	1	1	1	1	1	1	1	1
Shape – Non Fluid Even									
Taut	39	1	1	1	1	1	1	1	1
Tense [tense-tight]	40	1	1	1	1	1	1	1	1
Tight [tight-tense]	41	2	2	2	2	2	1	2	2
Wiry	42	1	1	1	1	1	1	1	1
Ropy	43	1	1	1	1	1	1	1	1
Shape – Non Fluid Uneven									
Choppy	44	2	1	2	2	1	2	2	2
Smooth vibration	45	1	2	1	1	2	2	2	2
Rough vibration	46	1	1	1	1	1	1	1	1
Shape – Miscellaneous									
Amorphous	49	1	1	1	1	1	1	1	1
Hard-leather	50	1	1	1	1	1	1	1	1
Split vessel	53	1	1	1	1	1	1	1	1
Modifiers									
Transient	54	2	2	1	2	2	2	1	2
Separating	55	1	1	1	1	1	1	1	1
Rough	56	1	1	1	1	1	1	1	1
Anomalies									
Fan Quan Mai/ San Yin Mai	57	2	2	2	2	2	2	2	2
Ganglion	58	1	1	1	1	1	1	1	1
Local trauma	59	1	1	1	1	1	1	1	1

Intra-rater reliability or the consistency within a tester over time[19] to assess the pulse was measured by comparing the four testers' results on day one, with their results on day two, for example, t1d1*t1d2, where t=tester and d=day. This method resulted in four kappa values for each of the 30 pulse positions or pulse categories in each of the 14 subjects, totalling 4 × 30 × 14 = 1680 individual kappa calculations. Each kappa value measured reliability in terms of pulse quality matches for that pulse position.

These data were then analysed in terms of tester and subject with respect to pulse position or category. To help identify trends and manage the numerous individual kappa calculations, values were averaged for four testers across the 14 subjects, resulting in 56 (4 × 14 subjects) average intra-rater kappa calculations. Average intra-rater kappas for individual testers were also analysed according to the subjects that they measured.

Inter-rater reliability, more accurately divided into inter-rater agreement (comparing testers within one day) and reliability (comparing testers over time or between days)[19] was determined by comparing the results of two testers at a time across both days of testing. This produced six tester combinations (t1*t2, t1*t3, t1*t4, t2*t3, t2*t4 and t3*t4) and four day combinations (d1*d1, d1*d2, d2*d2 and d2*d1). This resulted in 24 kappa values for each pulse position in all subjects totalling 24 × 30 × 14 = 10 080 individual kappa calculations.

These data were analysed in terms of testing day and subject with respect to pulse position or category. For ease of handling and reporting the vast number of calculations, individual inter-rater kappa values were averaged for each of the 30 pulse categories according to testing day combination, totalling 120 average inter-rater kappas (30 × four day combinations).

Results

INTRA-RATER RELIABILITY

Of the 1680 raw intra-rater kappa scores, 43.2% (726) showed kappas ≥0.75 or excellent agreement[29]; a further 42.5% (713) scored kappas 0.41–0.74, indicating moderate to good agreement[29]; while 14.3% (241) scored kappas ≤0.40, showing poor agreement.[29] In total 67% (1126) scored ≥0.60.

The averaged intra-rater kappas for individual testers and the subjects they assessed are shown in Table 4. Four of the testers (ta, tb, tc, td) attained excellent agreement between their repeated assessments from day one to day two in at least one of the subjects they tested, and two of these testers (tb, td) obtained average intra-rater kappa values ≥0.60 for all subjects tested. One tester (tf) scored average intra-rater kappas <0.60 in all subjects tested.

The 56 averaged intra-rater kappas are presented in terms of the 14 subjects in Table 5. Excellent agreement was demonstrated in the upper limit of the kappa ranges of nine subjects. Six of the subjects (1, 3, 5, 6, 9 and 15) demonstrated average intra-rater kappa values ≥0.60 for all four testers. Intra-rater disagreement was unevenly distributed with the two lowest average intra-rater kappa values of 0.44 and 0.45 appearing within subject 13, and 0.48 in subject 10. Examination of the individual kappas for these testers and subjects showed the unusual occurrence[24,25] of negative values, that is, agreement less than that expected by chance, in up to seven pulse categories.

INTER-RATER AGREEMENT AND RELIABILITY

Of the 10 080 raw inter-rater kappa scores, 23.5% (2366) showed kappas ≥0.75 or excellent agreement[29]; 46% (4642)

TABLE 4 Intra-rater reliability: Average kappa scores and kappa ranges according to tester

Tester	No. of subjects tested	Average kappa	Range of kappa
t a	14	0.66	0.44 – 0.76
t b	14	0.72	0.65 – 0.82
t c	7	0.68	0.54 – 0.78
t d	9	0.70	0.62 – 0.78
t e	8	0.62	0.49 – 0.72
t f	4	0.52	0.45 – 0.57

TABLE 5 Intra-rater reliability: Average intra-rater kappa ranges according to subject

Subject	Range of intra-rater kappas
Subject 1	0.66 – 0.76
Subject 2	0.54 – 0.76
Subject 3	0.61 – 0.78
Subject 4	0.59 – 0.77
Subject 5	0.62 – 0.75
Subject 6	0.64 – 0.77
Subject 7	0.57 – 0.74
Subject 8	0.49 – 0.78
Subject 9	0.68 – 0.74
Subject 10	0.48 – 0.70
Subject 12	0.57 – 0.70
Subject 13	0.44 – 0.78
Subject 14	0.57 – 0.82
Subject 15	0.62 – 0.73

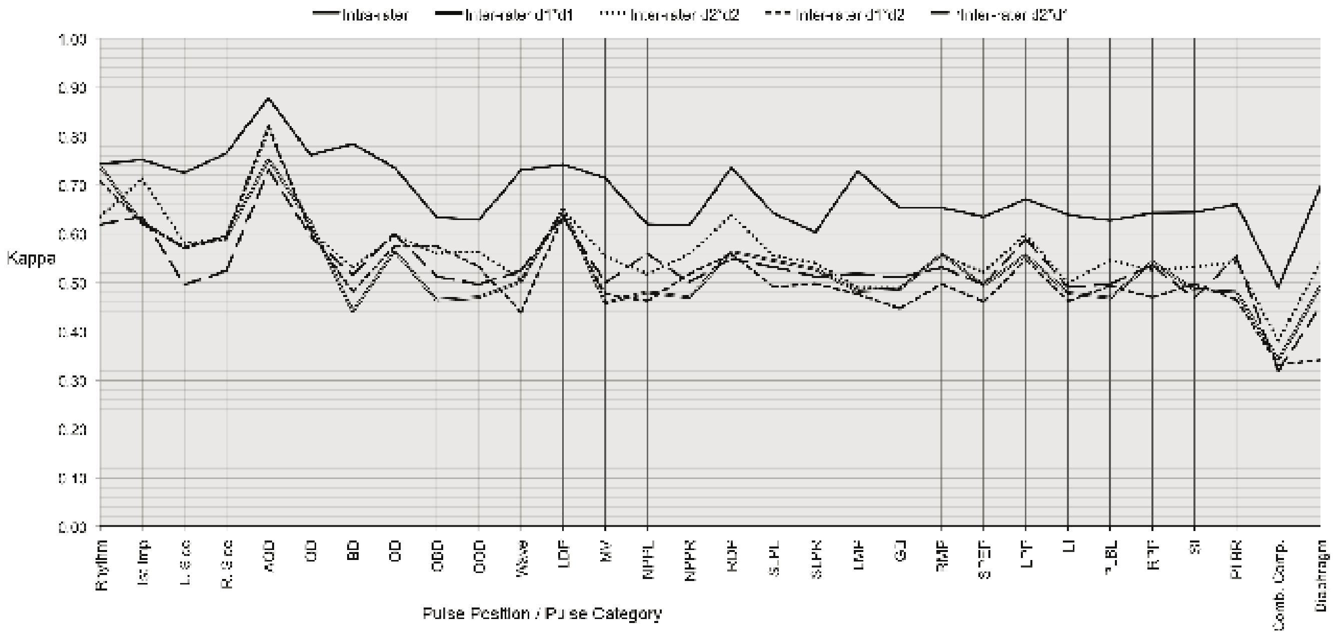
scored kappas 0.41–0.74, indicating moderate to good agreement[29]; while 30.5% (3072) scored kappas ≤0.40, showing poor agreement.[29] In total, 44.1% (4442) scored ≥0.60.

The kappa values averaged by subject for inter-rater agreement and reliability are presented in Table 6. The different testing day combinations (bottom row of Table 6) showed values between 0.52–0.56, indicating moderate to good agreement.[29] In terms of the 14 subjects, the average kappas for all testing day combinations (the last column of Table 6) ranged from 0.42–0.63, indicating moderate to good agreement.[29] The two lowest scores of 0.47 and 0.42 occurred in subject 13 and subject 14, reported previously as demonstrating a higher incidence of intra-rater disagreement. The averaged 24 inter-rater kappa scores for individual pulse categories for these two subjects showed poor agreement (kappas ≤0.40) in seven categories common to both. Negative individual inter-rater kappa values were again noted; however, they were not skewed according to pairs of testers.

Finally, Figure 2 presents the results for intra-rater and inter-rater reliability in terms of pulse position/category. The y-axis indicates the gross levels of agreement or kappa values averaged across all testers and subjects, while the x-axis indicates pulse position/category. The intra-rater results demonstrated moderate to good agreement or above in all but one category, the Combined Complementary Positions that scored <0.50. The averaged inter-rater kappas were lesser and followed a similar pattern; however, in this instance, the Combined Complementary Positions was significantly lower and rated poor agreement for all d*d combinations.

Discussion

The study employed the kappa coefficient in such a way that the calculations determined the level of agreement in terms of pulse quality matches for one pulse position in one subject. They were used as a descriptive measure, akin but superior to reporting percentage agreement, to identify pulse positions that seem to have reliable assessments, and those that appear to exhibit lower levels of reliability. The study did not use kappa as an inferential measure, generalise values to other pulse positions or subjects, test a hypothesis or explore sampling variation. Accordingly standard errors were not reported with the results.



Rhythm = Rhythm
1st Imp = First impressions
L. Side = Left side
R. Side = Right side
AQD = Above qi depth
QD = Qi depth
BD = Blood depth
OD = Organ depth
OBD = Organ blood depth
OOD = Organ organ depth

Wave = Waveform
LDP = Left distal position
MV = Mitral valve position
NPPL = Neuropsychological position left
NPPR = Neuropsychological position right
RDP = Right distal position
SLPL = Special lung position left
SLPR = Special lung position right
LMP = Middle position left
GB = Gall Bladder position

RMP = Right middle position
SPEP = Stomach pylorus extension
LPP = Left proximal position
LI = Large intestine position
PLBL = Pelvic lower body left
RPP = Right proximal position
SI = Small intestine
PLBR = Pelvic lower body right
Comb. Comp. = Combined complementary positions
Diaphragm = Diaphragm

FIGURE 2 Average intra-rater and inter-rater Kappa values

The statistical literature[19] discusses concerns regarding the misinterpretation of kappa in studies reporting rater reliability.[30-34] Intra-rater reliability addresses the extent to which raters produce essentially the same score[25,30,35] and represents an index of proportional consistency across raters or over time.[19] Here lies the first caveat encountered with the application of kappa to the data. With test–retest measuring of intra-rater reliability there will always be some degree of correlation or dependence between the responses from testers over time.[36,37] An attempt to validate the study design was made by reducing dependence[36] or allowing 28 days between ratings to reduce rater memory of the previous test[36] and optimise the parameters of stability[27] defined by the female subjects' pulses.

Another caution that proved relevant to the study relates to the reporting of mean or average kappa values and kappa ranges.[19] Although the literature expresses some concern that these may disguise variability that might be important to the research,[19] the raw or individual kappa totals stated justify the method of reporting kappa. Kappa values were used as strategic markers to direct the discussion as well as to report agreement.

Values presented by Jelles et al.[29] were used in the study as a guide to interpret the kappa scores. However, in determining the strength of kappa there are no fixed absolutes.[38] Statistics cannot replace clinical judgements,[37] so interpretation depends on the circumstances and variables being tested.[38] The current pulse diagnosis literature has no recommendations regarding kappa; however, several authors suggest values <0.40 may be unacceptable in clinical situations.[36,38]

TABLE 6 Inter-rater reliability: Average inter-rater kappa values according to subject and testing day

Subject	d1*d1	d2*d2	d1*d2	d2*d1	Average K for subject
Subject 1	0.62	0.66	0.63	0.60	0.63
Subject 2	0.46	0.53	0.47	0.50	0.49
Subject 3	0.54	0.54	0.58	0.47	0.53
Subject 4	0.56	0.58	0.58	0.57	0.57
Subject 5	0.51	0.56	0.53	0.53	0.53
Subject 6	0.59	0.62	0.51	0.64	0.59
Subject 7	0.55	0.54	0.57	0.50	0.54
Subject 8	0.51	0.57	0.52	0.52	0.53
Subject 9	0.59	0.58	0.59	0.56	0.59
Subject 10	0.47	0.49	0.53	0.47	0.49
Subject 12	0.51	0.50	0.45	0.51	0.49
Subject 13	0.44	0.58	0.40	0.43	0.47
Subject 14	0.38	0.46	0.40	0.44	0.42
Subject 15	0.56	0.56	0.52	0.58	0.56
Average K for d*d	0.52	0.56	0.52	0.53	0.55

*Where d1*d1 compares the results of day 1 with day 1; d2*d2 day 2 with day 2; d1*d2 day 1 with day 2 and d2*d1 day 2 with day 1.*

From a clinical viewpoint, the results of pulse assessments are an integral part of formulating diagnoses and plans of treatment. Accordingly, the kappa values that judge the reliability of the tool should be stringent. The rating of moderate to good agreement (kappa 0.41–0.74) recommended by Jelles et al.[29] is too large for the purposes of patient management and further differentiation is warranted. Therefore, the findings of this study suggest that kappa values ≤0.40 (poor or unacceptable agreement), 0.41–0.59 (moderate agreement), 0.60–0.74 (good agreement) and kappa ≥0.75 (excellent agreement), are appropriate indicators with respect to pulse diagnosis.

Based on these values, intra-rater reliability proved to be good to excellent (≥0.60) in 67% of the raw kappa calculations. When kappas were <0.60, both tester and subject were implicated. Lower levels of agreement evident in one tester may have resulted from varied experience affecting how decisions are made in the face of uncertainty.[39] Another consideration was that two of the four subjects assessed by that tester showed the lowest averaged intra-rater kappas.

Levels of inter-rater agreement were less than that of intra-rater reliability: good to excellent (≥0.60) in 44.1%, moderate (0.41–0.59) in 25.4% and poor or unacceptable (≤0.40) in 30.5% of the raw calculations. Variability according to the combination of testers or days was negligible, yet the distribution of inter-rater disagreement was skewed in terms of pulse category/position, as well as subject, suggesting a portion of the variability was intrinsic to the participants' pulses.

Variability in both intra-rater and inter-rater reliability was found to be concentrated in three subjects, all exhibiting uncommon[24,25] negative individual kappa values. Investigation of the statistical literature indicates that kappa scores <0.00 may reflect disagreement, but also paradoxes termed prevalence and bias.[36,40] Prevalence occurs when the incidence of a variable is very high or low, increasing the agreement expected by chance and decreasing the magnitude of kappa.[41] Prevalence proved to be an important factor influencing the data in many of the pulse categories. Despite this, disagreement remained skewed to three subjects, each of which were recorded by all testers as exhibiting a Fan Quan or San Yin pulse quality.

Pulses that are in a constant state of flux are encountered in the clinic, and may represent either the Fan Quan or San Yin pulse quality, or the Qi Wild condition.[16] In the first instance, blood shunted between an anomalous divergent vessel and the true radial artery results in varying sensations where the pulse is palpated. It affects all pulse positions, and depending on the degree of irregularity, may render the pulse exam invalid.[16] In the absence of anomaly, pulses with no fixed characteristic other than the change itself are thought to represent a situation of extreme deficiency and vulnerability to disease (the Qi Wild condition[16]). In both situations, lower levels of intra-rater and inter-rater agreement could be expected due to the constant fluctuation in pulse qualities.

Allowing for prevalence, the Combined Complementary Positions still exhibited lower levels of agreement for both intra-rater and inter-rater reliability (Figure 2). This category incorporated eleven complementary positions that are indicated on the CCPD pulse form as being present or absent and do not have specific pulse qualities recorded for them. These complementary positions are found in relation to a principal or main pulse position and represent yang organs or areas of the body. The sensations felt at these positions are often transient and difficult to access,[16] which may influence agreement; however, repeated kappa values ≤0.40, that is, poor or unacceptable agreement, indicate other factors may be implicated.

Variance within the technique of the testers is suggested by lesser intra-rater reliability and unacceptable inter-rater agreement (kappas ≤0.40) for these positions. Continued ambiguity existing in the CCPD terminology or instructions for accessing these positions may be implicated, emphasising the importance of operational definitions and theoretical frameworks in determining the reliability of clinical practice.

Conclusion

This paper reports the levels of intra-rater and inter-rater agreement of skilled practitioners employing CCPD to evaluate the pulse bilaterally at the radial artery. The results support the findings reported by King et al.,[11] suggesting that, when the system of pulse diagnosis is operationally defined and all that use the system understand the methods and interpret the pulse terminology in the same way, acceptable levels of reliability can be achieved.

Disagreement was found to be dependent on the skill of the tester, the stability of the subject's pulse, and the specific pulse position being assessed. In terms of pulse positions, further investigation is required to identify the source of variation demonstrated by unacceptable levels of agreement within the Combined Complementary Positions. This includes a detailed review of the terminology and instructions for techniques of palpation used to access the component positions of this category. If variation continues, then the reliability of this category within a clinical context needs to be re-evaluated entirely.

In addition, further analysis of the data needs to investigate and compare the reliability of the principal and complementary positions and the large segments (pulse categories assessed by simultaneous bilateral wrist palpation) and small segments (pulse positions assessed by unilateral wrist palpation) of the pulse. Individual pulse qualities should also be examined to isolate which qualities, if any, show unacceptable levels of agreement. This will bring to light further areas within CCPD that need to be revised.

It is essential to answer the questions regarding the diagnostic relevance of pulse diagnosis raised both within and external to the profession. To preserve the integrity of Oriental medicine these must be addressed with accepted analytical methods. Ethical and legal obligations of the profession call for responses that demonstrate the claims regarding its clinical efficacy. Future areas of investigation should include the capacity of the radial pulse to accurately indicate the state of health of the individual. This would address the claim in Oriental medicine that there are specific and predictable changes in pulse qualities that occur in the presence of dysfunction and disease, and that these changes are detectable via manual palpation of the pulse.

Clinical Commentary

Pulse diagnosis plays an important role in OM diagnosis. Case studies encountered in texts, the clinic or otherwise usually proffer a pulse presentation described according to the traditional literature. These pulse 'pictures' are widely accepted within the profession as credible patient commentaries despite there being little proof that it is either reliable, or a valid method for accurately predicting disease states. As such, critical evaluation must persist to provide evidence that either justifies or invalidates its continued use in diagnosis. In answer this study substantiates the reliability of skilled practitioners using an operationally defined system of pulse diagnosis.

References

1. Walsh S, King E. Pulse diagnosis: a clinical guide. Edinburgh: Churchill Livingstone; 2008.
2. Wang SH, Yang S (translator). The pulse classic: a translation of the Mai Jing. Boulder, CO: Blue Poppy Press; 1997.
3. Veith I (translator). The yellow emperor's classic of medicine, New ed. Berkeley: University of California Press; 1972.
4. Kuriyama S. The expressiveness of the body and the divergence of Greek and Chinese medicine, New York: Zone Books, 1999.
5. Ramholz J. An introduction to advanced pulse diagnosis: theory and clinical practice in light of the Nan Jing, Li Shi-zhen and Mai Jing. Pac J Orient Med 2001(June):37–41.

6. Walsh S, Cobbin D, Bateman K, Zaslawski C. Feeling the pulse: trial to assess agreement level among TCM students when identifying basic pulse characteristics. Eur J Orient Med 2001;3(5):25–31.
7. Cole P. Pulse diagnosis and the practice of acupuncture in Britain [PhD thesis]. Sussex: University of Sussex; 1977.
8. Krass R. Traditional Chinese medicine and pulse diagnosis in San Francisco health planning: implications for a Pacific Rim city [PhD thesis]. Berkeley: Social Welfare Department, University of California; 1990.
9. Craddock DS. Is traditional Chinese medical pulse reading a consistent practice? A comparative pilot study of four practitioners [Undergraduate research project]. Sydney: University of Technology, Sydney; 1997.
10. King E. Do the radial qualities of traditional Chinese medicine provide a reliable diagnostic tool? An examination of pulse relationships stated in modern and classical Chinese texts [MSc thesis]. Sydney: University of Technology, Sydney; 2001.
11. King E, Cobbin D, Walsh S, Ryan D. The reliable measurement of radial pulse characteristics. Acupunct Med 2002;20(4):150–9.
12. Walsh S. The radial arterial pulse: correlation of traditional Chinese medicine pulse characteristics with objective tonometric measures [PhD thesis]. Sydney: University of Technology, Sydney; 2003.
13. Bilton K. Personal observation. Dragon Rises Seminars; 2006.
14. Shen JHF. Chinese medicine. New York: Educational Solutions; 1980.
15. Scheid V. Currents of tradition in Chinese medicine, 1626–2006. Seattle: Eastland Press; 2007.
16. Hammer L. Chinese pulse diagnosis: a contemporary approach. Seattle: Eastland Press; 2005.
17. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20:37–46.
18. Cyr L, Francis K. Measures of clinical agreement for nominal and categorical data: the kappa coefficient. Comput Biol Med 1992:22(4):239–46.
19. Tooth LR, Ottenbacher KJ. The kappa statistic in rehabilitation research: an examination. Arch Phys Med Rehabil 2004:(85):1371–6.
20. Fleiss JC. Statistical methods for rates and proportions. 2nd ed. New York: Wiley; 1981.
21. Maclure M, Willett WC. Misinterpretation and misuse of the kappa statistic. Am J Epidemiol 1987;126:161–9.
22. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.
23. Uebersax JS. A generalized kappa coefficient. Educ Psychol Meas 1982;42(1):181–3.
24. Rae G. The equivalence of multiple rater kappa statistics and intraclass correlation coefficients. Educ Psychol Meas 1988;48:367–74.
25. Portney LG, Watkins MP. Foundations of clinical research: applications to practice. 2nd ed. Princeton: Prentice-Hall; 2000.
26. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess 1994;6:284–90.
27. Altman DG. Practical statistics for medical research. London: Chapman and Hall; 1991.
28. Shrout PE. Measurement reliability and agreement in psychiatry. Stat Methods Med Res 1998;7:301–17.
29. Jelles F, Van Bennekom CA, Lankhorst GJ, Sibbel CJ, Bouter LM. Inter- and intra-rater agreement of the rehabilitation activities profile. J Clin Epidemiol 1995;48(3):407–16.
30. Berk RA. Generalizability of behavioral observations: a clarification of inter-observer agreement and inter-observer reliability. Am J Ment Def 1979;83:460–72.
31. Suen HK. Agreement, reliability, accuracy, and validity: toward a clarification. Behav Assess 1988;10:343–66.
32. Maclure M, Willett WC. Misinterpretation and misuse of the kappa statistic. Am J Epidemiol 1987;126:161–9.
33. Rigby AS. Statistical methods in epidemiology. V. Towards an understanding of the kappa coefficient. Disabil Rehabil 2000;22:339–44.
34. Guggenmoos-Holzmann I. The meaning of kappa: probabilistic concepts of reliability and validity revised. J Clin Epidemiol 1996;49:775–82.
35. Ottenbacher K, Tomchek SD. Measurement in rehabilitation research: consistency versus consensus. Am J Phys Med Rehabil 1993;4:463–74.
36. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation and sample size requirements. Phys Ther 2006:85(3):257–68.
37. Thompson WD, Walter SD. Kappa and the concept of independent errors. J Clin Epidemiol 1988;41:969–70.
38. Devane D, Lalor J. Midwives' visual interpretation of intrapartum cardiotocographs: intra- and inter-observer agreement. J Adv Nurs 2005;52(2):133–41.
39. Gorelick MH, Yen K. The kappa statistic was representative of empirically observed inter-rater agreement for physical findings. J Clin Epidemiol 2006;59:859–61.
40. Feinstein AR, Cicchetti DV. High agreement but low kappa. I. The problems of two paradoxes. J Clin Epidemiol 1990;43:543–9.
41. Dormer A, Klar N. The statistical analysis of kappa statistics in multiple samples. J Clin Epidemiol 1996;49:1053–8.