JOURNAL OF THE
JAPANESE PHYSICAL
THERAPY ASSOCIATION

REPORT

Intra-Tester and Inter-Tester Reliability of Chest Expansion Measurement in Clients with Ankylosing Spondylitis and Healthy Individuals

Jagannath SHARMA[1,3], Hideaki SENJYU[1,2], Linda WILLIAMS[1] and Colin WHITE[3]

[1]*Curtin University of Technology, Western Australia*
[2]*Nagasaki University, Japan*
[3]*Infantry Training Centre, Catterick Garrison, England*

Abstract. The purpose of this study was to examine the intra-tester and inter-tester reliability of chest expansion (CE) using a tape measure, in people with ankylosing spondylitis (AS) and healthy subjects. Twenty-two subjects with AS with a mean age of 41.4 years and 25 healthy subjects with a mean age of 41.0 years were tested in two arm positions: hands on head and arms at the sides, the tape measure being placed at the level of xiphisternum. There were three testers for subjects with AS and two testers for healthy subjects. Three trials in both arm positions were recorded by each tester on two separate occasions which were 10 minutes apart. Results showed intraclass correlation coefficients (ICC) for intra-tester reliability good (0.85 to 0.97) across the occasions. Intraclass correlation coefficients for inter-tester reliability were also very good (0.93 to 0.97). As reliability is good it is suggested that CE can be used for monitoring disease progression and efficacy of intervention with confidence within tester and between testers.

Key words: ankylosing spondylitis, chest expansion measurement, intra-tester and inter-tester reliability

(J Jpn Phys Ther Assoc 7: 23–28, 2004)

Chest Expansion (CE) measurement with a tape measure has been commonly used in diagnosis, to monitor disease progression and as an evaluation tool for efficacy of intervention of ankylosing spondylitis (AS). Ankylosing spondylitis is a chronic, systemic inflammatory rheumatic disease of unknown etiology, mainly affecting the spine[1)2)] in which limitations of mobility of spinal joints and chest wall increase with duration of disease[3)]. Ankylosing spondylitis predominantly affects young adults and has about 1% prevalence in the population[4)]. Of people with AS, approximately 10% become significantly disabled within 20 years[5)]. As the precise natural history, optimal treatment strategies and outcome measures are not clear, AS requires long term management and regular monitoring. This indicates that AS places a high demand on health services and a burden on health budgets. Thus there are important economic implications for management of AS[6)].

There is currently no cure, but the disease can be managed adequately by non-steroidal anti-inflammatory drugs, to reduce pain and inflammation, and regular exercise to improve mobility, strength and fitness[1)]. Lubrano and Helliwell[7)] reported that unless the disease is treated aggressively in the early stage it may progress to total spinal ankylosis.

Various monitoring and assessment strategies have been used, including measurement of chest expansion (CE), vital capacity (VC), thoracolumbar flexibility and Schober's test. These have also been utilised to evaluate the benefits of exercise. However, there has been no consensus as to which are the most useful measurement tools for efficacy, accuracy, sensitivity or reliability of these outcome measures with AS[8)].

Most studies have focused on CE as an outcome measure[3)4)]. The usefulness of measurement depends on how accurate it is. Inter-tester and intra-tester reliability studies are required to ascertain whether changes are due to

Received: April 14, 2003
Accepted: December 6, 2003

Correspondence to: Jagannath Sharma, Physiotherapy Department, Headquarters, Infantry Training Centre, Vimy Barracks, Catterick Garrison, North Yorkshire DL9 3PS, England, UK.
e-mail: jaga_s9@yahoo.com

the effects of intervention or error of measurement. Without such reliability studies, one can neither have confidence in the collected data nor draw rational conclusions from those measurement. Thus usefulness of CE measurement as an outcome measure is questionable[9]. Nevertheless, the method continues to be used in the clinical setting because the technique is cost and time effective, requiring minimal equipment and being simple to perform.

There is a paucity of research into the reliability of CE measurements, and in those studies which have been performed results have proven conflicting[10][11]. In their reliability study Pile *et al.*[10] of 10 subjects with AS found a coefficient of reliability of CE was 0.15. However, Viitanen *et al.*[11] looked at intra-tester and inter-tester reliability in 52 subjects with AS. They found that ICC was 0.95 and 0.85 intra and inter tester reliability respectively. Therefore, reliability between arm positions and groups, are not only worthwhile, but essential to provide information to clinicians concerning this widely used method for management of AS.

The purpose of this study is to examine the inter and intra-tester reliability of CE measurement using a tape measure, subjects with AS and healthy subjects. This study aims to extend and contribute to the body of knowledge regarding the use of CE in the assessment and management of AS.

Methods

Subjects

Subjects were recruited from the metropolitan area of Perth, Western Australia. The 22 subjects with AS comprised 13 males and 9 females, and had a mean age of 41.4 years (SD=11.6; range=22 to 61): mean, SD and range of disease duration of 12.9, 11.9 and 1 to 46 years respectively. The 25 healthy subjects (16 males and 9 females) had a mean age of 41.0 years (SD=12.1, range=26 to 65). The subjects with AS were measured on two occasions by three investigators, while the healthy subjects were measured on two occasions by two investigators. Measurements were taken at the level of the xiphisternum in a standing position, using two different arm positions: hands on head and arms at the sides.

Three physiotherapist (A, B and C) were used for the subjects with AS and two (A and C) for the healthy subjects. This was due to tester B being not available to test for healthy subjects. Testers A and B both had eight years clinical experience in the management of people with AS, both used the CE test in routine clinical practice, whereas tester C was a new graduate with minimal experience in the rheumatology area.

The measurement instrument for this study was a simple, retracting, flexible metal tape of three metres. The tape was clearly marked in fractions of both inches (16ths) and centimeters (milimeters), with ten centimeter increments being marked in red.

This instrument was chosen for its simplicity, ease of use, inexpensiveness and wide use in clinical trials of this nature, being routinely used in clinics to assess subjects with AS.

Procedure

Each tester took three trials in each of the two arm positions on each occasion. The duration of trials were the time taken to complete the measurements of maximum inspiration and expiration to calculate the difference. This difference is the CE measurement in centimeters. This procedure was repeated immediately for three trials 1,2,3 on each occasion.

Subjects were seen on two occasions, ten minutes apart, on the same day, which was convenient for subjects and testers' by minimising attendance time. Ten minutes was chosen as the time period between CE measurement occasions in the expectation that this time interval would prevent experimental bias by minimising memory of the previous results. Within the given measurement period, each tester measured at least two subjects, and this was also intended to obviate the possibility of tester memory affecting results. Subjects were asked to wear one layer of clothing during testing.

The tape was placed in the centre of the xiphisternum, in such a way as to ensure that it remained horizontal with the xiphisternum by using a grid as a guide. The subject was then requested to hold the tape while the investigator check the level of the tape from anterior, posterior and lateral views. Once the investigator holds the tape, subjects were instructed to "fill your lungs right up with air and hold while I measure, then breathe out completely and then I will measure you again". Encouragement such as 'right in' was not given since providing encouragement might influence CE measurement from trial to trial[12].

The CE at maximal inspiration and at maximal expiration were recorded at each trial.

The measurement data sheets were passed on to the principal investigator, who calculated the CE, which is the difference between maximal inspiration and expiration measurements. This system was adopted to minimise the possibility of inaccuracies, which might have occurred had the testers calculated the CE on the spot, and also to reduce the likelihood of contamination of results due to the testers' memory of the previous record.

Statistical design

Data were analysed using SPSS Version 10.05 statistical package[13]. To evaluate intra-tester reliability, the average measurement of each occasion was calculated. Using these averages, the ICC was calculated by means of a two-way mixed effect model, average measure (model 3,K). This average was chosen to better represent whether an observed change represents real change or a fluctuation measurement[14]. Furthermore, this helps to minimise unpredictable measurement changes from trial to trial. The inter-tester reliability was tested using a two-way random effect model, average measure ICC (model 2, K)[13]. The 95% confidence interval and critical alpha level of p value of 0.05 were computed for every ICC to determine statistical significance.

Results

Intra-tester reliability of chest expansion

In order to answer intratester reproducibility, the three trials were averaged on each occasion in each arm position for each subject. The mean, SD and range of CE in cm are presented in Table 1. To calculate the ICC a two-way mixed effect model was used. As shows in Table 2 the ICCs of subjects with AS ranged from 0.85 to 0.97 and those of healthy subjects from 0.90 to 0.96 across occasions. The 95% CI in subjects with AS ranged from 0.64 to 0.98 and in healthy subjects from 0.79 to 0.98. Consideration of the ICC gives a p value of <0.001 for all testers, both groups and both arm positions.

The ICC, 95% confidence interval and p value showed a measure of reliability catagorised by Landis and Koch[15] as falling within the good-to-perfect range of reliability. The results therefore provide evidence to answer in the affirmative the research question "Is measurement of CE reproducible across two occasions?"

Inter-tester reliability of chest expansion

For inter-tester reliability a two-way random effect model, average measure[13] was used. The six trials in each arm position were averaged and the average value for each tester for each subject was used to calculate inter tester reliability. The mean, standard deviations and range of CE

Table 1. Mean, standard deviation and ranges of CE in cm across trials and testers

	Tester A		Tester B		Tester C	
	Mean (SD)	Range	Mean (SD)	Range	Mean (SD)	Range
AS Group						
Arms at sides in cm						
Occasion 1	6.26 (2.33)	2.37 to 10.53	5.66 (2.32)	1.67 to 10.50	6.45 (2.63)	2.33 to 12.70
Occasion 2	6.11 (3.32)	2.37 to 17.43	5.58 (2.81)	2.23 to 13.17	6.79 (3.01)	2.50 to 12.67
Hands on head in cm						
Occasion 1	5.76 (2.23)	2.47 to 10.63	5.32 (2.31)	1.73 to 9.67	6.44 (2.74)	2.33 to 14.27
Occasion 2	5.69 (2.49)	2.40 to 11.40	5.48 (2.24)	2.00 to 9.67	6.63 (2.64)	2.87 to 12.00
Healthy Group						
Arms at sides in cm						
Occasion 1	7.22 (2.43)	2.80 to 12.00			6.73 (2.14)	1.80 to 9.93
Occasion 2	6.83 (2.59)	2.47 to 11.43			6.95 (2.32)	2.67 to 10.67
Hands on head in cm						
Occasion 1	7.05 (2.27)	2.80 to 11.57			6.74 (2.21)	2.50 to 10.23
Occasion 2	6.93 (2.22)	2.00 to 11.17			6.96 (2.21)	2.67 to 10.27

Table 2. Intraclass correlation coefficients and 95 per cent confidence interval for intra-tester reliability across the occasions for each tester in both arm positions for both groups

	Tester A		Tester B		Tester C	
	ICC	95%CI	ICC	95%CI	ICC	95%CI
AS group						
Arms at sides	0.88	0.70 to 0.95	0.92	0.81 to 0.96	0.85	0.64 to 0.93
Hands on head	0.97	0.92 to 0.98	0.95	0.89 to 0.98	0.92	0.81 to 0.96
Healthy group						
Arms at sides	0.90	0.78 to 0.96			0.93	0.84 to 0.97
Hands on head	0.93	0.86 to 0.97			0.96	0.91 to 0.98

$p<0.001$

Table 3. Mean, standard deviation and range across six trials of each tester, both group and arm positions

	Arms at side in cm			Hands on head in cm		
	Mean	SD	Range	Mean	SD	Range
AS group						
Tester A	6.19	2.71	2.57 to 13.98	5.73	2.33	2.57 to 11.02
Tester B	5.62	2.49	2.17 to 11.83	5.40	2.23	2.03 to 9.67
Tester C	6.62	2.64	3.67 to 12.55	6.54	2.60	3.33 to 13.13
Healthy group						
Tester A	7.02	2.40	2.80 to 11.18	6.99	2.18	3.22 to 11.37
Tester B						
Tester C	6.84	2.16	2.23 to 10.07	6.85	2.17	2.95 to 10.25

Table 4. Intraclass correlation and 95 per cent confidence interval for inter-tester reliability in two arm positions for both groups

	Hands on head			Arms at sides		
	ICC	95%CI	p	ICC	95%CI	p
AS group	0.97	0.94 to 0.98	<0.001	0.96	0.92 to 0.98	<0.001
Healthy group	0.95	0.89 to 0.97	<0.001	0.93	0.84 to 0.97	<0.001

in cm for each tester are presented in Table 3.

Intraclass correlation coefficient, 95% CI and p value are given in Table 4. The ICC values for both groups and both arm positions range from 0.93 to 0.97. The 95% CIs were from 0.84 to 0.98. The resultant p value was <0.001 for both groups. The ICC, 95% CI and p value showed inter-tester reliability was good for both groups and both arm positions. These results therefore provide evidence to answer in the affirmative the research question “Is measurement of CE reproducible between three testers for the AS group and two testers for the healthy subjects on two occasions?”

Discussion

Intra-tester reliability of CE

The results in subjects with AS and healthy subjects indicate that CE measurement is reproducible by the same therapist on different occasions. The ICC across the occasions (0.85 to 0.97) suggests that the CE measurement using a tape measure has relatively high test-retest reliability. According to Keating and Matyas[14] it is possible that the data are more stable when averaged, there being no increase or decrease in CE measurement score.

The ICC for Tester C on the first occasion in the arms at side position was moderate. A possibility was that one subject found it difficult to bring the hands to the head due to shoulder pain and stiffness on the first occasion with Tester C. The stiffness might have dissipated by the time that subject went to the second tester or the second trial. This kind of occurrence might well be found in the clinical setting also.

The ICCs show very high reliability when calculated from mean rating. Generally, however, most clinicians take the best of three measurements or even take one measurement only. Clinically, it is possible, and, in the light of these results, it may be advantageous, to perform three trials and take a mean value for recording instead of a single rating.

This study compared favourably with findings by Viitanen *et al.*[11] who looked at intra-tester reliability in 52 subjects with AS. They found that intra-tester reliability ICC was 0.95. Another study by Viitanen *et al.*[3] of 38 subjects who were randomly selected from 151 subjects from an inpatient rehabilitation program found ICC values for intra-tester reliability of CE ranged from 0.72 to 0.92.

However, direct comparison of the results of the present study with Viitanen’s work was not possible due to lack of standardisation. In Viitanen’s studies, CE measurements were taken using landmarks at the fourth intercostal space. In addition, arm positions used by Viitanen were undefined. Furthermore, the ICC model used was not reported. Portney and Watkins[16] considered that the type of ICC should always be reported because of potential differences in results.

Similarly, Helliwell *et al.*[17] and Roberts *et al.*[12] found that CE measurement was reliable. Helliwell *et al.*[17] in a pilot study to examine intra-tester reliability in five subjects with AS, prior to their main study which investigated different physiotherapy regimes, found a mean difference of 2 mm and a standard deviation difference of 5 mm. The paired t-test shows differences are insignificant. The CE measurement was taken at the level of xiphisternum. The problems of this study are, the number of subjects was very small and CE data were not given in their findings.

A study by Roberts *et al.*[12] investigated intra-tester reliability in 10 subjects with AS and 10 healthy subjects. Chest expansion, measured at xiphisternum level in the hands on head position, found that intra-tester reliability was good. Pearson's Product Correlation Coefficient was 0.95 in subjects with AS and 0.86 in healthy subjects.

Studies by Roberts *et al.*[12], Viitanen *et al.*[11] did not report CE data. It is hard to interpret their results in order to compare them with the present study as only reliability was reported.

In contrast, Fisher *et al.*[18] and Pile *et al.*[10] found that the CE measurement was unreliable. Fisher *et al.*[18] investigated the relation between CE and exercise tolerance in 33 subjects with AS. They found a reliability of coefficient of variance percentage of 14.0, and concluded that it was unreliable. Direct comparison with their study, however, is difficult due to the fact that their methodology differs from that of the present study: there are differences in landmarks and statistical methods. In this study, the CE was measured with a tape placed circumferentially around the chest wall at the fourth intercostal space, whereas in the present study it was placed at the xiphisternum. The data were calculated using reliability of coefficient of variance percentage which is a ratio of standard deviation to mean [18].

Also, in contrast to the present study, Pile *et al.*[10] in their reliability study of 10 subjects with AS found a coefficient of reliability of CE was 0.15 and the intra-observer differences were highly significant. They reported that to be 90% confident of the measured change being significant, a CE measurement greater than 1.2 cm within observers must be found. Pile *et al.*[10] analysed their data using the absolute value of the paired differences between each of the four measurements taken by each tester.

Inter-tester reliability of CE

Inter-tester reliability using ICC is found to be very high in both groups and both arm positions. The ICC for hands on head positions in the AS group was 0.97 and arms at sides was 0.96. The ICC for the healthy subjects group on hands on head position was 0.95 and arms at side was 0.93.

This is comparable with study by Viitanen *et al.*[11] which found inter-tester reliability moderate to good. Viitanen *et al.*[11] found the ICC value for inter-tester reliability was 0.85.

On the other hand, Pile *et al.*[10] in their reliability study found inter-observer reliability to be poor. They calculated inter-observer variation using absolute value of the paired differences for each measurement between the five observers. Analysis of variance showed consistently significant differences between the five observers. Absolute mean difference between observers was 3 cm. A 90th percentile inter-observer variation indicates values greater than 3 cm are more likely due to a real change than observer variation. The coefficient of reliability was 0.15.

The possible explanation for the low intertester reliability in the study of Pile *et al.*[10] was their inclusion criteria of CE less than 2.5 cm. This inclusion criterion would have made the variance in CE measurement very small. According to Portney and Watkins[16] if the total variance is small the reliability coefficient will probably be low even if measurement is fairly consistent.

In the present study the ICC for inter-tester reliability appears higher than the ICC value for intra-tester reliability. This is, in part, possibly due to the method of calculation. The average from six trials of CE score on two occasions for each tester was used to calculate the inter-tester reliability. However, for intra-tester reliability across occasions with three trials of CE score on each occasion were averaged. This also may have contributed to the disparity between intra-tester and inter-tester reliability. According to Portney and Watkins[16] the ICC based on mean rating always shows higher reliability than one based on single ratings. Another possibility is that the difference is due to the different model used to calculate the ICC.

A similar result was obtained in an intra-tester and inter-tester reliability study by Bennell *et al.*[19]. They found an inter-tester reliability of 0.99 and intra-tester reliability 0.97 to 0.98. They analysed the data using Model Two for inter-tester and Model Three for intra-tester reliability which were the same models used for this study. The high reliability found in the present study may be due to the wide spread of CE measurement, which prevents range effects from producing spuriously low reliability coefficients[20]. Using the mean of six trials for data analysis may also have caused high reliability results. This may provide a better representation of the subject's performance than a single measurement[14]. Last, we tested subjects with AS who are actively involved in an exercise program. In such subjects, CE is likely to be limited by bony, ligamentous and muscular factors. In contrast to subjects with AS in acute situation where chest pain and/or joints swelling due to inflammation[21] may be the limiting factors and these may render the CE measurement less reliable.

Acknowledgement

This research was undertaken to fulfill the requirement of the honours project at Curtin University of Technology, Western australia. The author acknowledges Dr. Marie Blackmore, Statistician, Curtin Univesity of Technology,

for her assistance on data analysis.

References

1) Bakker C, Hidding A, Linden SVD, *et al.*: Cost effectiveness of group physical therapy compared to individualized therapy for ankylosing spondylitis. J Rheumatol 21: 264–268, 1994.
2) Moll JMH, Wright V: Objective clinical study of chest expansion. Ann Rheum Dis 31: 1–8, 1972.
3) Viitanen JV, Suni J, Kautiainen H, *et al.*: Effect of physiotherapy on spinal mobility in ankylosing spondylitis. Scand J Rheumatol 21: 38–41, 1992.
4) Kraag G, Stokes B, Groh J, *et al.*: The effects of comprehensive home physiotherapy and supervision on patients with ankylosing spondylitis- A randomized control trial. J Rheumatol 17: 228–233, 1990.
5) Carette S, Graham D, Little H, *et al.*: The natural disease course of ankylosing spondylitis. Arthritis Rheu 26: 186–190, 1983.
6) Ryall NH, Helliwell PS: A critical review of ankylosing spondylitis. Criti Rev Phys Rehabil Med 10: 265–301, 1998.
7) Lubrano E, Helliwell P: Detorioration in anthropometric measures over six years in patients with ankylosing spondylitis. Physiotherapy 85: 138–143, 1999.
8) Calin A: The individual with ankylosing spondylitis: Defining disease status and the impact of the illness. Br J Rheumatol 34: 663–672, 1995.
9) Badley EM, Wood P: The why and the wherefore of measuring joint movement. Clin Rheumatol Dis 8: 533–544, 1982.
10) Pile KD, Laurent MR, Salmond CE, *et al.*: Clinical assessment of ankylosing spondylitis: A study of observer variation in spinal measurements. Br J Rheumatol 30: 29–34, 1991.
11) Viitanen JV, Heikkila S, Kokko ML, *et al.*: Clinical assessment of spinal mobility measurements in ankylosing spondylitis: a compact set for follow-up and trials? Clin Rheumatol 19: 131–137, 2000.
12) Roberts WN, Liang MH, Pallozzi LM, *et al.*: Effect of warming up on reliability of anthrometric techniques in ankylosing spondylitis. Arthritis Rheu 31: 549–552, 1988.
13) Coakes S, Steed L: SPSS. Version 10.05, Sydney, John Wiley and Sons, 1999.
14) Keating J, Matyas T: Unreliabile inferences from reliable measurements. AJP 44: 5–10, 1998.
15) Landis JR, Koch GC: The measurement of observer agreement for categorical data. Biometrics 33: 159–174, 1977.
16) Portney LG, Watkins MP: Foundations of Clinical Research Applications to Practice. Norwalk, Appleton & Lange, 1993.
17) Helliwell PS, Abbott CA, Chamberlain MA: A randomised trial of three different physiotherapy regimes in ankylosing spondylitis. Physiotherapy 82: 85–89, 1996.
18) Fisher LR, Cawley MID, Holgate ST: Relation between chest expansion, pulmonary function, and exercise tolerance in patients with ankylosing spondylitis. Ann Rheum Dis 49: 921–925, 1990.
19) Bennell K, Talbot R, Wajswelner H, *et al.*: Intra-reter and inter-rater reliability of a weight bearing lunge measure of ankle dorsiflexion. Austral J Physiother 44: 175–180, 1998.
20) Rosner B: Fundamentals of Biostatistics. 4th ed, New York, Duxbury Press Publisher, 1995.
21) Khan MA: Ankylosing spondylitis: clinical features. In: Klippel JH, Dieppe PA (eds) Rheumatology. London, Mosby, 1998, pp 16.1–16.10.