East African Medical Journal Vol. 91 No. 10 October 2014
RELIABILITY AND VALIDITY OF SUBJECTIVE ASSESSMENT OF LUMBAR LORDOSIS IN CONVENTIONAL RADIOGRAPHY
E. Ruhinda, DMR, BMR, Senior Radiographer, Joint Clinical Research Centre (JCRC) P.O.Box 10005, Kampala Uganda, R. K. Byanyima, MBChB, MMed (Rad), MBA, Senior Consultant, Mulago Hospital/Visiting Consultant JCRC P.O.Box 7051, Kampala Uganda and H. Mugerwa MBChB, MSc, Medical Officer and Biostatistician, JCRC P.O.Box 10005, Kampala Uganda

RELIABILITY AND VALIDITY OF SUBJECTIVE ASSESSMENT OF LUMBAR LORDOSIS IN CONVENTIONAL RADIOGRAPHY

E. RUHINDA, R. K. BYANYIMA and H. MUGERWA

ABSTRACT

***Background*: Reliability and validity studies of different lumbar curvature analysis and measurement techniques have been documented however there is limited literature on the reliability and validity of subjective visual analysis. Radiological assessment of lumbar lordotic curve aids in early diagnosis of conditions even before neurologic changes set in.**
***Objective*: To ascertain the level of reliability and validity of subjective assessment of lumbar lordosis in conventional radiography.**
***Design*: A blinded, repeated-measures diagnostic test was carried out on lumbar spine x-ray radiographs.**
***Setting*: Radiology Department at Joint Clinical Research Centre (JCRC), Mengo-Kampala-Uganda.**
***Subjects*: Seventy (70) lateral lumbar x-ray films were used for this study and were obtained from the archive of JCRC radiology department at Butikiro house, Mengo-Kampala.**
***Results*: Poor observer agreement, both inter- and intra-observer, with kappa values of 0.16 was found. Inter-observer agreement was poorer than intra-observer agreement. Kappa values significantly rose when the lumbar lordosis was clustered into four categories without grading each abnormality**
***Conclusion*: The results confirm that subjective assessment of lumbar lordosis has low reliability and validity. Film quality has limited influence on the observer reliability. This study further shows that fewer scale categories of lordosis abnormalities produce better observer reliability.**

INTRODUCTION

Lordosis, by definition, refers to a curvature of the spine in the sagittal plane in which the convexity of the curve is directed anteriorly as seen in the cervical and lumbar spine of humans. (1) This curvature makes sustained bipedal walking possible and also provides shock absorbing resilience and flexibility to the axial skeleton (2).

The shape of the lumbar lordosis is a result of the shape of the lumbo-sacral intervertebral discs which are wedge-shape with posterior height approximately 6 – 7 mm less than the anterior height. This is most pronounced between the fifth lumbar and first sacral vertebrae (3, 4).

The second factor that generates the lorditic curve is the wedge shape of L5 vertebra. The height of its posterior surface is approximately 3 mm less than the anterior height. Lastly each vertebra above L5 is inclined slightly backwards in relation to the vertebra below thus stretching the anterior parts of the anuli fibrosi and anterior longitudinal ligament (4).

Abnormal lumbar lordosis may be due to factors within the lumbar vertebrae itself (intrinsic) or from elsewhere (extrinsic). Examples of intrinsic causes of abnormal lumbar curvature include: muscular weakness and spasms, structural changes within the lumbar vertebrae or intervertebral discs such as spondylolisthesis, crush fractures, disc herniation, prolapse or degeneration (2, 5-9).

Extrinsic causes of abnormal lumbar lordosis include postural compensation for an exaggerated thoracic kyphosis, lower limb deformities such as flexion contractures of the hips and increased weight of the abdominal contents like in the case of a gravid uterus (10-12).

Lumbar spine radiographs are the standard first line radiological investigations worldwide. Interpretation of these radiographs includes precise quantification of the vertebral curvature which contributes to patient management decisions. This management may range from simple conservative observation to complex surgeries. Early diagnosis and management is always of utmost importance and the ability to accurately measure the spinal curvatures is a step towards early diagnosis.

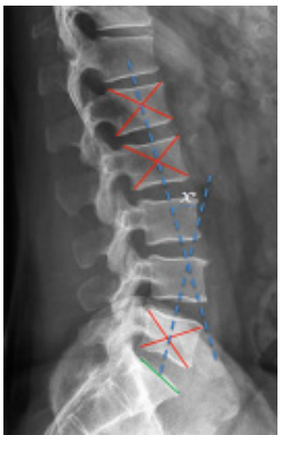
Roentgenometric analysis or Orthospinology x-ray analysis have been employed. These methods play an important role in film interpretation by allowing quantification of observed structural and biomechanical alterations.

Subjective analysis depends on the radiologist's experience and gives no room for precise comparative analysis on follow-up films. Most radiologists depend on subjective visual analysis to grade lumbar lordosis. The reliability and validity of the latter method of lumbar curvature evaluation are not known.

There are several roentgenometric techniques radiologists use to measure the lumbar lordosis. These include:

- *Vertebral centroid method (Figure 1)*
  The vertebral body centroids are approximated by the intersections of the body diagonals using all four vertebral body corners of L1, L2, and L5. The intersection of the perpendicular lines drawn from the proximal line connecting L1 and L2 centroid and distal line connecting L5 centroid and bisected sacral point provides the lumbar lordosis angle (13).

Figure 1
Lateral lumbar radiograph showing vertebral centroid method (Radiograph adapted from www.wikiradiography.com with permission from M. J. Fuller)

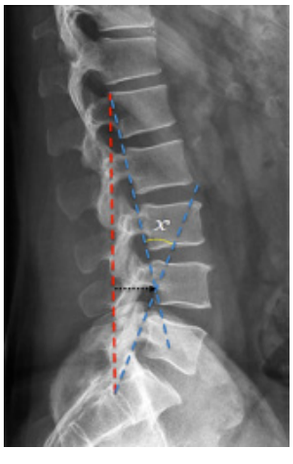


- *Harrison's posterior tangent method (Figure 2)*
  A line is drawn along the posterior body of L1 and a second line is drawn along the posterior body of L5. The superior or inferior angle of intersection is measured as the lumbar lordosis and averages 39.7 degrees with a standard deviation of 9.1 degrees (14).

Figure 2
Lateral lumbar radiograph showing Harrison's posterior tangent method

- *Tangential radiologic assessment of lumbar lordosis [TRALL] (Figure 3)*
  The largest perpendicular distance (black arrow) to the posterior longitudinal ligament from a line connecting the posterior–inferior corner of S1 and the superior–posterior body corner of L1 is used to locate the lumbar curve apex. This apex point is used as the vertex of the angle with the sides to L1 and S1 (13).

Figure 3
Lateral lumbar radiograph showing TRALL method (Radiograph adapted from www.wikiradiography.com with permission from M. J. Fuller)



- *Using fluoroscopic images and reflective markers*
  The co-ordinates of vertebral body center from fluoroscopic images are digitised. The vertebral

body center coordinates from fluoroscopic image and the coordinates from markers are then used to calculate lumbar lordotic angle.

- *Cobb's method (Figure 4)*
  With Cobb's method, the angle between the inferior endplate of L1 and the superior endplate of S1 is measured. Normal lumbar lordosis ranges from 31 to 79 degrees (15, 16).

  The Cobb's method was selected for this study because it is the most commonly used technique by clinicians for it provides a simple and quick measurement of lumbar lordosis (13). There is published literature about the accuracy of the Cobb's method compared to the other techniques.

Figure 4
Cobb's method (Radiograph adapted from www.wikiradiography.com with permission from M. J. Fuller)

MATERIALS AND METHODS

Lateral lumbar spine x-ray films obtained from the archive of JCRC radiology department were used. JCRC was founded in 1991 and serves as a national AIDS research center. JCRC Radiology department was established in 2006 and is equipped with a Philips Cosmos BS ® 2005 model x-ray machine with a kilovoltage range of 40 to 125kV and current range of 1.0 – 50mA. A table-top automatic film processor (OPTIMAX®) is used as well as a 45 x 80 cm Shenguang® film illuminator for film viewing. High speed green sensitive cassette film combination is used.

A sample size (n) of 66 was derived from sample size calculation for reliability studies (17). Assuming invalid-response rate of 15%, the sample size was adjusted to 77 films.

$$n = A^2 \left\{ \frac{[\pi(1-\pi)(k_1-k_0)]^2}{\pi^2 + \pi(1-\pi)k_0} + \frac{2[\pi(1-\pi)(k_1-k_0)]^2}{\pi(1-\pi)(1-k_0)} + \frac{[\pi(1-\pi)(k_1-k_0)]^2}{(1-\pi)^2 + \pi(1-\pi)k_0} \right\}$$

Where,

$A^2 = (Z_{1-a/2} + Z_{1-b})^2$

Type-I error (α) = 0.05

Type-II error (β) = 0.20 and thus Power = 0.80

Value of kappa characterized as representing substantial agreement (K_O) = 0.60

Assumed probability of abnormal lordosis (π) = 0.30

The radiographic film quality assessed included the following:

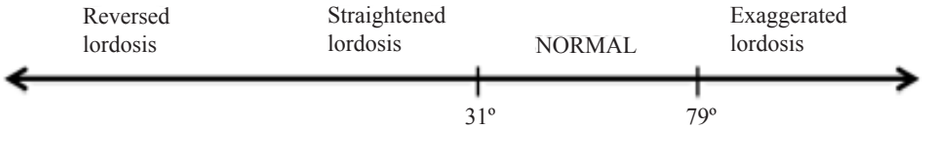
(i) The radiographic images including 12th thoracic vertebra down 1st sacral vertebrae which should be clearly visible
(ii) Clear view through the centre of the intervertebral disc spaces, with individual vertebral endplates superimposed.
(iii) Superimposed cortices at the posterior and anterior margins of the lumbar vertebral bodies.

The subjective grading for film quality was based on a Likert scale, where 1 = very poor, and 5 = very good.

Two observers selected for the study were practicing consultant radiologists in Kampala-Uganda, with at least ten years' experience in plain radiography film reporting. These two do not work for JCRC and they were also blinded of the original interpretation of these radiographs.

Subjective visual assessment was used to grade the lumbar lordosis of the film mounted on the standard film viewing box. A blinded repeat visual assessment of lumber lordosis was also done.

The Cobb's angle was measured and categorised based on a scale below:



The data were analysed using SPSS version 11.0.1 to obtain observer reliability based on kappa statistics. Validity was based on specificity and sensitivity using the Wilson score as obtained from OpenEpi, version 2 – open source calculator.

RESULTS

Seventy seven lumbar spine radiographs were used in this study though seven were excluded from analysis due to either invalid responses or incomplete data. There were 28 males (40 %) and 42 females (60%). Lowest age was 18 years and the highest 70 years, with a mean, median and standard deviation of 44.66, 45 and 12.06 years, respectively.

Each respondent assessed the same lumbar film twice at an interval of two weeks. The inter-observer agreement for each of the two occasions of film reading was poor with kappa values of 0.36 on the first reading and 0.16 on the second reading.

Intra-observer strength of agreement for Respondent-A was poorer ($k = 0.226$, $p < 0.001$) whereas respondent-B exhibited moderate strength of agreement ($k = 0.542$, $p < 0.001$).

When films of moderate-to-very good quality (3 to 5 on the Likert scale) were considered, there was no significant change intra- and inter-observer agreement.

Table 1
Observer agreement with respect to film quality

	All films	Only films of moderate to very good quality
Interobserver agreement		
1st episode of film reading	$\kappa = 0.36$	$\kappa = 0.36$
	($p < 0.001$), n = 70	($p < 0.001$), n = 58
2nd episode of film reading	$\kappa = 0.16$	$\kappa = 0.18$
	($p < 0.001$), n = 70	($p = 0.012$), n = 57
Intra-observer agreement		
Respondent-A	$\kappa = 0.226$	$\kappa = 0.285$
	($p < 0.001$), n = 70	($p < 0.001$), n = 61
Respondent-B	$\kappa = 0.542$	$\kappa = 0.561$
	($p < 0.001$), n = 70	($p < 0.001$), n = 60

Kappa values significantly rose more than two-fold when the lumbar lordosis was clustered into only four categories.

Table 2
Observer agreement with respect to the number of scale categories

	Observer agreement based on 9 category scale	Observer agreement based on 4 category scale
Inter-observer agreement		
1st episode of film reading	$\kappa = 0.36$	$\kappa = 0.64$
	($p < 0.001$), n = 70	($p < 0.001$)
2nd episode of film reading	$\kappa = 0.16$	$\kappa = 0.46$
	($p < 0.001$)	($p < 0.001$)
Intra-observer agreement		
Respondent-A	$\kappa = 0.23$	$\kappa = 0.49$
	($p < 0.001$)	($p < 0.001$)
Respondent-B	$\kappa = 0.54$	$\kappa = 0.63$
	($p < 0.001$), n = 70	($p < 0.001$)

Despite high sensitivity parameters (95% and 100% for respondent A and B respectively) the specificity parameters were low (12% and 2% for respondent A and B respectively).

Diagnostic accuracy, which is an incorporation of sensitivity and specificity, was 45.71% for respondent-A and 37.4% for respondent-B.

Table 3
Comparison of validity parameters

	Respondent A	Respondent B
1st episode of film reading		
Sensitivity (%)	95.24	100
Specificity (%)	24.29	10.2
Diagnostic Accuracy (%)	45.71	37.14
2nd episode of film reading		
Sensitivity (%)	95	100
Specificity (%)	12	2
Diagnostic accuracy (%)	35.71	30

DISCUSSION

Intra-observer reliability obtained in this study was below substantial level, with kappa value of 0.226 for respondent-A and 0.542 for respondent-B.

The better Intra-observer reliability exhibited by respondent-B could be due to difference in clinical experience. Respondent-B is employed at a National Referral and Teaching Hospital with bigger patient numbers compared to respondent A who works at a private hospital with lower workload. Respondent-B had five more years of clinical service compared to respondent-A. Not surprisingly, most studies comparing different groups of observers concluded that clinical experience had a significant bearing on ability to correctly interpret radiographic findings (18).

Intra-observer and inte-robserver agreement are both below substantial level, which indicates that difference in opinion is less likely to have been the cause of the interobserver variation. If this variation was attributed to genuine difference in opinion, one would expect significantly better Intra-observer reliability. Krupinski asserts that some of the errors that lead to observer variation can be attributed to technical difficulties concerning film quality, such as film underexposure (19).

In this study, a 5-point Likert scale was used by the observers to assess film quality. This was aimed at exploring whether film quality influenced observer consistence. The Intra-observer agreement about film quality was poor, $\kappa = 0.081$ ($p = 0.294$) for respondent-A and $\kappa = 0.302$ ($p < 0.001$) for respondent-B. Inter-observer agreement about film quality was equally poor with $\kappa = 0.151$ ($p = 0.023$) and $\kappa = 0.000$ ($p = 0.993$) on the first and second occasion of film reading respectively.

However, considering that some of the P-values are greater than 0.05, it is possible that this lack of agreement between the respondents is likely to have been due to chance alone.

When data from only films of moderate-to-very good quality (3 to 5 on the Likert scale) was analysed, the intra- and inter-observer agreement on lumbar lordosis grading was still below substantial. This comparison is elaborated in Table 4(b). This therefore downplays the effect of film quality on observer reliability.

It may be that there is no link between film quality and observer reliability. Alternately, it may be that there is an association but the study's design was not sensitive enough to identify the association due to extrinsic factors like the film reading environment and radiologists' mood.

Kappa values significantly rose more than two-fold when the lumbar lordosis was clustered into only four categories without grading the severity of each abnormality (into mild, moderate and gross). These findings are in agreement with those of Brennam & Silman (20).

The difference between subsequent grades is subtle, making it frequently difficult to subjectively differentiate "mildly straightened" from "normal". This perceptive difficulty in interpretation is the most plausible explanation for the higher kappa value obtained when smaller scale categories are used and poorer kappa with more categories.

The better agreement seen in the 4-category

scale of the study reflects that radiologists may agree about the presence of a certain abnormality but will not agree on its magnitude.

Most characteristics in this study were unevenly distributed with the majority of lumbar radiographs with lordosis in normal range (n = 50, 71.4%) and the smallest proportion being of exaggerated lumbar lordosis (n = 1, 1.4%).

Whereas this nature of the sample is probably a reflection of the population distribution of these abnormalities, the disadvantage of the kappa statistic is that it is affected by the prevalence of abnormality among the subjects and this, therefore, could have skewed the findings (20).

The observers unanimously agreed on films with exaggerated lumbar lordosis (Cobb's angle > 79 degrees). This subset of cases of exaggerated lordosis consisted of 1 film (1.4%) which both observers consistently assessed as so, on all occasions of film reading.

Surprisingly, interobserver agreement was poor (k = 0.333) when films with Cobb's angles less than 20 degrees (n = 4, 5.7%) were considered.

Hence, contrary to expectations, it could not be concluded that there is improved observer reliability at both extremes of abnormality.

It has been reported that where radiologists were aware of the clinical rationale for the study prior to interpreting any x-rays, the interpretation errors reduced (14).

This assertion is in tandem with an earlier study by Doubilet & Herman that showed an increase in the rate of true-positive readings when those doing film reading were availed a suggestive history (21).

However, for the purpose of eliminating a potential source of bias, the patients' history was excluded from this study, but in so doing, the effect this omission could have had on interpretation of lumbar lordosis was not represented in the study.

Lastly, the reliability of subjective analysis of lordosis obtained in this study was inferior to previously published series which compared various measurement techniques of lordosis assessment and showed high observer reliability with correlation coefficients ranging above 0.7. (13, 15, 21- 24)

A perfect test is never positive in a patient who is disease-free and is never negative in a patient who is diseased. However, it is well known that in radiology false positive and false negative decisions are occasionally made, both of which can impact on patient care and treatment (19).

From this study it can conclude that there is improved validity of subjective analysis at both extremes of abnormality of lumbar lordosis.

The single false-negative case by respondent-A was a radiograph with Cobb's angle of 25 degrees. Whereas one would presume that this error in diagnosis was because of a borderline angle of lordosis, it should be noted that there were seven (33.3%) other films with Cobb's angles ranging between 26 and 30 degrees, and another 12 (57.1%) films with Cobb's angles below 25 degrees and yet these were correctly identified as abnormal by the same respondent. The possible explanation for this observation could therefore not be deduced from the data collected.

The 24.5% (12 of 49) true-negatives scored by respondent-A on first occasion of film reading were films with mean and median Cobb's angles of 55.9 and 57.5 degrees respectively. On second occasion, respondent-A had a 12.2% (6 of 49) true-negative rate this time with films of mean and median Cobb's angles of 57.3 and 59 degrees respectively.

On the other hand, respondent-B had 10.2% (5 of 49) true-negatives on first occasion with mean and median Cobb's angles of 60.8 and 61 degrees respectively.

On second reading, respondent-B scored 2.0% (1 of 49) true-negatives with a Cobb's angle of 62 degrees.

These findings therefore indicate that specificity of subjective analysis of lumbar lordosis is improved at Cobb's angles above 55 degrees.

The overall diagnostic accuracy of Respondent-A was higher than that of respondent-B with 35% v 30% on first session of film reading and 45.71% v 37.4% on second session of film reading, despite poor reliability and lower sensitivity.

Despite high sensitivity parameters (95% for respondent-A and 100% for respondent-B) the specificity parameters were quite low (12% for respondent-A and 2% for respondent-B).

This lack of trade-off between sensitivity and specificity results in many patients with normal lumbar x-rays being told of the possibility that they have abnormal lumbar lordosis and are then subjected to further investigation or even unnecessary medication thus increased cost of health care. It may however be argued that false negative errors may be considered the most serious errors in the diagnosis because they are likely to have greater and more serious consequences than false positive errors.

In actual clinical practice, radiologists will seldom read seventy films in one day, let alone in a couple of hours. However, during this study, respondents were asked to assess more than seventy lumbar radiographs in a period of approximately two hours. This study does not entirely simulate a real-life situation of film reporting.

CONCLUSION

The findings of this study indicate that the task of assessing lordosis on lumbar radiographs and subsequently classifying the degree of curve abnormality is less reliable and less valid when using

subjective analysis even in the eyes of experienced radiologists.

Film quality does not influence observer reliability during interpretation, and what is perceived as good quality by one radiologist may not be seen as so by another.

Fewer scale categories of lordosis abnormalities produce higher observer reliability.

Caution has to be taken in interpreting prognosis of lumbar lordosis abnormalities and basing treatment decisions on radiographic findings derived by subjective assessment.

REFERENCES

1. Warner Jr. William. Kyphosis. In Raymond T. Morrissy & Stuart L. Weinstein, Lovell & Winter's pediatric orthopaedics, 6th ed., p.797. Lippincott Williams & Wilkins 2006.
2. Sammut Emanuel & Searle-Barnes Patrick, Osteopathic diagnosis. Stanley Thornes Ltd, pp. 139 – 142, 1998.
3. Matshes Evan W., Burbridge Brent, Sher Belinda, Mohamed Adel, Juurlink H. Bernhard. Human osteology & skeletal radiology: An atlas and guide, CRC Press, pp.192 – 207, 2005.
4. Bogduk Nikolai. Clinical anatomy of the lumbar spine and sacrum. 4th ed., Churchill Livingstone, pp.51-54, 2005.
5. McConnell Jonathan, Renata Eyres and Julie Nightingale. Interpreting trauma radiographs, Blackwell Publishing, p.217, 2005.
6. Bull Sheila. Skeletal radiography, a concise introduction to projection radiography. 2nd ed., Toolkit publications, , pp.154, 2005.
7. Adams Michael Anthony, Bogduk Nikolai, Burton Kim, Dolan Patricia, The Biomechanics of Back Pain, Churchill Livingstone, pp.160-78, 2004.
8. Sutton David. Textbook of radiology and imaging. 7th ed., Churchill Livingstone, p.1393, 2003.
9. Canale S. Terry & Beaty H. James. Canale & Beaty: Campbell's operative orthopedics, 11th ed., Mosby, 2007.
10. DePalma, Michael J. iSPINE – Evidence-based interventional spine care, DemosMedical (New York), pp.19 – 26, 2011.
11. McGraw, Kevin. Interventional radiology of the spine: image - guided pain therapy, Humana Press Inc.pp.75-6, 2004.
12. Snell, S. Richard. Clinical anatomy by systems. Lippincott Williams & Wilkins, p. 150, 2006.
13. Hong Jae Young, Suh Seung Woo, Modi N. Hitesh, Hur Chang Yong, Song Hae Ryong and Park Jong Hoon. Reliability analysis for radiographic measures of Lumbar lordosis in adults with scoliosis: a case control study comparing 6 methods. *Eur Spine J*, 2010; **19**: 551-57
14. Machiori Dennis. Clinical Imaging with skeletal, chest, and abdomen pattern differentials, 2nd ed., Mosby Elsevier, 2005 pp.175-204.
15. Harrison Deed E, Harrison Donald D, Cailliet Rene, Troyanovich Stephan J, Janik Tadeusz, Holland Burt. Cobb method or Harrison posterior tangent method: which to choose for lateral cervical radiographic analysis, *Spine*, 2000, **25**: 2072-2078
16. Warner Jr. William. Kyphosis. In Raymond T. Morrissy & Stuart L. Weinstein, Lovell & Winter's Pediatric Orthopaedics, 6th ed, p.797. Lippincott Williams & Wilkins, 2006.
17. Donner A & Eliasziw M. A goodness-of-fit approach to inference procedures for the kappa statistic: confidence interval construction, significance-testing and sample size estimation. *Stat Med*. 1992;**11**:1511-9.
18. Robinson P J, Radiology's Achilles' heel: error and variation in interpretation of Roentgen image, *Br J Radiol*, 1997; **70**:1085-1098
19. Krupinski A. Elizabeth, Berbaum S. Kevin, Caldwell T. Robert, Schartz M. Kevin, & Kim John, Long Radiology Workdays Reduce Detection and Accommodation Accuracy, *J Am Coll Radiol*, 2010; **7**: 698–704.
20. Brennan Paul & Silman Alan, (1992).Statistical methods for assessing observer variability in clinical measures, *BMJ*, 1992; **304**: 1491-94
21. Doubilet P & Herman PG, Interpretation of radiographs: effect of clinical history. *Am J Roentgenol*, 1981; **137**:1055-1058.
22. Taichi Tsuji, Yukihiro Matsuyama, Koji Sato, Yukiharu Hasegawa, Yu Yimin and Hisashi Iwata. Epidemiology of low back pain in the elderly: correlation with lumbar lordosis. *J Orthop Sci*, 2001; **6**: 307-311
23. Hicks E. Gregory, Steven Z. George, Michael A. Nevitt, Jane A. Cauley, Molly T. Vogt M. (October 2006). Measurement of Lumbar lordosis: inter-rater reliability, minimum detectable change and longitudinal variation. *J Spinal Disord Tech*, 2006; **19**:501-506.
24. Jin-Ho Hwang, Hitesh N. Modi, Seung-Woo Suh, Jae-Young Hong, Young-Hwan Park, Jong-Hoon Park, Jae-Hyuk Yang. (5th). Reliability of lumbar lordosis measurement in patients with spondylolisthesis: A case-control study comparing the Cobb, Centroid and Posterior Tangent methods, *Spine*, 2010; **35**: 1691-1700.