# Interobserver Discrepancies in Distance Measurements from Lumbar Spine CT Scans

G. Jerome Beers[1]
Anthony P. Carter[1]
Bruce E. Leiter[1]
Shripad P. Tilak[1]
Rangit R. Shah[1, 2]

**Lumbar spine computed tomographic (CT) scans of 10 patients were examined independently at two levels by five experienced radiologists. At each level the minimum midline sagittal diameter was measured, and at each intervertebral space the left foramen was measured for its minimum diameter. Statistically significant differences were found between the measurements of different observers, differences that in a number of cases could have led to disagreement over whether or not stenosis was present. There were reasonably strong correlations between different observers' readings of midline sagittal diameters but generally not of foraminal diameters. Reasons for discrepancies between observers in spine CT measurements are reviewed briefly.**

Our clinicians have requested that when we interpret spinal computed tomographic (CT) scans, we state exactly how large the foramina are, rather than reporting that they are "ample," "moderately to severely narrowed," etc. In fact, various authors have attempted to make the use of spinal CT as objective as possible in the diagnosis or treatment of spinal stenosis [1–6], lateral recess stenosis [4, 7], and disk protrusion [5] by using CT distance measurements. Yet it is known that distance measurements in spinal CT may not be reliable for a number of reasons, probably the most important being the strong effect of window settings on measurements [1, 3, 4, 8–12]. We conducted this study in the hope of helping to quantify how reliable distance measurements in spinal CT might be.

## Materials and Methods

Ten consecutive CT scans from the top of L4 to S1 obtained at University Hospital were selected from patients without previous surgery or residual contrast material. All scans were obtained on a 2002 Elscint scanner at Boston University Hospital using the "A" filter-function, a scan speed of 17 sec, the absorber (a beam hardener) out, "standard" collimation, and "normal" sample density. As recommended by the manufacturer for all scanning situations, 140 kVp and 43 mAs were used. The translate-rotate mode was used with a 140 mm reconstruction circle. Images were zoomed by a factor of 1.10. To obtain cuts approximately parallel to disks, typically a series of cuts was obtained from the top of L4 to the top of L5 and another from the top of L5 to the top of S1. No reformatting in other planes was performed in the 10 cases.

Five radiologists then measured the following distances on each scan: minimum midline sagittal diameter (MSD) at L4, MSD at L5, minimum width in the axial plane across the left L4–L5 foramen at a level where the root can be seen to traverse it (FD), and FD at L5–S1. To avoid being influenced by the window settings at which the technologists transferred the images to the floppy disks, all radiologists began by viewing the studies at window widths and centers of zero, and throughout the study had the "keep window" button pressed so that they saw only the window settings they chose themselves.

Apart from very specific instructions about what was to constitute MSD and FD for the purpose of the study, the radiologists were given no instructions on how to make measurements. Among other things, they were given no specific guidelines about window settings, whether to use black or white cursors, or how much attention to pay to the density readings

**This article appears in the November/December 1984 issue of *AJNR* and the February 1985 issue of *AJR*.**

Received March 12, 1984; accepted after revision May 16,1984.

[1] Department of Radiology, Boston University Medical Center, and Boston City Hospital, 818 Harrison Ave., Boston, MA 02118. Address reprint requests to G. J. Beers.

[2] Present address: Department of Radiology, Delaware County Memorial Hospital, Drexel Hill, DA 19026.

***AJNR* 5:787–790, November/December 1984**
0195–6108/84/0506–0787

that automatically appear whenever a cursor is used. It was suggested to each that the measurements should be completed in 60–90 min, but no specific time limits were set. Although the exact time taken to perform the measurements was not monitored, the time taken by different radiologists was reported to vary from about 90 min to about 3 hr. Some radiologists went through the studies only once, satisfied that they had thereby obtained the best possible measurements. Others went through some or all of the cases twice. Those who performed this repetition stated that virtually always they obtained the same reading the second time as the first. In measuring two foramina, one radiologist stated that he believed the correct measurement in millimeters was between one integer and the next; in these cases his measurements were tabulated as the mean between these two integers. (Thus, his 1–2 mm and his 2–3 mm were tabulated as 1.5 mm and 2.5 mm, respectively.)

Four of the five radiologists are staff neuroradiologists (G. J. B., A. P. C., S. P. T., and R. R. S.); one is a staff body scanner (B. E. L.). When the study was performed, all had shared at least about 1½ years working with CT in the same department, and two radiologists had worked together with CT for about 7 years. Three of the five, however, received substantial parts of their training in CT at other institutions. In our department the choice of bone and soft-tissue settings for obtaining spinal CT scans is usually left to the technologist. Thus, no formally prescribed set of window settings was anticipated by the radiologists in this study.

In one case, after four observers had obtained their measurements at the left L4–L5 foramen, the fifth observer found that the floppy disk had become degraded so that it would not display one cut necessary for evaluating the foramen.

For each MSD or FD, a mean measurement and standard deviation (SD) were determined. The ratio of each SD to its corresponding mean was also determined. The mean and median of this ratio were determined for both the FDs and the MSDs. For both the FDs taken together and the MSDs taken together, the mean of the means ($\overline{\text{mean}}$) and the mean of the SDs ($\overline{\text{SD}}$) were calculated. In calculating the $\overline{\text{mean}}$ and $\overline{\text{SD}}$ for the foramina, the means were weighted such that the means and SDs of the foramen that only four observers were able to measure were given only 80% of the weight of the means and SDs of the other 19 foramina. The maximum variations of each of the MSDs and the FDs were determined, and the means and medians of these maxima were calculated.

For both the FDs and the MSDs the means were determined from each observer's FD and MSD measurements. These means were compared with the other observers' mean FDs and MSDs, and the statistical significance of differences between different observers' readings was tested by the ranked sign test [13]. Here, as elsewhere in this study, if there was significance at the $p = 0.05$ level (two-tailed), the difference was considered "significant"; if the significance was at the $p = 0.01$ level (two-tailed), it was considered "highly significant."

The significance of the difference between $SD_{FD}$ and $SD_{MSD}$ was tested with the $t$ test [13]. In performing this particular calculation, that calculated for the foramen measured by only four observers was given as much weight as the others. The difference was calculated to be so far from significant that it is highly unlikely that this lack of weighting made any meaningful difference.

The regression coefficient was calculated for the MSDs and the FDs. When correlating the readings of the FDs made by one observer (observer E in table 1), who made only 19 sets of measurements, with those made by the other observers, the extra reading made by each of the others was discarded. Otherwise all 20 sets of measurements were used in determining correlations.

The regression coefficient was used to determine whether the presence of a positive correlation was statistically significant. The statistical signficance of the difference between the coefficients of the FDs on the one hand and the coefficients of the MSDs on the other was tested by means of the two-sample rank test [13].

TABLE 1: **Correlations of MSD and FD Measurements and Comparisons among Observers**

| Statistical Parameter: Observer | Observer A | B | C | D | E |
|---|---|---|---|---|---|
| Correlation of MSD measurements: | | | | | |
| A | ... | 0.6* | 0.8* | 0.7* | 0.7* |
| B | 0.6* | ... | 0.9* | 0.8* | 0.9* |
| C | 0.8* | 0.9* | ... | 0.9* | 0.7* |
| D | 0.7* | 0.8* | 0.9* | ... | 0.8* |
| E | 0.7* | 0.9* | 0.7* | 0.8* | ... |
| Correlation of FD measurements: | | | | | |
| A | ... | 0.5† | 0.4‡ | 0.5† | 0.3 |
| B | 0.5† | ... | 0.3 | 0.4‡ | 0.4 |
| C | 0.4‡ | 0.3 | ... | 0.8* | 0.6* |
| D | 0.5† | 0.4‡ | 0.8* | ... | 0.7* |
| E | 0.3 | 0.4 | 0.6* | 0.7* | ... |
| Differences (in mm) between observers' FD measurements: | | | | | |
| A | ... | B < A (0.2) | C > A (0.5) | D > A* (1.9) | E < A (0.3) |
| B | A > B (0.2) | ... | C > B (0.6) | D > B* (2.0) | E < B (0.2) |
| C | A < C (0.5) | B < C (0.6) | ... | D > C* (1.4) | E < C* (0.7) |
| D | A < D* (1.9) | B < D* (2.0) | C < D* (1.4) | ... | E < D* (2.2) |
| E | A > E (0.3) | B > E (0.2) | C > E* (0.7) | D > E* (2.2) | ... |
| Differences (in mm) between observers' MSD measurements: | | | | | |
| A | ... | B > A‡ (0.8) | C < A (0.4) | D > A (0.5) | E > A (0.1) |
| B | A < B‡ (0.8) | ... | C < B* (1.2) | D < B (0.4) | E < B†§ (0.8) |
| C | A > C (0.4) | B > C* (1.2) | ... | D > C* (0.9) | E > C‡ (0.5) |
| D | A < D (0.5) | B > D (0.4) | C < D* (0.9) | ... | E < D (0.4) |
| E | A < E (0.1) | B > E (0.8)†§ | C < E‡ (0.5) | D > E (0.4) | ... |

Note.—MSD = midline sagittal diameter; FD = foraminal diameter. Numeric correlations of MSD and FD measurements are expressed as regression coefficients. Differences between observers' FD measurements reflect differences between observers' average measurements; differences between observers' MSD measurements reflect differences between observers' mean measurements.

* Highly significant correlation.

† Statistically significant correlation.

‡ Correlation approaches being statistically significant.

§ Correlation approaches being highly significant.

## Results and Discussion

As shown in table 1, there was reasonably strong positive correlation (0.6–0.9; mean, 0.77) between different observers' measurements of the minimum MSDs of the canal. Moreover, the strength of the correlations was sufficiently strong that all correlations could be shown to be highly significant statistically. However, the correlations of measurements of the minimum FDs were not nearly so strong (0.3–0.8; mean, 0.48). Moreover, five out of 10 times the correlations of the readings of the FDs were sufficiently weak that a statistically significant positive correlation could not be shown.

The difference between the regression coefficients of the foramina and those of the MSDs was highly significant. Certain observers tended to arrive at higher values than others. Five pairs of observers demonstrated highly significant differences in foraminal readings. At one extreme, one observer's

average foraminal reading was 2.2 mm greater than another observer's. In the readings of the MSDs, there were two pairings of observers from which there were highly significant differences in measurements and one pairing from which there were significant differences. With the MSDs the greatest difference in average readings was 1.2 mm.

It is unclear to what extent an observer who tended to read high on the FDs would also read high on the MSDs. Observer D read significantly higher than three observers for MSDs and higher than all four other observers for the FDs. Observer E, on the other hand, had significantly higher readings for MSDs than did observer C, but had lower readings for FDs.

We had expected that the $\overline{SD}_{FD}$ would exceed the $\overline{SD}_{MSD}$ because of the relative difficulty in measuring foramina. In fact, the $\overline{SD}_{FD}$ at 1.2 was greater than the $\overline{SD}_{MSD}$, which was 1.0. This difference was not significant, however. A similar relation held between the average maximum disagreement between observers at a given foramen (3.1 mm) and the average maximum disagreement for an MSD (2.3 mm).

The SDs of the measurements of a given MSD or foramen, being on the average only about 1 mm, at first appear to suggest the discrepancies between different observers' measurements might be rather inconsequential. The average ratios of SD to mean suggest otherwise, however. In the case of the MSDs the mean ratio was 6.5% and the median 6.3%; in the case of the FDs the mean ratio was 33.2% and the median 30.8%. This means that in the case of the average MSD in our study, one had about a one in three chance of obtaining a reading that differed from the mean value by 6% or more; in the case of the average foramen in our study one had about a one in three chance of obtaining a measurement that varied from the mean value by 30% or more.

Another way of looking at the meaningfulness of the discrepancies between different observers' readings is to show their possible effect on diagnosis. For example, we often follow Verbiest [14] in using 13 mm as the lower limit of normal for MSDs. If this standard is applied to the 20 MSDs of the 10 spines used in our study, the five observers disagree on whether stenosis is present four times (or 20% of the time). Looking at different pairs of observers, one finds that different pairs disagreed zero to four times (0%–20% of the time). If 14 mm were considered the lower limit of normal there would be even greater disagreement.

Lacking a good single criterion for diagnosing foraminal stenosis, one of us uses for the foramina a criterion that has been used for diagnosing lateral recess stenosis [4, 7, 15]; namely, that while an FD of 4–5 mm is quite ambiguous, a diameter of 2–3 mm is highly suggestive of stenosis. Using this criterion, the five observers in this study would have disagreed in 16 cases (or 80% of the time) whether or not there was good evidence of foraminal stenosis. Pairs of observers would have disagreed in four to 10 cases (20%–50% of the time).

While we have no proof that our 10 cases are typical of all patients who are scanned, we suspect that in many practices many patients, like many of ours, have FDs or MSDs that are near the lower limit of normal, where 1–2 mm could lead to changes in therapy. Consequently, we find the degree of discrepancy in readings between observers disconcerting, but not very surprising, especially in the case of the foramina, given their small size (averaging 3.8 mm in our study) in relation to the smallest unit of measurement (1 mm) and the thickness of the slices (6 mm), the occasional tendency of their components to be oblique to the axial plane, and the tendency for some of their structures to differ little from each other in density.

This study was designed to quantify the discrepancies between different observers' measurements, not to quantify the relative importance of the causes of the discrepancies. Nevertheless, it is useful to consider the different types of error [16] that may have led to the discrepancies.

### *Blunders and Other Illegitimate Errors*

There were no clear-cut examples of blunders, although it is possible that, for example, an observer might have inadvertently skipped a relevant slice or might have recorded a measurement incorrectly. One possible example of a blunder was one level where one observer recorded 24 mm for one MSD whereas all the other observers measured it at 18 or 19 mm. Such a discrepancy might have arisen from a difference in understanding instructions in an unclear situation, for example, in a vertebra with asymmetric laminae with a deep anterior groove at their junction. There is no reason to suspect that there were any crucial equipment failures, except in the case where one observer was unable to obtain one measurement because of a problem with one slice on one floppy disk.

### *Random Errors*

In medicine, where one often relies on a single measurement to make a diagnosis, one often forgets that there is always a tendency to arrive at different figures when obtaining multiple measurements of the same unchanging entity. Random errors can arise from such things as small disturbances or fluctuating conditions in an environment (changes in line voltage, for example). They can also derive from errors of judgment and from ambiguities in the definition of what is to be measured.

That observers A and E had similar average foraminal measurements, but did not show a statistically significant correlation in their measurements, might be an example of a discrepancy arising largely from random error (or alternatively from complex combinations of nonrandom errors). On the other hand, participants volunteered that whenever they repeated measurements they usually obtained the same measurements the second time as the first. These comments suggest that random error might not have played a decisive role.

### *Systematic Errors*

These are errors in which all values are in error by a constant amount, which can be caused by errors of calibration, changes in experimental conditions, personal habits, or systematically imperfect techniques. That systematic errors played an important role in the discrepancies is strongly suggested by the fact that often one observer had a significantly higher average reading than another.

There are at least two systematic errors that could have led to discrepancies in our study. First, an observer might

have systematically misunderstood the rather specific instructions about what was to be measured. For example, he might have measured foramina from the disk or centrum anteriorly to the superior articular process posteriorly rather than to a joint capsule that was compressing a root.

Second, because of some particular preconception about the anatomy, an observer might have tended to see a given boundary (for example, the anterior border of the joint capsule) placed rather posteriorly whenever the image was ambiguous. Observer C tended to measure foramina as wider than those measured by E, but minimum MSDs were measured by C as narrower than by E, which suggests the existence of one or more such systematic errors.

Much of the discrepancy between observers likely resulted from systematic differences in choice of window settings, the choice of which strongly affects the apparent size of structures in CT images [1, 3, 4, 8–12] (as can be readily observed by looking at a foramen or lateral recess or a small structure while adjusting the window center). One probable example of the effect of window settings was the tendency of observer D, who generally likes to view all images with rather high window widths and centers, to report higher readings than the other observers both for FDs and for MSDs.

It has been reported that accurate distance measurements can be obtained if a cursor is placed on a given boundary when a fairly narrow window is centered halfway between the densities of the structures on either side of the boundary [1, 10]. Yet even very careful observers might be unable to use this rule to set windows similarly. First of all, one encounters circular reasoning, for, to determine the density of, say, a root, one must determine just where the root is. Yet, to determine where a root is requires determination of its boundaries, which is the point of the exercise. Second, use of the rule is further confounded by the heterogeneity of the densities of many of the relevant structures on spinal CT. Does one use the whole vertebral arch to determine the density of the arch, or just the cortex? If the latter, how does one readily determine where exactly the cortex is without knowing the right window settings? What settings does one use for a partly calcified joint capsule? Finally, even if one does not resort to histograms to help with these problems, using the rule would be overly time-consuming.

It should be noted that in several respects our study probably overestimated the reliability of spinal CT measurements. Different observers did not differ in scanning parameters such as filter function or slice thickness when making measurements on the same patient. (For example, perhaps if 3 mm cuts had been obtained rather than 6 mm cuts, smaller structures in the foramina might have been easier to differentiate. On the other hand, owing to increased mottle in the thinner slices, boundaries might have become less distinct). In addition, we did not examine the same patients on different machines.

Furthermore, our study dealt directly only with the discrepancies between different observers' measurements. It did not test directly the relation between the measurements and the actual value of what was measured. Even if we had all agreed on all measurements, we might have made similar errors derived, for example, from similar misconceptions of anatomy or from some consistent distortion of the data made by the equipment.

Our study only reports on our results using our equipment and our protocol on some of our patients. Our measurements agreed less than we would have liked. Other groups, drawing on different backgrounds and practicing in different settings, might obtain better or worse results. For example, it is easily conceivable that a group of radiologists, after extensively studying anatomic material together or after obtaining exhaustive and detailed surgical follow-up, might consistently obtain reliable measurements. Unless one has rigorously correlated one's distance measurements with anatomic or surgical findings, however, one should probably not put unquestioning faith in distance measurements made on spinal CT.

REFERENCES


1. Ullrich CG, Binet EF, Sanecki MG, Kieffer SA. Quantitative assessment of the lumbar spinal canal by computed tomography. *Radiology* **1980**;134:137–143
2. Postacchini F, Pezzeri G, Montanero A, Natali G. Computerized tomography in spinal stenosis. *J Bone Joint Surg [Br]* **1980**;62:78–82
3. Haughton VM, Williams AL. *Computed tomography of the spine*. St. Louis: Mosby, **1982**:210–215
4. Helms CA, Vogler JB III. Computed tomography of spinal stenosis and arthrosis. In: Genant HK, Chafetz N, Helms CA, eds. *Computed tomography of the lumbar spine*. San Francisco: University of California, **1982**:187–220
5. Glenn WV, Rothman SLG, Rhodes ML. Computed tomography/multiplanar reformatted (CT/MPR) examinations of the lumbar spine. In: Genant HK, Chafetz N, Helms CA, eds. *Computed tomography of the lumbar spine*. San Francisco: University of California, **1982**:87–124
6. Mall JC, Kaiser JA, Heithoff KB. Postoperative spine. In: Newton TH, Potts DG, eds. *Computed tomography of the spine and spinal cord*. San Anselmo, CA: Clavadel, **1983**:194–197
7. Ciric I, Mikhael MA, Tarkington JA. The lateral recess syndrome. *J Neurosurg* **1980**;53:433–443
8. Resjö M, Harwood-Nash DC, Fitz CR, Chuang S. Normal cord in infants and children examined with computed tomographic metrizamide myelography. *Radiology* **1979**;130:691–696
9. Koehler PR, Anderson RE, Baxter B. The effect of computed tomography view controls on anatomical measurements. *Radiology* **1979**;130:189–194
10. Seibert CE, Barnes JE, Dreisbach JN, Swanson WB, Heck RJ. Accurate CT measurements of the spinal cord using metrizamide physical factors. *AJNR* **1981**;2:75–85, *AJR* **1981**;136:777–780
11. Baxter BS, Sorenson JA. Factors affecting the measurement of size and CT number in computed tomography. *Invest Radiol* **1981**;16:337–341
12. DiChiro G, Schellinger D. Computed tomography of spinal cord after lumbar intrathecal introduction of metrizamide (computer assisted myelography). *Radiology* **1976**;120:101–104
13. Goldstein A. *Biostatistics: an introductory text*. New York: Macmillan, **1964**:51–59, 61–62, 144–146
14. Verbiest H. Results of surgical treatment of idiopathic developmental stenosis of the lumbar canal: a review of twenty seven years' experience. *J Bone Joint Surg [Br]* **1977**;59:181–188
15. Epstein JA, Epstein BA, Lavine L. Nerve root compression associated with narrowing of the lumbar spinal canal. *J Neurol Neurosurg Psychiatry* **1962**;25:165–176
16. Beers Y. *Introduction to the theory of error*, 2d ed. Reading, MA: Addison-Wesley, **1957**:1–6