Automatic Measurement of Vertebral Shape Using Active Shape Models

P.P. Smyth, C.J. Taylor and J.E. Adams
Depts. of Medical Biophysics & Diagnostic Radiology
University of Manchester
pps@sv1.smb.man.ac.uk

Abstract

In this paper, we describe how Active Shape Models (ASMs) have been used to accurately and robustly locate vertebrae in lateral Dual Energy X-ray Absorptiometry (DXA) images of the spine. DXA images are of low spatial resolution, and contain significant random and structural noise, providing a difficult challenge for object location methods. All vertebrae in the image were searched for simultaneously, improving robustness in location of individual vertebrae by making use of constraints on shape provided by the position of other vertebrae. We show that the use of ASMs with minimal user interaction allows accuracy to be obtained which is as good as that achievable by human operators, as well as high precision. Having located each vertebra, it is desirable to evaluate whether it has been located sufficiently accurately for shape measurements to be useful. We determined this on the basis of grey-level model fit, which was shown to usefully detect poorly located vertebrae, enabling accuracy to be improved by rejecting proposed search solutions whose grey-level fit was poorer than a threshold.

Introduction

Osteoporosis is a disease which affects a significant proportion of postmenopausal women. It is characterised by bone loss, resulting in vertebral, wrist and hip fractures. The most serious (and therefore important) type of fractures are hip fractures, which usually occur in very elderly patients. However, in trials of osteoporosis treatments, where treatment efficacy must be evaluated in as short a time as possible, vertebral fractures, which are less serious but occur in younger patients, are used as an indicator of osteoporosis. Rapid, accurate and reproducible measurement of vertebral deformity is therefore important in improving the statistical power of tests of treatment efficacy, and reducing the costs of large trials.

Until recently, measurements of vertebral deformity resulting from fracture were performed from 3 or 4 overlapping lateral radiographs covering the whole spine. However, Dual Energy X-ray Absorptiometry (DXA) scanners, which use a fan beam to obtain a digital image of the whole spine in a single pass, have now

BMVC 1996 doi:10.5244/C.10.51

become available [9]. These give a significantly lower x-ray dose to the patient, and eliminate projection effects due to non-parallel beam geometry. By imaging at two x-ray energies, the component of the x-ray absorption coefficient resulting from bone as opposed to soft tissue can be visualised, providing better bone-tissue contrast and a quantitative measure of the amount of bone present. However, this "bone" image is much lower in magnitude, and therefore has a poorer signal to noise ratio than either single energy image. The images from these scanners also exhibit poorer spatial resolution than standard radiography (1mm vs $\approx$0.3 mm), and include artefacts due to the patient's breathing. A typical scan is shown in figure 2.

The assessment of vertebral fracture has traditionally been performed by manually marking six points on each vertebra - at the four corners and at the midpoints of the upper and lower surfaces - then measuring anterior, mid, and posterior vertebral heights [8]. Ratios of these heights are used as the measure of vertebral fracture, in comparison to a normal reference population. A major disadvantage of this method is that it is extremely laborious and time consuming, taking an operator up to 15 minutes per patient to mark all points on a DXA image containing $\approx$11 vertebrae on a computer system using a mouse.

The problem of robustly locating regularly (but variably) shaped objects in a very noisy environment is generally best tackled using flexible template models. In such applications, model specificity is particularly important. Active Shape Models (ASMs) [4], which combine explicit models of object contours and grey-level appearance surrounding the contours, *only* allow realistic examples of shape to be generated, unlike other approaches. They have already been applied successfully to a range of medical image interpretation problems in 2D and 3D [1, 7].

In this particular application, the imaging environment is so noisy that, for ASM search to be robust, all available shape constraints must be used. For example, search for a single vertebra in the lower thoracic region (where breathing artefacts can occur) using a single vertebral model often fails completely. However, by searching for many vertebrae together, sufficient constraint is placed on the position and shape of each individual vertebra by the others, that satisfactory robustness can be obtained. We have analysed a large set of DXA scans and compared the accuracy of the ASM search method to that of human operators using the traditional approach.

An additional (and connected) problem in automatic segmentation of medical images is that of automatically assessing location accuracy. When a model shape and pose has been obtained which best matches the image evidence, it is often desirable to determine, from the model parameters and image evidence, how accurately the object has been located. In our application, if we could predict the accuracy of measurements based on the segmentation of a given vertebra we could, for example, exclude unreliable measurements from subsequent analyses. Knowledge of segmentation accuracy also affects the search process itself, as poorly located objects found early in the search process may hamper the location of subsequent objects.

We have attempted to find simple measures which correspond to changes in appearance resulting from poor visibility of the vertebrae. We show how our measures varied with "inherent" visibility of a vertebra, expressed by how repeatably

humans could locate its edges, and with segmentation accuracy, describing how far the solution obtained was from the true location of a vertebra.

We show that our measures can be used to successfully detect cases which have been poorly located, improving mean segmentation accuracy.

Active Shape Models

Active Shape Models(ASMs) use combined models of shape and grey-level appearance to search for objects in images. They are generated from a number of training images, which contain examples of the objects of interest.

Object shape is described by a Point Distribution Model (PDM), which is generated by performing principal components analysis on the variation in position of labelled landmark points over all the training examples. Each training example (containing n landmark points) is described as a vector of size $2n$. The PDM represents shape in terms of a mean shape and a set of linearly independant *modes* describing the main ways in which the training examples vary. A subset of the modes of variation is chosen so as to describe members of the training set to a chosen accuracy.

A new example object, $\boldsymbol{x} = (x_1, y_1, x_2, y_2, \dots)$, can then be generated by adding combinations of the subset of modes of variation $\boldsymbol{P}$ to the mean shape, $\bar{\boldsymbol{x}}$, with a vector of weights $\boldsymbol{b}$ controlling the influence of each mode:

$$\boldsymbol{x} = \bar{\boldsymbol{x}} + \boldsymbol{P}\boldsymbol{b} \tag{1}$$

Grey-level appearance is modelled by analysing the grey-level image profile at each landmark point in a direction perpendicular to the object contour. This appearance is modelled in a similar fashion to shape, using principal components analysis, describing the grey-level profile at each landmark point as a linear combination of a mean profile and linearly independant modes of variation. Full details of the training and use of these combined grey-level and shape models are given in [4].

Once a model has been constructed, ASM search can be used to locate the modelled object in new images. An initial approximation is projected into the image and iteratively refined. The grey-level landscape around the current position of each landmark point is probed for grey-level evidence which best matches the model. This suggests a better position for each landmark point. The PDM then attempts to deform itself to fit to these new suggested positions, within the constraints imposed by its modes of variation. This process is repeated until convergance. Because the PDM imposes global shape constraints, only objects of similar shape to those observed in the training set will be located in images.

A multiresolution approach, employing grey-level models trained on a Gaussian pyramid of images, has been shown to improve speed and robustness [3]. We have employed this approach in our experiments. In addition, methods for shape deformation which allow more flexibility to move *along* a contour rather than perpendicular to it [6], have been used throughout. This has been shown to be particularly effective in aiding location of objects with sharp corners, such as vertebrae.

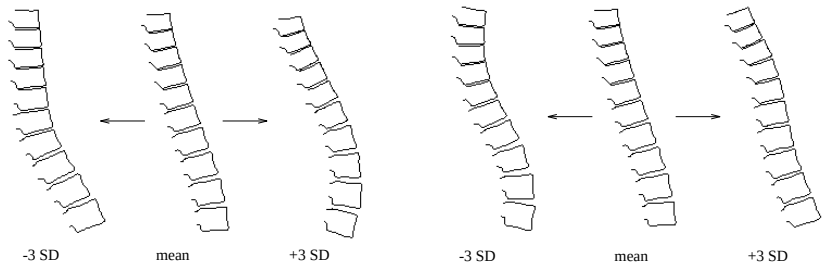

Figure 1: *The first two shape modes of the spine model.*

Methods

78 lateral spine DXA images of women aged over 45 were obtained using a Hologic QDR2000*plus* DXA scanner (Hologic Inc., Waltham MA). Landmark points were placed around the contours of ten vertebrae on each image by the author (PPS) on the advice of an experienced radiologist (JEA). Six thoracic vertebrae from T7 to T12, and four lumbar vertebrae from L1 to L4 were included. The first three modes of shape variation of the resulting spine model are shown in figure 1. It is worth noting that there is considerable interdependency in shape *between* vertebrae, emphasising the importance of considering a *combined* model of vertebrae for search, rather than for each vertebra in isolation.

For image search itself, a *moderately* good start position for the model is useful for eliminating complete search failures. The operator was asked to mark one point at the top of vertebra T7, one at the top of T12, and one at the bottom of L4. The model was then initialised from these three points, and multiresolution search started. The initial model position, and the search process through to convergence are shown in figures 2 to 4.

Comparison of Segmentation Accuracy With Manual Methods

In order to compare the accuracy of segmentation achieved using ASM search to the best manual methods, a random subset of 40 of the 78 images were marked up by 4 operators using the six point marking scheme descibed in [8].

A set of leave-one-out ASM search experiments was also carried out to provide an upper bound on the errors to be expected from the automatic method in real clinical use. The combined shape and grey-level appearance model was trained on all 78 examples except one. Its performance was then tested on the excluded example. This train-and-test process was repeated with each example in the training set left out in turn. Each search experiment was performed 20 times, with the position of the 3 manual start points randomly altered in accordance with operator variation. Precision values were obtained between these repeat experiments. The search results were characterised as a Gaussian distribution of successfully

Figure 2: *A DXA scan, with manually placed points marked.*

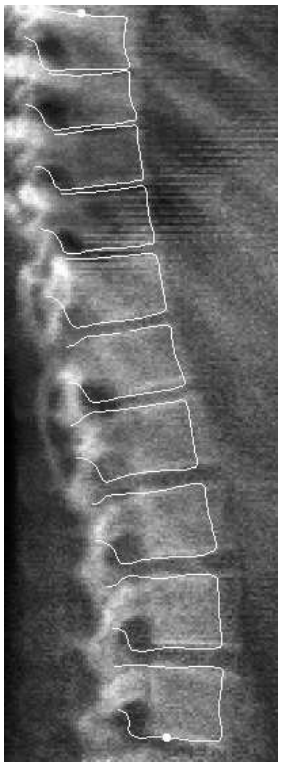

Figure 3: *Starting ASM search position, initialised from 3 points.*

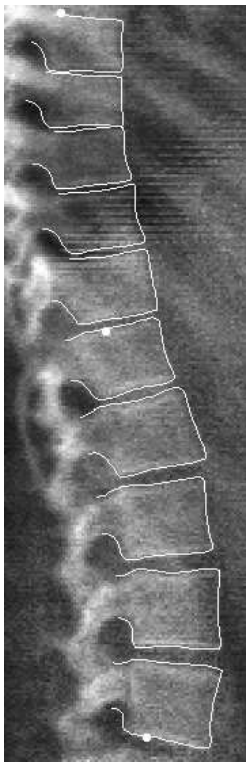

Figure 4: *Final ASM search position, after convergence.*

locationed vertebrae, and some failures. This was done by modelling the whole distribution as a mixture of Gaussians, using the EM algorithm. Cases were rejected as failures if their error lay above $\mu + 3\sigma$ of the distribution of "successful" cases.

The measures used to descibe accuracy were as follows: for the manual method, only the vertical component of error in the placement of vertebral midpoints is significant; however, both components of error in placement of vertebral corners are important. This is shown in figure 4. The accuracy was measured for each point by a Gaussian (2D, or 1D for midpoints) fitted to the distribution of the four operators' markings. For each vertebra, the root mean squared (rms) error (of the six points marked on each vertebra) was measured. The accuracy of the ASM segmentation was measured using the rms point-to-line distance errors from a landmark point at the "true" location of the object (as annotated during training) to the nearest point on the ASM located contour (figure 5).

Table 1 shows that the automatic ASM vertebra search performed as accurately as human operators. ASM search performed worse for the vertebrae at the extremes of the spine model, because these vertebrae were not surrounded by others in the model, and so the search was less constrained. Precision was very good,

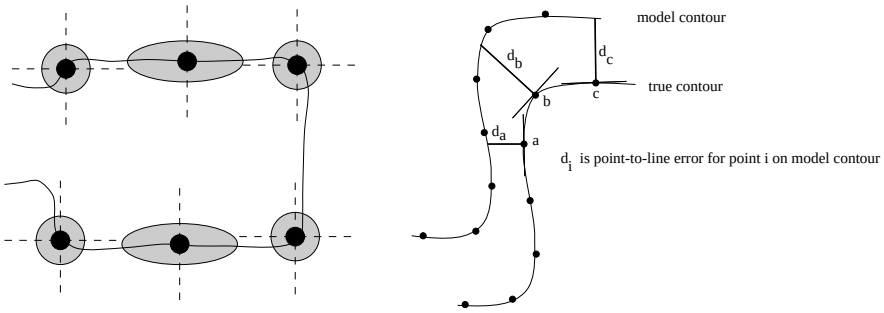

Figure 5: *(left) Manual vertebral markup error measurement. Dashed lines show directions in which errors are measured. Shaded areas give distribution of human points. (right) Measurement of point-to-line error from a true contour position to a proposed model search solution.*

	Manual	ASM		
Vertebral Level	Error	Error	Precision	% Failures
T7	1.66	1.60	0.17	0.8
T8	1.28	0.98	0.11	2.1
T9	1.15	0.98	0.12	3.4
T10	1.17	1.08	0.14	2.3
T11	1.18	1.10	0.21	3.7
T12	1.11	1.02	0.24	5.0
L1	1.12	1.17	0.26	5.3
L2	1.13	1.09	0.25	5.5
L3	1.15	1.08	0.25	6.0
L4	1.47	1.61	0.28	2.2

Table 1: *Manual and ASM errors (pixels rms), broken down by vertebral level.*

and the failure rate was low. Such good accuracy figures are very encouraging indeed, and offer the hope that this method may shortly replace manual operators at this particular task.

Automatic Assessment of Segmentation Accuracy

During ASM search, the image evidence is repeatedly compared to the grey-level model for the profile about each landmark point. A *quality-of-fit* measure is used to assess the extent to which the current evidence differs from the grey-level model. The search attempts to minimise this measure for each profile. The fit measure describes the departure of the evidence from the mean profile *along* modes of variation of the model (i.e. in ways already observed from the training data), and *perpendicular* to the model modes (i.e. in ways *not* observed in the training data).

This can be represented by a factor model [2]. The measure, f, is given by

$$f = \sum_{i=1}^{i=t} \frac{b_i^2}{\lambda_i} + \sum_{j=1}^{j=n} \frac{r_j^2}{v_j} \tag{2}$$

for a grey-level model with t modes, where b_i is the ith model parameter of the best fit of the model to the data, λ_i is the ith eigenvalue of the model, and r_j is the residual for the jth modelled pixel of n, whose variance (from jackknife estimations) is v_j. The measure reflects the degree to which a candidate profile differs from those observed in the training set, accounting for the fact that some pixels may be modelled better than others. The measure may therefore be useful for deciding how inherently visibile an object is. We expect this inherent visibility to be reflected in the size of manual or ASM errors in locating the object. In our experiments attempting to describe visbility, we use both of these measures as surrogates for visibility.

We have attempted to describe the visbility of vertebra using the fit measure, as a basis for deciding whether to accept or reject the final ASM search solution for each vertebra. ASMs contain separate grey-level models for the profile around *each* landmark point, so there is a need to combine the results for all the landmarks in some way. The simplest way to combine the fits for individual profiles is to add them - this is only valid if they are completely independent. However, a more principled probabilistic approach is to build a concatenated grey-level model for each vertebra. We built a concatenated grey-level model which described *all* of the profiles for a vertebra together: as a result it could model interdependencies *between* neighbouring profiles. It was created by concatenating the grey-level profiles for all the landmark points for each training example, and building a single model from the set of concatenated profiles [5].

In instances where one profile was near structural noise in the image, similar to its ideal profile, the absence of such structure in the neighbouring profiles would enable the noise to be ignored. With local grey-level models, the neighbouring profiles are assumed to be independent, so fitting to such noise becomes more likely. A concatenated grey-level model might therefore be fairer in its assignment of poor fit values to poorly visible structures. Search with the concatenated grey-level model can only be performed with a global optimiser, which is slow. We performed search with the concatenated model starting from the final solution obtained with a standard ASM. The fit measure for each vertebra were obtained from the final search position with the concatenated model. Importantly, the concatenated grey-level model allows true probability distributions to be obtained for the fit measures, enabling the measure to be treated in a principled way.

It would be expected that the grey-level fit measure of a particular search solution would vary with both the inherent visibility of a vertebra, and the nearness of the solution to the "true" one. To investigate the first behaviour, we plotted the grey-level fit measure of a good solution against the manual error (representing the inherent visibility of a vertebra). To examine how the fit measure varied as location accuracy worsened, we repeated our previous search experiments, but additionally varied the number of modes used in the shape model, in order to vary the accuracy of the final search solution.

Figures 6 and 7 show how the grey-level fit measure varied with "inherent" visibility (manual error) and with the location accuracy of the ASM search solution, for two example vertebra, using the concatenated and local grey-level models. They show that inherent visibility was very weakly correlated with fit measure ($r \approx 0.2$), while ASM accuracy was related more strongly. Search using concatenated grey-level models was too successful to obtain sufficient failures to test the response of the fit measure to poor location accuracy. Otherwise the two grey-level modelling methods appeared to perform similarly.

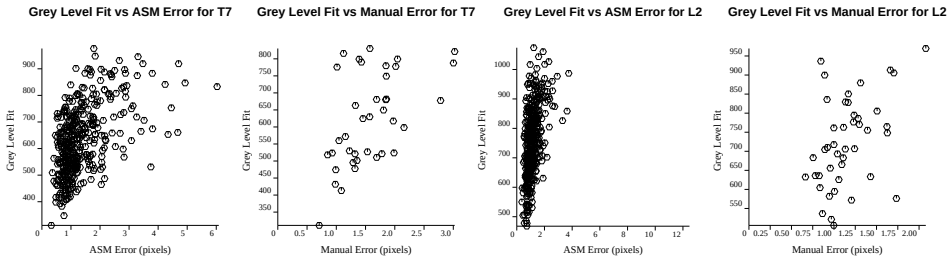

Figure 6: *Plot of grey-level fit measure (for a concatenated grey-level model) against manual and ASM error, for the T7 and L2 vertebrae.*

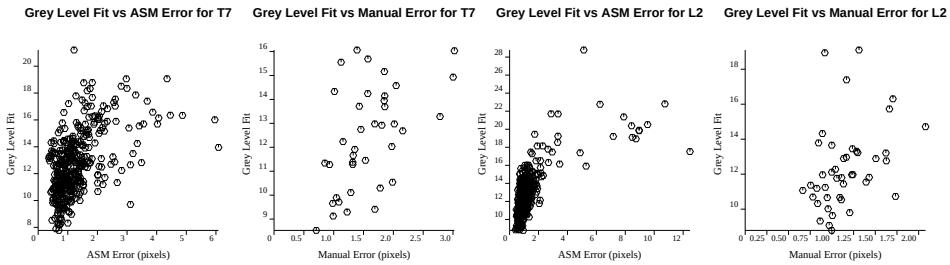

Figure 7: *Plot of grey-level fit measure (for local grey-level models) against manual and ASM error, for the T7 and L2 vertebrae.*

The fact that there is reasonable correlation between ASM accuracy and grey-level fit suggests that we may be able to reject poorly fitted solutions automatically. We investigated this using ROC analysis.

As one lowers the fit measure threshold for rejecting search solutions, the proportion of poorly located cases successfully detected (true positives) increases, as does the proportion of accuractely located cases misdetected as being poorly located (false positives). Having chosen a threshold on location error above which a case is judged as poorly located, one can plot an ROC curve describing the power of the fit measure to detect poorly located cases. Figure 8 shows such curves for vertebrae T7 and L2 using local grey-level models - solutions with rms error greater than 3 pixels were defined as poorly located.

For the L2 lumbar vertebra, using local grey-level models, large values for the fit measure were associated with poor location accuracy. This behaviour was

repeated for other vertebrae from T10-L4, which were also easily visible. However, for vertebrae such as T7, which was often poorly visualised, the correlation was weaker, and rejecting solutions on the basis of fit measure would be less successful. The example ROC curves show the fit measure to have real value in detecting poorly located cases at L2, while at T7 it might not be practically useful, because a significant number of false positives would have to be accepted in order to obtain a useful true positive rate.

One possible reason for the poorer performance of models which were trained from less visible vertebrae, and the lack of correlation between "inherent" visbility and grey-level fit measure, may lie in the fact that the fit measure describes how far an unseen example deviates in appearance from the *training data*, and not from good, clearly visible examples. It may be possible that the fit measure of a grey-level model built from just the most visible cases within a set would perform better.

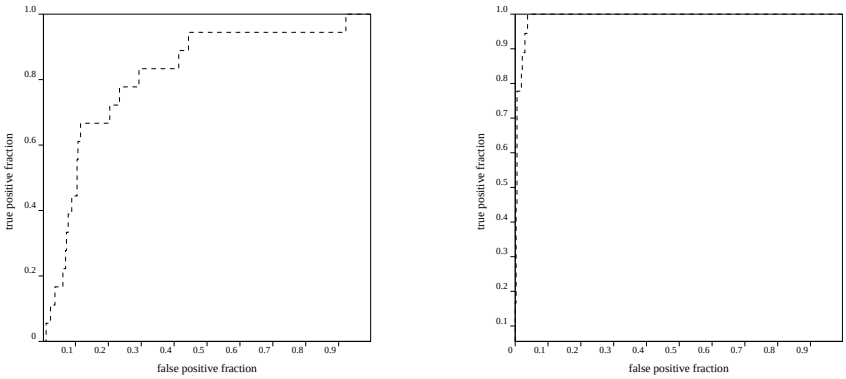

Figure 8: *ROC curves for detection of poorly located cases by grey-level fit measure, for vertebrae T7 (left) and L2 (right).*

Conclusions and Further Work

We have shown that ASM search is an accurate and robust tool for automatically analysing DXA spine images by testing it over a large set of images. It performed as well as manual observers at accurately locating the vertebrae from T7-L4. It was fast, taking approximately 30 seconds to analyse a scan that would take a human operator up to 15 minutes manually. It accurately located the *full* vertebral contour, rather than just six points. These results form the basis for its development for use as an clinical system.

Recognising that automatic assessment of segmentation accuracy is important to a clinical system's usefulness, we have investigated the description of the visibility in terms of the grey-level quality-of-fit measures, using both a local and concatenated grey-level model. We have shown that the local measures performed well in detecting poorly localised solutions for vertebrae which were generally well visualised. It was more difficult to detect poorly located solutions for vertebrae

which were usually poorly visualised. Improvements might be obtained if one excludes poorly visible examples from the training set. Too few failed cases were obtained with search using the concatenated grey-level model to properly evaluate its worth.

We next intend to train the model to perform fracture classification based upon the parameters of the shape model. Development into a working clinical system should then follow.

References

[1] T. F. Cootes, A. Hill, C. J. Taylor, and J. Haslam. The Use of Active Shape Models for Locating Structures in Medical Images. *Image and Vision Computing*, 12(6):276–285, 1994.

[2] T. F. Cootes, G. Page, C. Jackson, and C. J. Taylor. Statistical grey-level models for object location and identification. In 6^{th} *British Machine Vison Conference*, pages 533–542, Birmingham, England, Sept. 1995. BMVA Press.

[3] T. F. Cootes, C. Taylor, and A. Lanitis. Active Shape Models: Evaluation of a Multi-Resolution Method for Improving Image Search. In E. Hancock, editor, 5^{th} *British Machine Vision Conference*, pages 327–336, York, England, 1994. BMVA Press.

[4] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active Shape Models - Their Training and Application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

[5] J. Haslam, C. J. Taylor, and T. F. Cootes. A probabilistic fitness measure for deformable template models. In 5^{th} *British Machine Vison Conference*, pages 33–42, York, England, Sept. 1994. BMVA Press.

[6] A. Hill, T. F. Cootes, and C. J. Taylor. Active Shape Models and the Shape Approximation Problem. In 6^{th} *British Machine Vision Conference*, page (to appear), Birmingham, England, 1995. BMVA Press.

[7] A. Hill, A. Thornham, and C. J. Taylor. Model-Based Interpretation of 3D Medical Images. In J. Illingworth, editor, 4^{th} *British Machine Vision Conference*, pages 339–348, Guildford, England, 1993. BMVA Press.

[8] R. Smith-Bindman, P. Steiger, S. Cummings, and H. Genant. A comparison of the morphometric definitions of vertebral fracture. *J. Bone Miner. Res.*, 5:25–34, 1991.

[9] P. Steiger, S. Cummings, H. Genant, H. Weiss, and the Study of Osteoporotic Fractures Group. Morphometric x-ray absorptiometry of the spine: Correlation in vivo with morphometric radiography. *Osteoporosis Int.*, 4:238–244, 1994.