

Original Article

European Semi-anthropomorphic Spine Phantom for the Calibration of Bone Densitometers: Assessment of Precision, Stability and Accuracy The European Quantitation of Osteoporosis Study Group*

J. Pearson¹, J. Dequeker², M. Henley¹, J. Bright¹, J. Reeve¹, W. Kalender³, A. M. Laval-Jeantet⁴, P. Rügsegger⁵, D. Felsenberg⁶, J. Adams⁷, J. C. Birkenhäger⁸, P. Braillon⁹, M. Diaz Curiel¹⁰, M. Fischer¹¹, F. Galan¹², P. Geusens¹³, L. Hyldstrup¹⁴, P. Jaeger¹⁵, R. Jonson¹⁶, J. Kalef-Ezras¹⁷, P. Kotzki¹⁸, H. Kröger¹⁹, A. van Lingen²⁰, S. Nilsson²¹, M. Osteaux²², R. Perez Cano²³, D. M. Reid²⁴, C. Reiners²⁵, C. Ribot²⁶, P. Schneider²⁷, D. O. Slosman²⁸ and G. Wittenberg²⁷

¹MRC Clinical Research Centre and Northwick Park Hospital, Harrow, UK; ²Arthritic and Metabolic Bone Disease Research Unit, K.U. Leuven, Pellenberg, Belgium; ³Siemens Medical Systems, Erlangen, Germany; ⁴INSERM 18, Paris, France; ⁵Universität and ETH Zurich, Institut für Biomedizinische Technik, Zurich, Switzerland; ⁶Freie Universität Berlin, Klinik Steglitz, Berlin, Germany; ⁷University of Manchester, Department of Diagnostic Radiology, Manchester, UK; ⁸University Hospital Rotterdam, Department of Medicine III, Rotterdam, The Netherlands; ⁹Centre d'Absorptiometrie Osseuse, Hôpital E. Herriot, Lyon, France; ¹⁰Fundacion Jimenez Diaz, Madrid, Spain; ¹¹Kassel Städtische Kliniken, Zentrum für Radiologie, Kassel, Germany; ¹²Departamento de Medicina, Hospital Universitario, Seville, Spain; ¹³Arthritic and Metabolic Bone Disease Research Unit, K.U. Leuven, Pellenberg, Belgium; ¹⁴Hvidovre Hospital, Department of Endocrinology, Hvidovre, Denmark; ¹⁵University of Bern, Department of Diagnostic Radiology, Bern, Switzerland; ¹⁶Radiation Physics Section, Kliniskt Fysiologiska avd. Norra Alvsborgs Länssjukhus, Trollhattan, Sweden; ¹⁷Medical Physics Laboratory, University of Ioannina, Greece; ¹⁸Université de Montpellier, Hôpital Lapeyronie, Montpellier, France; ¹⁹Kuopio University Central Hospital, Department of Surgery, Kuopio, Finland; ²⁰Academisch Ziekenhuis Vrije Universitat Amsterdam, Department of Endocrinology, Amsterdam, The Netherlands; ²¹University of Uppsala, Department of Diagnostic Radiology, Uppsala, Sweden; ²²Akademisch Ziekenhuis, Vrije Universitat Brussels, Department of Radiology, Brussels, Belgium; ²³University Hospital 'Virgin Macaren', Departamento de Medicina, Seville, Spain; ²⁴Grampian Health Board, Geriatric and Specialist Services, Aberdeen, UK; ²⁵Universitätsklinikum Essen, Radiologisches Zentrum, Essen, Germany; ²⁶C.H.U. Purpan Service d'Endocrinologie, U.F. Maladies Osseuses et Metaboliques, Toulouse, France; ²⁷Klinik und Poliklinik für Nuklearmedizin, Würzburg, Germany; ²⁸University Hospital Geneva, Division of Clinical Pathophysiology, Geneva, Switzerland

Abstract. Up to now it has not been possible to reliably cross-calibrate dual-energy X-ray absorptiometry (DXA) densitometry equipment made by different manufacturers so that a measurement made on an individual subject can be expressed in the units used with a different type of machine. Manufacturers have adopted various procedures for edge detection and calibration, producing various normal ranges which are specific to each individual manufacturer's brand of machine. In this study we have used the recently

described European Spine Phantom (ESP, prototype version), which contains three semi-anthropomorphic "vertebrae" of different densities made of simulated cortical and trabecular bone, to calibrate a range of DXA densitometers and quantitative computed tomography (QCT) equipment used in the measurement of trabecular bone density of the lumbar vertebrae. Three brands of QCT equipment and three brands of DXA equipment were assessed. Repeat measurements were made to assess machine stability. With the large majority of machines which proved stable, mean values were obtained for the measured low, medium and high density vertebrae respectively. In the case of the QCT equipment these means were for the trabecular bone density, and in the case of the DXA equipment for vertebral body bone density in the posteroanterior projection. All DXA machines overestimated the pro-

* A Concerted Action of the European Community's COMAC-BME programme 1989–92. For further details of the Study Group see Appendix.

Correspondence and offprint requests to: J. Reeve, University Department of Medicine, Box 157 Addenbrooke's Hospital, Cambridge CB2 2QQ, UK. Fax: +44 (1223) 330105.

jected area of the vertebral bodies by incorporating variable amounts of transverse process. In general, the QCT equipment gave measured values which were close to the specified values for trabecular density, but there were substantial differences from the specified values in the results provided by the three DXA brands. For the QCT and Norland DXA machines (posteroanterior view), the relationships between specified densities and observed densities were found to be linear, whereas for the other DXA equipment (posteroanterior view), slightly curvilinear, exponential fits were found to be necessary to fit the plots of observed versus specified densities. From these plots, individual calibration equations were derived for each machine studied. For optimal cross-calibration, it was found to be necessary to use an individual calibration equation for each machine. This study has shown that it is possible to cross-calibrate DXA as well as QCT equipment for the measurement of axial bone density. This will be of considerable benefit for large-scale epidemiological studies as well as for multi-site clinical studies depending on bone densitometry.

Keywords: Cross-calibration; Dual X-ray absorptiometry (DXA); European Spine Phantom (ESP); Osteoporosis; Quantitative computed tomography (QCT)

Introduction

Over the last few years, the advent of equipment employing X-rays of two energies (dual-energy X-ray Absorptiometry, DXA) has replaced the previous generation of dual photon absorptiometry using isotope sources [1–5]. Meanwhile, other approaches to bone mineral measurement, including quantitative computed tomography (QCT), have continued to flourish. The aim of the Concerted Action 'Quantitative Assessment of Osteoporosis' was to achieve reliable cross-calibration and quality assurance in bone densitometry with DXA, QCT and other types of bone densitometry equipment in current use. Two recent studies have addressed the question of quality assurance in DXA densitometry [6,7].

At the time this study was initiated it was not possible to reliably convert measurements made, for example, on the spine of an individual subject using one manufacturer's DXA machine and interpret the results in the context of the reference ranges and spread of pathological values seen in contemporary studies conducted with machines made by other manufacturers [8–12]. This has hindered the development of protocols for multi-centre studies. In this report we described the use of a semi-anthropomorphic spine phantom in the assessment of the equipment of various manufacturers. Accuracy was assessed in comparison with the phantom, which was built to a precise specification. An

approach to calibrating machines using the phantom is described. At the present time, one other systematic attempt has been made at intercalibrating DXA densitometers made by different manufacturers [13] and a preliminary report has appeared of another [14]. Several reports have systematically examined the factors influencing the results obtained with QCT [15–20]. Our purpose was to explore the potential for cross-calibration of machines in routine clinical use with a variety of measurement protocols which for various reasons could not be altered to achieve better cross-calibration.

Methods

Phantom

The semi-anthropomorphic spine phantom used in this study was manufactured by QRM, Erlangen, Germany, and has already been described in detail [21]. Briefly it was made of several phases of epoxyresin, one calculated to have the attenuation coefficient and hence the apparent density of water (solid water), the others calculated to have the density of compact (cortical) bone (400 mg/cm³ added to the plastic material) and three densities of cancellous bone (50, 100 and 200 mg/cm³). These were achieved by adding varying quantities of powdered hydroxyapatite [21]. The shape of the phantom was designed for relative ease of manufacture so as to contain costs and was not perfectly anthropomorphic. It is illustrated in Figs 1 and 2. Within each phantom there were three simulated 'vertebrae'. Table 1 gives the specified densities of these three vertebrae.

Nineteen centres participated in measurements of the European Spine Phantom (ESP) using DXA equipment and nine centres made measurements using QCT equipment, a total of 25 phantoms being used in the present study. A finalized version of the ESP has recently been produced, so the ESP version used here is properly referred to as the ESP prototype, being identical with the specification described by Kalender [21]. For brevity, in what follows the ESP prototype is referred to as the ESP.

Table 1. Densities and areas for each vertebra of the ESP as specified to the manufacturer

Variable	Vertebra		
	Low	Medium	High
BMA/BMD (g/cm ²)	0.5	1.0	1.5
Trabecular density (mg/cm ³)	50	100	200
Projected area of vertebral body, AP view (cm ²)	9.0	9.0	9.0

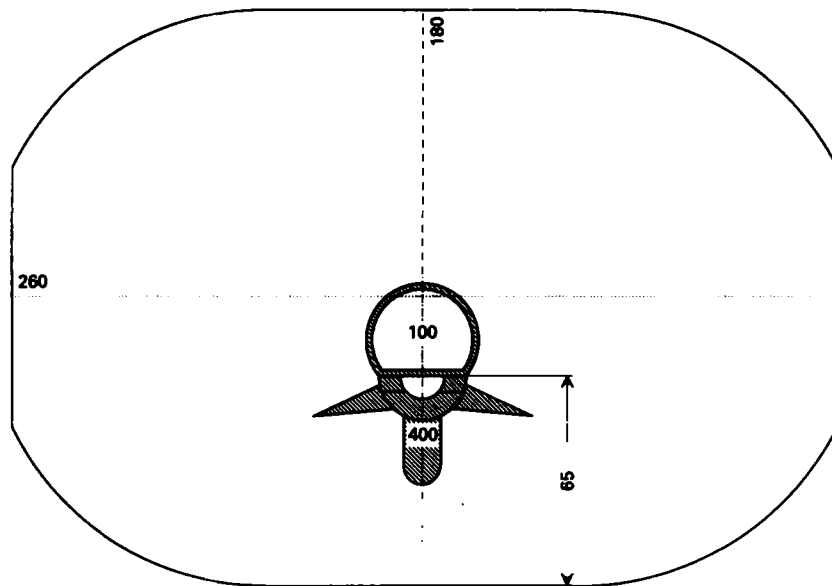


Fig. 1a-c. Drawings of one of the three phantom cross-sections with a vertebral insert of medium bone mineral density. (Compare with Table 1.)

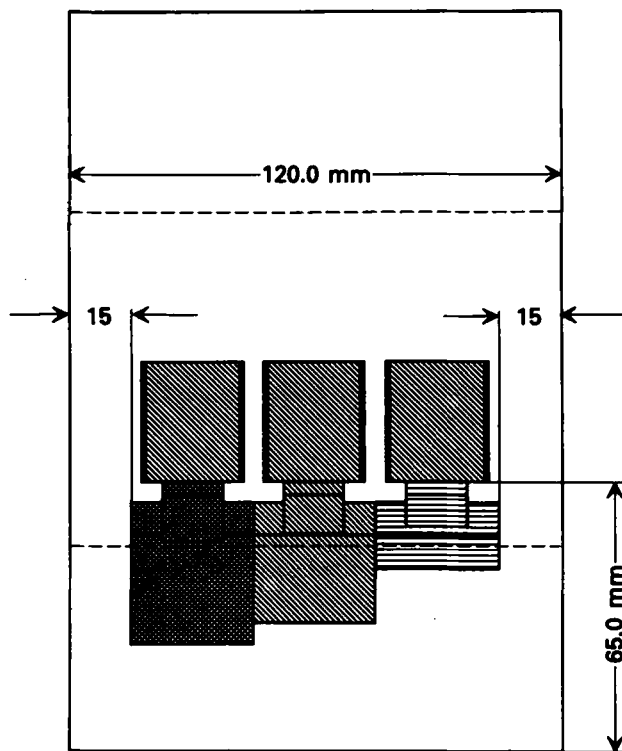


Fig. 2a,b. Lateral projection of the phantom.

Procedures for Measuring ESP by DXA (Table 2)

All scanning procedures were standardized between centres for each brand of machine. In the case of Hologic machines, the ESP was scanned in the postero-

anterior (PA) mode. The phantom was first located using the ESP carrying case as a spacer between the edge of the table and the ESP. The starting point was identified as some 5 mm "inferior" to the lower end plate of the densest "vertebra" (designated 'L4') and the region of interest box size was chosen to be 119×96 . The intervertebral spaces were marked by the operator. No bone edge modification was allowed. For the Lunar machine the spine phantom was scanned as though it was a patient, beginning and ending the scan within the phantom. The scan path was never allowed to extend into air. The analysis required that the intervertebral spaces were eliminated and this was done by bringing together the edge marks on the four scan lines of the two intervertebral regions using the "profiles option" in the autoanalysis software. With the Norland machine, which "expected" to see at least five vertebrae with its autoanalysis software, it was necessary to scan each vertebra separately and this was done with the phantom placed approximately centrally on the scanner table with the low-density vertebra towards the head. The standard posteroanterior (PA) spine scan was then performed in patient mode, but including only a single vertebra beginning with the low-density one. The scan width was set at the standard 12 cm. The diagram on the top of the phantom was used to mark start and end points about 2 mm above and below the vertebra of interest. Each region of interest was then 3.1–3.3 cm long. Results are presented for bone mineral content (BMC), projected area and bone mineral areal mass (BMA) [22]. The acronym BMA was proposed to avoid confusion with true densities expressed in g/cm^3 , because BMA values are expressed in g/cm^2 . Nevertheless many investigators, including the International DXA Standards Committee, hold to the BMD terminology, and to avoid misunderstandings the results

Table 2. DXA densitometers and their software used in the study

Densitometer brand and type	No. of centres using each machine	Software version	No. of centres using each software version
Hologic QDR-1000	(2)	6.1/4.47	(2)
Hologic QDR-1000W	(6)	6.1/4.47	(4)
		6.1/4.46	(2)
Hologic QDR-2000	(1)	7.10	(1)
Lunar DPX-L	(8)	1.1	(3)
		1.2	(4)
		1.3	(1)
Lunar DPX	(2)	3.1	(1)
		3.4	(1)
Norland XR26	(3)	2.2.4	(2)
		2.2.1	(1)

Table 3. QCT whole-body tomographic scanners and their software used in the study

Scanner type (no. of centres using each)	Single/dual energies (kVp)	Shape of region of interest ^a	Software version	Slice thickness (mm)
IGE 9000 (1)	120	C	11.02	5
IGE 9800 (1)	80	O	QC	10
IGE Highlight Advantage (1)	85	C	XA	10
Siemens Somatom Plus (4)	80	P	VD 30 (2)	8
			VD 1A (1)	
			VD 2B (1)	
Siemens Somatom DR (1)	80/125	C	VB 2	8
Siemens Somatom DRH (1)	85/125	P	C2	8
Elscont Elite Plus	120	O	3DST 2	10

^a C, circular; O, oval; P, Pacman.

obtained with DXA in the units g/cm^2 are expressed as 'BMA/BMD' (g/cm^2) in this paper.

Procedures for Measuring ESP by QCT (Table 3)

For whole-body QCT machines (Siemens, IGE and Elscint) the ESP was positioned on top of the bone calibrating system of the appliance centrally and parallel to the longitudinal axis of the table. In the case of the Siemens, this system consisted of the special mat with the calibrating phantom, and for the other equipment a special calibrating phantom (e.g. Cann and Genant, Image Analysis or Siemens phantom) was used as in clinical studies and the system set to the standard measurement parameters. Measurements were performed with low tube voltage (80–96 kV) or using dual-energy mode (voltages approximately 80 and 140 kV). In the evaluation of measurements with these computed tomography machines the methods corresponded to those used for patients. Where possible automatic procedures were employed and the trabecular density values of the three vertebrae obtained using regions of interest between 99 mm^2 and 900 mm^2 in area. Different centres used different-shaped regions of interest (circular, oval and "Pacman") according to their normal patient procedures.

Quality Control of Phantoms

Each phantom was measured three times on a DXA Hologic machine and a QCT Siemens machine before distribution to the participating centres. This allowed adjustment for differences between phantoms measured on the DXA machines.

Stability Analysis

In what follows, "density" refers to trabecular density (QCT) or BMA/BMD (g/cm^2) as appropriate. Each phantom was measured daily for 1 week and once weekly thereafter. The data for the first week's measurements and the subsequent weekly measurements were plotted against time for each separate vertebra on the ESP. Linear regression was used to determine the rate of change of density with time for each vertebra on each machine. Where changes in the X-ray tube occurred during the phantom measurement period, separate regression lines were fitted for each source. Regression was not performed for machines with fewer than 10 weeks of phantom measurements. For each machine which had a significant ($p < 0.01$) change in density with time, the change in density estimated to occur over a period of 13 weeks was calculated and expressed as a percentage of the mean density.

Comparison of Machines

The mean was calculated for each vertebra for each machine. Whenever a source change was made the mean was calculated separately for each source. This figure was used to determine the accuracy and precision of each machine's phantom measurement. Accuracy was defined (arbitrarily) as the ability to produce results that, on average, matched the specified values of the parameter of interest in the phantom, as given in Table 1. Precision was defined by the degree to which the machine obtained the same results in repeated measurements of the phantom, and was expressed as a standard deviation of the phantom measurements after calibration. The resulting standard deviations were then in the same scale of measurement for all machines.

Calibration

A cross-calibration formula was determined for each machine at each centre. Whenever a source change occurred during a measurement period a cross-calibration formula was determined for each source separately. For all machines with a significant drift in density with time and more than an arbitrary 3% change over 13 weeks, the calibration formulae were determined separately for each 7-week period. Simple linear regression was used to determine the following formula for each measurement of the phantom on each machine:

$$\text{Observed measurement} = A + B \times \text{Specified density} + \epsilon_i \quad (1)$$

where A and B are the regression constant and coefficient respectively and ϵ_i represents the statistical uncertainty of the estimate.

The values of specified densities used in the analysis are given in Table 1. The regression lines for each machine were summarized by taking the means of the coefficients A and B . Resulting estimated values of A and B were referred to as the intercept (regression constant) and the slope (regression coefficient) of the regression line.

The fit of this regression line was assessed by looking at the distribution of the residuals and the variability of the residuals. The residual was defined as the difference between the observed phantom measurement and the estimated density from the regression line. This analysis

assumed that these residuals were normally distributed and that the variation of residuals was constant with changing specified density. Histograms and normal probability plots were used to check the assumption of normality and plots of the residuals against the specified density were used to check the constant variability in the residuals.

The regression lines were converted into a cross-calibration formula by rearranging them, so that standardized (specified) densities could be predicted from observed densities.

Results

Table 4 summarizes the measurements of the phantom before distribution to the centres. The standard deviations (SDs) of trabecular densities between phantoms are of the same order of magnitude as those obtained from repeated measurements of the same phantom. For DXA measurements, the SDs between phantoms were larger than those obtained for repeated measurements on the same phantom, showing that some differences between the phantoms were measurable.

Thirty-nine machines measured the ESPs a total of 1289 complete measurements being made, 860 and 429 by DXA and QCT machines respectively.

Figures 3 and 4 show representative stability plots for DXA Hologic and QCT Siemens machines. However, there was evidence, with a minority of machines, of statistically significant drift with time affecting two DXA and three QCT machines. Three of these showed a drift of more than 3% in any vertebra over a 3-month period.

Table 5 contains the averages of the mean density for each manufacturer, weighted for numbers of measurements. These show differences between the machines of different manufacturers. The SDs in this table refer to the spread of results obtained between mean values from machines of a single manufacturer when used to measure the ESP.

Of the DXA machines, Hologic and Norland underestimate the specified densities for the high-density vertebra, Lunar giving on average nearly the specified value of 1.50. However, the Lunar machine overestimated the low- and medium-density vertebrae.

Table 4. Mean (and SD) between phantoms, measured using a DXA Hologic and a QCT Siemens machine before distribution to participating centres (n = 25)

Machine	Vertebra		
	Low	Medium	High
DXA BMA/BMD (g/cm ²)	0.46 (0.010)	0.88 (0.017)	1.22 (0.025)
QCT trabecular density (mg/cm ³)	50 (3.6)	99 (3.1)	202 (3.0)

Note: The standard deviation of the three measurements for each phantom typically was 0.005 g/cm² and 3.0 mg/cm³ for DXA and QCT machines respectively.

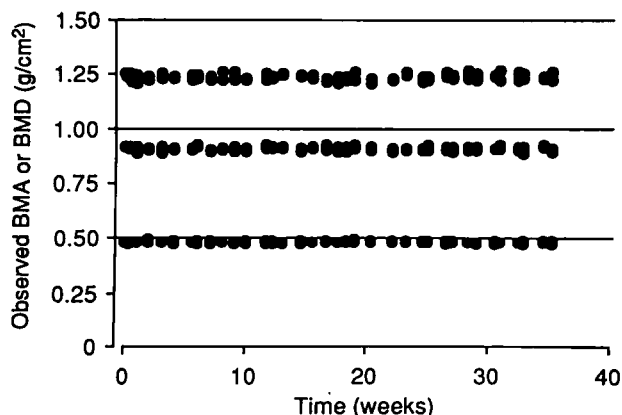


Fig. 3. Stability plot of BMA of the ESP measured by a DXA Hologic machine. The horizontal lines indicate the specified BMA for each of the vertebra in the ESP. In this and subsequent figures the two current, alternative terminologies (BMA and BMD) for DXA derived quantities in g/cm² are both represented since they are equivalent.

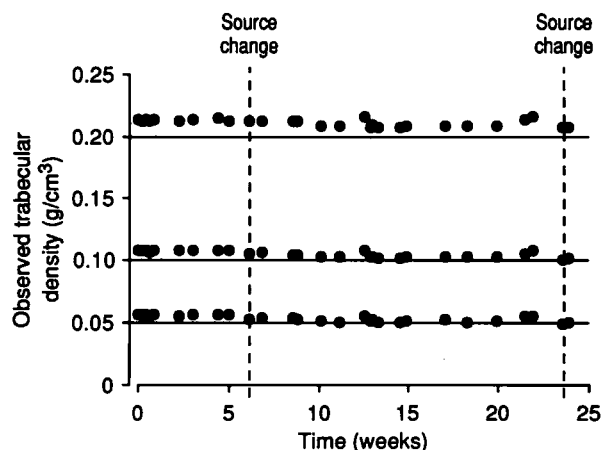


Fig. 4. Stability plot of trabecular densities of the ESP measured by a QCT Siemens machine. The horizontal lines indicate the specified trabecular density for each vertebra.

Except for the low-density vertebrae, the biggest difference between the observed densities and the specified densities occurs with the Hologic machines. Relative to the mean values for the Lunar machines, the Hologic and Norland machines gave results for the low-density vertebrae which were 0.07 g/cm² and 0.12 g/cm² lower respectively. For the medium-density vertebrae, the equivalent figures were 0.19 g/cm² and 0.18 g/cm² lower respectively. For the high-density vertebrae the equivalent figures were 0.25 g/cm² and 0.07 g/cm² lower respectively. For most of these instruments, the area detected was larger than the specified 9 cm² for each vertebral body (Table 5). For all vertebrae, a variable portion of the proximal part of the transverse processes was being measured in addition to the vertebral body. These processes have a lower density than the vertebral body, thus reducing the observed densities compared with the specified densities. Also parts of the intervertebral spaces were usually included. The projected

Table 5. Weighted mean (and SD) between machines for measurements of each vertebra in the ESP

Machine	No. of machines ^a	Vertebra		
		Low	Medium	High
<i>DXA BMA/BMD (g/cm²)</i>				
Hologic	11	0.47 (0.013)	0.90 (0.012)	1.24 (0.010)
Lunar	12	0.54 (0.010)	1.09 (0.016)	1.49 (0.030)
Norland	4	0.42 (0.002)	0.91 (0.012)	1.42 (0.007)
<i>DXA area (cm²)</i>				
Hologic	11	9.36 (0.14)	10.57 (0.14)	11.95 (0.22)
Lunar	12	8.06 (0.41)	9.52 (0.21)	10.49 (0.38)
Norland	4	11.37 (0.26)	10.47 (0.02)	10.54 (0.04)
<i>QCT trabecular density (mg/cm³)</i>				
IGE	4	49 (2.4)	96 (1.8)	191 (2.1)
Elscont	1	40 (-)	86 (-)	184 (-)
Siemens	6	50 (1.6)	98 (3.0)	199 (6.3)

^a'Machines' refers to a single machine to which a single period of stability analysis was applied. When, due to mechanical changes (e.g. X-ray tube) stability was assessed separately over two periods of time on a single machine, that machine counted twice in these analyses.

area of the vertebral bodies plus half an intervertebral space on either side was 10.4 cm².

The Siemens QCT machine produced measurements close to the specified densities, while the IGE and Elscint machines slightly underestimated the specified densities (Table 5).

Table 6 shows that all DXA machines measuring the ESP, the SD is smaller for the low-density vertebra, indicating that this is measured more precisely in terms of units of measurement than the high-density vertebra. Relative SDs (i.e. SDs divided by means) were more similar between the three vertebrae with these machines than SDs. Results obtained for QCT measurements of the ESP showed rather similar SDs for the three vertebrae in g/cm³.

In the calibration of DXA measurements of the ESP, the analysis of residuals showed that for the Hologic and Lunar machines linear regression did not explain the whole relationship between the observed measurements and the specified densities. Essentially, this was because the relationship between the observed measurements and the specified densities was curved. Typical plots of the two types of data and the fitted regression lines are shown in Figs 5 and 6. The fitted regression lines underestimated the observed measurements of the medium-density vertebra by an average of 3.2% and 4.5% for the Hologic and Lunar machines respectively. There was no problem with the fit of the linear regression for the Norland DXA machines or for any of the QCT machines.

For the DXA machines an alternative non-linear fit was therefore tried, with the aim of ensuring that at a specified density of zero, the machines would read zero as assumed in the built-in software. It was found that the data can be fitted by an exponential curve with the formula

Table 6. Precision of calibrated ESP measurements. Values are expressed as the mean (range) for all machines of each brand; means were weighted for the number of results included from each machine

Machine	No. of machines ^a	Vertebra		
		Low	Medium	High
<i>DXA BMA/BMD (g/cm²)</i>				
Hologic	11	0.006 (0.002–0.008)	0.012 (0.007–0.019)	0.026 (0.009–0.062)
Lunar	12	0.014 (0.009–0.021)	0.014 (0.005–0.020)	0.037 (0.009–0.065)
Norland	4	0.014 (0.012–0.016)	0.021 (0.013–0.024)	0.029 (0.017–0.033)
<i>QCT trabecular density (mg/cm³)</i>				
IGE	4	1.1 (0.0–0.16)	1.3 (1.1–1.7)	1.5 (0.7–1.9)
Elscont	1	1.0 (–)	1.2 (–)	1.8 (–)
Siemens	6	2.0 (0.7–3.1)	2.2 (0.5–3.7)	2.9 (0.8–5.0)

^a 'Machines' refers to a single machine to which a single period of stability analysis was applied. When, due to mechanical changes (e.g. X-ray tube) stability was assessed separately over two periods of time on a single machine, that machine counted twice in these analyses.

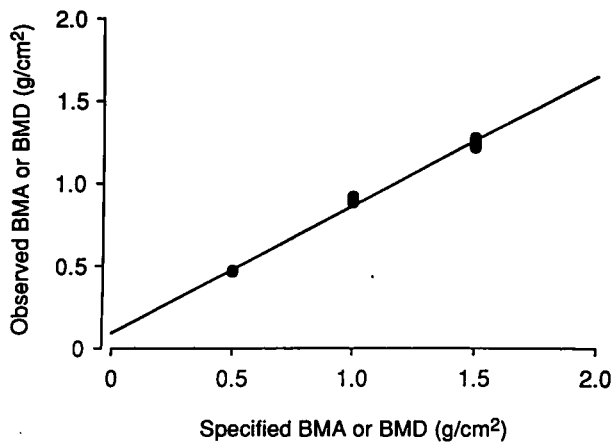


Fig. 5. Bone mineral densities of the ESP as measured by a typical Hologic machine, plotted against the specified BMA and the linear regression line.

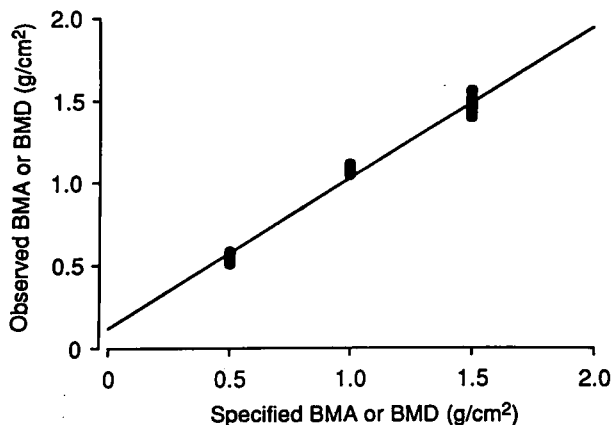


Fig. 6. Bone mineral densities of the ESP as measured by a typical Lunar machine, plotted against the specified BMA and the linear regression line.

Observed measurement =

$$\alpha \times [1 - \exp(-\beta \times \text{Specified density})] + \epsilon_i \quad (2)$$

where α is the asymptote and $\alpha \times \beta$ is the slope when the specified density is small. This curve-fitting formula was found to provide a better fit than the linear regression formula with all but six of the DXA machines, and was adopted except where otherwise stated. When the exponential fit was employed the standardized density was obtained by rearranging Eq. 2 to give:

Specified density =

$$\delta \times \ln(\alpha) - \delta \times \ln(\alpha - \text{Observed measurement}) \quad (3)$$

where $\delta = 1/\beta$ and \ln is the natural logarithm.

For six DXA machines the parameter estimates of this curve would not converge, due to the linearity of the relationship, in which case a straight line was fitted through the origin.

Example regression lines and data are shown in Figs 7 and 8 for selected DXA and QCT machines measuring the ESP. Weighted averages of the regression coefficients are shown in Tables 7, 8a and 8b. These show differences in the regression lines between the manu-

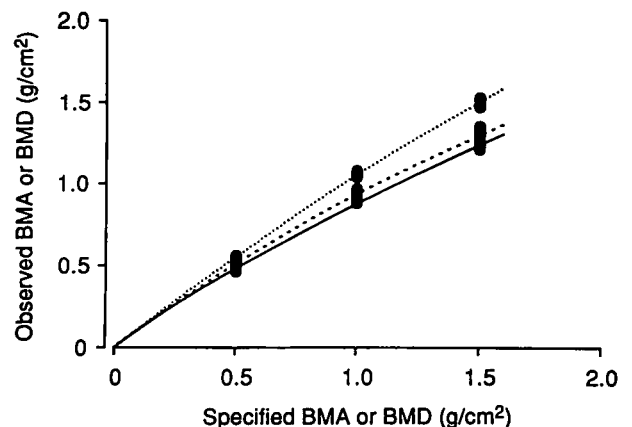


Fig. 7. Typical BMA measurements of the ESP using three DXA machines and their non-linear regression lines: *continuous line*, Hologic; *dashed line*, Lunar; *dotted line*, Norland.

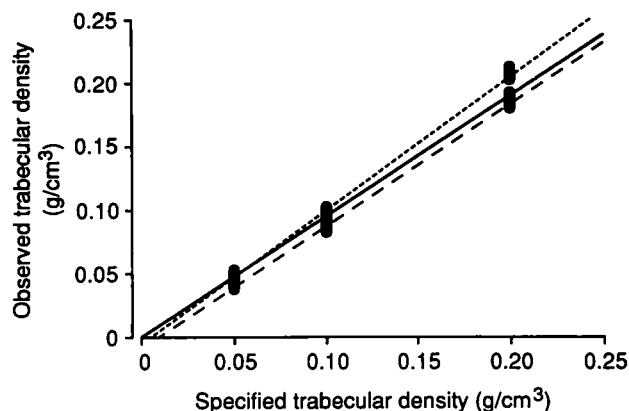


Fig. 8. Typical BMA measurements of trabecular density of the ESP using three QCT machines and their linear regression lines: dotted line, Siemens; continuous line, IGE; dashed line, Elscint.

Table 7. Weighted linear regression lines for QCT measurements of ESP (units: mg/cm³)

Machines	No. of machines	Intercept (A)	SD*	Slope (B)	SD*
<i>ESP QCT</i>					
IGE 9000	4	1.9	3.9	0.9463	0.0274
Elscont	1	-8.9	—	0.9605	—
Siemens	6	-0.1	2.9	0.9934	0.0407

*The standard deviations are a measure of variability between machines.

Table 8a. Weighted linear regression lines for DXA measurements of ESP (only for machines whose linear regressions were optimal) (units: g/cm²)

Machines	No. of machines	Slope	SD*
<i>ESP DXA</i>			
Lunar	2	1.0572	0.0085
Norland	4	0.9317	0.0049

*The standard deviations are a measure of variability between machines.

Table 8b. Weighted curves for DXA measurements of ESP (units: g/cm²)

Machines	No. of machines	α	SD*	β	SD*
<i>ESP DXA</i>					
Hologic	11	3.5129	0.6006	0.3044	0.0674
Lunar	10	4.8923	1.3079	0.2628	0.0620

*The standard deviations are a measure of variability between machines.

facturers, which reflect the differences in the means described previously. For the ESP, these regression lines fit the data well. The average R^2 value for both DXA and QCT machines was 99.9%.

Discussion

Over the last 10 years bone densitometry has become widely available for measurements on the axial skeleton [1-5,23,24]. The skeleton is complex in its spatial arrangement. Therefore there has been great scope for individual companies supplying densitometry equipment to develop different approaches as to how measurements should be made. Although commercial DXA manufacturers all use the principle of having two X-ray beams which simulate the monoenergetic emissions of the previous generation of dual photon absorptiometers employing ¹⁵³Gd, this is achieved in different ways. Manufacturers make different allowances in their software for the presence of fat in the axial skeleton, they use different edge detection algorithms and they have different procedures for identifying areas of interest. So when an individual subject is measured by machines supplied by different manufacturers, the results show considerable differences.

The problems of comparing equipment manufactured for the analysis of cancellous bone densities using QCT have been somewhat less intractable. However, the equipment supplied by some manufacturers, but not others, has in particular been sensitive to variations in the table height employed [25]. Both QCT and DXA measurements have therefore suffered up to now by a lack of adequate standardization [26].

Morita et al. [13] devised a geometrically simple phantom for use with DXA machines, containing cylinders of three diameters holding five different concentrations of hydroxyapatite, and sent 20 of these phantoms to different institutes throughout Japan. They were able to devise interconversion equations for relating BMA/BMD (g/cm²) measured on one machine with BMA/BMD measured on another, and used these equations to generate an age distribution of lumbar BMA/BMD for a large number of normal women. Like us, they found that it was necessary to standardize individual machines, finding measureable differences in the results obtained by machines of the same brand and model type.

Important components of DXA software are the edge detection algorithms. In the automatic delineation of 'bone' from 'not bone', such algorithms begin with a three-dimensional array in which each pixel is associated with three numbers, two of which refer to its position in the AP projection while the third is the calculated mass of bone in the pixel's volume. There are many possible ways of differentiating 'bone' from 'soft tissue' and wide scope for obtaining different results due to variations in the *a priori* criteria adopted, or to the complexities of the surface topography of the bone. This has been studied previously [27]. Some commercial

phantoms are designed to match closely the surface contours of one particular human original. If this led to bone edge detection which not only differed between machines but which differed idiosyncratically in a way that was dependent on the unique anatomical contours of the phantom, this could be a disadvantage. For this reason a semi-anthropomorphic phantom with smooth surfaces as well as edges that were less irregular than those seen *in vivo* seemed worth investigating.

An important feature of the ESP is that it can be used on both QCT and DXA machines. Also, it was designed to span about 90%–95% of the range of bone density values found in human subjects, including those with osteoporosis. This phantom has allowed us to assess the degree to which deviations from mathematical linearity with respect to densities specified for the phantom need to be corrected for with the machines of different manufacturers. The slightly curvilinear calibration equations with all Hologic and most Lunar machines do not directly imply that in a hypothetical dissection study, performed with a sufficient number of human cadavers, a non-linear relationship would be found between measured density and ash weight. A study of this nature of sufficient numerical size would probably be impractical. Since differences in curvature between individual machines are of only moderate statistical significance, it is predictable that the results obtained in a series of subjects measured on two different machines on the same day would fit a linear rather than a curvilinear regression [14]. Furthermore, since in other studies with different test material all DXA manufacturers found their machines to respond linearly to increased density in g/cm^2 , it is likely that these slightly curvilinear relationships relate to some property of the ESP which may or may not be associated with its anthropomorphic properties. The practical issue is that by using curves instead of straight lines, slightly better cross-calibration equations will be obtained with the ESP prototype.

The results shown in Table 5 demonstrate that all the DXA equipment under review, except that made by Lunar, underestimated specified densities when measuring the ESP. One reason for this is that Lunar have built into their software a correction for fat content. *In vivo* fat reduces the bone density measured by DXA equipment and Lunar adjusts for this arbitrarily by assuming that soft tissues contain 25% fat. These results reinforced the expectation of large differences between machines from individual manufacturers of DXA equipment. Also, standardized density values obtained after cross-calibration are not adjusted for fat, so that our approach to cross-calibration does not include any correction for the fat content of the marrow, which has been shown to lower measurement values obtained with QCT and DXA. Another finding (illustrated in Table 5) is that differences between machines of the same manufacturer may be expected with respect to both DXA and QCT equipment – which could be important,

particularly in longitudinal studies, if subjects were considered for transfer from one machine to another. The implication is that great care must be taken when transferring longitudinal studies between densitometers, even when they are being transferred to another manufactured by the same company. Either the results should be standardized with a phantom such as the ESP, or the second densitometer should be tuned very carefully to the first so that it gives similar results with a phantom such as the ESP and some representative human subjects [28].

An alternative approach to the cross-calibration of bone densitometers has been adopted by Genant et al. [14]. They made triplicate measurements with three brands of DXA machines on over 100 human subjects and used regression analysis to derive equations for interconversion between pairs of densitometers. The scatter about each of the two regression lines for the spine was moderate but higher for the measurement sites in the femur, leaving some statistical uncertainties associated with the regression constants and coefficients obtained (which were for practical purposes quite impressively small for the spine). Nevertheless these were higher than the uncertainties associated with the regressions we fitted to our phantom measurements for individual densitometers. The physical reasons underlying scatter such as that observed by Genant et al. have been investigated by Mazess et al. [27] and to a substantial extent reflect the alternative choices of a modified fixed threshold by some manufacturers (e.g. Hologic) and a derivatized thresholding procedure by others (e.g. Lunar). Hologic machines tend to generate larger areas of interest, particularly in patients with relatively dense transverse processes. Parts of these may become included in Hologic areas of interest.

Derivatization techniques invariably are more sensitive to random uncertainties (e.g. associated with non-ideal photon flux) and may therefore tend to give less precise estimates of area. Both these tendencies are evident from the results shown in Table 5. An important consideration with generalizing Genant et al.'s [14] approach to the problem of routine cross-calibration is that, as we have shown, individual machines give varying results even if they are made by the same manufacturer, due to differences in setting up or local measurement procedures. Therefore with a cross-calibration based initially on many human studies a secondary cross-calibration (e.g. between each Hologic, each Lunar or each Norland machine in a study) with a phantom such as the ESP covering the full range of expected bone densities may still be necessary to eliminate artifactual between-centre differences.

With regard to stability, there was one DXA machine and two QCT machines which showed a drift with time of more than 3% over a 3-month period in the results obtained with the ESP. This must have compromised the value of the clinical results obtained from these machines over the same time span. The value of using

phantoms such as the ESP for documenting stability cannot be overemphasized. Currently this is normally part of routine procedure with most dedicated densitometry systems, the importance of which became obvious when early work with radioisotopic sources (dual photon absorptiometry) showed problems which were attributable to declining source strength or contamination by other radioisotopes with different energies and half-lives [1]. Our results with a representative group of equipment from different manufacturers in everyday use give no grounds for complacency and emphasize that the current generation of equipment, while of improved stability, requires regular monitoring with an appropriate phantom to guarantee stability if the results are to be acceptable.

Given that the measurements procedure is rigorously standardized there are many other factors which can affect the precision with which a phantom is measured. These include both photon flux and photon energy, particularly with respect to the low-energy beam in DXA systems, and repositioning, which in patient measurement of the proximal femur can have a substantial influence on the BMD results obtained [8]. Differences in rotation were minimized with the ESP because of its flat bottom and standardized repositioning procedure. DXA machines also offer choices to operators which potentially could affect the application of standardization procedures [29], but these were minimized by the adoption of uniform measurement procedures in close consultation with the manufacturers. In a future paper the measurements of in vivo precision will be compared with the measurements of precision of phantom measurements. This will help assess the potential of improved machine precision for further improving in vivo precision.

In conclusion, we have developed and applied a semi-anthropomorphic phantom for use in quantitative axial bone densitometry. This allows the calculation of bone density values which can be related to the density specified in the phantom. Calibration curves have been derived for each machine in the study, which has allowed us to adopt a cross-calibration procedure in which 'standardized density' values are derived from individual patient measurements. In order to minimize the error in the calibrated measurements it is not sufficient to use the weighted regression line for each manufacturer to calibrate subject measurements; a regression line derived specifically for each machine must be used. Most of the machines in this study demonstrated acceptable stability and high levels of precision. The results contained in this paper have been applied in clinical studies of representative populations of normal subjects in a number of European countries to generate European normal ranges expressed in terms of specified densities [30]. This and other work documenting the results obtained in various patient groups will be presented elsewhere.

Appendix: The European Quantitation of Osteoporosis Study Group

Project Management Group

J. Dequeker (Project Coordinator), Leuven; J. Reeve, Harrow; J. Pearson, Harrow; D. Felsenberg, Berlin; W. Kalender, Erlangen; C. Langton, Sheffield; A. M. Laval-Jeantet, Paris; P. Ruegsegger, Zürich; G. Van der Perre, Leuven.

Statisticians

J. Pearson, Harrow; M. Henley, Harrow; J. Bright, Harrow.

Other Participating Centres

J. E. Adams, Manchester; J. C. Birkenhäger, Rotterdam; J-P. Bonjour, Geneva; P. Braillon, Lyon; M. Diaz Curiel, Madrid; M. Fischer, Kassel; F. Galan, Seville; C. Gennari, Siena; P. Geusens, Leuven; S. Nilsson and A. Hemmingsson, Uppsala; L. Hyldstrup, Hvidovre; P. Jaeger, Bern; R. Jonson, Trollhättan; J. Kalef-Ezra, Ioannina; P. Kotzki, Montpellier; H. Kröger, Kuopio; P. Lips and A. van Lingen, Amsterdam; M. Osteaux, Brussels; D. M. Reid, Aberdeen; C. Reiners, Essen; C. Ribot, Toulouse; P. Schneider and G. Wittenberg, Würzburg.

References

1. Kelly TL, Slovik DM, Neer RM. Calibration and standardization of bone mineral densitometers. *J Bone Miner Res* 1989;4:663-9.
2. Hochberg AM, Wahner HW, Dunn WL, Bevan J, Stein J. A performance comparison of x-ray and gamma-ray bone mineral analyzers utilizing a standard spine phantom. In: Dequeker J, Geusens P, Wahner HW, editors. *Bone mineral measurements by photon absorptiometry: methodological problems*. Leuven, Belgium: Leuven University Press, 1988: 242-5.
3. Kelly TL, Slovik DM, Schoenfeld DA, Neer RM. Quantitative digital radiography versus dual photon absorptiometry of the lumbar spine. *J Clin Endocrinol Metab* 1988;67:839-44.
4. Borders J, Kerr E, Sartoris DJ, Stein JA, Ramos E, Moscona AA, Resnick D. Quantitative dual-energy radiographic absorptiometry of the lumbar spine: in vivo comparison with dual-photon absorptiometry. *Radiology* 1989;170:129-31.
5. Wahner HW, Dunn WL, Brown ML, Morin RL, Riggs BL. Comparison of dual-energy x-ray absorptiometry and dual photon absorptiometry for bone mineral measurements of the lumbar spine. *Mayo Clin Proc* 1988;63:1075-84.
6. Orwoll ES, Oviatt SK, Biddle JA. Precision of dual-energy x-ray absorptiometry: development of quality control rules and their application in longitudinal studies. *J Bone Miner Res* 1993;8:693-9.
7. Glüer CC, Faulkner KG, Estilo MJ, et al. Quality assurance for bone densitometry research studies: concept and impact. *Osteoporosis Int* 1993;3:227-35.
8. Svendsen OL, Marslew U, Hassager C, Christiansen C. Measurement of bone mineral density of the proximal femur by two commercially available dual energy X-ray absorptiometric systems. *Eur J Nucl Med* 1992;19:41-6.
9. Pocock NA, Sambrook PN, Nguyen T, Kelly P, Freund J, Eisman

- JA. Assessment of spinal and femoral bone density by dual X-ray absorptiometry: comparisons of Lunar and Hologic instruments. *J Bone Miner Res* 1992;7:1081-4.
10. Lai KC, Goodsitt MM, Murano R, Chesnut CH. A comparison of two dual-energy X-ray absorptiometry systems for spinal bone mineral measurement. *Calcif Tissue Int* 1992;50:203-8.
 11. Laskey MA, Flaxman ME, Barber RW, Trafford S, Hayball MP, Lyttle KD, Crisp AJ, Compston JE. Comparative performance in vitro and in vivo of Lunar DPX and Hologic QDR-1000 dual energy x-ray absorptiometers. *Br J Radiol* 1991;64:1023-9.
 12. Laskey MA, Crisp AJ, Cole TJ, Compston JE. Comparison of the effect of different reference data on Lunar DPX and Hologic QDR-1000 dual energy X-ray absorptiometers. *Br J Radiol* 1992;65:1124-9.
 13. Morita R, Orimo H, Hamamoto I, Fukunaga M, Shiraki M, Nakamura T, et al. Some problems of dual energy X-ray absorptiometry in clinical use. *Osteoporosis Int* 1993;3(Suppl):S87-90.
 14. Genant HK, Grampp S, Glüer CC, Faulkner KG, Jergas M, Engelke K, et al. Universal standardization for dual X-ray absorptiometry: patient and phantom cross-calibration results. *J Bone Miner Res* 1994;9:1503-14.
 15. Goodsitt MM, Johnson RH, Chesnut CH. A new set of calibration standards for estimating the fat and mineral content of vertebrae via dual energy QCT. *Bone Miner* 1991;13:217-33.
 16. Steenbeek JCM, van Kuijk C, Grashuis JL. Influence of calibration materials in single and dual-energy quantitative CT. *Radiology* 1992;83:849-55.
 17. Faulkner KG, Glüer CC, Grampp S, Genant HK. Cross-calibration of liquid and solid QCT calibration standards: corrections to the UCSF normative data. *Osteoporosis Int* 1993;3:36-42.
 18. Healey J, Williams-Russo P, Szatrowski T, Schneider R, Paget S, Ales K, Schwarzberg P. Comparison of quantitative CT densitometry using liquid versus solid phantoms. *J Bone Miner Res* 1992;7(Suppl 1):S179.
 19. Suzuki S, Yamamuro T, Okumura H, Yamamoto I. Quantitative computed tomography: comparative study using different scanners with two calibration phantoms. *Br J Radiol* 1991; 64:1001-6.
 20. Whitehouse RW, Adams JE. Single energy quantitative computed tomography: the effects of phantom calibration material and kVp on QCT bone densitometry. *Br J Radiol* 1992;65:931-4.
 21. Kalender WA. A phantom for standardization and quality control in spinal bone mineral measurements by QCT and DXA: design considerations and specifications. *Med Phys* 1992;19:583-6.
 22. Jonson R. Mass attenuation coefficients, quantities and units for use in bone mineral determinations. *Osteoporosis Int* 1993;3:103-6.
 23. Genant HK, Faulkner KG, Glüer CC. Measurement of bone mineral density: current status. *Am J Med* 1991;91(Suppl 5B):49S-53S.
 24. Sandor T, Felsenberg D, Kalender WA, Clain A, Brown E. Compact and trabecular components of the spine using quantitative computed tomography. *Calcif Tissue Int* 1992;50:502-6.
 25. Steiger P, Block JE, Steiger S, Heuck AF, Friedlander A, Ettinger B, et al. Spinal bone mineral density measured with quantitative CT: effect of region of interest, vertebral level and technique. *Radiology* 1990;175:537-43.
 26. Kalender WA, Fischer M. Quality control and standardisation of absorptiometric and computed tomographic measurements of bone mineral mass and density. *Radiat Protection Dosimetry* 1993;49:229-33.
 27. Mazess RB, Trempe JA, Bisek JP, Hanson JA, Hans D. Calibration of dual-energy X-ray absorptiometry for bone density. *J Bone Miner Res* 1991;6:799-806.
 28. Faulkner KG, Glüer C-C, Estilo M, Genant HK. Cross-calibration of DXA equipment: upgrading from a Hologic QDR 1000W to a QDR 2000. *Calcif Tissue Int* 1993;52:79-84.
 29. Trevisan C, Gandolini GG, Sibilla P, Penotte M, Caraceni MP, Ortolani S. Bone mass measurements by DXA: influence of analysis procedures and interunit variation. *J Bone Miner Res* 1992;7:1373-82.
 30. Pearson J, Dequeker J, Reeve J, Felsenberg D, Henley M, Bright J, et al. Dual X-ray absorptiometry of the proximal femur: normal European values standardised with the European Spine Phantom. *J Bone Miner Res* (in press).

*Received for publication 24 November 1994
Accepted in revised form 19 September 1994*