

Cross-calibration of dual-energy X-ray densitometers for a large, multi-center genetic study of osteoporosis

D.M. Reid · I. Mackay · S. Wilkinson · C. Miller
D.G. Schuette · J. Compston · C. Cooper · E. Duncan
N. Galwey · R. Keen · B. Langdahl · A. McLellan
H. Pols · A. Uitterlinden · J. O’Riordan · J.A.H. Wass
S.H. Ralston · S.T. Bennett

Received: 7 December 2004 / Accepted: 18 April 2005 / Published online: 1 September 2005
© International Osteoporosis Foundation and National Osteoporosis Foundation 2005

Abstract Osteoporosis is a common disease with a strong genetic component characterized by reduced bone mass and an increased risk of fragility fractures. Bone mineral density (BMD) is the most important

determinant of osteoporotic fracture risk, but the genes responsible for BMD regulation and fracture are incompletely defined. To enable multi-center studies to examine the genetic influences on BMD there is a requirement to standardize measurements across different manufacturers of bone densitometers, different versions of machines and different normative ranges. This paper describes a method developed to allow near-identical subjects with low age-adjusted BMD (based on Z-scores) to be recruited in 17 centers using 27 different densitometers. Cross-calibration was based on measurements using a European spine phantom circulated to all centers and measured ten times on each individual machine. From these values an individual exponential curve, based on nominal versus observed BMD, was derived for each machine. As expected, there were large and significant variations in nominal BMD values, not only between scanners from different manufacturers but also between different versions of scanners from the same manufacturer. Hologic scanners tended to underestimate the nominal BMD, while Lunar scanners overestimated the value. Norland scanners gave mixed values over estimating BMD at the lower nominal value (0.5 g/cm²) while underestimating the value at the higher value (1.5 g/cm²). The validity of the exponential equations was tested using hip and spine measurements on 991 non-proband women from a familial osteoporosis study (FAMOS). After cross-calibration there was a considerable reduction in variation between machines. This observation, coupled with the absence of a similar reduction in variation attributable to a linear regression on age, demonstrated the validity of the cross-calibration approach. Use of the cross-calibration curves along with a standard normative range (in the case of this study, the Hologic normative range) allowed age-specific Z-scores to be used as an inclusion criterion in this genetic study, a method that will be useful for other trials where age-specific BMD inclusion criteria are required.

D.M. Reid (✉) · S. Wilkinson · S.H. Ralston
Department of Medicine and Therapeutics, Medical School,
University of Aberdeen, Foresterhill, Aberdeen, AB25 2ZD, UK
E-mail: d.m.reid@abdn.ac.uk
Tel.: +44-1224-551154
Fax: +44-1224-554615

I. Mackay · D.G. Schuette · N. Galwey · S.T. Bennett
Oxagen Ltd., Abingdon, Oxfordshire, UK

C. Miller
Bioimaging Inc., Lincolnshire, Illinois, USA

J. Compston
University of Cambridge School of Clinical Medicine, Cambridge,
UK

C. Cooper
MRC Environmental Epidemiology Unit, Southampton General
Hospital, Southampton, UK

E. Duncan · J.A.H. Wass
Department of Endocrinology, Oxford Centre for Diabetes,
Endocrinology and Metabolism, The Churchill Hospital, Oxford,
UK

R. Keen
Department of Medicine, University College, London, UK

B. Langdahl
Department of Endocrinology,
Århus Amtssygehus, Århus, Denmark

A. McLellan
Department of Medicine and Therapeutics, Western Infirmary,
Glasgow, UK

H. Pols · A. Uitterlinden
Department of Internal Medicine, Erasmus Medical Centre,
Rotterdam, The Netherlands

J. O’Riordan
Department of Endocrinology, Middlesex Hospital, London, UK

Keywords Dual-energy X-ray · Multi-center study · Osteoporosis

Introduction

Osteoporosis is a multi-factorial disease, determined by interplay between genetic and non-genetic factors. Genetic effects are subtle, and to elucidate these effects one requires large numbers of subjects in whom high quality clinical data are matched with a widespread genomic approach to disease or trait pathogenesis. We have set up the Family and Osteoporosis (FAMOS) study in eight European centers to ensure that there are adequate numbers of families to give statistical power to detect subtle genetic effects against a noisy non-genetic background. The primary aim of this study was to conduct a genome-wide scan for linkage to BMD as assessed by axial dual energy X-ray absorptiometry (DXA), a key component in determining the risk of future osteoporotic fracture.

Since their introduction in the late 1980s, DXA bone densitometers have been used extensively in clinical practice and epidemiological studies [1, 2]. Their clinical value in the management of osteoporosis has been reviewed recently [3]. A variety of different versions of these machines is now in common use, each with different operating characteristics. Our study, like previous large epidemiological and pharmaceutical studies, involves multiple centers and a range of different DXA densitometers. For genetic studies such as FAMOS, in which samples that have been collected across all centers will be combined, it is imperative that data are consistent across centers, and, therefore, cross-calibration of DXA devices is essential.

In previous studies various standardization techniques have been used, and cross-calibration equations have been derived from measurements either from healthy volunteers or from a variety of semi-anthropomorphic phantoms [3]. The European spine phantom (ESP) was developed to provide a standard that does not depend on the software algorithms of any particular manufacturer [4, 5]. The ESP provides three test densities (0.5 g/cm^2 , 1.0 g/cm^2 and 1.5 g/cm^2) and a semi-anthropomorphic shape. These characteristics allow non-linear calibration curves to be fitted, if required, but challenge the image analysis software of the manufacturers, which is a major and often inaccessible component of DXA machines.

The initial approach to cross-calibration using phantoms was challenged, and this led to an enhancement that minimized differences between the three major manufacturers, using only healthy women and producing cross-calibration equations for the lumbar spine and total hip [6]. Such an approach can be, and is, used to produce a standardized BMD and, hence, a standardized “T-score” for diagnostic and epidemiological purposes (e.g., osteoporosis = $T < -2.5$, i.e., 2.5 standard

deviations below the young normal mean) by use of a recognized normative range such as the National Health and Nutrition Examination Survey (NHANES) III database [7] for the total hip DXA site. The approach works reasonably well for T-score comparisons for the lumbar spine and total hip as well as for the femoral neck, Wards and trochanter regions [8].

In the FAMOS study we wished to recruit probands, not with a single cut-off value of BMD that could be assessed by a T-score, but to recruit on the basis of low age-matched BMD or Z-scores. This required not only production of careful cross-calibration equations but also the use of a standard normative range. The cross-calibration exercise reported here was used to cross-calibrate 27 DXA scanners located at 17 hospitals under the aegis of the eight collaborating centers comprising the FAMOS study. The target was to collect extended pedigrees from probands who had age-matched BMD values of 2 standard deviations below the age-matched mean (Z-score of -2.0 or less; lower 2.5% of the normal distribution) at the femoral neck and/or lumbar spine. Immediate relatives of these probands were also scanned, and if any of their BMD scores reached 1.28 standard deviations below the age-matched mean (Z-score of -1.28 or less; lower 10% of the normal distribution), their first-degree relatives were invited to participate, too. In that way it was possible to generate multi-generation pedigrees for further genetic analyses. Therefore, individual age-matched BMD values measured at each center had to include a close-to-identical proportion of the population to be selected at each center having a Z-score of -2 for probands and, for first-degree relatives, a Z-score of -1.28 . The approach we now report uses the ESP for the initial part of the cross-calibration, as described previously [9].

Methods

Seventeen centers were involved in this study, and measurements have been taken from 27 different DXA machines from three manufacturers (Norland, Cooper Surgical, Trumbull, Conn., USA; Lunar, GE Medical Systems, Madison, Wis., USA and Hologic Bedford, Mass., USA), with nine different versions of each scanner from the three manufacturers. Machines and their locations are given in Table 1. At the start of the study a single ESP prototype was used to calibrate all machines. The ESP was positioned on each DXA machine and scanned in air ten times, without being repositioned, using the machine-specific lumbar spine scanning protocol and analysis.

An analysis of differences between various versions of the same manufacturer was then done by treating machines of the same version as independent replicates, allowing data collected for calibration to be used also to study the nature of variation between these different versions. To this end, a linear mixed model was fitted,

Table 1 Hospitals and DXA scanners used in the FAMOS study (*DK* Denmark, *NL* The Netherlands)

Machine number	Center	Hospital	Manufacturer	Version
1	Aberdeen, UK	Aberdeen	Lunar	Expert
2	Aberdeen, UK	Aberdeen	Norland	XR26
3	Aberdeen, UK	Aberdeen	Norland	XR36
4	Århus, DK	Århus	Hologic	2000
5	Århus, DK	Århus	Hologic	4500
6	Århus, DK	Århus	Hologic	1000
7	Cambridge, UK	Cambridge	Lunar	Expert
8	Cambridge, UK	Cambridge	Hologic	4500
9	University College, London, UK	Eastbourne	Hologic	4500
10	University College, London, UK	Hertford	Hologic	4500
11	Århus, DK	Hvidovre	Hologic	4500
12	Rotterdam, NL	Ijsseland	Hologic	4500
13	Rotterdam, NL	Losser	Hologic	1000
14	Rotterdam, NL	Losser	Hologic	4500
15	Oxford, UK	Mount Vernon	Hologic	1000
16	Oxford, UK	Oxford	Hologic	1000
17	Oxford, UK	Oxford	Hologic	4000
18	Rotterdam, NL	AZR	Lunar	DPX
19	Rotterdam, NL	AZR	Lunar	DPXL
20	University College, London, UK	Royal Free	Hologic	4500
21	University College, London, UK	Stanmore mobile	Hologic	1000
22	University College, London, UK	Stanmore	Hologic	1000
23	Rotterdam, NL	St. Francis	Hologic	4500
24	Southampton, UK	Southampton	Hologic	2000
25	Southampton, UK	Southampton	Lunar	DPX
26	University College, London, UK	Middlesex	Hologic	4500
27	Glasgow, UK	Glasgow	Hologic	1000

using the residual maximum likelihood (REML) directive of GENSTAT [10]. Differences between different DXA versions, between the three ESP densities and the interaction of ESP density and DXA version were treated as fixed effects. Differences between machines of the same type and any interactions with this term were treated as random effects. Residual variation from this model comprises differences within ESP densities on each machine (i.e., between the ten replicate readings).

Since, at each scan, all three BMD densities were recorded together, estimates of variation within ESP density classes on the same machine may be correlated with each other. Such an effect could bias significance levels. However, estimates of the three correlations between the three pairs of BMD densities, pooled over machines, were small and statistically insignificant (0.054, -0.067, and 0.116). We therefore made the simplifying assumption that errors from replicate measurements on the same machine, but with different ESP density classes, are independent. We treated differences between similar versions of DXA scanner (up to 3% variation) as random effects, as we regarded these as replicates of the same type of machine, as might be observed in general clinical use.

Short-term stability of the machines used in this study was assessed by examining the standard deviation of the ten replicate readings on each machine. Calibration curves were fitted to all machines for the variable densities of the ESP, using the statistical package GENSTAT [10]. Initially, quadratic curves were fitted for each machine. Although these gave excellent fits to the data, extension of the curves outside the range of the ESP

produced a maximum BMD value, for some machines, which was lower than some of the values seen in practice using those machines. As a consequence, very high BMD values observed on these machines (typically for L3 or L4) could not be converted to ESP-adjusted values, since the observed BMD value would not fall on any part of the calibration curve. For this reason, an exponential curve was fitted to all machines. These gave as good a fit within the range of the phantom but behaved better at very high BMD values with the curve rises to an asymptote outside the range of BMD values seen in practice. An example calibration curve is given in Fig. 1.

Calibration curves were not constrained to pass through zero, as it would not produce a more accurate

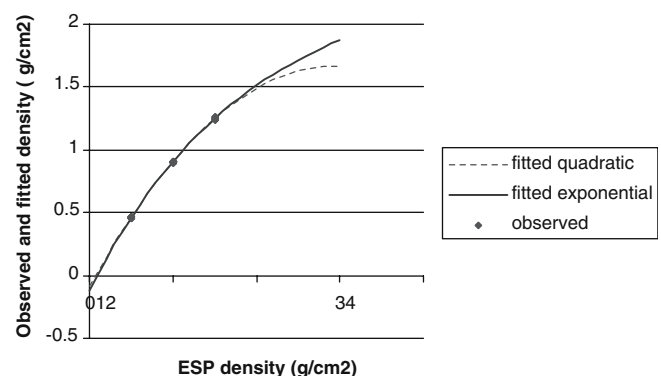


Fig. 1 Example of the quadratic and exponential calibration curves (Oxford Hologic QDR 1000). Each data point plotted represents ten observations at each of three densities

calibration curve for the majority of values that fall within the range of the ESP and may well produce less accurate calibrations over this range.

We validated the effectiveness of the calibration curves, using ESP-calibrated and non-calibrated BMD values from hip and spine measurements on 991 female family members of the probands scanned as part of the FAMOS collaboration. The validation allowed us to partition variation into components between and within scanners. If calibration has been effective in removing variation between machines, then an analysis on calibrated data should remove less variation than an analysis on raw, non-calibrated data. The effect of age on the calibrated and raw BMD values was examined, using simple linear regression.

Results

Results from the model fitting exercise are given in Table 2. Differences between various versions from the same manufacturers, between densities, and their interaction are all extremely significant ($P < 10^{-6}$). However, there was no evidence of significant variation between machines of the same version ($P = 0.213$). This implies that identical versions of machines from each manufacturer behave very similarly. However, variation that can be attributed to the interaction between ESP values and the machine type is large ($P < 10^{-6}$), approaching the replicate variance itself, hence confirming the need for separate calibration curves for each machine.

Mean values of ESP for all 27 machines, grouped by version, are given in Table 3. It can be seen that Hologic machines tended to underestimate nominal BMD values across the range of densities, while Lunar machines tended to overestimate the nominal value. The two Norland scanners showed a mixed picture, overestimating the density at 0.5 g/cm² but underestimating the higher BMD values, substantially so in the case of the 1.5 g/cm² nominal vertebra. Standard errors averaged 0.0102 g/cm² across all machines, with the minimum error being 0.0042 g/cm² and the maximum 0.0134 g/cm².

The analysis of the standard deviation of readings, used as a measure of repeatability for each of the machines, separately, is given in Table 4, and the mean

effects for different versions of the same machine are shown in Table 5. There are large and significant differences between versions, between ESP BMD values and in the interaction of these two effects. The low numbers of machines for many versions makes interpretation of the mean effects tenuous. However, on the basis of these data, Hologic machines are the most stable, followed by Lunar, then Norland (although data here are based on only two machines of different specifications). For Hologic and Lunar, repeatability decreases with increasing density, but, as there were only two Norland machines, no comment can be made on linearity with this manufacturer's machines. The significance of the "machines within versions" component of variance demonstrates that there is at least as much or more variation in repeatability among machines of the same version as that seen within machines of different versions and manufacturers. This is in contradistinction to the data on the results of ESP BMD, where variation between machines of the same version was much less (and insignificant) than the variation between different versions of the same manufacturer's machine and different manufacturers.

Means and variances for BMD at each of three lumbar vertebrae, at the femoral neck hip, and at the trochanter, for 991 female probands from the FAMOS study, are given in Table 6. Mean values after calibration were numerically greater than mean values before, and this produced similarly numerically greater variances. The percentage variation accounted for by scanner and, for comparison by age, is presented in Table 7. It is clear that substantially more variation is accounted for by machine when raw BMD is analyzed than when calibrated BMD is analyzed. This is consistent with the calibration's being effective in removing differences between different machines and different versions of the same machine. Percentage variation accounted for by regression on age is always highly statistically significant. The percentage variation accounted for is similar with both calibrated and raw BMD, and there is, therefore, no evidence that the calibration has introduced a new source of error variation into the measurement of BMD on these individuals.

The results of the cross-calibration were used to derive center- and machine-specific cut-offs based on Z-scores of -1.28 and -2.0. In summary, the cut-off,

Table 2 Analysis of differences between machines (*SE* standard error, *NS* not significant)

Random effect	Variance component	SE	Significance ^a
Machines within versions	0.000032	0.000030	NS
(Machines within versions) × ESP density	0.000132	0.000035	< 0.001
Error	0.000161	0.000009	
Fixed effect	df	Wald statistic^b	Probability
ESP BMD	2	66264.71	< 0.001
Version	8	979.58	< 0.001
Version × ESP BMD	16	599.83	< 0.001

^aSignificance tests based on change in $-2 \times \log$ -likelihood on dropping terms for random effects. Although the random effect "machines within versions" is not significant it has been kept in the final model fitted.

^bWald statistics for fixed effects, asymptotically distributed as chi-squared, are calculated ignoring terms fitted later in the model

Table 3 Average BMD at three ESP nominal values for 27 machines in the FAMOS study

Version	Location	0.5 g/cm ²	1.0 g/cm ²	1.5 g/cm ²
Hologic 1000	Århus	0.4637	0.9113	1.2409
Hologic 1000	Losser	0.4572	0.9017	1.2540
Hologic 1000	Mount Vernon	0.4656	0.9081	1.2267
Hologic 1000	Oxford	0.4611	0.9058	1.2487
Hologic 1000	Stanmore mobile unit	0.4715	0.908	1.2237
Hologic 1000	Stanmore	0.4568	0.9059	1.2365
Hologic 1000	Glasgow	0.4587	0.9098	1.2223
Hologic 2000	Århus	0.4925	0.9759	1.3583
Hologic 2000	Southampton	0.4969	0.9233	1.3346
Hologic 4000	Oxford	0.4597	0.9047	1.2284
Hologic 4500	Århus	0.4753	0.9399	1.3403
Hologic 4500	Cambridge	0.4541	0.9301	1.3452
Hologic 4500	Eastbourne	0.4715	0.9265	1.3413
Hologic 4500	Hertford	0.4767	0.9328	1.3466
Hologic 4500	Hvidore	0.4750	0.9302	1.3253
Hologic 4500	Ijsseland	0.4695	0.913	1.3210
Hologic 4500	Losser	0.4638	0.9241	1.3219
Hologic 4500	Royal Free	0.4678	0.9248	1.3316
Hologic 4500	St Francis	0.4732	0.9305	1.3353
Hologic 4500	Middlesex	0.5108	0.9718	1.3381
Lunar DPX	AZR	0.5258	1.0623	1.5252
Lunar DPX	Southampton	0.5417	1.0670	1.4955
Lunar DPXL	AZR	0.5088	1.0612	1.5318
Lunar Expert	Aberdeen	0.5214	1.0164	1.5090
Lunar Expert	Cambridge	0.4802	1.0478	1.4956
Norland XR26	Aberdeen	0.5133	0.9261	1.2780
Norland XR36	Aberdeen	0.5541	0.9464	1.3180

post-calibration BMD was calculated from the Hologic male and female normative ranges, as this was the most commonly used manufacturer of densitometer in our collaboration. Femoral neck and lumbar spine (L2–4) post-calibration BMD was calculated from the derived machine-specific exponential equations. For each age, a cut-off standardized BMD was used to calculate what the cut-off for measured BMD would be for both male and female subjects and at both chosen regions of interest. The Z scores for the cut-off values were then calculated by the equation:

$$Z \text{ score} = \frac{\text{Standardized BMD} - \text{Measured BMD}}{\text{ISD Standardized BMD}} \quad (1)$$

The individual center- and machine-specific Z-scores were used to recruit subjects for the FAMOS genetic study, as indicated in the introduction.

Discussion

Previous efforts to cross-calibrate DXA BMD have been based on measurements of phantoms [9], human volunteers [6] or a mixed approach [3]. These approaches produce reasonable agreement for absolute BMD, hence allowing clinical measurements on one scanner to be compared with those from another manufacturer. However, using standardized BMD for multi-site epidemiological studies is more problematic as there is a need to relate the findings to the expected age-associated bone loss, which may be center-specific. To enable such comparisons, a European normative range for both lumbar spine [11] and femoral neck [12] was developed, using the ESP for both lumbar spine [11] and femoral neck [12] was developed using the ESP. The same approach was used to determine real geographic differ-

Table 4 Analysis of repeatability across machines (*SE* standard error)

Random effects Effect	Variance component	SE	Significance ^a
Machines within versions	0.000012154	0.000004874	< 0.001
Error variance	0.00000711	0.000001677	
Fixed Effects	df	Wald statistic ^b	Probability
Version	8	53.65	< 0.001
ESP BMD	2	84.50	< 0.001
Version × ESP BMD	16	42.15	< 0.001

^aSignificance test based on change in log-likelihood on dropping terms for random effects

^bWald statistics for fixed effects, asymptotically distributed as chi-squared, are calculated ignoring terms fitted later in the model

Table 5 Average repeatability across DXA versions and ESP nominal BMDs. Standard errors vary with number of machines tested for each model

Version	No. of machines	ESP BMD ^a			
		0.5	1.0	1.5	Average
Hologic 1000	7	0.00440	0.00795	0.01104	0.00780
Hologic 2000	2	0.00510	0.01071	0.01386	0.00989
Hologic 4000	1	0.00216	0.00558	0.00611	0.00462
Hologic 4500	10	0.00743	0.00626	0.01069	0.00813
Lunar DPX	2	0.00943	0.01053	0.01887	0.01294
Lunar DPXL	1	0.01008	0.01182	0.01837	0.01342
Lunar Expert	2	0.01265	0.01238	0.01879	0.01460
Norland XR26	1	0.01723	0.01155	0.03125	0.02001
Norland XR36	1	0.03130	0.02478	0.04196	0.03268
Average		0.01109	0.01128	0.01899	

^aNominal values**Table 6** Mean and variance of BMD at each of three scan sites on female non-probands from the FAMOS study

BMD region of interest	No. of observations	Raw BMD		Calibrated BMD	
		Mean	Variance	Mean	Variance
L2	974	0.940	0.031	1.000	0.039
L3	977	0.984	0.030	1.053	0.040
L4	969	0.997	0.027	1.071	0.041
Femoral neck	975	0.759	0.020	0.787	0.021
Trochanter	952	0.657	0.014	0.678	0.014

Table 7 Percentage of variance in BMD accounted for by regression on scanner type (manufacturer and variety) and on age

BMD region of interest	Raw BMD		Calibrated BMD	
	Scanner	Age	Scanner	Age
L2	14.1	17.9	2.6	18.5
L3	17.2	16.0	4.9	16.6
L4	11.7	10.8	5.6	10.1
Femoral neck	20.1	21.6	8.8	23.6
Trochanter	14.2	14.1	7.6	14.8

ences between BMD at centers across Europe taking part in the European Vertebral Osteoporosis Study (EVOS) [13] and at different centers in subjects with osteoporotic fractures [14].

Our primary aim was to conduct a genome-wide scan for linkage to BMD as assessed by axial DXA. To advance this program, we required not only to have a method of standardizing BMD across eight different sites in Europe but also to relate the results to a normative range, thus enabling selection of approximately equal proportions of female probands at each site with equivalent age-specific standardized BMD. Accordingly, we first calibrated 27 DXA machines using an ESP and then validated our calibration equations in a collection of 991 female subjects collected as part of the FAMOS study. The considerable reduction in variation between machines after calibration, coupled with the absence of a similar reduction in variation that can be attributed to a linear regression on age, demonstrates not only the effectiveness of the calibrations but, moreover, the importance of carrying out this exercise. There was,

however, variation remaining between machines after calibration, and, while this could reflect inaccuracies in the calibration exercise, it is more likely that these residual effects were a result of real differences among local populations in our sample, as has been shown previously [13, 14].

Having derived the cross-calibrations, we then calculated center-specific cut-offs for age-corrected BMD (Z-scores) from a normative range of standardized BMD. We originally planned to use the previously ESP-derived European normative range, which give standardized BMD for the lumbar spine [11] and femoral neck [12], our chosen regions of interest for proband selection. Although the site recognized as the standard for diagnosis of osteoporosis is total hip BMD [15], this site was not available in this study as we chose to use historical BMD databases from which to select our probands. Our decision not to use the European normative range was based on the fact that the NHANES III normative range is now applied as standard on all the BMD devices used in this study; although this normal

range is available for only the hip sites [7]. We decided to standardize the data using the “Hologic” normative range, as the majority of scanners available in our centers were from this manufacturer. The new Hologic normative range is NHANES-compatible for the hip [16]. Unlike the old Hologic normative range for hip BMD [17], the normative range for spine BMD is consistent across the two major manufacturers.

Despite use of the most prevalent normative NHANES (for hip) and manufacturer-specific for spine ranges available in our study centers we still used our cross-calibration algorithm for each machine because, after modeling, we still detected significant differences between different versions from the same manufacturers after cross-calibration with the ESP, as shown in Table 2.

Given the relatively small number of versions of machines from some manufacturers in this data set, it is difficult to draw firm conclusions about the merits of the different versions or manufacturers. However, there are differences between machine versions in repeatability, and there are also significant differences in calibration equations between machines of the same version. This gives support to the cautious approach taken in this study of producing a separate calibration curve for each machine. However, not all scanners used in the FAMOS study collaboration were calibrated. In addition to the 27 scanners calibrated here, some scanners have been used for small numbers of individuals who were typically relatives of probands who lived a considerable distance away from one of the centers in this study. For these individuals, it was impracticable to calibrate all the scanners involved and is of lesser importance to our data collection, as the vast majority of these individuals were family members rather than probands. Fortunately, it is clear from the clustering of values given in Tables 3 and 4 that an average calibration curve for the version, or if the specific version is unknown then for the manufacturer, while not ideal, is better than excluding the data.

The problem of observed BMD values producing calibrated values that fall outside the range of the ESP merits further comment. There were 4,847 calibrated BMD values available for use in our validation study from the five measured scan sites (L2, 3, 4, femoral neck, and trochanter) in the cohort of 991 women. Of these 4,847 values, 1.55% fell below the 0.5 g/cm² lower ESP value, and 0.99% were above the upper 1.5 g/cm² ESP value. Hence, 2.54% of the observations fell outside the range of accurate calibration. In retrospect, a phantom with more extreme values than the ESP used here would have been preferable. For the FAMOS study we have assumed that the projection of the calibration curves outside their range does not introduce additional sources of error. In this respect, fitting an exponential curve has the effect of producing a calibrated value for all machine values observed to date. This implies more realistic behavior for this method than that produced by using quadratic calibration. Moreover, the results from the

validation exercise presented in Table 7 indicate that the calibration to an ESP has, indeed, worked in practice.

In conclusion, this cross-calibration exercise has been successful in allowing machine- and version-specific calibration curves to be produced for each instrument at each of the centers involved in the FAMOS study. The validation of this exercise demonstrates that, for multi-center studies where absolute BMD cut-offs are to be produced, this method of cross-calibration will be more successful than the previous methodology of producing standardized BMD values that do not allow adequately for machine and version variance. This method is thus effective for selecting comparative age-matched BMD cohorts in different geographical populations.

Acknowledgements We are grateful to Lon Cardon for his advice on the statistical design of the study. We are also extremely grateful to FAMOS research nurses for their help and dedication: Janet Ball, Irene Debiram, Lise Hejl, Vera Herd, Alison James, Debbie Nelson, Ria Skelton, Rita Smith, Ingrid Struik and Bridget Watkins.

References

1. Baran DT, Faulkner KG, Genant HK, Miller PD, Pacifici R (1997) Diagnosis and management of osteoporosis: guidelines for the utilization of bone densitometry. *Calcif Tissue Int* 61:433–440
2. Blake GM, Fogelman I (2002) Dual energy X-ray absorptiometry and its clinical applications. *Semin Musculoskelet Radiol* 6:207–218
3. Genant HK, Grampp S, Gluer CC, et al (1994) Universal standardization for dual X-ray absorptiometry: patient and phantom cross-calibration results. *J Bone Miner Res* 9:1503–1514
4. Kalender WA (1992) A phantom for standardization and quality control in spinal bone mineral measurements by QCT and DXA: design considerations and specifications. *Med Phys* 19:583–586
5. Kalender WA, Felsenberg D, Genant HK, Fischer M, Dequeker J, Reeve J (1995) The European spine phantom—a tool for standardization and quality control in spinal bone mineral measurements by DXA and QCT. *Eur J Radiol* 20:83–92
6. Hui SL, Gao S, Zhou XH, et al (1997) Universal standardization of bone density measurements: a method with optimal properties for calibration among several instruments. *J Bone Miner Res* 12:1463–1470
7. Looker AC, Orwoll ES, Johnston CC Jr, et al (1997) Prevalence of low femoral bone density in older US adults from NHANES III. *J Bone Miner Res* 12:1761–1768
8. Lu Y, Fuerst T, Hui S, Genant HK (2001) Standardization of bone mineral density at femoral neck, trochanter and Ward’s triangle. *Osteoporos Int* 12:438–444
9. Pearson J, Dequeker J, Henley M, et al (1995) European semi-anthropomorphic spine phantom for the calibration of bone densitometers: assessment of precision, stability and accuracy. The European Quantitation of Osteoporosis Study Group. *Osteoporos Int* 5:174–184
10. Genstat (2000) Genstat for Windows. Release 4.2, 5th edn. VSN International Ltd, Oxford
11. Dequeker J, Pearson J, Reeve J, et al (1995) Dual X-ray absorptiometry—cross-calibration and normative reference ranges for the spine: results of a European Community Concerted Action. *Bone* 17:247–254
12. Pearson J, Dequeker J, Reeve J, et al (1995) Dual X-ray absorptiometry of the proximal femur: normal European values standardized with the European spine phantom. *J Bone Miner Res* 10:315–324

13. Lunt M, Felsenberg D, Adams J, et al (1997) Population-based geographic variations in DXA bone density in Europe: the EVOS Study. *European Vertebral Osteoporosis. Osteoporos Int* 7:175–189
14. Kroger H, Lunt M, Reeve J, et al (1999) Bone density reduction in various measurement sites in men and women with osteoporotic fractures of spine and hip: the European Quantitation Of Osteoporosis Study. *Calcif Tissue Int* 64:191–199
15. Kanis JA (2002) Diagnosis of osteoporosis and assessment of fracture risk. *Lancet* 359:1929–1936
16. Looker AC, Wahner HW, Dunn WL, et al (1998) Updated data on proximal femur bone mineral levels of US adults. *Osteoporos Int* 8:468–489
17. Faulkner KG, Roberts LA, McClung MR (1996) Discrepancies in normative data between Lunar and Hologic DXA systems. *Osteoporos Int* 6:432–436