Reproducibility of bone mineral density measurement in daily practice

M C Lodder, W F Lems, H J Ader, A E Marthinsen, S C C M van Coeverden, P Lips, J C Netelenbos, B A C Dijkmans, J C Roos

Background: Bone mineral density (BMD) measurements are frequently performed repeatedly for each patient. Subsequent BMD measurements allow reproducibility to be assessed. Objective: To examine the reproducibility of BMD by dual energy x ray absorptiometry (DXA) and to investigate the practical value of different measures of reproducibility in a group of postmenopausal women.

Methods: Ninety five women, mean age 59.9 years, underwent two subsequent BMD measurements of spine and hip. Reproducibility was expressed as smallest detectable difference (SDD), coefficient of variation (CV), and intraclass correlation coefficient (ICC). Sources of variation were investigated by multilevel analysis.

Results: The median interval between measurements was 0 days (range 0–45). The mean difference (SD) between the measurements (g/cm²) was −0.001 (0.02) and −0.0004 (0.02) at L1–4 and the total hip, respectively. At L1–4 and the total hip, SDD (g/cm²) was ±0.05 and ±0.04 and CV (%) was 1.92 and 1.59, respectively. The ICC at spine and hip was 0.99.

Conclusions: Reproducibility in the postmenopausal women studied was good. In a repeated DXA scan a BMD change exceeding 2×CV (%), the least significant change (LSC), or the SDD should be regarded as significant. Use of the SDD is preferable to use of the CV and LSC (%) because of its independence from BMD and its expression in absolute units. Expressed as SDD, a BMD change of at least ±0.05 g/cm² at L1–4 and ±0.04 g/cm² at the total hip should be considered significant.

Abbreviations: BMD, bone mineral density; BMI, body mass index; CV, coefficient of variation; DXA, dual energy x ray absorptiometry; ICC, intraclass correlation coefficient; LSC, least significant change; SDD, smallest detectable difference.
interval between the two DXA measurements was 45 days. A change in BMD was not expected during this interval.

At the same time girls participating in a local study on the effects of puberty on BMD had their BMD measured twice.24

Ethics
All four study protocols were approved by the local medical ethics committee.

Methodology of BMD measurement
All BMD measurements were performed on a Hologic QDR 2000 machine (Hologic Inc, Waltham, Massachusetts). The software version used was V4.7. The DXA scans were obtained by standard procedures supplied by the manufacturer for scanning and analysis. The compare feature was used for the second scan. No records were kept of difficulties observed in the positioning of patients. Plain x rays documenting the presence of arthritic changes were not used. Daily quality control was carried out by measurement of a Hologic anthropomorphic spine phantom. At the time of the duplo measurements, phantom measurements showed stable results. The phantom precision expressed as the CV (%) was 0.82. The BMD measurements were carried out by experienced technicians.

Patient BMD was measured at the lumbar spine (L1-4) (posteroanterior projection) and at the left femur. At the femur the following sites were studied: total hip, femoral neck, and trochanter. When the two measurements were made on the same day, the patient was completely repositioned after the initial measurement.

T and Z scores were calculated using the reference population provided by the manufacturer. In the T score, the patient’s BMD value is expressed as SD as compared with the mean BMD of a reference population of young adults. For Z score calculation, the patient’s BMD is compared with the mean BMD of people of the same sex and age and also expressed as SD.

ΔBMD and the ΔT score were calculated by subtracting the results of the second measurement from the results of the first. The range of the difference in BMD as a percentage was calculated by dividing the difference between the first (a) and the second (b) measurement by the mean of those two figures, giving the fraction of difference between the two measurements as compared with the mean of the two measurements. The normally distributed variables are presented as mean (SD).

Precision
The measurement error was calculated using Bland and Altman’s 95% limits of agreement method. Other methods used to evaluate reliability and agreement are also described. These are the CV and the intraclass correlation coefficient (ICC).

Precision expressed according to the Bland and Altman’s 95% limits of agreement method gives an absolute and metric estimate of random measurement error, also called SDD. In this case, where there are two observations for each subject, the standard deviation of the differences (SDdiff) estimates the within variability of the measurements. Most disagreements between measurements are expected to be between limits called “limits of agreement” defined as

\[
d \pm z(1-\alpha/2) \times SD_{diff}
\]

where d is the mean difference between the pairs of measurements and \( z_{1-\alpha/2} \) is the 100(1-\( \alpha/2 \))th centile of the normal distribution.29 The value d is an estimate of the mean systematic bias of measurement 1 to measurement 2. d is expected to be 0 because we do not assume a true change in BMD to occur during the interval between the two BMD measurements. Defining \( z \) to be 5%, the limits of agreement are +1.96SDdiff and −1.96SDdiff. Thus, about twice the standard deviation (SD) of the difference scores gives the 95% limits of agreement for the two measurements by the machine. A test is considered to be capable of detecting a difference, in absolute units, of at least the magnitude of the limits of agreement.

The CV, the most commonly presented measure for BMD variability, is the SD corrected for the mean of paired measurements. CV, expressed as a percentage, was calculated as

\[
CV\% = \left( \frac{\sqrt{\sum (a-b)^2} / (n)}{\text{mean of all differences}} \right) \times 100
\]

where a and b are the first and second observations, Ma and Mb are the mean values for the two groups, and n is the number of paired observations.24 For two point measurements in time, a BMD change exceeding 2/2 times the precision error of the technique is considered a significant change (with 95% confidence).2 Gluer et al called this smallest change that is considered statistically significant, the least significant change (LSC).11 In the current study, the LSC (%) was computed for the different BMD measurement sites. In these calculations the precision error is expressed as the CV (%).

The ICC equals variance between patients divided by variance between patients plus variance between measurements. The value of the ICC ranges from 0 to 1, 1 representing perfect reliability of the measurement.

Multilevel analysis of variability
Because multiple linear regression analysis does not allow us to discern whether the observed variation in BMD is attributable to individual differences between patients or to the influence of interval length, multilevel analysis was used.27 Multilevel analysis can separate and quantify individual variability in contrast with measurement of interval length variability. The models used were two level variance component models, with measurement interval at the first level and patients at the second level. Demographic variables—for example, age (years) and body mass index (BMI) (kg/m2), and BMD variables, such as area (cm²) of BMD and the intraclass correlation coefficient, were included as sources of BMD variability—that is, possible confounders that need correction. Separate models were used for different measurement sites. The fixed parameters in the final models describe the average contribution of each confounder. Patients with their corresponding interval between measurements will vary around the predictive value of the model. The degree of variation at patient level and at interval level is estimated in the random part of the model.

Statistical analysis was carried out using SPSS, version 9.0 (SPSS, Chicago, Illinois) and MLwiN for multilevel analysis.25

RESULTS
Postmenopausal women
Patient characteristics
The BMD measurements of 95 postmenopausal women were collected during the recruitment period. The mean (SD) age of the women was 59.9 (8.1) years. Their mean (SD) height was 163.7 (6.6) cm and their mean (SD) weight 68.8 (10.5) kg. The mean (SD) BMI was 25.7 (3.8) (kg/m²). The interval between the first and the second measurement was 24 months (SD 0.60). The mean (SD) T scores of the first measurement were 1.30 at L1-4 and 1.84 at the total hip. A change in BMD to occur during the interval between the two BMD measurements. The range of the difference between the T score of the first and the second BMD measurement was 0.42 to +1.30 at L1-4 and −1.84 to +0.60 at the total hip.
Variability

Table 2 presents the results of the various methods of calculating variability for the three most frequently used measurement sites. Figures 1 and 2 show the scatter plots of the difference between the two measurements against their mean, for lumbar spine and total hip. The horizontal lines in these graphs show the mean of the differences and the limits of agreement. When Bland and Altman’s 95% limits of agreement were used, the mean of the differences approached zero, reflecting no systematic bias between measurements (the 95% CI included zero difference). In this method, random measurement error, is expressed as SD of the differences. Twice this value approaches the 95% confidence intervals at all measurement sites.

Multilevel analysis

The fixed part of the final model for the spine (L1–4), apart from the constant, contained the variables technician, age, and area of BMD measurement. Both age and area predict BMD, though these variables do not influence BMD variability. The random part of the model shows that most of the variance in BMD (98.6%) can be attributed to the technician in absolute units) of 0.04 g/cm² or the LSC (%) of 4.50%.

Reliability expressed by ICC was 0.99 with narrow 95% confidence intervals at all measurement sites.

Table 2
Reproducibility of BMD (g/cm²) measurement in postmenopausal women (n = 95) and children (n = 23)

<table>
<thead>
<tr>
<th></th>
<th>Postmenopausal women</th>
<th>Children</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L1–4</td>
<td>Femoral neck</td>
</tr>
<tr>
<td>Bland and Altman’s methods</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean difference (95% CI)</td>
<td>−0.001</td>
<td>−0.0008</td>
</tr>
<tr>
<td>(systematic bias)</td>
<td>(−0.05 to 0.05)</td>
<td>(−0.04 to 0.04)</td>
</tr>
<tr>
<td>SD difference</td>
<td>0.0238</td>
<td>0.0220</td>
</tr>
<tr>
<td>(random measurement error)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SDD 95% limits of agreement</td>
<td>±0.0466</td>
<td>±0.0432</td>
</tr>
<tr>
<td>(assuming no systematic bias)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CV (%)</td>
<td>1.92</td>
<td>2.25</td>
</tr>
<tr>
<td>LSC (%)</td>
<td>5.43</td>
<td>6.36</td>
</tr>
<tr>
<td>Random effects ICC (95% CI)</td>
<td>0.99 (0.98 to 0.99)</td>
<td>0.99 (0.98 to 0.99)</td>
</tr>
</tbody>
</table>

Mean difference, mean of the difference between the first and the second BMD measurement; SD difference, SD of the difference between the first and the second BMD measurement; SDD, smallest detectable difference (g/cm²); CV, coefficient of variation (%); LSC, least significant change (%); ICC, intraclass correlation coefficient. SD difference, and SDD are rounded off to two decimal places in the text of the article.
measurement sites in the hip, BMD variation is mainly explained by patient variability in BMD; the interval between measurements and technician variation in the case of the total hip explain only a small part of BMD variation (figures not shown).

**Children**

**Patient characteristics**
The mean (SD) age of the 23 girls investigated was 11.2 (1.3) years. In each girl, the first and second BMD measurements were carried out on the same day, with complete repositioning between the two measurements. The mean (SD) of the first BMD measurement (g/cm²) was 0.72 (0.11) at the spine, 0.66 (0.08) at the femoral neck, and 0.70 (0.11) at the total hip. Although the mean of the difference in BMD values for the children was of the same order as for the postmenopausal women, the SD and the range of the difference were smaller in children than in the postmenopausal women. At the lumbar spine, for example, the mean (SD) of the difference between the two measurements was −0.009 (0.009) (range −0.03 to +0.01) in children, whereas these figures were −0.001 (0.02) (range −0.05 to +0.14) in the women. At the total hip these figures were −0.003 (0.01) (range −0.04 to +0.02) in children and −0.0004 (0.02) (range −0.04 to +0.07) in the postmenopausal women, respectively.

**Variability**
Table 2 and figs 1 and 2 show that the SDD tended to be smaller in the children than in the women.

In the children CV% at L1-4 and the total hip were 0.84 and 1.19, whereas in the postmenopausal women these figures were 1.92 and 1.59, respectively. Consequently, the LSC in children is smaller. Hence, as compared with the women, a smaller change in BMD can be regarded as a significant change.

The ICC was as high in the children as in the postmenopausal women (table 2).

**Multilevel analysis**
The group of children described was considered too small for multilevel analysis to be performed.

**DISCUSSION**
This study shows the in vivo short term variability of BMD measurement by DXA in a group of postmenopausal women with a wide range of BMD values. In the group of women studied, reproducibility expressed by different means is good. The clinician interpreting a repeated DXA scan of a subject should be aware that a BMD change exceeding the LSC is significant, here arising from a BMD change of at least 4.5% at the total hip and 5.4% at the spine. Expressed as SDD, a BMD change should exceed 0.04 g/cm² at the total hip and 0.05 g/cm² at the spine before it can be considered a significant change.

Despite the many publications on BMD variability in different patient groups, only limited data from a large number of postmenopausal women are available on short term BMD variability.6–11 In the reports published, variability is usually expressed as CV and the figures for short term variability are lower than the ones we found.4–7 Two studies showed variability data more in line with our results.11 29 Two samples of healthy (n = 70) and elderly (n = 57) postmenopausal women showed a CV (%) of 0.9 and 1.8, respectively, at the spine, and of 0.9 and 2.3, respectively, at the total hip.13 Eastell showed an LSC (%) of 5.4 at the lumbar spine and 8 at the femoral neck, respectively, in osteoporotic postmenopausal women.17 The varying results of reproducibility studies might be explained by the “population” investigated; a phantom and healthy young subjects are likely to show more favourable variability than postmenopausal women, possibly in part because of easier positioning for measurement. The current study also shows better variability, expressed as CV (%), in children. Secondly, osteoarthrits in postmenopausal women may contribute to poorer variability than found in healthy young subjects. Besides, the majority of the studies mentioned had small patient samples, giving less precise results.

Alternative measures of variability are the SDD and ICC. The Bland and Altman plots visualise between measurement differences. The scatter plots of the current data show a random distribution of values, indicating the absence of a relationship between the measurement error and the true BMD value, as estimated by the mean of the two measurements. The SDD values found in the adult patients were slightly higher than the figures presented by Ravaud et al.13 In the first group of postmenopausal women (mean age 53 years) they describe, the SDD was 0.02 (g/cm²) at the total hip and 0.02 at the lumbar spine. In the second group described, women with a mean age of 80 years, these figures were 0.04 and 0.04, respectively. The SDD values of the children studied tended to be lower than the values in the postmenopausal women. Using the SDD one can state that a (BMD) change larger than the figure found is a true (BMD) change in 95% of the cases. The characteristics of the Bland and Altman method thus allow direct insight into the variability of the measurement under study. Previously published reports,13 18 as well the current data, show that reproducibility expressed in absolute units (SDD) is independent of the BMD value. Reproducibility expressed as a percentage (CV) and the derived LSC, however, depend on the BMD value. Because of therapeutic consequences, the clinician should be especially careful in judging an apparent BMD change in patients with osteoporosis. The use of the SDD in the evaluation of an apparent BMD change gives a more conservative approach than the use of the CV at low BMD. Because of its independence from the BMD level and its expression in absolute units, the SDD is a preferable measure for use in daily clinical practice as compared with the CV and the derived LSC.

The ICC found in postmenopausal women and children was high, indicating good overall reproducibility of BMD measured by DXA. However, it is important to note that a
large variability between patients automatically increases the ICC. The ICC and the Bland and Altman method yield complementary information; the presence of systematic bias cannot be found by estimating ICC.

Although the variability as expressed by the ICC, and especially the SDD, is reassuring, showing good short term variability at group level, the wide range of the differences in BMD and the derived T scores indicates considerable individual differences between two consecutive BMD measurements in some patients. The range in T scores, for example, indicates that in some patients the diagnosis, based on the diagnostic thresholds of the WHO, would change owing to the measurement variability. Measured as percentages, differences between BMD measurements in patients (table 1) should be interpreted with care because the difference as a percentage depends on the mean BMD of the two measurements. Similar BMD differences in the numerator of the fraction can yield a different percentage BMD change depending on the mean value of the BMD in the denominator of the fraction.

Before investigating sources of raw BMD variation we considered examining sources of ABMD variation. However, technician variability in ABMD could not be assessed because in some cases different technicians carried out the two consecutive DXA scans. To our knowledge, this is the first study using multilevel analysis to investigate the sources of BMD variation. It is difficult to interpret why the variable technician predicts BMD only at the spine and the total hip. The well known determinants of BMD, age and BMI, means is good at a group level. However, the clinician must not at the other measurement sites is also hard to understand. The known determinants of BMD, age and BMI, are also found in this study. BMD variation was mainly explained by patient variability, while interval length and technician in the case of the spine and total hip only explain a small part of the BMD variation. The local technicians thus have negligible influence on BMD variation. Whether this applies to other centres remains a question to be answered at the individual centre.

It should be noted that the presented variability figures hold exclusively for this patient group and this particular DXA device in our hands. The figures show that at our centre the favourable variability values presented in the literature cannot be reproduced in daily practice in postmenopausal women. Thus, for optimal clinical decision making, individual centres should establish the reproducibility figures based on routine DXA measurements in different patient groups.

In conclusion, reproducibility of BMD measurement by DXA in postmenopausal women expressed by different means is good at a group level. However, the clinician must remain aware that an apparent BMD change in an individual patient may represent a precision error. In daily practice, the use of the SDD is preferable to the use of the CV and LSC because of its independence from BMD level and its expression in absolute units.

REFERENCES