Osteoporos Int (2001) 12:500–505

Osteoporosis International

Original Article

Comparison of Two Methods for Measuring Ultrasound Properties of the Heel in Postmenopausal Women

B. M. Ingle, K. E. Sherwood and R. Eastell

Bone Metabolism Group, Section of Medicine, Division of Clinical Sciences, University of Sheffield

Abstract. Combined indices of ultrasound measurements have been proposed, such as ''stiffness index'' (SI) for the Lunar Achilles+ and 'quantitative ultrasound index' (QUI) for the Hologic Sahara ultrasound devices. We used the Bland and Altman approach and the kappa (κ) score (classifying women by tertile, independent of age) to compare these methods. We studied 105 postmenopausal women (ages 57 to 88 years). We measured the heel (in duplicate) using both devices. Single lumbar spine (LS) bone mineral density (BMD) measurements were also made using the same two manufacturers' densitometers. QUI values were higher than SI values with a mean difference of 2.4 units (95% CI, 1.5–3.2). This difference in SI and QUI was most marked at higher ultrasound values (r = 0.61, $p < 0.0001$). The kappa score between SI and QUI was 0.69 (95% CI, 0.57–0.80). When we calculated the kappa scores based on the mean of duplicate SI and QUI measurements, the kappa score increased to 0.90 (95% CI, 0.77–0.94). Lunar DPX LS-BMD values were higher than Hologic QDR 1000/W LS-BMD values with a mean difference of 0.18 g/cm^2 (95% CI, 0.17–0.19). The difference between the machines was most marked at higher BMD values (r = 0.38, $p < 0.001$). The kappa score between the DPX and QDR 1000/W was good (κ = 0.79, 95% CI = 0.66–0.88), and was similar to the agreement of SI and QUI. Based on a single measurement, some women would be classified in different tertiles using the two heel ultrasound machines (about 20%). However, this is not significantly greater than the misclassification rate using two machines to measure spinal BMD (about 15%). Although there are significant differences between SI and QUI measurements, the misclassification rates are similar to those observed measuring LS-BMD using two different manufacturers' DXA machines. The misclassification rate using quantitative ultrasound improves when based on duplicate measurements.

Keywords: Classification; Dual-energy X-ray absorptiometry; Equivalence; Method comparison; Osteoporosis; Ultrasound

Correspondence and offprint requests to: Professor R. Eastell, Bone Metabolism Group, Division of Clinical Sciences, Clinical Sciences Centre, Northern General Hospital, Sheffield S5 7AU, UK.Tel: +44 0114-271-4705. Fax: +44 0114-261-8775. e-mail r.eastell@sheffield.ac.uk

Introduction

There has been a recent increase in the use of quantitative ultrasound (QUS) devices for the assessment of fracture risk. QUS technology has the potential to meet the increased demand for bone densitometry services due to the progressive ageing of the world's population. Despite this, the widespread clinical application of QUS has been limited because dual-energy X-ray absorptiometry (DXA) remains the accepted gold standard for assessing skeletal status. However, DXA devices are relatively expensive and require patients to be referred to hospital-based facilities. QUS has several advantages including no patient or operator exposure to ionizing radiation, low cost and portability. It has been proposed that one of the main uses of ultrasound is to predict fracture risk. Large prospective fracture studies have demonstrated that both broadband ultrasound attenuation (BUA) and speed of sound (SOS) at the calcaneus can predict osteoporotic fracture as well as DXA at the spine and hip [1,2]. Recently the US Food

and Drug Administration (FDA) approved a number of commercial QUS systems, two of which are the Lunar Achilles+ and the Hologic Sahara. In addition to reporting SOS and BUA, both of these machines express their results as an index that combines SOS and BUA. The aim of these combined indices is to simplify the interpretation of these tests. For the Lunar Achilles+ this is called the ''stiffness index'' (SI) and for the Hologic Sahara it is called the ''quantitative ultrasound index'' (QUI).

It has been established that the absolute bone mineral density (BMD) at the lumbar spine (LS) differs between different manufacturers' densitometers by approximately 15% due to differences in calibration, correction for marrow fat, and edge detection algorithms between the different systems [3–6]. The Lunar Achilles+ and Hologic Sahara QUS machines also differ in their design, measurement site, coupling medium and calculation of measurement variables [7,8] and therefore (as with DXA) it may not be possible to make direct comparisons of results obtained using different systems. As the number of QUS devices in clinical use is set to increase it is important to know if the results from different manufacturers' devices are comparable.

The aims of this study were (1) to determine the relationship between combined heel indices (SI and QUI) when measured using two different manufacturers' heel QUS machines, (2) to determine the relationship between LS-BMD measured using two different manufacturers' densitometers; and (3) to compare the classification of individuals by tertile when measured by heel QUS and LS-BMD.

Materials and Methods

Subjects

We studied 105 postmenopausal women ages 57 to 80 years (mean 67 years) recruited from a population-based group. These women were randomly selected from General Practice lists. None had evidence of vertebral deformity on spinal radiographs. Biochemical screen excluded diseases likely to affect bone metabolism (thyroid function, serum calcium, phosphate, alkaline phosphatase, parathyroid hormone (PTH), and creatinine, and 24-h urine calcium excretion). The study protocol conformed to the Revised Helsinki Declaration of 1983, and was approved by the North Sheffield Local Research Ethics Committee. All subjects gave written informed consent.

Methods

Single lumbar spine measurements were made using the Lunar DPX (Lunar, Madison, WI; software version 3.6z) (L2–L4) and the Hologic QDR 1000/W (Hologic, Bedford, MA; software version 4.47) (L1–L4) on 95 of the 105 subjects. The short-term precision errors for LS-BMD measured by DXA have previously been reported by this group and were 1.02% for Lunar DPX and 1.06% for Hologic QDR [6]. Duplicate ultrasound measurements were made on all subjects (on the same day) using the Lunar Achilles+ (software version 1.51) and Hologic Sahara (software version 1.03) heel ultrasound machines. The non-dominant heel was measured in all subjects. Dominance was ascertained by asking the subjects which foot they would use to kick a ball. The short-term precision errors were estimated from the duplicate measurements with repositioning in between and were 1.5% for the Lunar Achilles+ SI and 2.4% for Hologic Sahara QUI. Measurements were performed on the Hologic Sahara before the Lunar Achilles+, this was to avoid any effects of wetting the heel during Lunar Achilles + measurements on the subsequent Hologic Sahara measurements.

Ultrasound Measurements

Hologic Sahara Clinical Bone Sonometer (HS): The Sahara Clinical Bone Sonometer consists of two 19-mm-diameter unfocused 0.5 MHz transducers mounted coaxially on a motorized caliper. One of the transducers acts as a transmitter the other as a receiver. The transducers are acoustically coupled to the heel using soft rubber angled pads and oil-based coupling gel. The Sahara device measures both BUA and SOS at a fixed region of interest in the mid-calcaneus and the results are combined to provide a single output measurement, the ''quantitative ultrasound index'' (QUI) using the following equation:

$$QUI = 0.41 \times (SOS + BUA) - 571$$

Lunar Achilles+ (LA+): The Lunar Achilles+ uses a water bath maintained at 37 °C. Two 25-mm-diameter 0.5 MHz unfocused transducers are mounted coaxially at a fixed separation of approximately 95 mm [9]. Water contains air bubbles, which affect the QUS measurement. Surfactant is added to disperse the air. The LA+ assumes a fixed heel width of 4 cm for the calculation of SOS; in contrast, BUA values are not normalized for heel width [10]. BUA and SOS are combined to provide a single output measurement, the ''stiffness index'' (SI), using the following equation:

$$SI = (0.28 \times SOS) + (0.67 \times BUA) - 420$$

Statistical Analysis

The precision error of QUS measurements was calculated as the ratio of the global estimate of standard deviation (root mean square error) of duplicate measurements to mean SI or QUI, expressed as a percentage. The

differences between the two QUS and DXA machines were assessed using a method comparison approach described by Bland and Altman [11]. This is where the difference between two measurements is taken as the dependent variable (as an estimate of the average bias of one measurement relative to the other) and the mean of the two measurements is taken as the independent variable (as the best estimate of the unknown true value). If two measurements are in close agreement, the differences should lie around zero, with no systematic relationship.

In order to apply the Bland and Altman approach correctly, the precision estimates (CV) of the methods being compared have to be similar. This is because the mean (of the two methods) should be the best estimate of the true value. In the current study, the data were of unequal precision for the QUS measurements (LA+ SI, 1.531% (L) and HS QUI, 2.427% (H)). The mean was therefore calculated after weighting the data statistically according to the reciprocal of the precision variances using the method described by Reeve and Lunt [12] and Eastell and Peel [13] (this is shown below). The weighting factor for the LA+ SI was 0.715 (CV_H^2/(CV_H^2 + CV_L^2) or $2.427^2/(2.427^2 + 1.531^2)$. The weighting factor for the HS QUI was 0.284 (CV_L^2/(CV_L^2 + CV_H^2) or $1.531^2/(1.531^2+2.427^2)$. Mean bias was then calculated as the mean of individual differences measured by Lunar and Hologic QUS and DXA machines. The 95% range of agreement was calculated as the mean ±2 SD of these individual differences. Linear regression model was fitted to the relationship between the difference and the mean level of the measurements.

The subjects were then classified into low, middle and upper tertiles. Kappa (κ) scores were then calculated to determine inter-rater agreement between the two QUS and DXA machines. Kappa scores have a maximum of 1.00 when agreement is perfect; a value of zero indicates no agreement better than chance, and negative values show worse than chance agreement. The guidelines for interpreting kappa scores are shown in Table 1.

Statistical analyses were performed using Statgraphics Statistical Graphics System version 5.0 (Statistical Graphics Corporation, Rockville, MD) and MedCalc Software version 4.15d (MedCalc Software, Mariakerke, Belgium).

Table 1. Guidelines for the interpretation of kappa (κ) scores.

Value of κ	Strength of agreement
Less than 0.20	Poor
0.21 to 0.40	Fair
0.41 to 0.60	Moderate
0.61 to 0.80	Good
0.81 to 1.00	Very good

Results

QUI values were slightly higher than SI values with a mean difference of 2.4 units, (95% CI, 1.5–3.2) (Fig. 1.). This difference was most marked at higher ultrasound values ($r = 0.46$, $p < 0.0001$). When we performed the Bland and Altman analysis using the mean of duplicate SI and QUI values this difference was reduced to 2.0 units, (95% CI, 1.4–2.8).

There was good agreement between SI and QUI values when examined by tertile (Fig. 2). For example, there was agreement in 29 of 35 subjects in the lower tertile, 24 of 35 middle tertile and 30 of 35 in the upper tertile. There were no gross misclassifications (i.e., lowest tertile with one method and highest tertile with the other). The kappa score between SI and QUI was 0.69 (95% CI, 0.57–0.80). There was even better agreement using the kappa scores based on the mean of duplicate SI and QUI measurements with a kappa score of 0.90 (95% CI, 0.77–0.94).

LS-BMD values were on average higher when measured by the Lunar DPX by 0.18 g/cm^2 (95% CI, 0.17–0.19) compared with the Hologic QDR 1000/W (Fig. 3). The difference between the machines was greater at higher BMD values ($r = 0.38$, $p < 0.001$).

There was good agreement between DPX and QDR values when examined by tertile (Fig. 4). For example, there was agreement in 27 of 31 subjects in the lower tertile, 24 of 32 middle tertile and 28 of 32 in the upper tertile. The kappa score between DPX and QDR was 0.79 (95% CI, 0.66–0.88).

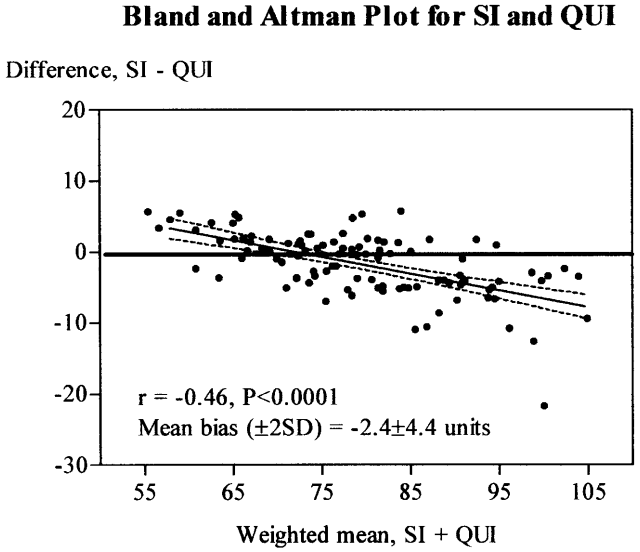
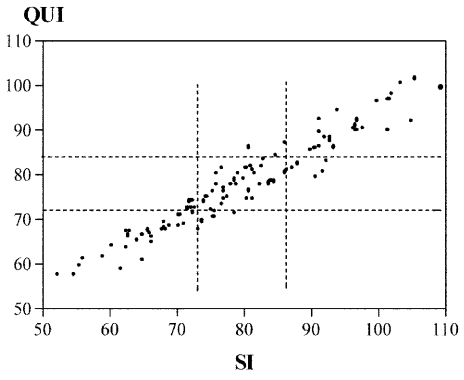


Fig. 1. Bland and Altman plot for ''stiffness index'' (SI) and ''quantitative ultrasound index'' (QUI) measured using two manufacturers' heel quantitative ultrasound (QUS) machines. Dotted lines represent the 95% CI of the slope.



QUI

U	0	5	**30**
M	6	**24**	5
L	**29**	6	0
SI	**L**	**M**	**U**

κ = 0.69
(95% CI, 0.57 to 0.80)

Fig. 2. Agreement by tertile between QUI and SI. Each cell in the frequency table (right panel) represents the data in the corresponding section of the scatter plot (left panel). The scatterplot is divided into low (L), middle (M) and upper (U) tertiles.

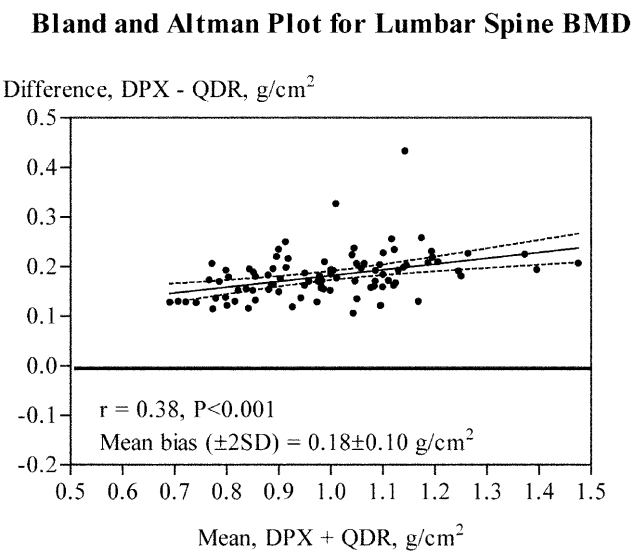
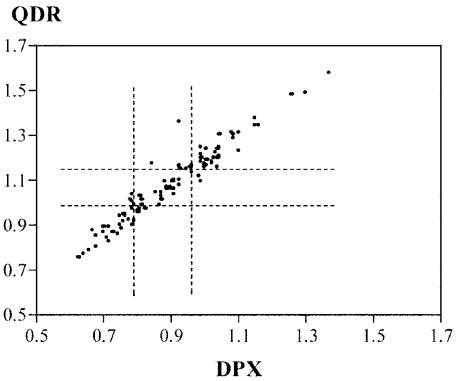


Fig. 3. Bland and Altman plot for lumbar spine bone mineral density measured using the Hologic (QDR) and Lunar (DPX) densitometers. Dotted lines represent the 95% CI of the slope.

Discussion

There is strong evidence from prospective studies that calcaneal QUS measurements can predict fracture among older postmenopausal women [12,14]. Although there is little comparative data, it appears that both BUA and SOS similarly predict fracture events [2]. Some manufacturers report a combined index utilizing BUA and SOS. This single output is used for a number of reasons: (1) it is the combined indices of the LA+ SI and HS QUI (and not BUA or SOS), which are approved by the FDA for assessing fracture risk; (2) to simplify interpretation of the BUA and SOS results; and (3), the temperature effects on combined indices are less than on BUA and SOS alone [15]. While it is generally accepted that QUS predicts hip fractures as well as DXA, it is important to know if there are differences between the QUS machines.

In order to test for agreement of SI and QUI we used the method of Bland and Altman [11] (Fig. 1). This approach is preferable to correlation between two variables because it determines how closely the two



QDR

U	0	4	**28**
M	4	**24**	4
L	**27**	4	0
DPX	**L**	**M**	**U**

κ = 0.79
(95% CI, 0.66 to 0.88)

Fig. 4. Agreement by tertile for lumbar spine bone mineral density measured on the Hologic (QDR) and Lunar (DPX) densitometers. Each cell in the frequency table (right panel) represents the data in the corresponding section of the scatter plot (left panel). The scatter plot is divided into low (L), middle (M) and upper (U) tertiles.

machines are in agreement. The difference in absolute measurement values between SI and QUI was 2.4 units. The use of correlation coefficients demonstrated a relationship between measurements made using the two machines, indicating that at higher values there is a bigger difference between SI and QUI. When we repeated the Bland and Altman analysis based on the mean of the duplicate SI and QUI the mean difference was reduced slightly to 2.0 units. This difference between SI and QUI is not surprising because the Lunar Achilles+ and Hologic Sahara QUS machines differ in their design, measurement site, coupling medium and calculation of measurement variables [7,8]. When assessing the agreement of individuals the approach is slightly different. Assigning individuals to tertiles of high, medium and low risk can assess fracture risk. The best approach to comparing agreement of classification is the kappa statistic. The kappa scores for SI and QUI showed good agreement (Fig. 2) and this improved when we calculated the kappa scores based on the average of duplicate ultrasound measurements.

There are many sources of error for QUS measurements in vivo; these include the surrounding soft tissue and foot positioning. Inter-subject variability and precision are influenced by soft tissue thickness [16], temperature [15,17] and composition [18,19] as well as the quality of the sound transmission from the coupling medium into the skin. Duplicate QUS measurements may have reduced some of these errors, consequently resulting in the improved agreement of the machines and classification of individuals.

The difference in absolute LS-BMD measurements between DPX and QDR of 18 g/cm^2 (18%) in this cohort was slightly higher than those reported by other groups [3–5]. The use of correlation coefficients demonstrated an association between measurements made using the two densitometers. The Bland and Altman plot showed that for any individual the difference was greater with the DPX than with the QDR (Fig. 3). The difference between the densitometers results from differences in calibration, correction for marrow fat and edge-detection algorithms between the different systems [6].

This study has demonstrated that there are systematic differences between heel QUS machines, resulting in about 20% of women being classified into different tertiles. However, this is not significantly greater than the misclassification rate using two different manufacturers' DXA machines to measure LS-BMD (about 15%). When the classification is based on the mean of duplicate SI and QUI measurements, it improved from 20% to 10% making them superior to a single LS-BMD measurement.

The differences between the two manufacturers' DXA machines have been well documented and are supported by this study. Despite the systematic differences between DXA machines it is considered reasonable practise to use both the Lunar DPX and Hologic QDR 1000/W to assess fracture risk. This study has established that although there are differences between the LA+ SI and HS QUI, the misclassification rate using SI and QUI is similar to LS-BMD measured on different manufacturers DXA machines. As prospective fracture studies have verified that QUS at the calcaneus can predict osteoporotic fracture as well as DXA at the spine and hip [1,2] this data would suggest that it is feasible that both the LA+ SI and HS QUI could be used in the same way as DXA (to assess fracture risk) in clinical practise. However, it is important to note that kappa scores classify women by tertile and consequently *T*-scores for the two machines may differ. *T*-scores are known to vary between measurement technique and measurement site [20,21]. This is a recognized weakness of the *T*-score approach, which is currently being addressed.

References

1. Bauer DC, Gluer CC, Cauley JA, et al. Broadband ultrasound attenuation predicts fractures strongly and independently of densitometry in older women: a prospective study. Study of Osteoporotic Fractures Research Group Arch Intern Med 1997; 157:629–34.
2. Hans D, Dargent-Molina P, Schott AM, et al. Ultrasonographic heel measurements to predict hip fracture in elderly women: the EPIDOS prospective study. Lancet 1996;348:511–14.
3. Laskey MA, Flaxman ME, Barber RW, et al. 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.
4. Pocock NA, Sambrook PN, Nguyen T, Kelly P, Freund J, Eisman JA. Assessment of spinal and femoral bone density by dual X-ray absorptiometry: comparison of Lunar and Hologic instruments J Bone Miner Res 1992; 7:1081–4.
5. Svendsen OL, Marslew U, Hassager C, Christiansen C. Measurements 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.
6. Peel NF, Eastell R. Comparison of rates of bone loss from the spine measured using two manufacturers' densitometers. J Bone Miner Res 1995; 10:1796-1801.
7. Njeh CF, Blake GM. Calcaneal quantitative ultrasound devices: water coupled. In: Njeh CF, Hans D, Fuerst TP, Gluer CC, Genant HK, editor. Quantitative ultrasound: assessment of osteoporosis and bone status. London: Martin Dunitz, 1999:109–24.
8. Cheng S, Hans D, Genant HK. Calcaneal quantitative ultrasound devices: gel coupled. In: Njeh CF, Hans D, Fuerst TP, Gluer CC, Genant HK, eds. Quantitative ultrasound: assessment of osteoporosis and bone status. London: Martin Dunitz, 1999:125–44.
9. Zagzebski JA, Rossman PJ, Mesina C, Mazess RB, Madsen EL. Ultrasound transmission measurements through the os calcis. Calcif Tissue Int 1991;49:107–11.
10. Njeh CF, Boivin CM, Langton CM. The role of ultrasound in the assessment of osteoporosis: a review. Osteoporos Int 1997;7:7–22.
11. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;i:307–10.
12. Reeve J, Lunt M. Assessing the rate of bone loss: a technique revisited. J Bone Miner Res 1999;14:1990.
13. Eastell R, Peel NFA. Rate of loss recalculated. J Bone Miner Res 1999;14:1991.
14. Porter RW, Miller CG, Grainger D, Palmer SB. Prediction of hip fracture in elderly women: a prospective study. BMJ 1990;301:638–41.
15. Pocock NA, Babichev A, Culton N, et al. Temperature dependency of quantitative ultrasound Osteoporos Int 2000;11:316–20.

16. Johansen A, Stone MD. The effect of ankle edema on bone ultrasound assessment of the heel. Osteoporos Int 1997;7:44–7.
17. Barkmann R, Heller M, Gluer CC. The influence of soft tissue and waterbath temperature on quantitative ultrasound transmission parameters: an in vivo study. Osteoporos Int 1996;6(Suppl 1).
18. Hans D, Schott AM, Arlot ME, Sornay E, Delmas PD, Meunier PJ. Influence of anthropometric parameters on ultrasound measurements of os calcis. Osteoporos Int 1995;5:371–6.
19. Barkmann R, Gluer CC. Factors influencing QUS parameters of the calcaneum: suggestions for and improved measurement procedure. J Clin Densitom 1988;1:93–4.
20. Frost ML, Blake GM, Fogelman I. Can the WHO criteria for diagnosing osteoporosis be applied to calcaneal quantitative ultrasound? Osteoporos Int 2000;11:321–30.
21. Faulkner KG, von Stetten E, Miller P. Discordance in patient classification using *T*-scores. J Clin Densitom 1999;2:343–50.

Received for publication 6 October 2000
Accepted in revised form 22 February 2001