Journal of Clinical Densitometry, vol. 10, no. 4, 365–369, 2007
© Copyright 2007 by The International Society for Clinical Densitometry
1094-6950/07/10:365–369/$32.00
DOI: 10.1016/j.jocd.2007.08.002

Original Article

A New Approach for Quantifying Change and Test Precision in Bone Densitometry

***William D. Leslie,*[*,1] *Alireza Moayyeri,*[2] *Mohsen Sadatsafavi,*[3] *and Liqun Wang*[1]**

*[1]University of Manitoba, Winnipeg, Manitoba, Canada; [2]Department of Public Health and Primary Care,
University of Cambridge, Cambridge, UK; and [3]Center for Clinical Epidemiology and Evaluation,
Vancouver Coastal Health Institute, Vancouver, British Columbia, Canada*

Abstract

The effect of precision study sample size is not considered in current recommendations for assessing bone mineral density (BMD) change. Intuitively, a larger sample size should provide greater confidence in the derived least significant change (LSC), which should translate into a more confident determination of change. We evaluated an empirical Monte Carlo simulation method for estimating the significance of an observed change in BMD that simultaneously considers the magnitude of the change, the LSC point estimate, and the precision study sample size. This method showed a progressive increase in the ability to identify BMD change using larger precision study sample sizes. Approaches that consider the error in the LSC estimate may provide more robust determinations of BMD change even when the precision study sample size is limited.

Key Words: Bone densitometry; dual-energy x-ray absorptiometry; osteoporosis; precision; sample size.

Introduction

Bone density measurements are widely used for serial monitoring of patients with suspected or confirmed osteoporosis *(1)*. Assessment of precision error in bone mineral density (BMD) testing is a prerequisite to characterizing longitudinal changes *(2)*. The International Society for Clinical Densitometry (ISCD) has a standardized methodology for conducting an in vivo precision study and recommends that this be performed by each densitometry center *(2,3)*. The ISCD procedure states that precision error should be obtained from an assessment with 30 degrees of freedom (df; e.g., 30 individuals with 2 scans each or 15 individuals with 3 scans each) drawn from the patient referral population and using the root mean square (RMS) approach.

A sample size of 30 df ensures that the upper limit for the 95% confidence interval (CI) of the calculated precision value is no more than 34% greater than the calculated precision value *(2)*. It is not widely appreciated that this degree of variation in the precision value can lead to considerable inconsistency (20% or greater) in classifying change when applied to patients undergoing BMD monitoring *(4)*. This relates to the fact that a conventional precision study only provides a point estimate of the precision error. The error in the precision estimate, which is a function of the sample size df, is not explicitly considered in current approaches for assessing BMD change. Intuitively, a larger sample size should provide greater confidence in the precision estimate and derived least significant change (LSC), which should in turn translate into a more confident determination of change. Conversely, if the observed change in a patient's BMD exceeds the LSC point estimate but still falls within the 95% confidence interval for the LSC, then it may not be possible to conclude that the patient's BMD has truly changed.

Methods

Theoretical Considerations

We developed an empirical method for estimating the significance of an observed change in a patient's BMD that

Received 07/05/07; Revised 08/06/07; Accepted 08/06/07.

*Address correspondence to: Dr. William D. Leslie, MD, MSc, Department of Medicine (C5121), University of Manitoba, 409 Tache Avenue, Winnipeg, Manitoba R2H 2A6, Canada. E-mail: bleslie@sbgh.mb.ca

simultaneously considered the observed magnitude of the change, the RMS error, and LSC as estimated by the ISCD procedure (denoted the observed RMS error and LSC), and the precision study sample size (df). Monte Carlo simulation was used to define a continuous metric of confidence that change had occurred, which is referred to as the sample size responsive *p* value (SSR-*p*). Assumptions underlying this model are that: (1) measurement error is normally distributed; (2) short-term precision error can be used to approximate long-term precision error (no baseline or machine drift); and (3) that there is equal measurement error for all patients (no spectrum bias). These assumptions are also fundamental to the ISCD procedure *(2,3)* and are consistent with empirical data *(5,6)*.

To implement this method, we simulated the results that would be obtained if 10,000 independent precision assessments had been performed and used to categorize the observed BMD change in a patient. When there is a consistent designation of change using these 10,000 LSC classifiers, then one can have a high degree of confidence that the change is real. To simulate 10,000 LSC estimates based on 30 df, it is necessary to generate 300,000 simulated scan-pair errors. These 300,000 measurements are then grouped into independent subsets of 30, which are then used to derive 10,000 independent LSC estimates using the ISCD procedure. The measurement error for each simulated scan-pair is generated as a random normal variate with a mean of 0 and standard deviation equal to the observed RMS error from the precision study. To determine if the observed change in patient BMD is significant, ΔBMD is compared with the 10,000 LSC estimates to determine the proportion in which ΔBMD exceeds the LSC. The larger this fraction, the more likely that the observed patient change is not due to measurement error. SSR-*p* is the fraction of the cases in which the LSC exceeds ΔBMD. When this fraction is <0.05, then it can be interpreted as an empirically derived *p* value consistent with significant patient change at the 95% confidence level. Computational details are given in the Appendix.

This procedure was implemented in Excel (Windows XP Version, Microsoft Inc.). The calculations can be readily integrated with the ISCD Bone Densitometry Precision Calculating Tool (www.iscd.org). Although the computation time is relatively slow as described in the Appendix (5−10 min), the spreadsheet only needs to be run once to fully analyze the data in terms of all possible BMD cutoffs and sample sizes. There are additional computational shortcuts that can be implemented that reduce the program's running time. For example, the simulated LSCs can be directly modeled by fitting a Chi-square distribution to the variance of the simulated precision studies. By using this modification and limiting the spreadsheet to a single sample size df, the run time is less than 1 min.

These principles are illustrated in Fig. 1 which looks at a hypothetical patient with an observed change in BMD of 0.060 g/cm^2 for a facility with LSC 0.048 g/cm^2. In this example, the change is significant (SSR-$p < 0.05$) if the precision study sample size is 100 df or larger, but not significant (SSR-$p > 0.05$) if the precision study sample size is 50 df or smaller. Precision study sample size was varied from 10 df to 500 df, using 10,000 simulated LSC values to determine the SSR-*p* for each monitored patient.

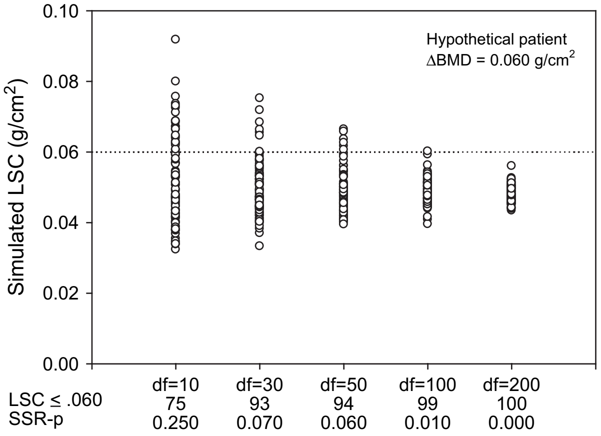

Fig. 1. Sample calculations (100 simulation runs, df from 10 to 200) for a patient with an observed change in BMD of 0.060 g/cm^2 and facility LSC of 0.048 g/cm^2. If the LSC is based on a precision assessment of only 10 df, then there is uncertainty in the LSC estimate with most (75 of 100) simulated LSC values falling below the observed patient change of 0.060 g/cm^2 but still a substantial number (25 of 100) falling above the observed patient change. It is not possible to conclude (with 95% confidence) that the patient's BMD has changed more than the LSC (SSR-$p = 25/100 = 0.250$). With 30 df, only 7 of 100 simulated LSC values exceed the observed patient change (SSR-$p = 0.070$) which would be considered a statistically borderline result. With 200 df, none of the simulated LSC values exceed the observed patient change (SSR-$p = 0.000$).

Study Populations

We evaluated our approach using a large "clinical monitoring population" and "precision population." These populations have been previously described in detail *(5)*. Data from the Manitoba Bone Density Program were used for the analyses and the current report was approved by the facility's Office for Clinical Research. The Program provides all bone density services to the population of Province of Manitoba, Canada (total population 1,119,583 according to 2001 Statistics Canada census data) *(7)*. The Program maintains an electronic database of all dual-energy X-ray absorptiometry (DXA) bone density tests performed because DXA testing was first offered in 1990 *(8)*. We identified all individuals who had baseline and follow-up BMD measurements on the same instrument between 1994 and 2002. We excluded cases where scanning was performed on different instruments and those that did not report the lumbar spine and the total hip BMD. This left 1420 scan-pairs for the clinical monitoring population.

Replicate measurements of the spine and hip were obtained from a convenience sample of female individuals referred for bone density testing who were agreeable to undergo a repeat assessment, usually on a separate day (mean interval between scans 6 ± 5 d) and by 2 different technologists. The RMS method was used to calculate the standard deviation of the precision error (units of g/cm^2). The 95% confidence LSC was defined for each RMS precision error by multiplying by 2.77. Absolute LSC (units of g/cm^2) was used for assessing significance in the absolute change between 2 BMD measurements. The precision population consisted of 198 spine scan-pairs and 193 hip scan-pairs.

BMD Measurements

The DXA scans in our densitometry clinics are performed and analyzed in accordance with the manufacturer's recommendations. All densitometers underwent daily assessment of stability using an anthropomorphic spine phantom and each showed excellent long-term phantom stability (coefficient of variation < 0.5%). A pencil-beam instrument (Lunar DPX, GE Lunar, Madison, WI) was the primary instrument used before 2000 and a fan-beam instrument (Lunar Prodigy, GE Lunar, Madison, WI) was used after that date. Precision error calculated separately for these scanner configurations did not demonstrate any significant difference as assessed by the *F*-ratio test, and therefore results were pooled.

Data Analysis

The performance of our approach was compared with that of the current ICSD procedure in the clinical monitoring population. We simulated precision studies across a wide range of sample size df (df = 10−500). For each sample size, we derived the smallest change in BMD required to indicate a significant change using the new method (SSR-$p < 0.05$). The proportion of the clinical monitoring population having significant change based on this BMD cutoff was then calculated. This proportion was then compared with the proportion of subjects having change using the LSC point estimate based on the ICSD procedure. The smallest change in BMD required to achieve significance in our method is always higher than LSC point estimate. Therefore, the difference between the 2 proportions reflects the number of subjects with significant change according to the ICSD procedure but not significant when precision study sample size limitations are considered.

Results

The characteristics of the separate precision and clinical monitoring populations are summarized in Table 1. BMD change for the former was close to 0 whereas the latter showed an average increase for both the lumbar spine and total hip. The lumbar spine LSC point estimate derived from the 198 spine-pairs was 0.048 g/cm^2. When this cutoff was applied to the clinical monitoring population, 30.7% were found to have absolute change in lumbar spine BMD that exceeded this value. The total hip LSC point estimate from the 193 hip-pairs was 0.026 g/cm^2, and classified 40.1% of the clinical monitoring population as showing significant change.

Table 1
Characteristics of the Precision Study and Clinical Monitoring Populations

Characteristic	Precision	Clinical monitoring
Age at baseline (yr)	54 ± 10	56 ± 14
Female gender (%)	198 (100)	1220 (86)
Lumbar spine		
n	198	1420
Baseline BMD (g/cm^2)	1.085 ± 0.174	0.983 ± 0.178
Change (g/cm^2)	0.001 ± 0.025	0.016 ± 0.050
Total hip		
n	193	1420
Baseline BMD (g/cm^2)	0.920 ± 0.133	0.812 ± 0.136
Change (g/cm^2)	−0.002 ± 0.013	0.007 ± 0.038

Compared with the LSC point estimate derived from all 198 spine-pairs, individual simulated LSC values from 30 df showed inconsistent patient categorization in spine BMD change (95% CI from 8.1% underestimation to 13.5% overestimation of the proportion of subjects with significant change). For 100 df, these values were substantially smaller (5.3% and 7.1%, respectively). The SSR-*p* showed a progressive increase in the ability to identify BMD change using larger precision study sample sizes (Table 2). A sample size of 100 df was needed to give results within 5% of the reference value.

Fig. 2 illustrates how confidence in identifying change from the SSR-*p* (confidence = 100% × (1 − SSR-*p*)) varies as a function of precision study sample size and observed change in BMD. Each precision study sample size defines a confidence curve based on the observed change in BMD, and with larger sample sizes, the curves fall upward and to the left indicating a higher level of confidence for the same observed change in BMD. With the largest sample size in the figure (200 df), the curve approaches but does not quite reach the case where the LSC point estimate is assumed to be without error.

To further illustrate this procedure with an example, again consider the case where the observed change in spine BMD for a patient is 0.060 g/cm^2 which would be considered a significant change using the conventional LSC point estimate. Of the 10,000 simulated spine LSCs generated using 30 df, 9350 were less than 0.060 g/cm^2 for an empirically derived SSR-*p* of 0.0635 which fails to achieve the criterion for significant change but suggests a trend in that direction. For a precision study sample size of 100, 9977 of the 10,000 simulated spine LSCs were less than 0.060 g/cm^2 for an empirically derived SSR-*p* of 0.0023 which would be classified as significant change. For 30 df, any observed change in spine BMD larger than 0.062 g/cm^2 would have an SSR-*p* below 0.05, whereas for 100 df this value is 0.055 g/cm^2 and for 10 df it is 0.077 g/cm^2.

Table 2
Proportion of the Clinical Monitoring Population Classified as Showing BMD Change Using the Reference LSC (from the Precision Population Using the ISCD Method) or Where the Cutoff BMD was Defined Using the SSR-$p < 0.05$ for Precision Study Sample Sizes from 10 to 500 df

	Lumbar spine			Total hip		
Sample size (df)	BMD cutoff for change (g/cm^2)	Proportion with change exceeding cutoff (%)	Difference from reference method (%)	BMD cutoff for change (g/cm^2)	Proportion with change exceeding cutoff (%)	Difference from reference method (%)
Reference LSC	0.048	30.7	—	0.026	40.1	—
10	0.076	12.4	−18.3	0.041	21.4	−18.7
20	0.065	17.2	−13.5	0.035	27.5	−12.6
30	0.061	19.2	−11.5	0.033	29.8	−10.3
40	0.058	20.7	−10.0	0.032	31.1	−9.0
50	0.057	22.5	−8.2	0.031	32.9	−7.3
75	0.055	24.9	−5.8	0.030	34.6	−5.5
100	0.054	25.6	−5.1	0.029	36.5	−3.7
125	0.053	26.5	−4.2	0.029	36.5	−3.7
150	0.053	26.5	−4.2	0.028	37.7	−2.4
175	0.052	27.3	−3.4	0.028	37.7	−2.4
200	0.052	27.3	−3.4	0.028	37.7	−2.4
300	0.051	28.1	−2.6	0.027	38.7	−1.4
400	0.051	28.1	−2.6	0.027	38.7	−1.4
500	0.050	28.9	−1.9	0.041	38.7	−1.4

Discussion

We have presented an empirical method for assessing BMD change that extends the ISCD procedure by also considering the effect that the precision study sample size has on confidence in classifying change. Smaller precision study sample sizes classify significantly fewer individuals as showing change when compared with larger sample sizes, reflecting the reduced confidence in LSCs estimated from a smaller population. With precision study sample sizes of approximately 100 df, the proposed method is within 5% of the ISCD reference method. Larger sample sizes gave greater confidence for a given observed change in patient BMD.

The proposed method is more conservative in terms of classifying change than the ISCD procedure, especially for smaller precision study sample sizes, because an observed BMD change (ΔBMD) that is only slightly greater than the LSC point estimate will be associated with a substantial number (more than 5%) of simulated LSC values that exceed ΔBMD. Falsely identifying BMD change, especially a decrease, could be deleterious in patient care because it may lead to initiation of treatment that is unnecessary, or to the incorrect conclusion that a treatment has failed. Although failing to detect a decrease in BMD that is just outside of the LSC limit is not inconsequential, the reality is that ongoing BMD loss will be detected with additional follow-up measurements.

Our approach is empirical and could be criticized for the lack of a mathematically rigorous theoretical framework. However, it is more complete than the current ISCD procedure which does not directly consider the effect of precision study sample size. More formal model-based methods to estimate response rate in populations have been proposed, but these cannot currently be used to determine if a particular patient has an increase in BMD over a period of therapy *(9)*.

A limitation of our method is the assumption that precision errors are normally distributed, though this underlies the entire LSC approach as advocated by others *(1−3,10,11)*. One group has reported that outlying residuals can give rise to statistically significant results for kurtosis and has proposed a method for trimming outliers to better fit the Gaussian function *(6)*. Patient weight or body mass index (BMI) have also been reported to affect BMD precision *(12)*, though we did not find this to be the case in our precision population. Finally, we did not have access to treatment information on our clinical monitoring population. Current recommendations for assessing change in BMD do not distinguish treated from untreated patients, because it is the absolute change in BMD rather than the direction of change that is important.

We believe that the effect of precision study sample size on classifying change in monitored patients is an important element of the precision assessment that is neglected in current recommendations. Sample sizes larger than 30 df are required if low levels of categorization error are to be achieved when using the ISCD procedure. There is always a trade-off between the desire for greater statistical power, feasibility, and resource constraints. Although statistical confidence in a precision estimate increases with larger sample sizes, there are practical limits to what can realistically be achieved.

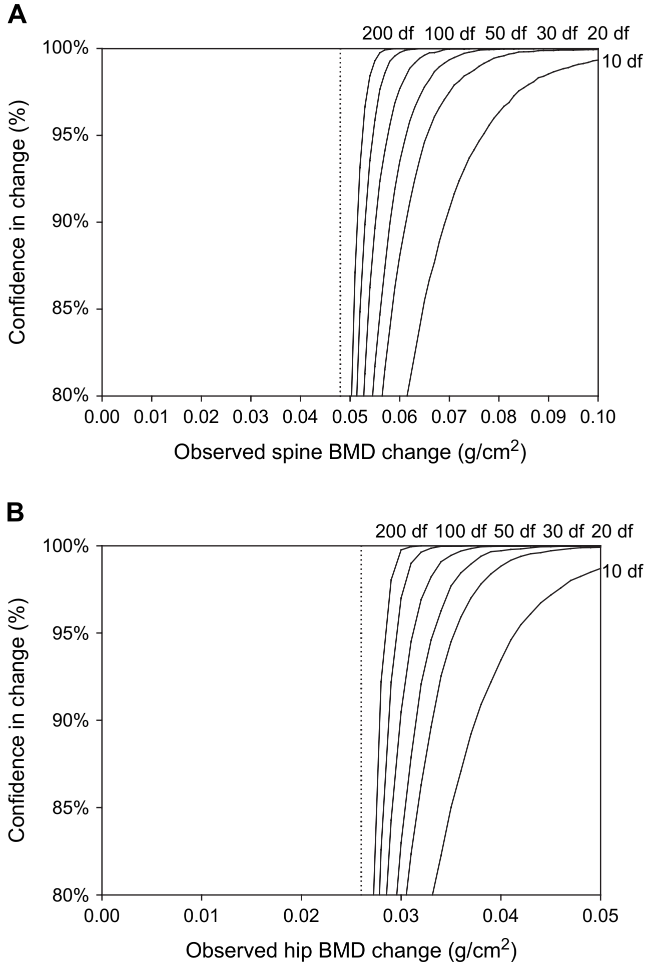

Fig. 2. Effect of sample size (10−200 df) on confidence in identifying change based on the observed absolute BMD change during clinical patient monitoring. The dotted line indicates the point estimate for the LSC. (**A**). Spine L1−L4 (observed LSC 0.048 g/cm^2). (**B**). Total hip (observed LSC 0.026 g/cm^2).

Approaches that consider the error in the LSC estimate, such as the one presented here, may provide more robust determinations of BMD change even when the precision study sample size is limited.

Acknowledgments

This article has been reviewed and approved by the members of the Manitoba Bone Density Program Committee. The authors and Committee would like to express their gratitude to Manitoba Health, the Winnipeg Regional Health Authority, and the Brandon Regional Health Authority for their vision, trust, and support in the establishment of this Program.

Appendix

The SSR-*p* is estimated using the following procedure. Given an observed RMS precision error for sample size (df) n, denoted $\mathrm{RMS}_{\mathrm{Obs,n}}$, and observed patient change in BMD, denoted ΔBMD:

(1) generate n independent and identically distributed random values from a normal distribution defined by Normal(0, $\mathrm{RMS}_{\mathrm{Obs,n}}$)
(2) compute $\mathrm{RMS\text{-}SD}_{\mathrm{sim,n}}=\sqrt{\left(\sum X_i^2/\mathrm{n}\right)}$
(3) compute $\mathrm{LSC}_{\mathrm{sim,n}}=2.77\times\mathrm{RMS}_{\mathrm{sim,n}}$
(4) define the change function,

$$C\left(\Delta\mathrm{BMD},\mathrm{LSC}_{\mathrm{sim,n}}\right)=\left\{\begin{matrix}1 & \text{if } \Delta\mathrm{BMD}\geq\mathrm{LSC}_{\mathrm{sim,n}}, \\ 0 & \text{otherwise}\end{matrix}\right\}$$

(5) repeat (1)−(4) 10,000 times to generate 10,000 different values for $\mathrm{RMS}_{\mathrm{sim,n}}$ and $\mathrm{LSC}_{\mathrm{sim,n}}$, and the related change function, $C(\Delta\mathrm{BMD},\mathrm{LSC}_{\mathrm{sim,n}})$ functions
(6) compute the sample size responsive *p* value as:

$$\begin{aligned}\text{SSR-}p&=\Pr(\mathrm{LSC}_{\mathrm{sim,n}}>\Delta\mathrm{BMD})\\&=1-\left(\sum C(\Delta\mathrm{BMD},\mathrm{LSC}_{\mathrm{sim,n}})\right)/\ 10,000\end{aligned}$$

For 95% confidence, a value for SSR-$p\leq0.05$ is interpreted as significant change in ΔBMD. The confidence levels for interpreting the SSR-*p* and for generating the LSC are inherently arbitrary, and can be modified to be either more restrictive or more liberal in classifying change.

References

1. Lenchik L, Kiebzak GM, Blunt BA. 2002 What is the role of serial bone mineral density measurements in patient management? J Clin Densitom 5 Suppl:S29−S38.
2. Bonnick SL, Johnston CC Jr, Kleerekoper M, et al. 2001 Importance of precision in bone density measurements. J Clin Densitom 4:105−110.
3. Baim S, Wilson CR, Lewiecki EM, Luckey MM, Downs RW Jr, Lentle BC. 2005 Precision assessment and radiation safety for dual-energy X-ray absorptiometry: position paper of the International Society for Clinical Densitometry. J Clin Densitom 8:371−378.
4. Leslie WD, Moayyeri A. 2006 Minimum sample size requirements for bone density precision assessment produce inconsistency in clinical monitoring. Osteoporos Int 17:1673−1680.
5. Leslie WD. 2006 The importance of spectrum bias on bone density monitoring in clinical practice. Bone 39:361−368.
6. Patel R, Blake GM, Rymer J, Fogelman I. 2000 Long-term precision of DXA scanning assessed over seven years in forty postmenopausal women. Osteoporos Int 11:68−75.
7. Leslie WD, Metge C. 2003 Establishing a regional bone density program: lessons from the Manitoba experience. J Clin Densitom 6:275−282.
8. Leslie WD, Caetano PA, MacWilliam LR, Finlayson GS. 2005 Construction and validation of a population-based bone densitometry database. J Clin Densitom 8:25−30.
9. Qu Y, Kulkarni PM, Sanger TM. 2007 Estimating the response rate in the presence of measurement error. Stat Med 26:197−211.
10. Gluer CC, Blake G, Lu Y, Blunt BA, Jergas M, Genant HK. 1995 Accurate assessment of precision errors: how to measure the reproducibility of bone densitometry techniques. Osteoporos Int 5: 262−270.
11. Gluer CC. 1999 Monitoring skeletal changes by radiological techniques. J Bone Miner Res 14:1952−1962.
12. Patel R, Blake GM, Herd RJ, Fogelman I. 1997 The effect of weight change on DXA scans in a 2-year trial of etidronate therapy. Calcif Tissue Int 61:393−399.