Osteoporosis Int (1995) 5:262–270
© 1995 European Foundation for Osteoporosis

Osteoporosis
International

Original Article

Accurate Assessment of Precision Errors: How to Measure the Reproducibility of Bone Densitometry Techniques

C.-C. Glüer[1], G. Blake[2], Y. Lu[1], B. A. Blunt[1], M. Jergas[1] and H. K. Genant[1]

[1]Department of Radiology, University of California, San Francisco, USA; and [2]Department of Nuclear Medicine, Guy's Hospital, London, UK

Abstract. Assessment of precision errors in bone mineral densitometry is important for characterization of a technique's ability to detect logitudinal skeletal changes. Short-term and long-term precision errors should be calculated as root-mean-square (RMS) averages of standard deviations of repeated measurements (SD) and standard errors of the estimate of changes in bone density with time (SEE), respectively. Inadequate adjustment for degrees of freedom and use of arithmetic means instead of RMS averages may cause underestimation of true imprecision by up to 41% and 25% (for duplicate measurements), respectively. Calculation of confidence intervals of precision errors based on the number of repeated measurements and the number of subjects assessed serves to characterize limitations of precision error assessments. Provided that precision error are comparable across subjects, examinations with a total of 27 degrees of freedom result in an upper 90% confidence limit of +30% of the mean precision error, a level considered sufficient for characterizing technique imprecision. We recommend three (or four) repeated measurements per individual in a subject group of at least 14 individuals to characterize short-term (or long-term) precision of a technique.

Keywords: Bone densitometry; Heteroscedasticity; Precision error; Reproducibility; Statistics

Introduction

Normal changes of the mineral content of skeletal tissue proceed at a relatively slow pace ranging from 0.5–2% per annum for most of the adult lifespan of healthy individuals to 2–5% in early postmenopausal women [1–5]. The upper portion of these ranges reflects changes of trabecular bone as assessed by quantitative computed tomography (QCT), reflecting a high responsiveness of this technique to change in bone mineral density (BMD). The lower portion is more typical for changes of cortical or integral (i.e. cortical plus trabecular) bone as assessed by projection-type techniques such as dual-energy X-ray absorptiometry (DXA) or single photon absorptiometry (SPA).

To detect changes of small magnitude, bone densitometry techniques with very high reproducibility had to be developed. State-of-the-art approaches for monitoring the progression of disease or the efficacy of treatment, such as DXA of the spine or the femoral neck or QCT of the spine, have been reported to achieve reproducible results in vivo within approximately ±1%, 1.5% and 2–3%, respectively [6–9]. Recently, newer approaches such as lateral DXA, peripheral QCT (pQCT) and quantitative ultrasound (QUS) have been introduced and, except for pQCT, with precision errors of 0.5–1% [10,11], the reported reproducibility errors for these techniques generally have been similar to or larger than those of the established bone densitometry approaches [12,13]. For judging a technique's ability to monitor changes in BMD, agreement on how to measure and calculate reproducibility is required. Precision errors have been used to characterize reproducibility, but the applied methodology has been inconsistent or ill defined.

Correspondence and offprint requests to: Dr. Claus-C. Glüer, Osteoporosis Research Group, University of California, San Francisco, CA 94143–0628, USA.

In this article we propose and discuss a concept on how to measure, calculate and report precision errors. Some of the concepts presented are fairly basic (albeit apparently not common knowledge) but we feel obliged to include them to present a coherent framework. We will address the following questions:

What is an appropriate definition of precision errors in the individual subject?
The definition should reflect differences of short- versus long-term precision and be applicable to both patient and phantom studies. The outcome may depend on the subject group but it should be unbiased, i.e. reflect the true population mean independent of the number of repeated measurements. Secondary criteria include efficiency (i.e. having narrow confidence intervals even when estimated from a relatively small subject group), and robustness with respect to the shape of the distribution of the precision errors in the subject group.

How should the precision of a technique be computed?
Precision errors measured in individual subjects need to be pooled to obtain a statistic that appropriately describes the precision of a technique. Expressing precision errors in absolute units or on a percentage basis requires use of different mathematical concepts.

How many measurements are required for a reliable characterization of precision errors?
Calculation of confidence intervals of precision errors allows one to judge the significance of the reported precision errors. We present formulae that allow one to calculate the numbers of measurements and subjects required to obtain precision errors with 'adequately narrow' confidence limits.

Methods

The Concept of 'Precision'

Precision errors have been defined to characterize the reproducibility of a diagnostic technique. *Accuracy errors* (here used as equivalent to the term bias), on the other hand, reflect the degree to which the measured results deviate from the 'true' values. The example that has frequently been presented to illustrate the difference between precision and accuracy is that of the performance of archers. If an archer consistently hits the target board close to the bull's-eye, but with the arrows spread out around it, this would be regarded as good accuracy but poor precision. If he consistently hits the board far off the bull's-eye, but with all of his arrows of approximately the same location, we would speak of poor accuracy but good precision. Similarly, the errors of repeated BMD assessments for a specific technique and a given subject can be characterized by the difference of the true versus the mean measured BMD (i.e. the accuracy error), and the spread of the individual readings around the mean measured BMD (i.e. the precision error).

Even for a given technique the precision error may vary from patient to patient (e.g. it is usually higher in osteoporotic patients than in normal subjects). Therefore, it may be misleading to determine the technique's precision by measuring just normal subjects (analogous to having just the better archers compete) or by measuring only a small number of patients (which would probably either over- or underestimate the technique's performance). The calculation of a confidence interval of the measured precision error will tell how many subjects and repeated measurements are needed to achieve a preset goal of exactness of this measure. Furthermore, even if the sample of the subjects is representative of the typical study population great care has to be taken to apply the correct statistical concept to characterize the overall precision of the technique. As we will demonstrate, simple averaging of the individuals' precision errors as has been done in many publications is inadequate.

Finally, precision errors also depend on the time interval that elapsed between the repeated measurements. Generally, short-term precision errors (measurements performed on the same day) are considerably smaller than long-term precision errors. In fact, since true changes in BMD can be expected to occur over longer periods of time short-term precision errors and long-term precision errors require different mathematical definitions which will be given in the following section.

Definition of 'Precision'

Short-Term Precision of an Individual Subject. Assuming that the random variations of repeated measurements in an individual are normally distributed, precision is represented by the estimate of the parameter σ in the Gaussian probability distribution. Short-term precision (SD) is then defined as the standard deviation of $i=1 \ldots n_j$ repeated measurements on a given subject j:

$$\mathrm{SD}_j = \sqrt{\sum_{i=1}^{n_j} \frac{(x_{ij} - \bar{x}_j)^2}{n_j - 1}} \qquad (1)$$

where n_j is the number of measurements performed, x_{ij} is the result of the ith measurement for subject j, and $\bar{x}_j$ is the mean of all x_{ij} for this subject j. The patient should be repositioned between measurements to include this source of reproducibility error unless machine imprecision is investigated. Since the true mean of the measurements is unknown and has to be estimated from the mean of the n repeated measurements, the denominator has to be represented as $(n_j - 1)$ in order to make SD^2 an unbiased estimate of the parameter σ^2 in the Gaussian probability distribution. This adjustment insures that SD is independent of the number of repeated measurements. The denominator $(n_j - 1)$, i.e. the number of repeat measurements minus one, is the number of degrees of freedom, df_j associated with this estimate.

Precision errors may be expressed in absolute numbers, or as coefficient of variation (CV) of repeated measurements, typically given on a percentage basis:

$$\mathrm{CV}_{\mathrm{SD}_j} = \frac{\mathrm{SD}_j}{\bar{x}_j} \cdot 100\% \tag{2}$$

where $\bar{x}_j$ is the mean of all x_{ij}.

Short-Term Precision of a Technique. As noted above, measurements of precision on a single subject may not be representative of the performance of the technique in general. A representative group of subjects needs to be assessed (for a comprehensive characterization it may even be necessary to calculate separate precision errors for different groups of patients, e.g. normals versus osteoporotics). Consequently, the question arises as to how to pool precision data obtained on several subjects. Contrary to intuition and common practice, the correct estimate of a technique's precision error is *not* given by the (arithmetic) mean of the individual subject's precision errors [14]. Instead, the technique's squared precision error SD^2 (i.e. the variance) is given by the arithmetic mean of the individual subject's $j = 1 \ldots m$ variances SD_j^2 [14]:

$$\mathrm{SD}^2 = \sum_{j=1}^{m} \mathrm{SD}_j^2/m, \text{ noting that } \mathrm{SD} \neq \sum_{j=1}^{m} \mathrm{SD}_j/m \tag{3}$$

Statistically speaking, the reason is that the measured variance SD^2, but not the measured standard deviation, can be considered an unbiased estimate of the parameter σ^2 of the Gaussian normal distribution. Therefore, only the former can be averaged arithmetically. Consequently, the technique's precision error is given by the root-mean-square (RMS) average of the precision errors calculated by Eq. 1 for each of the m subjects:

$$\mathrm{SD} = \sqrt{\sum_{j=1}^{m} \mathrm{SD}_j^2/m} \tag{4a}$$

which for duplicate measurements on each subject (demonstrating a difference d_j between the first and the second result) is equivalent to

$$\mathrm{SD} = \sqrt{\sum_{j=1}^{m} d_j^2/2m} \tag{4b}$$

When expressing precision on a percentage basis we propose to use the following formula:

$$\mathrm{CV}_{\mathrm{SD}} = \left(SD/\sum_{j=1}^{m} \bar{x}_j/m \right) \cdot 100\% \tag{5}$$

The missing subscript index j on the left-hand side of the equations indicates that the data are based on an average obtained on a group of patients. Alternatively, $\mathrm{CV}_{\mathrm{SD}}$ could be calculated according to the formula

$$\mathrm{CV}_{\mathrm{SD}} = \sqrt{\sum_{j=1}^{m} \mathrm{CV}_j^2/m}$$

i.e. by first calculating the individual CVSD$_j$ and then taking the RMS average. This estimate will produce slightly larger results due to the fact that $1/\bar{x}_j^2$ is not an unbiased estimator of $1/\mu^2$ (μ is the population mean). However, as can be shown by simulation the magnitude of this difference is negligible for practical densitometry purposes.

Equations 3–5 are strictly valid only if the number of repeated measurements per patient is identical for all patients. If this is not the case, the following generic formula needs to be applied:

$$\mathrm{SD} = \sqrt{\sum_{j=1}^{m} \sum_{i=1}^{n_j} \frac{(x_{ij} - \bar{x}_j)^2}{\mathrm{df}}} \tag{6}$$

The denominator is the total number of degrees of freedom (df) for the estimate of standard deviation formed by combining all the data. For a technique it is simply the sum of the degrees of freedom df_j of the measurements in the individuals:

$$\mathrm{df} = \sum_{j=1}^{m} \mathrm{df}_j = \sum_{j=1}^{m} (\mathrm{n}_j - 1) \tag{7}$$

Equation 5 as an estimate of the pooled variance is valid if the subjects have comparable precision errors (compare, e.g., chapters 8–3 in [15] or [16]). Equations 3–5 follow if $n_j = n$, i.e. if the number of repeat measurements is constant across subjects and, hence, $\mathrm{df} = m \cdot (n - 1)$.

Confidence Intervals of Short-Term Precision of a Technique. Understanding the need to obtain precision data on a group of subjects leads to the questions: How many subjects and how many repeat measurements per subject would be sufficient to characterize the performance of a technique accurately? These questions can be answered by calculation of confidence interval of precision. Large confidence intervals would indicate insufficient numbers of subjects and/or numbers of repeat measurements.

Contrary to intuition, the correct estimate of the 95% confidence interval of the technique's precision error is *not* given by ±2 times the observed standard deviation of all of the individual subjects' precision errors. It is apparent that this would be inappropriate since the confidence interval needs to be asymmetric, reflecting that precision errors while having no upper limit cannot become negative. Instead, the confidence interval of variances are commonly calculated using the (asymmetric) chi-square (χ^2) distribution. This distribution depends on the total number of degrees of freedom (df) formed by combining all the data from all of the subjects given by Eq. The $(1 - \alpha) \cdot 100\%$ confidence intervals of the true precision error σ is given by (chapter 12.2 in [17]):

$$\frac{df}{\chi^2_{1-\frac{\alpha}{2},df}} \mathrm{SD}^2 < \sigma^2 < \frac{df}{\chi^2_{\frac{\alpha}{2},df}} SD^2 \tag{8}$$

where χ^2 (df) is the chi-square distribution with df degrees of freedom; it is tabulated in many statistics textbooks (e.g. [18]) and df is calculated by way of Eq.

7. Note that alternatively the confidence limit could be calculated with identical results using the F-distribution (with $\mathrm{SD}^2 \cdot F_{\frac{\alpha}{2}}(\infty, df) < \sigma^2 < \mathrm{SD}^2 F_{1-\frac{\alpha}{2}}(\infty, df)$). Equation 8 can be used to calculate the sample size required to specify precision errors with an upper confidence limit not exceeding a given level – the lower confidence limit being of less practical relevance.

Equataion 8 specifies the confidence interval for the precision error σ of the technique – not to be confused with the quantiles of the distribution of the subjects' individual precision errors SD_j that typically are much wider and given by:

$$\frac{\sigma^2 \chi^2_{\frac{\alpha}{2}, df_j}}{df_j} < \mathrm{SD}_j^2 < \frac{\sigma^2 \chi^2_{1-\frac{\alpha}{2}, df_j}}{df_j} \tag{9}$$

where σ^2 can be estimated by SD^2. Equation 8 is the χ^2-analogon to two times the standard error of the mean of a normal distribution; Eq. 9, on the other hand, represents the χ^2-equivalent of two times the standard deviation of individual data points.

If true precision errors vary substantially between subjects (this situation is termed "heteroscedasticity") the confidence interval based on the χ^2-distribution (Eq. 8) will be conceptually inappropriate. Moreover, the mean precision error as calculated from Eqs 3–5 will also be less meaningful since a single summary measure may be too simplistic for this situation. How can one determine whether the observed distribution of precision errors is compatible with the assumption of a common true precision error? One suggested procedure would be an ANOVA of BMD values versus subject IDs. Bartlett's test for equal variance among subjects is available in commercial statistics packages (e.g. SAS, JMP from SAS Institute, Cary, NC; or SPSS from SPSS Inc., Chicago, IL). Bartlett's test assumes a normal distribution of the underlying data, which in our case of BMD measurements appears reasonable. If this assumption is not fulfilled O'Brien's test would be applicable instead (available also e.g. in SAS, JMP).

If unequal variances among individuals are suspected it is strongly recommended to examine the sources of heteroscedasticity, which could include technique specific, operator specific, machine specific, or subject-group specific factors. To test this the aformentioned ANOVA procedure could be carried out for BMD* versus subject group (grouped by error source, with BMD* representing a BMD value that has been adjusted for differences between individuals). For example, if the subject group included both healthy as well as osteoporotic indviduals a test for unequal variances between those groups would be advisable since for most densitometric techniques osteoporotic subjects generally tend to have higher precision errors. If heteroscedasticity is observed precision error analysis results should be reported separately for each subject group.

Long-Term Precision of an Individual Subject. The assessment of long-term reproducibility is complex since the variability of the data may be due to imprecision of the technique as well as to true changes in the mineral density. Even for phantom measurements where variability of true mineral content does not occur, systematic long-term changes (e.g. due to scanner drift, recalibrations) may be encountered. Applying Eq. 1 for expression of long-term precision in vivo would therefore result in an overestimation of the precision errors of the technique.

A parameter that quantifies variability due to reasons other than (true) linear changes is available from regression analysis. When repeated measurements are taken on the same subject over time, the variability about the regression curve (i.e. the *standard error of the estimate* (*SEE*) or *root mean square* (*RMS*) *error*) is taken as an estimate of that person's long-term precision error. It does, however, still include variability due to non-linear loss of bone.

Suppose we have a set of data from a long-term precision study in which y_{ij} is the ith measurement on a subject j and $\hat{y}_{ij}$ is the corresponding result predicted from the regression model, then the long-term precision error of $i = 1 \ldots n_j$ repeated measurements is given by:

$$\mathrm{SEE}_j = \sqrt{\frac{\sum_{i=1}^{n_j} (y_{ij} - \hat{y}_{ij})^2}{n_j - 2}} \tag{10}$$

or when expressed on a percentage basis:

$$\mathrm{CV}_{\mathrm{SEE}_j} = \mathrm{SEE}_j / \bar{y}_j \times 100 \tag{11}$$

where $\bar{y}_j$ is the mean of all y_{ij} of a given subject j and the suffix is used to differentiate $\mathrm{CV}_{\mathrm{SEE}_j}$ from the $\mathrm{CV}_{\mathrm{SD}_j}$ defined in Eq. 2.

The denominator $(n_j - 2)$ is the number of degrees of freedom associated with the standard error of the estimate. It is adjusted by subtracting 2 instead of 1 because two parameters of the fitted model (i.e. slope and intercept) are unknown. For estimating short-term precision only one parameter (i.e. the mean) needed to be estimated (resulting in degrees of freedom of $n - 1$).

Long-Term Precision of a Technique. Similar to the situation above, the correct estimate of long-term precision for a group of $j = 1 \ldots m$ patients is then given by:

$$\mathrm{SEE} = \sqrt{\frac{\sum_{j=1}^{m} \sum_{i=1}^{n_j} (y_{ij} - \hat{y}_{ij})^2}{\sum_{j=1}^{m} (n_j - 2)}} \tag{12}$$

where the denominator $\sum_{j=1}^{m} (n_j - 2)$ is the total number of degrees of freedom for the standard error of the estimate formed by combining all the data. Again, for the special case that the number of measurements n is the same for all subjects, the long-term precision error of the group of subjects is equal to the root-mean-square of the individual subjects' long-term precision errors:

$$\mathrm{SEE} = \sqrt{\sum_{j=1}^{m} \mathrm{SEE}_j^2/m} \tag{13}$$

or when expressed on a percentage basis

$$\mathrm{CV_{SEE}} = \mathrm{SEE}\Big/\left(\sum_{j=1}^{m} \bar{x}_j/m\right) \tag{14}$$

The confidence intervals can be calculated using formulae derived analogously to Eq. 8 and 7:

$$\frac{df}{\chi^2_{1-\frac{\alpha}{2},df}}\,\mathrm{SEE}^2 < \sigma^2 < \frac{df}{\sigma^2_{\frac{\alpha}{2},df}}\,\mathrm{SEE}^2 \tag{15}$$

with

$$df = \sum_{j=1}^{m} df_j = \sum_{j=1}^{m} (n_j - 2) \tag{16}$$

Heteroscedasticity of the SEEs can be handled analogously to the procedure described above for SDs. As a first step the BMD values have to be transformed according to the following equation:

$$\mathrm{BMD}'_{ij} = \mathrm{BMD}_{ij} - \mathrm{BMD}_{1j} - \mathrm{b}_j(\mathrm{date}_{ij} - \mathrm{date}_{1j}) \tag{17}$$

Dropping BMD'_{Ij} and using the remaining BMD'_{ij} in the same way that BMD values were used for the assessment of short-term precision heteroscedasticity will allow application of the same methods proposed above. Equation 17 is based on a model of linear changes in BMD over time with a slope b_j for each subject j. Eliminating the first data point of each indvidual and analyzing the deviations of follow-up measurements from the value expected on the basis of the model allows for appropriate adjustment of the degrees of freedom.

Results

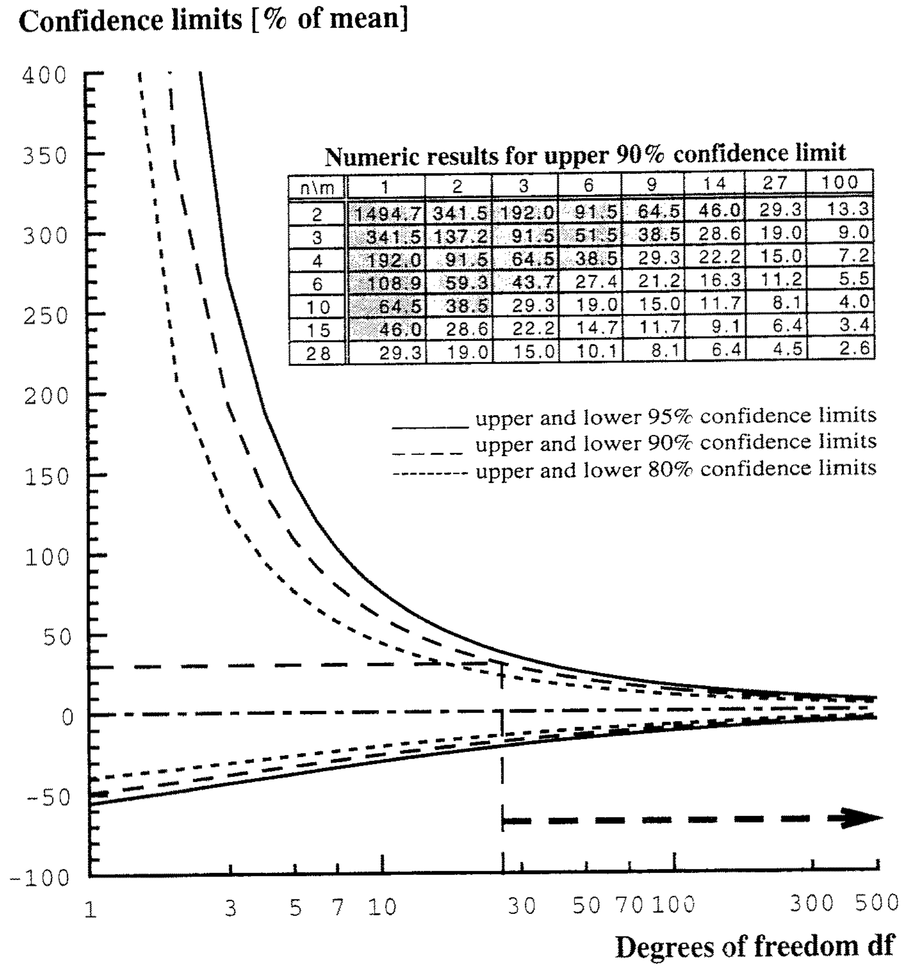
Incorrect incorporation of degrees of freedom in the formulae for calculating precision errors leads to underestimation of precision errors. Dividing by n instead of $n - 1$ in Eq. 1 for short-term precision errors would cause an error in the estimation of the precision errors. True precision errors, i.e. the population variance σ, would be higher than measured precision errors by a factor of $\surd[n/(n-1)]$, i.e. an additional 41%, 22% and 16% for 2-point, 3-point and 4-point measurements, respectively. Division by $n - 1$ instead of $n - 2$ in Eq. 10 for long-term precision errors would cause corresponding underestimation of long-term imprecision by again 41%, 22% and 16% for 3-point, 4-point and 5-point measurements, respectively.

If the mean precision error is mistakenly calculated as the arithmetic mean instead of as the root-mean-square average the precision is also made to appear better than it really is. Table 1 contains results of a computer simulation where estimates of precision are based on paired measurements, on 3-point measurements, 4-point measurements and so on. The underlying 'true' parameters of this simulation were based on data typical for BMD measurements by DXA of the spine in a healthy population. The population was characterized by a BMD of 1.000 ± 0.164 g/cm^2 and the "true" precision error level was set to 0.01 g/cm^2. The expectation value for the CV thus was 1.0%. The data of Table 1 demonstrate that only precision errors as defined in Eqs 4a, 4b or 5 yield unbiased estimates of the true parameters. The problem is most severe when individual data are simple pairs of measurements. As shown in Table 1 for 2-point measurements the true precision errors are about 25% higher than the measured precision errors. What is demonstrated with this simulation can also be shown mathematically. As shown in the Appendix the error introduced by using the arithmetic means varies with the total degrees of freedom, i.e. the number of subjects and repeat measurements carried out. Results of the simulation and those based on this theory show basically identical results.

Both of the noted errors, i.e. incorrect degrees of freedom and erroneous use of the arithmetic mean, would lead to biased estimates of precision: the precision result would differ depending on the number of

Table 1. Simulated "measured" levels of precision errors for n "repeat" measurements on m "subjects" with true σ^2 of 1.0 units (i.e. independent of the BMD level). All seven different experimental designs have approximately 27 degrees of freedom (df) and each was run 200 times for the estimates shown in the last four columns. Calculation based on arithmetic means demonstrates dependence of the estimated mean on the number of repeat measurements. For example, the true precision error of 1.0 is 25% higher than the estimated value based on 2-point measurements. Root-mean-square (RMS) estimates yield accurate results within ±1% independent of the number of repeat measurements

Precision estimate based on				Arithmetic average		RMS average	
m subjects	with *n*-point repeat exams.	No. of exams.	df	SD	CV	SD	CV
27	2-point	54	27	0.797	0.799	0.987	0.990
14	3-point	51	28	0.889	0.894	0.995	1.000
9	4-point	36	27	0.925	0.929	1.003	0.997
6	6-point	36	30	0.962	0.958	1.001	0.997
3	10-point	30	27	0.972	0.978	0.990	0.996
2	15-point	30	28	0.982	0.992	0.990	1.001
1	28-point	28	27	0.995	0.995	0.995	0.995



Numeric results for upper 90% confidence limit

n\m	1	2	3	6	9	14	27	100
2	1494.7	341.5	192.0	91.5	64.5	46.0	29.3	13.3
3	341.5	137.2	91.5	51.5	38.5	28.6	19.0	9.0
4	192.0	91.5	64.5	38.5	29.3	22.2	15.0	7.2
6	108.9	59.3	43.7	27.4	21.2	16.3	11.2	5.5
10	64.5	38.5	29.3	19.0	15.0	11.7	8.1	4.0
15	46.0	28.6	22.2	14.7	11.7	9.1	6.4	3.4
28	29.3	19.0	15.0	10.1	8.1	6.4	4.5	2.6

Fig. 1. Upper and lower 80%, 90% and 95% confidence intervals as a function of the degrees of freedom (df). Results are given as percentage difference to the mean precision error. The *tabular insert* shows numeric results for the upper 90% confidence limit for selected typical combinations of number of measurements $n = 2 \ldots 28$ obtained on each of $m = 1 \ldots 100$ subjects. Combinations located in the *unshaded area* of the table (also indicated in the figure by the *arrow*) yield upper 90% (or one-sided 95%) confidence intervals of less than +30%, a level considered sufficient for characterizing technique imprecision.

repeat scans carried out for each subject, asymptotically approaching the true value with increasing numbers of repeat scans.

Figure 1 plots the upper and lower 80%, 90% and 95% confidence limits of the estimate of the precision error as a function of degrees of freedom. It takes a large number of repeated measurements and/or a large group of patients to obtain an accurate estimate of precision errors within reasonably small error margins. For instance, for 10 repeated measurements on one subject (df = 9) the 90% confidence interval (i.e. the 5 to 95 percentile) is approximately −27% to +65% of the precision error (e.g. for a precision error of 1% the 90% confidence interval would by 0.73–1.65 mg/ml). It would take 100 repeated measurements on one subject (df = 99) to narrow down this confidence interval to approximately −10% to +13%. To determine an estimate of precision errors with an upper 90% confidence limit of below 30% (e.g. a precision error of 1% with confidence intervals from 0.83% to 1.3%) would require 28 repeated measurements on the subject (df = 27). The probability that the true precision error is greater than 1.3 × CV_{SD} would then be less than 5%. Figure 1 shows 80%, 90% and 95% confidence levels in graphical and, for selected typical combinations of numbers of subjects and measurements, in numeric form (the latter for the upper 90% confidence limit only). In theory, an infinite number of measurements would allow one to determine the mean precision error perfectly accurately, i.e. with confidence intervals of zero width.

In the model of Table 1 and Fig. 1 the SD was assumed to be independent of BMD. In reality the precision error typically increases with decreasing BMD (e.g. due to the difficulty of defining the bone edges in osteoporotic individuals). How large would the confidence interval of σ^2 be if, however, the distribution of precision errors across subjects showed significant heteroscedasticity? Modeling the SD with a decrease by 0.002 to 0.004 g/cm^2 per 1 g/cm^2 (close to data published by Ryan et al. [19]) showed only minimal increases in the confidence interval of σ^2.

Discussion

Our results demonstrate that precision errors have to be calculated using the correct degrees of freedom and averaging based on root-mean-square averages. Only this approach ensures unbiased estimates of precision errors. Otherwise precision errors would be underestimated by as much as 25% (using arithmetic means), and by up to an additional 41% (using incorrect degrees of freedom), if results are based on duplicate measurements per individual. In the literature, the selected methodology and definition employed in studies is rarely spelled out in detail. The magnitudes of the noted errors are substantial, thus warranting greater attention.

Short-term precision studies require repositioning of the patient between measurements unless machine imprecision is to be assessed individually. For long-term precision the SEE represents a worst-case estimate of technique-related imprecision because it also incorporates any non-linearity in skeletal changes. Here, separating true changes in mineral density from technique-related imprecision is difficult in vivo since there is no technique more accurate than bone densitometry that could serve as a gold standard. The two sources of variability could be differentiated to some extent by repeated measurements at each time point during a study, thus reducing technique-related imprecision. However, the added radiation exposure makes this approach ethically problematic. Fortunately, for most purposes, it may not be necesary to differentiate between technique-related imprecision and true variability of mineral density. Both true variability and technique imprecision will commonly be encountered in patient studies and they both limit the ability to detect changes in similar ways. The SEE of a linear regression analysis as a parameter that summarizes these two limiting factors thus may often be sufficient, particularly for sample size evaluations. While a detailed discussion of sample size estimates is beyond the scope of this contribution, it is evident that long-term precision errors may have a substantial impact on the power of research studies.

For most situations, using *linear* regression models for calculating the RMS error will be sufficient or at least represent a good approximation. However, a linear pattern of change over time may not be appropriate for describing response to treatment, particularly short-term gains with anti-resorbers, nor for accelerated postmenopausal bone loss. Here, one might instead observe a rapid change during the initial period of treatment followed by a plateau. For that kind of situation the RMS error would have to be calculated for a more complex model (e.g. non-linear or split into several linear time periods).

Many studies present precision data expressed on a percentage basis. This is only appropriate if absolute precision errors are proportional to the BMD. Reports in the literature [19] as well as our own unpublished data demonstrate that this is typically not the case. Absolute errors may actually increase in severely osteoporotic subjects due to edge detection problems caused by low density or degenerative changes. Due to the added effect of a decreasing denominator precision errors expressed on a percentage basis increase even more rapidly with decreasing BMD. It is for this reason that the subject group on which precision estimates are obtained should be characterized sufficiently (at least stating standard deviations of age and BMD). This will help the reader to determine whether precision data reported will be applicable to their own subject groups. In addition, for some measurements such as ultrasound velocity mean values are much higher than for, for example, DXA. The resulting very small CVs cannot directly be compared with the apparently much larger CVs of DXA. Consequently, in general, we would discourage expressing precision errors solely on a percentage basis. Principally, errors should be given in absolute units, with percentage values added for completeness. Other methods for standardizing precision errors are required.

Very few, if any, precision studies have specified confidence intervals of their estimates of precision errors. Without them, however, it is difficult to judge whether reported results are generalizable. The formulae presented in this paper allow one to estimate confidence intervals if at least the number of repeat measurements and subjects are available. As can be seen, the confidence interval of the precision error is not symmetric, indicating that precision errors have a lower limit of 0 but no upper limit. It is the upper confidence limit that in a practical evaluation of a technique's limitations is of particular concern. How many measurements (or, more precisely, degrees of freedom) are considered sufficient? The results presented show that 27 degrees of freedom may be considered sufficient to establish precision errors with what we consider to be a reasonably small confidence limit: an upper confidence limit that is 30% higher than the mean precision error with a probability of less than 5% that the true precision error would be larger (one-sided test). One would still have to decide whether to achieve this goal by performing a large number of measurements on each of a small number of subjects, or a small number of measurements on each of a large number of subjects. For examples, if one were to aim for the noted limit of at most +30%, one could perform 28 measurements on one subject, 3 measurements on each of 14 subjects, or 2 measurements on 27 subjects (compare the non-shaded area of the table in Fig. 1). The statistical value of these different choices would be the same provided all subjects in a given group can be characterized by the same precision. In practice, it might be best to spread one's desired degrees of freedom over a large number of subjects to average out differences between individuals. Taking 2 or 3 and 3 or 4 measurements per individual for assessment of short- and long-term precision, respectively, may be optimal.

The proposed approach will allow one to test whether precision errors are comparable across subjects; i.e. whether Eqs 3–6, 8 and 12–14, 15 provide valid estimates of means and confidence limits of short- and long-term precision errors, respectively. One caveat should be mentioned: the suggested tests for unequal variances are notorious for not being very powerul – a substantial degree of heteroscadisticity has to be present until it will be picked up by these tests. Fortunately, this limitation may not be very problematic for the purpose discussed here. Moderate degrees of heteroscedasticity lead to only minor increases in the confidence intervals of the mean precision error and thus can be tolerated. For example, increases of precision errors within the range of data reported in the literature [19] would only minimally affect the width of the confidence interval of the mean precision error. Substantial variability among subject groups, on the other hand, will more likely be picked up by the suggested tests and separate precision error analyses are indicated under these circumstances. When assessing precision errors it is advisable to plot the distribution of precision errors across subjects and study the shape of this distribution. Are shape and width of the distribution compatible with a chi-squared distribution of the given degrees of freedom? Specifying the 95% confidence interval of measured individual precision errors and comparing these with the expected values for the given degrees of freedom will give an indication as to what extent results fall within the expected range. Running tests for heteroscedasticity of precision errors across subjects is helpful with the caveat noted above. Does the distribution of measured individual precision errors show any signs of a bimodal shape? In that case, particularly if confirmed by a statistical test such as Bartlett's, it may be advisable to report precision errors separately for each subgroup.

The proposed method not only proved to be accurate but also appeared to be reasonably efficient (requiring 30–50 measurements) and robust with regard to shape and width of the distribution of precision errors. However, as noted, the lack of power of the mentioned statistical approaches imposes some limitations. Further research is warranted on the development of better tests as well as on the impact of width and shape of the distribution of individual precision errors on the confidence interval of the estimated mean precision error.

Conclusion

We conclude with the following recommendations:

Precision errors of individuals should be based on standard deviations (for short-term precision) or standard errors of the estimate (for long-term precision) employing formulae with correctly calculated degrees of freedom. Using degrees of freedom of n instead of $(n-1)$ causes underestimation of the precision errors (by up to 41% for duplicate measurements).

For short-term precision measurements, the patient should be repositioned between measurements.

Averaging of precision errors of several individuals should be based on root-mean-square averages. The commonly used arithmetic mean underestimates true precision errors (by up to 25% for duplicate measurements).

The subject group on which precision estimates are obtained should be characterized sufficiently (at least stating standard deviation of age and BMD).

The characterization of precision errors should include their estimated confidence interval according to Eqs 8 and 15. In addition, the observed 5% and 95% quantile of the individual precision errors should be compared with the expected 90% confidence interval of individual precision errors given by Eq. 9. Homogeneity of precision errors across subjects should be tested applying the proposed tests.

Three (or four) repeated measurements per individual in a subject group of at least 14 individuals (or other strategies that provide at least 27 degrees of freedom) are considered sufficient to characterize short-term (or long-term) precision of a technique. In the case of heteroscedasticity more measurements are required.

Acknowledgement. We acknowledge the assistance of Deljuanna Williams and Alan Herrick in preparing the manuscript.

References

1. Nilas L, Christiansen CC. Rates of bone loss in normal women: evidence of accelerated trabecular bone loss after the menopause. Eur J Clin Invest 1988;18:529–34.
2. Block J, Smith R, Glüer CC, et al. Models of spinal trabecular bone loss as determined by quantitative computed tomography. J Bone Miner Res 1989;4:249–57.
3. Kalender WA, Felsenberg D. Louis O, et al. Reference values for trabecular and cortical vertebral bone density in single and dual-energy quantitative computed tomography. Eur J Radiol 1989;9:75–80.
4. Harris S, Dawson-Hughes B. Rates of change in bone mineral density of the spine, heel, femoral neck, and radius in healthy postmenopausal women. Bone Miner 1992;17:87–95.
5. Davis JW, Ross PD, Wasnich RD, MacLean CJ, Vogel JM. Long-term precision of bone loss rate measurements among postmenopausal women. Calcif Tissue Int 1991;48:311–8.
6. Pacifici R, Rupich R, Vered I, et al. Dual energy radiography (DER): a preliminary comparative study. Calcif Tissue Int 1988;43:189–91.
7. Glüer CC, Steiger P, Selvidge R, et al. Comparative assessment of dual-photon-absorptiometry and dual-energy-radiography. Radiology 1990;174:223–8.
8. Steiger P, Block JE, Steiger S, et al. Spinal bone mineral density by quantitative computed tomography: effect of region of interest, vertebral level, and technique. Radiology 1990;175:537–43.
9. Lilley J, Walters BG, Heath DA, Drolc Z. In vivo and in vitro precision of bone density measured by dual-energy x-ray absorption. Osteoporosis Int 1991;1:141–6.
10. Schneider P, Börner W, Mazess RB, Barden H. The relationship of peripheral to axial bone density. Bone Miner 1988;4:279–87.
11. Rüegsegger P, Durand E, Dambacher MA. Localization of regional forearm bone loss from high resolution computed tomographic images. Osteoporosis Int 1991;1:76–80.
12. Glüer CC, Vahlensieck M, Faulkner KG, et al. Site-matched calcaneal measurements of broadband ultrasound attenuation and single x-ray absorptiometry: do they measure different skeletal properties? J Bone Miner Res 1992;7:1071–9.

13. Devogelaer JP, Baudoux C, Nagant de Deuxchaisnes C. Reproducibility BMD measurements on the QDR–2000 Hologic, Inc. In: Proceedings of Ninth International Workshop on Bone Densitometry, Traverse City, USA, 1992.
14. Kotz S, Johnson NL, Encyclopedia of statistical sciences. Vol. 8, New York: Wiley, 1982.
15. Harnett DL. Statistical methods. Reading, MA: Addison-Wesley, 1982.
16. Kuzma JW. Basic statistics for the health sciences. Mountain View, CA: Mayfield, 1984.
17. Olson CL. Statistics: Making sense of data. Boston, Mass.: Allyn and Bacon, 1987.
18. Wonnacott TH, Wonnacott RJ. Introductory statistics for business and economics. New York: Wiley, 1984.
19. Ryan PJ, Blake GM, Herd R, Parker J, Fogelman I. Spine and femur BMD by DXA in patients with varying severity spinal osteoporosis. Calcif Tissue Int 1993;52:263–8.
20. Abramowitz M, Stegun IA, Handbook of mathematical functions. Washington, DC: National Bureau of Standards, 1964.

Appendix

For a random sample from a given population, the sample variance s^2 is an unbiased estimate of the population variance σ^2, i.e. the expected value of s^2 is equal to σ^2. It is a simple arithmetical fact that if the mean of a set of s^2 values equals the constant σ^2 then the mean of the corresponding set of the s-values does not generally equal σ. In other words, s is a biased estimate of σ, because the expected value of s does not equal σ [17].

How large is the bias? In order to answer this question it is helpful to restate that

$$\chi^2_{df} = \frac{df \cdot s^2}{\sigma^2} \qquad \text{(A1)}$$

The χ^2-distribution can be expressed by means of the Gamma function Γ [20]:

$$\chi^2_{df} = \left[2^{\frac{df}{2}} \cdot \Gamma\left(\frac{df}{2}\right)\right]^{-1} \cdot \int_0^{x^2} (t)^{\frac{df}{2}-1} e^{-\frac{t}{2}} dt \qquad \text{(A2)}$$

One can show that the expectation value of the χ^2-distribution and, more generically, a power function of the χ^2-distribution, are given by [15]:

$$E[(\chi^2_{df})^p] = 2^p \cdot \frac{\Gamma\left(\frac{df}{2} + p\right)}{\Gamma\left(\frac{df}{2}\right)} \qquad \text{(A3)}$$

Given Eq. A1, the expectation value of s^2 would be given by

$$E[s^2] = \frac{\sigma^2}{df} \cdot E\,[\chi^2_{df}] \qquad \text{(A4)}$$

Since by definition $\Gamma\ (df + 1) = df \cdot \Gamma\ (df)$, it follows from Eq. A3 that

$$E[s^2] = \frac{\sigma^2 \cdot \Gamma\left(\frac{df}{2} + 1\right)}{\frac{df}{2} \cdot \Gamma\left(\frac{df}{2}\right)} = \sigma^2 \qquad \text{(A5)}$$

Hence s^2 and thus any RMS average of s_j represent unbiased estimators of σ^2, independent of the degrees of freedom of the measurement.

By contrast s or any arithmetic means of several measurements of s_j has the expectation value of

$$E[s] = \sqrt{\frac{\sigma^2}{df}}\, E[(\chi^2_{df})^{\frac{1}{2}}] \qquad \text{(A6)}$$

which is given by (compare also p. 626 in [14])

$$E[s] = \frac{\sqrt{2}\sigma}{\sqrt{df}} \cdot \frac{\Gamma\left(\frac{df}{2} + \frac{1}{2}\right)}{\Gamma\left(\frac{df}{2}\right)} \qquad \text{(A7)}$$

The value of $E\,[s]$ varies with df and only equals σ in the asymptotic limit as df becomes infinite.

As an example, take the case of 2 measurements on each subject. Here we have $E[s] = \sqrt{2} \cdot \sigma \frac{\Gamma(1)}{\Gamma(\frac{1}{2})}$. By definition, $\Gamma\ (1) = 1$ and $\Gamma\ (\frac{1}{2}) = \sqrt{\pi}$ and thus $E[s] = \sqrt{\frac{2}{\pi}}\,\sigma = 0.798$. Similarly, for $n = 3, 4, 6$ and 12, i.e. $df_j = 2, 3, 5$ and 11, we obtain $E\,[s] = 0.886$, 0.921, 0.952 and 0.978, respectively – essentially the results displayed in Table 1. These figures directly represent the degree of underestimation of σ when using arithmetic means for averaging of standard deviations.

Received for publication 15 April 1994
Accepted in revised form 27 December 1994