JOURNAL OF BONE AND MINERAL RESEARCH
Volume 14, Number 11, 1999
Blackwell Science, Inc.

Monitoring Skeletal Changes by Radiological Techniques

CLAUS-C. GLÜER

ABSTRACT

The longitudinal sensitivity of a technique, i.e., its ability to monitor skeletal changes, is affected by two parameters: the long-term precision error (PE_{lt}) and the subject group-specific response rate (i.e., annual rates of change). Both need to be considered to avoid misinterpretation of measured changes. A new concept to aid clinical decision making for longitudinal measurements is proposed which is based on three types of measures: criteria for detecting changes—the "least significant change" (LSC) is the smallest change to be considered statistically significant, but for certain clinical questions a smaller margin, the "trend assessment margin" (TAM), can be sufficient for decision making; follow-up time intervals—for follow-up exams the patient should be called in at about the time interval specified by the (population specific) "monitoring time interval" (MTI) or, about one-third of the time earlier, after the "trend assessment interval" (TAI), depending on whether the decision can be based on the LSC or the TAM; and the standard precision error (stdPE)—the smaller stdPE, the more sensitive the technique to monitor skeletal changes. Together, these three measures yield a good characterization of a technique's ability to monitor skeletal changes. Compared with previous concepts, the proposed standardization by a response ratio instead of measures of spread or response rates makes the stdPE substantially less subject group dependent. It allows comparison of stdPE across different studies and could replace the misleading concept of expressing precision as a coefficient of variation. Application of this concept should facilitate the interpretation of measured skeletal changes. (J Bone Miner Res 1999;14:1952–1962)

INTRODUCTION

FOR THE EVALUATION of disease progression, response to treatment, and the estimation of fracture risk it is important to interpret measured changes in bone mineral density (BMD) and other skeletal parameters in a sensible fashion. Bone densitometry is an accurate and precise method, but due to limitations of the technique, the measured results only approximate the true changes. The longitudinal sensitivity of a technique, defined as the ability to monitor changes in skeletal status,[1] is limited by technique imprecision. To allow a comparison of the imprecision of techniques specified in different units, precision errors are commonly reported on a percentage basis, calculated, e.g., as a coefficient of variation (CV) of repeated measurements. However, it is known that the apparent comparability of percentage units can be misleading and therefore different ways of standardizing precision errors have been proposed[2–8]: division of precision errors by population variance, 10–90% range, normal age-related decline per annum, etc. To date, none of these methods for standardization has been investigated thoroughly, compared with other approaches, let alone been unanimously adopted. There is a lack of standardization of the methods of how to standardize precision. To judge the advantages and limitations of the competing approaches, one needs to define the problems and point out the goals for standardization: What sense is in standardization of precision errors?

When evaluating longitudinal changes over time, the three following issues frequently need to be addressed in clinical decision making:

- The interpretation of measured changes: Are the changes calculated meaningful and clinically relevant?—Random fluctuations are sometimes mistaken as real changes.
- Scheduling of a follow-up visit to determine rates of

Arbeitsgruppe Medizinische Physik, Klinik für Diagnostische Radiologie, Universitätsklinikum an der Christian-Albrechts-Universität zu Kiel, Kiel, Germany.

change: What time interval is required to allow accurate assessment of response to treatment or progression of disease?—Follow-up measurements performed too early do not allow to judge the significance of measured changes.

- Comparison of techniques: Which technique is suited best to detect changes accurately and quickly?—A confusing number of insufficient methods for standardizing precision errors have been used.

To motivate the concept proposed, a few explanations regarding the difficulty to answer the third question will be given. The simple answer that the technique with the best reproducibility or smallest precision error would be best suited for monitoring changes over time is flawed. First of all, it is obvious that precision errors of different measurement parameters cannot be directly compared when specified in absolute units (e.g., in g/cm^2 vs. mg/cm^3 vs. dB/MHz vs. m/s). Expressing the precision error on a percentage basis is quite popular but it does not solve this problem. Quite the contrary: superficially implying that the results are now readily comparable, this common percentage unit can be highly misleading; just change the definition of a parameter (e.g., by adding an offset) and any desired level of percentage precision can be achieved. For example, original directly calculated broadband ultrasound attenuation (BUA) values typically range between 30 dB/MHz and 80 dB/MHz and a precision error, e.g., 2 dB/MHz would yield percentage precision errors of 2.5% (2/80 × 100) to 6.7% (2/30 × 100). On some quantitative ultrasound (QUS) devices, the original values are subjected to an offset. If this were, for simplicity's sake, taken to be 50, the resulting range would be 80–130 dB/MHz. This simple manipulation term would reduce the precision error to a range of 1.5% (2/130 × 100) to 2.5% (2/80 × 100), without any true improvement in longitudinal sensitivity. These examples also illustrate a second problem, that percentage precision errors appear to be better (i.e., lower) in healthy subjects, simply because of the larger denominator (e.g., 2.5% vs. 6.7%). Again, this does not reflect any true difference in longitudinal sensitivity.

More fundamentally, however, one has to recognize that precision errors by themselves tell little about the ability of a technique to monitor changes: to characterize longitudinal sensitivity the responsiveness of the monitoring parameter needs to be considered as well. QUS parameters represent good examples to demonstrate what happens if this aspect is neglected. Changes in speed of sound (SOS) are typically in the range of only a few meters per second per annum out of perhaps 1500–2000 m/s, reflecting responses of less than 1% per annum; paralleling changes in BUA amount to a few decibels per megahertz per annum out of perhaps 50–100 dB/MHz, reflecting responses of several percentage points. Not surprisingly, the percentage precision error for SOS is typically smaller than that of BUA by at least an order of magnitude. This does not, however, necessarily represent an equivalent advantage in the ability of SOS to monitor changes because the lower responsiveness of SOS has not been taken into account. Dividing percentage errors by some measure of responsiveness to obtain a measure of longitudinal sensitivity called standardized precision error thus appears to be useful, and various methods have been proposed. So, what is the problem with this kind of standardization?

One issue is that different measures of responsiveness result in differently standardized precision errors, most of which are still substantially biased, e.g., affected by the cohort effect. Let's take, just as a typical example, one of the common methods to standardize precision errors, i.e., division by the SD of the readings of the subject group. In case of a narrowly defined subject group (e.g., young normals) the resulting "standardized precision error" will be larger than that calculated for a mixed group of healthy and osteoporotic individuals even if the technique has identical precision (expressed in absolute units) for healthy and osteoporotic subjects. Sample sizes for precision studies are typically quite small. Subject selection therefore can easily introduce a substantial bias. In fact, one could easily "improve" the standardized precision by simply adding a few more extreme cases (very healthy or very osteoporotic) to the subject group. As long as one limits the comparison of precision of techniques to measurements all obtained in the same subject group, this kind of standardization is helpful. But the moment this "standardized" precision is regarded as a universal characteristic that describes the ability of a technique to monitor changes in any subject group, there is room for misinterpretation. Clearly, standardized precision errors calculated from different subject groups cannot be directly compared; they have in fact not been standardized at all in a meaningful fashion.

What needs to be done to resolve this problem? Four requirements for useful ways of standardization can be named: the measure(s) should reflect both imprecision and responsiveness; the measure(s) should make it possible to directly compare the performance of techniques tested in different studies; the measure(s) should have an intuitive clinical meaning; and the measure(s) should be as insensitive to subject selection bias as possible.

Responsiveness varies for different genders, age groups, and therapies, and reproducibility may also differ. Thus, one needs to investigate those different cohorts separately in order to determine the respective levels of standardized precision. Consequently, standardized precision as a measure that reflects reproducibility as well as responsiveness can no longer be represented by a single number.

Finally, the quality of the estimate of standardized precision will depend on the type of study design employed. As usual, results derived from longitudinal studies are preferable to cross-sectional data.

Keeping these caveats in mind, the following proposed concepts should facilitate an objective assessment of a technique's ability to monitor longitudinal changes.

MATERIALS AND METHODS

The interpretation of measured changes: Introducing change criteria

For clinical decision making it is important to know what magnitude of measured change is required to be sure that the patient has truly lost bone. In other words, which change is statistically significant, taking into account the limitations of instrument performance? As other authors have previously shown,[9] for two point measurements over time only changes exceeding 2.8 times the precision errors of a technique can be considered as a criterion for true changes (with 95% confidence). The corresponding change criterion has been termed "least significant change" (LSC):

$$\mathrm{LSC} = 2.8 \times \mathrm{PE}_{\mathrm{lt}}$$

where $\mathrm{PE}_{\mathrm{lt}}$ is the largest precision error of the technique used.

However, clinicians have to balance the desire to attain statistical certainty with the patient's need to get treated as quickly as possible if there is a valid indication to do so. To not withhold potentially important medication, the clinician may be satisfied with confidence levels < 95%.[10] Indeed, it is conceivable that varying confidence limits may be appropriate under different clinical situations. For example, when identifying someone who has indeed responded to therapy in a situation where response is expected, the required confidence may be somewhat less. However, in a situation where a change in a course of therapy is being considered, the clinician may require the 95% confidence in order to change the intervention. Statistically, intervals for any level of confidence can be defined. Avoiding a plethora of different confidence levels, we propose to introduce one additional, less stringent change criterion, the "trend assessment margin" (TAM). Given by

$$\mathrm{TAM} = 1.8 \times \mathrm{PE}_{\mathrm{lt}}$$

it can be considered as a criterion for true changes at a confidence level of 80% for two-sided tests or a level of 90% for single-sided tests. The word "trend" should imply that less strict requirements have to be met than for the test for significance at the 95% confidence level.

Both change criteria, LSC and TAM, should be calculated using the long-term, not the short-term, precision error specified for measurements in vivo in a comparable subject group.

Scheduling of follow-up visits: Introducing follow-up time intervals

After establishing the baseline status of a parameter (e.g., BMD), the rate of change of that parameter needs to be determined to assess progression of disease or response to treatment. What time interval between that baseline and a follow-up measurement is sufficient to allow for an accurate and valid assessment? When answering this question one will face the dilemma of having to settle for either a quick answer at an early follow-up visit associated with greater statistical uncertainty when estimating the true change from the measured change or a more solid answer at a later visit with the risk of substantial bone loss and fractures in the meantime. Therefore, analogous to the preceding section, two different follow-up time intervals can be defined.

The "monitoring time interval" for assessment of disease progression ($\mathrm{MTI}_{\mathrm{P}}$) is an estimate of the time period after which half of the patients with normal bone loss will show a measured change exceeding the change criterion LSC. It is given by:

$$\begin{aligned}\mathrm{MTI}_{\mathrm{p}} &= \mathrm{LSC/median\ response}\\ &= 2.8 \times \mathrm{PE}_{\mathrm{lt}}/\mathrm{median\ change\ per\ annum}\end{aligned}$$

Similarly, the "trend assessment interval" (TAI) is an estimate of the (shorter) follow-up time period, after which half of the patients with normal bone loss will demonstrate a change exceeding the change criterion TAM. It is given by:

$$\begin{aligned}\mathrm{TAI}_{\mathrm{p}} &= \mathrm{TAM/median\ response}\\ &= 1.8 \times \mathrm{PE}_{\mathrm{lt}}/\mathrm{median\ change\ per\ annum}\end{aligned}$$

For example, for a technique with a long-term precision error of $\mathrm{PE}_{\mathrm{lt}} = 1.5\%$ and a patient for whom an annual change of 1% per annum could be expected, the $\mathrm{TAI}_{\mathrm{P}}$ and $\mathrm{MTI}_{\mathrm{P}}$ would be 2.7 years and 4.2 years, respectively. For a subject with a faster expected annual loss rate of 3%, the $\mathrm{TAI}_{\mathrm{P}}$ and $\mathrm{MTI}_{\mathrm{P}}$ would be 0.9 years and 1.4 years, respectively.

When scheduling a patient to assess response to treatment, a similar strategy could be followed. In the previous section, the criteria by which patients can be considered to have responded positively to treatment were established: those for whom the measured change was larger than the normal pretreatment loss by at least TAM or LSC. At what point in time is that expected to happen for the majority of the treated patients? $\mathrm{MTI}_{\mathrm{T}}$ and $\mathrm{TAI}_{\mathrm{T}}$ for treatment could be defined as:

$$\begin{aligned}\mathrm{MTI}_{\mathrm{T}} &= \mathrm{LSC/median\ treatment\ response}\\ &= 2.8 \times \mathrm{PE}_{\mathrm{lt}}/\mathrm{median\ improvement\ vs.\ placebo\ per\ annum}\end{aligned}$$

$$\begin{aligned}\mathrm{TAI}_{\mathrm{T}} &= \mathrm{TAM/median\ treatment\ response}\\ &= 1.8 \times \mathrm{PE}_{\mathrm{lt}}/\mathrm{median\ improvement\ vs.\ placebo\ per\ annum}\end{aligned}$$

The index "T" stands for treatment but it should be specified according to the treatment investigated. For example, if estrogen is expected to improve bone by 3% per annum (median), while untreated individuals would lose bone at a median rate of –1% per annum, the median treatment effect would be 4% per annum, and for a technique with a 1.5% precision error the recommended $\mathrm{TAI}_{\mathrm{estr}}$ and $\mathrm{MTI}_{\mathrm{estr}}$ would be 8.1 months and 12.6 months, respectively. After these time periods, the median gain in BMD will be 2.02% and 3.15%, respectively, which represents the levels at which one can have 80% or 95% confidence that the subject is indeed losing bone at less than the normal rate (2.02% = –0.68% + 2.7% and 3.15% = –1.05% + 4.2%).

Comparison of techniques: Introducing redefined precision errors

Both the MTI and the TAI defined in the previous section would be measures appropriate to characterize a technique's ability to monitor skeletal changes: the shorter the MTI and the TAI, the better the longitudinal sensitivity.

Alternatively, longitudinal sensitivity could also be expressed by precision errors if these are corrected for differences in responsiveness. Such a standardization procedure would allow one to stay with the familiar concept of expressing precision errors in percentage units (instead of the units of years for TAI and MTI). This facilitates the interpretation since it is the kind of measure most researchers and clinicians are used to.

Standardization can be achieved by correcting the precision error of the technique A investigated by the response ratio (rr); rr is given as the ratio of the response rate of the reference technique R divided by the response rate of the technique A:

$$\mathrm{rr(AvsR)} = \text{response rate (R)/response rate (A)}$$

The "standardized precision error," $\mathrm{sPE_{lt}}$, of a technique A that has been standardized relative to the reference technique R is then given by:

$$\begin{aligned}\mathrm{sPE_{lt}(AvsR)} &= \mathrm{PE_{lt}(A)} \times \mathrm{rr(AvsR)}\\ &= \text{response rate (R)/response rate (A)}\end{aligned}$$

Once standardized in this fashion, the standardized precision error can now directly be compared with the precision error of the reference technique R.

This method of standardization transforms the precision error of technique A to the scaling of the reference technique R. The multiplication by the rr makes standardization precision errors truly comparable across techniques. All precision errors of techniques A, B, C. . . that have been standardized in this fashion can now directly be compared among each other and also with the precision error of the reference technique (which, by definition is equal to the standardized precision error because it is standardized to itself).

For example, if a QUS device has a (long-term) precision error for SOS of 0.3%, and if one wishes to compare this with the reported BUA performance of this device, in this example set to 1.5%, one would standardize one or the other parameter versus the second parameter. Let us (arbitrarily) denote BUA as the reference technique. The precision error of SOS would be standardized by multiplication with the rr of BUA versus SOS. If this were, for example, found to be equal to 5 (i.e., the annual change of BUA is five times larger than that for SOS), the standardized precision error of SOS would be 1.5%, i.e., equal to the precision error of BUA. Both devices would, in this example, have the same longitudinal sensitivity.

If we had instead set SOS as the reference technique, the precision error of BUA would have to be standardized. In this case, the rr is 0.2 and the standardized precision error of BUA would be equal to $1.5 \times 0.2 = 0.3\%$, i.e., again equal to the precision error of SOS. No matter which technique was selected as the reference technique, the result "equal standardized precision" remains the same. However, the scaling of the standardized precision error depends on the selection of the reference technique. In the first example, we calculated a standardized precision error of 1.5% for both techniques, whereas, if we switched the reference technique, the standardized precision error was 0.3%.

Consequently, one has to agree on the choice of a universal reference technique to really make techniques comparable across studies. We propose to use posterior–anterior dual-energy X-ray absorptiometry of the lumbar spine ($\mathrm{DXA_{sp}}$) as the reference technique because it is most widely used in longitudinal studies. To denote clearly this choice, we propose to call a standardized precision error of a technique A that has been standardized versus $\mathrm{DXA_{sp}}$ the "standard precision error," $\mathrm{stdPE_{lt}(A)}$:

$$\mathrm{stdPE_{lt}(A)} = \mathrm{sPE_{lt}(AvsDXA_{sp})}$$

Use of standard precision error should be preferred over standardized precision errors whenever possible, i.e., whenever the technique investigated and $\mathrm{DXA_{sp}}$ can be measured in the same subjects. The response rate of the technique investigated and that of $\mathrm{DXA_{sp}}$ should be obtained on the same subjects.

If the techniques A and R have different units (e.g., m/s and $\mathrm{g/cm_2}$), both the precision error and the response rates need to be expressed on a percentage basis. If the techniques A and R have the same units, standardized precision could alternatively also be evaluated in absolute units, but then both the precision errors and the response rates need to be expressed consistently in absolute units. To make all equations as universally applicable as possible, the precision errors are all expressed on a percentage basis throughout the remainder of this manuscript.

Response ratios are likely to be less subject group dependent than response rates (part of the cohort-bias cancels out). Still, gender and ethnic group, health status (healthy, osteopenic, osteoporotic, etc.), and—if applicable—type and dosage of treatment may have an impact and should therefore be specified. Standard precision errors thus may differ and the ranking of longitudinal sensitivity could depend on the cohort. Therefore, the following scenarios should be investigated before a generic statement on the ranking of techniques can be made:

- Longitudinal sensitivity for detecting normal aging processes;
- Longitudinal sensitivity for detecting disease progression in osteoporotic individual; and
- A standardized longitudinal precision error for detecting changes due to treatment which is treatment specific and thus type of treatment and dosage need to be specified.

Application of the concepts

To illustrate their utility, the concepts derived are being applied using data from the literature. Short-term precision errors (since long-term precision errors are not established for QUS, yet) and typical response rates have been gathered for two DXA and two QUS parameters. Since it is not the focus of this paper to compare techniques but to present the concept, the numbers given should only be taken as an example of the application of the concept, not an assessment of the longitudinal sensitivity of the four parameters.

Table 1. Hypothetical Example for the Concepts Derived

Technique	*Response rate (% p.a.)*	*Precison error (%)*	*Change criteria (%)*		*Follow-up times (years)*		*Standard precision error (%)*
		PE_{st}	*TAM*	*LSC*	*TAI*	*MTI*	*$stdPE_{st}$*
BMD_{spine}	0.9	0.7	1.3	2	1.4	2.2	0.7
BMD_{femtot}	0.6	0.7	1.3	2	2.2	3.3	1.1
SOS_{calc}	0.07	0.16	0.29	0.45	4.1	6.4	2.1
BUA_{calc}	0.3	1.2	2.2	3.4	7.3	11.3	3.6

Skeletal parameters include bone mineral density (BMD) measured by posterior–anterior dual-energy X-ray absorptiometry (DXA) of the lumbar spine (BMD_{spine}), DXA of the total proximal femur (BMD_{femtot}), speed of sound (SOS_{calc}), and broadband ultrasound attenuation (BUA_{calc}) of the calcaneus. Despite large differences in uncorrected short-term precision errors (PE_{st}) and response rates, parameters reflecting the longitudinal sensitivity, such as trend assessment interval (TAI), monitoring time interval (MTI), and standard short-term precision error ($stdPE_{st}$), can be compared directly across techniques. Change criteria such as the trend assessment margin (TAM) and the least significant change (LSC) provide threshold levels for assessing whether significant changes at the 95% and 80% confidence level, respectively (two-sided tests), have occurred.

RESULTS

The concepts proposed have been applied to hypothetical performance for two DXA approaches (BMD of posterior–anterior DXA of the lumbar spine, BMD_{spine}, and DXA of the total proximal femur, BMD_{femtot}) and two QUS approaches (SOS and BUA of the calcaneus) presented in Table 1.

DISCUSSION

The need for a new concept

"The search for difference seems to be, for current research, what the search for the philosophers' stone was for alchemy, or the Holy Grail for the knights of legend–beguiling, elusive and, all too often, illusory."[(11)] A dozen years after Robert Heaney raised the issue, his assessment remains largely true. Important contributions have been made in the meantime, but in clinical practice still today considerable confusion about the interpretation of measured changes and the comparative performance of techniques remains. With the increasingly widespread use of ultrasound techniques, these problems are amplified since the precision of QUS and bone densitometry techniques cannot easily be compared because of different units and the fallacies of the expression on a percentage basis. In addressing these issues, a new concept was developed to aid the clinician in making decisions when following and treating individual patients. Researchers should benefit from getting a tool for more objective ways of comparing the longitudinal responsiveness of technique. The concept centers around the three issues listed in the introduction section. Those issues and the components of the concept proposed as a solution are:

- The interpretation of measured changes: Are the changes calculated meaningful and clinically relevant? Proposed answer: yes, if they exceed the change criteria LSC or TAM.
- Scheduling of a follow-up visit to determine rates of change: What time interval is required to allow accurate assessment of response to treatment or progression of disease? Proposed answer: re-examine patient after the follow-up time intervals MTI or TAI.
- Comparison of techniques: Which technique is suited best to detect changes accurately and quickly? Proposed answer: the technique with the lowest standard precision error (stdPE).

Also, the four requirements for useful ways of standardization listed in the introduction are largely fulfilled. Subject selection bias is still an issue for the MTI but only because it is meant to be specific for populations with differing rates of changes. For stdPE, this problem is minimal as long as the response rates used to calculate the rr have been obtained on the same individuals for both techniques. If this is not the case, care has to be taken to compare similar populations. stdPE is best suited for direct comparisons of different techniques, even across different studies.

All three parameters have fairly intuitive meanings. A change less than the TAM cannot be interpreted as clinically relevant; a change less than the LSC is not a statistically proven change. A follow-up time interval shorter than the TAI or MTI, respectively, will yield such insufficient changes in the majority of cases. The stdPE can be easily interpreted since the scaling is simple and familiar: a performance of a stdPE of 1–1.5% is to be considered as fairly good. This is similar to the level of precision reported in many studies for DXA_{sp}, which is familiar to most researchers.

The concepts derived are not limited to radiographic diagnostic approaches. Change criteria, follow-up time intervals, and standard precision errors could, for example, also be calculated for markers of bone turnover. The huge difference in the response rates and precision errors for markers versus radiographic parameters does not represent a hurdle, since they cancel out when calculating follow-up time intervals or standard precision errors. Therefore, the standard precision errors of a marker of bone resorption can be put in perspective directly with the corresponding results for radiographic parameters.

Scheduling of follow-up visits

The interpretation of the MTI (or TAI) as a measure of longitudinal sensitivity is intuitive and simple: it represents the follow-up time required to test whether clinically relevant changes have occurred. The shorter the MTI, the more sensitive the technique.

Still, a few caveats should be noted. First, there is no single MTI (or TAI) for each technique. The magnitude of the parameter is likely to be different for studies on disease progression (and again between normal and fast losers) and response to treatment (here it may also depend on the type of treatment investigated). When pursuing the latter issue, one should also note that response to treatment is quite variable even for an established effective medication like estrogen.[12,13] By definition, half of the patients will show a response, which is less than the median response, and, consequently, their measured improvement during the MTI_T (or TAI_T) will be smaller than the LSC (or TAM).

Patients that do not reach the level of change expected after the MTI_T (or TAI_T) may still have benefited from treatment, albeit at a somewhat lower level. How do we interpret such a "negative" insufficient response? How do we detect true nonresponders? As long as a patient's measured change is "better" than the loss expected without treatment, the patient is more likely to benefit from the treatment than not. However, the statistical uncertainty would be unacceptably high. Depending on the health status of the patient, one could still take the upward trend as encouraging and schedule another follow-up visit at twice the MTI_T (or TAI_T). At this point in time, even patients with only half the median response rate (comparing treated and untreated patients) can be expected to show a change that exceeds the LSC (or TAM). According to published studies, this would be met by ~60% of the patients on estrogen[12,14] and ~80% of the patients on alendronate,[15] assuming normal distributions of the response. Further reductions in the change criteria appear to be clinically questionable, not only because the response is smaller, but because the follow-up time intervals required to test responsiveness would become prohibitively long.

Alternatively, it may also be justified to schedule follow-up visits at time intervals shorter than the TAI or MTI for the purpose of identifying patients that continue to lose bone at a rapid rate. Bone losses exceeding the TAM or LSC would represent appropriate test criteria.

(Re-)Defined standardized precision errors

In the appendices, a number of different definitions for standard precision errors have been developed. To avoid confusion, one should use the term stdPE_L only if the standard precision error has been obtained from truly longitudinal data. stdPE_{lt} is preferable to other approaches. If normative data of all manufacturers would be of equally good quality, the stdPE_N might be a good estimate of longitudinal sensitivity to detect aging changes. However, it is known that differences have been reported recently for DXA,[16] and discrepancies again may be encountered for newly introduced devices and methods. Therefore, this type of standardization should be used carefully.

Compared with previously proposed approaches, the new definitions of standard (and standardized) precision errors presented here offer the advantages of ease of interpretation (all parameters), suitability for comparison of any two techniques (standardized precision errors), comparability across different studies (standard precision errors), minimal cohort bias (corrections by rr's rather than response rates), and applicability to radiographic as well as biochemical approaches.

These advantages will be discussed, and afterward the limitations of definitions previously proposed by other authors will be outlined.

Why introduce two concepts of standard and standardized precision errors?

The advantage of the concept of the standardized precision error is that is can readily be used to compare the precision errors of any two techniques, provided that the uncorrected precision errors and response rates are known for both of the techniques. This will allow comparisons in a variety of research situations, whereas the concept of standard precision errors requires researchers to determine both precision errors and response rates of DXA_{sp} in their population, which may not always be feasible. However, agreeing on a common reference standard—as required for the standard precision error—yields a well defined robust measure for comparison of the longitudinal sensitivity of techniques, even across different studies.

To facilitate assessment of standard precision errors for a large number of techniques, publication of the rr's themselves would be helpful. Such data would provide researchers with a methodology to determine the stdPE for a new technique, even if no direct comparison with posterior–anterior dual-energy X-ray absorptiometry (PA-DXA) of the lumbar spine can be carried out at the center. It would only be necessary to compare the new technique with a reference technique for which rr versus PA-DXA of the spine is already available from the literature.

Why use BMD of PA-DXA of the spine as the reference standard?

Previous standardization approaches failed to achieve the goal of standardization because the result was still very subject-group dependent. The proposed concept reduces the impact of this error source. Still, other forms of bias needed to be considered. BMD of PA-DXA of the spine is substantially affected by degenerative changes. Subjects affected by degenerative changes need to be excluded when calculating standard precision errors, specifically when evaluated from cross-sectional data. The choice of DXA_{sp} as the reference technique for calculation of the standardized precision error does not mean that DXA_{sp} is the technique with the best longitudinal sensitivity; it was only considered to be the best reference standard.

Why correct sPE and stdPE using response ratios rather than response rates?

Correcting precision errors by division by response rates yields a good measure of longitudinal sensitivity but this measure is very sensitive to the population sample studied (cohort bias). This approach was used to define the MTI (or TAI) because in the context of estimating follow-up times the impact of the population is of critical importance. For a generic comparison of techniques, a more robust measure like the stdPE is preferable. As long as different techniques measure a similar aspect of bone, their response rates will be partially correlated. Therefore, a substantial fraction of the impact of the population studied is eliminated when using rr's instead of response rates. Moreover, multiplication by the rr, which will be unity for the reference technique, leaves percentage precision errors in the range of values (typically 1–5%) that researchers and clinicians are familiar with, increasing the likelihood of acceptance and facilitating the interpretation. This is not the case for most definitions of standardized precision errors proposed previously.

Why adjust for annual rates of change and not for a measure of intersubject variability?

When calculating response rates for standardized precision errors, the measure "annual rates of change" was proposed. If standardized precision errors are meant to be used as a measure of longitudinal sensitivity, it seems logical that response should be defined as a change over time. This should be the most intuitive approach to quantitate a technique's ability to monitor longitudinal changes. For longitudinal studies it is the obvious choice anyway, but for cross-sectional estimates of longitudinal sensitivity one might consider other measures of responsiveness. However, standardization by annual rates of change was selected here as well, in order to make the (short-term, cross-sectional) definition of $\mathrm{stdPE}_{\mathrm{st}}$ as similar as possible to that of (the longitudinal) $\mathrm{stdPE}_{\mathrm{lt}}$ (see Appendix 2). Moreover, one should note that any measure of spread or dynamic range includes an error component caused by the precision (and accuracy) errors. Therefore, for two techniques of comparable true responsiveness, the one with the larger precision error would show the larger apparent responsiveness. Consequently, estimates of stdPE that are based on measures of spread underestimate the differences in longitudinal sensitivity between techniques. Techniques with poorer precision will demonstrate an artificially enlarged dynamic range and, consequently, their calculated standardized precision error looks better than it really is (precision bias). One can correct for this, i.e., remove the precision error from the measure of spread, by two-way nested analysis of variance.

Why should parameters of longitudinal sensitivity be based on long-term rather than short-term precision data?

The assessment of skeletal changes via radiological techniques such as bone densitometry or QUS usually requires time intervals between follow-up measurements of 1 year or longer. Therefore, the reproducibility of techniques has to be based on long-term precision errors, which are usually larger than short-term precision errors.[17] There are additional error sources (e.g., long-term stability of equipment, variability of body temperature for SOS measurements, etc.) that can only be determined from longitudinal data.[18] Precision errors derived from short-term repeat measurements only approximate true reproducibility errors. Still, their calculation can be helpful, particularly if one can assume that the ratio of short-term and long-term precision errors (i.e., the precision error ratio) of the technique investigated and that of the reference technique would be similar. Then, the ranking of the sensitivities of the techniques would not be affected (but the absolute magnitudes of longitudinal sensitivity will be overestimated).

Previous concepts

The previously published concepts of standardization all have some of the aforementioned problems. Miller et al. introduced the standardized CV based on normalization by the dynamic range given by 90% interpercentile range,[3] and Greenspan et al. used a similar approach but standardized with the 95% interpercentile range.[8] Both measures of population spread depend on subject selection criteria and thus are affected by the noted cohort and precision biases. Langton proposed the concept of ZSD, i.e., the standard deviation of the Z score which is taken as a measure of standardized precision.[5] Here, the problems with sampling bias are less severe since the population variance used to calculate the Z score is usually obtained from large populations measured to derive normative data. However, the current debate about the validity and comparability of normative data provided by the manufacturers puts some question marks on this approach. More importantly, however, rather than being a good measure of responsiveness, the larger population variance could also be due to technique problems (precision bias) and to diversity in subjects which is unrelated to osteoporosis (accuracy bias). In fact, a technique with a large age-related decline relative to its population variance is more likely to allow monitoring of skeletal changes compared with a technique that—in the extreme—would show no age-related change, even if that second technique had an equally large or even larger population variance. Population variance does not appear to be a reliable measure of responsiveness over time, and the ZSD may be more suitable for characterizing diagnostic sensitivity.

In another approach, Blumsohn et al. have proposed the index of individuality[4] which is affected by the noted sampling bias because it incorporates a measure of intersubject variability. The problems are similar to those noted by Quan and Shih for another measure of standardized precision, the intraclass CV.[19] Both of these measures are perhaps better suited to assess diagnostic sensitivity. Machado and colleagues have standardized precision by dividing precision errors by the average difference between healthy and osteoporotic individuals.[7] This measure is affected by cohort bias due to the ambiguities in the degree of osteoporosis, which makes it impossible to compare standardized precision errors across different studies. A cross-sectional comparison of subjects with and without osteoporosis is problematic for assessing longitudinal sensitivity for another reason: the osteoporotic individuals may have had a low peak skeletal status to start with and therefore under these circumstances standardization based on the average difference of healthy and osteoporotic individuals would overestimate true longitudinal responsiveness.

Extensions of the concepts

While the proposed concepts avoid a number of the problems addressed above, a few caveats need to be noted. First of all, it is impossible to characterize longitudinal sensitivity by a single universally applicable figure of merit. Only together will the change criteria LSC (or TAM), the follow-up time intervals MTI (or TAI), and the standardized precision error provide the answers sought. More fundamentally, one could criticize the proposed approach because it does not consider whether the observed change in a bone parameter, even if highly significant, would relate to a relevant change in fracture risk. Ross et al. have alluded to this problem.(20) This does represent a limitation; however, the relationships between changes in a bone parameter and subsequent changes in fracture risk have not been well established to date. Increased bone loss can be a risk factor in itself or because of the expected extrapolated long-term reductions in BMD. Moreover, such a concept would reduce the relevance of bone loss measurements to simply the risk-related aspect, whereas for clinical decision making, assessment of the efficacy of therapy or compliance may play a more important role.

There are a number of assumptions to using the proposed concept. The underlying bone parameters are considered to be normally distributed. For calculating long-term precision errors and response rates, the changes are assumed to be linear with time. For response to treatment, this is usually not the case. However, the proposed concepts could be easily adapted. Nonlinear changes can be divided into piecewise linear segments. Compared with later responses, the large early response to treatment would result in shorter MTIs (or TAIs). Long-term precision errors could also be calculated from nonlinear models, should this make biological and statistical sense. Whether this offers advantages remains to be seen. Also, one needs to acknowledge that, irrespective of the type of model selected, the standard error of the estimate (SEE; see Appendix 1) includes two components of variability, i.e., technique imprecision and true deviations from the fit. Therefore, prospectively defined standardized precision errors do not solely represent true technique limitations. In this regard, the term "precision error" may be considered misleading and the alternative term "longitudinal sensitivity" may be preferable. However, for most clinical applications, this ambiguity does not represent a problem. If one is, for example, interested in estimating the follow-up time required to establish success of treatment, the power to detect this will depend both on the technique's imprecision and the true variability over time.(21) Thus, the SEE can be considered to represent a good approximation of overall diagnostic, biological, and therapeutic variability.

Statistical tests like the ones proposed in this paper might be incorporated in the device's operating software. For example, in serial measurements, an automatic indication whether a change from previous exams is significant could aid the clinician in the process of decision making.

Application of the concept

The above mentioned advantages and disadvantages of the parameters of the concept are demonstrated by the data shown in Table 1. As can be seen, the performance (i.e., ability to monitor changes) of the technique cannot be judged based on uncorrected precision errors since the response rates vary substantially. The change criteria can be used directly to determine which changes reflect trends (TAM) or significant changes (LSC). The follow-up times TAI and MTI and stdPE reflect both precision errors as well as responsiveness to changes. stdPE is less dependent on the subject group than MTI (or TAI) and thus is closer to the goal of defining a single parameter that characterizes the overall performance of a technique. MTI (TAI) will usually be different for each subject group and technique since they are meant to be direct indicators of follow-up times and will be subject group dependent.

The results of Table 1 are based on short-term precision errors and thus need to be interpreted with caution since they will likely underestimate long-term stdPE_{lt}. In this hypothetical example, the longitudinal sensitivity of $\text{BMD}_{\text{spine}}$ or $\text{BMD}_{\text{femtot}}$ is better than that of either of the two QUS parameters.

CONCLUSIONS

A comprehensive assessment of the longitudinal sensitivity of a technique should be based on calculation of a change criterion (like TAM or LSC), a follow-up time interval (like TAI or MTI), and a standard precision error (stdPE). Together these three measures yield a good characterization of a technique's ability to monitor skeletal changes: LSC is the smallest change to be considered statistically significant, the patient should be called in at about the time interval specified by the (population specific) MTI, and the smaller the stdPE the more sensitive the technique. For matters of clinical decision making that require or allow earlier judgement at lower levels of statistical significance, i.e., trend assessment, shortening the follow-up time interval by 36% (next visit after TAI instead of MTI) may be adequate.

Some of the previous methods of standardization of precision have been shown to represent cases of flawed application (amplification of the cohort effect) of a useful concept (standardization) to a parameter that has sometimes been misinterpreted in the past (precision, as a parameter that for the purposes discussed here is not valuable in itself but only in conjunction with good responsiveness). The presented concept should improve the ability to investigate, characterize, and compare the ability of techniques to monitor changes in skeletal status.

ACKNOWLEDGMENT

I would like to acknowledge the helpful discussions with Richard Eastell, M.D., Sheffield, U.K., Ying Lu, Ph.D., San Francisco, CA, U.S.A., and Harry Genant, M.D., San Francisco, CA, U.S.A.

REFERENCES

1. Genant HK, Engelke K, Fuerst T, Glüer CC, Grampp S, Harris ST, Jergas M, Lang T, Lu Y, Majumdar S, Mathur A, Takada M 1996 Noninvasive assessment of bone mineral and structure: State of the art. J Bone Miner Res **11:**707–730.
2. Davis JW, Ross PD, Wasnich RD, MacLean CJ, Vogel JM 1991 Long-term precision of bone loss rate measurements among postmenopausal women. Calcif Tissue Int **48:**311–318.
3. Miller CG, Herd RJM, Ramalingam T, Fogelman I, Blake GM 1993 Ultrasonic velocity measurements through the calcaneus: Which velocity should be measured? Osteoporos Int **3:**31–35.
4. Blumsohn A, Hannon RA, Al-Dehaimi AW, Eastell R 1994 Short-term intraindividual variability of markers of bone turnover in healthy adults. J Bone Miner Res **9** (Suppl 1)**:**S153.
5. Langton CM 1997 ZSD: A universal parameter for precision in the ultrasonic assessment of osteoporosis. Physiol Meas **18:**67–72.
6. Glüer CC, Blunt B, Engelke K, Jergas M, Grampp S, Genant HK 1994 'Characteristic follow-up time'—A new concept for standardized characterization of a technique's ability to monitor longitudinal changes. Bone Miner **25** (Suppl 2)**:**S40.
7. Machado ABC, Hannon R, Henry Y, Eastell R 1997 Standardized coefficient of variation for dual energy x-ray absorptiometry (DXA), quantitative ultrasound (QUS) and markers of bone turnover. J Bone Miner Res **12** (Suppl 1)**:**S258.
8. Greenspan SL, Bouxsein ML, Melton ME, Kolodny AH, Clair JH, DeLuca PT, Stek M, JrFaulkner KG, et al. 1997 Precision and discriminatory ability of calcaneal bone assessment techniques. J Bone Miner Res **12:**1303–1313.
9. Cummings SR, Black D 1986 Should perimenopausal women be screened for osteoporosis? Ann Intern Med **104:**817–823.
10. Genant HK, Block JE, Steiger P, Glüer CC, Ettinger B, Harris ST 1989 Appropriate use of bone densitometry. Radiology **170:**817–822.
11. Heaney RP 1986 En recherche de la différence (P < .05). Bone Miner **1:**99–114.
12. Lufkin EG, Wahner HW, O'Fallon WM 1992 Treatment of postmenopausal osteoporosis with transdermal estrogen. Ann Intern Med **117:**1–9.
13. Riis BJ, Thomsen K, Strøm V, Christiansen C 1987 The effect of percutaneous estradiol and natural progesterone on postmenopausal bone loss. Am J Obstet Gynecol **156:**61–65.
14. Riis B, Christiansen C 1987 Prevention of postmenopausal osteoporosis by estrogen/ gestagen substitution therapy. Med Klin **82:**238–241.
15. Liberman UA, Weiss SR, Bröll J, Minne HW, Quan H, Bell NH, Rodriguez-Portales J, Downs RW, et al. 1995 Effect of oral alendronate on bone mineral density and the incidence of fractures in postmenopausal osteoporosis. N Engl J Med **333:**1437–1443.
16. Faulkner KG, Roberts LA, McClung MR 1996 Discrepancies in normative data between Lunar and Hologic DXA systems. Osteoporos Int **6:**432–436.
17. Fuleihan GE-H, Testa M, Angell JE, Porrino N, LeBoff MS 1995 Reproducibility of DEXA absorptiometry: A model for bone loss estimates. J Bone Miner Res **10:**1004–1014.
18. Nguyen TV, Sambrook PN, Eisman JA 1997 Sources of variability in bone mineral density measurements: Implications for study design and analysis of bone loss. J Bone Miner Res **12:**124–135.
19. Quan H, Shih WJ 1996 Assessing reproducibility by the within-subject coefficient of variation with random effects models. Biometrics **52:**1195–1203.
20. Ross PD, Davis JW, Wasnich RD, Vogel JM 1991 The clinical application of serial bone mass measurements. Bone Miner **12:**189–199.
21. Blake GM, Fogelman I 1997 Technical principles of dual energy x-ray absorptiometry. Semin Nucl Med **27:**210–228.
22. Glüer 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.

Address reprint requests to:
Claus-C. Glüer
Arbeitsgruppe Medizinische Physik
Klinik für Diagnostische Radiologie
Universitätsklinikum an der
Christian-Albrechts-Universität zu Kiel
Michaelisstrasse 9
D-24105 Kiel, Germany

Received in original form May 11, 1998; in revised form April 2, 1999; accepted April 5, 1999.

APPENDIX 1. GLOSSARY OF TERMS, ABBREVIATIONS, AND DEFINITIONS

LSC: Least significant change: $LSC = 2.8 \times PE_{lt}$
Criterion for smallest change in measurement results that can be considered to be statistically significant with 95% confidence (two-sided test). For statistical assumptions see.[9] Compare TAM.

MTI: Monitoring time interval: LSC/median response
Follow-up time interval after which the majority of patients can be expected to show a change exceeding the LSC, i.e., time interval recommended between follow-up visits if high 95% confidence level (two-sided test) is required. MTI is a characteristic of a technique but it depends on the subject group, e.g., disease progression (MTI_p), response to treatment (e.g., MTI_{estr} or MTI_{VitD}). Compare TAI.

PE_{st}: Short term precision error, expressed on a percentage basis: $PE_{st} = RMS(SD_i/mean_i)$
Derived from two or more measurements repeated at short time intervals and obtained on $i = 1..m$ individuals; see Appendix 2.

PE_{lt}: Long term precision error, expressed on a percentage basis: $PE_{lt} = RMS(SEE_i/mean_i)$
Derived from longitudinal studies of $i = 1..m$ individuals with a minimum of three repeated measurements per individual over time. See Appendix 2.

RMS: Root-mean-square average; averaging method appropriate for averaging of variances (e.g., precision errors) which are not normally distributed, but according to the F-distribution. See Appendix 2.

rr: Response ratio: rr(AvsR) = response rate (reference technique R)/response rate (technique A investigated), where response rates reflect %changes per annum or %changes per year of age for a given technique.

SD: Standard deviation of repeated measurements; measure of short term precision. Compare SEE.

SEE: Standard error of the estimate: measure of scatter around the regression line and, therefore, of long term precision. Compare SD.

sPE: Standardized precision error: sPE = PE × rr(AvsR) Precision error adjusted for response ratio rr of reference technique versus technique A investigated. Expressed on a percentage basis. As a result of the standardization procedure, the scaling of the standardized precision error is now equivalent to the scaling of the precision error of reference technique. Consequently, standardized precision errors of both techniques can now directly be compared.

stdPE: Standard precision error: StdPE = PE × rr(PA-DXAspine vs. technique A)
Standardized precision error for which PA-DXA of the spine was selected as the reference technique. Expressed on a percentage basis.

TAI: Trend assessment interval: TAM/median response
Follow-up time interval after which the majority of patients can be expected to show a change exceeding the TAM, i.e., time interval recommended if somewhat relaxed tests for change are sufficient. TAI is a characteristic of a technique but it depends on the subject group, e.g., disease progression (TAI_P), response to treatment (e.g., TAI_{estr} or TAI_{VitD}). Compare MTI.

TAM: Trend assessment margin: TAM = 1.8 × PE_{lt}
Criterion for smallest change in measurement results that can be considered to be statistically significant with 80% confidence (two-sided test) or 90% confidence (single-sided test). For statistical assumptions see.[9] Compare LSC.

APPENDIX 2. CALCULATION OF STANDARDIZED PRECISION ERRORS

Short term precision errors are calculated in the following way. For an individual, the absolute precision error is given by the standard deviations (SD) of repeated measurements. Expressed on a percentage basis, the short-term precision error $PE_{st,i}$ for the ith individual is given by:

$$PE_{st,i}[\%] = \frac{SD_i}{\bar{x}_i} = 100 \times \sqrt{\frac{\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2}{n_i - 1}} \Bigg/ \bar{x}_i$$

where x_{ij} is the bone parameter from the jth measurement of the ith individual and $\bar{x}_i$ the mean of n_i repeated measurements on that subject.

The average short-term precision error of a group of m individuals is not given by the arithmetic mean but rather by the root-mean-square average (RMS) of the precision errors of the individuals[22]:

$$PE_{st} = RMS\,(PE_{st,i}) = \sqrt{\frac{\sum_{i=1}^{m} (n_i - 1)\, PE_{st,i}^2}{\sum_{i=1}^{m} (n_i - 1)}}$$

Long-term precision errors can be calculated from linear regression analysis of an individual's measurements over time. The standard error of the estimate (SEE_i), which reflects the deviations of repeated measurements from the fitted curve, can be taken as a measure of the absolute long-term precision error of the ith individual. The individual's long-term precision error, $PE_{lt,i}$, when expressed on a percentage basis is given by:

$$PE_{lt,i}[\%] = \frac{SEE_i}{\bar{x}_i} = 100 \times \sqrt{\frac{\sum_{j=1}^{n_i} (x_{ij} - \hat{x}_{ij})^2}{n_i - 2}} \Bigg/ \bar{x}_i$$

where $\hat{x}_{ij} = a + bt_{ij}$ is the predicted value of the jth measurement in the ith individual at the time t_{ij} according to the fitted line with intercept a and slope b.

For a group of m individuals the average long term precision error PE_{lt} is then given by:

$$PE_{lt} = RMS\,(PE_{lt,i}) = \sqrt{\frac{\sum_{i=1}^{m} (n_i - 2)\, PE_{lt,i}{}^2}{\sum_{i=1}^{m} (n_i - 2)}}$$

Standard (or standardized) precision errors are derived from the precision errors defined above by first multiplying the individual's short- or long-term precision error with the response ratio, and then, second, calculating the RMS average across all subjects. The response ratio, rr, can be derived in the following fashion.

(1) Longitudinal studies (preferred approach):

$$rr_{L,i} = \frac{(\%\text{slope per annum of reference technique})_i}{(\%\text{slope per annum of technique investigated})_i}$$

% slope is the slope of the regression line (i.e., the precent change per annum) of an individual's measurements over time. As a measure of the response observed for this individual, it is calculated for the technique investigated and the reference technique—both obtained in this individual—to calculate the response ratio. For this method, unlike for the two following ones, the response ratio is specific to each individual and it can be used for estimating longitudinal sensitivity for response to treatment.

(2) Normative data:

$$rr_N = \frac{\begin{array}{c}\%\text{slope per year of normative data}\\ \text{for reference technique}\end{array}}{\begin{array}{c}\%\text{slope per year of normative data}\\ \text{for technique investigated}\end{array}}$$

Here, the response rates are based on the cross-sectional fit of age-related changes in normative data.

(3) Cross-sectional data (least desirable approach):

$$\mathrm{rr}_C = \frac{\%\text{slope per year of age for reference technique}}{\%\text{slope per year of age for technique investigated}}$$

This approach can be used if neither longitudinal response studies nor normative data are available. If the technique investigated and the reference technique have been obtained in the same individuals one would, separately for each of the two techniques, regress the parameter of the technique versus the age of the subjects included in order to obtain the slope per year of age as a measure of the response rate.

Standard precision errors are then calculated from either

$$\mathrm{stdPE_i} = \mathrm{PE_i} \times \mathrm{rr_{L,i}} \text{ for longitudinal studies or}$$

$$\mathrm{stdPE_i} = \mathrm{PE_i} \times \mathrm{rr_{N(or\ C)}} \text{ for cross sectional studies}$$

The standard precision error averaged across a group of *m* individuals stdPE is then given by:

$$\mathrm{stdPE} = \mathrm{RMS}(\mathrm{stdPE_i}) = \sqrt{\frac{\sum_{\mathrm{i}=1}^{m} (n_\mathrm{i} - 2)\mathrm{stdPE_i}^2}{\sum_{\mathrm{i}=1}^{m} (n_i - 2)}}$$

Depending on the data type, i.e., short-term or long-term precision errors, longitudinal or cross-sectional study design, different types of stdPE could be calculated. The preferred approach, in which both long-term precision errors and response rates derived from longitudinal studies, is denoted as $\mathrm{stdPE_{lt}}$:

$$\mathrm{stdPE_{lt}} = \mathrm{RMS}(\mathrm{stdPE_{L,lt,i}}) = \sqrt{\frac{\sum_{\mathrm{i}=1}^{m} (n_\mathrm{i} - 2)\mathrm{stdPE_{L,lt,i}}^2}{\sum_{\mathrm{i}=1}^{m} (n_\mathrm{i} - 2)}}$$