Osteoporos Int (2008) 19:1589–1596
DOI 10.1007/s00198-008-0615-1

ORIGINAL ARTICLE

Optimal decision criterion for detecting change in bone mineral density during serial monitoring: A Bayesian approach

M. Sadatsafavi · A. Moayyeri · L. Wang · W. D. Leslie

Received: 17 October 2007 / Accepted: 5 February 2008 / Published online: 22 April 2008

Abstract
Summary Interpretation of change in serial bone densitometry using least significant change (LSC) may not lead to optimal decision making. Using the principles of Bayesian statistics and decision sciences, we developed the Optimal Decision Criterion (*ODC*) which resulted in 11–12.5% higher rate of correct classification compared with the *LSC* method.
Introduction The interpretation of change in serial bone densitometry emphasizes using least significant change (LSC) to distinguish between true changes and measurement error.
Methods Using the principles of Bayesian statistics and decision sciences, we developed the optimal decision criterion (ODC) based on maximizing a 'utility' function that rewards the correct and penalizes the incorrect classification of change. The relationship between LSC and ODC is demonstrated using a clinical sample from the Manitoba Bone Density Program.
Results Under certain conditions, it can be shown that using LSC at the 95% confidence level implicitly equates the benefit of 39 true positive diagnoses with the harm of one false positive classification of BMD change. ODC resulted in an 11% higher rate of correct classification for lumbar spine BMD change and a 12.5% better performance for classifying total hip BMD change compared with LSC with this method.
Conclusions ODC has the same clinical interpretation as LSC but with two major advantages: it can incorporate prior knowledge of the likely values of the true change and it can be fine-tuned based on the relative value placed on the correct and incorrect classifications. Bayesian statistics and decision sciences could potentially increase the yield of a BMD monitoring program.

Keywords Bone densitometry · Dual-energy X-ray absorptiometry · Osteoporosis · Precision

M. Sadatsafavi
Center for Clinical Epidemiology and Evaluation,
Vancouver Coastal Health Institute,
Vancouver, BC, Canada

A. Moayyeri
Department of Public Health and Primary Care,
University of Cambridge,
Cambridge, UK

L. Wang
Department of Statistics, University of Manitoba,
Winnipeg, MB R2H 2A6, Canada

W. D. Leslie (✉)
Department of Medicine (C5121), University of Manitoba,
409 Tache Avenue,
Winnipeg, MB R2H 2A6, Canada
e-mail: bleslie@sbgh.mb.ca

Introduction

Bone density measurement has a primary clinical role in the initial diagnostic and fracture risk assessment of osteoporosis [1, 2] and is also widely used for serial monitoring of patients with suspected or confirmed osteoporosis [3]. Serial measurement of bone mineral density (BMD) is recommended for detecting bone loss in susceptible individuals and for monitoring the efficacy of treatment [3, 4]. Despite the fact that methods like dual energy X-ray absortiometry (DXA) have very good precision compared with many other biologic measurements, the rate of change in BMD is slow. Therefore, the longitudinal skeletal changes that occur over a short follow-up time (1–3 years) are confounded by the measurement error of the technique [5, 6].

The International Society of Clinical Densitometry (ISCD) recommends that measurement error be estimated

by an in vivo study at each densitometry center. Based on the root mean square standard deviation (SD) of the measurement error, a threshold value, called the least significant change (LSC) should be calculated and used to distinguish between real change and measurement noise [3, 4]. In statistical terms, LSC is the critical value to test the frequentist null hypothesis that no change has occurred at a given confidence level. Although the frequentist theory of statistics is well developed and has deep roots in experimental sciences, it is now widely acknowledged that decisions based on frequentist hypothesis testing do not result in the most efficient use of the available evidence [7]. For instance, consider a clinical trial in which the treatment effect of the new drug has not reached statistical significance. Under classical frequentist approach, the new treatment has failed to show any superior effect and as such it should not replace the standard treatment. Nevertheless, the new treatment might still be a better alternative when the information from similar clinical trials (or expert opinion) is considered and the treatment effect, costs, and adverse events are analyzed in a cohesive framework. Consequently, it has been advocated that medical research should move from the idea of frequentist significance testing to the interpretation of the findings in the context of the experiment in order to maximize the health outcomes per unit of resource used [8]. This can be achieved using Bayesian analysis. A Bayesian approach to the interpretation of an observation would use all relevant information available before the observation, which can be derived from the literature, expert opinion, or from the records of previous patients with similar characteristics, to construct a priori knowledge. This knowledge is then updated according to the current observation(s) to build a posteriori knowledge. The combination of previous information and current observation often results in more powerful inference regarding the unobserved parameter of interest [7].

Although the majority of arguments criticizing the classical frequentist theory in clinical research have stemmed from inferences in clinical trials, we argue that the same logic applies to BMD monitoring. Decisions based on an inherently arbitrary confidence level (such as the LSC) do not guarantee that maximum benefit is achieved from BMD testing. The ISCD recommendation on the choice of 95% as the confidence level (compared with other options like 80%) was based on the argument that it is 'the most widely used in clinical densitometry' and because 'it represents a higher standard, and it potentially decreases the number of patients being considered for diagnostic evaluation because of apparent loss of BMD while on therapy' [3]. Both arguments are subjective, and the position does not consider the fact that setting a higher confidence level increases the chance that a true change would be missed. Of course, the main purpose of monitoring is to detect those with unfavorable BMD change (decrease); hence the sensitivity of a policy in detecting BMD decline is of paramount importance, a fact that is not reflected in current recommendation.

An inclusive search for the optimal interpretation of an observed change entails quantification of the outcome of BMD monitoring in terms of both costs (e.g., further tests, medications, and fractures) and health effects (e.g., the impact of fracture on survival and the quality of life of patients). Unfortunately, there are major obstacles in the valuation of such outcomes in BMD monitoring. For example, it has been shown that the beneficial impact of treatment on fracture risk is not necessarily mediated through an increase in BMD [9]. Lack of clarity in the benefits of serial BMD monitoring should not, however, preclude a rigorous analysis based on the best information available. Reasonable assumptions and expert opinion could lead one closer to adopting an optimal policy.

In the present work we look at this problem from a broad perspective: we quantify the performance of a BMD monitoring program by defining a 'utility function' that assigns a numerical score to correct and incorrect classification, and then find the policy that maximizes the overall utility. The utility function can be based on expert opinion, or if the benefits and costs of serial monitoring are quantified it can be updated to reflect realistic gains from BMD monitoring. The analysis is inherently Bayesian, as quantifying the probability of correct and incorrect classification requires that the true change in BMD be considered as a random variable. The resulting decision threshold, which we call the optimal decision criterion (ODC, units g/cm^2), has the same clinical interpretation as LSC: if the magnitude of the observed change (defined as the second minus the first BMD) is below this criterion, the subject is deemed to have had an unfavorable change in BMD.

Methods

General context

Our proposed method for classification of BMD change is based on the following assumptions:

1. Measurement error has a normal distribution with zero mean, and its variance is independent of BMD change.
2. Given a 'true' change in a patient, we can classify this change either as favorable or unfavorable.
3. The magnitude of the measurement error (SD) is known with reasonable accuracy.
4. We are able to express our prior knowledge of the likely change in the BMD of a patient as a normal distribution.

The current ISCD recommendation is based on assumptions 1 and 3; in the case of n_1 baseline and n_2 follow-up measurements and σ_e the measurement error estimated from the in vivo precision study, the ISCD-recommended LSC would be [5, 10]:

$$LSC = Z'\left(\frac{1+\alpha}{2}\right)\sigma_e\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \quad (1)$$

where Z′ is the inverse of the cumulative density function of the standard normal distribution, and α is the desired two-tailed confidence level (equivalently *1-*α is called the significance level). For the most common situation where n_1=n_2=1 and α=95% this amounts to LSC=2.77 x σ_e.

We denote the true change in BMD as x, and the cutoff that we would use, had we known the true change, to distinguish the favorable change from unfavorable change as T_x. Unfortunately measurement error does not let us utilize T_x in our decision. Instead, we observe a BMD change (denoted by y) which is confounded by measurement error. We will make our decision based on T_y, a threshold on the observed change. It might be suggested that as the measurement error has zero mean, the thresholds on the observed and true changes should be equal (T_y=T_x). This is generally not the case, however, as the following analyses suggest. Also, unlike frequentist analysis, there is no null hypothesis defined as 'no change in BMD'. In the Bayesian context, zero change is not a special condition and it can be unfavorable or favorable depending on T_x.

Specific context

The first step in finding the optimal criterion is to construct a posterior distribution of the true change given the observed change Y and our prior information on x. This is the conditional distribution of x on Y. Using the principles of Bayesian statistics [11], we have

$$x|Y \sim N\left(\frac{\sigma_p^2\mu_x+\sigma_x^2 Y}{\sigma_x^2+\sigma_p^2}, \sqrt{\frac{\sigma_p^2\sigma_x^2}{\sigma_x^2+\sigma_p^2}}\right) \quad (2)$$

with

$$\sigma_p = \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}.\sigma_e \quad (3)$$

σ_p is the SD of the measurement error in current experiment and n_1 and n_2 are the number of baseline and follow-up measurements for the present subject, respectively. In practice often n_1=n_2=*1*; hence $\sigma_p = \sqrt{2.\sigma_e}$.

Eq. 2 is the distribution introduced by Nguyen et al. to construct Bayesian credible interval around the point estimate of change in BMD [11]. One can use the tail probabilities of the normal distribution with the above parameters to calculate the critical value of Y such that if the observed Y is below this value, the probability that BMD has truly decreased is above a certain confidence level. The absolute value of the critical value on Y is always smaller than the LSC. Therefore, a smaller observed change in BMD in the ODC approach is needed to be considered significant, indicating the additional power provided by incorporating prior knowledge.

The next step in determining the optimal policy is to calculate the probability of correct and incorrect decision given an observed change. Defining an unfavorable true change as the one that falls below the actual threshold (x<Tx), and a positive detection (of BMD decline) as an observed change below our defined threshold (y<Ty), four conditions can arise:

True positive (TP): $(x < Tx, y < Ty)$
False positive (FP): $(x > Tx, y < Ty)$
True negative (TN): $(x > Tx, y > Ty)$
False negative (FN): $(x < Tx, y > Ty)$

In order to calculate an overall score for an approach in categorization of an observed BMD change, we need to assign a numerical score to each of the above conditions. Obviously, TP and TN are favorable and receive higher scores while FP and FN are unfavorable situations and are given lower values. The utility function is then defined as the weighted sum of scores across quadrants with weights being the probability that the subject will fall into that quadrant. Namely:

$$U = P_{TP}{*}S_{TP} + P_{FP}{*}S_{FP} + P_{TN}{*}S_{TN} + P_{FN}{*}S_{FN} \quad (4)$$

S and P correspond to the score of each quadrant and probability that a subject would be in that quadrant, respectively. Calculation of P_{TP}, P_{FP}, P_{TN}, and P_{FN} are presented in Appendix 1. The optimal decision criterion is the T_y that maximizes the above utility function. It is shown in the Appendix 2 that ODC can be written as:

$$ODC = T_x + \frac{\sigma_p^2}{\sigma_x^2}(T_x-\mu_x) - Z'\left(\frac{K}{K+1}\right)\frac{\sigma_p}{\sigma_x} \times \sqrt{\sigma_x^2+\sigma_p^2} \quad (5)$$

K in the above formula is defined as $\frac{S_{TN}-S_{FP}}{S_{TP}-S_{FN}}$ and can be interpreted as the number of true positive diagnoses needed to compensate for the harm of one false positive classification of change. For example, K=*2* if we accept that the harm caused by incorrectly identifying and treating a favorable change as an untoward change is equal to the benefit of correctly identifying two unfavorable changes. Alternatively, the term $\frac{K}{K+1}$ in Eq. 5 can be interpreted as the 'treatment threshold' as it is known in clinical decision making [12], and is the threshold on the probability of the disease (unfavorable BMD change in this context) above

which treatment is justified. Therefore, to find the optimal threshold, it turns out that instead of four scores we only need a single metric, K, with a straightforward clinical interpretation, though it is still necessary that a score be assigned to each quadrant so that the absolute utility for a given policy can be calculated.

The above equation can also be solved for K to find the implicit value of K for a policy based on an arbitrary threshold T_y. For example, one can use LSC as T_y in the above equation to find out the implicit value of K for a BMD monitoring policy based on that LSC.

Relationship between ODC and LSC

Assuming that the treatment threshold is zero (T_x=0) and we have no prior information about the true change ($\sigma_x \to +\infty$), Eq. 4 reduces to the *LSC* formula (Eq. 1) with the only difference being that the confidence level is now a function of K:

$$K = \frac{1+\alpha}{1-\alpha} \tag{6}$$

This connects the arbitrary confidence level of LSC to the utility function and enables us to examine the benefit of recommendations based on LSC at a certain confidence level. For example, the value of K that corresponds to 95% confidence level, the most widely used value for LSC is 39. In other words, by using LSC with 95% confidence level, and assuming no prior information on BMD change, one is implicitly equating the harm of one false positive detection of change with the benefit of true detection of change in 39 patients. The value of K for 80% confidence level is 9.

A very important situation is when K is set to 1, in which case we are implicitly assuming that the benefit of one true positive categorization is equal to the harm of one false positive categorization. A policy based on such utility function actually maximizes the proportion of correct classification. The proportion of correct classification for such a policy is estimable from the equations provided in Appendix 1 by assigning a score of 1 for TP and TN and 0 for FP and FN. A sample HTML calculator for Bayesian interpretation of BMD change using the ODC approach is available from the authors on request.

Example using a clinical dataset

We evaluated the ODC approach to classification of unfavorable BMD change (decrease) using data from the Manitoba Bone Density Program [13]. The program provides all bone density services to the population of Province of Manitoba, Canada, and maintains an electronic database of all clinical DXA examinations performed since 1990 [14]. The DXA scans in our densitometry clinics are performed and analyzed in accordance with the manufacturer's recommendations. A pencil-beam instrument (Lunar DPX, GE Lunar, Madison, WI) was the primary instrument used prior to 2000 and a fan-beam instrument (Lunar Prodigy, GE Lunar, Madison, WI) was used after that date. All densitometers underwent daily assessment of stability using an anthropomorphic spine phantom and each showed excellent long-term phantom stability (coefficient of variation [CV] <0.5%).

Prior information on the probability of true change was elicited by analyzing data from all individuals who had baseline and follow up BMD measurements on the same instrument between 1994 and 2002 (mean interval between scans 21±9 months). We excluded cases where scanning was performed on different instruments and those that did not report the lumbar spine and the total hip BMD. A total of 1420 scan-pairs were available for the analysis and is referred to as the 'clinical monitoring population'. For simplicity, analysis is performed for the overall patient population and important factors affecting the prior knowledge (e.g., interval between scans, gender, and treatment) were ignored. Measurement error was estimated based on 198 lumbar spine scan-pairs and 193 total hip scan-pairs obtained from an independent convenience sample of female individuals referred for bone density testing (mean interval between scans 6±5 days) and usually done by two different technologists, referred to as the 'reproducibility population'. The clinical monitoring and reproducibility populations have been previously described [15]. This report was reviewed and approved by the facility's Office for Clinical Research.

Results

The SD of the measurement error estimated from the monitoring sample was 0.0174 g/cm^2 for the lumbar spine and 0.0094 g/cm^2 for the total hip, which gives a 95% confidence level *LSC* of 0.0482 g/cm^2 and 0.0260 g/cm^2, respectively. The distribution of BMD change in both lumbar spine and hip conformed to the assumption of normality (Fig. 1), with mean 0.0148 g/cm^2 and SD 0.0482 g/cm^2 for the lumbar spine and mean 0.0067 g/cm^2 and SD 0.0376 g/cm^2 for the hip. However, it should be kept in mind that the distribution of change in the clinical monitoring database is also affected by measurement error, and therefore to estimate the SD of the true change the effect of measurement error should be removed from the estimate. Since we assume the measurement error and BMD change in clinical population are independent, and according to the theorem that the variance of the sum of independent variables is the sum of their variances [16], such adjustments can be performed by simply subtracting

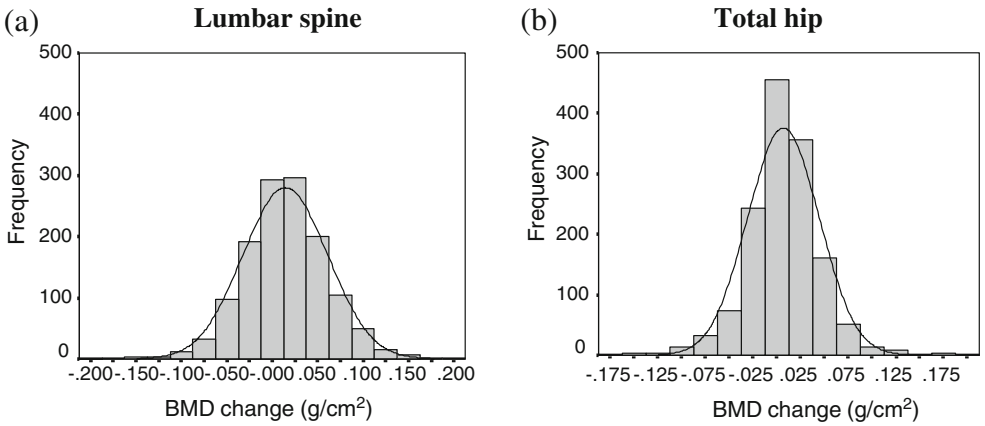

Fig. 1 Distributions of the observed BMD change (g/cm^2) in the clinical monitoring population for (**a**) the lumbar spine and (**b**) the total hip with fitted normal distribution curves

twice (due to two measurements) the variance of measurement error from the variance of change in the clinical sample. The adjusted SDs for the spine and hip are therefore 0.0414 g/cm^2 and 0.0352 g/cm^2, respectively.

Setting the parameters from our dataset in Eq. 5, with $K=1$, the *ODC* would be -0.0052 g/cm^2, and based on equation in Appendix 1 with $S_{TP}=S_{TN}=1$ and $S_{FP}=S_{FN}=0$, results in 84.1% correct classification. The classic 95% confidence level LSC results in 73.1% correct classifications. Therefore, compared with LSC, ODC has 11% higher rate of correct classification. The same analysis on the hip site results in ODC of -0.0010 g/cm^2, with 88.9% correct classification, which represents12.5% better performance compared with LSC.

The relationship between *ODC*, *LSC*, and different values of K for the monitoring population is illustrated in Fig. 2. The left and right panels show the relationships for lumbar spine and hip, respectively. The intersection of the curve and horizontal lines is the value of K that corresponds to *LSC* at that level. For the lumbar spine, the 95% *LSC* and *ODC* are equal when $K=14.1$, and 80% *LSC* and *ODC* are equal when $K=4.6$. In other words, by using *LSC* at 95% confidence level, one is accepting one false positive classification of change for each 14.1 true positive classification. Similarly for the hip, the 95% and 80% *LSC* are equal to *ODC* at $K=24.8$ and 6.8, respectively.

Discussion

The conceptual framework for the interpretation of serial BMD measurements presented here deviates from the LSC approach as recommended by several groups, including the ISCD, in two major ways: first, it is based on incorporating the 'prior' knowledge of the true change, and second, the cutoff point placed on the observed change is based on an optimization algorithm. As suggested in our clinical examples, both changes could help improve the yield of BMD monitoring policies. Using the proportion of correct classification of increase/decrease in BMD as the utility function, we showed that ODC resulted in more than 10% higher rate of correct classification for both lumbar spine and total hip BMD change in our clinical monitoring sample. In addition, the stronger theoretical foundations of

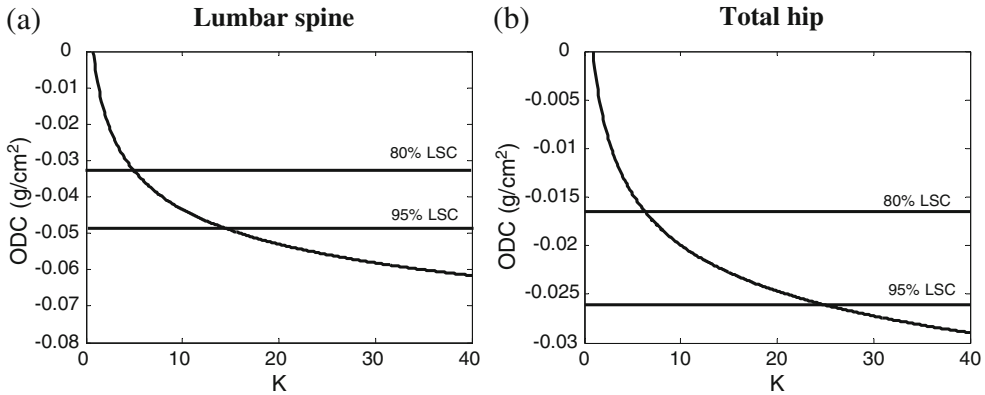

Fig. 2 Relation between the optimal decision criterion (ODC) and K (number of true positive diagnoses needed to compensate for the harm of one false positive classification of change) for (**a**) lumbar spine BMD change and (**b**) total hip BMD change. The curve illustrates the value of *ODC* corresponding to each value of K in the clinical monitoring population of the Manitoba bone density program. The intersection of the curve with horizontal lines is the implicit value of K for the 80% and 95% confidence-interval least significant change (LSC)

ODC compared with LSC, and the ability of ODC to incorporate clinicians' prior knowledge as well as their tradeoff between correct and incorrect diagnosis, can lead to more insightful and confident decisions.

The ability to incorporate the prior knowledge on the likely outcome of the experiment has been simultaneously considered the strength and weakness of Bayesian statistics [17]. On one hand, at many times we already know much about the likelihood of the outcome we are going to observe and considering such knowledge would enhance our power in making inference based on the results of the observation. On the other hand, the heuristic principles implemented by the decision maker to generate subjective probability (namely the availability, representativeness, and anchoring mechanism) are known to be prone to several biases [18]. However, there are at least three arguments in favor of incorporating prior knowledge in the context of a BMD monitoring program. The first and foremost is the crucial difference between a BMD screening practice and a large sample epidemiological experiment. Unlike formal epidemiological studies where the large sample size often attenuates the impact of prior belief, each BMD measurement amounts to a very small study with very limited power, and an efficient use of prior knowledge could significantly increase the power of the decision maker. Second, there is often enough data, either from past patients with the similar characteristics or from the published literature, which could be used to derive prior distribution without losing the credibility of results as if they are based on subjective and debatable priors. Finally, it should be noted that the current frequentist approach is not assumption-free. As shown in this work, it amounts to using a uniform distribution that can take any value for the prior knowledge of BMD change. While an informative prior might be biased at some instances, a distribution that can accommodate extreme and biologically implausible values is definitely suboptimal.

Many clinical densitometry centers keep records of the past measurements along with important characteristics of patients. These data could be used to check the assumption of normality and to estimate the mean and variance of the prior distribution in different subgroups of patients (e.g., early and late postmenopausal women, patients on glucocorticoids, or patients on antiosteoporotic treatments) or in the whole population if practicality is a concern. Such an approach requires that future and past subjects belong to the same population so that the parameters of the distribution estimated from the past subjects are applicable to the prospective cases. In the absence of any major change in the clinical setting, this seems to be a reasonable assumption. On the other hand, prior distributions could be based on the results of epidemiologic findings or expert opinion. For example, Nguyen et al. [11] have estimated the mean and SD of change in the clinical population from a meta-analysis of alendronate clinical trials [19], and an another study [20] has rigorously evaluated the mean and SD of the annual rate of change in BMD by gender and age groups from the Dubbo Osteoporosis Epidemiology Study [21]. These data can be used as parameters of the prior distribution for the method described in this work. This has the added advantage of bringing homogeneity to the reports from different centers. Care should be taken, however, as various factors such as differences in patients' characteristics could potentially invalidate such inferences.

The Bayesian approach in interpretation of BMD and its change has previously been proposed by Nguyen et al. [8], who proposed combining the distribution of change in the population and measurement error to construct Bayesian confidence intervals (credible intervals). However, the calculated interval is still used to test the hypothesis that no change has occurred, and the authors do not proceed towards defining a cutoff on the observed change based on the relative valuation of the correct and incorrect classification.

One of the findings of our study was that under certain conditions there is a direct relation between the significance level in LSC and implicit assumption on the relative value of true positive versus false positive classification. The two-tailed 95% confidence level currently recommended results in a policy that under the implicit assumption of no prior information equates the benefit of correct detection of change (K in our notations) in 39 patients with the harm of misdetection of one truly negative change. When a more rational prior distribution was selected based on our clinical database, the current ISCD recommendation equated 14 and 25 true positives with one false positive classification for the lumbar spine and the hip, respectively. Although it seems that in most clinical scenarios the harm of false positive detection outweighs the benefit of the true positive detection, and therefore K is more than one, these ratios are quite high and the decision maker might be willing to accept higher false positive rates in return for greater sensitivity to identify BMD change.

The aim of this work was to develop the methods and theoretical framework for a Bayesian approach to classification of BMD change, and any recommendation based on these results should also consider practicality and ease of use. Recognizing the difficulties of using different approaches for different subgroups, the ISCD expert panel recommended using *LSC* with fixed significance level for all scenarios [3]. Here it is also possible that a guideline based on a unique ODC be proposed that applies to most clinical contexts. This, however, should not discourage a competent clinician from using the full power of this approach based on his or her own prior knowledge and utility function.

We are aware of shortcomings in our analysis. Although the assumption of a normal distribution on the prior knowledge of change does not seem to be restrictive for a subjective assessment, the assumption of normality might not hold if prior distribution is estimated from real data. This was not the case, however, in our clinical data set. Another limitation of this approach is that it assigns the same score to all subjects falling in a specific quadrant. It might be argued that for patients whose BMD decline is much faster than average, correct identification of change and prompt management is more critical and the penalty of misclassification of these subjects should be higher than those with modest decline in BMD. This can theoretically be incorporated in a Bayesian approach. However, under such assumptions, a closed form solution for ODC might not exist, though numerical simulation could still be used especially in a one-time analysis that would inform a guideline. Another limitation is that it was assumed that measurement error is accurately known, while in practice it is usually estimated from an in vivo precision study with limited power. The impact of uncertainty in the measurement error was not analyzed in this study.

While different utility functions could be used for the optimization procedure, the most rational choice seems to be the one based on the net benefit approach [22]. In this method, the score of the true and false classification of changes is equal to their net benefit, which incorporates both the costs and health outcomes in a unified framework and thus is the most inclusive answer to the choice of the best policy [22]. It enables estimating the monetary value of using the optimal versus a non-optimal criterion and will also make possible comparing the benefit of the monitoring policy with other health interventions. As mentioned earlier, however, such approach requires that the effect of BMD monitoring practice on the important outcomes (e.g., fractures) be identified. In the absence of definitive evidence, a reasonable alternative is for an expert panel to make subjective judgments on the relative benefits/harms of detecting true positive and false positive changes to derive ODC for important subgroups of individuals undergoing BMD monitoring. Various techniques, such as the Delphi method [23], can be used to elicit the expert opinion in a rigorous way.

In summary, frequentist hypothesis testing in serial BMD monitoring using the LSC method has significant limitations. A more rigorous interpretation based on the principles of Bayesian statistics and decision sciences has the potential to significantly improve the outcome of BMD monitoring practice.

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

Funding none

Conflicts of interest None

Appendix

Appendix 1. Calculation of the probability of TP, TN, FP, and FN

Using the notations describes in the text and the general assumptions of our approach (see Methods), we have

$$x \sim N(\mu_x, \sigma_x)$$
$$y|x \sim N(x, \sigma_p)$$

Based on the properties of the normal distribution [24], and that e and x are assumed independent, the joint distribution of the x and y is a bivariate normal:

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim BVN\left(\begin{bmatrix} \mu_x \\ \mu_x \end{bmatrix}, \begin{bmatrix} \sigma_x^2 & \sigma_x^2 \\ \sigma_x^2 & \sigma_x^2 + \sigma_p^2 \end{bmatrix} \right)$$

The first and second parameters are the mean and covariance matrix of the bivariate normal distribution, respectively. The probability that a subject falls into each category can, therefore, be calculated using the cumulative distribution function of the bivariate normal distribution from the above equation. Defining the function $\Phi(a,b,\rho)$ as the cumulative distribution function of the standard bivariate normal distribution with at point (a, b) with correlation function ρ, we will have:

$$P_{TP} = \Phi\left(\frac{x-T_x}{\sigma_x}, \frac{y-T_y}{\sqrt{\sigma_x^2+\sigma_y^2}}, \frac{\sigma_x}{\sqrt{\sigma_x^2+\sigma_y^2}} \right)$$
$$P_{FP} = \Phi\left(\frac{x-T_x}{\sigma_x}, \frac{T_y-y}{\sqrt{\sigma_x^2+\sigma_y^2}}, \frac{\sigma_x}{\sqrt{\sigma_x^2+\sigma_y^2}} \right)$$
$$P_{TN} = \Phi\left(\frac{T_x-x}{\sigma_x}, \frac{T_y-y}{\sqrt{\sigma_x^2+\sigma_y^2}}, \frac{\sigma_x}{\sqrt{\sigma_x^2+\sigma_y^2}} \right)$$
$$P_{FN} = \Phi\left(\frac{T_x-x}{\sigma_x}, \frac{y-T_y}{\sqrt{\sigma_x^2+\sigma_y^2}}, \frac{\sigma_x}{\sqrt{\sigma_x^2+\sigma_y^2}} \right)$$

Appendix 2. Derivation of ODC

One way to derive ODC is to solve the derivate of the utility function in Eq. 4 for T_x. However, an easier way is to consider the utility of a decision based on observing change Y

$$U = \begin{cases} Y < T_y & z.S_{TP} + (1-z)S_{FP} \quad \text{(change detected)} \\ Y > T_y & z.S_{FN} + (1-z)S_{TN} \quad \text{(no change detected)} \end{cases}$$

while from Eq. 2:

$$z = Z\left(\frac{T_x - \frac{\sigma_p^2\mu_x + \sigma_x^2 Y}{\sigma_x^2+\sigma_p^2}}{\sqrt{\frac{\sigma_p^2\sigma_x^2}{\sigma_x^2+\sigma_p^2}}}\right)$$

The goal is to find T_y that results in the selection of the maximum of the two terms in U for all Y. This would be achieved by matching T_y to a Y for which the utilities for positive and negative classification are equal. For any Y below this critical value, our policy would detect an unfavorable change and the utility of our decision equals the first term, which is also the maximum of the two for the same Y given the fact it is a descending function of Y as long as $S_{TP}>S_{FP}$ and $S_{TN}>S_{FN}$. Setting the two terms equal and solving for Y, we will have:

$$z.S_{TP} + (1-z)S_{FP} = z.S_{FN} + (1-z)S_{TN} \Rightarrow z = \frac{S_{TN} - S_{FP}}{S_{TN} - S_{FP} + S_{TP} - S_{FN}}$$

Solving for Y yields Eq. 6 for ODC.

References

1. Cummings SR, Bates D, Black DM (2002) Clinical use of bone densitometry: scientific review. JAMA 288:1889–1897
2. Lewiecki EM, Watts NB, McClung MR et al (2004) Official positions of the international society for clinical densitometry. J Clin Endocrinol Metab 89:3651–3655
3. Lenchik L, Kiebzak GM, Blunt BA (2002) What is the role of serial bone mineral density measurements in patient management? J Clin Densitom 5(Suppl):S29–S38
4. Baim S, Wilson CR, Lewiecki EM et al (2005) Precision assessment and radiation safety for dual-energy X-ray absorptiometry: position paper of the International Society for Clinical Densitometry. J Clin Densitom 8:371–378
5. Bonnick SL, Johnston CC Jr, Kleerekoper M et al (2001) Importance of precision in bone density measurements. J Clin Densitom 4:105–110
6. Cummings SR, Palermo L, Browner W et al (2000) Monitoring osteoporosis therapy with bone densitometry: misleading changes and regression to the mean. Fracture Intervention Trial Research Group. JAMA 283:1318–1321
7. Sterne JA, Davey SG (2001) Sifting the evidence-what's wrong with significance tests? BMJ 322:226–231
8. Lilford RJ, Braunholtz D (1996) The statistical basis of public policy: a paradigm shift is overdue. BMJ 313:603–607
9. Chesnut CH III, Rosen CJ (2001) Reconsidering the effects of antiresorptive therapies in reducing osteoporotic fracture. J Bone Miner Res 16:2163–2172
10. Gluer CC, Blake G, Lu Y et al (1995) Accurate assessment of precision errors: how to measure the reproducibility of bone densitometry techniques. Osteoporos Int 5:262–270
11. Nguyen TV, Pocock N, Eisman JA (2000) Interpretation of bone mineral density measurement and its change. J Clin Densitom 3:107–119
12. Pauker SG, Kassirer JP (1980) The threshold approach to clinical decision making. N Engl J Med 302:1109–1117
13. Leslie WD, Metge C (2003) Establishing a regional bone density program: lessons from the Manitoba experience. J Clin Densitom 6:275–282
14. Leslie WD, Caetano PA, MacWilliam LR et al (2005) Construction and validation of a population-based bone densitometry database. J Clin Densitom 8:25–30
15. Leslie WD (2006) The importance of spectrum bias on bone density monitoring in clinical practice. Bone 39:361–368
16. Weiss NA (eds) (2005) A Course in Probability. Addison-Wesley
17. Van DS (2006) Prior specification in Bayesian statistics: three cautionary tales. J Theor Biol 242:90–100
18. Tversky A, Kahneman D (1974) Judgment under Uncertainty: Heuristics and Biases. Science 185:1124–1131
19. Karpf DB, Shapiro DR, Seeman E et al (1997) Prevention of nonvertebral fractures by alendronate. A meta-analysis. Alendronate Osteoporosis Treatment Study Groups. JAMA 277: 1159–1164
20. Nguyen TV, Sambrook PN, Eisman JA (1997) Sources of variability in bone mineral density measurements: implications for study design and analysis of bone loss1. J Bone Miner Res 12:124–135
21. Jones G, Nguyen T, Sambrook P et al (1994) Progressive loss of bone in the femoral neck in elderly people: longitudinal findings from the Dubbo osteoporosis epidemiology study. BMJ 309: 691–695
22. Hoch JS, Briggs AH, Willan AR (2002) Something old, something new, something borrowed, something blue: a framework for the marriage of health econometrics and cost-effectiveness analysis. Health Econ 11:415–430
23. Dalkey N, Helmer O (1963) An experimental application of the Delphi method to the use of experts. Manage Sci 9:458–467
24. Patel JK, Read CB (eds) (1982) Handbook of the normal distribution. New York, Dekker, M