J Musculoskel Neuron Interact 2004; 4(1):64-74

Hylonome

Perspective Article

Statistical validation of surrogate endpoints: Is bone density a valid surrogate for fracture?

Z. Li[1], A.A. Chines[2], M.P. Meredith[1]

[1] Department of Biometrics and Statistical Sciences,
[2] Research and Clinical Development, The Procter & Gamble Company, Mason, OH, USA

Abstract

In the treatment of osteoporosis using anti-resorptive agents there has been increasing interest in quantifying the relationship between fracture endpoints and surrogates such as bone mineral density (BMD) or bone turnover markers. Statistical methodology constitutes a critical component of assessing surrogate validity. Depending on study designs, data resources, and statistical methods used for analyses, one has to use caution when interpreting results from different analyses, especially when results are disparate. For example, analyses based on individual patient data reported that only a limited proportion of the anti-fracture efficacy was explained by BMD increases for agents such as alendronate, risedronate and raloxifene. Analyses employing meta-regression based on summary statistics, however, indicated that most of the anti-fracture benefits were due to improvements in BMD. In this paper, we review definitions of surrogate endpoints and requirements for their statistical validation. We evaluate whether BMD meets these requirements as a possible surrogate for fracture. Our review indicates that the actual BMD value is correlated with fracture risk and thus BMD is useful in identifying patients that might need treatment. There is limited evidence to support BMD increase with anti-resorptive agents as a reliable substitute for fracture risk reduction. Strengths and limitations for various statistical methods are discussed.

Keywords: Bisphosphonate, Correlation, Meta-analysis, Prediction, Risedronate

The fragility fracture problem

In clinical research for chronic diseases, the clinical endpoint is often disease occurrence or death. Such trials typically require a follow-up of numerous patients for years before yielding useful results. For example, in the treatment of osteoporosis for postmenopausal women using anti-resorptive agents, assessment of the effect of a new drug regimen on the incidence of new vertebral fractures is of primary importance in judging efficacy[1]. Trials with fractures as the primary endpoint require a study design that is either very large or very long in order to demonstrate the anti-fracture efficacy. Consider a population with fracture incidence of 10%-20%, a study requires approximately 470-1000 analyzable patients per group in order to have 90% power to detect 40% risk reduction compared to placebo. For an event with lower incidence such as hip fractures, the number of patients required to detect a clinically meaningful difference increases dramatically. This requirement has been reflected in the study designs of the pivotal studies for alendronate[2-3], risedronate[4-6] and raloxifene[7]. In contrast, bone mineral density (BMD) changes can be demonstrated in a study having relatively short duration and a much smaller number of patients. Investigators and clinicians may have an interest in making inferences about treatment effects of therapies on fractures based on the BMD outcome. Consequently, there has been increasing interest in trying to validate BMD change as a surrogate endpoint for fracture[8-12]. Postmenopausal osteoporosis is not the only area in which researchers are interested in exploring potential surrogates. For example, in the area of HIV infection and AIDS, CD4 counts have been studied as a possible surrogate endpoint[13]. Fleming and DeMets[14] provide a summary for additional therapeutic areas.

Over the last 50 years statistics have provided a frame-

All authors have corporate appointments with Procter & Gamble Pharmaceuticals. Dr. Arkadi Chines also has stock in Procter & Gamble Pharmaceuticals.

Corresponding author: Zhengqing Li, PhD, Department of Biometrics and Statistical Sciences, The Procter & Gamble Company, 8700 Mason-Montgomery Rd., Box 2199, Mason, OH 45040, U. S. A.
E-mail: li.z@pg.com

Accepted 7 November 2003

work for designing and analyzing clinical trials to ascertain the benefits of a treatment in clinical endpoints as well as to determine its effect on surrogate endpoints. Statistical efforts, however, for evaluating whether a biological parameter is a valid surrogate endpoint only began in earnest around 1989[15]. Various statistical approaches have been developed in an attempt to validate surrogate endpoints[16]. These approaches have both strengths and limitations. Most importantly, different approaches may yield different results. For example, to possibly provide more robust and precise estimates, meta-analyses have been recommended[17]. However, results from a meta-regression based on group-level summary statistics gleaned from published literature may not necessarily be consistent with results derived from a meta-analysis employing individual patient data. Analyses based on individual patient data have reported that only a limited proportion (4%-28%) of the anti-fracture efficacy is explained by BMD across three anti-resorptive agents including alendronate, risedronate and raloxifene[8-10]. In contrast, a meta-regression based on summary statistics across multiple agents from published literature reported that most of the anti-fracture effects for osteoporosis were due to improvements in BMD[11-12]. To understand this and other differences among various statistical approaches, a thorough review of relevant statistical methods is deserved.

In this paper, we provide a general review of definitions and relevant statistical validation methods for surrogate endpoints. The strengths and limitations of various statistical methods are discussed. Specifically, we review statistical analyses conducted to quantify the relationship between BMD and fractures. Perspectives are provided on whether BMD change may serve as a valid surrogate for fracture risk reduction based on evidence from these analyses and from the results of some new analyses we have conducted.

What is a surrogate?

Definition. Randomized clinical trials are the gold standard scientific method for evaluating a new drug, device, or procedures for prevention or treatment of disease in humans. In a clinical trial, one has to specify endpoints in the study protocol in order to answer the questions that investigators wish to explore. A clinical endpoint is a characteristic or variable that reflects how a patient feels or functions, or how long a patient survives[15]. Based on this definition, it is clear that a clinical endpoint must unequivocally reflect tangible benefit to patients[18], regardless of the therapeutic area. In the treatment of osteoporosis for postmenopausal women using anti-resorptive therapies, regulatory guidelines clearly indicate that an agent which preserves or enhances bone mass only provides suggestive evidence that it may reduce fracture risk; fracture studies must be run to document reduction of fracture incidence[1]. A study using clinical outcomes such as death or fracture as the primary endpoint typically requires either long study duration or a large sample size in order to demonstrate any meaningful clinical benefit. Researchers, however, want effective new treatments available to patients as quickly as possible, provided safety is adequately demonstrated. Surrogate endpoints constitute an effort to realize this latter goal.

Various definitions for surrogate endpoints have been proposed over the past 15 years. As defined by Temple[19],

> *a surrogate endpoint is a laboratory measurement or a physical sign used as a substitute for a clinically meaningful endpoint that measures directly how a patient feels, functions, or survives. Changes induced by a therapy on a surrogate endpoint are expected to reflect changes in a clinically meaningful endpoint.*

In a workshop organized by NIH[16], the following definition was recommended.

> *A biomarker intended to substitute for a clinical endpoint. A clinical investigator uses epidemiologic, therapeutic, pathophysiologic, or other scientific evidence to select a surrogate endpoint that is expected to predict clinical benefit, harm, or lack of benefit or harm.*

These definitions require that a valid surrogate endpoint should not only correlate to the clinical endpoint, but also be able to predict the clinical endpoint. As pointed out by Fleming and DeMets[14], "a correlate does not a surrogate make".

From the point view of statistical validation, Prentice[15] provides a definition of a valid surrogate. By his definition, a valid surrogate is

> *a response variable for which a test of the null hypothesis of no relationship to the treatment groups under comparison is also a valid test of the corresponding null hypothesis based on the true endpoint.*

Based on this definition, two primary conditions have been proposed by Prentice to provide guidance for how one might approach using empirical evidence to assess validation. The first condition of ensuring the validity of a surrogate is the "correlate" requirement. That is, a valid surrogate endpoint must be correlated with the true clinical endpoint. This condition usually holds since potential surrogates are often selected by searching for measures that are strongly correlated with clinical efficacy endpoints. The second condition, which is very restrictive, requires the surrogate to fully capture the treatment effect on the true clinical endpoint. This means that if one knows the value of a surrogate for a patient, one would be able to determine the clinical outcome with great certainty; i.e., knowing the treatment allocation would provide no additional information on the clinical outcome.

Statistical validation. Validation of a surrogate endpoint is a complex issue. It requires not only the empirical evidence from clinical trials documenting treatment effects on both surrogate and clinical endpoints but also a thorough biologic understanding about the mechanisms of treatment effect[18]. Statistical analyses on the existing data provide valuable insight into the relationship between surrogate and clin-

ical endpoints. While various statistical approaches have been proposed, it appears that all validation methods focus on the following three requirements.

(1) A valid surrogate must be correlated with the clinical endpoint.

(2) A valid surrogate should capture a reliable and sufficiently large portion of the treatment effect on the clinical endpoint.

(3) A valid surrogate should be able to predict the treatment effect on the clinical endpoint.

Validation of these three requirements requires distinct statistical approaches.

To be a valid surrogate, a measure has to first be correlated with the clinical endpoint. In practice, it is hard to imagine a surrogate as valid if not highly correlated with the clinical endpoint. A strong correlation with the clinical endpoint, however, does not automatically validate the endpoint as a surrogate, since even a strong correlation does not necessarily indicate a cause-effect relationship[14]. The statistical validation of this requirement is straightforward. One just conducts a regression or correlation analysis using appropriate statistical methods (e.g., parametric or non-parametric regression or correlation). Importantly, one must have the individual patient data to be able to explore the relationship between the surrogate and clinical endpoints. It is critical to recognize that the group-level summary statistics from published literature provide no information about the underlying association between the surrogate and clinical endpoints for patients[20].

The second condition requires the surrogate endpoint to reliably and sufficiently explain a large portion of the treatment effect on the clinical endpoint. The relevant statistical approach was first proposed by Freedman et al.[21]. While the definition and criteria of Prentice provide valuable guidance for validating a surrogate, it has been recognized by researchers that the criterion requiring a surrogate to fully capture the treatment effect on the clinical endpoint is too stringent and not straightforward to verify[21,22]. To overcome this difficulty, Freedman et al. proposed to calculate the proportion of treatment effect explained by a surrogate as the ratio of regression coefficients for the treatment indicator from two separate models with or without adjusting for the surrogate. In practice, a surrogate would be deemed acceptable if the lower limit of the confidence interval for the proportion was sufficiently large. While quantifying the proportion of the treatment effect explained by a surrogate is intuitively appealing, there are some limitations associated with this concept. First, this quantity is typically subject to large variability unless large sample sizes are available or a very strong effect of treatment on the clinical endpoint is observed. For this reason meta-analytical approaches have been recommended[17]. Second, the two models used to calculate the proportion cannot hold simultaneously. Third, the proportion of treatment effect could take values outside the range of [0, 1]. To help surmount these difficulties, Li et al.[10] proposed an alternative measure that is calculated within the same model and is interpretable even when the measure exceeds unity. In quantifying the proportion, the approach using individual patient data has been shown to be the preferred approach compared to meta-regression based on summary statistics since it takes into account the variability of individual patients and is a necessary approach for valid inference for the underlying relationship for patients[20].

The third requirement focuses on the ability of the surrogate endpoint to predict the treatment effect on the clinical endpoint. Since the focus of this condition is on treatment effects and requires between-group comparisons, the group-level summary statistics are necessary for the validation of this requirement. Typically one would conduct a so called meta-regression in which the observed treatment differences on clinical outcomes from an array of clinical studies are used as response and the treatment effect on surrogates are treated as a covariate. One must recognize that this type of analysis cannot capture the underlying association between surrogates and clinical outcomes for patients[20]. The primary use of this analysis should be on the trial-level treatment effect prediction rather than the causal association between surrogates and clinical outcomes. Molenberghs et al.[22-24] proposed to evaluate the prediction by the ratio between the effect of treatment on the clinical and surrogate endpoints.

Is BMD change a valid surrogate for fracture?

There are several reasons that BMD change is considered as a potential surrogate for fracture. Bone mass is an important determinant of bone strength and has been shown to be strongly correlated with elastic modulus and ultimate strength. Small changes in BMD could dramatically influence bone material properties since bone strength increases in proportion to the square of BMD[25]. Similarly, increases in BMD observed with bisphosphonate treatment also significantly contribute to bone strength. These studies were conducted using laboratory animals as well as bones obtained from cadavers[26]. The World Health Organization's (WHO) definition of osteoporosis is based on the relationship between low BMD and the consequent increase in bone fragility and susceptibility to fracture. Several randomized trials have demonstrated that anti-resorptive drugs improve BMD and reduce the risk of fractures. Because of these reasons, there has been great interest in knowing whether BMD change can be used as a surrogate for fracture. Herein, we provide some perspective using the validation criteria outlined in "What is a surrogate?"

Are BMD and fractures correlated? The correlation between BMD and fracture is the first validation requirement that one needs to assess. However, as phrased by Guyatt et al.[27], "the surrogate must be linked causally to the outcome". A strong correlation implies that a high fracture risk is strongly associated with a low BMD value and changes in BMD affect the fracture risk substantially assuming other risk factors such as age, gender, and treatment are the same. Likewise, a weak correlation suggests that the fracture risk is little changed by the associated change in BMD. To understand the correlation, one has to keep other risk factors the same. For example, it would be difficult to understand the correlation if one uses two patients from different treatment groups since one will

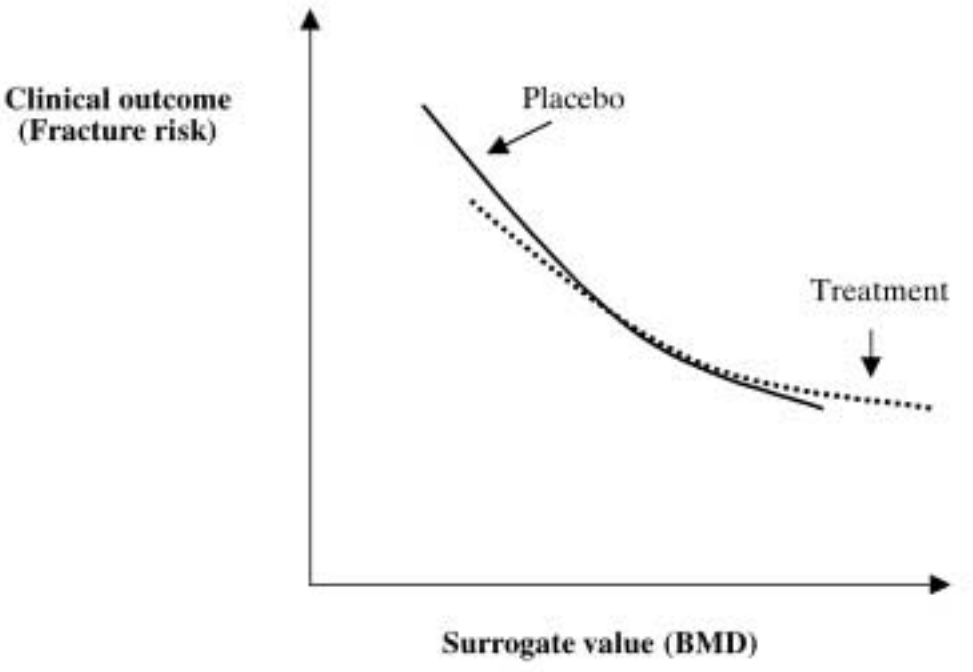

Figure 1. Expected relationship when the surrogate explains all or most of the treatment effect on the clinical outcome.

not know whether the change in risk is due to the difference in BMD or difference in treatment. Similarly, it would be difficult to interpret the correlation between fracture risk and BMD if patients have different baseline BMD values. To study the correlation, individual patient data for both BMD and fractures provide the most comprehensive information and should be the basis for statistical analyses.

The correlation between fracture and BMD has been studied using different data resources. Observational studies suggest a two-fold increase in fracture risk per SD reduction in BMD[28]. The relationship was also examined for men and women separately based on the data from a prospective study and the results suggested a positive association[29]. Based on placebo data from the MORE trial for raloxifene, the analysis suggested that 1 SD decrease in baseline femoral neck BMD and baseline lumbar spine BMD significantly increased the risk of new vertebral fracture 1.5-fold and 2-fold, respectively, at 3 years[9]. All these data indicate that the fracture risk associated with BMD decreases and fracture risk associated with BMD increases under treatment are unlikely to be the same magnitude. To investigate how the BMD increases affect the fracture risk, Hochberg et al. conducted an analysis based on alendronate treated patients in the FIT study[30]. In their analysis, they presented 3-dimensional graphics for the incidence of vertebral fractures for three subgroups of patients defined by post-baseline BMD increases (BMD percent change <=0%; >0 but <3%; >= 3%) stratified by BMD tertiles at baseline. Based on this analysis, the authors concluded that greater increases in BMD are associated with lower risk of new vertebral fractures. Their conclusions and the interpretation of this analyses clearly deserve further clarification since their published graphs serve to punctuate not only the post-baseline BMD increases but also the importance of the baseline BMD. For example, patients in the lowest baseline BMD tertile had a relatively high fracture risk even though they had more than 3% BMD increase. The fracture risk for these patients was even higher than those in the highest baseline BMD tertile that had no post baseline BMD gain. Therefore, this analysis actually suggests that fracture risk depends on both the baseline BMD value and the post-baseline BMD increase. It is clearly premature to draw conclusions based on the post-baseline BMD increases only. In this sense, the actual post-baseline BMD value appears to be more relevant to the fracture risk compared to the BMD increases since the actual values consist of the baseline value plus post-baseline increase. A plot based on risedronate data also suggested that there was a threshold in BMD increases above which BMD increases would no longer translate into fracture benefit[31].

In summary, current data indicate that the actual BMD value is strongly related to fracture risk. This is not surprising since BMD would not have been considered as a surrogate candidate if its correlation with fracture were not established.

Fracture risk reduction explained by BMD increases. As pointed out in the earlier discussion, correlation is a necessary but not sufficient condition for a valid surrogate endpoint. If BMD change is a valid surrogate, one would expect that two patients having the same baseline risk factors have the same risk of fractures as long as they reach the same post-baseline BMD value. That is, the same BMD increase should have the same effect on the fracture risk for the two patients, regardless of whether the BMD increase is achieved through active treatment or placebo with standard calcium and vitamin D supplementation. In a clinical trial setting comparing an active treatment vs. placebo, the baseline risk factors are approximately

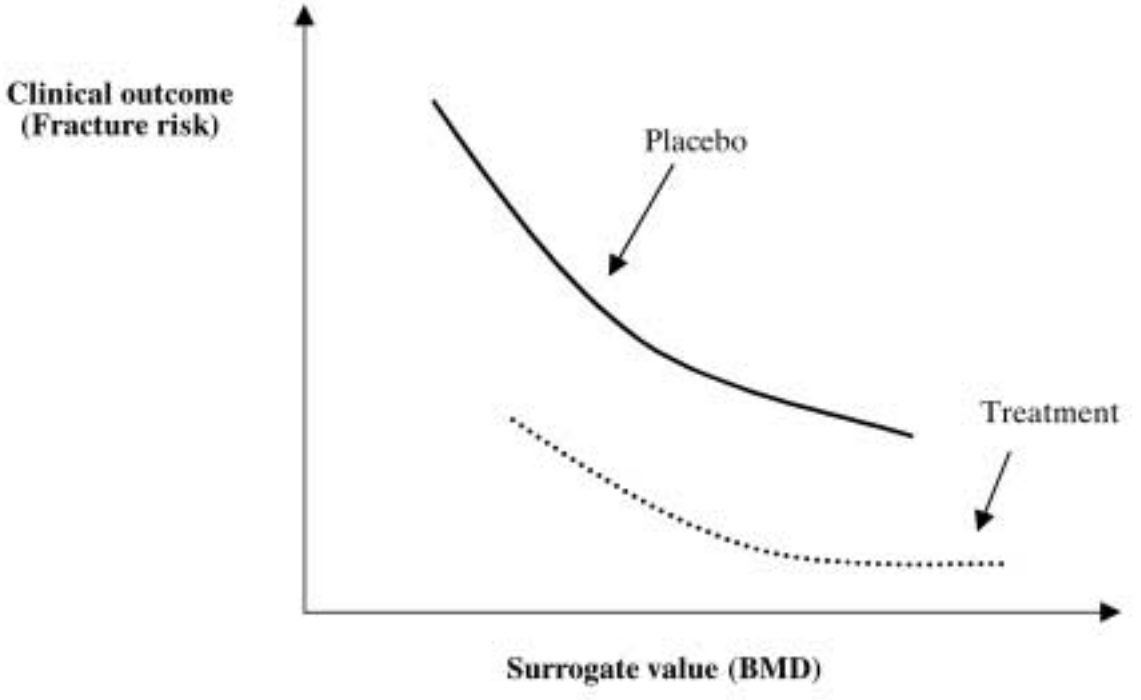

Figure 2. Expected relationship when the surrogate explains only a small portion of the treatment effect on the clinical outcome.

balanced via randomization. Therefore the only difference between the two treatment groups is the treatment assignment. This is the fundamental reason that one can assess the treatment efficacy using randomized clinical study. Graphically, one can plot the fracture risk vs. BMD value (post-baseline) for the two treatment groups separately using individual patient data. If BMD is a valid surrogate endpoint, one would expect the two curves overlay closely for the same BMD value (Figure 1). A large separation between the two curves indicates a substantial difference in fracture risk unexplained by BMD change (Figure 2); the separation reflects the impact of factors other than BMD due to the treatment. To assess this, one has to know the underlying relationship between BMD and fractures (the curve). Through this relationship, one can quantify the effect of BMD change on fracture risk (the slope of the curve). By coupling the BMD difference between the two treatment groups with the slopes of curves, one can estimate the fracture difference due to the BMD increase over placebo. If the change in fracture risk associated with the change in BMD accounts for most of the fracture risk difference in the study, then this would be a good indication that BMD change is a valid surrogate endpoint for fractures.

Statistical methods for estimating the effect of a surrogate on fracture endpoint for both binary outcomes and time-to-first event have been developed[10,21]. In these analyses, one has to check whether BMD affects the fracture risk in the same way. Namely, BMD should be related to the fracture risk in the same way between the treatment and placebo groups. Any violation of this may imply that BMD is not the only causal pathway of the disease and cast some doubts on the validity of the surrogate. To check this, one can test the interaction effect between BMD and treatment. In statistical modeling, one critical issue that people have largely ignored is whether the actual value of BMD or increase should be used as a covariate for fracture risk. To answer this question, one has to answer the question of which variable, actual value or increase, has a direct effect on the risk of fractures. If one believes the quality of bone is measured by the actual BMD value, linking the actual BMD values to fracture risk would provide a direct measurement of the causal relationship. Through this relationship, the effect of BMD increases on fracture risk can be assessed via the slope of the curve. If one just uses percent changes and ignores the baseline value, one may draw some incomplete and biased conclusions. For example, a patient may have a very high BMD percent change from baseline and a very low baseline value. The actual value for this patient could be very low because of the low baseline value. In this case, the fracture risk for the patient could be very high. If one just simply models the association between fractures and BMD percent changes, one may draw the conclusion that a high fracture risk is associated with a large increase of BMD. The 3-dimensional bar-charts based on alendronate data strongly support that the fracture risk is affected by the combination of baseline BMD and post-baseline BMD increases[30].

Statistical analyses based on individual patient data from different therapeutic agents have been performed to explore the anti-fracture efficacy explained by BMD increases. Based on the individual patient data from the FIT study for alendronate, Cummings et al.[8] reported that 16% of the vertebral fracture risk reduction that resulted from treatment with alendronate was explained by increases in BMD. For

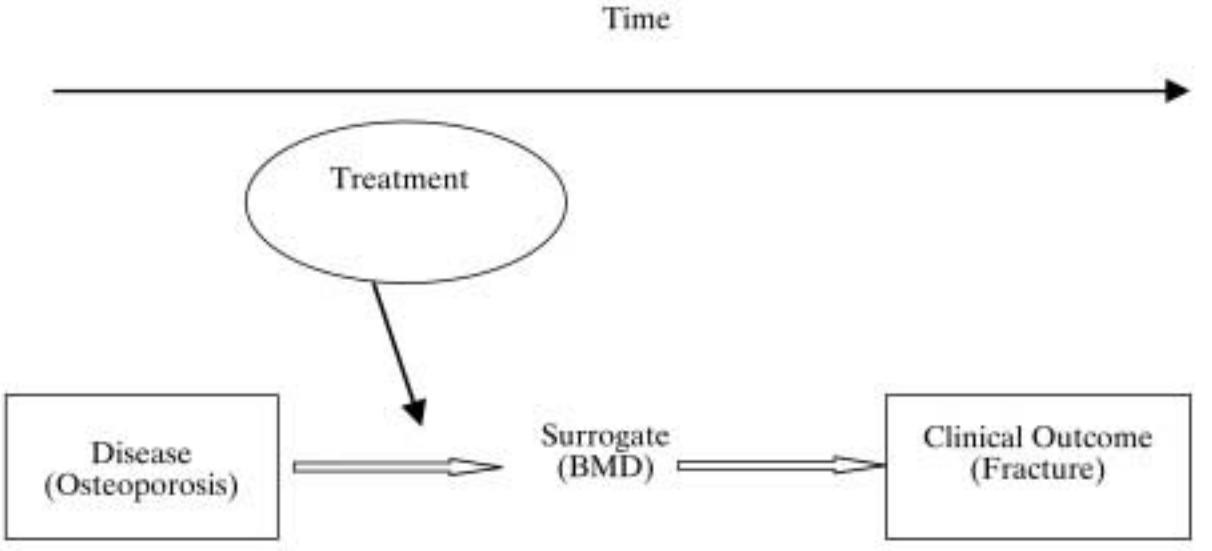

Figure 3. The setting in which the surrogate is the only causal pathway of the disease process and the entire treatment effect is mediated through its effect on the surrogate.

risedronate, Li et al.[10] conducted an analysis using individual patient data from the VERT/NA[4] and VERT/MN[5] studies and reported that 28% of the vertebral fracture risk reductions was explained by BMD increases. Sakar et al.[9] reported only 4% of the fracture risk reduction with raloxifene was explained by BMD increases based on the individual patient data from the MORE study. It should be pointed out that there are some differences in terms of statistical modeling among these three analyses. For example, the analysis by Cummings et al.[8] and Sakar et al.[9] used logistical regression models since they only used the binary outcomes for fractures while the analysis by Li et al.[10] used time-to-event methodology. Cummings et al.[8] used the approach by Freedman et al.[20] in calculating the fracture risk reduction explained by BMD while the analyses by Sakar et al.[9] and Li et al.[10] used modifications of the Freedman's approach. Nevertheless, results from the three analyses using individual patient data support the same conclusion. That is, BMD increases explain only a limited proportion of the anti-fracture efficacy observed with anti-resorptive therapies.

Can BMD increases predict fracture risk reductions over placebo? While it is not appropriate to use summary statistics to quantify the underlying relationship between BMD and fracture for patients[20], the group-level summary statistics are useful in evaluating whether a trial-level BMD increase over placebo is predictive of the fracture risk reduction for a study[32]. Specifically, one can develop a regression model by treating the relative risk of treatment vs. placebo as the response variable and the BMD improvement over placebo as a covariate. Each trial can also be weighted appropriately by using a weighting factor such as the inverse of variance for the relative risk. The intercept of this regression model represents the treatment effect associated with no improvement in BMD over placebo. To assess the predictability, one can build a prediction model for a study by using the data from all other studies. Since the prediction model is independent of the data from this study, one can assess the predictability by comparing the predicted result versus the observed result for the study. A formal statistical technique has also been developed[24].

Various statistical analyses have been conducted to evaluate the predictability of vertebral fracture risk reduction using lumbar spine BMD increases based on the group-level summary statistics[8,11,25]. The conclusions from these analyses are disparate. The analysis by Wasnich and Miller[11] concluded "treatments that increase spine BMD by 8% would reduce the risk by 54%; most of the total effect of treatment was explained by the 8% increase in BMD." They also concluded that "the small but significant reductions in risk that were not explained by measurable changes in BMD might be related to publication bias, measurement error, or limitations of current BMD technology". On the other hand, a similar analysis conducted by Guyatt et al.[27] concluded that "the fact that the model predicts a substantial relative risk reduction in vertebral fractures even with no change in bone density may be problematic". The analysis by Cummings et al.[8] also concluded that "improvement in spine bone mineral density during the treatment with anti-resorptive drugs accounts for a predictable but small part of the observed reduction in the risk of vertebral fractures". Although the conclusions were different, a careful review of these three analyses indicated some similarity of their results. The fracture risk reductions associated with no BMD improvement over placebo from the three analyses were estimated to be 22-25%, very consistent across all analyses. This implies that for a trial with 45% vertebral fracture risk reduction, at least half of the observed treatment effect cannot be predicted by BMD increases.

Regarding the accuracy of prediction using meta-regression analyses, Guyatt et al.[27] compared predicted results vs.

Study	Sample Size (Placebo, 2.5 mg, 5.0 mg)	Treatment Group	RR (95% CI)	BMD Difference Over Placebo
VERT-NA	(660, 618, 669)	2.5 mg 5.0 mg	0.54 (0.32, 0.91) 0.35 (0.19, 0.62)	2.05% 2.98%
VERT-MN	(334, 329,333)	2.5 mg 5.0 mg	0.50 (0.30, 0.84) 0.39 (0.22, 0.68)	3.14% 3.92%
HIP (Group 1)	(1028, 1027, 1027)	2.5 mg 5.0 mg	0.58 (0.36, 0.93) 0.53 (0.32, 0.86)	2.50% 3.26%
Risedronate CIO Treatment	(61, 59, 58)	2.5 mg 5.0 mg	0.42 (0.14, 1.29) 0.31 (0.09, 1.09)	1.24% 2.68%
Risedronate CIO Prevention	(54, 35, 56)	2.5 mg 5.0 mg	0.64 (0.21, 1.92) 0.38 (0.12, 1.16)	2.98% 3.81%

Table 1. Relative risk and BMD increases over placebo at 1 year.

the observed estimates for calcium, vitamin D, risedronate, etidronate, calcitonin, raloxifene and HRT. In their analyses, the authors predicted the treatment effect of one therapy based on the efficacy data on both BMD and fractures from all other therapies using a meta-regression model. However, the analysis for predicting the treatment efficacy of alendonate based on the data from all other therapies was not provided. As stated by the authors, the amount of data contributed by alendronate was so large that the remaining data based on all other drugs did not allow a robust prediction for alendronate. This implies that the prediction model was dominated by data from alendronate. Consequently, the robustness of the prediction model is compromised. While the authors indicated that the regression model predicted the fracture risk reductions very well, one should notice that the predicted vertebral fracture risk reductions for vitamin D and calcitonin were off the target by 30% and 39%, respectively. Cummings et al.[8] built a prediction model using a different approach. They predicted the fracture risk reduction based on the improvement in bone mass using data from the placebo group of the FIT study. Based on this analysis, it was estimated that each 0.10 g/cm^2 decrease in baseline spine BMD was associated with a 1.5-fold increase in the risk of vertebral fracture. They also concluded that the fracture risk reduction resulting from the anti-resorptive treatment was greater than what was predicted from the improvement in spine BMD. For example, for a trial with 45% fracture risk reduction, the predicted risk reduction by BMD improvement is only 20%.

The analyses reported so far included data from different agents. Further, the studies included had different study durations. Since BMD increases typically are not distributed linearly over the course of study, one may want to know whether using 1-year BMD data would increase the predictability. We performed an analysis based on 1-year data from risedronate studies since the anti-fracture efficacy over the first year has been demonstrated in multiple studies[4-6,33-34]. The relative risks and BMD increases over placebo at 1 year are summarized in Table 1 for both the risedronate 2.5 mg and 5.0 mg groups. We conducted a regression analysis using relative risk as the response variable and BMD increase over placebo as the covariate. Each treatment group was weighted by the inverse of variance for the relative risk. We obtained the following regression equation:

Relative Risk = 0.587 - 0.047*ΔBMD.

In this equation, the BMD increase did not show a statistically significant effect (p-value=0.346). In contrast, the intercept was statistically significant (p-value=0.003). That means, for a trial with 60% risk reduction during the first year, approximately 41% risk reduction is not predicted by BMD increases.

We also evaluated whether one could predict the fracture risk reduction based on BMD increases over placebo using this model. In this analysis, we omitted one treatment group at a time and used the remaining data to build a regression model. Using this regression model we predicted the fracture risk reduction of the treatment group that was left out using the observed BMD increase. A comparison of the observed fracture risk reductions and the predicted risk reductions is summarized in Table 2. On average, the predicted value was off the observed value by 11%, which was about 20 percentage points of the observed fracture risk reduction. Specifically, the predicted fracture risk reductions in the VERT-NA and VERT-MN were greater for the 2.5 mg group than for the 5 mg group while the 5 mg group actually demonstrated a higher fracture risk reduction in all studies. All these suggest BMD increases over placebo are inadequate for predicting fracture risk reductions.

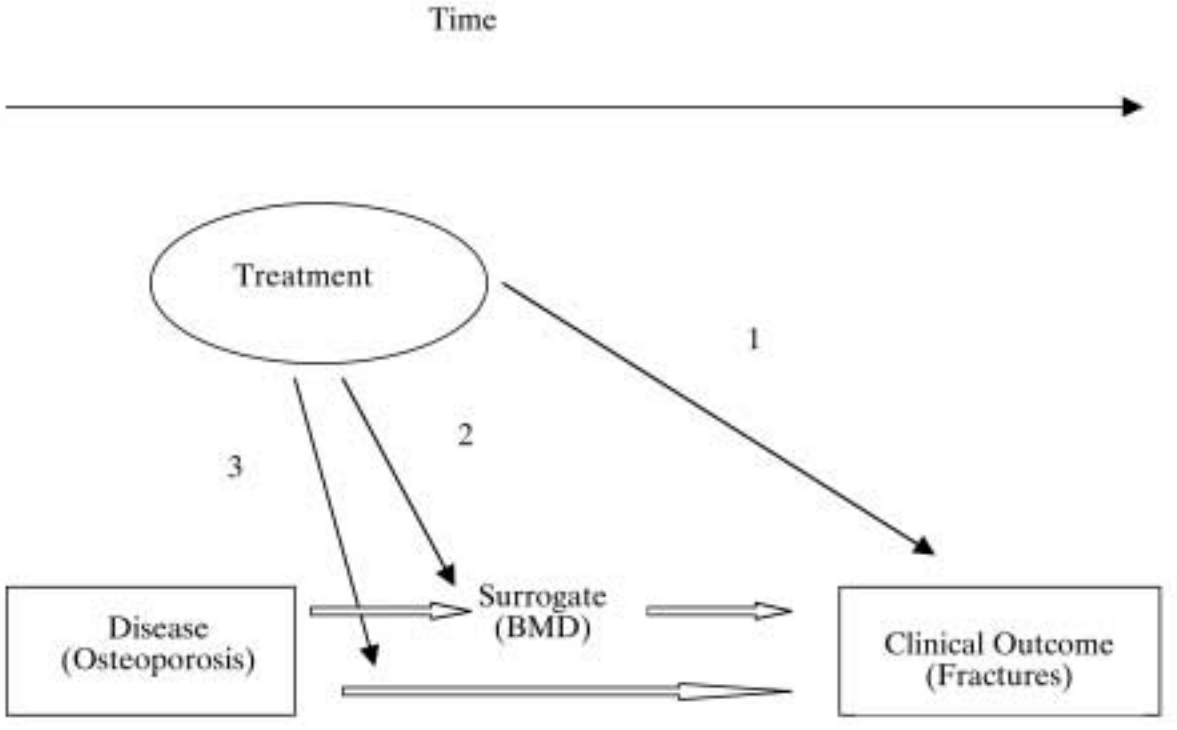

Figure 4. Multiple pathways that the treatment may impact the clinical outcome. Pathway 1: A pathway that is independent of the disease process. Pathway 2: A pathway where the treatment impacts the clinical outcome through its effect on the surrogate endpoint. Pathway 3: A pathway where the treatment impacts the clinical outcome through the disease process but independently of the surrogate endpoint.

What may affect the validity of BMD as a surrogate?

While the actual BMD value is correlated with fracture, there is no sufficient evidence to support BMD increase as a valid surrogate because of the small proportion of the treatment effect explained by BMD and because of the inadequacy of BMD increases in predicting fracture risk reductions. Figure 3 illustrates the setting that provides the greatest potential for BMD to be a valid surrogate. In this setting, BMD is the only causal pathway of the disease process and the entire treatment effect on fracture is mediated through its effect on BMD. The fact that a substantial amount of treatment effect is associated with no BMD increase casts doubt on this hypothesis.

The mechanism of action of a treatment on a surrogate endpoint and its clinical outcome is complex. The treatment may affect the disease process through multiple pathways. For example, the treatment may affect the clinical outcome by unintended mechanisms of action that are independent of the disease process. The effects of the treatment mediated through intended mechanism could be substantially offset by unintended, unanticipated or unrecognized mechanisms. Unfortunately it is impossible to separate the treatment effect mediated through intended mechanisms from those through unintended mechanisms. Figure 4 illustrates a hypothetical setting in which a treatment affects the diseases of postmenopausal osteoporosis in multiple pathways. In this setting, the treatment affects the clinical outcome through 3 different pathways, a pathway mediated through BMD, a pathway that is independent of the disease process and a pathway that is mediated through the disease process but independent of BMD. The clinical benefit of a drug is a net outcome of the three pathways. For example, a treatment may increase the bone quality by changing the microstructure of bone in the absence of a change in BMD. This mechanism of action will reduce the risk of fracture without increasing the bone density. Recent work by Eastell et al. also suggested that part of the anti-fracture efficacy might be mediated through changes in bone turnover[35-38]. Reduction of bone turnover and the associated reduction in osteoclastic perforative resorption may improve trabecular microarchitecture by preventing trabecular perforation and thus preserving trabecular connectivity without significant or just moderate increases in BMD[36].

A question that one may also ask is whether treatment effects that are mediated through BMD or other factors such as bone microstructure are the same across all anti-resorptive agents. The empirical evidence appears to support the hypothesis that different agents act differently. For example, three large clinical trials have shown that alendronate, risedronate and raloxifene reduced the risk of vertebral fractures over 3-years by 47%, 49% and 46%[2,5,7], respectively, very similar across the three agents. However, the corresponding BMD increases over placebo were 6.2%, 5.8% and

Study	Treatment Group	Observed Risk Reduction	Model Predicted Risk Reduction
VERT-NA	2.5 mg 5.0 mg	46% 65%	53% 52%
VERT-MN	2.5 mg 5.0 mg	50% 61%	57% 42%
HIP (Group 1)	2.5 mg 5.0 mg	42% 47%	55% 58%
Risedronate CIO Treatment	2.5 mg 5.0 mg	58% 69%	42% 53%
Risedronate CIO prevention	2.5 mg 5.0 mg	36% 62%	55% 59%

Table 2. Comparison of model predicted fracture risk reductions vs. observed fracture risk reductions.

2.7%, reflecting a wide range of BMD responses from agent to agent. In addition, significant increases in BMD could be accompanied by formation of bone of poor quality that in turn would decrease bone strength and negate the positive effect on BMD increases. Treatment with fluoride has been shown to increase spinal bone density by 35% while vertebral fracture risk remained unchanged[37]. Doses of fluoride in the latter study were relatively high leading to the development of osteomalacia[38]. Thus it is important to discriminate between increases in BMD with formation of bone of normal quality versus increases in BMD accompanied by formation of pathological bone when using BMD as a surrogate marker for fracture risk prediction. It should be noted that the fracture risk could also be affected by factors that are independent of treatment such as falls. Quantifying the effects of these factors is even more challenging because of the difficulty of separating these factors from the treatment-responsive factors in clinical studies.

Predicting the fracture risk reduction based on the observed BMD increases over placebo for a clinical trial may also be complicated by the possible non-linear relationship between fracture risk reduction and BMD increases. In the meta-regression analysis, studies with different populations and different baseline fracture risks are included. It is likely that the same BMD increases over placebo may translate into different fracture benefit for patients with different baseline BMD value and fracture risk. For example, the same BMD increase over placebo may not mean the same fracture benefit for a healthier population compared to a more osteoporotic population. This would also increase the uncertainty in predicting the fracture risk reduction using BMD increases. Therefore, one has to be cautious when using BMD increases to measure or compare the clinical efficacy of therapeutic agents. To increase the predictability and usefulness of surrogates, one potential research area that deserves attention is to identify surrogates for different pathways and study the joint effect of several intermediate endpoints on fracture efficacy.

Conclusion and discussion

In this paper, we have reviewed the concept of surrogate endpoints and relevant statistical validation requirements and methods. Specifically, we focused on whether BMD change from baseline can be validated as a surrogate endpoint for fracture in the treatment of postmenopausal osteoporosis. Our review indicates that the actual BMD value is correlated with fractures. The combined existing evidence, however, appears to not support BMD increase from baseline as a valid surrogate for fractures. The high proportion of the treatment effect unexplained by BMD and the lack of accuracy in predicting the fracture risk reduction for a clinical trial using BMD provide strong support for the hypothesis that BMD is not the only causal pathway through which a treatment affects the fracture outcome. This has important clinical implications, especially when treating patients with anti-resorptive agents: estimation of fracture risk reduction in individual patients based solely on BMD changes is not supported by the current body of data.

One may ask how you should use BMD in clinical and scientific research. Our review has indicated that the actual BMD value is correlated with fracture risk. Regulatory guidelines also indicate that if a drug has been approved for the treatment of osteoporosis by demonstrating anti-fracture efficacy, BMD may serve as an appropriate efficacy endpoint in trials for prevention of osteoporosis[1]. Research based on risedronate data has indicated a non-linear relationship between BMD changes and fracture risk; there appears to be

a threshold above which BMD increases no longer translate into fracture benefits[31]. Recent analyses with ibandronate data support the same conclusion[39]. Therefore, although BMD remains an important clinical measure, one needs to understand its limitations. It is also important to keep in mind that the treatment effects on BMD and fracture risk reduction, as well as the treatment effects explained by BMD increases for different agents, were not derived from head-to-head comparison trials. Consequently, one should be cautious when comparing the efficacy of different agents using their BMD increases, fracture risk reductions, and the proportions of treatment effect explained by BMD increases.

Surrogate validation is a complex task. While statistical analyses provide useful insight into the impact of BMD on the fracture endpoint, one has to consider the biologic mechanisms through which a treatment may affect the fracture risk. Research in identifying different pathways through which a treatment acts on the disease is very limited. The fact that the three anti-resorptive agents, alendronate, risedronate and raloxifene, demonstrated similar vertebral fracture risk reduction over 3 years but markedly different BMD increases also warrants research on whether the mechanisms of action are the same across all anti-resorptive agents. In addition to the differences that may exist among different agents, factors such as variability in clinical measurements and data collection, heterogeneity in study populations and study designs, and differences in statistical methods confound the interpretation of the results. The relationship between BMD and fractures and the ability of BMD to predict fracture risk is also complicated by a variety of non-skeletal factors for fractures. Among them are propensity to falls, neuromuscular responses to falls, severity and direction of falls, and the amount of fat padding around the bone[25].

We have paid special attention to the two statistical approaches used in surrogate validation: analyses based on individual patient data and meta-regression using summary statistics. All information contained in summary statistics is also contained in the relevant individual patient data. Thus, whenever possible, one should conduct analyses based on individual patient data. An analysis using individual patient data from one agent, however, is typically limited by its sample size, especially when quantifying the treatment effect explained by BMD. To overcome this difficulty, a collaborative effort across all sponsors for the relevant agents is desirable.

References

1. Guidelines for preclinical and clinical evaluation of agents used in the prevention or treatment of postmenopausal osteoporosis. Division of Metabolic and Endocrine Drug Products, Food and Drug Administration, 1994.
2. Black DM, Cummings SR, Karpf DB, Cauley JA, Thompson DE, Nevitt MC, Bauer DC, Genant HK, Haskell WL, Marcus R, Ott SM, Torner JC, Quandt SA, Reiss TF, Ensrud KE. Randomized trial of effect of alendronate on risk of fracture in women with existing vertebral fractures. Lancet 1996; 348:1535-1541.
3. Cummings SR, Black DM, Thompson DE, Applegate WB, Barret-Connor E, Musliner TA, Palermo L, Prineas r, Rubin SM, Scott JC, Bogt T, Wallace R, Yates AJ, LaCroix AZ. Effect of alendronate on risk of fracture in women with low bone density but without vertebral fractures: results from Fracture Intervention Trial. JAMA 1998; 280:2077-2782.
4. Harris ST, Watts NB, Genant HK, McKeever CD, Hangartner T, Keller M, Chesnut C, Brown J, Eriksen EF, Hoseyni MS, Axelrod DW, Miller PD. Effects of risedronate treatment on vertebral and nonvertebral fractures in women with postmenopausal osteoporosis. JAMA 1999; 282:1344-1352.
5. Reginster JY, Minne HW, Sorensen OH, Hooper M, Roux C, Brandi ML, Lund B, Ethgen D, Pack S, Roumagnac I, Eastell R. Randomized trial of the effects of risedronate on vertebral fractures in women with established postmenopausal osteoporosis. Osteoporos Int 2000; 11:83-91.
6. McClung MR, Geusens P, Miller PD, Zippel H, Bensen WG, Roux C, Adami S, Fogelman I, Diamong T, Eastell R, Meunier PJ, Reginster JY; Hip Intervention Program Study Group. Effect of risedronate on the risk of hip fracture in elderly women. N Engl J Med 2001; 344:333-340.
7. Ettinger B, Black DM, Mitlak BE, Knickerbocker Rk, Nickelsen T, Genant HK, Christiansen C, Delmas PD, Zanchetta JR, Stakkestad J, Gluer CC, Krueger K, Cohen FJ, Eckert S, Ensrud KE, Avioli LV, Lips P, Cummings SR. Reduction of vertebral fracture risk in postmenopausal women with osteoporosis treated with raloxifene: results from a 3-year randomized clinical trial. JAMA 1999, 282:637-645.
8. Cummings SR, Karpf DB, Harris F, Genant HK, Ensrud K, LaCroix AZ, Black DM. Improvement in spine bone density and reduction in risk of vertebral fractures during treatment with anti-resorptive drugs. Am J Med 2002; 112:281-289.
9. Sarkar S, Mitlak BH, Wong M, Stock JL, Black DM, Harper KD. Relationship between bone mineral density and incident vertebral fracture risk with raloxifene therapy. J Bone Miner Res 2002; 17:1-10.
10. Li Z, Meredith MP, Hoseyni MS. A method to assess the proportion of treatment effect explained by a surrogate endpoint. Stat Med 2001; 20:3175-3188.
11. Wasnich RD, Miller PD. Anti-fracture efficacy of anti-resorptive agents are related to changes in bone density. J Clin Endocrinol Metabol 2000; 85:231-236.
12. Hochberg MC, Greenspan S, Wasnich RD, Miller P, Thompson DE, Ross PD. Changes in bone density and turnover explain the reductions in incidence of nonvertebral fractures that occur during treatment with anti-resorptive agents. J Clin Endocrinol Metabol 2002; 87:1586-1592.
13. Lin DY, Fischl MA, Schoenfeld DA. Evaluating the

role of CD4-lymphocyte counts as surrogate endpoints in human immunodeficiency virus clinical trials. Stat Med 1993; 12:835-842.
14. Fleming TR, DeMets DL. Surrogate endpoints in clinical trials: are we being misled? Ann Intern Med 1996; 125:605-613.
15. Prentice RL. Surrogate endpoints in clinical trials: Definitions and operational criteria. Stat Med 1989; 8:431-440.
16. DeGruttola VG, Clax P, DeMets DL, Downing GJ, Ellenberg SS, Friedman L, Gail MH, Prentice R, Wittes J, Zeger SL. Considerations in the evaluation of surrogate endpoints in clinical trials: summary of a National Institute of Health Workshop. Control Clin Trials 2001; 22:485-502.
17. Daniel MJ, Hughes MD. Meta-analysis for the evaluation of potential surrogate markers. Stat Med 1997; 16:1965-1982.
18. De Gruttola V, Fleming TR, Lin DY, Coombs R. Validating surrogate markers – Are we being naive? J Infect Dis 1997; 175:237-246.
19. Temple RJ. A regulatory authority's opinion about surrogate endpoints. In: Nimmo WS, Tucker GT (eds) Clinical Measurement in Drug Evaluation. J Wiley, New York; 1995.
20. Li Z, Meredith MP. Exploring the relationship between surrogates and clinical outcomes: analysis based on individual patient data versus meta-regression on group-level summary statistics. J Biopharm Stat 2003; 13:777-792.
21. Freedman LS, Graubard BI, Schatzkin A. Statistical validation of intermediate endpoints for chronic diseases. Stat Med 1992; 11:167-178.
22. Buyse M, Molenberghs G. Criteria for the validation of surrogate endpoints in randomized experiments. Biometrics 1998; 54:1014-1029.
23. Molenberghs G, Geys H, Buyse M. Evaluation of surrogate endpoints in randomized experiments with mixed discrete and continuous outcomes. Stat Med 200; 20:3023-3038.
24. Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H. The validation of surrogate endpoints in meta-analyses of randomized experiments. Biostatistics 2000; 1:49-67.
25. Bouxsein ML, Myers ER, Hayes WC. Biomechanics of age-related fractures. In: Marcus R, Feldman D, Kelsey J (eds) Osteoporosis. Academic Press 1996:373-393.
26. Balena R, Toolan BC, Shea M, Markatos A, Myers Er, Lee SC, Opas EE, Seedor JG, Klein H, Frankenfield D. The effect of 2-year treatment with aminobisphosphonate alendronate on bone metabolism, bone histomorphometry, and bone strength in ovariectomized nonhuman primates. J Clin Invest 1993; 92:2577-2586.
27. Guyatt GH, Cranney A, Griffith L, Walter S, Krolicki N, Favus M, Rosen C. Summary of meta-analyses of therapies for postmenopausal osteoporosis and the relationship between bone density and fractures. Endocrinol Metab Clin North Am 2002; 31:659-679.
28. Marshall D, Johnell O, Wedel H. Meta-analysis of how well measures of bone mineral density predict occurrence of osteoporotic fractures. Brit Med J 1996; 312:1254-1259.
29. The relationship between bone density and incident vertebral fractures in men and women. The European Prospective Osteoporosis Study Group (EPOS). J Bone Miner Res 2002; 17:2214-2221.
30. Hochberg MC, Ross PD, Black D, Cummings SR, Genant HK, Nevitt MC, Barrett-Connor E, Musliner T, Thompson D. Larger increases in bone mineral density during alendronate therapy are associated with a lower risk of new vertebral fractures in women with postmenopausal osteoporosis. Arthritis Rheum 1999; 42: 1246-1254.
31. Watts N, Bockman R, Smith C, Li Z, Eastell R, Pack S Lindsay R. BMD change explains only a fraction of the observed fracture risk reduction in risedronate-treated patients. Osteoporos Int 2000; 11:S203.
32. Molenberghs G, Buyse M, Geys H, Renard D, Burzykowski T, Alonso A. Statistical challenges in the evaluation of surrogate endpoints in randomized trials. Control Clin Trials 2002; 23:607-625.
33. Reid DM, Hughes RA, Laan RF, Sacco-Gibson NA, Wenderoth DH, Adami S, Eusebio RA, Devogelaer JP. Efficacy and safety of daily risedronate in the treatment of corticosteroid-induced osteoporosis in men and women: a randomized trial. J Bone Miner Res 2000; 15:1006-1013.
34. Cohen A, Levy RM, Keller M, Boling E, Emkey RD, Greenwald M, Zizic TM, Wallach S, Sewell KL, Lukert BP, Axelrod DW, Chines AA. Risedronate therapy prevents corticosteroid-induced bone loss. Arthritis Rheum 1999; 42:2309-2318.
35. Eastell R, Barton I, Hannon RA, Chines A, Garnero P, Delmas PD. Relationship of early changes in bone resorption to the reduction in fracture risk with risedronate. J Bone Miner Res 2003; 18:1051-1056.
36. Riggs BL, Melton LJ. Bone turnover matters: the raloxifene treatment paradox of dramatic decreases in vertebral fractures with commensurate increases in bone density. J Bone Miner Res 2002; 17:11-14.
37. Riggs BL, Hodgson SF, O'Fallon WM, Chao EY, Wahner HW, Muhs JM, Cedel SL, Melton LJ. Effect of fluoride treatment on the fracture rate in postmenopausal women with osteoporosis. N Engl J Med 1990; 322(12):802-809.
38. Lundy MW, Stauffer M, Wergedal JE, Baylink DJ, Featherstone JD, Hodgson SF, Riggs BL. Histomorphometric analysis of iliac crest bone biopsies in placebo-treated versus fluoride-treated subjects. Osteoporos Int 1995; 5:115-129.
39. Wasnich R, Miller PD Chesnut CH, Huss H, Wilson K, Schimmer RC. Changes in bone mineral density as a predictor for vertebral fracture efficacy with ibandronate: results from a phase III fracture study. J Bone Miner Res 2003; 18:SA353.