IBMS BoneKEy. 2009 August;6(8):279-294
http://www.bonekey-ibms.org/cgi/content/full/ibmske;6/8/279
doi: 10.1138/20090391

PERSPECTIVES

Interpretation of Randomized Controlled Trials of Fracture Prevention

Tuan V. Nguyen

Osteoporosis and Bone Biology Program, Garvan Institute of Medical Research, Sydney, Australia

Abstract

The question that a reader of a randomized controlled trial (RCT) is interested in is whether therapy is effective. However, prevailing methodology addresses the opposite question: if the therapy is not effective, what is the chance of obtaining the present (or more extreme) data? This current methodology has generated considerable confusion and misinterpretation in the literature. In this *Perspective*, an alternative interpretation of major data from RCTs of fracture prevention is offered in light of Bayesian inference, with the hope that this approach will be adopted more often in future clinical research studies of osteoporosis.
IBMS BoneKEy. 2009 August;6(8):279-294.

The randomized controlled trial (RCT) is now universally accepted as the gold standard design to learn about the efficacy and safety of an intervention. The efficacy of an intervention is summarized in terms of the effect size and a measure of sampling variability. In anti-fracture clinical trials, the effect size is commonly expressed by a relative risk (RR) accompanied by a *P*-value and 95% confidence interval (CI). A result where the *P*-value is less than 0.05 (or, equivalently, when the 95% CI excludes 1) is interpreted as "significant." On the other hand, a result where the *P*-value is higher than 0.05 (or when the 95% CI includes 1) is considered "non-significant." In RCTs, a significant result is often interpreted as evidence that the therapy is effective, and experience abounds with examples that editors and reviewers are not keen to publish papers if results are not significant ($P > 0.05$).

However, the *P*-value may be misunderstood. In a survey of 397 clinicians, only 19% correctly understood the meaning of the *P*-value (1). One of the most common misunderstandings of the *P*-value is that it is the probability that the finding is due to chance. Thus, a *P*-value of 0.05 is thought to indicate that the probability that the therapy is ineffective is 5%, or that there is a 95% chance of a real treatment effect. However, such an interpretation is wrong; the *P*-value tells us nothing about treatment effect. When the *P*-value is used in the context of multiple tests of hypothesis, it generates additional problems of interpretation (2;3). Furthermore, it has been shown that most published research findings are false (4), and that about 25% of all findings with "$P < 0.05$," if viewed in a scientifically agnostic light, can be regarded as meaningless (5) or as nothing more than chance findings (6). In an analysis of 45 original, highly-cited papers claiming treatment efficacy published in the *New England Journal of Medicine*, *Lancet*, and *JAMA,* it was found that 14 (or 32%) were subsequently shown to be either contradictory or exaggerated (7). Thus, a prominent scientist has suggested that "[t]he most important task before us…is to demolish the *P*-value culture, which has taken root to a frightening extent in many areas of both pure and applied science, and technology" (8).

The Problem of the *P*-Value and the Confidence Interval (CI)

In order to understand the real meaning of the *P*-value in the context of clinical research, it is useful to consider the RCT as

a procedure of hypothesis falsification (9). In this procedure, a null hypothesis of no treatment effect (*e.g.*, a RR equal to exactly 1) is assumed; data are collected in a clinical trial; and, using the data collected, a test statistic (*e.g.*, t-test or Chi-square test) is computed and compared to the value the statistic would take if the null hypothesis *were* true. Thus, the *P*-value is the probability of observing the current data (plus other data that are at least as extreme as the current data but not yet observed) if there was no treatment effect.

The above procedure is similar to the process of inductive inference by "disproving" a null hypothesis, with three premises: (a) if the intervention has no effect, then the data (*e.g.*, test statistic) cannot occur; (b) the data have occurred; (c) therefore, the hypothesis of no effect is unlikely. This awkward inference has been amusingly illustrated by a hypothetical example (10): "If a person is a Martian, then he is not a member of Congress. This person is a member of Congress. Therefore, he is not a Martian." This one-dimensional reasoning can be likened to making a diagnosis that a person has a disease just because one test result falls outside of a reference range!

Recognizing that this one-dimensional strategy (of significance testing) is untenable, Neyman and Pearson introduced the concept of hypothesis testing (11). In Neyman and Pearson's approach, it is necessary to define a null hypothesis and alternative hypotheses, and error rates are established for falsely deciding to reject the null hypothesis (type I error or α level) or alternative hypotheses (type II error or β level) such that "…in the long run experience, we shall not often be wrong" (11). A test statistic is then calculated from the observed data, and compared to the critical value from the expected value of the test statistic under the assumption that the null hypothesis is true. In this formulation, the α and β levels are measures of the long-term behavior of a test statistic; they are *not* thresholds for deciding whether a hypothesis is plausible.

The current model of the RCT is, conceptually, a hybrid of Fisher's test of significance and Neyman-Pearson's test of hypothesis. This synthesis reconciles two differing perspectives on how research hypotheses are defined and tested. It adopts the Neyman-Pearson convention of two competing hypotheses, but one is always labeled as the null hypothesis as in Fisher's test of significance. In this realization, Neyman-Pearson's α level is often arbitrarily set at 0.05, and a *P*-value from Fisher's test of significance is compared to the α level. This arbitrary α threshold creates confusion regarding the real meaning of the *P*-value. Indeed, many problems with the current paradigm result from the mixture of two incompatible approaches (*e.g.*, test of significance and test of hypothesis) (12).

P-value-based inference has been criticized for more than 70 years (13;14). Among the criticisms that are cited include illogical formulation, arbitrariness, and lack of clinical relevance. In terms of logic, recall that the *P*-value is the probability of getting a result as least as extreme as the one observed plus the data that are not yet observed. The problem with this formulation is that the long-run condition of such testing is a fiction (because in the real world, nobody would repeat the RCT an infinite number of times) relative to actual scientific inquiry and decision-making. Moreover, the null hypothesis is always false, because there are no truly zero effects in nature (3). For example, in a placebo-controlled trial, it is very likely that there is some difference in fracture rate between the placebo and the treated group. The chance that the two groups have the same outcome value is virtually zero. Therefore, testing a null hypothesis is illogical and impractical.

In terms of arbitrariness, the arbitrary threshold of $P = 0.05$ creates a false dichotomy between "significance" and "non-significance." Statistical significance or insignificance is often equated to clinical significance or insignificance. However, there is no real scientific basis for declaring that $P = 0.051$ is not acceptable while $P = 0.049$ is a reason to be satisfied. Ronald Fisher, who developed the test of significance and the *P*-value, selected the

cut-off value of 0.05 for convenience: "It is convenient to draw the line at about the level at which we can say: either there is something in the treatment, or a coincidence has occurred such as does not occur more than once in twenty trials" (15). However, Fisher also said that "[w]e shall not often be astray if we draw a conventional line at 0.05..." (16). However, in the real-world setting of clinical research, outcomes with statistical significance ($P < 0.05$) have higher odds of being published than outcomes where $P > 0.05$ (17;18). As a result, many researchers tend to regard data analysis as a means to obtain a statistically significant end. However, some authors suggest that reliance on the *P*-value and significance testing "retards the research enterprise by making it difficult to develop cumulative knowledge." (19).

It is not uncommon to find that when a finding does not reach a statistically significant level of 0.05, inadequate sample size or low power (*e.g.*, high β) is often blamed for the failure. However, while it is quite plausible that the smaller a sample size is, the less reliable its conclusion is, this argument overlooks the fact that small sample size studies with large effect sizes can yield the same *P*-value as large studies with very small effect sizes. Consider the following hypothetical studies (Table 1), each involving two patient groups of equal sample size, but with different magnitudes of effect as measured by an odds ratio (OR).

Table 1. The relationship between effect size, sample size, and *P*-value through 7 hypothetical studies

Study	Sample size per group	Incidence group 1	Incidence group 2	OR (95% CI)	*P*-value
1	100	2 (2%)	10 (10%)	5.44 (1.16 – 25.52)	0.03
2	500	10 (2%)	22 (4.4%)	2.26 (1.06 – 4.81)	0.03
3	1,000	20 (2%)	36 (3.6%)	1.83 (1.05 – 3.18)	0.03
4	5,000	100 (2%)	132 (2.6%)	1.33 (1.02 – 1.73)	0.03
5	10,000	200 (2%)	245 (2. 5%)	1.23 (1.02 – 1.49)	0.03
6	100,000	2000 (2%)	2138 (2.1%)	1.07 (1.00 – 1.14)	0.03
7	1,000,000	20000 (2%)	20430 (2.04%)	1.02 (1.00 – 1.04)	0.03

All studies yielded a *P*-value of approximately 0.03. Study 1 shows a substantial effect, but the sample size is rather modest to warrant a definite conclusion. Yet, as the sample size is increased and the effect size is decreased, the *P*-value remains unchanged. Indeed, study 7 seems to demonstrate conclusively no real effect (the incidence rates of the two groups are virtually equal), but the *P*-value is 0.03! This simple demonstration shows that any small difference, no matter how clinically unimportant, will be statistically significant ($P < 0.05$) when the sample size is large enough. On the other hand, any large difference, no matter how clinically important, will be statistically insignificant ($P > 0.05$) when the sample size is small. Thus, a low *P* value in a small study provides stronger evidence than the same *P* value in a larger study. In short, the *P*-value is not a rational measure of the weight of evidence for the treatment effect.

The question is not whether there is no effect, but rather how large the effect size is. Unfortunately, the *P*-value as computed from the above procedure does not provide any information on effect size. Recognizing the fallacy of the *P*-value and the limitation of hypothesis testing, confidence intervals have been introduced (20). In clinical research, the 95% CI is increasingly becoming a measure of effect size. CIs are useful because they provide the lower and upper limits of effect size consistent with a study's data. However, the meaning of CIs is also rather awkward and not easy to understand. Clinicians often understand that a 95% CI means that there is a 95% probability that the effect size lies within the interval. However, such an understanding is incorrect. The meaning of the CI is based on the following logic: if a study is repeated an infinite number of times, then 95% of all calculated "95% CIs" would be expected to contain the true, but unknown value of the parameter – assuming the null hypothesis is correct. This is often referred to as the "frequentist CI" because it is based on long-term frequency. Thus, the CI approach is based on a fictitious assumption that the

study will be repeated an infinite number of times, when in reality, the study has been done only once. The CI has often been used as a *de facto* significance test when researchers examine whether the confidence limit overlaps the null value (3). For example, a 2006 study stated that supplementation with vitamin C and E during pregnancy does not reduce the risk of death or other serious outcomes in infants (21). This conclusion appeared to be based on the RR of 0.79 and a 95% CI: 0.61-1.02. A CI thus provides no more information about the likelihood of chance as an explanation for the finding than does a *P*-value. In a clinical research setting, what we want to know is: given the data that we have from prior and current studies, what is the chance that there is a treatment effect? Only a Bayesian approach can provide answers to this question.

Bayesian Inference

In recent years, an alternative model of scientific inference has been proposed and increasingly applied in medical research (22-25). This "new" model of inference is based on the idea put forth by Thomas Bayes (1702-1761) in the 18th century (26). The Bayesian approach offers what researchers want to know: the likelihood of a hypothesis given some data that have been observed. It allows researchers to combine new data with their existing data or knowledge to arrive at more refined data and knowledge.

At the simplest level, the inference of treatment effect from an RCT can be likened to the reasoning that is used in clinical diagnosis (27). In clinical diagnosis one is interested to know whether a patient has a disease if his or her test result is positive. This probability – also referred to as the positive predictive value (PPV) – is a function of the disease's prevalence in the population, as well as the test's sensitivity and specificity. This is basically a Bayesian approach to diagnosis. In fact, all clinicians are Bayesians.

Similar to the diagnostic scenario, in an RCT one is interested to know whether there is a treatment effect given "positive" data (not the probability of getting the data given no effect as conveyed by the *P*-value). According to the Bayesian theorem (26), the probability of treatment effect given the data can be formally expressed as a function of the probability of effect *prior* to conducting the study, and the probability of obtaining the data if there is an effect. Expressed formally, the theorem states that the probability of hypothesis *H* conditioned on data *D* – denoted by P(H | D) – is proportional to the probability of H before the study – P(*H*) – and the probability of data *D* given the hypothesis *H* – P(D | H). In other words: $P(H \mid D) \propto P(H) \times P(D \mid H)$. This formulation is equivalent to saying that "what we knew beforehand + what we learn from the present data = what we know afterwards." It is analogous to the clinical setting, where prior to making a diagnosis, the clinician usually knows the background risk of the disease in the general population (such as lifetime risk) and clinical history (prior information), and when the test provides a result (actual data), the clinician can update his or her initial estimate of risk of disease for the patient (posterior information).

Thus, there are three basic elements in a Bayesian inference: prior data, likelihood of present data, and posterior distribution. Prior information describes clinical opinion, a *priori* or previously observed results from other studies. Prior information is expressed in terms of a probability distribution related to the effect size. This can be based on information available in the literature, existing knowledge, and clinical databases, and even experts' opinions. For example, the RR reduction and its 95% CI of past clinical trials can be considered "prior information" in a Bayesian inference. The *likelihood* captures how the present data modify prior knowledge of the therapeutic effect. The *posterior distribution* synthesizes both prior knowledge and likelihood function to provide the ultimate answer to the question: what is the likelihood that the treatment has an effect given the observed data? The Bayesian approach is therefore a process of cumulating research findings, which is also an essential feature of the scientific enterprise.

Vitamin D Supplementation and Fracture: a Bayesian Interpretation

As an example of the Bayesian approach, consider the association between vitamin D and fracture risk. Bischoff-Ferrari and colleagues (28) have undertaken a meta-analysis of the available data on the anti-fracture efficacy of vitamin D supplementation, and concluded that vitamin D supplementation in ambulatory individuals at doses between 700 and 800 IU/d may reduce hip fracture by 26% (pooled RR = 0.74; 95% CI: 0.61-0.88) and non-hip fracture risk by 23% (RR = 0.77; 95% CI: 0.68-0.97) (28). However, results from the RECORD study, published almost at the same time (29), suggested that routine oral supplementation with vitamin D with or without calcium did not significantly reduce fracture risk in elderly men and women either at the hip (RR = 1.14; 95% CI: 0.75-1.71) or non-vertebrae (RR = 1.07; 95% CI: 0.90-1.29). The question of interest is: given the results of meta-analysis and the latest data, what is the probability that vitamin D supplementation could reduce hip or non-vertebral fracture risk?

In the presence of conflicting findings, the Bayesian approach offers an attractive method for synthesizing various data into a coherent summary and a more reliable conclusion regarding treatment effect. The distribution of RR from the meta-analysis can be considered as prior information (Fig. 1 and Fig. 2, top panel). The current (RECORD) data are represented by the middle panel of Fig. 1 and Fig. 2. The shape of the distribution shows that the prior information has much more information than the RECORD data. The posterior distribution (bottom panel of Fig. 1 and Fig. 2) provides the expected RR given the prior information and current likelihood of the RECORD data. The posterior data indicate that the RR of hip fracture was 0.79 (95% CreI: 0.67 – 0.94), and of non-hip fracture was 0.91 (95% CreI 0.80 – 1.03).

It should be noted that an interval estimate in Bayesian methods is referred to as the "credible interval" (CreI) (30). The meaning of the CreI is much different than the meaning of the CI. A 95% Bayesian CreI means that there is 0.95 probability that the effect size value being estimated will fall between the lower and upper bounds of the set. Thus, in the above example, there is a 95% chance that vitamin D supplementation reduces hip fracture by between 6% and 33%. This interpretation is conditional on prior information and the confidence one has in that prior information.

A major advantage of Bayesian analysis is that it can estimate *any* magnitude of efficacy by computing the area under the curve between any two points on the distribution. For example, if "clinical efficacy" is defined as a RR reduction of at least 25%, then the probability of clinical efficacy can be estimated by computing the area under the curve where RR ≤ 0.75. Accordingly, the probability that 700-800 IU/d of vitamin D supplementation reduced hip fracture and non-vertebral fracture risk by at least 25% was only 0.25 and 0.002, respectively. Thus, in the ambulatory elderly, given these latest data, it seems that the evidence for the effect of vitamin D supplementation on fracture risk is still inconclusive. If there is an effect, the effect size is likely to be modest.

In Bayesian inference, there is no dichotomy of "significance" versus "non-significance." The *P*-value has no interpretational meaning in Bayesian inference; instead, the result is expressed in terms of the posterior probability of treatment effect given the observed data. For example, we can make a direct statement concerning the effect as follows: "There was a 90% probability that the intervention reduced fracture risk by at least 10%" (31). By moving away from a dichotomous decision, the Bayesian approach avoids drawing conclusions from a single study.

Bayesian Interpretation of Anti-Fracture Efficacy

During the past two decades, several RCTs of anti-fracture efficacy have been conducted (32-42). All of these RCTs were designed with some initial guesses of effect sizes; however, therapeutic efficacy has been interpreted in terms of the *P*-value and CI rather than in terms of the initial hypothetical effect size. For example, the

IBMS BoneKEy. 2009 August;6(8):279-294
http://www.bonekey-ibms.org/cgi/content/full/ibmske;6/8/279
doi: 10.1138/20090391

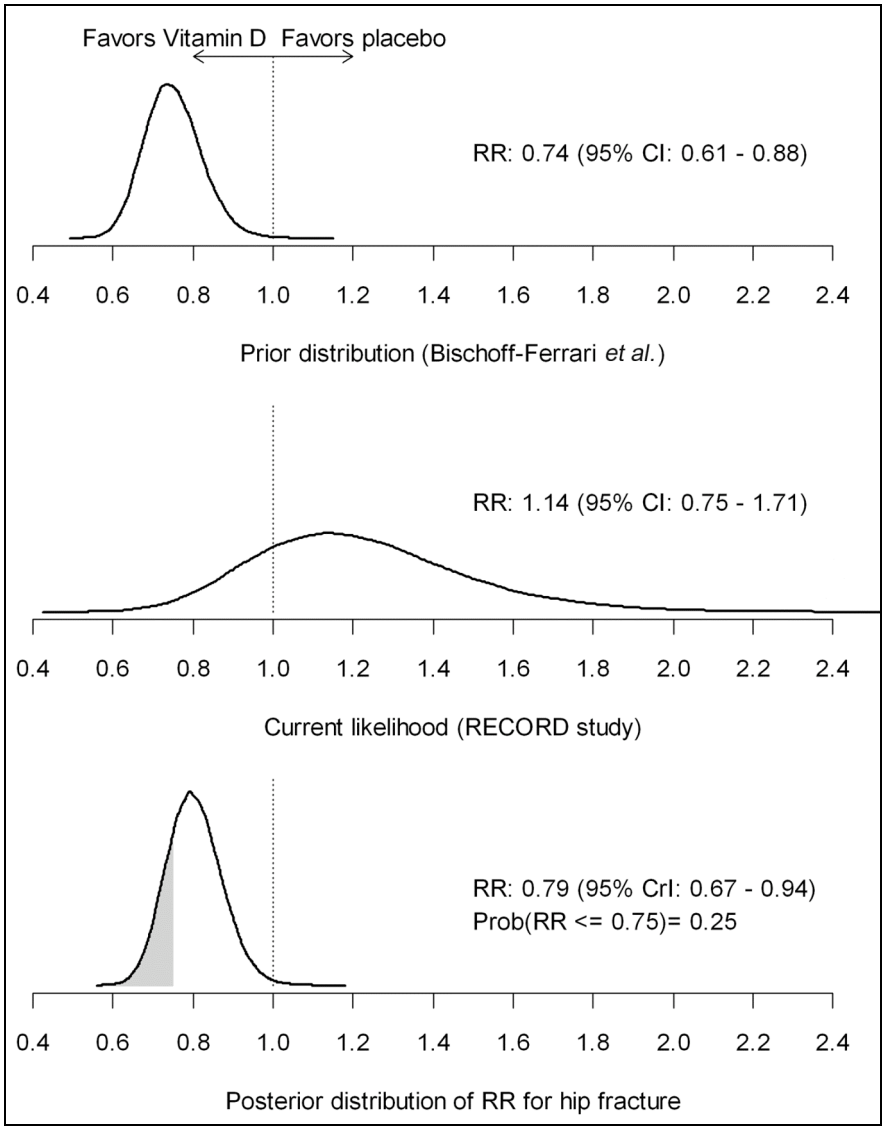

Fig. 1. Prior distribution (top panel), likelihood of RECORD data (middle), and the resulting posterior distribution (bottom) of relative risk (RR) associated with vitamin D supplementation's effect on hip fracture. The highlighted area under the posterior distribution curve represents the probability that vitamin D reduced fracture by at least 25%. This probability was estimated at 0.25 for hip fracture.

FIT-1 trial (32) was designed to detect a 40% RR reduction of vertebral fractures, and the actual RR reduction was 47% (95% CI: 32% to 59%), but the investigators do not formally comment on whether the data are consistent with the initial hypothesis. Similarly, the TROPOS study (36) was designed to detect a 25% RR reduction of non-vertebral fracture, but this anticipated effect size was not formally corroborated with the actual RR of 0.89 (95% CI: 0.67–1.19). The Bayesian approach can be helpful in assessing the consistency between observed data and the anticipated effect size.

As mentioned above, in order to estimate the probability of effect given the observed data, it is necessary to specify the prior information of the effect size. In the context of anti-fracture clinical trials, the primary measure of effect size is the RR. It is well-known that logarithmic RR follows a normal distribution characterized by its mean and variance. Previous RCTs (32-42) suggest that there is substantial uncertainty about

IBMS BoneKEy. 2009 August;6(8):279-294
http://www.bonekey-ibms.org/cgi/content/full/ibmske;6/8/279
doi: 10.1138/20090391

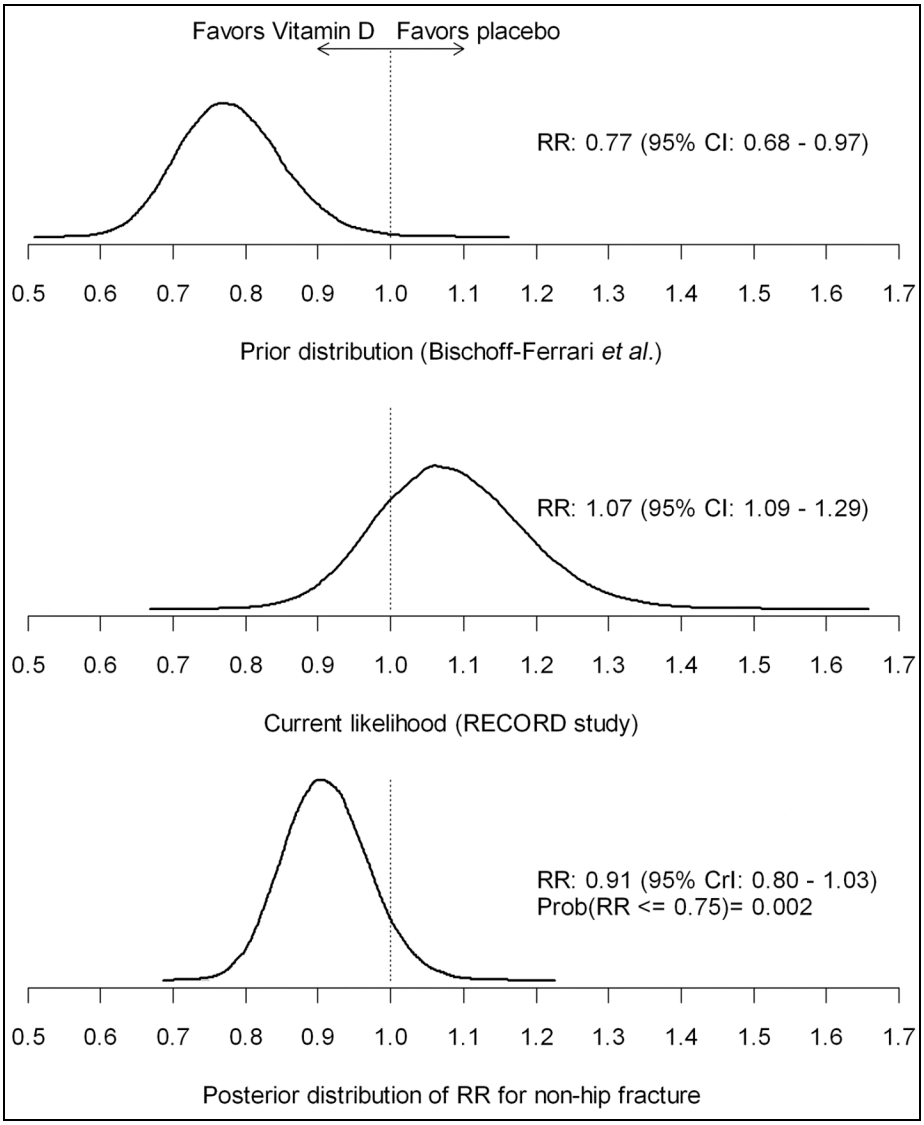

Fig. 2. Prior distribution (top panel), likelihood of RECORD data (middle), and the resulting posterior distribution (bottom) of relative risk (RR) associated with vitamin D supplementation's effect on non-hip fracture. The highlighted area under the posterior distribution curve represents the probability that vitamin D reduced fracture by at least 25%. This probability was estimated at 0.002 for non-hip fracture.

the RR reduction among study populations. This uncertainty can be quantified by three types of prior distributions: vague, skeptical, and enthusiastic priors. In the vague prior, it is assumed that the mean log RR is 0 (or a RR of 1) with a large variance (10,000). The vague prior implies a state of ignorance, in the sense that the effect could be negative as well as positive with equal probability. In the skeptical prior, based on previous meta-analysis, it is hypothesized that there is little chance (*i.e.*, 5%) that a treatment can reduce fracture risk by more than 60% (RR ≤ 0.4) or increase the risk by more than 2.4-fold. In the "positive" (enthusiastic) scenario, it is assumed that on average a treatment could reduce fracture risk by 20% (*i.e.*, RR = 0.8) with the same variance of skeptical prior. The distribution of the three scenarios is shown in Fig. 3.

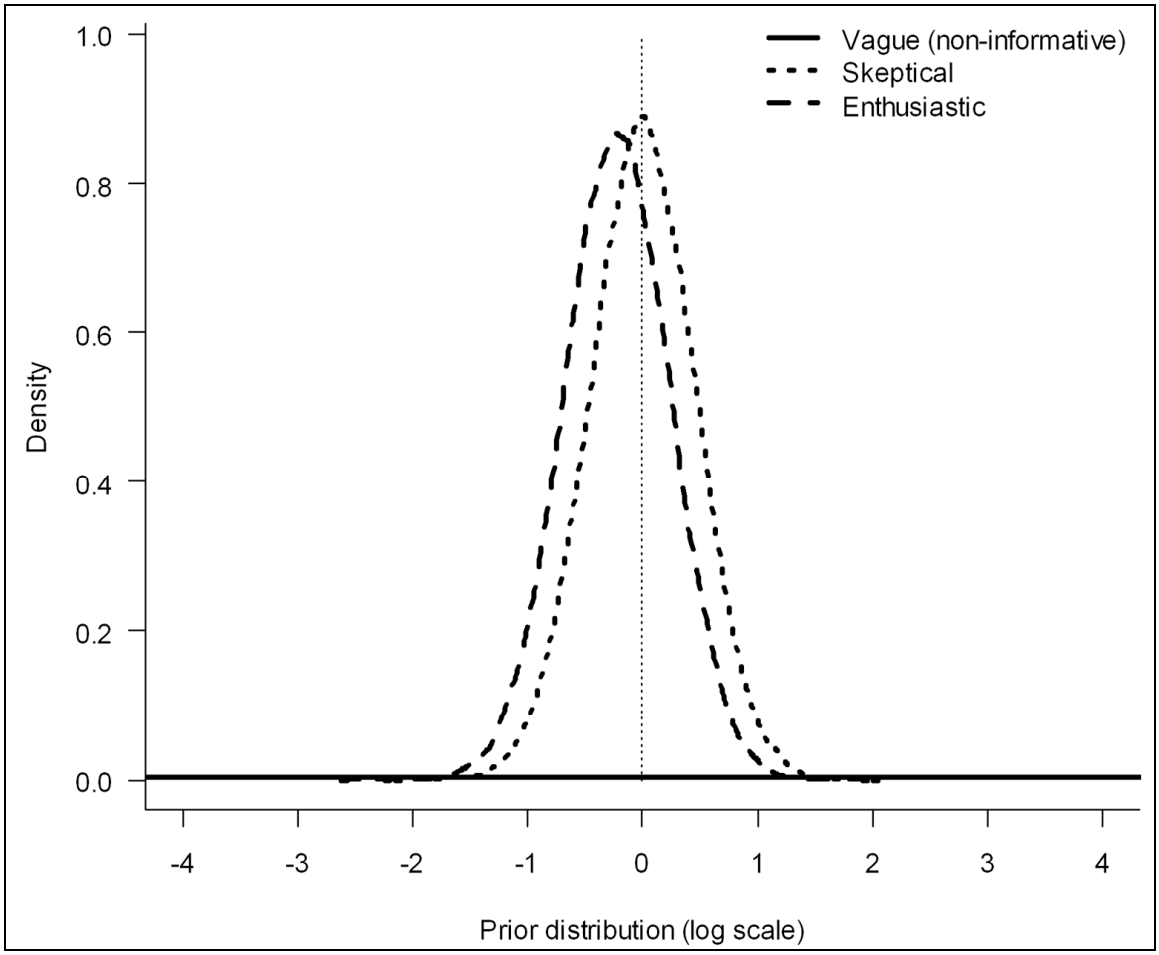

Fig. 3. Distribution of prior data. The x-axis represents log relative risk (RR) and the y-axis is the probability.

With the three priors and the actual RRs obtained from major clinical trials of anti-resorptive and anabolic therapies, the posterior probabilities of RR reduction of vertebral fracture and non-vertebral fracture were estimated (Table 2 and Table 3). If a RR reduction of 25% or more is thought to demonstrate "clinical efficacy" (*e.g.*, clinically worthwhile), then these results suggest that there is a high probability of clinical efficacy (>0.95) for alendronate, risedronate, zoledronic acid, teriparatide, strontium ranelate, and raloxifene. However, for hip fracture, apart from zoledronic acid, none of these drugs have clear clinical efficacy. Indeed, none of the probabilities of efficacy reached 0.90 (Table 3). A remarkable feature of these results is that the magnitude of posterior probability is not substantially dependent on the choice of particular prior.

In Bayesian analysis, it is possible to estimate the probability of efficacy for any threshold. Table 4 presents the posterior probability of RR reduction of >10%, >20%, >30%, >40% and >50% for vertebral and hip fractures. As expected, the posterior probability of efficacy drops as the threshold of efficacy rises. The observed RCT data are quite consistent with a RR reduction of 10% or more. The data are also consistent with the hypothesis that alendronate, risedronate, raloxifene, teriparatide and strontium ranelate (but not calcitonin) reduce vertebral fracture risk by 30% or more. On the other hand, for hip fracture, the data are largely consistent with a RR reduction of 10% or less for alendronate, risedronate and strontium ranelate. The data provide weak evidence that these drugs reduce hip fracture risk by more than 30%. However, the probability that zoledronic acid reduces

Table 2. Posterior probability of anti-vertebral fracture efficacy

Study	Relative risk reduction and 95% CI	Posterior probability of relative risk reduction of vertebral fracture by at least 25%		
		Vague prior[1]	Skeptical prior[2]	Enthusiastic prior[3]
Alendronate (5/10 mg), FIT-1 study (32)	47 (32-59)	0.996	0.992	0.995
Alendronate (5/10 mg), FIT-2 study (33)	44 (20-61)	0.944	0.893	0.923
Risedronate (5 mg), VERT-US study (34)	49 (27-64)	0.984	0.961	0.974
Risedronate (5 mg), VERT-MN study (35)	41 (18-57)	0.927	0.876	0.908
Zoledronic acid (5 mg), HORIZON (42)	70 (62-76)	1.000	1.000	1.000
Raloxifene (60 mg), MORE-1 study (37)	50 (20-60)	0.989	0.972	0.982
Raloxifene (60 mg), MORE-2 study (37)	30 (10-40)	0.748	0.695	0.732
Calcitonin (200 IU), PROOF study (38)	33 (3-53)	0.729	0.629	0.695
Teriparatide (20 mg) (39)	65 (45-78)	0.999	0.996	0.998
Strontium ranelate (40)	41 (27-52)	0.987	0.979	0.984

Notes: [1]In the vague prior, it is assumed that the log relative risk (RR) is normally distributed with mean 0 (on average, there is no effect) and variance of 10,000. [2]In the skeptical prior, based on previous meta-analysis, it is hypothesized that there is a 95% chance that the RR varies from 0.4 to 2.4, with the average being 1 (no effect). This is equivalent to the statement Pr(log RR ≤ -0.916) = 0.025, and by symmetry, Pr(log RR ≥ 0.875) = 0.025. With this skeptical assumption and by normal distribution, it can be shown that the prior variance is 0.209. Therefore the skeptical prior distribution can be specified with mean 0 and variance of 0.209. [3]In the "positive" (enthusiastic) scenario, it is assumed that on average a treatment could reduce fracture risk by 20% (*i.e.*, RR = 0.8), with the same variance of skeptical prior. Under this enthusiastic assumption, it can be shown that the prior distribution is characterized by a mean of -0.223 and variance of 0.209.

hip fracture by 30% or more is almost 1, which is highest among the drugs.

It is commonly stated that current antiresorptive and anabolic agents reduce fracture risk by 50%. However, Fig. 4 shows that the probability of this effect size is modest. Only teriparatide and zoledronic acid have a posterior probability (of RR reduction of 50%) higher than 0.9; none of the remaining drugs could achieve this probability. For hip fracture, the probability that any of the drugs results in hip fracture risk reduction of >50% is consistently less than 0.7.

In summary, results from the Bayesian analysis suggest that currently available pharmacologic therapies could reduce vertebral fracture risk by 30% or less, and that some drugs could reduce hip fracture risk by at most 10%. Current RCT data are not consistent with the hypothesis that most drugs, except zoledronic acid and teriparatide, reduce vertebral fracture risk by 50% or more. Of course, this analysis does not in any way invalidate the conclusions of these original trials, but rather provides an alternative and complementary interpretation of probable effect sizes.

Toward a Bayesian interpretation of RCTs

The call for evidence-based practice has driven the growing use of RCTs as a scientific means to establish evidence. Since its formal introduction in 1951 by Bradford Hill (43), the RCT is now considered the apotheosis of scientific advances in clinical medicine. In osteoporosis research, the RCT has been used as a tool to identify effective therapeutic and preventive treatments that are currently used in clinical practice. Increasingly, RCTs are designed with large to very large sample sizes to detect ever

Table 3. Posterior probability of anti-hip fracture efficacy

Study	Relative risk reduction and 95% CI	Posterior probability of relative risk reduction of hip fracture by at least 25%		
		Vague prior	Skeptical prior	Enthusiastic prior
Alendronate, FIT-1 study (32)	51 (1-77)	0.873	0.687	0.787
Alendronate (5/10 mg), FIT-2 study, T-scores < -2.5 (33)	56 (3-82)	0.893	0.681	0.790
Alendronate (5/10 mg), FIT-2 study, T-scores < -1.6 (33)	21 (+44 to -57)	0.433	0.311	0.413
Risedronate (2.5 and 5 mg), HIP study, 70-80 years with osteoporosis (41)	40 (10-60)	0.860	0.765	0.822
Risedronate (2.5 and 5 mg), HIP study, >80 years (41)	20 (+20 to -40)	0.357	0.285	0.348
Zoledronic acid (5 mg), HORIZON (42)	41 (17-58)	0.916	0.857	0.892
Calcitonin, PROOF study (38)	50 (+60 to -80	0.778	0.509	0.652
Raloxifene (60/120 mg), MORE study (37)	+10 (+90 to -40)	0.096	0.076	0.123
Strontium ranelate, ITT analysis (36)	11 (+19 to -34)	0.127	0.101	0.130
Strontium ranelate, high risk group (36)	43 (3-67)	0.841	0.703	0.783

Notes: See notes from Table 2.

smaller effect sizes. In this setting, it is important to explicitly quantify the hypothesis and effect size being tested in these trials. The Bayesian approach presented here offers an attractive method for such a quantification.

Clinical trial data represent an important source of medical knowledge, and knowledge should be accumulated or updated when new data become available. Regrettably, the issue of how to formally update knowledge has received little attention from clinical researchers, as results of an RCT are often considered in isolation from previous results. The Bayesian method allows for the synthesis of existing knowledge, including expert opinions, with previous data into a more coherent and more reliable conclusion. As a result, Bayesian inference is much less likely to be prone to “significant” results and provides protection against false positive findings that are highly prevalent in the medical literature, especially in studies with low plausibility (4).

A reader of an RCT is confronted with three questions: what should one do, what does one believe, and how should one interpret the result as evidence (44)? In the face of uncertainty, the Bayesian approach can help to address the second and third questions. It is particularly useful in studies where a statistically significant result is too small to be clinically relevant, or a statistically non-significant result is large enough to be clinically important. For example, in the study of the supplementation of vitamin C and E during pregnancy, although the traditional analysis did not reveal a statistically significant risk reduction (RR = 0.79; 95% CI: 0.61 – 1.02) (21), the data are also consistent with a probability of 0.96 that vitamin C and E reduced the risk of death or other serious outcomes in infants. Perhaps the most helpful inference the Bayesian model can offer is its estimates of direct probability statements about any differences that are of clinical interest (45;46). Therefore, the Bayesian approach does not create a positive result from a “negative” trial, because if an intervention has no

IBMS BoneKEy. 2009 August;6(8):279-294
http://www.bonekey-ibms.org/cgi/content/full/ibmske;6/8/279
doi: 10.1138/20090391

Table 4. Posterior probability of relative risk (RR) reduction for vertebral and hip fractures

Study	Relative risk reduction and 95% CI	Posterior probability of relative risk reduction by				
		More than 10%	More than 20%	More than 30%	More than 40%	More than 50%
Vertebral fracture						
Alendronate (5/10 mg), FIT-1 study (32)	47 (32-59)	0.999	0.999	0.984	0.831	0.325
Alendronate (5/10 mg), FIT-2 study (33)	44 (20-61)	0.995	0.974	0.888	0.647	0.268
Risedronate (5 mg), VERT-US study (34)	49 (27-64)	0.999	0.994	0.960	0.816	0.456
Risedronate (5 mg), VERT-MN study (35)	41 (18-57)	0.994	0.968	0.850	0.541	0.157
Zoledronic acid (5 mg), HORIZON (42)	70 (62-76)	1.000	1.000	1.000	1.000	0.999
Raloxifene (60 mg), MORE-1 study (37)	50 (20-60)	1.000	0.999	0.999	0.993	0.500
Raloxifene (60 mg), MORE-2 study (37)	30 (10-40)	0.992	0.902	0.500	0.068	0.000
Calcitonin (200 IU), PROOF study (38)	33 (3-53)	0.944	0.831	0.594	0.275	0.057
Teriparatide (20 mg) (39)	65 (45-78)	0.999	0.999	0.998	0.989	0.936
Strontium ranelate (40)	41 (27-52)	0.999	0.998	0.945	0.562	0.061
Hip fracture						
Alendronate, FIT-1 study (32)	51 (1-77)	0.948	0.905	0.830	0.707	0.521
Alendronate (5/10 mg), FIT-2 study, T-scores < -2.5 (33)	56 (3-82)	0.952	0.918	0.860	0.765	0.617
Alendronate (5/10 mg), FIT-2 study, T-scores < -1.6 (33)	21 (+44 to -57)	0.664	0.516	0.347	0.186	0.068
Risedronate (2.5 and 5 mg), HIP study, 70-80 years with osteoporosis (41)	40 (10-60)	0.975	0.918	0.772	0.500	0.189
Risedronate (2.5 and 5 mg), HIP study, >80 years (41)	20 (+20 to -40)	0.747	0.500	0.225	0.052	0.004
Zoledronic acid (5 mg), HORIZON (42)	41 (17-58)	0.992	0.960	0.837	0.538	0.170
Calcitonin, PROOF study (38)	50 (+60 to -80)	0.866	0.812	0.737	0.634	0.500
Raloxifene (60/120 mg), MORE study (37)	+10 (+90 to -40)	0.247	0.139	0.062	0.019	0.004
Strontium ranelate, ITT analysis (36)	11 (+19 to -34)	0.530	0.239	0.056	0.004	0.000
Strontium ranelate, high risk group (36)	43 (3-67)	0.952	0.891	0.772	0.573	0.317

effect, then the posterior probability of effect will be less than 0.5.

However, a major criticism of the Bayesian approach is the use of what may be considered subjective prior information, which can vary from one study to another. While subjective prior information can be problematic, the explicit specification of such information in the analysis creates an environment for better communication of results. The Bayesian approach requires a much more extensive computation that was an obstacle to its application in practice.

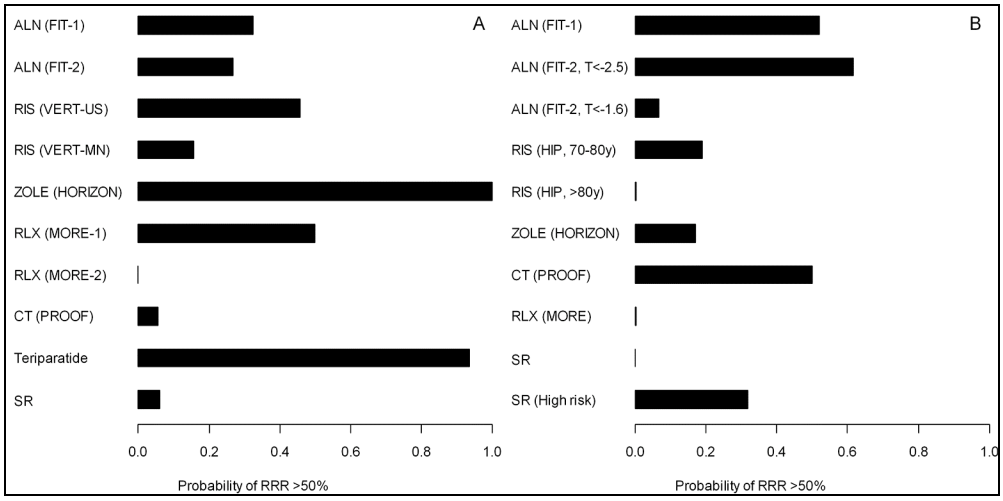

Fig. 4. Posterior probability that a drug reduced the risk of (A) vertebral fracture and (B) hip fracture by at least 50% (based on "vague" prior information). SR: Strontium ranelate; CT: Calcitonin; ZOLE: Zoledronic acid; RLX: Raloxifene; RIS: Risedronate; ALN: Alendronate.

However, with the advent of user-friendly software (47) and powerful personal computers, this computational complexity is no longer an issue.

Despite attempts over the past ten years to introduce Bayesian inference into the field of osteoporosis (48-53), the approach is still under-utilized. The recent revived interest in the Bayesian approach has been prominent in medical research and risk assessment, and it is expected that the 21st century research method will be a synthesis of frequentist and Bayesian methods. Currently, several medical journals have encouraged the application of Bayesian approaches in the analysis and reporting of clinical trials (54-56). In this environment, osteoporosis research could benefit from alternative methods offered in the Bayesian approach, and could adopt them with an amount of vigor similar to that which has characterized other research settings.

Acknowledgments

The author is supported by an Australian National Health and Medical Research Council Research Senior Fellowship. The author thanks Dr. Nguyen D. Nguyen for his technical assistance in R simulation and the Bayesian analysis of the data shown in Fig. 1 and Fig 2. The author also thanks Professor Ego Seeman for his critical and stimulating comments on an early draft of this manuscript.

Conflict of Interest: Dr. Nguyen reports that he has received lecture fees or consulting fees from Merck Sharp & Dohme (Vietnam), Novartis, sanofi-aventis, Roche Diagnostics (Vietnam), Bridge Health Care Ltd. (Australia), and Solvay Pharmaceuticals.

Peer Review: This article has been peer-reviewed.

References

1. Wootton R, Bryson E, Elsasser U, Freeman H, Green JR, Hesp R, Hudson EA, Klenerman L, Smith T, Zanelli J. Risk factors for fractured neck of femur in the elderly. *Age Ageing*. 1982 Aug;11(3):160-8.

2. Diamond GA, Kaul S. Prior convictions: Bayesian approaches to the analysis and interpretation of clinical megatrials. *J Am Coll Cardiol*. 2004 Jun 2;43(11):1929-39.

3. Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. *Ann Intern Med*. 1999 Jun 15;130(12):995-1004.

4. Ioannidis JP. Why most published research findings are false. *PLoS Med*. 2005 Aug;2(8):e124.

5. Matthews RJ. Why should clinicians care about Bayesian methods? J *Stat Inf Plan*. 2001;94:43–58.

6. Berger JO, Sellke T. Testing a point null hypothesis: the irreconciliability of p values and evidence (with discussion). *J Amer Statist Assoc*. 1987;82:112–22.

7. Ioannidis JP. Contradicted and initially stronger effects in highly cited clinical research. *JAMA*. 2005 Jul 13;294(2):218-28.

8. Nelder JA. From statistics to statistical science. *Statistician*. 1999;48:257-69.

9. Senn SJ. Falsificationism and clinical trials. *Stat Med*. 1991 Nov;10(11):1679-92.

10. Cohen J. The earth is round (P < .05). *Am Psychol*. 1994 Dec;49(12):997-1003.

11. Neyman J, Pearson ES. On the problem of most efficient tests of statistical hypotheses. *Philos Trans Roy Soc A*. 1933;231:289-337.

12. Gigerenzer G. Mindless statistics. *Journal of Socio-Economics*. 2004 Nov;33(5):587-606.

13. Berkson J. Some difficulties of interpretation encountered in the application of the chi square test. *J Amer Statist Assoc*. 1938;33:526-42.

14. Berkson J. Tests of significance considered as evidence. *J Am Statist Assoc*. 1942;37:325–35.

15. Fisher RA. The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain*. 1926;33:503-13.

16. Fisher RA. *Statistical Methods for Research Workers*. London: Oliver and Boyd; 1950.

17. Chan AW, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA*. 2004 May 26;291(20):2457-65.

18. Chan AW, Krleza-Jerić K, Schmid I, Altman DG. Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. *CMAJ*. 2004 Sep 28;171(7):735-40.

19. Schmidth FL, Hunter JE. Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In Harlow LL, Mulaik SA, Steiger JH, eds. *What If There Were No Significance Tests?* London: Lawrence Erlbaum; 1997.

20. Rothman KJ. A show of confidence. *N Engl J Med*. 1978 Dec 14;299(24):1362-3.

21. Rumbold AR, Crowther CA, Haslam RR, Dekker GA, Robinson JS; ACTS Study Group. Vitamins C and E and the risks of preeclampsia and perinatal complications. *N Engl J Med*. 2006 Apr 27;354(17):1796-806.

22. Lilford RJ. Ethics of clinical trials from a bayesian and decision analytic perspective: whose equipoise is it anyway? *BMJ*. 2003 May 3;326(7396):980-1.

23. Berry DA. Bayesian clinical trials. *Nat Rev Drug Discov*. 2006 Jan;5(1):27-36.

24. Spiegelhalter DJ, Abrams KR, Myles JP. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. West Sussex, England: John Wiley & Sons Ltd.; 2004.

25. Spiegelhalter DJ, Freedman LS, Parmar MB. Bayesian approaches to randomized trials. *J R Stat Soc*. 1994;157(3):357-416.

26. Bayes RT. Essay toward solving a problem in the doctrine of chances. P*hilos Trans R Soc*. 1763;53:370-418.

27. Browner WS, Newman TB. Are all significant P values created equal? The analogy between diagnostic tests and clinical research. *JAMA*. 1987 May 8;257(18):2459-63.

28. Bischoff-Ferrari HA, Willett WC, Wong JB, Giovannucci E, Dietrich T, Dawson-Hughes B. Fracture prevention with vitamin D supplementation: a meta-analysis of randomized controlled trials. *JAMA*. 2005 May 11;293(18):2257-64.

29. Grant AM, Avenell A, Campbell MK, McDonald AM, MacLennan GS, McPherson GC, Anderson FH, Cooper C, Francis RM, Donaldson C, Gillespie WJ, Robinson CM, Torgerson DJ, Wallace WA; RECORD Trial Group. Oral vitamin D3 and calcium for secondary prevention of low-trauma fractures in elderly people (Randomised Evaluation of Calcium Or vitamin D, RECORD): a randomised placebo-controlled trial. *Lancet*. 2005 May 7-13;365(9471):1621-8.

30. Winkler RL. *An Introduction to Bayesian Inference and Decision*. New York: Holt, Rinehart and Winston; 1972.

31. Milling T, Holden C, Melniker L, Briggs WM, Birkhahn R, Gaeta T. Randomized controlled trial of single-operator vs. two-operator ultrasound guidance for internal jugular central venous cannulation. *Acad Emerg Med*. 2006 Mar;13(3):245-7.

32. 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. Randomised trial of effect of alendronate on risk of fracture in women with existing vertebral fractures. Fracture Intervention Trial Research Group. *Lancet*. 1996 Dec 7;348(9041):1535-41.

33. Cummings SR, Black DM, Thompson DE, Applegate WB, Barrett-Connor E, Musliner TA, Palermo L, Prineas R, Rubin SM, Scott JC, Vogt 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 the Fracture Intervention Trial. *JAMA*. 1998 Dec 23-30;280(24):2077-82.

34. Harris ST, Watts NB, Genant HK, McKeever CD, Hangartner T, Keller M, Chesnut CH 3rd, Brown J, Eriksen EF, Hoseyni MS, Axelrod DW, Miller PD. Effects of risedronate treatment on vertebral and nonvertebral fractures in women with postmenopausal osteoporosis: a randomized controlled trial. Vertebral Efficacy With Risedronate Therapy (VERT) Study Group. *JAMA*. 1999 Oct 13;282(14):1344-52.

35. Reginster J, 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. Vertebral Efficacy with Risedronate Therapy (VERT) Study Group. *Osteoporos Int*. 2000;11(1):83-91.

36. Reginster JY, Felsenberg D, Boonen S, Diez-Perez A, Rizzoli R, Brandi ML, Spector TD, Brixen K, Goemaere S, Cormier C, Balogh A, Delmas PD, Meunier PJ. Effects of long-term strontium ranelate treatment on the risk of nonvertebral and vertebral fractures in postmenopausal osteoporosis: Results of a five-year, randomized, placebo-controlled trial. *Arthritis Rheum*. 2008 Jun;58(6):1687-95.

37. Ettinger B, Black DM, Mitlak BH, Knickerbocker RK, Nickelsen T, Genant HK, Christiansen C, Delmas PD, Zanchetta JR, Stakkestad J, Glüer 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. Multiple Outcomes of Raloxifene Evaluation (MORE) Investigators. *JAMA*. 1999 Aug 18;282(7):637-45.

38. Chesnut CH 3rd, Silverman S, Andriano K, Genant H, Gimona A, Harris S, Kiel D, LeBoff M, Maricic M, Miller P, Moniz C, Peacock M, Richardson P, Watts N, Baylink D. A randomized trial of nasal spray salmon calcitonin in

postmenopausal women with established osteoporosis: the prevent recurrence of osteoporotic fractures study. PROOF Study Group. *Am J Med*. 2000 Sep;109(4):267-76.

39. Neer RM, Arnaud CD, Zanchetta JR, Prince R, Gaich GA, Reginster JY, Hodsman AB, Eriksen EF, Ish-Shalom S, Genant HK, Wang O, Mitlak BH. Effect of parathyroid hormone (1-34) on fractures and bone mineral density in postmenopausal women with osteoporosis. *N Engl J Med*. 2001 May 10;344(19):1434-41.

40. Meunier PJ, Roux C, Seeman E, Ortolani S, Badurski JE, Spector TD, Cannata J, Balogh A, Lemmel EM, Pors-Nielsen S, Rizzoli R, Genant HK, Reginster JY. The effects of strontium ranelate on the risk of vertebral fracture in women with postmenopausal osteoporosis. *N Engl J Med*. 2004 Jan 29;350(5):459-68.

41. McClung MR, Geusens P, Miller PD, Zippel H, Bensen WG, Roux C, Adami S, Fogelman I, Diamond T, Eastell R, Meunier PJ, Reginster JY; Hip Intervention Program Study Group. Effect of risedronate on the risk of hip fracture in elderly women. Hip Intervention Program Study Group. *N Engl J Med*. 2001 Feb 1;344(5):333-40.

42. Black DM, Delmas PD, Eastell R, Reid IR, Boonen S, Cauley JA, Cosman F, Lakatos P, Leung PC, Man Z, Mautalen C, Mesenbrink P, Hu H, Caminis J, Tong K, Rosario-Jansen T, Krasnow J, Hue TF, Sellmeyer D, Eriksen EF, Cummings SR; HORIZON Pivotal Fracture Trial. Once-yearly zoledronic acid for treatment of postmenopausal osteoporosis. *N Engl J Med*. 2007 May 3;356(18):1809-22.

43. Hill AB. The clinical trial. *Br Med Bull*. 1951;7(4):278-82.

44. Royall R. *Statistical Evidence: A Likelihood Paradigm*. Boca Raton: Chapman & Hall/CRC; 1997.

45. Lilford RJ, Braunholtz D. The statistical basis of public policy: a paradigm shift is overdue. *BMJ*. 1996 Sep 7;313(7057):603-7.

46. Burton PR. Helping doctors to draw appropriate inferences from the analysis of medical studies. *Stat Med*. 1994 Sep 15;13(17):1699-713.

47. Spiegelhalter DJ, Thomas A, Best NJ, Lunn D. *WinBUGS User Manual Version 1.4*, MRC Biostatistics Unit; 2003. http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/manual14.pdf

48. Nguyen TV. Pharmacogenetics of anti-resorptive therapy efficacy: a Bayesian interpretation. *Osteoporos Int*. 2005 Aug;16(8):857-60.

49. Nguyen TV, Pocock N, Eisman JA. Interpretation of bone mineral density measurement and its change. *J Clin Densitom*. 2000 Summer;3(2):107-19.

50. Nguyen ND, Eisman JA, Nguyen TV. Anti-hip fracture efficacy of bisphosphonates: a Bayesian analysis of clinical trials. *J Bone Miner Res*. 2006 Feb;21(2):340-9.

51. Nguyen ND, Wang CY, Eisman JA, Nguyen TV. On the association between statin and fracture: A Bayesian consideration. *Bone*. 2007 Apr;40(4):813-20.

52. Tran BN, Nguyen ND, Eisman JA, Nguyen TV. Association between LRP5 polymorphism and bone mineral density: a Bayesian meta-analysis. *BMC Med Genet*. 2008 Jun 27;9;55.

53. Sadatsafavi M, Moayyeri A, Wang L, Leslie WD. Optimal decision criterion for detecting change in bone mineral density during serial monitoring: a Bayesian approach. *Osteoporos Int*. 2008 Nov;19(11):1589-96.

54. Appendix. Information for authors. *Ann Intern Med*. 2002;136:A1-A5.

55. Rennie D. Fourth International Congress on Peer Review in Biomedical Publication. *JAMA*. 2002 Jun 5;287(21):2759-60.

56. Guyatt GH, Rennie D. Users' guides to the medical literature. *JAMA*. 1993 Nov 3;270(17):2096-7.