Evidence-based medicine and common sense: practical and ethical issues in clinical trials for osteoporosis

Robert P Heaney, MD
Creighton University Medical Center, 601 N. 30th St., Suite 4841, Omaha, NE 68131, USA
Tel.: +1 402 280 4029; Fax: +1 402 280 4751; rheaney@creighton.edu

For ethical reasons, testing the efficacy of new treatments for chronic diseases such as osteoporosis has hit a brick wall. Furthermore, existing trial designs, although satisfying requirements for drug registration, typically produce results with limited generalizability. Compounding the problem, these results often serve as the basis for treatment guidelines, which, in turn, are assembled by policy-makers and analysts who often do not understand the biology of the systems concerned. In this perspective, the details of these problems are briefly described and the broad outlines of some solutions suggested.

Clinical research today is enmeshed in an orgy of what is called evidence-based medicine (EBM). By that name, EBM's advocates and practitioners claim the moral high ground, implying that other approaches to clinical practice are not based on evidence. Service has commented on the fallacy of that premise, noting most tellingly that EBM has not met its own criterion for efficacy; specifically, there are no randomized controlled trials (RCTs) demonstrating that medicine practiced according to EBM-based guidelines produces better patient outcomes than those it purports to displace [1]. This awkward fact, like the emperor's lack of clothes, is widely ignored by EBM practitioners and enthusiasts.

EBM ranks the evidential value of various clinical study designs, giving primary credence to RCTs and least weight to case reports and expert opinion. It is true that the experimental design is the only investigational type that supports strong causal inference. However, while acknowledging this, one must add that, when applied in real clinical situations, RCTs suffer severe limitations which are inadequately recognized and may greatly weaken the conclusions that can be correctly drawn from them.

In this brief perspective, I discuss some of these limitations, describe the asymmetry of the evidence used to establish benefit and harm, point out examples of the biological errors in many of the published meta-analyses and systematic evidence reviews (on both of which EBM battens) and highlight substantial ethical and cost problems regarding the use of RCTs for the evaluation of possibly improved treatments for established diseases, such as osteoporosis. Finally, I suggest some possible solutions to these problems.

Keywords: antiresorptive drugs, bone remodeling, calcium, cohort study, evidence-based medicine, fracture, osteoporosis, randomized controlled trial, nutrition, vitamin D

future medicine part of fsg

Generalizability & the controlled trial

The most obvious problem with the RCT is limited generalizability from its conclusions [2]. While the efficacy of the agent tested in a positive RCT can be confidently asserted (with a definable chance of being wrong; i.e., the P-value), it is difficult to decide to which patients the results might be applicable. This is because study patients often differ substantially from patients seen in physicians' offices, exhibiting both healthy volunteer bias and relative absence of the many comorbidities and co-therapies that are typical of older patients. As an illustration of this often overlooked point, Dowd and colleagues analyzed all new diagnoses of osteoporosis from our center in a 40-month period, testing them against the inclusion and exclusion criteria for four multicenter, industry-sponsored trials being conducted at the time [3]. As it turned out, the most inclusive of those trials would have admitted only 25% of our real patients, and the most restrictive, just 4%. Various comorbidities and their treatments were the principal reasons for exclusion. However, whatever the reason, somewhere from 75 to 96% of real patients would not have been included in the trials concerned. Application of these trial results, both efficacy and safety, to such patients is therefore uncertain, particularly in view of frequent comorbidity.

I do not criticize the design of such trials, since the pharmaceutical industry must consider economic factors as it attempts to meet the efficacy requirements for drug registration. Excluding extraneous sources of variability in response to treatment greatly reduces the size of the already large samples required, for example, to demonstrate antifracture efficacy. The fault lies not so much in the industry, or even in the

regulatory process, as in the subsequent use of such studies to develop treatment guidelines for patients who differ substantially from those who were studied, specifically in EBM as practiced.

A second problem with the conduct of RCTs is the loss of sampling units. Such loss is understood to reduce the power of an investigation, and designers compensate by over-recruiting. However, this approach ignores the fact that the inferential power of the experimental design lies mainly in the fact that factors potentially influencing the outcome are randomly distributed between the contrasting groups. Since drop-outs from trials are manifestly not random (and, in any case, can almost never be shown to be so), the allocation of treatment to those remaining can no longer be said to be random. Thus, inferences drawn from random processes (the basis for strong causal inference) are no longer secure. Recognizing this, Peto and colleagues go so far as to state that investigators should allow no drop-outs [4]. As that counsel is both unfeasible and unethical, it is simply ignored; but ignoring the problem does not make it go away.

The approach to drop-outs usually taken is intention-to-treat (ITT). ITT analyzes outcomes on the basis of group assignment, including within a given treatment group all assigned participants even if they dropped out, stopped taking medication or voluntarily crossed over to the other treatment. This inevitably biases towards the null hypothesis and, when loss of sampling units or non-adherence is substantial, ITT virtually guarantees that only strongly efficacious agents have a reasonable chance of being detected in a real-world clinical trial. Failure to find positive results in a trial analyzed by ITT can never prove that the agent was not itself efficacious. It is not simply that it is never possible to prove the null hypothesis. Rather, a null effect by ITT usually reflects an indeterminate mix of the agent's efficacy and the effectiveness of its means of delivery (e.g., voluntary daily pill taking). Nevertheless papers using this approach commonly draw precisely such an unsupportable (and incorrect) conclusion. For example, in the past 18 months, the results of three large trials of vitamin D supplementation were reported [5–7]. Each found no significant difference in fracture rate, and despite adherence rates of 50% or less, the study authors dogmatically asserted that vitamin D was not effective. And this was despite the fact that another large trial, but one with high adherence, had reported efficacy just 2 years earlier [8].

Investigators defend such conclusions by stating that a regimen which does not elicit substantial compliance is effectively equivalent to a regimen which is inherently ineffective; specifically, if you cannot get people to take the medicine, then that is not different from a medicine that is itself without benefit. That, of course, is incorrect. At the very least, if the treatment agent is efficacious, then it works in those who take it. However, more importantly, a conclusion of ineffectiveness distracts from the challenge involved in devising optimal deployment of efficacious agents. For example, of the four vitamin D trials just mentioned [5–8], three required daily pill-taking and all three had both poor compliance and a null effect. By contrast, the fourth trial used only three large doses per year, implemented by a special mailing, had high, documented treatment adherence and found a significant treatment effect. In a sense, all four trials constituted tests not so much of vitamin D, but of ways of improving vitamin D status in the population.

It might seem that per protocol (PP) analysis (analyzing results only from those who had adhered to the treatment regimen and completed the study) would provide the information actually required to determine whether the agent itself had the desired effect. Unfortunately, as adherence itself is not randomly distributed, inferential capacity is limited. Even when PP analysis is specified in advance of the study, it cannot avoid the huge effect that loss of sampling units has on investigative power. While a loss of as few as 10% of the participants might be planned for, it can be shown that the corresponding reduction in power (since losses are nonrandom) could easily be greater than 50% [9].

In brief, positive RCTs have limited generalizability, and null-effect RCTs cannot support a conclusion of inefficacy. For these reasons, RCTs, as conducted, often provide a very shaky foundation on which to erect the edifice of clinical practice guidelines.

Asymmetry of evidence for benefit & harm

For drug treatments to be approved, RCTs are required. But for drugs to be withdrawn after prior approval, the evidence usually comes from case reports and cohort studies. It is curious that both the regulatory and the clinical investigative communities use much more relaxed criteria for imputation of harm than for evidence of benefit. Only very rarely is harm established by an RCT.

Indeed, most experts are of the view that it would not be ethically permissible to mount an RCT designed to test a hypothesis of harm.

One might counter that this is precisely as it should be. The Hippocratic maxim *primum, non nocere* gives priority to avoiding harm. However, harm also accrues to the larger community when an efficacious agent is not made available, its deployment inordinately delayed or its development cost exaggerated because of the efficacy standards to which it is held. In fact, accountability itself is asymmetrical. Regulatory authorities can be criticized and companies sued if approved agents have harmful side effects, whereas failure to develop, approve or market products carries far fewer public consequences. I do not suggest that the standards for evidence of harm be changed. Rather, I believe we should re-evaluate why we are more skeptical about benefit than harm. Do we take the cynical view that new treatments, even innovative ones, have only marginal utility, and that the benefits to perhaps millions of patients are somehow trumped by the risk of harm to a few? What calculus ought we use? And what are its assumptions?

Systematic reviews & the biology of the system concerned

A third problem lies in the meta-analyses and systematic evidence-based reviews that are much loved within EBM ranks. To date, so far as one can judge from their credentials as well as from how they treat the data, the analysts who assemble these reviews seem often to be unencumbered by knowledge of the relevant biology. Two common examples will serve to illustrate what I mean. They relate to the measure of efficacy for skeletal response and to how one ought to pool studies.

Bone mass change is one commonly used measure of effect. It is produced by a shift in the balance between bone formation and resorption. Since these activities are cell-based and do not respond simultaneously to various interventions, the time course of measured bone density (or mass) is curvi-linear. Antiresorptives all reduce resorption first, followed weeks or months later by a secondary reduction in bone formation. For this reason, pooling data from studies of varying durations is not a simple matter [9].

To characterize the effects of antiresorptives in a treatment trial, two numbers are needed: the degree of remodeling suppression and the induced change in tissue-level bone balance. Both can be derived from analysis of the time course of the change in bone mass following onset of treatment. As an illustration, analysis of data from a trial by Peacock and colleagues revealed that the intervention concerned reduced bone remodeling by 8% and caused a 0.3%/year improvement in bone balance [10,11]. However, if the study had been analyzed simply on the change in bone mineral density (BMD), the effect of treatment, expressed in percent change per year, would have been +0.7% per year at 1 year, +0.1% per year at 2 years, -0.01% per year at 3 years, and -0.05%/year at 4 years. In other words, a single study could be interpreted to show that bone was being gained, lost or was essentially unchanged, depending upon when one conducted the analysis. This is why meta-analyses pooling various studies must use the two critical numbers (which together describe the entire time course), not the time-dependent changes in BMD. To date, no meta-analysis of bone mass change has done this.

Another example, more egregious still, is the failure of both study designers and study analysts to reckon with certain features of the interventions concerned. Calcium and vitamin D are good cases in point, as both, like iron, function as threshold nutrients. For all such nutrients, the threshold is the intake at which the effect reaches its maximal value, with further increases in intake producing no further change in the measured outcome. Biological response (e.g., change in bone mass or in calcium absorption) can be expected to be seen only when one compares low intakes with adequate or high intakes. Studies which have no low calcium intake control group (e.g., the Women's Health Initiative [5]) would be predicted to show no effect [12]. To pool such studies with others which do have a low calcium control group in some sort of a systematic review or meta-analysis borders on wrong-headedness. Unfortunately, I have yet to see a meta-analysis of a nutritional intervention in the field of bone biology that lists the presence of a low-intake control group as a criterion for inclusion.

A part of the reason may be the general tendency to treat nutrients as if they were pharmacological agents. Drugs can be contrasted with a drug-free state whereas nutrients cannot. Drugs are designed, for the most part, to act alone, whereas nutrients virtually always act in concert with one another. Drugs produce one, or at most a few, major effects, while nutrients produce many small beneficial effects across multiple body systems. To evaluate nutrient effects using a drug model is thus often procrustean.

Nevertheless, several meta-analytic approaches to these issues have systematically excluded studies that used more than one nutritional agent, for example, the combination of calcium and vitamin D. Doing so ignores a world of quantitative physiology which demonstrates that calcium supplementation does not yield much calcium unless vitamin D status is normal, and that vitamin D will not produce much of an increase in calcium absorption if calcium intakes are low. It is precisely the combination of the two that gets more calcium into the body. Similarly, based in the biologic fact that bone is 50% protein by volume, a growing body of evidence also indicates that even calcium plus vitamin D are without much effect on bone mass if protein intake is not adequate [13,14]. In brief, unless steps are taken to ensure fully adequate intakes of all the other nutrients important for a given body system, tests of single nutrient effects may be useless and will often be literally meaningless.

These are only a few examples of many that could be cited. Meta-analysts are trained as epidemiologists, not as bone biologists and, hence, they might be expected to be unaware of the unique features of the system they are studying. However, that does not excuse or justify using inappropriate assemblages of evidence as a basis for clinical practice guidelines.

Efficacy requirements, costs & ethics

Currently, efficacy requirements for registration of antiosteoporosis agents include the demonstration of reduced fracture risk. That criterion is so inherently logical as to appear beyond criticism. After all, the principal consequence of osteoporosis is increased fragility, and what we want of an effective agent is precisely a reduction in fragility.

The problem is that, while common at a population level, fractures are rare in the lives of individuals, with a typical osteoporotic patient having a fracture risk of about one chance in 20 in a given year. Most of the patients in a trial will, therefore, not sustain a fracture during the course of the study, which means that studies must be large and usually of several years' duration. In comparison with an end point such as a change in bone mass or remodeling, a fracture end point raises the cost to demonstrate efficacy by an order of magnitude.

Of even greater import is the fact that such a demonstration of efficacy requires that there be a control group which must sustain otherwise preventable fractures if the agent being tested is to be shown to be efficacious. Since osteoporotic fractures are nontrivial events, with significant morbidity, mortality and cost, the very use of a placebo-control group in a disease for which there is now recognized efficacious therapy becomes ethically unacceptable [15]. Unfortunately, equivalence or noninferiority trials, no matter how impeccably designed and executed, do not have the same persuasive power (or marketing attractiveness) as true placebo-controlled designs [16]. Thus, in the field of osteoporosis, as in most chronic diseases in medicine, placebo-controlled designs become unacceptable once there is a single efficacious agent available to treat the disorder concerned. The problem is not confined to evaluation of newly developed pharmacological agents. One of the likely reasons for the absence of a low calcium control group in so many trials is the ethical barrier of imposing on vulnerable participants a nutrient intake defined as inadequate in myriad public policy statements.

In brief, the very logic of the RCT leads to an unfortunate conclusion, that is, that improvements in therapy (which are, typically, incremental rather than dramatic) become difficult or impossible to demonstrate once the first agent for the disorder concerned has been approved. For both of these reasons, a substitute for the fracture end point would be highly desirable.

Future perspective

While the problems outlined above are substantial, they are not, I believe, unsolvable. As has been discussed, they all center around, or flow from, exclusive reliance on the RCT as currently implemented. So far as can be discerned, the ideal solution will have two principal components: a study design that retains as many of the features of an RCT as possible, but with improved generalizability, and a surrogate for the fracture end point. Good solutions to these challenges are not currently available. In the remainder of this perspective I sketch out not what I believe to be the ideal solution, but an illustration of how the desired solution features might be approached.

Currently the most logical candidate for a fracture end point surrogate is some combination of bone mass change and bone remodeling change. Bone mass change, by itself, has previously been shown to be a poor predictor of fracture risk reduction [17], which is a principal reason why fractures are currently a required end point. However, there is a growing consensus that remodeling

reduction, more than mass change, is responsible for much of the fracture risk reduction produced by antiresorptives [18,19]. While remodeling biomarkers exhibit sufficient biological and measurement variability to provide limited value in individual patients, they appear to exhibit better predictive value in groups of individuals than does bone mass [20], and certainly better prediction than bone mass change. The markers that may be best in this regard, and the mode of combining them with bone mass need to be validated for this purpose, but sufficient data already exist in the pharmaceutical industry databases to permit beginning exploration of various options. Substitution of an end point depending upon bone mass and remodeling change could shorten study duration and reduce sample sizes, resulting in substantial cost savings for efficacy trials. Furthermore, by not requiring fractures in a placebo-treated group, such an end point would greatly lower the ethical barriers to continued development of efficacious agents.

The inferential power of the RCT resides not solely in the ability to use experience with random chance to evaluate the difference produced by an experimental treatment, but in its blinding, which equalizes investigative interference in the outcome (i.e., the placebo effect that, as is not always recognized, affects the response of both the treatment and the placebo arms in a controlled trial). There is one other investigational design that not only equalizes that interference across treatment groups, but eliminates it entirely. That is the nonconcurrent cohort study, the design that is, today, the principal basis for the detection and evaluation of harm.

Here is an illustration of how a nonconcurrent cohort design might be used to establish efficacy. Prior to initial introduction, I suggest a new agent would have to show both human safety and efficacy in two or more subhuman species. Then, in what amounts to a variant of postmarketing surveillance, human efficacy would be assessed by comparing cohorts of treated and untreated individuals who are suitably matched (see later). Such initial use of the new agent would have to be controlled in many ways for which no clear precedent or experience exists, such as requiring that the drug, while marketed, be sold at a price competitive with already approved agents (so as to minimize selection bias), and would require programmed, periodic measurements of the desired end points (e.g., bone mass and remodeling). Then, after the fact, contrast groups could be assembled blindly by applying the desired inclusion and exclusion criteria, but matching for various confounding factors, such as disease severity and comorbidity. Finally, the outcome differences, if any, would be evaluated in the usual way.

There are existing large cohorts, such as National Osteoporosis Risk Assessment (NORA) [21] and Study of Osteoporotic Fractures (SOF) [22], that doubtless contain information that would be helpful both in designing such a study and in selecting the cofactors that would need to be incorporated into the inclusion and exclusion criteria.

I do not propose abandoning the RCT entirely. Some judicious combination of RCT designs and larger nonconcurrent cohort studies may constitute an optimal compromise. But both the RCT itself, as it must be executed in the real world, and primary reliance upon its results for treatment guidelines, appear to be less and less acceptable today.

Adopting such an approach will require change from the way of doing things with which we have become familiar and, to a certain extent, comfortable, but failure to find an alternative leaves us stuck in a blind alley that both obstructs investigative progress in the treatment of osteoporosis and escalates its costs.

Executive summary

- Registration of new treatments for osteoporosis currently requires evidence from randomized controlled trials with a fracture end point.
- Once a single efficacious agent is identified, placebo controls become ethnically unacceptable for all subsequent agents.
- Moreover, randomized controlled trials, while establishing the efficacy of a treatment agent, have limited generalizability as a result of resource limitations that constrain which patients are studied.
- Accordingly, treatment guidelines that dependend upon such evidence are of limited applicability; furthermore, they are commonly assembled by individuals who pool the evidence without knowledge of the underlying biology, and, thus, often come to erroneous conclusions.
- Alternative approaches that obviate these problems have been identified and must be evaluated for their suitability.

Bibliography

Papers of special note have been highlighted as either of interest (•) or of considerable interest (••) to readers.

1. Service FJ: Idle thoughts from an addled mind. *Endocr. Prac.* 8, 135–136 (2002).
2. Feinstein AR: Epidemiologic analyses of causation: the unlearned scientific lessons of randomized trials. *J. Clin. Epidemiol.* 42, 481–489 (1989).

•• **Introduction to the issue of limited generalizability of randomized controlled trials.**

3. Dowd R, Recker RR, Heaney RP: Study subjects and ordinary patients. *Osteoporos. Int.* 11, 533–536 (2000).

•• **Study documenting the fact that most patients with osteoporosis are not eligible for inclusion in clinical trials of bone active drugs.**

4. Peto R, Pike MC, Armitage P *et al.*: Design and analysis of randomized clinical trials requiring prolonged observation of each patient. *Br. J. Cancer* 34, 585–612 (1976).
5. Jackson RD, LaCroix AZ, Gass M *et al.*: Calcium plus vitamin D supplementation and the risk of fractures. *N. Engl. J. Med.* 354, 669–683 (2006).
6. Porthouse J, Cockayne S, King C *et al.*: Randomised controlled trial of supplementation with calcium and cholecalciferol (vitamin D3) for prevention of fractures in primary care. *Br. Med. J.* 330, 1003–1005 (2005).
7. Grant AM, Avenell A, Campbell MK *et al.*: For the writing 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 randomized placebo-controlled trial. *Lancet* 365, 1621–1628 (2005).
8. Trivedi DP, Doll R, Khaw KT: Effect of four monthly ORal vitamin D3 (cholecalciferol) supplementation on fractures and mortality in men and women living in the community: randomized double blind controlled trial. *Br. Med. J.* 326, 469–474 (2003).
9. Heaney RP: Design considerations for clinical investigations of osteoporosis. In: *Osteoporosis (3rd Edition).* Marcus R, Feldman D, Nelson D, Rosen C (Eds). Elsevier Inc., CA, USA (2007) (In Press).

•• **Exhaustive description of issues arising in the design and conduct of efficacy trials in osteoporosis.**

10. Peacock M, Liu G, Carey M *et al.*: Effect of calcium on 25OH vitamin D3 dietary supplementation on bone loss at the hip in men and women over the age of 60. *J. Clin. Endocrinol. Metab.* 85, 3011–3019 (2000).
11. Heaney RP: The bone remodeling transient: interpreting interventions involving bone-related nutrients. *Nutr. Rev.* 59, 327–333 (2001).
12. Heaney RP, Bachmann GA: Interpreting studies of nutritional prevention. A perspective using calcium as a model. *J. Women's Health* 14, 990–897 (2005).

•• **Discussion of principal design-related issues relevant to studies of nutrients.**

13. Dawson-Hughes B, Harris SS: Calcium intake influences the association of protein intake with rates of bone loss in elderly men and women. *Am. J. Clin. Nutr.* 75, 773–779 (2002).
14. Heaney RP: Effects of protein on the calcium economy. In: *Nutritional Aspects of Osteoporosis.* Burckhardt P, Dawson-Hughes B, Heaney RP (Eds). Elsevier Inc., Amsterdam (2007) (In Press).
15. Levine RJ: Placebo controls in clinical trials of new therapies for osteoporosis. *J. Bone Miner. Res.* 18, 1154–1159 (2003).

•• **Analysis of the ethics of the use of placebo controls in trials with a fracture end point.**

16. Ellenberg SS: Scientific and ethical issues in the use of placebo and active controls in clinical trials. *J. Bone Miner. Res.* 18, 1121–1124 (2003).
17. Cummings SR, Karpf DB, Harris F *et al.*: Improvement in spine bone density and reduction in risk of vertebral fractures during treatment with antiresorptive drugs. *Am. J. Med.* 112, 281–289 (2002).
18. Heaney RP: Is the paradigm shifting? *Bone* 33, 457–465 (2003).
19. Chapurlat RD, Palermo L, Ramsay P, Cummings SR: Risk of fracture among women who lose bone density during treatment with alendronate. The Fracture Intervention Trial. *Osteoporos. Int.* 15, 842–848 (2005).
20. 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.* 18, 1051–1056 (2003).
21. Siris ES, Miller PD, Barrett-Connor E *et al.*: Identification and fracture outcomes of undiagnosed low bone mineral density in postmenopausal women. *JAMA* 386, 2815–2822 (2001).
22. Cummings SR, Black DM, Nevitt MC, Browner WS, Cauley JA, Genant HK: Appendicular bone density and age predict hip fracture in women. *JAMA* 263, 665–668 (1990).

Affiliation

• *Robert P Heaney, MD*
Creighton University Medical Center
601 N. 30th St., Suite 4841, Omaha,
NE 68131, USA
Tel.: +1 402 280 4029;
Fax: +1 402 280 4751;
rheaney@creighton.edu