A White Paper on behalf of the

Center for Medical Technology and Policy

Understanding Heterogeneity of Treatment Effects in Pragmatic Trials

With an Example of a Large, Simple Trial of a Drug Treatment for Osteoporosis

Jodi B. Segal MD, MPH, Carlos Weiss MD, MHS, Ravi Varadhan, PhD

Heterogeneity of treatment effect in effectiveness trials should be celebrated as informative rather than averaged and ignored.

Pragmatic trials should be designed as explorations of heterogeneity of treatment effect rather than as evaluation of average treatment effect.

Abstract

Individuals differ in their response to therapies. This can be called heterogeneity of treatment effect (HTE). Traditional trials aimed at understanding the efficacy of an intervention seek to answer the question as to whether an intervention works under optimal circumstances in a carefully chosen, treatment-adherent patient population. Variation in outcomes is reduced by excluding people with characteristics that may cause variations in responses to treatment, and sometimes even by analyzing only treatment adherent individuals. These trials typically report a single summary measure of treatment effect, the *average treatment effect*. Pragmatic trials, in contrast, are typically more inclusive and more closely replicate practice in a usual care setting. Our aims in this white paper are three-fold: (1) to characterize HTE, (2) to explore how HTE is particularly prominent in pragmatic trials due to their design, and (3) to explore how this heterogeneity can be useful. To illustrate, we use an example of a suggested design for a pragmatic trial investigating the effectiveness of a drug for the treatment of osteoporosis. Our premise is that trialists should not be aiming to eliminate HTE from trials; they might welcome HTE in pragmatic trials as a source of important data. We recommend that pragmatic trials *always* have a goal of informing the users of the trial about HTE so that multiple stakeholders including patients, clinicians and policy-makers can all benefit from the evidence. We recommend that some hypotheses be specified as the confirmatory hypotheses. Other hypotheses will be specified as being exploratory and provide important information for testing in future studies, although not for decision-making. There should be attention to ascertainment of multiple outcomes (including harms) as heterogeneity in responses may differ by outcome.

Table of Contents

I. Introduction

Individuals differ in their response to therapies. Clinicians know that prescription of a beta-blocker for hypertension control may or may not provide the desired response in blood pressure in an individual patient; the prescription of a serotonin reuptake blocker for an individual with depression may or may or may not relieve the depressive symptoms. In contrast, most individuals treated with a HMG-CoA reductase inhibitor have a reduction in low density lipoproteins, and most individuals vaccinated against the varicella virus avoid chicken pox. Complex interactions between an individual's genes, diet, environment, stressors, concurrent medical conditions, other medications, and behaviors, including adherence to treatment, influence the response to an intervention. While this is well-appreciated clinically, traditional randomized controlled trials, upon which most practice recommendations are based, are designed to minimize these differences between enrolled individuals rather than to learn from them.

Traditional trials aimed at understanding the efficacy of an intervention seek to answer the question as to whether an intervention works under optimal circumstances in a carefully chosen, treatment-adherent patient population. Variation in the population is unwelcome in efficacy trials and is reduced by excluding people with characteristics that may cause variations in responses to treatment, and sometimes even by analyzing only treatment adherent individuals. These trials typically report a single summary measure of treatment effect, the *average treatment effect*. This is a summary of individual treatment effects and obscures differences in how individuals or subgroups of individuals respond to the treatment. Thus, the results of efficacy trials (or "explanatory trials") are often difficult to apply to an individual patient unless that patient is very like those enrolled in the trial in measured and unmeasured ways.

Since the 1960's, trials designed to include patients more representative of a clinical population, in a usual care setting, have been used to answer questions important to decision makers.(1) These pragmatic clinical trials (PCTs) are prospective studies designed specifically with the objective of generating data to inform decisions about an intervention. These studies aim to produce evidence that is applicable to the broad range of patients in usual care settings and conditions. Although clinicians really want to know "Will this intervention work for *this* individual?"; individual prediction is challenging. Pragmatic trials, however, may be particularly well- suited to generating information about subgroup effects that may be more useful to a clinician than an average treatment effect.

A prominent criticism of pragmatic trials is that they are "noisy"; the variability of responses within treatment groups is so great that it precludes the observation of differences in response *between* treatment groups. (2) A null treatment effect in a pragmatic trial may be due to this variability within treatment groups. We argue that this variability is highly informative, and if used appropriately, can identify subgroups of individuals that *can* benefit from the treatment. We define heterogeneity of treatment effect (HTE) as this non-random, explainable variability in the direction and magnitude of the treatment effect. HTE is not noise that needs to be filtered out; it is explanatory and needs to be understood to make appropriate policy and care decisions. HTE is particularly important in studies of effectiveness and studies of safety that are broadly inclusive in their enrollment, such as pragmatic trials.

Goal of this White Paper

Our aims in this white paper are three-fold: (1) to characterize HTE, (2) to explore how HTE is particularly prominent in pragmatic trials due to their design, and (3) to explore how this heterogeneity can be useful. To illustrate, we will use as an example a suggested design for a pragmatic trial investigating the effectiveness of a drug for the treatment of osteoporosis. This example will illustrate how design choices may increase HTE and how several analytic options can use this HTE to generate information relevant to the decision-makers who will want to use the results of this trial. Our premise is that trialists should not be aiming to eliminate HTE from trials; they may welcome HTE in pragmatic trials as a source of important data. In this white paper we will make recommendations about how we can learn from HTE in pragmatic trials.

II. An Introduction to Heterogeneity of Treatment Effects

A. Definitions

HTE is not simply variability in outcomes. Variability is part of any study. *Random* variability is uncorrelated with explanatory variables and can be handled well with basic statistical approaches for bounding uncertainty. We focus, here, on the *non-random* variability in treatment effects that can be attributed to patient, treatment, provider or environmental factors. Therefore, we define HTE as non-random variability in the direction or magnitude of a treatment effect, where the effect is measured using clinical outcomes.

For a formal definition of HTE, let an individual or a targeted subgroup with specific levels of characteristics be denoted by *i*. Let z stand for treatments (3). The potential outcomes are {Y_i(z=1), Y_i(z=2)},assumed to be binary, 0 or 1. The individual treatment effects can be measured using an absolute or relative risk model. For the absolute risk model the individual treatment effect, θ_i = Prob(Y_i(2)=1) – Prob(Y_i(1)=1). For the relative risk model θ_i = log [Prob(Y_i(2)=1)/Prob(Y_i(1)=1)]. Variability of treatment effect occurs if the variance of $\theta_i > 0$. More explicitly, HTE is present if the variance of individual treatment effect is non-random. HTE comprised of individual treatment effects in different directions, i.e. benefit or harm, is sometimes called a qualitative treatment interaction, whereas differences in the magnitude of treatment effect are called quantitative interactions.(4)

HTE may be present on the absolute or relative risk scales of treatment effect, or both. Treatment effect cannot be homogeneous in *both* scales, unless the baseline risk of the outcome is constant. (Figure 1) Suppose that the treatment effect is constant on the absolute risk scale, that is Prob(Y_i(2)=1) – Prob(Y_i(1)=1) = a, for all individuals *i*, then the individual treatment effect on the relative risk scale is equal to 1 + (a/ Prob(Y_i(1)=1)) and varies with the individual's baseline risk. Conversely, suppose that the individual treatment effect is constant on the relative risk scale, that is Prob(Y_i(2)=1)/ Prob(Y_i(1)=1) = b, for all individuals *i*, then the individual treatment effect on the absolute risk scale is equal to (1 – b) * Prob(Y_i(1)=1), which also varies with the individual's baseline risk.

Figure 1. Heterogeneity of Treatment Effect Is Present in Absolute And/Or Relative Scales

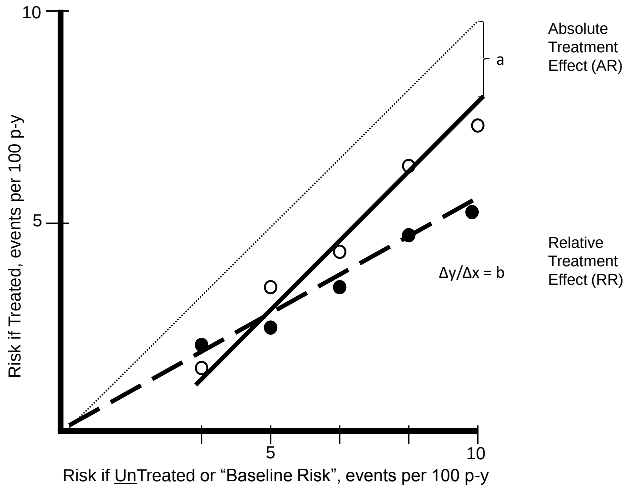


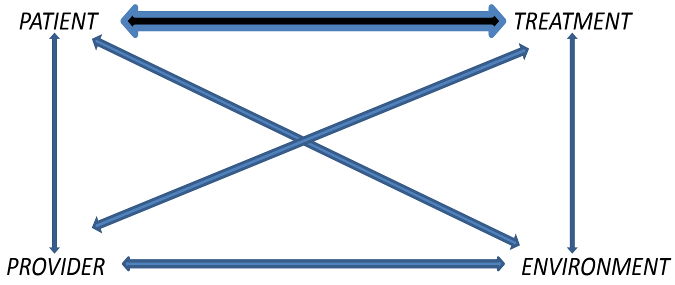
Figure 1 Legend: Absolute baseline risk of the primary outcome is on the x axis and is the source of HTE in this example. Risk if treated is on the y axis and the dotted line indicates no treatment effect. Treatment effects according to quintiles of baseline risk are presented for 2 studies (closed, open circles). Treatment effect can be calculated according to absolute risk (solid line) or a relative risk (dashed line) or both. Corresponding effect models are that the absolute risk reduction equals `a' and relative risk equals `b'. Treatment effect cannot be homogeneous on both scales. Heterogeneity may be apparent in one effect model and not in the other effect

model. The study represented by closed circles would not show HTE on a relative risk scale, but would show HTE related to baseline risk on an absolute risk scale; the study represented by open circles would not appear to have HTE on an absolute risk scale, but would on a relative risk scale.

B. Sources of HTE Arising from Patient, Treatment, Provider and Environment Sources

We distinguish between sources of HTE because information about each source may be informative in interpreting study results.(Figure 2). Patient characteristics are typically the main source of HTE in most situations because the unit of treatment is commonly the individual patient. Pragmatic trials sometimes use cluster-randomization in which case the clinic or health plan may be an even more important source of variability.

Figure 2. Sources of Heterogeneity of Treatment Effect.



An ntroduction

Figure 2 Legend: Heterogeneity refers to non-random variability and can come from several types of sources. An emphasis is placed (thicker arrow) on patient and treatment sources of heterogeneity of treatment effect because these are fundamental. Arrows reflect interactions between these sources of heterogeneity.

Partitioning the sources of HTE is essential to understanding the results of a trial and appropriately using its results. Consider the example of a health care system that conducts a trial of a hand disinfectant and finds markedly different infection rates among its hospitals. It is essential to know whether these differences in outcomes are the results of patient- level differences. If the patient-level differences explain the different infection rates convincingly, then interventions at the level of the provider or hospital may not be appropriate. Stated differently, not understanding patient-level sources of HTE could result in a faulty intervention aimed at changing the health care system.

We consider patient-level sources of HTE to be variables that are measured on individual participants, such that the same measurement could not be taken from more than one person simultaneously. These are quantified for an individual, for example, serum cholesterol concentration or health-related quality of life. In contrast, other sources of HTE are not necessarily measured separately on each study participant. These include characteristics of the treatment, provider and environment such as housing type, insurance plan, or region; even though this information may be gathered from an individual participant. We propose that patient-level sources of heterogeneity may be approached according to four patient-treatment relationships: baseline risk, competing risks, treatment responsiveness and treatment harm. An example for each is given in the Glossary. We describe these here in detail as we propose that these are major sources of HTE within pragmatic trials and highly informative.

Baseline risk is the risk, without treatment, of experiencing the primary outcome the treatment intends to prevent. This is usually predicted at baseline on the basis of previously validated equations that draw on patient level characteristics, but can also be observed during the course of a study in an untreated control arm, hence the synonym 'control event rate'. Baseline risk has been extensively explored as a source of HTE among patients(5) and studies.(6) Since the treatment effect is commonly defined as either a ratio of or the difference between baseline risk and treated risk, it is correlated with baseline risk, and this correlation should be considered in analysis.(6-8)

Competing risks are the risks, either with or without treatment, of experiencing any outcome that renders impossible the occurrence of primary outcome that the treatment intends to prevent. These are called classical competing risks. In addition, any outcome that alters the meaning of the primary outcome may be a semi-competing risk.(9) An example of a semi-competing event for the outcome of hip fracture is a stroke causing loss of the ability to walk. Like baseline risk, competing risks may be predicted at baseline on the basis of previously validated equations that draw on patient-level characteristics and can also be observed during the course of a study. Trial reports often provide findings using relative risk reductions that ignore competing risks, although careful accounting of all outcomes in all treatment arms often reasonably diminishes the threat of competing risks to the inferences drawn from the trial.

Treatment harms experienced by patients can take several forms that are not mutually exclusive. We define *primary treatment harms* as those that are manifest on the primary outcome of the study, that is when the treatment causes the event that it was supposed to prevent (e.g. antihypertensive medication can prevent or cause heart attacks; carotid endarterectomy can prevent or cause stroke). In this manner, *primary treatment harms* can comprise qualitative HTE and the variables that predict risk of treatment harm may sometimes be correlated with baseline risk of the primary outcome. However, there are often other types of treatment harm. *Competing treatment harms* can take the form of increasing the risk of a competing event (e.g., increasing the risk of lethal stroke while decreasing the risk of heart attack). In other words, not all competing risks are treatment harms, but sometimes the treatment increases competing risks by increasing treatment harms which are competing events. Finally, *non-competing treatment harms* can take the form of other harmful effects that are neither the primary outcome nor competing events, such as hallucinations.

Treatment responsiveness is another factor that modifies treatment effect and is not captured by baseline risk, treatment harms or competing risks. Treatment responsiveness may be exemplified by genetic differences in drug metabolism, leading to marked differences in clinical effectiveness, as has been suggested for the drug clopidogrel.(10) We emphasize that for treatment responsiveness to be considered a source of HTE, according to this framework, effectiveness must be measured with clinical outcomes. In contrast, *surrogate* treatment responsiveness refers to the magnitude of the impact of treatment on a well-established biomarker that is targeted by treatment (e.g., reduction in tumor size). Unlike treatment effect, which is based on a clinical outcome, surrogate treatment responsiveness refers to an intermediate outcome that can be used to judge whether a person or subgroup is likely to go on to experience the desired treatment effect.

Beyond these four patient-treatment interfaces, the treatment itself is also a basic source of HTE. We set aside study designs that manipulate treatment to characterize different levels of a treatment effect, such as variable-dosing studies, or designs that deliberately give different treatments to different participants.(11,12) We consider unplanned differences in the way treatment is received to be highly relevant to HTE and particularly relevant to pragmatic trials. (9)Unplanned variability in treatment can arise for many reasons, including inadequate standardization across or within study sites (investigator adherence) and reported or unreported failure of study participants to follow protocols (participant adherence). This non-random variability caused by unplanned differences in treatment implementation and adherence can detected and addressed through quality control efforts during conduct of the trial.

III. Pragmatic Trials and Heterogeneity of Treatment Effect

A. Pragmatic Clinical Trials

With the above as background, we now describe how HTE is particularly prominent in pragmatic trials, but not unwelcome. We begin with a brief review of key features of pragmatic trials.

The influential early paper about pragmatic trials was published in 1967. (1)In this paper, Schwartz and Lellouch suggested that the design of a trial needs to be directed by the goal of the investigation. Is the goal to acquire information about the true effects of a treatment (i.e. to verify a biological hypothesis) or is the goal to gather information needed to make a decision about a treatment? They described the former goal as requiring an “explanatory” trial and the latter as requiring a “pragmatic trial.” Schwartz and Lellouch described in detail the appropriate selection of outcomes to be evaluated in trials, contrasting explanatory and pragmatic trials. They used as an example the outcome of “returning to work”. They acknowledged that this is an important outcome to patients and appropriate for evaluation in a pragmatic trial, but as it conveys little biological information, it may not be a relevant outcome in an explanatory trial. Schwartz and Lellouch described how closely linked the analytic decisions are to the selection of individuals for inclusion in the trial. With an explanatory approach, a strict patient selection criterion may be used in order to render the population homogenous, which they suggested reduces the withdrawal rate. However, in a pragmatic trial, a heterogeneous population with more withdrawals is acceptable. Patients should not be turned away from a trial for reasons that would not preclude use of the intervention in usual practice. “The trial must represent as far as possible the population to which the results are to be extrapolated.”

Dr. Alvan Feinstein at Yale University appreciated the limitations of traditional RCTs. In a Perspective piece in the Annals of Internal Medicine in 1983, Feinstein wrote on challenges in trial design that stem from conflicting goals of trials. (13) He said that individuals who want answers to pragmatic questions in clinical management want trials that incorporate heterogeneity and ambiguity, and other “messy” aspects of clinical practice. He offered many suggestions to improve the usefulness of clinical trials for decision making including patient populations that include a separate heterogeneous group as well as a “pure” group, a pragmatic treatment arm which could be added if the chosen comparator treatments seem clinically unsuitable, a “double-observer” procedure to allow flexible dosing; and efforts to “harden” the softer patient-relevant outcomes (such as functional status and quality of life).

By the mid-1990’s, the term “mega-trial” was used to describe large, simple randomized trials analyzed on an intent- to-treat basis.(14)(12) Early proponents of these “large simple” trials were Peto and colleagues at the University of Oxford. They asserted that there are some underlying assumptions when using this trial design.(15) The first assumption is that the real differences between two treatments in some important outcome will *probably not be large,* but that even a moderate difference in an important outcome may be worthwhile [to detect]. The second is that if there is, for some readily identifiable category of patients, a moderate difference between two treatments in their effects on some specific outcome, then this difference might be larger or smaller in other readily identifiable categories of patient, but it is *unlikely to be reversed.* Detractors argue that the between-subject variation, within each treatment group in large simple trials, makes the results of these trials difficult to apply to an individual patient. (16)

In the next decade, additional terms were used to describe more pragmatic trial designs including “naturalistic trials” and “effectiveness trials”. Efficacy studies are closest to what Schwartz and Lellouch described as explanatory trials. These are studies that aim to investigate whether an intervention works under *optimal* circumstances, or in other words “can it work?” Effectiveness studies are closer in their goals to those of pragmatic studies. They aim to evaluate whether an intervention works under usual circumstances, or in other words “does it work?” Others have since carefully articulated the differences between efficacy studies and effectiveness studies, such as is shown in this table from Bombardier and Maetzel.(17)

Table 1 Efficacy versus effectiveness studies

	Efficacy studies	*Effectiveness studies*
Objective	Does it work under optimal circumstances?	Does it work under usual circumstances?
Motivation	Regulatory approval	Formulary approval
Intervention	Fixed regimen / forced titration	Flexible regimen
Comparator	Placebo	Usual care

	Arbitrarily chosen comparator	Least expensive / most efficacious
Design	Randomized controlled trial - strict control	Randomized controlled trial - or open label - minimum control
Subjects	Selected or "eligible" subjects	Any subjects
	High compliance	Low compliance
Outcomes	Condition-specific	Comprehensive (for example, quality of life utilities)
	Strong link to mechanism of action	Weak link to mechanism of action
	Short-term horizon	Short and long term horizon
Analysis	Protocol adherers	Intent to treat

A recent advance in differentiating pragmatic and explanatory trials was the exercise by Thorpe and colleagues who devised a graphical method by which an investigator, or reader, can evaluate where a study lies on the explanatory – pragmatic continuum.(18) This grew out of discussion among investigators involved in the PRACTiHC project, a Canadian and European Union initiative to promote pragmatic trials in low and middle-income countries. They called this the Pragmatic-Explanatory Continuum Indicator Summary (PRECIS).The key domains which distinguish explanatory and pragmatic trials are shown in Table 2.

Table 2. Domains for the PRECIS Graphic

1. The eligibility criteria for trial participants.
2. The flexibility with which the experimental intervention is applied.
3. The degree of practitioner expertise in applying and monitoring the experimental intervention.
4. The flexibility with which the comparison intervention is applied.
5. The degree of practitioner expertise in applying and monitoring the comparison intervention.
6. The intensity of follow-up of trial participants.
7. The nature of the trial's primary outcome.
8. The intensity of measuring participants' compliance with the prescribed interventions, and whether compliance-improving strategies are employed.
9. The intensity of measuring practitioners' adherence to the study protocol, and whether adherence-improving strategies are employed.

10. The specification and scope of the analysis of the primary outcome.

Among participants in the PRACTiHC project there was dissension about definitions. Karinicolas and his colleagues challenged the prevailing explanatory-pragmatic framework.(19) They maintained that this framework "ignores the varying perspectives of those using RCT results" to make decisions. They thought that most trialists have severed the link between the goals of the trial – answering questions relevant to decision-makers – and the design of the trial. They feared that the design of pragmatic trials, as discussed above, exclusively answers questions from a public health perspective, and provides little information that is relevant to clinicians caring for individual patients. They felt that a practical trial can legitimately seek to enroll highly compliant patients managed by skilled specialists if this is the setting in which the intervention is intended for use. In other words, this may be the usual care setting for some interventions (perhaps cancer therapies). The "pragmatists", lead by Oxman and colleagues, countered that Karinicolas' use of the term "practical" to describe trials which may be performed in carefully selected patients in optimal clinical settings distorts the idea of a pragmatic trial.(20) They feared that these trials would look much like explanatory trials, except perhaps for the choice of outcomes. They conceded, however, that there are sometimes reasons for *not* having broad inclusion criteria, but cautioned that trial results are *always* average results and there is never information specific to an individual patient outside of an n-of-1 trial. They said:

> "Although explanatory trials may help to understand mechanisms of treatment effects, they are primarily designed to test whether interventions have hypothesized effects under optimal circumstances, not necessarily to investigate 'possible mechanisms of effect.' The strength of explanatory trials is that a 'negative' result can directly inform practice, because an intervention that does not work under optimal circumstances is unlikely to work under usual circumstances. The weakness of explanatory trials is that 'positive' results do not directly inform practice, although they may directly inform practice under a narrow set of optimal circumstances and they can inform decisions about future research. Pragmatic trials, on the other hand have the opposite strengths and weaknesses. The weakness of pragmatic trials is that with 'negative' results it is unclear whether the intervention is 'worthless' or whether it might, in fact, be worthwhile under some (more optimal) circumstances or for a subgroup of patients. The strength of a pragmatic trial is that 'positive' results can directly inform decisions under the 'usual' conditions for which the trial was intended to be applicable." (21)

B. *Apparent* HTE in Pragmatic Trials

In section C, below, we describe the design choices in pragmatic trials that amplify HTE. However, we want to caution in this section that there is some variability which we do not consider HTE but that might be considered just apparent HTE. We differentiate actual HTE from apparent HTE based on how informative the variability is to the users of the trial results. To illustrate, we do not consider systematic measurement effort to be a source of actual HTE. Although measurement error contributes to variability in the reported treatment effect, there is little to be gained from exploring this contribution – measurement error should be identified and eliminated. An example may be a systematic error in using a quality of life measure. Perhaps all individuals over age 70 years in the trial were queried using an outdated data collection form. These forms are systematically graded two points lower than the contemporary form – hence, there is a systematic likelihood that those older than 70 will be described as having a lower quality of life. This is not informative – this needs to be identified and corrected. Similarly, the measurement error might be made when measuring baseline characteristics of the participants which are treatment effect modifiers. These errors should be rectified before inferences are made based on the observed treatment effects.

C. Design Choices in Pragmatic Trials Contribute to Heterogeneity of Treatment Effect

We explore here how the choices made in the design and analyses of pragmatic trials amplify HTE and how this can be highly informative for decision-makers using the results of these trials. We use key elements from the framework of Bombardier and Maetzel to organize our presentation of sources of HTE in pragmatic trials. (17)

Objective: How does the objective of a pragmatic trial introduce HTE? An effectiveness study (or a pragmatic trial) seeks to address whether an intervention will work under usual circumstances. This is typically understood as an intervention working as applied by diverse clinicians caring for diverse patients in diverse settings, although the context is specific to the intervention. Just from knowing the objective of the study, it is apparent that when “usual circumstances” is operationalized at the time of the design of the trial, this will be a design that permits a lot of choices that will contribute to HTE.

Motivation: Similarly, how does the motivation to conduct a pragmatic trial introduce HTE? The motivation of effectiveness studies has been described as for “formulary approval” rather than regulatory approval. In other words, the trial may be conducted to learn if the drug should be made widely available for use. This narrow definition of motivation really just illustrates that the results of the trial may be used for deciding about unrestricted access to the intervention for clinicians and their patients and the "usefulness of the intervention". Given that the users may then be very diverse, heterogeneity is introduced even by the motivation for conducting the study.

Intervention: Bombardier and Maetzel, as well as the PRACTiHC project participants, suggest that in effectiveness studies, the primary intervention of interest can be applied with a flexible regimen. This seems somewhat overstated as even in pragmatic trials there is typically a starting dose and sometimes even a prescribed titration regimen. However, by design, the clinicians who administer the treatment are not tightly regulated in the application of the intervention. By design, there is limited contact between the clinicians and the study investigators, with minimal reinforcement about correct application of the interventional treatment. This has the potential for introducing important heterogeneity as the exposures of the patients, even with the same treatment group, may be different These differences may be driven by patient-level factors (like adherence to the intervention), by clinician-level factors (like use of concomitant therapies), or by environmental or social factors that influence access to the intervention such as availability of transportation to the site to receive the intervention.

Some treatments, more than others, are subject to variable implementation that will contribute to HTE. The variation in implementation is likely to increase as the complexity of treatment increases and becomes more dependent on patient participation or treatment operators. For example, long-term anticoagulation therapy currently uses warfarin, which requires frequent monitoring of blood values and dosing adjustments.(22) This therapy is more operator-dependent than a pill that requires no monitoring and no dosage adjustments. Invasive surgical procedures are an obvious example of complex treatments that are operator-dependent and may therefore introduce HTE.

Comparator: In a pragmatic trial, the choice of comparator may be extremely important to the treatment effect reported. In pragmatic trials that are aiming to answer questions of comparative effectiveness of interventions, the comparator will typically be an active therapy or will be "usual care". Pragmatic trials that compare two active treatments may require extra attention to *treatment harms* as a source of HTE. In active comparison trials, either treatment can cause harm (a qualitative interaction), whereas in placebo-controlled trials only the active treatment arm can cause harm. The presence of harms will affect the conclusions about overall treatment effectiveness.

Additionally, when there is an active comparator, a greater number of participants are needed to demonstrate an importantly different average treatment effect. Presumably each treatment arm has been demonstrated to be superior to placebo and therefore the treatments may be more alike in their treatment effects than either one is like the placebo, thus requiring a greater sample size to demonstrate a difference. This larger sample size, however, allows potentially more exploration of HTE as there may be a greater number of participants in the subgroups of interest.

However, pragmatic trials need not necessarily have active comparators. Particularly relevant to a discussion of heterogeneity in pragmatic trials is the heterogeneity that arises from using a comparator group that includes subjects receiving a mixture of different regimens. This is frequently the case in studies of an active intervention compared to treatments offered in "usual care". One example is a trial comparing lipid outcomes for patients randomized to a specialty lipid clinic compared to outcomes for patients receiving usual care from their cardiologist.(23) One anticipates that the group of patients receiving usual cardiology care had a diverse set of interventions administered depending on the strengths and interests of the cardiology practice and the cardiologist. As would be expected, the variability in the outcomes (change in low density lipoproteins (LDL) and total cholesterol) were greater in the usual care group as indicated by the large standard deviations for the post-intervention LDL, allowing for differences in the sizes of the treatment groups. Another example is a cluster-randomized trial of an education intervention for weight loss in children. Some clinics were given an intensive educational program to administer to eligible children, while other clinics were randomized to continue their usual care of these children. One anticipates that the usual care clinics administered vastly different interventions to the children.(24) In neither study was there explicit description of the interventions provided in the usual care group; we suspect that this is a frequent reporting deficiency in studies that have a usual care arm. Finally, we illustrate with a hypothetical comparison of a new treatment for benign prostatic hypertrophy that is compared to *usual care* of this condition. The usual care group will include some men using an alpha-antagonist (terazosin), some using an 5-alpha reductase inhibitor (finasteride) and some managing without drug therapy. The active treatment has a reasonable chance of being more effective than one or more of the existing treatments and less effective than another, making an average treatment effect uninterpretable. This null treatment effect will have little value to decision makers. In summary, the choice of comparator will markedly affect the inferences drawn from a pragmatic trial and trialists should explicitly address this at time of the design of the trial and in their plans for analysis.

Design: How do design choices in a pragmatic trial introduce HTE? Many have been described above. Design choices establish the participants, the providers, the treatments (including comparators) and the environment of the trial. Pragmatic trials are typically designed with less control over the individuals administering the treatments, which are often clinicians in practice. The diversity in application of the interventions and comparisons that is introduced by allowing this flexibility contributes to HTE. We highlight, in this section, the impact of the providers and the environment chosen for the trial.

A hallmark of pragmatic trials is that they are conducted in a usual care setting; this may be a primary care clinic, or an obstetrical ward, or an emergency department. It surprises no one that practice settings differ substantially and that this can affect treatment effects in a trial. To illustrate, a respected pragmatic trial is the ALLHAT trial of hypertension treatment. ALLHAT enrolled more than 24,000 individuals from 625 centers in the U.S., including Puerto Rico and the U.S. Virgin Islands, and in Canada.(25) It is inconceivable that clinicians' approaches to managing hypertension, diabetes, and cerebrovascular disease are identical for Canadian-trained physicians treating universally insured-patients, and nurse practitioners practicing in rural, underserved regions of the United States where the patients are uninsured or underinsured. Although patients were randomized to an initial treatment of one of four drugs and a titration scheme, practitioners had a great deal of flexibility with concomitant therapies (such as lipid-lowering drugs, aspirin, and use of invasive therapies). Although randomization will have minimized the differences between groups, it is likely that the within group differences in treatments were large. ALLHAT convincingly demonstrated a best treatment response for the chlorthalidone-treated patients, despite the heterogeneity induced by the design choices. It is conceivable that had this large trial not demonstrated a difference, that exploration of the heterogeneity of effects across sites would have been essential. Indeed, the ALLHAT investigators have further explored subgroups of their population to better understand who benefits and who is harmed by the intervention, although many of these analyses were done post-hoc.(26)

Pragmatic trials frequently allow the addition of other interventions as needed for a clinical response during the course of a trial. Although in ALLHAT, this scheme was carefully dictated, in other pragmatic trials there is a great deal more flexibility, with the choices left to the judgment of the treating clinician. This is a hallmark of pragmatic trials. The introduction of other therapies which may interact with the primary therapy under evaluation contributes greatly to HTE in pragmatic trials.

Subjects: As pragmatic trials seek to replicate conditions of usual care they are often more inclusive in their enrollment criteria than explanatory trials. Pragmatic trials often aim to have the most liberal inclusion criteria possible that still protect the safety of the enrolled participants. The Clinical

Antipsychotic Trials of Intervention Effectiveness (CATIE) Project study of schizophrenia is a good example of liberal inclusion criteria. The investigators specified that the participants had to be between 18 and 65 years, with schizophrenia, in need of treatment with an oral medication, and able to consent to participate.(27) The few exclusion criteria were largely for safety or for documented treatment-resistant disease. A less pragmatic trial might have restricted the duration of disease, intensity of symptoms (disease spectrum), number of previous hospitalizations, number of failed therapies, concurrent medical illnesses, or other criteria. As can be expected, when the inclusion and exclusion criteria for study participants are not rigid, there is a more diverse group of participants whose heterogeneous characteristics may influence treatment effect in several ways. The characteristics of the participants, before randomization, can act as treatment effect moderators. There is a broad literature about the baseline risk of the primary outcome being a treatment effect moderator.(28-33) The baseline differences in risk of the outcome could impact the results if the groups are not balanced. In sufficiently large trials, the randomization process will balance the risk of the outcomes.

Correspondingly, a treatment effect mediator is a variable that is observed *after* randomization and is presumably on the causal pathway between the treatment and the outcome. Medication adherence is a treatment effect mediator – it is known only after treatment. Given that pragmatic trials are less directive about adherence, this is often a powerful effect mediator in pragmatic trials. Another example of a mediator may be blood pressure response to therapy in a trial of a drug intervention to prevent stroke. Blood pressure response may be highly variable in a trial based on genetic differences in individuals; the blood pressure variability may mediate the effect of the intervention on stroke outcomes. (34)

Finally, patient conditions that are known to have variability of expression can be sources of HTE. This is often the case with chronic diseases, which generally result from multiple risk factors and have variable incidence and severity. As Glasziou and Irwing said, “some diseases are more than one disease process, each of which may respond differently to the treatment.”(35) In addition, a majority of older adults have multiple chronic conditions.(36) To the extent that major chronic diseases such as diabetes, hypertension, congestive heart failure and chronic renal disease can influence each other, then the pattern of diseases and clinical conditions presenting in an individual, and across subpopulations, can be relevant to understanding HTE.

Outcomes: The outcomes selected for study in a pragmatic trial are often those that are patient-relevant but which may have a weak link to mechanism of action. An example is a quality of life measure – while this is a highly relevant outcome to a patient, the link between this measure and the pathophysiological mechanism of the treatment may be tenuous. Measures that are highly patient-relevant are often *patient-reported.* Patient reported outcomes fall into several domains, as described by Wu, et al. (37) These may be health-related quality of life measures, symptom measures, ratings of the quality of care or satisfaction with care, description of use of a medication or device, or description of participation in health behaviors. Others have categorized the *tools* used to measure patient-reported outcomes. The tools may be generic (useful across conditions), disease-specific, domain specific (such as pain only), or preference based (individualized to what the patient thinks is most important), or utility-based as used in economic assessments.(38) We would argue that the inclusion of patient-reported outcomes should be a goal in any study that aims to be maximally patient relevant. Wu and colleagues suggest that these outcomes are often not included in trials due to competing demands placed on research sites and relative scarcity of resources.

We know of no studies that have empirically evaluated the heterogeneity within patient-reported outcomes that is separate from measurement error due to use of faulty tools. A recent international consensus panel came to agreement on the taxonomy, terminology, and definitions of measurement properties that are preferred for health-related patient reported outcomes with the goal of working towards better measurement properties.(39) There is, however, a body of literature to support that characteristics of the patients before intervention can affect the report of outcomes. One example is that patients' literacy affects how they report the self-management support received to manage their diabetes.(40) Given this, we suggest that there is likely to be great heterogeneity in patient-reported outcomes that is driven by baseline heterogeneity in the enrolled population. Exploring this in pragmatic trials may be an important methodological contribution.

How does the choice of outcomes contribute to HTE in a pragmatic trial? We *do not* think that these outcomes are measured with less precision than biological outcomes. If they are, this is a study flaw and quality control processes are needed. We *do not t*hink that these measures inherently have more heterogeneity than biological measures. We suggest, however, that because these outcomes may be more "distal" to the treatment than biological outcomes (i.e. intermediate outcomes, surrogate outcomes), there is the potential for greater heterogeneity because of variability in the causal mechanisms between the intermediate outcomes and these patient-relevant outcomes. The pathway of interest, that which is affected by the intervention, may be just one of many pathways to the outcome, and there may be subgroups of individuals for which the targeted pathway might be more importantly affected by treatment than other subgroups. To illustrate, investigators who conducted a study of rosuvastatin to prevent cardiac events and deaths in older people with heart failure found good surrogate treatment responsiveness (lowering of LDL) but no significant benefit.(41) It may be that in a diverse group of individuals, there is unevenness in the relationship between lower LDL and cardiac events. While for a homogenous group of individuals with controlled hypertension and regular aspirin use, this linkage may exist. In a more diverse population of patients, this linkage may be weaker.

Using biomarkers as surrogates for treatment responsiveness is based on the assumption that these markers are modified by treatment and are in the causal pathway between treatment and outcome. It can be argued that these *can* be legitimately used as outcomes when there is sufficient scientific understanding of mechanisms and pathways responsible for mediating the treatment effect on clinical outcomes, as is often the case in effectiveness trials. The relationship between the biomarker and the outcomes will have been established in earlier studies. However, a given biomarker may reflect only one of several pathways affected by treatment, so several biomarkers may be required to demonstrate surrogate treatment responsiveness particularly when the population is diverse as is the case in pragmatic trials.

We suggest that the issue of outcomes which are competing risks for the primary outcome may also be particularly prominent in pragmatic trials. Because pragmatic trials often use patient-relevant, clinical outcomes and therefore may be longer than many efficacy trials with surrogate outcomes, the opportunity for competing outcomes to affect results is heightened. (9)

Analysis: Pragmatic trials are typically analyzed with intent-to-treat methods. Participants contribute outcomes to the group to which they were randomized, regardless of treatment received or concomitant treatments received. The treatment experience of any individual subject may be very different even within a group defined by the randomization process. These differences within groups, however, are the basis for exploration of treatment effects in subgroups. Pragmatic trials are an ideal setting in which to apply stringent and pre-specified methods for conducting subgroup analyses. What follows are our recommendations for how analyses of pragmatic trials can take advantage of the HTE introduced by the choices above.

We suggest that pragmatic trials should be regularly used for both *confirmatory* and *exploratory* analyses of subgroup effects to take advantage of the rich data acquired in these trials. We suggest that analyses can be considered *confirmatory* if all of the below characteristics are satisfied by the trial.

a) There is biological plausibility and prior evidence for suspecting HTE

b) There is pre-specification of (a small number of) HTE hypotheses in a manner consistent with prior evidence (including anticipated direction of subgroup effects)

c) Definition of subgroups is pre-specified including how continuous variables would be categorized

d) Outcomes are clearly specified including how multiple endpoints will be addressed.

e) The study power is adequate for examination of HTE

f) There is high quality measurement of outcome, exposure, and prognostic variables

g) Biases due to measurement error, confounding, and missing information are minimized

h) The analytic protocol includes prespecification of important data modeling methods including methods for covariate adjustment, and the handling of censored and missing data.

An HTE analysis may be placed somewhere along the exploratory-confirmatory continuum according to the extent to which it satisfies the above requirements. A trial that cannot adequately satisfy many or all of the above requirements is exploratory. It should be noted that within a study, some HTE analyses can be confirmatory and others can be exploratory. In either a confirmatory or exploratory analysis, there should be tests for interaction and then subgroup analyses if appropriate.

The appropriate technique for assessing confirmatory HTE, before subgroup analysis is performed, is a statistical test for interaction that detects differences in treatment effects in subgroups. If the interaction test is positive, subgroup analyses should be done to report stratified treatment effects along with their confidence intervals. Treatment effects are estimated separately in the subgroups for which the interaction test was significant. Even if the test is negative, treatment effect heterogeneity might still be a concern because the power of a test of interaction is lower than for an overall treatment effect of the same magnitude. Therefore, treatment effects may still be estimated in each subgroup and reported along with 95% confidence intervals in a forest plot.

There are two types of subgroups that may be used: (a) subgroups based on treatment effect moderators, e.g. sex, age, genetic polymorphism, ventricular ejection fraction, disease severity, etc.; and (b) subgroups based on *patient/treatment* characteristics, such as baseline risk score. For the treatment effect moderators, standard interaction testing methods apply. However, if HTE is expected to be a function of a baseline risk, two approaches are available. Individual baseline risk can be estimated if a valid multivariate risk model for baseline risk of the primary outcome is available and if the predictors need by the model are measured in the study.(5) Subgroups are defined by using cut-offs (e.g. median, tertiles, or quartiles of risk) to divide the continuous risk score into mutually exclusive categories. Care needs to be taken to ensure that the model provides a good fit to the observed baseline risk in the control arm of the RCT. Alternatively, a multivariate interaction test was developed by Follman, wherein one estimates and tests the baseline risk as a treatment effect moderator in one step.(42) This approach is likely to be more generally useful than the first approach in that it does not depend on a pre-existing model of baseline risk and on all the predictors in that model being measured in the current RCT, although it does not provide explicit estimates of individual baseline risk.

In contrast to confirmatory HTE analysis, the primary goal of *exploratory HTE* analysis is to generate promising hypotheses for further study. Compared to confirmatory HTE analyses, exploratory analyses enjoy more flexibility for detecting baseline characteristics that are likely sources of HTE. Exploratory analyses can also use more sophisticated analytic approaches to account for practical data-related challenges including missing data, imperfect adherence to treatment, and loss to follow-up. In many cases, valid analytic methods to overcome these challenges are sufficiently challenging that it is difficult to pre-specify or to rigidly adhere to the pre-specified analytic plan (required for confirmatory analyses). While multiplicity adjustments may be considered here for reporting p-values that more accurately reflect the statistical significance of the ad-hoc finding, merely adjusting a p-value does not inherently ensure the scientific validity of the findings. In addition, it is not trivial to know how many hypotheses were tested in exploratory analyses, and hence to know how to adjust for multiplicity. In many cases, limitations in study design and poor measurement quality vis-à-vis the specific scientific question may render an HTE analysis to be exploratory. Therefore, the results from exploratory analyses, irrespective of biological plausibility or statistical significance, should be regarded as preliminary until they are replicated. We recommend that exploratory HTE analyses not be mentioned at all in the main body of a research article. Even if the authors state that their analyses are post-hoc and exploratory, there is a high likelihood that the results will generate much controversy. Therefore, the results of exploratory HTE should be clearly labeled as exploratory and reported only as supplemental material.

Interesting tools are being developed by which to explore differences between enrolled trial participants before the start of treatment. One such tool is the risk predictor developed by Kaplan, et al. which aims to quantify an individual's "Potential for Benefit" from a treatment. The authors suggest that the use of a composite scale, such as that they developed, may identify individuals with a greater likelihood of benefit and allow for *a priori* specification of important subgroups for analysis. (43) This model that they describe, however, does not allow for interactions between the baseline characteristics as measured with their scale and the treatment. We anticipate that other tools might incorporate baseline characteristics as treatment effect mediators.

Finally, Bayesian analytic approaches may be useful. An important feature of Bayesian approaches is that they allow estimation of subgroup treatment effects by combining the observed subgroup effect and the overall treatment effect using weights that reflect the a priori view of the degree and direction of heterogeneity. We find the following description to be helpful:

> "Many experienced clinical trial statisticians recommend performing separate analysis of subsets only if a statistically significant treatment-by-subset interaction is found. When important qualitative interactions are unlikely, this advice generally leads to correct results because the standard interaction tests and qualitative interaction tests (Gail & Simon 1985) have low power and will rarely lead to rejection of hypothesis of no interaction. Nevertheless, this analytic strategy is inadequate because it is likely to miss important treatment differences when they do exist. Frequentist analysis is very satisfactory for many clinical trial problems, but situations in which the frequentist methods force the statistician into a dichotomous choice based on a preliminary test of inadequate power are situations where frequentist methods may lead to the wrong answer and where Bayesian methods can be of considerable value. The subset analysis problem is one such situation..." (44)

IV. Clinical Background–Osteoporosis Treatment

We will later use, as an example, the design of a pragmatic trial of an osteoporosis treatment. Therefore we provide here background information about osteoporosis, its treatments, and the questions still requiring answers to provide context for that discussion.

Osteoporosis is a systematic skeletal disease characterized by low bone mass and microarchitectural deterioration of bone tissue with a consequent increase in bone fragility and susceptibility to fracture.(45) Low bone mineral density, measured with dual-energy X-ray absorptiometry (DXA), is the established reference standard for the diagnosis of osteoporosis. A diagnosis of osteoporosis is made in an individual, in a specific region of bone, if the DXA measurement is 2.5 standard deviations below the average for a young, female population. This cutoff was selected because with each standard deviation decrease in bone density the risk of fracture increases approximately 1.6 fold, and 2.5 standard deviations below normal is considered a clinically important increase in risk.(46) This definition relies heavily on the choice of the reference standard, which has typically been defined as 20-29 year old white women. However as nicely stated in this paper by Kanis et al, "There is a growing awareness that treatments should be targeted on the basis of fracture risk rather than solely on the information provided by a BMD test. … The measurement of risk most suited for their integration is the absolute risk, expressed as the probability of fracture within a given time frame, e.g. the 10- year fracture probability in %. Thus, intervention thresholds will be based on fracture risk and differ, therefore, from diagnostic thresholds."(46) This comment is particularly relevant to our discussion as it illustrates that the need for information for a decision-maker – the person instituting an intervention – may take precedence over the need to make a diagnosis of osteoporosis. Other determinants of fracture risk are captured nicely in the well-validated World Health Organization's FRAX tool which is a fracture risk predictor that incorporates bone density along with other clinical risk fractures which are predictors of fracture including age, weight, alcohol use, smoking status and family history, and others.(47) Some of these elements are predictors of falls, which is a recognized major risk for fracture.

Adequate intake of calcium and vitamin D is required for bone formation although neither has specific FDA approval for treatment of osteoporosis. Weight bearing exercise is recommended for encouraging bone formation, as is moderation in alcohol use. Additionally, treatment of medical conditions that contribute to low bone mineral density is recommended, including correction of hyperthyroidism and glucocorticoid excess, and removal of drugs that contribute, if possible, such as some anticonvulsants. Currently approved pharmacotherapies for treating osteoporosis are in Table 3.

Table 3. U.S. Food and Drug Administration Approved Therapies for Osteoporosis (48)

Class	Drugs within class	Mechanism and Notes
Calcitonin	calcitonin-salmon	Calcitonin-salmon is a potent synthetic polypeptide hormone that has similar effects to calcitonins of mammalian origin. It reduces the number of osteoclasts and prevents resorptive activity of the bone resulting in a reduced bone turnover rate. It also temporarily improves bone formation by increasing osteoblastic activity.
Bisphosphonates	alendronate, zoledronate, risedronate, ibandronate	Bisphosphonates bind to bone hydroxyapatite and inhibit osteoclast-mediated bone resorption. Other bisophosphonates are approved for treatment of hypercalcemia and for Paget's disease but are not approved for treatment of osteoporosis.

Selective estrogen receptor modulator	raloxifene	Raloxifene hydrochloride selectively activates and blocks estrogenic pathways by binding to specific estrogen receptors. It acts by reducing bone resorption.
Parathyroid hormone	teriperatide	The skeletal effects of teriparatide depend on the pattern of systemic exposure, when administered once daily teriparatide stimulates new bone formation on trabecular and cortical bone surfaces by preferential stimulation of osteoblastic activity over osteoclastic activity.
Monoclonal antibody	denosumab	Denosumab is a receptor activator of nuclear factor kappa-B ligand (RANKL) inhibitor. Binding to the transmembrane or soluble protein RANKL inhibits the formation, function, and survival of osteoclasts resulting in decreased bone resorption and increased bone mass and strength. This is the most recently approved osteoporosis therapy in the U.S. (August, 2010)

In Europe, there are a few other drugs approved for use including strontium ranelate, and another selective estrogen receptor modulators, lasofoxifene. A newly developed selective estrogen receptor modulator underdevelopment is bazodoxifene which is being tested in combination with estrogen. Several preclinical agents in development are representative of new classes underdevelopment. One is a cathepsin-K inhibitor (odanacatib) that inhibits a protease expressed by osteoclasts which degrades type I collagen. Another is a monoclonal antibody that may prove to be an anabolic agent for bone through its effect on the Wnt signaling pathway.(49)

Despite the array of options now for treatment of osteoporosis there are unanswered questions. Recent data suggest that bisphosphonates have waning usefulness for fracture prevention after approximately 5 years of use. It is uncertain whether the other agents will act similarly. There is uncertainty as to what optimal treatment of osteoporosis should be for very old individuals, who may be at highest risk for falls and fractures but may have substantial competing risks of other events rendering treatments of minimal effectiveness. Should treatments be different in men? There remains uncertainty about the treatment of osteoporosis that differs by cause – should osteoporosis induced by glucocorticoid excess be treated differently than that induced by estrogen deficiency? Perhaps there will be utility to cycling the treatments for osteoporosis that work via different mechanisms to reduce potential toxicities and/or to increase the durability of response to the treatments.

V. A Pragmatic Trial of an Osteoporosis Treatment with Attention to HTE

A. Existing Pragmatic Trials

There have been a handful of pragmatic trials conducted to answer questions about osteoporosis treatment. We highlight two of them here. (50,51) Interestingly, both were trials of non-pharmacological interventions.

One was a trial of hip protectors for community dwelling women. (50)The inclusion criteria were liberal: "Women were eligible to take part in the trial if they were aged 70 years and over and had at least one of the following risk factors: a history of any prior fracture, low body weight (<58 kg), family history of hip fracture (i.e., mother or sibling), or current smoker...Current use of any antifracture medication was not an exclusion criterion, nor did we exclude women who had any form of illness (unless they were bedridden)." The intervention was simple and administered to the women at their homes – women were mailed three pairs of hip protectors along with instructions on how to use them and a leaflet describing other methods of reducing fracture risk. The control group received only the leaflet in the mail. There was no other intervention and no contact with the participants or their physicians. Every six months the participants were contacted by mail and asked to report if they had fallen and/or fractured a hip. The results showed no difference in hip fractures between groups and very poor compliance with the intervention. The only subgroup analysis was of compliant individuals and it is not clear that this was prespecified. There was not a significant treatment benefit for the compliant women either.

Another pragmatic trial under way is to test the effect of Tai Chi on osteopenic women.(51) The authors describe that a key feature of their study design is the use of a network of screened Tai Chi schools to provide Tai Chi interventions in a naturalistic setting and manner. Only the design paper is published to date. The exclusion criteria are fairly extensive for a trial that considers itself to be pragmatic – women will be excluded for having secondary causes of osteopenia, for tobacco use, for malignancies, and for very low body mass indices, among other reasons. The intervention will be, indeed, pragmatic in that women will participate in Tai Chi instruction at any one of many established Tai Chi schools throughout Boston. The outcome measures will be largely biomarkers and tests of strength and balance, although quality of life measures are to be assessed as well. The authors specify that there will be some exploratory analyses in their pilot study. The methods will involve fitting ordinary least squares regression models to evaluate the association between the primary outcomes and a variety of possible predictors (e.g., baseline T-scores, use of calcium and vitamin D supplements, activity level, Tai Chi compliance, exposure to specific types of Tai Chi exercises, etc.). Their goals are to identify factors that will be included in the analysis of the definitive study (i.e., stratification factors) and to identify interesting associations for further investigation (i.e., hypothesis generation.) This suggests that they will consider relevant subgroups in their definitive study.

B. A Proposed Pragmatic Trial

If we were to design a pragmatic trial of a drug for treatment of osteoporosis, there are a number of recommendations we would make. For the sake of discussion we will call the drug that we are testing *Strongbone*; it is an oral agent administered once daily with minimal adverse effects noted in early phase trials. We turn again to the framework of Bombardier and Maetzel to organize our recommendations.(17)

Table 4 Pragmatic study of an osteoporosis drug: An example

	Recommendations
Objective	Is *Strongbone* effective and safe when used by clinicians in their usual practice, this includes practices of internists, family physicians, endocrinologists, geriatricians, endocrinologists, gynecologists, and advanced practice nurses in settings across the United States?
Motivation	This trial is motivated by the need to answer the question as to whether *Strongbone* is effective at preventing fracture and improving quality of life, and safe from major adverse effects, when prescribed in a usual practice setting to appropriate patients.
Intervention	Patients will be randomized by clinic site to one of two starting doses of *Strongbone* (two separate treatment arms.) It is not expected that this drug will be titrated to effectiveness, so clinicians will be asked to continue the enrolled subject on the starting dose, and they are free to add supplemental therapies, such a calcium and vitamin D, as they wish. They are free to recommend exercise, and any other bone healthy lifestyle recommendations. They will *not* be permitted to use the drug concurrently with another active therapy as this is *not* how the drug is expected to be used in practice and its product labeling is (or will be) for montherapy. Clinicians are also permitted to discontinue therapy if they are not satisfied with the response or if the patient develops side effect. The clinician will chose any another therapy at that point. This will *not* be considered a protocol violation.

Comparator	Eligible individuals at the other participating trial sites will be randomized to one of two active comparators. We will use two active therapies that are in widespread use: alendronate once weekly (70mg) and raloxifene (60mg) dosed daily.
Design	This will be a randomized controlled trial with randomization at the level of the clinic. (We may also have chosen randomization at the level of the clinician or participant). This study will be blinded. The clinic site will distribute capsules that are overencapsulations of one of the four regimens. The alendronate recipients will receive an inert substance on 6 days of the week.
Subjects	We want the participants in the trial to form a population that is exchangeable with the population that is likely to receive this therapy. In other words, we want our inclusion and exclusion criteria to be sufficiently liberal that the individuals in the study may be viewed as a random sample of the individuals who will be treated by clinicians in a usual care setting. To this end, the age distribution of the subjects should approximate the age distribution of likely recipients in the population, and the distribution of other medical illnesses should as well. The goal is that this distribution of baseline characteristics will yield a study population that has a distribution of baseline risk for the outcome (fracture) that is comparable to that in the population, and sufficiently diverse to allow for subgroup analyses. We will not specify the cause of osteoporosis, we will not specify any inclusion criteria regarding age, sex, or race. We will neither require nor forbid enrollment of individuals with past fractures. We will not require any specified bone density measurement – it will be sufficient if the clinician believes that the patient would benefit from medication for osteoporosis. (We will use bone density information at baseline in subgroup analyses).
Outcomes	We will assess as the primary outcomes: vertebral fracture, hip fracture, and functional status; and the patient reported outcomes of quality of life and pain. Outcome measures will be after 3 years and after 5 years of therapy. Secondary outcomes will be biomarker information – primarily bone density at 3 years and at 5 years, as well as information about tolerability of the intervention (primarily adherence to therapy) by the patient and the clinician
Analysis	The results will be presented as intent-to treat analyses, accounting for clustering of patients by clinic site. We will also do our confirmatory and exploratory analyses for HTE. We will test our confirmatory analysis that women with a bone mineral density T-score less than 3.5 are particularly good candidates for *Strongbone* therapy.

a) There is biological plausibility and prior evidence about these severely affected women from early phase trials.
b) We have prespecified that women with the lowest bone density (T-score <3.5) will have a more beneficial response to therapy than women with higher T-scores.
c) The study power is adequate for examination of HTE – we will use a blocked randomization process to assure sufficient numbers of low bone density women in each group.
d) There is adequate measurement of outcome, exposure, and prognostic variable.
e) Biases due to measurement error, confounding, and missing information have been minimized – clinical sites will need to submit documentation about the accuracy and reliability of the DXA devices to which they most commonly refer patients; the quality of life instruments will be validated instruments and suitable for self- administration or administration by a lightly trained staff member.
f) The analytic protocol includes prespecification of important data modeling methods, including whether an interaction term for the low bone mineral density subgroup by treatment will be included in the models or that analyses will be stratified by bone density.

All other subgroup analyses will be considered exploratory and results inappropriate for decision making. These may include subgroups defined by sex, race, age, and concurrent use of corticosteroids.

C. Other Considerations about Sources of Heterogeneity

As the trial is designed, there would be additional considerations about the range of *sources* of heterogeneity of treatment effect, as these might be considered for subgroup analyses. As described in Section II, we would recognize that heterogeneity could arise from differences in the patients' treatment responsiveness, from the baseline differences in their risks of the outcome, from the presence of competing risks in some of the treated individuals, and from harms from treatment that differ across individuals. In a pragmatic trial, these may be all more pronounced because of the flexible inclusion criteria.

We begin by discussing treatment responsiveness. These are patient characteristics that are assessable at baseline that influence the *biological* response to the treatment. We have identified in the literature very few baseline characteristics that have been demonstrated to predict treatment response to any osteoporosis therapy and it is challenging to know if the subgroup effects are due to biological differences in response or due to differences in baseline risk of the outcome. One study described the marked heterogeneity in response to parathyroid hormone therapy but found no baseline predictors to explain this heterogeneity.(52) Perhaps most proven is a differential response to estrogen therapy based on an estrogen receptor polymorphism.(53) The items listed in Table 5 are possible predictors of biological treatment responsiveness but are largely unproven.

Table 5. Factors that Might be Evaluated as Contributors to Biological Treatment Responsiveness

Biological Factor (References)
Age (53-55)
Sex(55)
Menopausal status (56)
Bone density (57,58)
Bone architecture (59)
Bone turnover at baseline
Quantitative measures of bone remodeling
Treatment naivety (60,61)
Calcium deficiency(62)
Vitamin D concentration (63,64)
Genetic polymorphisms (65,66)
Concomitant therapies (67)
Renal impairment(68)

The other categories that contribute to heterogeneity of treatment response are factors that influence whether the treatment is *effective* when used in that individual. An individual with metastatic lung cancer will not benefit from treatment of osteoporosis; he or she is unlikely to survive long enough to benefit. In this situation, the cancer diagnosis is a *competing risk* that makes treatment of osteoporosis ineffective. Any condition that importantly reduces life span so that the patient will not live to benefit from the therapy is an important competing risk. The relevance also depends on the time to benefit from the therapy, which may not be well known, particularly for novel therapies.

The importance of the baseline risk of the outcome is that it determines the absolute benefit from the intervention. If the outcome is very rare, even a highly effective intervention will likely have little importance to the individual because the absolute risk reduction will be low (even if the relative risk reduction is promising). This is important in the use of cardiac defibrillators – the risk of events is sufficiently low in patients with normal left ventricular ejection fractions that their absolute benefit is small relative to patients with low ejection fractions. Baseline bone density is a good predictor of fractures; therefore, in a treatment trial of an osteoporosis therapy, bone density may determine the absolute benefit. Women with adequate bone density and low risk of fracture at baseline are unlikely to have an absolute benefit from treatment of osteoporosis. In addition to bone density, any strong predictor of fracture might be similarly useful.

Finally, the risk of harm from an intervention may negate the benefit from an intervention. An effective intervention is one where the benefits outweigh the harms. If the risk of harm is low, a treatment may only need to confer a small absolute benefit to still prove worthwhile. For the most part, therapies for osteoporosis do not pose a mortality risk; however the side effects may negate their benefit. The harms attributable to a treatment may depend on a patient's baseline characteristics. One study reported little difference in harms from ibandronate among women over 70 years relative to those younger than 70 years.(69) There are many methods for balancing benefits and harms when considering the effectiveness of a treatment, but none is used widely for interpreting the results of an individual trial. Additionally, the incorporation of patient preferences and risk tolerance is important when weighing risks and benefits from an intervention but this information may not be practical to collect in a pragmatic trial.

Finally, we discuss post-randomization mediators of effect. Certainly a prominent mediator in medication trials is adherence to therapy. In most treatment trials it is powerful predictor of outcomes, even in the placebo treated arms of trials. Treatment adherent individuals have been shown to have better outcomes regardless of the intervention, including in the Women's Health Initiative hormone therapy trial.(70) There are tools that can be used to predict adherence including the Adherence Estimator which is a 3-item instrument designed to estimate a patient's likelihood of adhering to a medication prescribed for a chronic disease.(71) A study by McHorney et al, which validates this instrument, reports that there have been at least 25 instruments developed for "screening" patients on their propensity to adhere to medications although few have been well validated.(72) Additionally, the use of these instruments in clinical practice likely differs from their use in clinical trials. Nonetheless, balancing treatment arms by response to an instrument that predicts adherence may be an additional tool to balance treatment arms and reduce some of the heterogeneity in treatment effect.

VI. Discussion, Conclusions and Next Steps

We recommend that pragmatic trials *always* have a goal of informing the users of the trial about HTE. Although these trials are typically designed to generate an average treatment effect in a diverse population in a usual care, we argue that this is inadequate and sub-optimal use of information. While the trial's results are useful if there is a strong, convincing average treatment difference between groups, a pragmatic trial has an opportunity to provide so much additional decision-relevant information. Data dredging is never right – the risks of type I error is too high – but carefully specified hypotheses about interaction effects and subgroup effects chosen before the trial is conducted are essential. We recommend that some hypotheses be specified as the confirmatory hypotheses and meet the criteria described above. These subgroup effects will be useful for decision making if the treatment effect is convincing. Other hypotheses will be specified as being exploratory and provide important information for testing in future studies, although not for decision-making.

Because adherence to the intervention is such an important treatment effect mediator in pragmatic trials (more so than in explanatory/efficacy trials), it deserves special mention. We would recommend that in pragmatic trials, particular attention is paid to analyses of treatment effect in those who comply with treatment. "On-treatment" analysis is fraught with biases as the adherent patients are often importantly different than the non-adherent patients. Advanced methods that account for these difference are available and should be considered in the analysis of pragmatic trials.(73-75) Additionally, because we suggest that analysis of adherent individuals be a "confirmatory analysis", the "checklist" of items needs to be met. There should be biological plausibility that adherence to the treatment affects the treatment effect; this analysis should be prespecified and with anticipated results prespecified in the direction expected from prior evidence (e.g. adherent individuals will have a lower odds of outcome than non-adherent individuals); there should be an expectation that there will be a sufficient number of adherent individuals for adequate power; there will be adequate tools for assessment of adherence (a criteria that may be fairly challenging in pragmatic trials); steps are taken to reduce measurement error and missing information; and there is prespecification of how adherence will be modeled. We acknowledge that the greatest challenge for this in a pragmatic trial is having a good measure of adherence when there is less contact with participants that there is in a more controlled study. Nevertheless, we suggest that this will increase the value of pragmatic trials.

We also recommend that sufficient data be collected at baseline so that hypotheses regarding HTE can be explored. We would not want the goal of making pragmatic trials "simple" interfere with the collection of adequate information at baseline for exploration. The baseline data will not be used to make the inclusion and exclusion criteria more stringent and will not play a role in the randomization process, but will be used informatively in the analyses for exploration of HTE. We acknowledge that there will likely be a tension between the desire to make the trials streamlined and acceptable to clinicians in the community, and the desire to have this informative baseline data. We also acknowledge the challenges of measurement differences across sites. However, data that is inexpensive to collect, that is uncomplicated to standardize across sites, and that may identify treatment effect modifiers, should be collected.

In summary, this is an exciting time as pragmatic trials gain value for comparative effectiveness research and the tools of understanding HTE continue to improve. We hope that pragmatic trials proceed with testing treatments that benefit individuals in the settings in which they use these treatments, but that they are analyzed *as informatively as possible* out of respect to the patient, clinician participants, and the funders of these large trials.

Reference List

1. Schwartz D, Lellouch J: Explanatory and pragmatic attitudes in therapeutical trials. *J Chronic Dis* 20:637-648, 1967

2. Bell KJ, Irwig L, Craig JC, Macaskill P: Use of randomised trials to decide when to monitor response to new treatment. *BMJ* 336:361-365, 2008

3. Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2. ISIS-2 (Second International Study of Infarct Survival) Collaborative Group. *Lancet* 2:349-360, 1988

4. Gail M, Simon R: Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics* 41:361-372, 1985

5. Kent DM, Hayward RA: Limitations of applying summary results of clinical trials to individual patients - The need for risk stratification. *Jama-Journal of the American Medical Association* 298:1209-1212, 2007

6. McIntosh MW: The population risk as an explanatory variable in research synthesis of clinical trials. *Stat Med* 15:1713-1728, 1996

7. Schmid CH, Lau J, McIntosh MW, Cappelleri JC: An empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials. *Statistics in Medicine* 17:1923-1942, 1998

8. Dohoo I, Stryhn H, Sanchez J: Evaluation of underlying risk as a source of heterogeneity in meta-analyses: a simulation study of Bayesian and frequentist implementations of three models. *Prev Vet Med* 81:38-55, 2007

9. Varadhan R, Weiss CO, Segal JB, Wu AW, Scharfstein D, Boyd C: Evaluating health outcomes in the presence of competing risks: a review of statistical methods and clinical applications. *Med Care* 48:S96-105, 2010

10. Shuldiner AR, O'Connell JR, Bliden KP, Gandhi A, Ryan K, Horenstein RB, Damcott CM, Pakyz R, Tantry US, Gibson Q, Pollin TI, Post W, Parsa A, Mitchell BD, Faraday N, Herzog W, Gurbel PA: Association of cytochrome P450 2C19 genotype with the antiplatelet effect and clinical efficacy of clopidogrel therapy. *JAMA* 302:849-857, 2009

11. Tighiouart M, Rogatko A, Babb JS: Flexible Bayesian methods for cancer phase I clinical trials. Dose escalation with overdose control. *Stat Med* 24:2183-2196, 2005

12. Piantadosi S, Liu G: Improved designs for dose escalation studies using pharmacokinetic measurements. *Stat Med* 15:1605-1618, 1996

13. Feinstein AR: An additional basic science for clinical medicine: II. The limitations of randomized trials. *Ann Intern Med* 99:544-550, 1983

14. Charlton BG: Mega-trials: methodological issues and clinical implications. *J R Coll Physicians Lond* 29:96-100, 1995

15. Peto R, Collins R, Gray R: Large-scale randomized evidence: large, simple trials and overviews of trials. *Ann N Y Acad Sci* 703:314-40.:314-340, 1993

16. Charlton BG: Megatrials are based on a methodological mistake. *Br J Gen Pract* 46:429-431, 1996

17. Bombardier C, Maetzel A: Pharmacoeconomic evaluation of new treatments: efficacy versus effectiveness studies? *Ann Rheum Dis* 58 Suppl 1:I82-5.:I82-I85, 1999

18. Thorpe KE, Zwarenstein M, Oxman AD, Treweek S, Furberg CD, Altman DG, Tunis S, Bergel E, Harvey I, Magid DJ, Chalkidou K: A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. *J Clin Epidemiol* 62:464-475, 2009

19. Karanicolas PJ, Montori VM, Devereaux PJ, Schunemann H, Guyatt GH: A new 'mechanistic-practical" framework for designing and interpreting randomized trials. *J Clin Epidemiol* 62:479-484, 2009

20. Oxman AD, Lombard C, Treweek S, Gagnier JJ, Maclure M, Zwarenstein M: Why we will remain pragmatists: four problems with the impractical mechanistic framework and a better solution. *J Clin Epidemiol* 62:485-488, 2009

21. Oxman AD, Lombard C, Treweek S, Gagnier JJ, Maclure M, Zwarenstein M: A pragmatic resolution. *J Clin Epidemiol* 62:495-498, 2009

22. Mant D: Can randomised trials inform clinical decisions about individual patients? *Lancet* 353:743-746, 1999

23. Birtcher KK, Greisinger AJ, Brehm BJ, Wehmanen OA, Furman LM, Salinas CC, Mirzai-Tehrane M, Nayak A, Rashid H, Mortazavi A: A secondary prevention lipid clinic reaches low-density lipoprotein cholesterol goals more often than usual cardiology care with coronary heart disease. *J Clin Lipidol* 4:46-52, 2010

24. Taveras E, Gortmaker S, Kleinman K, Mitchell K, Price S, Prosser L, Rifas-Shiman S, Gillman M: C-c2-02: the high five for kids study: an intervention to improve primary care to prevent childhood obesity. *Clin Med Res* 8:204, 2010

25. Major cardiovascular events in hypertensive patients randomized to doxazosin vs chlorthalidone: the antihypertensive and lipid-lowering treatment to prevent heart attack trial (ALLHAT). ALLHAT Collaborative Research Group. *JAMA* 283:1967-1975, 2000

26. Wright JT, Jr., Probstfield JL, Cushman WC, Pressel SL, Cutler JA, Davis BR, Einhorn PT, Rahman M, Whelton PK, Ford CE, Haywood LJ, Margolis KL, Oparil S, Black HR, Alderman MH: ALLHAT findings revisited in the context of subsequent analyses, other trials, and meta-analyses. *Arch Intern Med* 169:832-842, 2009

27. Stroup TS, McEvoy JP, Swartz MS, Byerly MJ, Glick ID, Canive JM, McGee MF, Simpson GM, Stevens MC, Lieberman JA: The National Institute of Mental Health Clinical Antipsychotic Trials of

Intervention Effectiveness (CATIE) project: schizophrenia trial design and protocol development. *Schizophr Bull* 29:15-31, 2003

28. Brand R, Kragt H: Importance of trends in the interpretation of an overall odds ratio in the meta-analysis of clinical trials. *Statistics in Medicine* 11:2077-2082, 1992

29. McIntosh MW: The population risk as an explanatory variable in research synthesis of clinical trials. *Stat Med* 15:1713-1728, 1996

30. Walter SD: Variation in baseline risk as an explanation of heterogeneity in meta-analysis. *Stat Med* 16:2883-2900, 1997

31. Follmann DA, Proschan MA: A multivariate test of interaction for use in clinical trials. *Biometrics* 55:1151-1155, 1999

32. Ioannidis JPA, Lau J: Heterogeneity of the baseline risk within patient populations of clinical trials - A proposed evaluation algorithm. *American Journal of Epidemiology* 148:1117-1126, 1998

33. Kent DM, Hayward RA: Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. *JAMA* 298:1209-1212, 2007

34. Williams TA, Mulatero P, Filigheddu F, Troffa C, Milan A, Argiolas G, Parpaglia PP, Veglio F, Glorioso N: Role of HSD11B2 polymorphisms in essential hypertension and the diuretic response to thiazides. *Kidney Int* 67:631-637, 2005

35. Glasziou PP, Irwig LM: An evidence based approach to individualising treatment. *BMJ* 311:1356-1359, 1995

36. Weiss CO, Boyd CM, Yu Q, Wolff JL, Leff B: Patterns of prevalent major chronic disease among older adults in the United States. *JAMA* 298:1160-1162, 2007

37. Wu AW, Snyder C, Clancy CM, Steinwachs DM: Adding the patient perspective to comparative effectiveness research. *Health Aff (Millwood)* 29:1863-1871, 2010

38. Garratt A, Schmidt L, Mackintosh A, Fitzpatrick R: Quality of life measurement: bibliographic study of patient assessed health outcome measures. *BMJ* 324:1417, 2002

39. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HC: The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol* 63:737-745, 2010

40. Wallace AS, Carlson JR, Malone RM, Joyner J, Dewalt DA: The influence of literacy on patient-reported experiences of diabetes self-management support. *Nurs Res* 59:356-363, 2010

41. Kjekshus J, Apetrei E, Barrios V, Bohm M, Cleland JG, Cornel JH, Dunselman P, Fonseca C, Goudev A, Grande P, Gullestad L, Hjalmarson A, Hradec J, Janosi A, Kamensky G, Komajda M, Korewicki J, Kuusi T, Mach F, Mareev V, McMurray JJ, Ranjith N, Schaufelberger M, Vanhaecke J, van

Veldhuisen DJ, Waagstein F, Wedel H, Wikstrand J: Rosuvastatin in older patients with systolic heart failure. *N Engl J Med* 357:2248-2261, 2007

42. Follmann DA, Proschan MA: A multivariate test of interaction for use in clinical trials. *Biometrics* 55:1151-1155, 1999

43. Kaplan SH, Billimek J, Sorkin DH, Ngo-Metzger Q, Greenfield S: Who can respond to treatment? Identifying patient characteristics related to heterogeneity of treatment effects. *Med Care* 48:S9-16, 2010

44. Simon R: Bayesian subset analysis: application to studying treatment-by-gender interactions. *Stat Med* 21:2909-2916, 2002

45. Consensus development conference: diagnosis, prophylaxis, and treatment of osteoporosis. *Am J Med* 94:646-650, 1993

46. Kanis JA, McCloskey EV, Johansson H, Oden A, Melton LJ, III, Khaltaev N: A reference standard for the description of osteoporosis. *Bone* 42:467-475, 2008

47. World Health Organization Collaborating Centre for Metabolic Bone Diseases UoSU: WHO Fracture Risk Assessment Tool. [article online], 2010. Accessed 28 December 2010

48. MICROMEDEX (R) 1.0. [article online], 2010. Available from http://www.thomsonhc.com/hcs/librarian/ND_T/HCS/ND_CPR/ProductList/ND_PR/Drugs/CS/F759CC/DUPLICATIONSHIELDSYNC/716B77/ND_PG/PRIH/ND_B/HCS/SBK/2/ND_P/Drugs/PFActionId/hcs.Home. Accessed 28 December 2010

49. Honig S: Osteoporosis - new treatments and updates. *Bull NYU Hosp Jt Dis* 68:166-170, 2010

50. Birks YF, Porthouse J, Addie C, Loughney K, Saxon L, Baverstock M, Francis RM, Reid DM, Watt I, Torgerson DJ: Randomized controlled trial of hip protectors among women living in the community. *Osteoporos Int* 15:701-706, 2004

51. Wayne PM, Buring JE, Davis RB, Connors EM, Bonato P, Patritti B, Fischer M, Yeh GY, Cohen CJ, Carroll D, Kiel DP: Tai Chi for osteopenic women: design and rationale of a pragmatic randomized controlled trial. *BMC Musculoskelet Disord* 11:40, 2010

52. Sellmeyer DE, Black DM, Palermo L, Greenspan S, Ensrud K, Bilezikian J, Rosen CJ: Hetereogeneity in skeletal response to full-length parathyroid hormone in the treatment of osteoporosis. *Osteoporos Int* 18:973-979, 2007

53. Brodowska A, Starczewski A, Brodowski J, Szydlowska I, Nawrocka-Rutkowska J: The bone mass density in postmenopausal women using hormonal replacement therapy in relation to polymorphism in vitamin D receptor and estrogen receptor genes. *Gynecol Endocrinol* 25:315-323, 2009

54. Boonen S, Marin F, Mellstrom D, Xie L, Desaiah D, Krege JH, Rosen CJ: Safety and efficacy of teriparatide in elderly women with established osteoporosis: bone anabolic therapy from a geriatric perspective. *J Am Geriatr Soc* 54:782-789, 2006

55. Eastell R, Black DM, Boonen S, Adami S, Felsenberg D, Lippuner K, Cummings SR, Delmas PD, Palermo L, Mesenbrink P, Cauley JA: Effect of once-yearly zoledronic acid five milligrams on fracture risk and change in femoral neck bone mineral density. *J Clin Endocrinol Metab* 94:3215-3225, 2009

56. Langdahl BL, Marin F, Shane E, Dobnig H, Zanchetta JR, Maricic M, Krohn K, See K, Warner MR: Teriparatide versus alendronate for treating glucocorticoid-induced osteoporosis: an analysis by gender and menopausal status. *Osteoporos Int* 20:2095-2104, 2009

57. Boonen S, Adachi JD, Man Z, Cummings SR, Lippuner K, Torring O, Gallagher JC, Farrerons J, Wang A, Franchimont N, San MJ, Grauer A, McClung M: Treatment with Denosumab Reduces the Incidence of New Vertebral and Hip Fractures in Postmenopausal Women at High Risk. *J Clin Endocrinol Metab* 2011

58. Ensrud KE, Black DM, Palermo L, Bauer DC, Barrett-Connor E, Quandt SA, Thompson DE, Karpf DB: Treatment with alendronate prevents fractures in women at highest risk: results from the Fracture Intervention Trial. *Arch Intern Med* 157:2617-2624, 1997

59. Wehrli FW, Song HK, Saha PK, Wright AC: Quantitative MRI for the assessment of bone structure and function. *NMR Biomed* 19:731-764, 2006

60. Obermayer-Pietsch BM, Marin F, McCloskey EV, Hadji P, Farrerons J, Boonen S, Audran M, Barker C, Anastasilakis AD, Fraser WD, Nickelsen T: Effects of two years of daily teriparatide treatment on BMD in postmenopausal women with severe osteoporosis with and without prior antiresorptive treatment. *J Bone Miner Res* 23:1591-1600, 2008

61. Borggrefe J, Graeff C, Nickelsen TN, Marin F, Gluer CC: Quantitative computed tomographic assessment of the effects of 24 months of teriparatide treatment on 3D femoral neck bone distribution, geometry, and bone strength: results from the EUROFORS study. *J Bone Miner Res* 25:472-481, 2010

62. Burnell JM, Baylink DJ, Chesnut CH, III, Teubner EJ: The role of skeletal calcium deficiency in postmenopausal osteoporosis. *Calcif Tissue Int* 38:187-192, 1986

63. Antoniucci DM, Vittinghoff E, Palermo L, Black DM, Sellmeyer DE: Vitamin D insufficiency does not affect response of bone mineral density to alendronate. *Osteoporos Int* 20:1259-1266, 2009

64. Dawson-Hughes B, Chen P, Krege JH: Response to teriparatide in patients with baseline 25-hydroxyvitamin D insufficiency or sufficiency. *J Clin Endocrinol Metab* 92:4630-4636, 2007

65. Marini F, Falchetti A, Silvestri S, Bagger Y, Luzi E, Tanini A, Christiansen C, Brandi ML: Modulatory effect of farnesyl pyrophosphate synthase (FDPS) rs2297480 polymorphism on the response to long-term amino-bisphosphonate treatment in postmenopausal osteoporosis. *Curr Med Res Opin* 24:2609-2615, 2008

66. Marc J, Prezelj J, Komel R, Kocijancic A: VDR genotype and response to etidronate therapy in late postmenopausal women. *Osteoporos Int* 10:303-306, 1999

67. Ste-Marie LG, Schwartz SL, Hossain A, Desaiah D, Gaich GA: Effect of teriparatide [rhPTH(1-34)] on BMD when given to postmenopausal women receiving hormone replacement therapy. *J Bone Miner Res* 21:283-291, 2006

68. Miller PD, Schwartz EN, Chen P, Misurski DA, Krege JH: Teriparatide in postmenopausal women with osteoporosis and mild or moderate renal impairment. *Osteoporos Int* 18:59-68, 2007

69. Ettinger MP, Felsenberg D, Harris ST, Wasnich R, Skag A, Hiltbrunner V, Wilson K, Schimmer RC, Miller PD: Safety and tolerability of oral daily and intermittent ibandronate are not influenced by age. *J Rheumatol* 32:1968-1974, 2005

70. R Curtis J, Larson JC, Delzell E, Brookhart MA, Cadarette SM, Chlebowski R, Judd S, Safford M, Solomon DH, Lacroix AZ. Placebo Adherence, Clinical Outcomes, and Mortality in the Women's Health Initiative Randomized Hormone Therapy Trials. *Medical Care* 2011 [Epub]

71. McHorney CA: The Adherence Estimator: a brief, proximal screener for patient propensity to adhere to prescription medications for chronic disease. *Curr Med Res Opin* 25:215-238, 2009

72. McHorney CA, Victor SC, Alexander CM, Simmons J: Validity of the adherence estimator in the prediction of 9-month persistence with medications prescribed for chronic diseases: a prospective analysis of data from pharmacy claims. *Clin Ther* 31:2584-2607, 2009

73. Efron B, Feldman D: Compliance As An Explanatory Variable in Clinical-Trials. *Journal of the American Statistical Association* 86:9-17, 1991

74. Sommer A, Zeger SL: On estimating efficacy from clinical trials. *Stat Med* 10:45-52, 1991

75. Follman D: On the Effect of Treatment among Would-Be Treatment Compliers: An Analysis of the Multiple Risk Factor Intervention Trial. *Journal of the American Statistical Association* 95:1101-1109, 2000