American Journal of Clinical and Experimental Medicine
2015; 3(4): 178-188
Published online July 2, 2015 (http://www.sciencepublishinggroup.com/j/ajcem)
doi: 10.11648/j.ajcem.20150304.19
ISSN: 2330-8125 (Print); ISSN: 2330-8133 (Online)

Clinical Trials of New Drug Products: What Gets Compared to Whom

Charles Joseph Kowalski, Adam Joel Mrdjenovich

Health and Behavioral Sciences Institutional Review Board, Office of Research, University of Michigan, Ann Arbor, MI, USA

Email address:
chuckk@umich.edu (C. J. Kowalski)

To cite this article:
Charles Joseph Kowalski, Adam Joel Mrdjenovich. Clinical Trials of New Drug Products: What Gets Compared to Whom. *American Journal of Clinical and Experimental Medicine*. Vol. 3, No. 5, 2015, pp. 178-188. doi: 10.11648/j.ajcem.20150304.19

Abstract: Two of the most controversial aspects of phase III clinical trial design are the choice of the control group(s) and the choice of the outcome variable(s). Each of these choices has overlapping scientific and ethical ramifications, and the tension between maximizing scientific validity on the one hand, and protecting the rights and welfare of the human participants in the trial on the other, is the main source of the controversy. The intensity of the debate is increased whenever these choices are motivated not by scientific or ethical principles, but are driven by conflicts of interest. And so it comes to pass that in testing the safety and efficacy of new drug products, when study design choices are made more to achieve rapid market approval than to accurately assess safety and efficacy, thereby putting the welfare of both the trial participants and future patients at risk, the public and its advocates will recoil. In this paper, we study two issues of this kind: the use of placebo controls when an established intervention for the condition under consideration exists, and the use of surrogate outcome measures. There is a rich and growing literature on both of these topics and we will have little to contribute to a greater theoretical understanding of either of them. Our aim is to point to the ethical ramifications of these choices in the context of clinical trials of new drug products, and to suggest how these choices may be better made to serve public health interests. What is to come is portended by the observation that, "In the United States, the long tradition of leaving to the pharmaceutical industry the task of evaluating the efficacy and safety of its products has permitted manufacturers to make study design choices that largely determine the shape of the answers eventually provided by the trials" (Psaty and Weiss, 2007, p. 330).

Keywords: Phase III Clinical Trials, Trial Design, Drug Safety and Efficacy

1. Clinical Trials of New Drug Products: What Gets Compared to Whom

The pharmaceutical industry has now surpassed the defense industry as the largest defrauder of the federal government, as measured by payments made for violations of the False Claims Act. While most of these violations were for promotion of off-label uses of prescription drugs and overcharging for drugs under various federal programs, these problems have pointed to the need to examine other practices of the large drug companies, whose motives seem more directed towards profits than to the principles they publically espouse. The disconnect between the pharmaceutical industry's rhetoric and behavior has been pointed out, e.g., by Abramson (2004), Angell (2005), Avorn (2005),Brody (2007), Goozner (2004), Kassirer (2005), and Silverman and Lee (1974). These focus largely on the business/advertising methods employed by the pharmaceutical industry (e.g., gifts to physicians, sponsoring Continuing Medical Education, direct-to-consumer advertising, development of me-too drugs at the expense of real innovation, selling techniques of drug detailers, extending patent protection, the use of lobbyists, misuses of the Bayh-Dole Act, etc.). Following Safer (2002), we pursue a more methodological tack.In particular, this paper focuses on two features of the design of phase III clinical trials to assess new drug efficacy and safety, namely, the choice of the control group, and the choice of the outcome measure. In short, if we treat a group of patients with a new, experimental drug, *whatwill we compare to whom*?

These questions are considered in turn below.We begin with the choice of the control group(s). The focus is on the use of a placebo control when an effective intervention for the condition under consideration already exists. We argue that the appropriate comparator in this situation is the drug considered as standard of care; these trials are called *active control trials* (ACTs).ACTs may be appropriately designed as either a superiority trial, an equivalence trial, or a non-inferiority trial. The choice will depend on trial purpose. We then turn to the choice of the outcome variable(s). We argue that the primary outcome should be either length or quality of life (or both), and, in most cases, the use of a surrogate outcome entails serious risks. These risks can be to the validity of the conclusions drawn and/or to the participants in the study, as well as future patients. Examples are given.

2. The Choice of the Control Group

In her critique of the pharmaceutical industry, Angell (2005) made a series of recommendations for change in the ways the drug companies do business andinteract with the Food and Drug Administration (FDA). Although she thought all therecommendations important, if she could choose only one, it would be: "Food and Drug regulations should require that new drugs be compared not just with placebos but with old drugs for the same condition" (p. 240). She was mainly concerned here with the development of "me-too" drugs, and the industry's penchant for developing these at the expense of real innovation. See also Abramson (2004, p. 102), Avorn (2005, p. 53), Goozner (2004, p. 213), and Silverman and Lee (1974, p. 39). We recognize, as did these authors, that there is one place where placebo controls are entirely appropriate, viz., when no effective treatment for the condition exists. Indeed, following Freedman et al (1996, p. 243), we believe that "first-generation treatments, designed for previously untreatable conditions, *should* be tested against placebo before receiving regulatory approval," and that since *treatability*involves a clinical judgment, "this principle judges placebo controls to be acceptable for testing of refractory populations, and for treatments with an overall unfavorable therapeutic index." However, with this exception, we will argue that, in general, placebo controls are to be avoided whenever second-generation drugs are being evaluated. The pragmatic sound-bite for this stance is *what we really want to know is how the new drug compares with the old, not how the new drug compares to nothing.*This is surely not a new, novel idea. Indeed, almost fifty years ago, Sir Bradford Hill (1963, p. 1048) answered the question of whether it is ethical to use a placebo by noting,

"The answer to this question will depend, we suggest, upon whether there is already available an orthodox treatment of proved or accepted value. If there is such an orthodox treatment the question will hardly arise, for the doctor will wish to know whether a new treatment is more, or less, effective than the old, not that it is more effective than nothing."

Current practice has not been guided by this observation and, indeed, not only are placebos often used, they are often used *repeatedly*, trusting that regulatory approval will be based on one or two successful better-than-nothing demonstrations and not on similar previously done trials unable to surpass even this minimal threshold. Petersen (2008) reported that Pfizer did several studies that failed to show its antidepressant Zoloft worked better than placebo before finally completing two trials that did, and this was enough to gain regulatory approval, although not without some consternation on the part of one of the FDA physicians reviewing Pfizer's application. According to Petersen (2008, p. 49), this reviewer remarked, "… the sponsor could just do studies until the cows come home until he gets two of them that are statistically significant by chance alone, walks them out and says he had met the criteria." This observation was not enough to slow the approval of Zoloft, and apparently this strategy was followed by a number of manufacturers of antidepressant drugs. Khan et al (2002)studied trials of fifty-two antidepressants, involving more than ten thousand patients, and found that placebo did just as well or better than the antidepressant more than half the time.

It should also be noted here that in addition to avoiding comparisons to *absolutely* nothing, *what we really want to know is how the new drug compares with the old, and not some watered-down version of the old.* Angell (2005, p. 78ff) gives several examples (Nexium vs. Prilosec, Lipitor vs. Pravachol) where non-equivalent doses of drugs were compared in order to show that the new was superior to the old. See also Safer (2002, p. 583-4). Petersen (2008, p. 48) gives an example going in the opposite direction: A trial involving Claritin administered the drug at a very low dosage (but apparently high enough to outdo placebo) so as to support the advertising claim that it is the first "nonsedating" antihistamine. The results of pharmaceutically sponsored trials are often predictable solely from sponsor identification (Miller and Brody, 2005; Heres et al, 2006; Safer, 2002). Head-to-head comparative trials must be designed so as to give each arm its best chance to succeed. Thompson and Temple (2001, p. 37) recommended that,

"Trial studies should be designed in such a way as to compare the experimental procedures and products to one or more robust alternative treatments, if such exist, and not to weak versions of them. When the alternative treatment is 'standard medical practice,' it should not be some weak version of it, but rather one which the physicians and other personnel are carefully instructed in what is recommended practice. If this standard is not observed, it is impossible to determine how much, if any, the intervention adds to what can already be accomplished by following recommended practice."

In any event, the face-validity of comparing new drugs with the drugs currently considered standard of care (and not with placebo) seems so high that one might well question just how the idea of using placebo controls for testing new drugs ever took root. Freedman et al (1996, p. 258) provide an answer: "When, in 1962, the U.S. Congress extended FDA's mandate to ascertain effectiveness as well as safety of drugs, what exactly was it requiring? The undisputed legal interpretation of effectiveness has been absolute effectiveness – better than nothing – rather than a more useful measure, such as clinical effectiveness.[1] For this reason, FDA has required placebo controls rather than active controls." But the FDA has gone even further, supplementing legal mandate with scientific motivation. As pointed out by Weijer (1999, p.

222), "Substantial scientific criticism … has been leveled against the use of active controls in the evaluation of new drugs. The views of Robert Temple in the US Food and Drug Administration have convinced many clinician-investigators that the use of active controls is associated with serious drawbacks (Temple 1982, 1983)." In brief, these views include that ACTs have raised several unanswered questions and have certain scientific and practical problems:

- In an ACT, should we be testing for equivalence, superiority or non-inferiority?
- ACTs do not have an agreed upon statistical test to establish equivalence;
- ACTs have the wrong incentives;
- ACTs do not provide a direct measure of effect size;
- ACTs require larger sample sizes; and
- ACTs do not always possess 'assay sensitivity.'

The International Conference on Harmonization (ICH 2000) makes it clear that, in their opinion, placebo controlled trials (PCTs) do not suffer from these problems, e.g., to the last problem listed above they attach, "whereas PCTs do" (section 1.5). This is of considerable importance since the ICH document was produced and is endorsed by the regulatory agencies of the United States, the European Union, and Japan. Thus the feeling of many researchers that *the FDA all but requires PCTs* may extend well beyond American borders. The onus is therefore on the proposer of an ACT to argue that either these problems do not exist, or are unimportant, or can be managed in such a way that the ACT will yield clinically useful information.

A number of papers have addressed all or some of these issues (e.g., Anderson (2006, 2009a, 2009b), Freedman, Weijer and Glass (1996), Freedman, Glass and Weijer (1996), Howick (2009), Weijer (1999)), and we believe theseprovide good reasons for saying that the use of ACTs cannot be precluded on methodological grounds. We do not summarize all of the results of each of these papers here. Rather, brief refutations of each of the points listed above are made with an eye toward providing an annotated guide to this literature and to establish the background necessary to present our view that trials of new drug products should be designed to answer clinically relevant questions. PCTs can be valuable under certain (circumscribed) circumstances, but the alleged methodological superiority of PCTs to ACTs is unsupported. Rather than choose between these approaches on the basis of methodological considerations, we should be concentrating on the purpose of the trial and associated ethical constraints to choose the appropriate trial structure. One might think this to be self-evident, but the following two quotes describe prevailing practices: "The driving forces behind the superiority claim of PCTs are not reasoned arguments, but power and economics" (Rennie and Stürmer, 2009, p. 63); and "The present popularity of placebo controlled trials is easy to explain by marketing considerations and regulatory needs, but difficult to justify on scientific grounds" (Enkin, 2009, p.66). Keeping in mindthe following thought experimentwould severely limit the use of placebo controls: "It is hard to think of television's Marcus Welby, M.D., handing a patient a prescription and saying warmly, 'Here take this; it's probably better than nothing'" (Avorn, 2005, p. 61).

2.1. Superiority, Equivalence, or Non-Inferiority

We focus on comparing the response to the experimental treatment, T, to that of the control, C. We suppose that large values of the response are desirable so that T will be considered better than C if T > C. Our view is usually one-sided, i.e., we generally wish to determine whether T is *better* than C, T > C. However, there will be times when we wish to determine if T is equivalent to C, or at least that T is no worse than (non-inferior to) C. When considering equivalence, we need to specify an *equivalence margin*, δ, such that T and C will be considered equivalent whenever $|C – T| <\delta$. For non-inferiority, we need to specify the so-called *non-inferiority margin*, Δ, i.e., how much C can exceed T with T still being considered non-inferior to C. Symbolically, $C – T <\Delta$ or $T > C –\Delta$. If $T < C –\Delta$, T is inferior to C since C exceeds T by more than Δ ($C > T + \Delta$).

Most randomized controlled trials (RCTs) are designed to determine whether one intervention is superior to another, and these trials are called *superiority trials*. There is no difference in the conduct and analysis of superiority trials in which the comparator is placebo and superiority trials in which the comparator is an established treatment. The CONSORT statement may be consulted to see how one best reports, and hence designs and conducts superiority trials of either type. Without attempting detailed symbolic representation, the null hypothesis in a superiority trial is that the treatments are equal (T = C where T is the true mean response to the new treatment and C the true mean response to the control) and this is tested versus the alternative that the new treatment is better than the standard (T > C). Trials are designed so as to have adequate power to reject the null hypothesis whenever the new treatment is better than its comparator by a clinically meaningful amount. Whenever the observed difference is statistically significant, the null hypothesis is rejected, and one concludes (subject to the probabilistic constraints) that the new treatment is indeed superior. If the difference does not attain statistical significance, the null hypothesis is not rejected, but some care is required in interpretation. In particular, it would be a mistake to conclude that the treatments are *equivalent* (let alone equal). Not having enough evidence to reject a proposition does not mean that the proposition is true. For a good discussion, see Altman and Bland (1995); the title of their paper, "Absence of evidence is not evidence of absence" is often quoted in hypothesis testing contexts.

In an equivalence trial, the aim is to determine whether the new treatment is therapeutically similar to an existing treatment, while in a non-inferiority trial the aim is to show that the new treatment is no worse than the old. I show below how one can test for equivalence and/or non-inferiority, but the remainder of this subsection is limited to noting that three different kinds of ACTs are *possible*, and the choice between them will be driven by the purpose of the trial.

The three kinds have already been mentioned: superiority, equivalence and non-inferiority. Temple and Ellenberg (2000, p. 204) thought that superiority trials were, well, superior:

"A well-designed study that shows superiority of a treatment to a control (placebo or active therapy) provides strong evidence of the effectiveness of the new treatment, limited only by the statistical uncertainty of the result. No information external to the trial is needed to support the conclusion of effectiveness."

Note that the control here may be either placebo or active therapy, so that the statement applies to superiority tests generally (and not only to PCT vs. ACT comparisons). Equivalence tests were seen to be less desirable: "in contrast, a study that successfully shows 'equivalence' – that is, little difference between a new drug and a known active treatment – does not by itself demonstrate that the new treatment is effective. 'Equivalence' could mean that the treatments were both effective in the study, but it could also mean that both treatments were ineffective in the study." A similar statement could be made for non-inferiority tests: the new treatment might be no worse than the standard, but this, by itself, does not mean either was effective. This is the so-called *assay sensitivity* problem and this is dealt with below.

2.2. Testing for Equivalence or Non-Inferiority

Both of these trial types can be cast into a hypothesis-testing framework, but it is perhaps more meaningful to approach these questions by constructing confidence intervals for the difference between the new treatment and its referent (Rothman, 1986). Recall that in a superiority trial the confidence interval approach to testing the null hypothesis consists of constructing (say) a 95% confidence interval for the difference between the treatments (we take this difference to be C – T, where C is the control and T the new treatment). One concludes that the treatments differ if this interval does not contain the value zero. If we let δ denote a quantity such that T will be considered equivalent to C if $| C – T | <\delta$, one concludes the two treatments are equivalent if the confidence interval lies completely within the interval from $–\delta$ to $+\delta$. It is not often that one is interested in this "two-sided equivalence" hypothesis: It is more usual to test for non-inferiority. Assuming that larger values of the response are better, one concludes non-inferiority if the upper limit of the confidence interval is less than the non-inferiority margin, Δ (if this lower limit is actually greater than Δ, the new treatment is superior). The CONSORT statement has been extended to non-inferiority and equivalence trials (Piaggio et al, 2006) where details of the testing procedures are well-documented.

2.3. Wrong Incentives

This alleged problem with an ACT arises only in the case where equivalence is (erroneously) being tested using a superiority test framework. In this formulation, the null hypothesis is one of no difference, and the alternative is that one formulation is better than the other. Anything done here that will increase the variance of the response (e.g., just being sloppy) will make rejection of the null hypothesis more difficult, i.e., will bias the test toward the conclusion of no difference. Thus, it is argued that, if one is aiming to establish equivalence, there is no incentive to run a tight ship. However, in a properly structured ACT, staying within the hypothesis testing framework, the null and alternative hypotheses in a superiority trial are reversed. That is, the null hypothesis is now that the treatments differ (by a specified amount, δ) and the alternative is that the treatment difference is less than δ, making them "substantially equivalent." Thus any sloppiness in the trial will make it more difficult to conclude equivalence, not easier to do so. See Weijer (1999) for a more detailed discussion.

2.4. Direct Measure of Effect Size

Let T be the response to the new drug, C the response to the standard drug, and P be the response to placebo. In an ACT, the effect size will be (a function of) T – C, whereas the "absolute effect size" would be T – P. Given only the results from an ACT, then, we will not know the absolute effect size and this has been judged to be of value: "The placebo-controlled trial measures the total pharmacologically mediated effect of treatment. In contrast, an active controlled trial … measures the effect relative to another treatment … The absolute effect size information is valuable" (ICH, 2000, p. 18). While we would agree that it would be nice to know the value of $T – P^*$ where P^* denotes the *true* placebo effect, this is not necessarily given by T – P. The idea of using T – P to estimate a drug's biological effect is based on the assumption that there are two causes for the observed response – biological and psychological – that act independently so that subtracting the psychological (P) leaves the biological component. Freedman et al (1996) questioned the independence of these effects, and Howick (2009, p. 35) argued that, "Actual 'placebo' controls used in clinical trials are often 'illegitimate' in the sense that they do not accurately measure the 'placebo' effect." It has long been known that placebos possess their own pharmacological profiles, including measurable peak times, carry-over effects, cumulative effects, and toxicities. Freedman et al (1996, p. 244) put it simply: "If the very point of a placebo comparison is to provide a baseline against which we may discern a drug's biological effectiveness, which placebo should be used? – Which color, which dose, which dosing schedule, and so forth?" The use of 'illegitimate' placebos is also an issue in the assay sensitivity problem as will be discussed below.

2.5. Larger Sample Sizes

It is widely believed that PCTs require smaller sample sizes than do ACTs, e.g., ICH (2000, section 2.4). This is surely true if the ACT is run as a superiority trial, since one expects the T – P difference to exceed the T – C difference, and so to be easier to detect. Whether or not this is in fact true in a non-inferiority trial depends on the relative sizes of the non-inferiority margin and the treatment difference that the PCT is designed to detect. In particular, the choice of the non-inferiority margin, Δ, is germane to the question. The ICH (2000) guidelines suggest two criteria:

- The determination of the margin in a non-inferiority trial is based on both statistical reasoning and clinical judgment, and should reflect uncertainties in the evidence on which the choice is based, and should be suitably conservative.
- This non-inferiority margin cannot be greater than the smallest effect size that the active drug would reliably be expected to have compared with placebo in the setting of a placebo-controlled trial.

If we follow the second of these recommendations, i.e., If the equivalence margin is smaller than the active drug's effect size relative to placebo($\Delta < C - P$),the sample size will in fact be larger for the non-inferiority trial, and the increase will depend on the choice of Δ. This 'disadvantage' may be offset by other practical considerations, however (Howick, 2009, p. 43). First, it should be easier to attract participants to a trial in which randomization will lead to an active treatment, and not a placebo. Second, participants in a PCT may guess that they are getting the placebo and either drop-out of the study or covertly seek treatment elsewhere. This possibility should not be overlooked as subjects have proven quite adept at making this determination (Fergusson et al, 2004; Hrobjartsson et al, 2007). Finally, the PCT will prove inadequate to provide a firm basis for deciding whether or not the new treatment should actually be used. Howick noted, "In order to make an informed choice about whether to use the new treatment, the patient, the practitioner, or policy maker must know how the new treatment compares with the best existing treatments not merely how it compares with placebo."

2.6. Assay Sensitivity

The assay sensitivity question was thoughtfully addressed by Anderson (2006) and Howick (2009). Anderson's treatment is the more theoretical, focusing on the "inferential self-containment" that is claimed for PCTs (but not for ACTs) in Temple and Ellenberg's (2000, p. 204) statement, "No information external to the trial is needed to support the conclusion of superiority" (whereas "equivalence does not by itself demonstrate that the new treatment is effective"). Arguing that "the meaningful interpretation of the results of any empirical test depends on a host of background information and assumptions concerning the test conditions" (p. 72), Anderson (2006) shows clearly that "no test, placebo-controlled or not, possesses the property in inferential self-containment" (p. 73). He concludes,

"… there is no reason to believe that we can trust the findings of PCTs to a greater degree than the findings of ACTs. Nor is there an absolute contrast between ACTs and PCTs with respect to self-containment. In both ACTs and PCTs the ontological question concerning whether the trial possesses the property of 'assay sensitivity' is underdetermined by the (internal) evidence. Both ACTs and PCTs depend on external information for their meaningful interpretation" (p. 79).

Howick (2009) points to one way that PCTs suffer from their own assay sensitivity problems, namely, that actual 'placebo' controls can be more, or less, effective than 'real' placebos. After citing several examples where actual 'placebo' controls do not accurately measure the true 'placebo' effect, he places PCTs on equal footing with ACTs as far as assay sensitivity is concerned: "In order to claim that ACTs possess assay sensitivity we must assume that the established treatment control was effective; in order to claim that PCTs have assay sensitivity we must assume that a particular 'placebo' effect is 'correct'" (p. 37).

We believe that the above literature successfully refutes the challenges leveled at ACTs; and that the choice of trial architecture should be driven by the purpose of the trial which will, in most cases, be to assess the comparative efficacy of the new treatment to an existing standard. If there are doubts about the efficacy of the currently favored treatment, Angell (2005, p. 241) pointed out that the choice of the control group need not necessarily be dichotomous: "If there is really doubt about whether a standard treatment is effective, the FDA should require that clinical trials of new treatments have three comparison groups – new drug, old drug, and placebo." We close this section with a reprise of its beginning – namely that trial purpose should drive trial design and, note that, when one tests a second generation drug, ACTs are best in the sense that they better fulfill what is (or should be) the trial's purpose. As put by Anderson (2009a, p.61):

"Clinical trials cannot be conducted in isolation from the clinical context in which the intervention will be used. Clinical trial design, thus, must *begin* with the identification of what the average patient, clinician, and policy maker need to know." … "given the availability of an established therapy, well-designed ACTs (of both the non-inferiority and superiority variety) are *methodologically* preferable to PCTs precisely because they answer the question that is scientifically relevant in this context: is a new therapy better than what we already have?"

3. The Choice of an Outcome Measure

Temple and Thompson (2001, p. 31) advise, "In determining the value of a medical intervention two questions should be asked: 'Does it improve the quality of the person's life?' and 'Does it extend the person's life?'" If the answer is *yes* to both questions, most patients would want to undergo the intervention; if *no* to both, few, if any, would want to; a yes/no or no/yes mixed response would force a weighing of quality and quantity which would involve our being able to assess (among other things) quality of life (Kowalski et al 2008, 2012). What we will choose as the outcome variable in a clinical trial of a new drug product will depend upon what the new drug is intended to accomplish, but it seems clear that the two most important outcomes are length and quality of life (QoL). We take statins and/or blood pressure medications in the hope that they will extend life, but only if unwanted side effects are acceptable to achieve this. We take narcotics to relieve pain, but would be reluctant to take one that significantly shortened life expectancy. In order to make an informed choice about whether to take a newly proposed drug, we need to have information regarding *both* QoL and length of life extension, if any. The ethical imperative to gather and provide this information was noted by Levine (1996, p. 491):

"Patients do not consult physicians because they are offended by sphygmomanometer readings. They have a much broader view of what it means to secure their well-being now and in their personal futures. What they really want to know is this: 'What is Our life likely to be like if we take these drugs? What will we be able to do? How will we feel? And what if we don't?' These are, of course, questions about not only the quantity but also the quality of their lives." … "Research that neglects to develop information, upon which answers to such questions may be based, when the relevance of such questions can be foreseen, must be regarded as inadequate [in fulfilling ethical requirements for research]."

There are a number of ways to measure QoL (Kowalski et al, 2008, 2012), but if the intervention is meant also to extend life, there is a gold standard, and we are able to measure it:[3] According to Thompson and Temple (2001, p. 32), "For any intervention intended to extend life the primary question is are lives truly extended, and, if so, for how long? Any event short of death, no matter how important in itself in other ways, is a surrogate criterion."

It would seem, then, that for any medical intervention, including drugs, there is an ethical imperative to gather both QoL and mortality data. This is especially clear if the intervention is meant to lengthen life, but it is also important to ensure that new interventions do not unwittingly *shorten* life. There are many examples where adverse effects on mortality were not noted until the drugs were in general use and caused excess deaths among users. However, to include mortality as a primary outcome variable will, in many cases, require a long follow-up time, which in turn increases the costs associated with the trial, and the pharmaceutical industry has vigorously searched for ways to cut corners. They found that one effective way to contain costs is to use *surrogate* outcomes, outcomes that often consist of a laboratory measure or physical sign that is relatively easy to measure and can be expected to predict clinical outcomes. For example, Psaty et al (1999) noted that using lipid levels as an outcome variable, trials of lipid-lowering therapy typically include approximately 100 patients followed for 3 to 12 months. If the outcome is the incidence of cardiovascular events, the trials often require several thousands of patients followed for 4 to 5 years. The logic behind the approach seems reasonable: Suppose a risk factor causes morbidity and mortality, and the intervention reduces the risk factor. Then, the intervention will reduce the risk of morbidity and mortality. That this argument fails has been well-established (see, e.g., Fleming and DeMets, 1996, Fig 1, p. 606), but this has not had much effect on the use of surrogates; apparently, saving money trumps ensuring safety. We sketch below some thoughts on the uses and limitations of surrogate outcomes.

Surrogate Outcomes

The word *surrogate* is not an emotionally neutral descriptor (Wittes et al 1989, p. 415) thought that, "Arguing for a surrogate endpoint often entails a hint of disreputability, for the very word 'surrogate' evokes images of distorted motherhood),nor is there a single, universally accepted definition of what counts as a *surrogate outcome*. We begin, then, with a sampling of some of the definitions that have been proposed, showing the range of specificity of definition we have to work with. The first of these is from Wittes et al (1989, p. 416): "We define a surrogate endpoint as one that we elect to measure as a substitute for some other variable." This is an extremely loose, flexible definition, saying nothing about the election process and what properties we wish the substitute to have. Ellenberg and Hamilton (1989, p. 405) propose a definition in terms of the uses to which surrogates may be put: "Investigators use surrogate endpoints when the endpoint of interest is too difficult and/or expensive to measure routinely and when they can define some other, more readily measurable, endpoint, which is sufficiently well correlated with the first to justify its use as a substitute." This definition focuses on the correlation between the true and surrogate outcomes which, as will be seen below, is inadequate to "justify its use as a substitute."

A more descriptive definition was given in the preamble to the FDA's proposed accelerated approval rule (FDA, 1992): "A surrogate end point, or 'marker,' is a laboratory measurement or physical sign that is used in therapeutic trials as a substitute for a clinically meaningful end point that is a direct measure of how a patient feels, functions or survives, and is expected to predict the effect of therapy." This was discussed in detail by Schatzkin and Gail (2002) in the context of cancer research and Temple (1999) for cardiovascular drugs. The critical point here is that the surrogate "is expected to predict the effect of therapy" on the clinically meaningful end point. It is to be noted that *expected to predict* is not the same as *predicts*. Some more stringent requirements are considered next.

Prentice (1989, p. 432) defines a surrogate endpoint to be "a response variable for which a test of the null hypothesis of no relationship to the treatment groups under comparison is also a valid test of the corresponding null hypothesis based on the true endpoint." Schatzkin and Gail (2002) noted that this criterion could be checked by asking three questions (T is the true outcome, S the surrogate and E the experimental intervention or exposure): (i) Is S associated (correlated) with T?, (ii) Is E related to S?, and (iii) does S mediate the relationship between E and T (given S, are E and T independent)? Another way to describe this is that the S and T must not only be correlated, S "must fully capture the net effect of treatment on the clinical outcome" (Fleming and DeMets, 1996, p. 606). Given S, we must be able to accurately predict T.

This last set of definitions is the most demanding and the use of surrogates that could live up to these standards would cause few concerns. The problem is that to validate a surrogate to this extent would take a study as large and expensive as one involving the true target outcome.As noted by Schatzkin and Gail (2002, p. 26), "the large, long, expensive studies required to fully evaluate potential surrogates are precisely the studies that surrogates were designed to replace." The result has been that many surrogates have been used simply on the basis that they are correlated with the clinical outcome of interest.[4]However, as noted by Fleming and DeMets (1996, p. 605), "A correlate does not a surrogate make." This fact is widely appreciated, has been quoted by many, and in fact can be made even more emphatically: A *perfect* correlate does not a surrogate make. Baker and Kramer(2003) showed that no change in the surrogate outcome could correspond to either an increase or a decrease in the true outcome, and that an increase in the surrogate outcome could lead to a decrease in the true endpoint. In any case, there is no shortage of examples where surrogates have misled (e.g., Fleming and DeMets, 1966; Gotzsche et al, 1996; Psaty et al, 1999) and their use can rarely be justified. As stated by Fleming and DeMets (1996, p. 605):

"Surrogate end points can be useful in phase 2 screening trials for identifying whether a new intervention is biologically active and for guiding decisions about whether the intervention is promising enough to justify a large definitive trial with clinically meaningful outcomes. In definitive phase 3 trials, except for rare circumstances in which the validity of the surrogate end point has already been rigorously established, the primary end point should be the true clinical outcome."

4. Discussion

We have, in the above, argued against the use of placebo controls in the evaluation of second generation drugs, noted that there are no methodological penalties to be incurred if one adopts – in accordance with trial purpose – a non-inferiority trial design, and suggested that surrogate outcomes should rarely, if ever, be used in phase III trials. In doing so, we have relied on the extensive literature in these areas and pointed to convincing arguments for each of these points. The evidence seems to me to be overwhelmingly in favor of them. Still, placebo controls are often used in trials, along with surrogate outcome measures. Pharmaceutical companies design such trials and the FDA accepts their results in making licensing determinations. It's easy to see why pharmaceutical companies do this, and who pays the associated costs: As put by Rothman and Michels (1994, p. 396):

"The small placebo-controlled studies fostered by the FDA benefit drug companies, which can more easily obtain approval of an inferior drug by comparing it with placebo than they can by testing it against a serious competitor. Smaller studies are also cheaper. Unfortunately,the costs saved by the drug company are borne by patients, who receive placebos instead of effective treatments, and by the public at large, which is supplied with a drug of undetermined efficacy."

It is more difficult to see what stake the FDA has in all this,but it is impossible to ignore the possible effects of the Prescription Drug User Fee Act (PDUFA) according to which drug companies pay a fee for each new drug application in return for a speedier timetable for the new drug approval process. It might seem this in effect puts the FDA in the employ of drug industry (more than half the cost of reviewing new drug applications is funded by user fees from the drug industry). It is also possible that the FDA feels it lacks the regulatory authority to ban the use of placebos. O'Connor (2010, p. 979) noted, "These regulations can be interpreted as meaning that the FDA may not reject an application on the basis of efficacy if the treatment demonstrates a statistically significant benefit over placebo (justifying label claims of efficacy), *even if the new treatment appears to be less efficacious than established treatments or produces benefits of questionable clinical significance*"[Our italics]. This may be what the law says; but, if so, it should be changed to require reporting what is known about the comparative effectiveness of a new treatment in its label and marketing materials (Stafford et al, 2009). Over forty years ago, Stolley and Goddard (1970, p. 479) went even further: they proposed that Congress enact legislation to give the FDA authority to require that a new drug must be shown, before it is approved, to be "*as safe and more efficacious*than any drug presently on the market for the same indication." Since Congress sets the agenda for the FDA and provides for its operation, this seems to put the responsibility for change where it belongs. In any case, it seems clear that the FDA – whether or not they are in favor of any change – will go to considerable lengths to defend their guidelines with regard to placebos. They state:

"It is often possible to design a successful placebo-controlled trial that does not cause investigator discomfort nor raise ethical issues. Treatment periods can be kept short; early 'escape' mechanisms can be built into the study so that subjects will not undergo prolonged placebo-treatment if they are not doing well. In some cases randomized placebo-controlled therapy withdrawal studies have been used to minimize exposure to placebo or unsuccessful therapy; in such studies apparent responders to a treatment in an open study are randomly assigned to continued treatment or to placebo. Subjects who fail (e.g., blood pressure rises, angina worsens) can be removed promptly, with such failure representing a study endpoint."

Apparently, if none of these fixes can be applied, one will just have to live with investigator discomfort and the associated ethical concerns. Rothman and Michels (1994, p.396 - 7) retort to defenders of placebos follows:

"First, one can argue that withholding an accepted treatment may not lead to serious harm. For example, treating pain or nausea with a placebo may cause no long-term adverse effects, and the patient can call attention to any treatment failure or even chose to drop out of the study.

Nevertheless, although withholding an accepted treatment may occasionally seem innocuous, allowing investigators to do so runs counter to the ethical principle that every patient, including those in a control group, should receive either the best available treatment or a new treatment thought to be as good or better. Instead, it concedes to individual investigators and to institutional review boards the right to determine how much discomfort or temporary disability patients should endure for the purpose of research. Ethical codes in medical experimentation have been developed expressly to shield patients from such vulnerability."

The major counter to such concerns is that placebo-controlled studies can take less time to complete, thus allowing beneficial drugs to reach the market earlier, which is important to consumers. Pharmapologists[5]are quick to point to AIDS activists' insistence that drugs be made available in a timely fashion, especially when these drugs are potentially life-saving and there are no alternatives. There is no gainsaying such arguments when in fact there are *no alternatives*, but it is difficult to defend speed-at-all-costs to gain approval for another statin or blood pressure control medication. Indeed, it would seem prudent to insist on assurances that new drugs of this type are at least as effective, and at least as safe, as those already on the market. If we want to conduct trials of this kind, it needs to be realized that they may well be expensive. Zivin (2000) noted that trials can be described in three ways: *trustworthy*, *fast*, or *cheap*; and that a given trial can have only two of these characteristics. To insist that they be both fast and cheap is to rule out the most important of these characteristics. We need to be able to trust that new medications do more good than harm, and that they are at least as good as those already available. With apologies, we turn to a paraprosdokian (Wikipedia that!): It may be true that the early bird gets the worm, but it's the second mouse that gets the cheese. Another way to say *speed can kill* was suggested by Fleming (2005, p. 77):

"Why is it in patients' best interest to have more drugs from which to choose, if there are less reliable insights to guide their caregivers and themselves in making these choices? And why is it in patients' best interest to have earlier access to biologically active interventions, if these therapies may be inconvenient to receive, costly, and potentially more toxic than effective? And might earlier access to ineffective treatments delay or chill the development and proper testing of other interventions that really do work?"

So, what should we do? Many have suggested various reforms for the pharmaceutical industry (the books cited earlier all have their own lists; see also Califf, 2002; O'Connor, 2010; Ray and Stein, 2006; Stafford et al, 2009; Strom, 2006), but most of these have concentrated on strengthening the FDA, or by creating new, independent agencies for certain related tasks, so that they can "crack down." Some of these reforms could work, e.g., the current climate suggests that the ten conflict-of-interest action points outlined by Kassirer (2005, p. 211) might be feasible, but they will be vigorously resisted by PhRMA and its arOur of lobbyists – there is more than one lobbyist for each member of the House and Senate (Abramson, 2004, p. 90).This resistance, of course, does not mean that these proposals are not worthwhile, but we suggest that we choose our battles wisely. Rather than simply adopting an adversarial approach, we should seek a partnership that acknowledges the accomplishments of the pharmaceutical industry,[6] while recognizing that much remains to be done, and that many current practices can be improved. Regulatory reform is vitally needed but will, in-and-of-itself, prove inadequate to ensure that the drug approval process protects important public health concerns. There may be a way to cooperate with the pharmaceutical companies to achieve a result that will be acceptable to both parties. A carrot might be better than confrontation. Just such an approach was taken by Wood (2006).

Wood (2006, p. 619) proposed a number of changes to the drug approval process designed to provide incentives to, among other things,(i) demonstrate a drug's long-term safety, (ii) perform head-to-head drug-comparison studies,(iii) convert end points of surrogate markers or biologic markers to clinically meaningful ones, and (iv) encourage drug development with high commercial risk. The second and third of these items have been the explicit target of this paper. The first is also relevant to Our arguments in that the shorter, smaller trials made possible by the use of placebo controls and surrogate outcomes, preclude obtaining longer-term safety data; data which is of obvious public health importance. The last point also fits in: If we want to encourage real innovation, to spur development of drugs for which there is great need, we need to provide incentives to get away from the less risky, but currently more profitable, strategy of developing more and more drugs for a given condition by simply showing new formulations to be better than placebo. The ultimate goal of Wood's reforms "is to reward true, high-risk innovation that improves medical care … in contrast to our current system, which rewards duplicative, relatively low-risk drug development and encourages the use of new, expensive, heavily marketed drugs at the expense of older, equally effective drugs of the same pharmaceutical class." The changes proposed by Wood might actually be acceptable to the pharmaceutical industry since they are all based on providing incentives (in the form of extended exclusivity – the right to sell a drug without competition from manufacturers of generic drugs) to do things the right way. For example, to encourage the generation of long-term safety data, an extended period of exclusivity would be granted to drug manufacturers after they had completed FDA-approved studies demonstrating safety. This privilege would depend on the FDA approving the study design, including the choice of comparator and the margin of safety required. Failure to successfully complete safety studies in a timely matter would result in the loss of extended exclusivity. This approach has the virtue of simplicity and does not require the establishment of new agency. It would, however, depend on strengthening the FDA and assuring its independence from industry influences.

Wood's recommendation is close to one previously made in the area of cardiovascular research by Psaty et al (1999). They suggested, "a requirement for the regular use of phase 4 trials in the approval of new drug therapies for cardiovascular risk factors … These required phase 4 studies should be large, long-term clinical trials designed to assess the effects of drug therapies on major disease end points over 3 to 5 years" (p. 789). Wood believes that providing incentives for conducting these phase 4 trials is more likely to succeed than by making them a requirement; but there can be no denying that these trials have vital public health importance. The assessment of adverse effects in large trials over extended periods is necessary to detect rare[7] and/or delayed problems. This is sometimes expressed (e.g., Temple, 1999, p. 792) as "there is no surrogate for safety."

In any event, all four of Wood's reforms are important and doable if we accept that doing things properly will cost more money and take more time. Financial support is available from a combination of industry incentives and Congress recognizing that it has the responsibility to ensure that the FDA has the resources (and legal mandates) it needs to do its job. Patience may be more difficult to achieve in a society bent on progress, especially when one is faced with a diminishing expected time of survival with no prospect for effective intervention. Accelerated approval is already a possibility in such cases, but these arise in only a small fraction of the industry's total output, and there is no shortage of examples of serious harms that have occurred because safety concerns were shortchanged. We need to recognize that all drug approval is provisional and take whatever steps necessary to acquire the data needed to make informed risk/benefit assessments in clinical decision making contexts. This will take time; but some things are worth waiting for.

References

[1] Abramson, J. 2004. Overdo$ed America: The broken promise of American Medicine. New York: Harper.

[2] Altman, D.G., and J.M. Bland. 1995. Absence of evidence is not evidence of absence. British Medical Journal 311: 485.

[3] Anderson, J.A. 2006. The ethics and science of placebo-controlled trials: Assay-sensitivity and the Duhem-Quine thesis. Journal of Medicine and Philosophy 31: 65-81.

[4] Anderson, J.A. 2009a. Who's in control of the choice of control? American Journal of Bioethics-Neuroscience 9: 60-2.

[5] Anderson, J.A. 2009b. Contextualizing clinical research: the epistemological role of clinical equipoise. Theoretical Medicine and Bioethics 30: 269-88.

[6] Angell, M. 2005.The Truth about the Drug Companies: How they deceive us and what to do about it. New York: Random House Trade Paperbacks.

[7] Avorn, J. 2005. Powerful Medicines: The benefits, risks, and costs of prescription drugs. New York: Random House.

[8] Baker, S.G., and B.S. Kramer. 2003. A perfect correlate does not a surrogate make. BMC Medical Research Methodology 3: 16- 20. Available from: http://www.biomedcentral.com/1471-2288/3/16

[9] Bero, L.A., and D. Rennie. 1996. Influences on the quality of published drug studies. International Journal of Technology Assessment in Health Care 12: 209-37.

[10] Brody, H. 2007. Hooked: Ethics, the Medical Profession, and the Pharmaceutical Industry. Lanham MD: Rowman & Littlefield.

[11] Brody, H. 2011. Clarifying conflict of interest. American Journal of Bioethics 11: 23-8.

[12] Califf, R.M. 2002. The need for a national infrastructure to improve the rational use of therapeutics. Pharmacoepidemiology and Drug Safety 11: 319-27.

[13] D'Agostino, R.B., J.M. Massaro, and L.M. Sullivan. 2003. Non-inferiority trials: design concepts and issues – the encounters of academic consultants in statistics. Statistics in Medicine 22: 169-86.

[14] De George, R.T. 2009. Two cheers for the pharmaceutical industry. In: Ethics and the Business of Biomedicine, D.G. Arnold, ed. Cambridge University Press. pp 69-97.

[15] Ellenberg, S.S., and J.M. Hamilton. 1989. Surrogate endpoints in clinical trials: Cancer. Statistics in Medicine 8: 405-13.

[16] Ellenberg, S.S., and R. Temple. 2000. Placebo-controlled trials and active-control trials in the evaluation of new treatments. Part 2: Practical issues and specific cases. Annals of Internal Medicine 133: 464-70.

[17] Enkin, M.W. 2009. Questioning the methodological superiority of 'placebo' over 'active' controlled trials. American Journal of Bioethics – Neuroscience 9: 66-7.

[18] FDA. 1992. New drug, antibiotic and biological drug product regulations: accelerated approval. Proposed Rule. 57 Federal Register 13234-13242.

[19] Fergusson, D., K.C. Glass, D. Waring, and S. Shapiro. 2004. Turning a blind eye: The success of blinding reported in a random sample of randomized, placebo controlled trials. BioMedical Journal 328: 432.

[20] Fleming, T.R. 2005. Surrogate endpoints and FDA's Accelerated Approval process. Health Affairs 24: 67-78.

[21] Fleming, T.R. and D.L. DeMets. 1996. Surrogate endpoints in clinical trials: Are we being misled? Annals of Internal Medicine 125: 605-13.

[22] Freedman, B., C. Weijer and K.C. Glass. 1996. Placebo orthodoxy in clinical research: Empirical and methodological Ourths. Journal of Law, Medicine & Ethics 24: 243-51.

[23] Freedman, B., K.C. Glass, and C. Weijer. 1996. Placebo orthodoxy in clinical research: Ethical, legal, and regulatory Ourths. Journal of Law, Medicine & Ethics 24: 252-9.

[24] Goozner, M. 2004. The $800 Million Pill: The truth behind the cost of new drugs. Berkeley: University of California Press.

[25] Gottlieb, S.S. 1997.Dead is dead – artificial definitions are no substitute. Lancet 349: 662-3.

[26] Gotzsche, P.C., A. Liberati, V. Torri, and L. Rossetti. 1996. Beware of surrogate outcome measures. International Journal of Technology Assessment in Health Care 12: 238-46.

[27] Heres, S., J. Davis, M. Katja, E. Jetzinger, W. Kisserling, and S. Leucht. 2006. Why olanzapine beats risperidone, risperidone beats quetiapine, and quetiapine beats olanzapine: An exploratory analysis of head-to-head comparison of second-generation antipsychotics. American Journal of Psychiatry 163: 185-94.

[28] Hill, A.B. 1963. Medical ethics and controlled trials. British Medical Journal 1: 1043-9.

[29] Howick, J. 2009. Questioning the methodologic superiority of 'placebo' over 'active' controlled trials. American Journal of Bioethics-Neuroscience 9: 34-48.

[30] Hrobjartsson, A., E. Forfang, M.T. Haahr, B. Als-Nielsen, and S. Brorson. 2007. Blinded trials taken to the test: An analysis of randomized clinical trials that report tests for the success of blinding. International Journal of Epidemiology 36: 654-63.

[31] International Conference on Harmonization (ICH). 2000. Harmonized tripartite guideline. Choice of control group and related issues in clinical trials. International Conference on Harmonization, E10. San Diego CA: Centre for Biologics Evaluation and Research.

[32] Kassirier, J.P. 2005. On the Take: How medicine's complicity with big business can endanger your health. Oxford University Press.

[33] Khan, A., S. Khan, and W.A. Brown. 2002. Are placebo controls necessary to test new antidepressants and anxiolytics? International Journal of Neuropsychopharmacology 5: 193-7.

[34] Kowalski, C.J. 2010. Pragmatic problems with clinical equipoise. Perspectives in Biology and Medicine 54: 161-73.

[35] Kowalski, C.J., J. Bernheim, N.A. Birk, and P. Theuns. 2012. Felicitometric hermeneutics: Interpreting quality of life measurements. Theoretical Medicine and Bioethics 33: 207-20.

[36] Kowalski, C.J., S. Pennell, and A. Vinokur. 2008. Felicitometry: Measuring the 'quality' in quality of life. Bioethics 22: 307-13.

[37] Levine, R.J. 1996. Quality of life assessments in clinical trials: An ethical perspective. In: Quality of Life and Pharmacoeconomics in Clinical Trials, B. Spilker, ed. Philadelphia: Lippincott-Raven. 489-95.

[38] Miller, F.G., and H. Brody. 2005. Professional integrity in industry-sponsored clinical trials. Academic Medicine 80: 899-904.

[39] O'Connor, A.B. 2010. Building comparative efficacy and tolerability into the FDA approval process. JAMA 303: 979-80.

[40] Petersen, M. 2008. Our Daily Meds. New York: Picador.

[41] Piaggio, G., D.R. Elbourne, D.G. Altman, S.J. Pocock, S.J.W Evans for the CONSORT Group. 2006. Reporting of non-inferiority and equivalence randomized trials: An extension of the CONSORT statement. JAMA 295: 1152-60.

[42] Prentice, R.L. 1989. Surrogate endpoints in clinical trials: Definition and operational criteria. Statistics in Medicine 8: 432-40.

[43] Psaty, B.M., and N.S. Weiss. 2007. NSAID trials and the choice of comparators – Questions of public health importance. NEJM 356 (4): 328-30.

[44] Psaty, B.M., N.S. Weiss, C.D. Furberg, et al. 1999. Surrogate end points, health outcomes, and the drug-approval process for the treatment of risk factors for cardiovascular disease. JAMA 282: 786-90.

[45] Ray, W.A., and C.M. Stein. 2006. Reform of drug regulation – Beyond an independent drug-safety board. NEJM 354: 194-201.

[46] Rennie, S., and T. Sturmer. 2009. Strengthening Howick's argument against the alleged superiority of placebo-controlled trials. American Journal of Bioethics – Neuroscience9: 62-4.

[47] Rothman, K.J. 1986. Significance questing. Annals of Internal Medicine 105: 445-7.

[48] Rothman, K.J., and K.B. Michels. 1994. The continuing unethical use of placebo controls. NEJM 331: 394-8.

[49] Sackett, D.L., W.S. Richardson, W. Rosenberg, and R.B. Haynes. 1997. Evidence-based Medicine. New York: Churchill Livingstone.

[50] Safer, D.J. 2002. Design and reporting modifications in industry-sponsored comparative psychopharmacology trials. Journal of Nervous and Mental Disease 190: 583-92.

[51] Schatzkin, A., and M. Gail. 2002. The promise and peril of surrogate end points in cancer research. Nature Reviews Cancer 2: 19-27. Available at www.nature.com/reviews/cancer

[52] Schwartz, D., R. Flamant, and J. Lellouch (translated by M.J.R. Healy). 1980. Clinical Trials. New York: Academic Press.

[53] Silverman, M., and P.R. Lee. 1974. Pills, Profits & Politics. Berkeley: University of California Press.

[54] Stafford, R.S., T.H. Wagner, and P.W. Lavori. 2009. New, but not improved? Incorporating comparative-effectiveness information into FDA labeling. NEJM 361: 1230-3.

[55] Stolley, P.D., and J.L. Goddard. 1970. A "relative efficacy" system for new drugs. Annals of Internal Medicine 73: 479-80.

[56] Strom, B.L. 2006. How the US drug safety system should be changed. JAMA 295: 2072-5.

[57] Temple, R.J. 1995. A regulatory authority's opinion about surrogate endpoints. In: Clinical Measurement in Drug Evaluation, W.S. Nimmo and G.T. Tucker, eds. New York: Wiley.

[58] Temple, R. 1999. Are surrogate markers adequate to assess cardiovascular disease drugs? JAMA 282: 790-5.

[59] Temple, R., and S.S. Ellenberg. 2000. Placebo-controlled trials and active-control trials in the evaluation of new treatments. Part 1: Ethical and scientific issues. Annals of Internal Medicine 133: 455-63.

[60] Thompson, A., and N.J. Temple, eds. 2001. Ethics, Medical Research, and Medicine. Dordrecht: Kluwer.

[61] Weijer, C. 1999. Placebo-controlled trials in schizophrenia: Are they ethical? Are they necessary? Schizophrenia Research 35: 211-8.

[62] Wittes, J., E. Lakatos, and J.Probstfield. 1989. Surrogate endpoints in clinical trials: Cardiovascular diseases. Statistics in Medicine 8: 415-25.

[63] Wood, A.J.J. 2006. A proposal for radical changes in the drug-approval process. NEJM 355: 618-23.

[64] Zivin, J.A. 2000. Understanding clinical trials. Scientific American. 282: 69-75.

[1] Context makes it clear that Freeman et al define 'clinical effectiveness' as an *improvement* in the currently available armamentarium. Their opinion of the "better than nothing" threshold for effectiveness is clear from their statement that, "flinging a glass of water at a burning building is an effective fire-fighting strategy because it is better than nothing."

[2] Donald Rumsfeld even got into the act on this one, in the context of Iraq's weapons of mass destruction, stating "There's another way to phrase that and that is that the absence of evidence is not the evidence of absence. It is basically saying the same thing in a different way. Simply because you do not have evidence that something does exist does not mean that you have evidence that it doesn't exist."

[3] Wittes et al (1989, p. 419) noted that, "Well-run adequately funded clinical trials can usually ascertain the vital status of over 99 per cent of the participants. Therefore one can measure all-cause mortality without bias." This stands in contrast to surrogate outcomes which often suffer from informative censoring. Also see Gottlieb (1997).

[4] Simple correlation is far from sufficient. As noted by Schwartz et al (1980, p. 51), "in a trial of weight-reducing diets, it will scarcely do to replace weight as a criterion by the closely associated measurement, height."

[5] We first ran into this term in Brody (2011, p. 23). Abramson, Angell, Avorn, et al would be termed "pharmascolds."

[6] Our sentiments toward big PhRMA's accomplishments are nicely captured by the title of a paper by De George (2009), "Two cheers for the pharmaceutical industry." We should recognize that drugs are now available that save, prolong, and enhance life; and favorably impact the overall cost of healthcare by keeping people at work and out of the hospital. That third cheer, however, is yet to be realized.

[7] The rule of three (Sackett et al,1997, p. 107) provides some insight here: If a drug causes an adverse reaction once in every N persons who take it, then to be 95% certain to see it at least once you need to follow 3N subjects. It may also be helpful to know the converse: If you don't see any adverse events among N subjects, a one-sided 95% confidence interval for the probability of an adverse event is approximately (0, 3/N).