Considerations for Comparing a Test Drug With Standard of Care in Phase 2 Clinical Trials of Central Nervous System Disorders

Peer-reviewed author version

Considerations for Comparing a Test Drug with Standard of Care In Phase II Clinical Trials of Central Nervous System Disorders

Craig H. Mallinckrodt, Ph.D.[1], Michael J. Detke, M.D., Ph.D.[1,2],

William R Prucka, Ph.D.[1], Stephen J. Ruberg, Ph.D.[1],

and Geert Molenberghs, Ph.D.[3]

1. Eli Lilly & Co., Lilly Corporate Center, Indianapolis, IN 46285

2. Departments of Psychiatry, Harvard Medical School, Boston MA; McLean Hospital, Belmont MA; and, Indiana University School of Medicine, Indianapolis, IN

3. I-BioStat, Hasselt University, , Diepenbeek, Belgium, and, Katholieke Universiteit Leuven, Leuven, Belgium.

Abstract

An increasing need exists to understand the risks and benefits of a test drug compared with standard of care (SoC) earlier in development. Even if a drug is superior to placebo, it may not be worthwhile to continue development unless it has advantages over SoC. However, efficacy comparisons versus SoC in early phase studies can be challenging. Simulation studies were conducted to illustrate that in common scenarios simply randomizing a few patients to SoC will frequently yield misleading results. It may take samples sizes at least 5-fold greater to achieve reliable comparisons of a test drug with SoC than when comparing versus placebo. Therefore, it is important that the rate of false positive and false negative results be quantitatively evaluated before determining the sample size and the criteria upon which the test drug and SoC will be compared. Because test drugs often have no benefit, comparing a test drug with SoC may unnecessarily use resources that could be devoted to investigating other drugs. Moreover, it can be difficult to construct valid comparisons of a test drug versus SoC without experience with the test drug regarding appropriate dosing, patient population, etc. An example with actual clinical trial data is used to illustrate how the trade-off between the need to and the difficulties in comparing a test drug with SoC in Phase II can be mitigated using a literature database of placebo-controlled studies to construct an historical comparison.

Key words: Clinical trial, Active Comparator

Introduction

The goals of Phase II development has traditionally included: exploring use for the targeted indication (establishing Proof-of-Concept, PoC) and estimating dosage for subsequent studies (dose-response) [1, 2]. One of the changes in recent years in the health care industry is that key stake holders are demanding evidence not only that drugs are safe and effective, but that they are safer or more effective than alternatives. Therefore, greater need exists to understand the risks and benefits of a test drug compared with a standard of care (SoC). Ideally, this information would be available early in development. For example, even if a test drug were superior to placebo in a proof of concept (PoC) study, it may not be worthwhile to continue developing the drug unless it also provides benefit beyond a currently available cheaper, generic therapy.

However, efficacy comparisons versus SoC in a PoC study can be challenging[3,4]. For example, reliably establishing a difference between a test drug and SoC may require an appreciably larger sample size to achieve a given level of power than when comparing the test drug with placebo because the SoC is superior to placebo.

The larger sample size required to compare a test drug with SoC has important implications for PoC studies. Given that many drugs tested in PoC studies will have no benefit, that is, they are not better than placebo; does it make sense to do a big trial to compare the test drug with SoC when a small study would show the test drug was not different from placebo? Alternatively, does it make sense to do a smaller study to

compare with placebo when SoC is the benchmark that really matters? Is it useful to include some patients in an active comparator arm even if that arm were not adequately powered for a reliable comparison?

These are questions routinely faced in drug development today. The purpose of this paper is to illustrate a framework for addressing these questions. The intent is not to provide specific recommendations for specific scenarios. Rather, the focus is on establishing a framework that provides the bases for addressing the questions.

Developing a Quantitative Framework

It may be temping to believe that having some data is better than no data; that, all else equal, including even a few patients in an active comparator arm in a PoC trial is useful for benchmarking the test drug versus the SoC. But that may not necessarily be the case because an unreliable comparison can be a misleading comparison, which may not be better than no comparison. And, all else is not equal because including an active comparator adds cost, time, and potentially logistic or methodological difficulties to the study.

A useful way to approach this question is to consider how reliable the comparison with SoC must be for it to be useful. How many patients should be included in the test drug and SoC arms of the trial to have an acceptable rate of false positive and false negative findings?

Definitive comparisons based on a hypothesis test for superiority of the test drug versus SoC may be prohibitively large for a PoC study. Therefore, alternative decision making criteria may need to be evaluated. Consider the following example based on simulated data. The intent is not to specifically advocate a certain sample size or a specific criteria, but rather to illustrate general principles.

Characteristics of a simulation study for a first scenario were patterned after schizophrenia clinical trials. An effect size of 0.6 is generally considered moderate, and approximately equal to the effect size seen from a recent meta-analysis of atypical antipsychotics [5]. The test drug was hoped to be superior to SoC by a margin of 0.20 effect size. This effect size would represent a relative gain over SoC of 33% (0.2 / 0.6).

In scenario 1A, the effect size for the test drug versus placebo was 0.80, and the effect size for the SoC versus placebo was 0.60, with the effect size for test drug versus SoC equal to the hoped for advantage of 0.20. In scenario 1B, the test drug was in truth not different from the SoC as the test drug and SoC each had an effect size versus placebo of 0.60.

With an effect size of 0.8, given certain other assumptions about dropout rates and temporal profiles, typical of neuroscience research, a sample size of 40 per arm would yield approximately 90% power for the contrast between the experimental treatment and placebo.

Tables 1 and 2 summarize the results of simulations that assessed the operational characteristics of various comparisons between the experimental drug and SoC at varying sample sizes. The tables include three criteria that may be used to determine whether further development of the test drug is warranted based on how the test drug compares with SoC. These criteria include: 1) a superiority test using alpha = .05; 2) a non-inferiority test where non inferiority is declared if the lower bound of the 95% confidence interval is greater than an effect size of -0.2. This lower bound is chosen because it preserves 2/3 of the advantage of the active drug over placebo. In other words, if the test drug is no worse than having an effect size 0.2 less than the SoC, it is concluded that the test drug is not inferior to the SoC; and, 3) a rank test for whether or not the test drug is numerically superior to SoC. (Test drug has greater mean change than SoC by any amount.) A fourth comparison is also included in the tables to assess whether or not the estimated superiority of the test drug falls within the range of one half of the hoped for advantage (1/2 * 0.2 = 0.1) to twice the hoped for advantage (2 * 0.2 = 0.40). This comparison is included merely to assess how often the estimated treatment difference fell within a range close to the hoped for difference. As such, it is an intuitive assessment of whether or not the study yielded a reliable result.

In scenario 1A, the experimental drug was superior to SoC; therefore, the rates in Table 1 are the frequency of correct positive results. Recall that N = 40 per arm yields approximately 90% power for the test drug versus placebo if the test drug has the hoped for effect. Therefore, 40 per arm would be adequate to compare the test drug with placebo. This sample size resulted in correctly ranking the treatments in 88% of the

simulated trials. With 100 per arm, non-inferiority was established in over 95% of the simulated trials. With 200 per arm, only about 72% of the trials established superiority.

Table 1. Simulation results when the true difference between test drug and SoC is an effect size of 0.20.

N/arm	Power Superiority	Power Non-inferiority	Estimate with in range (%)[1]	Ranked Correctly (%)
40	21.1	64.8	61.0	88.1
100	42.1	95.1	76.9	96.7
200	72.4	100	88.0	99.5

[1] Estimated effect size for advantage of test drug over SoC is between of 0.1 and 0.4, which is 1/2x – 2x the hoped for advantage of a 0.20 effect size. .

Validate, move to production, and rerun pgm SOC_1. Update results

In scenario B, the experimental drug was not different from SoC; therefore, the rates in Table 2 are the frequency of false positive results. Regardless of sample size, superiority testing yielded false positive results at approximately the 2.5% rate that was expected from using a 5% level of significance and a two-tailed test.

It was still possible to establish non-inferiority since the two treatments were in fact equal. With 200 per arm, non-inferiority was established in over 78% of the simulated trials. This demonstrates a difficulty in using non-inferiority as the basis for continuing development when superiority is the goal. Namely, if the test drug is not really any better than SoC, non-inferiority tests can lead to many false positive results. Moreover, the larger the sample size, the more likely the false positive result.

Using rankings also leads to many false positive results. The probability of getting a point estimate of exactly zero difference between the experimental drug and SoC is

negligible, and each drug will be ranked as better than the other about half the time. Of course, in a real situation, the true difference is not known and basing continued development on having the test drug be numerically superior to SoC would lead to a false positive rate of 50% regardless of sample size.

Interestingly, the estimated advantage of the experimental drug fell within the hoped for range, in over 26% of the simulated trials with N = 40, and with 200 per arm rate was almost 10%. In other words, when there was no difference between drugs, there was still an appreciable probability that the estimated difference was consistent with the hoped for difference, but this probability decreased with increasing sample size.

Table 2. Simulation results when test drug and SoC are equal with the hoped for advantage over SoC is an effect size of 0.20.

N/arm	Power Superiority	Power Non-inferiority	Estimate with in range (%)[1]	Ranked Correctly (%)
40	2.5	23.2	26.6	NA
100	2.8	47.4	16.6	NA
200	2.4	78.6	8.8	NA

[1] Estimated effect size for advantage of test drug over SoC is between of 0.1 and 0.4, which is 1/2x – 2x the hoped for advantage of a 0.20 effect size. .

Validate, move to production, and rerun pgm SOC_1. Update results

Consider another set of scenarios. The same general parameters were input into the simulations except effect sizes were smaller, mimicking scenarios in depression where a recent meta-analysis showed the average effect size to be 0.31 [6]. In scenario 2A, the effect size for the test drug versus placebo was 0.40, and the effect size for the SoC

versus placebo was 0.30, with the effect size for test drug versus SoC equal to the hoped for advantage of 0.10. In scenario 1B, the test drug was in truth not different from the SoC as the test drug and SoC each had an effect size versus placebo of 0.30.

With an effect size of 0.4, given certain other assumptions about dropout rates and temporal profiles typical of neuroscience research, a sample size of 100 per arm would yield approximately XX% power for the contrast between the experimental treatment and placebo. In scenario 2B (Table 4), the test drug was not different from the SoC, but the test drug was hoped to be superior to SoC by a margin of 0.10 effect size.

Given that in this alternative scenario a smaller effect size was to be detected, the same sample size resulted in poorer operational characteristics, with lower rates of correct positive results and higher rates of false positive results. Even with 500 per arm, there was only about 50% power for a superiority test. Again, increasing sample size when basing decisions on non-inferiority increased the rate of false positive findings. And basing decisions on rankings again yielded a 50% false positive rate if the experimental drug were no better than SoC.

Interestingly, it took 500 per arm to yield an 80% probability that the point estimate fell within the hoped for range when a difference existed. However, the point estimate fell within the range in nearly 14% of the simulated trials when no difference existed.

Table 3. Simulation results when the true difference between test drug and SoC is an effect size of 0.10.

N/arm	Power Superiority	Power Non-inferiority	Estimate with in range (%)[1]	Ranked Correctly (%)
200	25.8	98.2	65.6	89.8
500	51.7	100	81.8	97.5

[1] Estimated effect size for advantage of test drug over SoC is between of 0..05 and 0.2, which is 1/2x – 2x the hoped for advantage of a 0.10 effect size.

Validate, move to production, and rerun pgm SOC_2. Update results

---

Table 4. Simulation results when test drug and SoC are equal with the hoped for advantage over SoC is an effect size of 0.10.

N/arm	Power Superiority	Power Non-inferiority	Estimate with in range (%)[1]	Ranked Correctly (%)
200	2.4	78.6	25.1	NA
500	2.8	98.8	13.9	NA

[1] Estimated effect size for advantage of test drug over SoC is between of 0..05 and 0.2, which is 1/2x – 2x the hoped for advantage of a 0.10 effect size. .

Validate, move to production, and rerun pgm SOC_2. Update results

---

Obviously, many scenarios could be considered, along with many possible criteria for comparing a test drug versus SoC to determine if further development is warranted. Therefore, the main point of the examples above is not the specific results. Rather, the focus is on the general point that comparisons versus SoC take much larger studies than comparisons with placebo. In these examples, non-inferiority testing and rankings yielded unacceptably high rates of false positive findings at all sample sizes and that superiority testing yielded unacceptably high rates of false negative results even when sample sizes were 5-fold greater than what was needed to compare versus placebo.

A Portfolio Perspective

Consider the following hypothetical scenario where research and development costs are thought of as the opportunity to buy outcomes, with only a fixed amount that can be spent. Obviously, various strategies might be leveraged to buy outcomes more efficiently, but ultimately only so many outcomes can be bought. Further, assume that some outcomes are more expensive than others. For example, in the simulation studies above the sample size required to obtain a reliable contrast of the experimental drug versus SoC was at least 5-fold greater than contrasts versus placebo. Given certain costs associated with clinical trials are fixed regardless of sample size, assume contrasts versus SoC are 3-fold more costly than contrasts versus placebo. In other words, assume placebo outcomes cost 1 unit and SoC outcomes cost 3 units. Also assume that the research budget allows purchase of 20 outcome units.

If all PoC studies contain SoC, it would cost 4 units to evaluate each compound (1 unit for the placebo outcome and 3 units for the SoC outcome). Hence, 5 compounds could be evaluated in total. If no PoC studies contained SoC, 12 compounds could be initially screened versus placebo, costing 12 units, for example. Then, assuming two compounds were positive they would then be evaluated versus SoC, costing 8 units. In this hypothetical situation, would it be better to rigorously compare 5 compounds versus placebo and SoC, or would it be better to preliminarily evaluate 12 drugs versus placebo and only proceed to comparing versus SoC for those compounds that beat placebo, or is some combination of the approaches optimal?

Answering this question depends on many factors and a full discussion is beyond our present scope. However, the optimal solution is likely strongly influenced by the probability that the test drug is effective. For example, for test drugs with proven mechanisms of action, it is more likely that the compounds have some benefit and thus first testing versus placebo would not screen out many compounds. And there is likely greater need to compare versus SoC with the already proven compounds of the same mechanism as early as possible. Conversely, for test drugs with novel mechanisms, it is less likely that they will have any beneficial effect and a small trial versus placebo will screen out many compounds. Moreover, if the novel test drug happens to beat placebo, it is likely to differ in some meaningful way from the SoC,because it has a different mechanism. The key point is that comparisons versus SoC are costly and in a resource-constrained environment will result in the ability to evaluate fewer drugs.

One potential method to mitigate the trade off that needing to compare with SoC results in the ability to evaluate fewer drugs could be to obtain information versus SoC from sources other than a concurrent control in a PoC study. The following section provides a real data example of using placebo-controlled studies of an experimental drug and comparing the results from earlier placebo-controlled studies of the SoC to benchmark an experimental drug versus SoC.

Using a Literature Database to Benchmark Versus SoC

Prior to approval of duloxetine for major depressive disorder, 11 clinical trials were conducted that included 15 treatment arms of duloxetine tested versus placebo. In 7 of

these studies, which included 11 duloxetine treatment arms, a positive control (SSRI) was also included. Among these 7 SSRI arms, 5 had equal randomization to duloxetine and 2 had half as many patients as duloxetine. These studies have been published individually [7-14] and in summaries [15,16,] with additional details being available at Lillytrials.com [17].

If each duloxetine arm is viewed as a PoC trial, the value of the active comparator in regards to benchmarking versus SoC can be compared to what would have been inferred based on an historical control. In other words, duloxetine can be compared head-to-head versus the SSRI and comparisons can be inferred by comparing the advantage of duloxetine over placebo to the historical advantage of selective serotonin reuptake inhibitors (SSRIs), the standard of care, over placebo.

It is important to realize that historical advantage in this context is different than the historical control as often described. Historical controls are often based on single-arm studies, comparing the response rate from a test drug to the response rate of a known effective drug[1, 18]. This is historical control based on uncontrolled studies. In the present context, historical comparisons versus SoC are based on placebo-controlled studies.

Use of historical control trades bias for precision. The potential biases of non head-to-head comparisons in drug development, such as using historical controls as opposed to concurrent controls, are well known and understood[1]. In situations such as clinical trials in MDD, where placebo response is highly variable [19-23] historical controls may be

assumed to not be useful for definitive testing. But does that mean historical controls are useless?

In MDD, SSRIs have been tested extensively, so the effect size is essentially known. From the historical data, the average effect size of SSRI versus. placebo = 0.31 [6]. The advantage of duloxetine versus SSRI was estimated to be an effect size of 0.11 [16]. For the sake of these retrospective comparisons, it can be assumed that this difference is the true advantage over SoC, or because the evidence for superiority of duloxetine was not definitive it can be assumed that the true difference is 0. Results from the duloxetine studies are summarized in Table 5.

First, consider the SSRI effect sizes and the duloxetine effect sizes from the studies that had an SSRI arm. The unweighted average SSRI effect size was 0.291, very close to the true value of 0.31. However, the range in results was 0.09 to 0.637 and in only one of the 7 studies was the observed SSRI effect size within ± 0.10 of the true value of 0.31. If assuming duloxetine is not different from SSRI, 4/11 concurrent control SSRI contrasts were within ± 0.10 of the true effect size of 0, whereas 6/11 historical contrasts fell within the same interval, and historical control was closer to the true value than concurrent control in 8/11 contrasts.

If assuming duloxetine is different from SSRI by the 0.11 estimated from the overall pooled data, 6/11 concurrent control SSRI contrasts were within ± 0.10 of the true effect size of 0.10, whereas 4/11 historical contrasts fell within the same interval. This

comparison is biased in favor of the concurrent contrasts because of the assumption that the average concurrent contrast is the true difference. However, even with this bias, historical control was closer to the true value than concurrent control in 6/11 contrasts..

Duloxetine effect sizes were greater on average in the 4 studies without an active comparator. Papakostas and Fava [20] reported that studies having more patients randomized to placebo led to greater drug-placebo differences. Based on historical control from the 4 studies that did not include an active comparator the difference between duloxetine and SSRI would be an effect size of about 0.21. This result suggests two important points. First, when trading bias for precision in using historical controls, efforts to control bias, such as matching on key trial design features, can be beneficial. And, sensitivity of results to various approaches should be evaluated.

Table 5. Effect sizes for duloxetine and SSRI

Study	Duloxetine dose	Duloxetine Effect size	SSRI Effect size	Concurrent Control Difference	Historical Control Difference
HMAT-A	40mg	0.249	0.467	-0.218	-0.061
	80mg	0.272	0.467	-0.195	-0.038
HMAT-B	40mg	0.378	0.191	0.187	0.068
	80mg	0.564	0.191	0.373	0.254
HMAY-A	80mg	0.49	0.637	-0.147	0.18
	120mg	0.726	0.637	0.089	0.416
HMAY-B	80mg	0.302	0.253	0.049	-0.008
	120mg	0.359	0.253	0.106	0.049
HMAQ-A	60mg	0.52	0.19	0.33	0.21
HMAQ-B	60mg	0.15	0.09	0.06	-0.16
HMCR	60mg	0.273	0.209	0.064	-0.037

HMBH-A	60mg	0.727	NA	NA	0.417
HMBH-B	60mg	0.321	NA	NA	0.011
HMBV	60mg	0.52	NA	NA	0.21
HQAC	60+120mg	0.55		NA	0.24

In addition, 12 of the 15 duloxetine treatment arms arose from having 2 identical studies run via the same protocol. Each arm was independently and adequately powered, but designed to be pooled to increase precision. Using the effect sizes for the duloxetine versus placebo contrast from the pooled data from each of the 6 study pairs and contrasting those 6 effect sizes versus historical control showed an effect size of -0.01 (40mg HMAT, lowest dose), 0.108 (80mg HMAT), 0.086 (80mg HMAY), 0.232 (120mg HMAY, highest dose), and 0.214 (60mg HMBH, two-arm study).

In other words, 3 of the 6 pairs yielded an estimate extremely close to the final (overall) estimate; 2 of the less accurate estimates may have been influenced by dosing as these came from the highest and lowest dose; and, 1 was because from the favored 2-arm design. If applying the adjustment factor reported in Papakostas and Fava [20], the "matched historical control advantage" of duloxetine from the two arm studies is reduced to XX, which is again reasonably close to the overall estimate of 0.11.

Simply put, the efficacy of duloxetine versus SoC could have been inferred from historical control with approximately equal accuracy as from the concurrent control.

Discussion

There is an increasing need to understand the risks and benefits of a test drug compared with the standard of care (SoC) early in development. Even if a drug is superior to placebo it may not be worthwhile to continue development unless it has advantages over SoC. However, efficacy comparisons versus SoC in early phase studies can be challenging.

Simulation studies were conducted to illustrate that in common scenarios it may take samples sizes at least 5-fold greater to achieve reliable comparisons of a test drug with SoC than when comparing versus placebo. Given that test drugs often have no benefit, studies that focus on comparing a test drug with SoC may unnecessarily use resources that could be devoted to investigating other drugs. It is also important to consider that constructing valid comparisons versus SoC may require extensive experience with the test drug[4]. In a PoC study, the dose of the test drug may not be the most appropriate dose to compare with SoC; or, the most relevant patient population may not have been enrolled; or, the most relevant outcomes on which to focus may not be known. Therefore, the value of comparing a test drug versus SoC in a PoC study is difficult to ascertain. Conversely, why focus on comparing a drug with placebo when SoC is the relevant comparator?

The example with clinical-trial data in major depressive disorder illustrated how this trade-off can be mitigated using a literature database of placebo controlled studies to construct an historical comparison of a test drug with SoC. However, this retrospective illustration should be considered in the light of several factors. The potential for bias in

non head-to-head comparisons is well known and can be considerable[1]. Therefore, historical comparisons cannot be used as a substitute for adequate and well-controlled concurrent comparisons. However, historical comparisons may still be useful in the early evaluations of a test drug when adequate and well controlled comparisons versus SoC may not be feasible.

Use of historical control as opposed to concurrent control is essentially a trade-off between bias and precision. If the bias in the historical control can be minimized and sufficient historical data exist, as would typically be the case for a standard of care, so that the effects of SoC can been estimated precisely, then historical control can be useful.

While the depression example appears to be one wherein historical control may be useful, it involved a small number of studies and was retrospective. Evaluations across larger databases would be useful. Furthermore, major depressive disorder is a scenario where abundant historical data exists and individual clinical trial results are highly variable. Therefore this was an ideal area to consider historical data and may not be indicative of other scenarios.

In addition, other means of utilizing historical data to improve early evaluations of a test drug with SoC should be considered. For example, Bayesian statistical approaches that explicitly incorporate prior data into the analysis may be useful INSERT CITATION ON BAYESIAN APPROACHES IN CLINICAL TRIALS BOOK BY JOE IBRAHIM AND BOOK BY DON BERRY AND DALENE STANGLE MAY BE USEFUL. It may also

be useful to consider adaptive designs INSERT CITATION ON ADAPTIVE DESIGNS IN CLINICAL TRIALS, CHOW and CHANG?. For example, patients could be randomized to placebo, test drug, and SoC, with an interim analysis conducted when the test drug versus placebo contrast is adequately powered. If the interim result is positive, enrollment would continue until the test drug versus SoC contrasts is sufficiently reliable. However, a discussion of Bayesian approaches and adaptive designs is beyond our present scope.

It is also beyond the present scope to discuss the myriad ways in which a test drug might be different from SoC. Nevertheless, if the sample sizes required for sufficiently reliable contrasts versus SoC on a primary efficacy outcome are not feasible, it is useful to consider all the potential ways a novel drug might differ from SoC in regards to safety outcomes, patient subgroups, etc. As such, proof of concept for a novel drug could be demonstrated versus placebo, with comparisons versus SoC deferred until the larger phase III studies are conducted, assuming that the novel mechanism is likely to yield meaningful differences from SoC on some important outcome.

And thus many issues remain unresolved and more work is needed to understand how to optimize early comparisons of a test drug versus SoC. However, some points are clear. A little data is not necessarily better than no data. Powering a study versus placebo and then randomizing the same number of patients to an SoC and contrasting SoC with the test drug can yield very misleading results. Therefore, it is crucially important that the

rate of false positive and false negative results be quantitatively evaluated before deciding upon the sample size and the criteria upon which the test drug and SoC will be compared.

References

1. ICH guidelines accessed at: http://www.ich.org/cache/compo/276-254-1.html

2. Keene, Oliver N. (December 14, 2007). Phase II Trials. In Wiley Encyclopedia of Clinical Trials (Ralph D'Agostino, Lisa Sullivan, and Joe Massaro, eds). Hoboken: John Wiley & Sons, Inc., dx.doi.org/ 10.1002/9780471462422.eoct368

3. Temple R, Ellenberg SS: Placebo-controlled trials and active-control trials in the evaluation of new treatments. 1. Ethical and scientific issues. Ann Intern Med 2000; 133:455–463

4. Lieberman JA, Greenhouse J, Hamer RM, Krishnan KR, Nemeroff CB, Sheehan DV, Thase ME, Keller MB: Comparing the effects of antidepressants: consensus guidelines for evaluating quantitative reviews of antidepressant efficacy. Neuropsychopharmacology 2005;30: 445-460.

5. Woos, S. Consistency of atypical antipsychotic superiority to placebo in recent clinical trials. *Biological Psychiatry*, Volume 49, Issue 1, Pages 64-70

6. Turner EH, Matthews AM, Linardatos E, Tell RA, Rosenthal R: Selective publication of antidepressant trials and its influence on apparent efficacy. N Engl J Med 2008; 358:252–260

7. Goldstein DJ, Mallinckrodt C, Lu Y, Demitrack MA: Duloxetine in the treatment of major depressive disorder: a double-blind clinical trial. J Clin Psychiatry 2002;63: 225-231.

8.Goldstein DJ, Lu Y, Detke MJ, Wiltse C, Mallinckrodt C, Demitrack MA: Duloxetine in the treatment of depression: a double-blind placebo-controlled comparison with paroxetine. J Clin Psychopharmacol 2004;24: 389-399.

9. Detke MJ, Wiltse CG, Mallinckrodt CH, McNamara RK, Demitrack MA, Bitter I: Duloxetine in the acute and long-term treatment of major depressive disorder: a placebo and paroxetine-controlled trial. Eur Neuropsychopharmacol 2004;14: 457-470.

10. Perahia DG, Wang F, Mallinckrodt CH, Walker DJ, Detke MJ: Duloxetine in the treatment of major depressive disorder: a placebo- and paroxetine-controlled trial. Eur Psychiatry 2006;21: 367-378.

11. Nierenberg AA, Greist JH, Mallinckrodt CH, Prakash A, Sambunaris A, Tollefson GD, Wohlreich MM. Duloxetine versus escitalopram and placebo in the treatment of patients with major depressive disorder: onset of antidepressant action, a non-inferiority study. Current Medical Research and Opinion. 2007;23(2):401-416.

12. Detke MJ, Lu Y, Goldstein DJ, Hayes JR, Demitrack MA: Duloxetine, 60 mg once

daily, for major depressive disorder: a randomized double-blind placebo-controlled trial. J Clin Psychiatry 2002;63: 308-315.

13. Detke MJ, Lu Y, Goldstein DJ, McNamara RK, Demitrack MA: Duloxetine 60 mg once daily dosing versus placebo in the acute treatment of major depression. J Psychiatr Res 2002;36: 383-390.

14. Raskin J, Wiltse CG, Siegal A, Sheikh J, Xu J, Dinkel JJ, Rotz BT, Mohs RC: Efficacy of duloxetine on cognition, depression, and pain in elderly patients with major depressive disorder: an 8-week, double-blind, placebo-controlled trial. Am J Psychiatry 2007;164:900-909.

15. Nemeroff CB, Schatzberg AF, Goldstein DJ, Detke MJ, Mallinckrodt C, Lu Y, Tran PV: Duloxetine for the treatment of major depressive disorder. Psychopharmacol Bull 2002;36: 106-132.

16. Mallinckrodt CH, Prakash A, Houston JP, Swindle R, Detke MJ, Fava M. Differential antidepressant symptom efficacy: Placebo-controlled comparisons of duloxetine and SSRIs (fluoxetine, paroxetine, escitalopram). Neuropsychobiology. 2007;56:73-85.

17 INSERT REFERENCE FOR LILLY TRIALS

18. Mallinckrodt, Craig H, and Virgil Whitmyer (December 14, 2007). Phase III Clinical Trials. In Wiley Encyclopedia of Clinical Trials (Ralph D'Agostino, Lisa Sullivan, and Joe Massaro, eds). Hoboken: John Wiley & Sons, Inc., dx.doi.org/ 10.1002/9780471462422.eoct328

19. Fava, M., Evins, A.E., Dorer, D.J., Schoenfeld, D.A., 2003. The problem of the placebo response in clinical trials for psychiatric disorders: culprits, possible remedies, and a novel study design approach. Psychother. Psychosom. 72, 115–127

20. Khan A, Detke M, Khan SR, Mallinckrodt C. Placebo response and antidepressant clinical trial outcome. Journal of Nervous and Mental Disease 2003; 191:211-218.

21. Papakostas. George I. and Maurizio Fava. Does the probability of receiving placebo influence clinical trial outcome? A meta-regression of double-blind, randomized clinical trials in MDD. European Neuropsychopharmacology (2009) 19, 34–40

22. Walsh BT, Seidman SN, Sysko R, Gould M (2002). Placebo response in studies of major depression: variable, substantial, and growing. JAMA 287: 1840–1847.

23. Stein, D.J., Baldwin, D.S., Dolberg, O.T., Despiegel, N., Bandelow, B., 2006. Which factors predict placebo response in anxiety disorders and major depression? An analysis of placebo-controlled studies of escitalopram. J. Clin. Psychiatry 67 (11), 1741–1746.