Rate of Termination in Clinical Depression Trials

by

Sai Priya Marrapu

Research Project

Submitted to the School of Health Sciences

Eastern Michigan University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Clinical Research Administration

Submitted to:

Irwin Martin, Ph.D.

March 2020

Ypsilanti, Michigan

Abstract

Clinical trials are essential for determining the efficacy and safety of a treatment, and many drugs fail this process due to trial termination. Among all trials, those for clinical depression have yielded higher termination rates. This study investigated reasons for trial failure by comparing terminated clinical depression trials against all terminated trials across conditions in order to ascertain whether three specific issues—placebo effect, sample size, and the use of composite scores—were unique concerns for clinical depression trials. Major findings show that higher termination rates in clinical depression trials were directly correlated with the use of composite scores as primary outcome measures. Common reasons cited for trial termination were toxicity, insufficient accrual rate, and lack of funding. While these results followed expectations, one issue not proven to contribute to higher clinical depression trial termination was sample size, which was found to be an issue in all types of terminated trials.

Table of Contents

List of Tables

Introduction

Clinical Trials

Clinical trials are necessary for ascertaining the efficacy and safety of a drug. However, each trial is often costly and can take a long time, and not all clinical trials are able to reach completion. False leads and unverifiable impressions represent principal reasons that many drugs do not successfully make it through the rigorous process of clinical trials and onto pharmacy shelves (Leon et al., 2011).

The existence of the perfect clinical trial design remains elusive. In fact, only a small proportion of clinical trials actually reach a marketing application with the Food and Drug Administration (FDA), with many trials ending in termination for a variety of reasons (Khan et al., 2018). Moreover, trial termination and flawed clinical trial designs and methodologies are especially common among trials for pharmacological therapies geared towards patients that have depression (Khan et al., 2018). Therefore, in order to further investigate the reasons for high termination rates among clinical depression trials, this study critically examined the history of clinical trials in depression, their termination and specific concerns leading to termination, and a comparison of clinical trials across conditions.

History of Clinical Depression Trials

There exists a long history of clinical trials for depression, dating back to the 1950s in Switzerland (Brown et al., 2015). In the late 1950s, the psychiatrist Roland Kuhn administered a test drug that was later named imipramine. This had differential effects on the treatment of schizophrenia, but some of Kuhn's patients exhibited an alleviation of their depression symptoms. While these effects were not tested in a controlled setting, their observation set the stage for the continued examination of the efficacy of drugs in treating psychiatric conditions

such as depression (Berna et al., 2017). Further momentum was gained in the early 1960s, with additional pharmaceutical compounds being developed and tested in patients with depressive symptoms. In fact, at this time, several antidepressants were regularly being prescribed to patients, although none of these treatments had been clinically investigated for efficacy or safety (Khan et al., 2018). Furthermore, many depression diagnoses being treated through these drugs may have been determined solely based on the subjective evaluation of individual psychiatrists and other clinicians. Despite the progress psychiatrists were making in treating some patients' depressive symptoms, there was still no systematic way of addressing the safety of these experimental compounds (Williams et al., 2015).

Loosely defined guidelines for the treatment of depression during this era meant that patients may have suffered from negative effects. Side effects of antidepressant drugs at this time were not well documented, and potential outcome measurements were not well defined (Berna et al., 2017). Similar issues abounded in the treatment of other types of conditions, as well, and in 1962 these unsafe practices of prescribing drug treatments to patients without appropriate testing were halted. Following this time, the FDA began mandating that a drug be proven not only safe but also effective before it could be marketed to the public, and this included the requirement that all sponsoring pharmaceutical companies first obtain FDA approval on their trial design before pressing forward with a new pharmacological agent (Berna et al., 2017; Khan et al., 2018). Naturally, this led to a significant change in how clinical trials were conducted for antidepressants, as well.

In 1965, one specific clinical depression trial utilized a double-blind, randomized, placebo-controlled trial design to demonstrate the advantages of several active drug treatments over a placebo as well as differential responses across treatment arms (British Medical Research

Council, 1965). Despite the limitations later observed in this study, the trial's FDA approval and success made a large impact on the way antidepressant trials would be conducted from then on. Importantly, sponsors seemed to equate this trial design with receiving FDA approval and thus proving efficacy in their experimental treatments. Consequently, alternate trial designs were largely abandoned, and another issue—sample size—still remained to be addressed (Khan et al., 2018). Sample size was not being given adequate attention and having fewer participants in each treatment group would negatively affect a study's ability to detect differences between the treatment group and the control group (Blackford, 2017; Khan et al., 2018; Williams, 2015).

Issues with Clinical Depression Trial Termination

While trial termination is not an unusual occurrence among experimental treatments across all conditions, the rate of termination is especially high among clinical depression trials (Khan et al., 2018). Frequently across the literature, lack of efficacy due to a placebo effect has been cited as the primary reason for this higher trial termination rate. Randomized controlled trials in clinical depression have historically utilized placebo arms as a way of determining the efficacy of new drug treatments (Walsh et al., 2002). Patients who are assigned to the placebo arm of a clinical depression trial believe that they are receiving treatment, and the placebo effect occurs when some of these patients respond with a relieving of their depression symptoms despite not receiving the active drug treatment (Frank et al., 1991; Leuchter et al., 2014). However, not all participants benefit from the placebo effect, and those patients who do not are then receiving no treatment for a condition that requires it (Bird et al., 2010; Nugent et al., 2017). Those who oppose the use of the placebo in clinical trials claim it is unethical to assign any patient to a placebo rather than to a known treatment proven to be effective (Kraemer, 2000; Michels, 2000; World Medical Association, 2000).

Beecher (1955) first documented the trend of placebo response among clinical trial patients, and the scientific community assumed for more than half a century that a large proportion of patients would experience an improvement in their depression symptoms through placebo treatment. In fact, clinical depression trials saw a varied placebo response, ranging from about 10% to over 50% (Walsh et al., 2002). With this variable and often significant response to placebo treatment, it became expected that patients' placebo response would be similar to that of an antidepressant treatment (Ghosh & Kramer, 1999; Leber, 1989; Robinson & Rickels, 2000). By the early 2000s, the 50% termination rate among clinical depression trials was fully attributed to placebo effect (Walsh et al., 2002; Khan et al., 2003).

Furthermore, the degree of placebo response in clinical depression trials has increased over time. From 1981 to 2000, there was a 7% increase each decade (Walsh et al., 2002), and a continuation of this trend was observed after 2000, as well (Khan et al., 2017). This latter investigation, which compared pre-2000 and post-2000 antidepressant clinical trials, reviewed 85 acute, parallel-group, double-blind, placebo-controlled trials for investigational antidepressants that had been registered for new drug application (NDA) in adult patients diagnosed with clinical depression. After excluding any treatment arms for active comparators or investigational antidepressants not approved by the FDA, the investigators compared placebo response to antidepressant response in the 115 active treatment arms that remained: 67 from before 2000, and 48 from after 2000. Based on their findings, placebo response increased from 29.8% in pre-2000 trials to 36.2% in post-2000 trials, or 6.4% (Khan et al., 2017). Researchers have not been able to pinpoint an exact association between this rise and clinical trial design (Khan et al., 2011; Khan et al., 2017). On the other hand, it is known that the length of randomized controlled trials has increased over time, making it possible that the proportion of patients exhibiting the

placebo effect is directly associated to the length of the study (Khan et al., 2000; Quitkin et al., 1987; Quitkin et al., 1996; Walsh et al., 2002). In a meta-analysis by Hrobjartsson and Gotzsche (2001), little evidence was found to support the idea that patients receiving a placebo fared much differently over time than those who received no treatment. A longer study may simply provide patients with an overall benefit accumulated throughout the study's interventions as well as a greater amount of time in which to achieve spontaneous recovery (Walsh, 2002).

Khan et al. (2018) argue, however, that this focus on the placebo effect is misplaced when there are other factors that should be considered as the reasons for the high rate of termination among clinical depression trials. In particular, modern clinical trial design and its assumption that all conditions and drug classes can be tested through the same methods may be a greater source of concern for clinical depression trial termination. According to Khan et al. (2018), the sample size and primary outcome measures used in many clinical depression trials are flawed—and thus the leading cause of their trial termination rate—and these will be discussed below.

Sample size. Underpowering, or when a clinical trial has too few participants in each treatment arm for the results to reach the statistical power necessary for reliable results, is a major concern for clinical depression trials (Khan et al., 2018). Smaller studies are by nature less precise and have been clearly linked to results with greater placebo effect (Fletcher, 2008; Gibertini et al., 2012). In comparison, larger trials show a more stable, less variable placebo response (Fletcher, 2008; Khan et al., 2018). In fact, the success rate of clinical depression trials has increased alongside sample sizes, and even when larger trials have yielded a larger placebo response, the active drug treatment maintained an approximate 10% superiority over the placebo treatment (Fletcher, 2008). Khan et al. (2018) asserts that if smaller sample size is causing this

trend toward lack of efficacy due to placebo effects, it should actually be sample size and not placebo response that receives greater attention.

Composite scores. Depression outcome measurement based on composite scores is another major concern for clinical depression trial terminations (Khan et al., 2018). Clinical depression trial design has seen a historical trend away from using single-factor outcome measures to complex composite outcome measures (Khan et al., 2018). Single-factor outcome measures are discrete measurements taken to summarize the patient's condition. In contrast, composite scores, such as the HAM-D or Montgomery-Asberg Depression Rating Scale, combine multiple symptom measurements to come to a total score aimed at capturing the patient's overall condition. The clinical depression trial shift toward using composite outcome measures assumes that patients with similar composite scores have a uniform severity in their overall condition, which would appear flawed given that composite scores themselves are not uniform in the depression symptoms that they measure (Khan et al., 2018). According to Fried and Nesse (2015), it is possible for patients with similar composite scores to represent as many as 1,000 different symptom combinations, and some of those patients may not even share a single symptom with others in the group. Moreover, any one symptom can come with varying degrees of impairment effects on the patient (Fried et al., 2014). Furthermore, antidepressants do not necessarily exhibit uniform effects on all patients' symptoms throughout a clinical trial, and this can influence the composite score, as well (Ballard et al., 2018). For instance, a single antidepressant may lead to sedation effects in a patient with hypersomnia or even more severe sleep loss for a patient with insomnia (Ballard et al., 2018).

Another vulnerability in using composite outcome measures in clinical depression trials is that this kind of measurement relies on patient recall of information and self-reporting (Fried,

2015; Khan et al., 2018). This requires the use of complicated, indirect interview questions, and yet these questions are being asked of individuals who are already cognitively impaired by the depression for which they are being included in the trial. It cannot be expected that all patients will be able to recall their symptoms accurately and report them uniformly (Fried, 2015).

Essentially, clinical depression trials have been expected to measure a single condition with a diversity of symptom patterns using a technique that is unlikely to be measuring the same thing across patients, making the use of composite outcome measures incompatible with clinical depression trials (Fried and Nesse, 2014; Fried et al., 2016). If a composite score were to be used appropriately, it would require a significantly larger sample size in order to yield meaningful analysis and results (Ballard et al., 2018), yet as stated previously, underpowering is an issue among clinical depression trials. For these reasons, Khan et al. (2018) argue that it is sample size and the use of composite outcome measures—not lack of efficacy due to placebo effect—that are responsible for the high rate of termination among clinical trials for depression.

Purpose of this Study

Various evidence has been presented as to the reasons for the higher rate of termination among clinical depression trials. These potential reasons include lack of efficacy due to placebo effects, insufficient sample size, and the appropriateness of the primary outcome measures being used. Then, this study investigated these and other reasons for trial termination by comparing terminated clinical depression trials against all terminated trials across an array of conditions in the human population. Since February 2000, the National Institute of Health (NIH) has maintained a registry of all clinical trials at ClinicalTrials.gov. It is mandated that the sponsor or principal investigator for every clinical trial submit an initial protocol as well as summary results of the trial within one year of the completion date in order to keep an accurate record of all trials

on the registry (Zarin et al.,2007; Zarin et al., 2011). By quantifying the degree with which the three concerns—placebo effect, sample size, and the use of composite outcome measures—are unique to clinical depression trials in this registry, this study seeks to ascertain which of the reasons are truly significant to the rate of termination in clinical depression trials.

Methods

Study Sample

Clinical trials to be examined in this study were collected from ClinicalTrials.gov, which contains all clinical trials regardless of their recruitment status and organizes them within nine different categories. See Table 1 for all recruitment status categories.

Table 1: ClinicalTrials.gov Recruitment Status Categories.

Recruitment Status Categories
1. Not yet recruiting 2. Recruiting 3. Enrolled by invitation 4. Active but not recruiting 5. Suspended 6. Terminated 7. Completed 8. Withdrawn 9. Unknown status

Inclusion criteria for this study sample were that the clinical trial both have data collected from study participants and also be terminated; thus, focus was placed on terminated clinical trials only. Terminated studies, by definition, are those where the trial was halted prematurely, will not resume, and the participants are no longer being examined or receiving intervention. This implies that the study included at least one or more participants and involved data collection but was terminated or ended without any final results that are meaningful for advancement.

Suspended and withdrawn studies were also initially considered for inclusion, although it was eventually determined that they should be excluded from the sample. Suspended trials are those that were halted prematurely but may potentially resume in the future. These trials do not reflect those that have a final termination status; therefore, it cannot yet be known if they will be terminated at a later stage. Thus, any conclusions regarding these trials are indeterminate. Similarly, withdrawn trials are those where the study was halted prematurely prior to the enrollment of any participants. With this type of trial, no results could be obtained due to there being no study participants; therefore, trials with this recruitment status also fail the inclusion criteria for further consideration. Since only terminated trials were considered for this particular study, it was possible to compare the failure rate and reasons for failure among clinical depression trials with those of trials across all conditions.

Study Attributes

For trials that met the inclusion criteria, the reasons for termination were examined. The entry for each terminated clinical trial includes an optional field with a limit of 160 characters, where the sponsor or principal investigator can provide more details regarding the reasons for termination. For this study's data analysis, information was extracted from these 160 characters and then organized into three potential categories. The first category included scientific reasons, under which there were both "lack of efficacy" and "toxicity" as sub-categories. Second, some trials provided no reason or the information could not be codified. In the third category, all other reasons were grouped into six sub-categories: lack of funding, insufficient accrual rate, low funding and insufficient accrual, strategic reason/business decision, existence of competing clinical trials, and trial administration issues (see Table 2).

Table 2: Reasons for Termination and Further Categorizations.

Reason for Termination	Sub-Categories
Scientific Reasons	1. Lack of efficacy 2. Toxicity
Reason not provided/not available (NA)	Not sub-categorized
All Other Reasons	1. Lack of funding 2. Insufficient accrual rate 3. Low funding and insufficient accrual 4. Strategic reason/business decision 5. Existence of competing clinical trials 6. Trial administration issues

In addition, for each clinical trial, the actual number of participants enrolled in the study was obtained in order to determine whether there may have been potential issues related to conducting a trial that have low accrual rate. The type of primary outcome measure used, either composite score or single outcome measure, was also obtained for further analysis.

Sample size determination for each of the clinical trials was compared between those that were clinical depression trials with those across all conditions. Here, sample size refers to the actual enrollment, and the clinical trials were divided into the following groups based on their actual enrollment: <10, 10 – 50, 50 – 100, 100 – 150, and >150. The frequency distribution in the proportion of all clinical depression trials meeting sample size categories were compared to those

across all condition groups. Primary outcome measures were categorized as either composite scores or single outcomes.

Potential Statistical Method Implemented

For calculating significant differences, this study used Cohen's h for standardized mean differences (ES) in proportions between groups. For relatively large sample groups, as with the large group of terminated clinical trials across all conditions, statistical significance can occur and present apparent differences by overpowering a smaller sample group, which in this study was the group of terminated clinical depression trials. In response to this, Cohen's h was used with a difference cutoff of 0.1. Comparison of group proportions that exceed the difference cutoff are considered statistically significant (Cohen et al., 2013). The effect size calculation uses an arcsine and square root function approximation to determine the difference in proportions, assuming that the proportions are within the interval from 0 to 1.

Results

A total of 218 terminated clinical depression trials were identified. For the comparison group, there were 4,625 terminated trials across all indications. For both clinical depression trials and all trials, insufficient accrual rate and trial administration issues topped the reasons for termination. When clinical depression trials were compared to all trials, a significant difference was observed in the proportion terminated due to insufficient accrual rate. It was also found that a significantly higher percentage of all trials were terminated due to toxicity. However, termination attributable to lack of funding was substantially higher for clinical depression trials, and clinical depression trials also had a significantly higher proportion whose termination reason was not provided. See Table 3 below for all comparisons between groups.

Table 3: Comparison of Reasons for Trial Termination.

Termination Category	Depression Number of Trials (%)	All terminated trials (%)	Effect size	Effect size > 0.1
Total Terminated	**218 (100%)**	**4625 (100%)**		
1. Scientific reasons				
a. Lack of efficacy	20 (9.17%)	409 (8.84%)	0.012	0
b. Toxicity	6 (2.75%)	391 (8.45%)	0.557	1
2.Termination reason not provided	37 (16.97%)	482 (10.42%)	0.192	1
3.All other reasons				
a. Lack of funding	26 (11.92%)	345 (7.45%)	0.151	1
b. Insufficient accrual rate	70 (32.11%)	1630 (35.24%)	1.158	1
c. Low funding and insufficient accrual	7 (3.21%)	93 (2.01%)	0.076	0
d. Strategic reason/business decision	14 (6.42%)	392 (8.47%)	0.078	0
e. Existence of competing clinical trials	1(0.46%)	70 (1.51%)	0.233	1
f. Trial administration issues	37 (16.97%)	813 (17.57%)	0.016	0

With respect to the sample sizes in terminated trials, greater than 70% of all terminated trials—and also of clinical depression trials only—had fewer than 50 participants, or what is considered a small sample size. Conversely, though there was a general decrease in the percentage of trials terminated as sample size increased, very large trials with more than 150 subjects participants represented over 10% of terminated trials for both clinical depression trials and all trials. Therefore, both groups followed a similar frequency distribution of termination across sample size/actual enrollment. See Table 3 for additional details.

Table 4: Comparison of Frequency Distribution.

Sample size / Actual Enrollment	Depression (N=218)	All Terminated Trials (N=4625)
<10	81 (37.15%)	1750 (37.83%)
10 – 50	79 (36.23%)	1833 (39.63%)
50 – 100	23 (10.55%)	393 (8.49%)
100 – 150	10 (4.58%)	173 (3.74%)
>150	25 (11.46%)	476 (10.29%)

A comparison of primary outcome measures yielded perhaps the clearest distinction between terminated clinical depression trials and all terminated trials. A view of all terminated trials shows that few used composite scores as their primary outcome measure, while a relatively much larger proportion of these terminated trials used a single outcome measure. Conversely, terminated clinical depression trials saw a much larger proportion using composite scores rather than a single outcome measure. This indicates that far fewer clinical depression trials were terminated if they had a single outcome measure, despite the opposite being true among clinical trials across all conditions.

Table 5: Comparison of Primary Outcome Measures.

Primary Outcome Measure	Depression (N=218)	All Terminated Trials (N=4625)
Composite Scores	160 (73.39%)	574 (12.41%)
Single outcome measure	50 (22.93%)	4049 (87.54%)
Not available	8 (3.66%)	2 (0.04%)

Discussion

Several key findings were obtained through this study, the principal ones being that the use of composite scores as primary outcome measures is widely more prevalent among terminated clinical depression trials than among all other terminated trials, as well as the fact that sample size was not significantly more frequent a reason for termination among clinical depression trials, specifically. Additionally, while lack of efficacy was cited as the reason for termination in nearly 10% of terminated clinical depression trials, the data provided does not prove why efficacy was not shown and thus cannot sufficiently link this lack of efficacy to placebo effect. Apart from these specific considerations under study, it was also found that toxicity, insufficient accrual rate, and lack of funding represented statistically significant differences between terminated clinical depression trials and all terminated trials. As well, despite the high termination rate among clinical trials for depression, it was unfortunate that no specific reason for termination was provided for a large portion of these trials.

First, it was found that the use of composite scores as the primary outcome measure was substantially more common among terminated clinical depression trials than among all terminated trials. This may indicate that using composite scores to define the outcome is more likely to lead a clinical depression trial to eventual termination, and that single outcome measures may be more effective in leading a clinical depression trial towards completion. In order to further examine this point, future studies could closely scrutinize the frequency of composite scores and single outcome measures in terminated clinical depression trials against those in non-terminated trials.

With regards to sample size, there are no salient features particular to terminated clinical depression trials, and this implies that all clinical trials are subject to issues with sample size or

insufficient enrollment. Though it was not explicitly examined in this study, there could be potential differences in termination reasons between small and large clinical trials. For instance, it may be possible that scientific and financial reasons for termination among smaller clinical trials are different than those for termination among large clinical trials. Future studies could consider examining these differences and drawing further speculation on design, organizational, and structural reasons for termination.

Lack of efficacy was not found to be a statistically significant reason for the termination of clinical trials, although it was slightly more common among clinical depression trials than among trials across all conditions. Interestingly, nowhere across the terminated trials at ClinicalTrials.gov is it specifically mentioned that the appearance of a placebo effect was the reason for lack of efficacy. This may in fact be the case; however, this cannot be known for certain without examining the prevalence of the placebo effect among non-terminated trials that yielded unsuccessful results in the active drug treatment. Within the scope of this study, examining all the non-terminated trials at ClinicalTrials.gov was not feasible.

Finally, coming to the reasons for termination, toxicity, insufficient accrual rate, and lack of funding to continue the trial each presented interesting differences between terminated clinical depression trials and terminated trials across all conditions. Toxicity, in particular, represented a much lower percentage among terminated clinical depression trials than among all terminated trials. Since these trials are shown to be less toxic, it would appear that exposing additional participants to antidepressants in clinical trials poses minimal risk. If clinical researchers can manage to improve sample size or increase the accrual rate of their trials, it could lead to a greater number of clinical depression trials progressing to completion without risk of toxicity to the participating subjects.

As well, though insufficient accrual rate represented a slightly lower percentage among terminated clinical depression trials than among all terminated trials, it was still the single most cited reason for trial termination in both groups. With insufficient accrual rate comes the higher probability of smaller sample sizes, which are already known to more frequently lead to trial termination.

This study's strengths included the use of the ClinicalTrials.gov registry, a trusted data source representative of all clinical trials across the nation. Furthermore, the study compared clinical depression trials against all others for a strong reference category, and the sample size of trials was sufficiently large enough to perform significant comparisons. Alongside these strengths, the study also faced two specific limitations. The first was the lack of statistical power behind each termination reason, given the relatively small sample size of clinical depression trials (N=218) compared with the sample size of all trials across conditions (N=4625). The second and more important limitation was the data source's lack of information regarding why efficacy was not demonstrated in the clinical depression trials that cited lack of efficacy as the reason for termination. It may be that a large placebo effect was the primary reason for termination in those trials; however, additional data sources would be necessary to support this position.

Conclusion

This study examined terminated clinical trials collected from ClinicalTrials.gov, concentrating on sample size, primary outcome measures, and identified reasons for termination across both clinical depression trials and all trials across conditions. Based on these comparisons, one major finding among terminated clinical depression trials was the prevalence of composite scores being used as primary outcome measures. Conversely, sample size was not proven to be correlated to termination in clinical depression trials specifically. Significant reasons identified for the termination of clinical depression trials were found to be insufficient accrual rate and lack of funding, but one reason for termination of clinical depression trials that did not yield conclusive findings for this study was lack of efficacy due to placebo effects. Future investigation into the exact nature of lack of efficacy exhibited in clinical trials for depression would be required in order to ascertain a clearer picture on this matter. Clinical depression trials remain an important part of addressing and developing drugs for mental health conditions, and more research is necessary to support these trials through to completion.

References

Ballard, E.D., Yarrington, J.S., Farmer, C.A., Lener, M.S., Kadriu, B., Lally, N., . . . Zarate, C.A. (2018). Parsing the heterogeneity of depression: An exploratory factor analysis across commonly used depression rating scales. *Journal of Affective Disorders, 231,* 51-57. doi: 10.1016/j.jad.2018.01.027

Beecher, H.K. (1955). The powerful placebo. *JAMA: The Journal of the American Medical Association, 159*(17), 1602-1606. doi: 10.1001/jama.1955.02960340022006

Berna, C., Kirsch, I., Zion, S.R., Lee, Y.C., Jensen, K.B., Sadler, P., Kaptchuk, T.J., & Edwards, R.R. (2017). Side effects can enhance treatment response through expectancy effects: an experimental analgesic randomized controlled trial. *Pain, 158*(6), 1014-1020. doi: 10.1097/j.pain.0000000000000870

Bird, S.J. (2010). Responsible research: What is expected? Commentary on: "Statistical power, the Belmont Report, and the ethics of clinical trials". *Science and Engineering Ethics, 16*(4), 693-696. doi: 10.1007/s11948-010-9248-9

Blackford, J.U. (2017). Leveraging statistical methods to improve validity and reproducibility of research finding. *JAMA Psychiatry, 74*(2), 119-120.

British Medical Research Council. (1965). Clinical trial of the treatment of depressive illness. *British Medical Journal, 1*(5439), 881-886. doi: 10.1136/bmj.1.5439.881

Brown, W.A., & Rosdolsky, M. (2015). The clinical discovery of imipramine. *American Journal of Psychiatry, 172*(5), 426-429. doi: 10.1176/appi.ajp.2015.14101336

ClinicalTrials.gov. (2014). "Basic Results" data element definitions (draft). *Bethesda (MD): National Library of Medicine (US), National Center for Biomedical Information.* Retrieved January 6, 2020, from http://prsinfo.clinicaltrials.gov/results_definitions.html

ClinicalTrials.gov. (2014). Protocol data element definitions (draft). *Bethesda (MD): National Library of Medicine (US), National Center for Biomedical Information.* Retrieved January 6, 2020, from http://prsinfo.clinicaltrials.gov/definitions.html.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences.* New York: Routledge. Retrieved from https://doi.org/10.4324/9780203771587

Frank, J.D., & Frank, J.B. (1991). *Persuasion and healing: a comparative study of psychotherapy.* Baltimore, MD: The Johns Hopkins University Press.

Fried, E.I. (2015). Problematic assumptions have slowed down depression research: Why symptoms, not syndromes are the way forward. *Frontiers in Psychology, 6*, 309. doi: 10.3389/fpsyg.2015.00309

Fried, E.I., & Nesse, R.M. (2014). Depression is not a consistent syndrome: An investigation of unique symptom patterns in the STAR*D study. *Journal of Affective Disorders, 172*, 96-102. doi: 10.1016/j.jad.2014.10.010

Fried, E.I., & Nesse, R.M. (2015). Depression sum-scores don't add up: Why analyzing specific depression symptoms is essential. *BMC Medicine, 13(72), 1-11.* doi: 10.1186/s12916-015-0325-4

Fried, E.I., van Borkulo, C.D., Epskamp, S., Schoevers, R.A., Tuerlinckx, F., & Borsboom, D. (2016). Measuring depression over time… or not? Lack of unidimensionality and longitudinal measurement invariance in four common rating scales of depression. *Psychological Assessment, 28*(11), 1354-1367. doi: 10.1037/pas0000275

Ghosh, K., & Kramer, M.S.. (1999). A review of placebo controlled efficacy trials for antidepressants as evaluated by the FDA. *New Clinical Drug Evaluation Unit Program Abstracts*, poster No. 36.

Gibertini, M., Nations, K.R., & Whitaker, J.A. (2012). Obtained effect size as a function of sample size in approved antidepressants: A real-world illustration in support of better trial design. *International Clinical Psychopharmacology, 27*(2), 100-106. doi: 10.1097/YIC.0b013e32834f504f

Hrobjartsson, A., & Gotzsche, P.C. (2001). Is the placebo powerless? An analysis of clinical trials comparing placebo with no treatment. *New England Journal of Medicine, 344*, 1594-1602. doi: 10.1056/NEJM200105243442106

Khan, A., Detke, M., Khan, S., & Mallinckrodt, C. (2003). Placebo response and antidepressant clinical trial outcome. *The Journal of Nervous and Mental Disease, 191*(4), 211-218. doi: 10.1097/01.NMD.0000061144.16176.38

Khan, A., Fahl Mar, K., Faucett, J., Khan Schilling, S., & Brown, W.A. (2017). Has the rising placebo response impacted antidepressant clinical trial outcome? Data from the US Food and Drug Administration 1987-2013. *World Psychiatry: Official Journal of the World Psychiatric Association (WPA), 16*(2), 181–192. doi: 10.1002/wps.20421

Khan, A., Faucett, J., & Brown, W.A. (2014). Magnitude of change with antidepressants and placebo in antidepressant clinical trials using structured, taped and appraised rater interviews (SIGMA-RAPS) compared to trials using traditional semi-structured interviews. *Psychopharmacology, 231*(22), 4301-4307. doi: 10.1007/s00213-014-3584-4

Khan, A., Khan, S.R., Walens, G., Kolts, R., & Giller, E.L. (2003). Frequency of positive studies among fixed and flexible dose antidepressant clinical trials: An analysis of the Food and Drug Administration Summary Basis of Approval reports. *Neuropsychopharmacology, 28*(3), 552-557. doi: 10.1038/sj.npp.1300059

Khan, A., Mar, K.F., & Brown, W.A. (2018). The conundrum of depression clinical trials:

one size does not fit all. *International Clinical Psychopharmacology, 33*(5), 239–248. doi: 10.1097/YIC.0000000000000229

Khan, A., Warner, H.A., & Brown, W.A. (2000). Symptom reduction and suicide risk in patients treated with placebo in antidepressant clinical trials: An analysis of the Food and Drug Administration database. *Archives of General Psychiatry, 57*(4), 311-317. doi: 10.1001/archpsyc.57.4.311

Khin, N.A., Chen, Y., Yang, Y., Yang, P., & Laughren, T.P. (2011). Exploratory analyses of efficacy data from major depressive disorder trials submitted to the US food and drug administration in support of new drug applications. *The Journal of Clinical Psychiatry, 72*(4), 464-472. doi: 10.4088/JCP.10m06191

Kraemer, H.C. (2000). Statistical analysis to settle ethical issues? *Archives of General Psychiatry, 57*(4), 327-328. doi: 10.1001/archpsyc.57.4.327

Leber, P.D. (1989). Hazards of inference: The active control investigation. *Epilepsia, 30* (Suppl 1): S57-S63. doi: 10.1111/j.1528-1157.1989.tb05816.x

Leon, A.C., Davis, L.L., & Kraemer, H.C. (2010; 2011). The role and interpretation of pilot studies in clinical research. *Journal of Psychiatric Research, 45*(5), 626-629. doi: 10.1016/j.jpsychires.2010.10.008

Leuchter, A.F., Hunter, A.M., Tartter, M., & Cook, I.A. (2014). Role of pill-taking, expectation and therapeutic alliance in the placebo response in clinical trials for major depression. *The British Journal of Psychiatry, 205*(6), 443-449. doi: 10.1192/bjp.bp.113.140343

Michels, K.B. (2000). The placebo problem remains. *Archives of General Psychiatry, 57*(4), 321-322. doi: 10.1001/archpsyc.57.4.321

Nugent, A.C., Miller, F.G., Henter, I.D., & Zarate, C.A. (2017). The ethics of clinical trials research in severe mood disorders. *Bioethics, 31*(6), 443-453. doi: 10.1111/bioe.12349

Quitkin, F.M., Rabkin, J.D., Markowitz, J.M., Stewart, J.W., McGrath, P.J., & Harrison, W. (1987). Use of pattern analysis to identify true drug response: A replication. *Archives of General Psychiatry, 44*(3), 259-264. doi: 10.1001/archpsyc.1987.01800150071009

Quitkin, F.M., McGrath, P.J., Stewart, J.W., Ocepek-Welikson, K., Taylor, B.P., Nunes, E., Deliyannides, D., Agosti, V., Donovan, S.J., Petkova, E., & Klein, D.F. (1996). Chronological milestones to guide drug change: When should clinicians switch antidepressants? *Archives of General Psychiatry, 53*(9), 785-792. doi: 10.1001/archpsyc.1996.01830090031005

Robinson, D.S., & Rickels, K. (2000). Concerns about clinical drug trials. *Journal of Clinical Psychopharmacology, 20*(6), 593-596. doi: 10.1097/00004714-200012000-00001

Walsh, B.T., Seidman, S.N., Sysko, R., & Gould, M. (2002). Placebo response in studies of major depression: Variable, substantial, and growing. *JAMA: The Journal of the American Medical Association, 287*(14), 1840-1847. doi: 10.1001/jama.287.14.1840

Williams, R.J., Tse, T., DiPiazza, K., & Zarin, D.A. (2015). Terminated trials in the ClinicalTrials.gov results database: Evaluation of availability of primary outcome data and reasons for termination. *PLoS ONE, 10*(5), e0127242. Retrieved from https://doi.org/10.1371/journal.pone.0127242

World Medical Association. (2000). World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects. *JAMA: The Journal of the American Medical Association, 284*(23), 3043-3045. doi: 10.1001/jama.2013.281053

Zarin, D.A., Ide, N.C., Tse, T., Harlan, W.R., West, J.C., Lindberg, D.A.B. (2007). Issues in

the registration of clinical trials. *JAMA: The Journal of the American Medical Association, 297*(19), 2112-2120. doi: 10.1001/jama.297.19.2112

Zarin, D.A., Tse, T., Williams, R.J., Califf, R.M., & Ide, N.C. (2011). The ClinicalTrials.gov results database — Update and key issues. *The New England Journal of Medicine, 364*(9), 852-860. doi:10.1056/NEJMsa1012065