1997

Excess sample size and the 'Delta Wobble' in randomized controlled trials

Michael Adam Fischer
Yale University

EXCESS SAMPLE SIZE AND THE 'DELTA WOBBLE' IN RANDOMIZED CONTROLLED TRIALS

Michael Adam Fischer

Yale University

1997

Excess Sample Size and the 'Delta Wobble' in Randomized Controlled Trials

A Thesis Submitted to the
Yale University School of Medicine
in Partial Fulfillment of the Requirements for the
Degree of Doctor of Medicine

by

Michael Adam Fischer

1997

ABSTRACT

EXCESS SAMPLE SIZE AND THE 'DELTA WOBBLE' IN RANDOMIZED CONTROLLED TRIALS.

Michael A. Fischer and Alvan R. Feinstein. Department of Internal Medicine, Yale University, School of Medicine, New Haven, CT.

To determine the occurrence and consequences of excess sample sizes in large randomized controlled trials, we reviewed 158 randomized controlled trials, each containing more than 100 patients, published in *Lancet, Journal of the American Medical Association,* and *New England Journal of Medicine* during the three years 1990-1992.

Of 98 trials with statistically significant differences between control and experimental groups, the reported P values were less than 0.001 in 27 (28%) and less than 0.01 in an additional 35 (36%). Since sample sizes are usually calculated to provide P values of 0.05, the occurrence of values less than 0.01, and particularly below 0.001, suggests either that sample size was excessive or that the investigators found differences much larger than δ (the anticipated difference for "clinical importance"). The original anticipations were difficult to determine, however, because sample size calculations were not reported consistently: among the 158 trials, the details were presented completely in 78 (49%), but wholly omitted in 58 (37%). Of 54 trials that stated the value of δ and that found "statistical significance," 31 had P values below 0.01, but only 10 of these trials had observed differences that were at least 25% larger than δ; in the remaining 21, the small P value was attained only by excess sample size. On the other hand, 15 trials found $P<0.05$ and claimed statistical significance although the observed difference was at least 25% smaller than δ.

The problems of excessive sample size (and resources) probably arise from the customary Neyman-Pearson strategy, which tries to satisfy two (contradictory) statistical hypotheses, thereby making sample size much larger than what is needed for a single null hypothesis. The excessive sample size may then allow "statistical significance" to be found and emphasized for differences much smaller than what was originally anticipated.

ACKNOWLEDGMENTS

I gratefully acknowledge the advice and guidance of Dr. Alvan Feinstein throughout this research. He provided the initial insights that shaped the first stages of this project, and he has been intimately involved with its evolution, providing detailed advice and criticism throughout. Moreover, his willingness to be a mentor, providing insight into the entire field of clinical epidemiology, has made our collaboration enjoyable and inspirational.

Dr. Harlan Krumholz provided valuable criticism and suggestions for the final draft. I am indebted to Donna Cavaliere for her assistance in making appointments and getting various drafts of material back and forth between Dr. Feinstein and me, often with short notice on my part. Lori Moran and Linda Soha were very gracious about providing me with a variety of manuscripts.

Lastly, I am grateful to Jill Fischer for her constant support, her invaluable proofreading, and her patience in listening as I worked through the issues of this research. Thanks also to Jacob Fischer for not eating the manuscript.

TABLE OF CONTENTS

LIST OF TABLES

I. INTRODUCTION

> "The value for which P=.05, or 1 in 20, ... is convenient ... as a limit in judging whether a deviation is to be considered significant or not." [1, p. 44]

Sir Ronald Fisher originally made the above definition in 1925. Today the threshold of P below 0.05 remains the principal criterion for statistical significance, often representing the difference between research that changes current practice and research that does not. In many clinical trials, however, the authors report P values that are much smaller than 0.05. Very small P values can arise either because the difference in the event rates reported is much larger than originally anticipated or because the sample size was excessively large.

This research was aimed at documenting the frequency with which randomized controlled trials in major medical journals report very small P values, and to suggest possible reasons for the phenomenon. The research data were obtained by reviewing the sample size calculations and the subsequent results described in the published trials.

The remainder of this section contains a review of previous literature on sample size calculation in randomized controlled trials. Section II describes the methods used to assemble a group of randomized controlled trials for review and to abstract data from those articles. Section III presents the main results of the study, including the range of P values in the reviewed articles, the extent of reporting for sample size calculations, and the differences between the reported and the originally anticipated event rates. Section IV

discusses the implications of the results in Section III and shows some illustrative examples of calculations of sample size. Section V contains the conclusions.

A. Literature Review of Studies Reporting on Sample Size Calculation

Previous studies of the reporting of sample size calculation in randomized controlled trials have repeatedly shown that most authors do not report the details of their sample size calculations. In 1978 Ambroz et al. reviewed 172 randomized controlled trials and found that none of the publications reported a sample size calculation [2]. A 1982 review of 67 randomized controlled trials found that 12% of the articles reported the details of the sample size calculation and 3% provided partial information about the sample size calculation [3]. In 1986, a survey of the breast cancer literature by Liberati et al. revealed that 20 (32%) of 63 articles had reported sample size calculations in the text [4]. In follow-up phone calls, 13 more sets of authors provided the details of sample size calculation that had not been presented in the text [4]. In a 1987 review, 5 (11%) of 45 articles reported sample size calculations [5]. In 1990 Altman and Dore found some improvement in completeness of reporting: 31(39%) of 80 trials reported details of the sample size calculations and only 27(34%) of 80 articles made no mention of advance consideration of sample size [6].

Two large literature reviews in 1994 showed that reporting of sample size calculation methodology had become more widespread than in 1978 but was still far from universal. In a review of the obstetrics and gynecology literature from 1990 and 1991,

Schulz and co-authors found that only 50 (24%) of 206 articles reported sample size calculations [7]. These authors also found considerable variation between journals in the extent of reporting of sample size calculation [7]. Moher et al., in a sample of articles from 1975 to 1990, found that 33(32%) of 103 reported sample size calculation [8]. The proportion of articles reporting sample size calculation had increased over time, from 0% in 1975 to 43% in 1990 [8].

B. Studies Arguing for Larger Sample Sizes

The attempt to determine the proper sample size for a randomized controlled trial can be seen as both a practical and an ethical concern:

> A study with an overlarge sample may be deemed unethical through the unnecessary involvement of extra subjects and the correspondingly increased costs. ... On the other hand, a study with a sample that is too small will be unable to detect clinically important effects. Such a study may be scientifically useless, and hence unethical in its use of subjects and other resources. [9, p. 1336]

Over the last two decades, the predominant argument made in the medical literature has been for larger sample sizes.

The importance of large samples to avoid type II error, or false negative conclusions, in randomized trials was brought to prominence by Freiman et al. [10] in an influential article, nearly 20 years ago, that noted many trials reporting no difference between control and experimental groups were in fact not able to rule out differences as large as 25 or even 50 percent. Freiman et al. urged that much larger sample sizes would

be needed to state conclusively that there was no difference between control and experimental groups. In a recent article that revisited the issue originally raised by Freiman et al., Moher et al. found that many clinical trials reporting no difference still did not have samples large enough to exclude effects of 25 or 50 percent[8].

On the other hand, during informal reviews of the literature, several readers had noted extremely small P values, suggesting that sample sizes were substantially larger than needed to achieve the boundary of 0.05. The current study was evoked by questions about the frequency and sources of the "too-large-sample" phenomenon.

II. METHODS:

A. Assembling the Articles

For this review, I chose randomized controlled trials published in *Lancet, New England Journal of Medicine,* and *Journal of the American Medical Association* during the three year period from 1990 through 1992, inclusive. This choice follows the method of two prior studies. In one of the most widely cited reviews of statistical methods in the medical literature, Freiman et al. [10] examined articles from several journals, but more than one-half of the articles came from these three journals. In their later review of the same topics, Moher et al. [8] also examined articles in those same three journals. With the

emphasis on large clinical trials, I restricted my search to articles with a sample size of at least 100 patients.

Table 1 summarizes the literature search and the criteria for exclusion of articles. Using the Medline computer program in the summer of 1994, I restricted the search to *"Clinical Trial," "Multicenter Study,"* and *"Randomized Controlled Trial."* The search produced 1003 articles, which were reviewed to determine appropriateness for this study. I made certain simple exclusions by inspecting the abstracts cited by Medline, but other exclusions required review of the text of the articles. The search produced many articles, including letters (276), editorials (56), reviews (8), meta-analyses (9), and news summaries (16), all of which I excluded. The 238 articles that described trials containing fewer than 100 patients were also excluded, as well as 102 articles that were not randomized controlled trials, having been obtained via the headings *"Clinical Trial,"* and *"Multicenter Study"*. Additional exclusions were 84 trials whose primary outcome measure was not a rate or proportion, 21 trials that were designed to show equivalence (rather than efficacy) between control and experimental groups, 11 trials that were designed to demonstrate vaccine efficacy, eight trials that had multi-stage randomization schemes or other complexities that made them inappropriate for this analysis, and six that were follow-up cohort studies of patients from prior randomized controlled trials. Appendix A lists full citation information on the included articles.

Table 1: Details of literature search

	NEJM	Lancet	JAMA	Total
Articles Identified	396	490	117	1003
Excluded because:				
Letter to editor	82	190	4	276
N<100	84	131	23	238
Not rand. controlled trial	36	50	16	102
Not measuring event rate	38	30	26	84
Editorial	47	8	1	56
Trial for equivalence	14	6	1	21
"News"	1	2	13	16
Vaccine trials	4	7	0	11
Meta-analysis	2	1	6	9
Review article	1	6	1	8
Complex randomization	3	3	2	8
Follow-up studies	0	2	4	6
TOTAL EXCLUDED	312	436	97	845
TOTAL KEPT	84	54	20	158

After the exclusions cited in Table 1, the remaining 158 articles were each reviewed and suitably excerpted for descriptions of the sample size calculations. For trials that reported a statistically significant difference between experimental and control groups, the magnitudes of the main difference between groups, and the corresponding P value, were recorded. The remainder of this section describes the recorded components of the sample size calculation.

B. Introduction to Neyman-Pearson Equation

The most widely accepted method for calculating sample size is the Neyman-Pearson equation, shown below:

$$n \geq \frac{(Z_\alpha + Z_\beta)^2 \times \left[2 \times \pi_c \times (1 - \pi_c)\right]}{\delta^2} \qquad (1)$$

In this equation, n represents the number of subjects that will be required in each of two groups. Z_α represents the Z-score that corresponds to the designated value of α, which is the risk of type I error that the authors are willing to accept. Z_β represents the Z-score that corresponds to the designated value of β, which is the risk of a type II error. π_c represents the estimated value for the event rate expected in the control group, and the quantity $[2 \times \pi_c \times (1-\pi_c)]$ represents the variance of that rate. δ represents the anticipated difference that the authors hope to find between rates in the control and experimental groups.

The Neyman-Pearson sample size calculation requires two basic decisions: the first is defining the levels of significance that will be used as cutoff points for α and β; and the second is estimating event rates. If both of these decisions are fully described, a reader can replicate the sample size calculation and, aware of statistical assumptions made in trial design, can understand the importance of the subsequently reported P value.

The level of significance for rejecting the null hypothesis is defined by the designation of α, typically 0.05 and traditionally two-tailed. It corresponds to a Gaussian Z_α value of 1.96. The α value of 0.05 implies a 5% risk of a type I error, in rejecting a true null hypothesis, so that the observed finding arises from chance alone. The risk of type II error, in rejecting a true alternate hypothesis of a large difference between groups, is defined by β, which can have various values, but is often designated at 0.10. β can be either one- or two-tailed. The assigned value of β is often stated implicitly as the power of the study, which is calculated as $1-\beta$, so that the most commonly assigned power for a study is 90%.

For the studies that provided a sample size calculation, I recorded whether the α or β designations were described, and also listed as one- or two-tailed. Presentation of the power of a study was considered equivalent to presenting the value of β.

The other important decision in sample size calculation is a prior estimation of event rate in the control group(π_c) and the change(δ), i.e. delta, that the authors believe

would represent a clinically significant finding. In Equation 1 the required components are the variance $[2\pi_c(1-\pi_c)]$ of the rate in the control group in the numerator and δ in the denominator. Many authors do not present both of these designations. When authors present only the δ that they hoped to find, it is helpful for reviewing the final outcome of the trial, but does not provide enough information for the reader to re-create the sample size calculation. Many authors do not cite the absolute difference which they hoped to find for δ, but instead describe θ, the desired proportional (or relative) change, which would usually be calculated as δ/π_c. The presentation of only θ gives some information about the authors' assumptions, but does not allow re-creation of the sample size calculation. The sample size calculation can be replicated only if authors provide their prior designation of π_c together with any citation of π_e (the anticipated rate in the experimental group), δ, or θ.

For the articles that provided sample size calculations, I recorded which of these features were reported.

III. RESULTS

A. Range of P Values

Of the 158 articles in this sample, 104 reported statistically significant differences between the control and experimental groups. Of those 104 articles, the six that did not use P values in discussing the results were not included in this section. Table 2 shows that among the 98 articles with statistically significant outcomes, 27 (28%) reported P values less than or equal to 0.001, 35 articles (36%) had P values that were between 0.01 and 0.001, and the remaining 36 articles reported P values between 0.05 and 0.01.

Given that $P<0.05$ is the commonly accepted threshold for statistical significance, it seems surprising that over 25% of the articles reported P values 50 times smaller than the threshold value (i.e. ≤ 0.001). Two possible explanations could account for these extremely small P values: the observed difference (hereafter referred to as d_0) found to be statistically significant might have been much larger than the difference (δ) that the authors expected to find; alternatively, the number of patients in the trials might have been much larger than needed to achieve significance at the $P<0.05$ level* . To assess the frequency of these explanations, the original sample size calculations must be examined to determine the event rates that were estimated when the trial was designed.

---

* Section III.C will point out a third explanation - altered variance - for small P values, and the example in Section IV.B.3 will expand on that explanation

Table 2: Range of P values in trials with statistically significant outcomes

P value	Number of articles(%)
$P \leq 0.001$	27 (28%)
$0.001 \leq P \leq 0.01$	35 (36%)
$0.01 \leq P \leq 0.05$	36 (37%)

B. Sample Size Calculations

1. Presentation of Sample Size Calculations

Table 3 summarizes the ways in which authors reported their sample size calculations. Of the 158 large, randomized controlled trials under analysis, 58 (37%) were reported with no description of how sample size was calculated and with no reference to a previously published calculation. Of 100 (63%) that provided at least some description of the calculation, 12 required reference to a prior publication to find some or all of the main components in the sample size calculation. The cited 100 articles were stratified into the several groups shown in Table 3. The first column shows that 78 reports provided prior designations of event rates. The remaining 22 articles provided less detailed descriptions that would limit a reader's ability to fully understand the assumptions that went into the sample size calculation.

The rows of Table 3 show the extent to which authors noted their prior designations of α and β. For the sake of simplicity this table does not include whether authors indicated if their designations of α and β were one- or two-tailed. Almost half of the authors who presented a value of α classified it as one- or two-tailed (42/87), while very few authors noted whether their values of β were one- or two-tailed (6/91). The first row shows that 83 articles (53% of the total sample) presented both α and β designations. Only four authors (3%) presented only α values (2^{nd} row) and eight (5%) reported only β values (3^{rd} row). The fourth row of Table 3 shows that in addition to the 58 articles that

Table 3: Elements reported for sample size calculation

	Designation of π_e and δ described *	Designation of δ described, but π_e not described	Designation of θ described, but neither π_e nor δ described	No event rate designations described	TOTAL
Designations of α and β described	67	5	9	2	83
Designation of α alone described	2	1	0	1	4
Designation of β alone described	4	1	3	0	8
Neither α nor β designation described	5	0	0	58	63
TOTAL	78	7	12	61	158

* π_e or θ were acceptable substitutes for δ

did not describe any sample size calculation, five others (with some description) did not include their designations of either α or β.

The columns of Table 3 show the frequency with which authors reported their prior estimations of outcome rates in the control and experimental groups. As noted earlier, proper interpretation of results requires knowledge of π_c and of δ, although the value of δ could be calculated by subtraction if π_c and π_e are described, or by appropriate multiplication if π_c (or π_e) and θ are described. The first column shows that 78 articles (49% of the total sample) either provided both π_c and δ, or provided π_c and enough information for δ to be easily calculated. Smaller numbers of articles presented either δ alone (2nd column, 7/158, 4%) or θ alone (3rd column, 12/158, 8%). No articles in the sample reported π_c alone without some indication of the desired change, but the fourth column shows that, beyond the 58 articles with no information on the sample size calculation, only three articles gave no indication of the estimated event rates or desired differences.

2. *Completeness of Reporting for Sample Size Calculation*

As noted earlier, a complete description of the sample size calculation would allow understanding of the methods used by the investigators. The 67 trials in the upper left-hand corner of Table 3 provided complete or near-complete descriptions of sample size calculations, limited only by inconsistent reporting of whether α and β were one- or two-

tailed (four articles reported this information for both α and β). Similarly, for the 11 articles in the rest of the first column of Table 3, values of π_c and δ are described (although α and β are not both reported) so that a reader could appreciate the connotation of the P values reported by the authors.

For the 19 articles in the middle two columns of Table 3, the reader would be hard-pressed to understand the sample size calculation. A determined reader could make multiple guesses at the designation of π_c, perhaps getting a sense of the prior estimates but the process would be quite laborious. For the 61 articles in the right-hand column of Table 3, there is no way for a reader to understand the sample size calculation, especially for the 58 articles that provided no details at all.

In summary, only 4 of the 158 articles (3%) reported all of the information needed to understand the sample size calculation, but an additional 63 articles (40%) offered almost complete information. Eleven articles (7%) provided incomplete information but described the critical rates that would allow the reader to evaluate the outcome of the trial. Nineteen additional articles (12%) offered information that might allow for a general sense of the sample size calculation, but was too limited for full evaluation of the results. Sixty-one articles (39%) did not provide information that would allow any realistic attempt at understanding the sample size calculation. There was no correlation between the actual sample sizes in the trials and the extent of reporting of the sample size calculations.

Overall, 78 of the articles (49%) provided enough information for readers to understand the important components of the sample size calculation.

For the 78 articles which presented their prior designations of event rates, if significant differences were found, readers could evaluate the magnitude of the P values reported Table 4 shows that of the 158 total articles reviewed, 54 both described π_c and δ and reported statistically significant outcomes. The following sections will compare prior estimates and reported results for these 54 articles.

C. Example of P Value Calculation

The P value used to determine statistical significance is based on a Z-score derived from the following equation:

$$Z = \frac{d_0}{SED} = \frac{p_c - p_e}{\sqrt{\bar{p}(1-\bar{p})\left[\frac{1}{n_c} + \frac{1}{n_e}\right]}} \qquad (2)$$

In this equation, the numerator, d_0, is equal to p_c (the outcome rate in the control group) minus p_e (the outcome rate in the experimental group). The denominator for this calculation is the standard error of the difference between groups (SED). For its calculation, $\bar{p}$ is the average outcome rate (i.e. the average of p_c and p_e), n_c is the number of patients in the control group, and n_e is the number of patients in the experimental group.

Table 4: Criteria for inclusion of articles in analysis of δ versus d_0

	Articles reporting a statistically significant d_0 with a P value		
	Yes	No	TOTAL
Articles that reported both π_c and δ	**54**	24	78
Articles that did not report both π_c and δ	44	36	80
TOTAL	98	60	158

Higher Z-scores correspond to lower P values; for example a Z-score of 1.645 corresponds to a two-tailed P value of 0.10 while a Z-score of 1.96 yields the familiar two-tailed P value of 0.05. Equation 2 shows that a Z-score could increase in three ways. An increase in d_0 would enlarge the numerator; alternatively, either an increase in n_c or n_e or a decrease in the variance would reduce the denominator. The next section will compare the range of observed d_0 values and the anticipated δ values in this group of articles.

D. Differences Between d_0 and δ

Table 5 shows that, of the 54 articles which reported both prior designations of event rates and statistically significant outcomes, the observed value of d_0 was greater than or equal to the prior assignment of δ in 29 cases, but 25 articles reported statistical significance for a d_0 that was smaller than the prior designation of δ. We have named this latter phenomenon "δ (delta) wobble" and will explain the term more completely later in the Discussion. In four extreme instances, the observed d_0 was at least 50% smaller than the anticipated δ. As an example of "δ wobble," in one of the articles[11] included in the second-to-last row of Table 5, a δ value of 0.30 was designated for the purposes of sample size calculation, but a statistically significant d_0 of 0.152 was presented in the results section. Thus for this article d_0 was smaller than δ by 49% [i.e. $(d_0 - \delta)/\delta = (0.152-0.30)/0.30 = -0.49$]. At the other extreme, in one of the articles[12] included in the first row of Table 5, a δ of 0.20 was

Table 5: Frequency of values for the proportionate difference $(d_0-\delta)/\delta$ (Negative percentage for $d_0<\delta$, positive percentage for $d_0>\delta$)

Percent difference between d_0 and δ	Number of articles
>50%	10
25% to 50%	6
0% to 25%	12
0%	1
-25% to 0%	10
-50% to -25%	11
<-50%	4

designated for sample size calculation, but a d_0 of 0.37 was presented in the results section. For this article d_0 was greater than δ by 85% [(0.37-0.20)/0.20 = 0.85].

The final row of Table 5 show that of the 25 articles with "δ wobble," i.e. a d_0 value less than the prior designation of δ, four presented a statistically significant d_0 that was less than half as large as δ (first row). An additional 11 articles presented d_0 values that were smaller than δ by a proportionate increment between one-quarter and one-half (second-to-last row). Seven of the 25 articles with "δ wobble" presented a statistically significant d_0 value that was less that δ by an absolute increment of more than 0.10.

Table 6 shows the relationship of discrepancies between d_0 and δ to the magnitude of P values reported. As noted previously, increased d_0 values could cause very small P values. The first two rows of Table 6 show that of the 31 trials which reported δ and had P values less than or equal to 0.01, 10 had d_0 values more than 25% larger than the prior designation of δ, but 12 of the 31 trials with $P \leq 0.01$ had d_0 values that were smaller than the prior designation of δ. Table 6 demonstrates that the very small P values are not restricted to trials reporting d_0 much larger than δ. Indeed, the final three rows show that even some of the trials that commit "δ wobble" report very small P values.

Table 6: Frequency of values for the proportionate difference $(d_0-\delta)/\delta$, categorized by magnitude of reported P-value in statistically significant trials. Last row contains those articles which did not report original designation of δ.

Percent difference between d_0 and δ	$P \leq 0.001$	$0.001 \leq P \leq 0.01$	$0.01 \leq P \leq 0.05$	Total
>50%	4	2	4	10
25% to 50%	3	1	2	6
0% to 25%	2	6	4	12
0%	1	0	0	1
-25% to 0%	0	6	4	10
-50% to -25%	1	4	6	11
<-50%	0	1	3	4
Total reporting δ	11	20	23	54
No δ described	16	15	13	44
Grand Total	27	35	36	98

IV. DISCUSSION

A. Reporting of Sample Size Calculations

The current finding, that 43% (67/158) of a group of published randomized controlled trials presented full details of sample size calculations, was only slightly higher than the 39% found in a 1990 review [6] and was identical to the rate noted in an analogous review published in 1994. [8] The finding that 39% (58/158) of articles reported no details of sample size calculation was also similar to the rate noted in a 1990 review [6].

Although the details of sample size calculation are now reported more often than noted in the first such reviews almost 20 years ago [2], reporting is still far from complete. In a 1994 review, Moher et al. suggested:

> that authors should report sample size calculations and that the following information should be contained in all published reports of RCTs: (1) The primary dependent measure(s) should be clearly identified. (2) A clinically important treatment effect should be specified. (3) The treatment effect should be clearly indicated as being an absolute or a relative difference. (4) The statistical test, directionality, α level, and statistical power used to estimate sample size should be reported. [8, p. 124]

This suggestion was re-iterated in an article later in 1994 by The Standards of Reporting Trials Group, an international committee established to address reporting of randomized controlled trials [13]. In light of the consequences that are about to be discussed, editors might follow the recommendations of Moher et al. and become more

demanding in asking authors to report their pre-trial assumptions when sample size was calculated.

B. Very Small P Values

As noted in Table 2, more than 25% of the trials with statistically significant results reported P values that were ≤0.001. As shown in Equation 2, these excessively small P values could have come from unexpectedly large values of d_0, but Table 6 shows that only 10 (32%) of the 31 articles with P≤0.01 found d_0 to be substantially larger than δ. The remainder of this section will show, however, that even for those 10 articles, the large d_0 values were not likely to have caused the excessively small P values.

1. An Example of an Article with d_0 Much Larger Than δ

In one article where δ was designated as 0.20 but d_0 was found to be 0.37, this distinction was reported as having a P<0.001 [12]. The Z-score for this result can be calculated using Equation 2. In the article cited, p_c was 0.77, p_e was 0.40, n_c was 104 and n_e was 95 [12]. $\bar{p}$ can be approximated as the weighted average of 0.77 and 0.40, which yields 0.593*. The Z-score is thus:

---

* This value is calculated as an average weighted by the number of subjects in each group:

$$\bar{p} = \frac{n_c p_c + n_e p_e}{n_c + n_e} = \frac{(104 \times 0.77) + (95 \times 0.40)}{104 + 95} = 0.593$$

$$Z = \frac{0.77 - 0.40}{\sqrt{0.593(1-0.593)\left[\frac{1}{104} + \frac{1}{95}\right]}} = 5.31 \qquad (3)$$

Since a Z-score of 3.80 corresponds to a P value of $1x10^{-4}$, this finding would not only result in P<0.001, but would yield an infinitesimally small P value in the range of 1.0×10^{-7} [14, p. 31].

2. *Effect of Discrepancy Between d_0 and δ on P Values*

If the d_0 in the above example had been the expected 0.20 instead of 0.37, i.e., if p_c was 0.60, the Z-score would be calculated as follows:

$$Z = \frac{0.60 - 0.40}{\sqrt{0.505(1-0.505)\left[\frac{1}{104} + \frac{1}{95}\right]}} = 2.82 \qquad (4)$$

The corresponding P value would be less than 0.005[15, p. 281]. Although much larger than the previous result for P, this value is still 10 times smaller than the boundary of 0.05 for achieving statistical significance.

3. *Effect of Discrepancy Between p_c and π_c on P Values*

In the cited article, however, not only was d_0 much larger than δ, but the observed event rates themselves were considerably larger than the values assigned prior to the study. The authors had previously estimated rates of 0.25 for the control group and 0.05

for the experimental group [12]. If these outcome rates had actually been found, the Z-score calculation would have shown:

$$Z = \frac{0.25 - 0.05}{\sqrt{0.155(1-0.155)\left[\frac{1}{104} + \frac{1}{95}\right]}} = 3.89, \qquad (5)$$

which corresponds to a P value of 1×10^{-4} [14, p. 31].

In this example the discrepancy between the observed and estimated event rates affected the variance term in the denominator. The closer $\bar{p}$ is to 0.50, the larger the variance will be and the smaller the Z-score will be. The discrepancy between predicted and reported rates in this example moved $\bar{p}$ very close to 0.50, but the P value was still reported as statistically significant by a wide margin. The more striking point, however, is that the Z-score calculated in Equation 5 represents the clinical outcome expected by the authors, but the corresponding P value is extremely small. If the very small P value is due neither to large d_0 values nor to discrepancies between p_c and π_c, the only remaining cause for the small P values is excessively large sample sizes.

C. The Phenomenon of "δ Wobble"

In addition to showing that increased d_0 values seldom cause very small P values, Table 5 and Table 6 reveal an additional phenomenon. As noted in Section III.D, 25 of the 54 articles in Table 5 reported d_0 values smaller than δ; indeed 15 of the "statistically significant" d_0 values were at least 25% smaller than δ. If the δ value entered into the

Neyman-Pearson equation represents the minimum boundary to be regarded as clinically significant, the frequent citation of "significance" for d_0 values much smaller than the pre-assigned δ calls into question the initial design of the trial. The remainder of this section will show a hypothetical sample size calculation to demonstrate that the Neyman-Pearson equation converts δ into a "wobbly" parameter. The calculated sample sizes will allow values of d_0 much smaller than δ to be declared statistically significant.

The values in the example below are chosen arbitrarily, but the results will hold for any set of values if readers want to replicate the exercise. The Neyman-Pearson calculation shown in Equation 1 in Section II.B is the standard method used for calculating sample sizes. The elements of the calculation were described in Section II.B and will not be repeated here.

1. Calculation of the Sample Size

For this example, I will assume that the mortality rate of 0.20 with current therapy for disease X is to be tested against a new treatment that is expected to reduce the mortality rate to 0.10. Following convention, the researchers designate an α value of 0.05 (two-tailed) and a β value of 0.10 (one-tailed) for the purposes of sample size calculation. The sample size calculated using the Neyman-Pearson equation will be as follows:

$$n \geq \frac{(1.96+1.282)^2 \times [2 \times 0.2 \times (1-0.2)]}{0.10^2} = 336 \qquad (6)$$

The researchers therefore recruit a total of 672 patients to their study, 336 to receive current therapy and 336 to receive experimental treatment. With this backdrop, I will now consider possible outcomes.

2. *Scenario 1: d_0 Equal to δ*

In the first scenario, the researchers find exactly what they had expected. Mortality in the control group is 67/336, or 0.199, while mortality in the experimental group is 34/336, or 0.101. To test the statistical significance of these findings, the Z score is calculated with Equation 2 to show:

$$Z = \frac{0.199 - 0.101}{\sqrt{0.15(1-0.15)\left[\frac{1}{336} + \frac{1}{336}\right]}} = 3.56 \qquad (7)$$

The corresponding two-tailed P value for this result is 0.0002[15, p. 280], which is much smaller than the anticipated P=0.05, although the d_0 found by the researchers almost exactly equals the prior designation of δ.

3. *Scenario 2: d_0 smaller than δ, but is Statistically Significant*

In the second scenario, mortality in the control group is again 67/336, or 0.199, but mortality in the experimental group is 47/336, or 0.140. This d_0 of 0.059 is much smaller than what was hoped, but calculation of a Z score reveals:

$$Z = \frac{0.199 - 0.140}{\sqrt{0.170(1-0.170)\left[\frac{1}{336} + \frac{1}{336}\right]}} = 2.04 \qquad (8)$$

This corresponds to a P value of 0.0414[15, p. 280]. Armed with this result, the investigators can now present this d_0 of 0.059 as statistically significant, although it is almost half of the δ value of 0.10 which they designated as a difference worth finding before the trial began.

4. Scenario 3: d_0 Smaller than δ, but not Statistically Significant

In a third scenario, control group mortality remains at 67/336 (0.199), but one additional experimental group patient dies, so that experimental group mortality is 48/336, or 0.143. The d_0 of 0.056 is again smaller than the prior designation of δ. Calculation of the Z-score shows:

$$Z = \frac{0.199 - 0.143}{\sqrt{0.171(1-0.171)\left[\frac{1}{336} + \frac{1}{336}\right]}} = 1.93 \qquad (9)$$

This corresponds to a P value that is slightly greater than 0.05[15, p. 280], so that the investigators cannot claim the result is significantly different from zero. Persevering, the researchers recall that the Z_β term in the sample size calculation was 1.282, corresponding to a one-tailed β of 0.10. A Z-score for the alternate hypothesis can be calculated using the following formula:

$$Z_A = \frac{\delta - d_0}{SED} \qquad (10)$$

This is almost the same Z-score formula shown in Equation 2, but differing in the use of the quantity $\delta - d_0$ in the numerator*, and it results in the following calculation:

$$Z_A = \frac{0.10 - 0.056}{\sqrt{0.171(1-0.171)\left[\frac{1}{336} + \frac{1}{336}\right]}} = 1.51 \qquad (11)$$

Since this value is greater than the threshold value of 1.282 used in the sample size calculation, the investigators can reject the alternate hypothesis of a large difference between groups. The investigators can now claim they have proven that there is no important difference between the two treatments, as their results excluded a difference of 0.10.

5. *A Zone of Double Significance*

A particularly interesting result arises, however, if we return to the scenario in Section IV.C.3, and to the Z-score calculated in Equation 8. If, for that same result, the researchers had calculated a Z_A for the alternative hypothesis:

$$Z_A = \frac{0.10 - 0.059}{\sqrt{0.170(1-0.170)\left[\frac{1}{336} + \frac{1}{336}\right]}} = 1.41 \qquad (12)$$

---

* The SED for the alternate hypothesis is properly calculated as follows:

$$SED_A = \sqrt{\frac{p_c q_c}{n_c} + \frac{p_e q_e}{n_e}}.$$

In this example, both the above calculation and the calculation shown in the denominator of Equation 11 produce a value of 0.029. The standard calculation of SED will be used for the remaining examples.

This result is greater than the prior Z_β value of 1.282, and the investigators can declare that they have proven that there is no important difference between groups, as this result excludes a difference of 0.10. For this result, therefore, the investigators have achieved double statistical significance. The difference between groups is both statistically significantly greater than zero (Equation 8) and is also statistically significantly smaller than the difference initially defined as clinically significant (Equation 12).

6. Neyman-Pearson Equation Shifts Threshold for Statistical Significance

Although the investigators stated initially that a difference of 0.10 between treatments represented the boundary for clinical importance, the sample size calculated with the Neyman-Pearson equation would in fact allow them to declare statistical significance for a d_0 as small as 0.059, or any larger value. At that same level of d_0, and at any smaller value, the investigators could declare that their result was statistically significantly smaller than 0.10. The crucial observation here is that when sample size is calculated with the Neyman-Pearson equation for this example, the investigators will find a statistically significant result no matter what value emerges for d_0. The particular values shown in Section IV.C.5 will even achieve double significance; thus, the threshold for significance is no longer the clinical threshold that was used in trial design. In addition, when the final result does equal the threshold defined as clinically significant for trial design, the P value is orders of magnitude smaller than the 0.05 value that conventionally determines statistical significance.

D. Calculation of Implicit Thresholds

The example in the preceding section demonstrated some of the consequences of using a sample size generated by the Neyman-Pearson calculation. In this section I will show how those consequences arise. The crux of this argument is that the Neyman-Pearson calculation uses both Z_α and Z_β, thereby incorporating both null and alternate hypotheses in one formula. In the analysis of results, however, these hypotheses are evaluated separately, each with an individual calculation.

If one calculated a sample size considering only the possibility of type I error, the sample size calculation would be as follows:

$$n \geq \frac{(Z_\alpha)^2 \times [2 \times \pi \times (1-\pi)]}{\delta^2} \qquad (13)$$

This differs from the Neyman-Pearson calculation in that Z_β is not included. Conversely, if one calculated a sample size with concern for type II error only, the following equation would be used:

$$n \geq \frac{(Z_\beta)^2 \times [2 \times \pi \times (1-\pi)]}{(\delta - d_0)^2} \qquad (14)$$

In this case, Z_α has been excluded from the numerator. Additionally, the denominator term (δ - d_0) reflects the increment at which a difference of δ will be ruled out at the Z_β level of significance.

Combining the previous two equations with the results of the example from the previous section, I will now demonstrate that the thresholds for statistical significance are not those demarcated in the Neyman-Pearson equation. In Equation 6, the calculated sample size was 336 patients per group. With this sample size, consider Equation 13. Inserting 336 for N, 1.96 for Z_α, and keeping the same term in the denominator, the result is as follows:

$$336 = \frac{(1.96)^2 \times [2 \times 0.2 \times (1-0.2)]}{\delta^2} \qquad (15)$$

Rearranging equation 15 to solve for δ produces:

$$\delta = \sqrt{\frac{(1.96)^2 \times [2 \times 0.2 \times (1-0.2)]}{336}} = 0.060 \qquad (16)$$

The implication of the above calculation is that although a δ value of 0.10 was designated when the Neyman-Pearson equation was used to calculate sample size, the result in fact represents an implicit δ designation of 0.060. In other words, although the study design designates a difference of 0.10 as clinically important, differences as small as 0.060 will be found to be statistically significant. This result explains both the findings summarized in section III.D and the outcome of the example trial presented in section IV.C.

The same process can be utilized for the calculation shown in Equation 14. Again combining that equation with the result of Equation 6, but omitting the intermediate steps, the final result will be as follows:

$$\delta - d_0 = \sqrt{\frac{(1.282)^2 \times [2 \times 0.2 \times (1-0.2)]}{336}} = 0.040 \qquad (17)$$

Since the prior designation of δ in this example was 0.10, this result shows that for any d_0 value of 0.060 or smaller, the Z_A value for the alternate hypothesis will be greater than 1.282 and the alternate hypothesis of a large difference between control and experimental groups will be rejected.

The results of Equations 16 and 17 converge around a d_0 value of 0.060. The preceding paragraphs show that for the sample size determined with the Neyman-Pearson calculation in Equation 6, any value larger than 0.060 will be statistically significantly greater than 0, while any value smaller than 0.060 will be statistically significantly smaller than 0.10. In fact, since the SED is used for Z-score calculation (see Equation 2) and the variance of the control group rate is used in the standard form of the Neyman-Pearson formula (see Equation 1), the actual numbers in Section IV.C stretch further, so that values somewhat smaller than 0.060 will be statistically significantly greater than 0, and some of these values will also be statistically significantly smaller than 0.10. In practical terms, then, the Neyman-Pearson sample size calculation has guaranteed a statistically significant result in one direction or the other, and has even created a zone of double significance. The value 0.060 has now become the threshold at which the therapy being studied will be declared effective or ineffective, although this value is barely half of the value originally designated by the investigators. This result is not particular to the numbers chosen for this example, and will be found with any other set of values that are chosen for the illustration.

V. CONCLUSION

This paper has attempted to show consequences of excess sample size in large randomized controlled trials. Since a landmark paper by Freiman et al. almost 20 years ago, most discussion of sample sizes has focused on the need for larger samples. The results of the present study, however, suggest that enlargement of sample sizes, and in particular the use of the Neyman-Pearson equation to calculate these large samples, may have two important unintended consequences.

The first consequence is that with the large sample sizes generated by the Neyman-Pearson equation, many results will produce extremely small P values, despite the general acceptance of 0.05 as the threshold for a statistically significant finding. As noted in the introduction, an overly large sample generates excess cost and requires an excessive number of patients. As research funding dwindles, the excessive costs of oversized trials represent a substantial overuse of resources.

The second consequence of sample sizes calculated with the Neyman-Pearson equation is that the quantitative threshold for an impressive difference is reduced to almost half of the initial level. Thus, whatever δ investigators designate at the outset of the trial, the d_0 that can be declared statistically significant will be considerably smaller than the original value of δ; we have named this phenomenon "δ wobble." The "wobbliness" of the boundary for clinical significance defined for sample size calculation undermines the clinical judgment used in originally defining δ.

The problems of excessively small P values and "δ wobble" suggest that the Neyman-Pearson strategy of sample size calculation requires serious reevaluation. It is beyond the scope of this paper to suggest alternative methods for sample size calculation. Until such methods are developed, however, editors can address the problem of "δ wobble" by requiring investigators to state both δ and d_0, and to justify reporting of statistically significant d_0 values which are smaller than the δ initially designated as clinically significant.

REFERENCES

1. Fisher RA. Statistical methods for research workers. (14 ed.) Darien, Conn: Hafner Publishing Company, 1970.

2. Ambroz A, Chalmers TC, Smith H, Schroeder B, Freiman JA, Shareck EP. Deficiencies of randomized controlled trials. Clinical Research 1978;26:280A.

3. DerSimonian R, Charette LJ, McPeek B, Mosteller F. Reporting on methods in clinical trials. New England Journal of Medicine 1982;306(22):1332-7.

4. Liberati A, Himel HN, Chalmers TC. A quality assessment of randomized controlled trials of primary treatment of breast cancer. Journal of Clinical Oncology 1986;4(6):942-951.

5. Pocock SJ, Hughes MD, Lee RJ. Statistical problems in the reporting of clinical trials. New England Journal of Medicine 1987;317(7):426-32.

6. Altman DG, Dore CJ. Randomisation and baseline comparisons in clinical trials. Lancet 1990;335:149-53.

7. Schulz KF, Chalmers I, Grimes DA, Altman DG. Assessing the quality of randomization from reports of controlled trials published in obstetrics and gynecology journals. Journal of the American Medical Association 1994;272(2):125-128.

8. Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. Journal of the American Medical Association 1994;272(2):122-124.

9. Altman DG. Statistics and ethics in medical research. British Medical Journal 1980;281:1336-1338.

10. Freiman JA, Chalmers TC, Smith H, Jr., Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. New England Journal of Medicine 1978;299(13):690-694.

11. UK Gabapentin Study Group. Gabapentin in partial epilepsy. Lancey 1990;335:1114-17.

12. Crawford J, Ozer H, Stoller R, et al. Reduction by granulocyte colony-stimulating factor of fever and neutropenia induced by chemotherapy in patients with small-cell lung cancer. New England Journal Of Medicine 1991;325(3):164-70.

13. The Standards of Reporting Trials Group. A proposal for structured reporting of randomized controlled trials. JAMA 1994;272(24):1926-1931.

14. Diem K, Lentner C, eds. Scientific Tables. Ardsley, New York: Geigy Pharmaceuticals, 1974.

15. Jekel JF, Elmore JG, Katz DL. Epidemiology, biostatistics, and preventive medicine. Philadelphia: W.B. Saunders, 1996.

APPENDIX A: ARTICLES REVIEWED

1. AIMS Trial Study Group. Long-term effects of intravenous antistrpelase in acute myocardial infarction:final report of the AIMS study. Lancet 1990;335:427-31.
2. Applegate WB, Miller ST, Graney MJ, Elam JT, Burns R, Akins DE. A randomized, controlled trial of a geriatric assessment unit in a community rehabilitation hospital. N Engl J Med 1990;322(22):1572-8.
3. Arshad SH, Matthews S, Gant C, Hyde DW. Effect of allergen avoidance on development of allergic disorders in infancy. Lancet 1992;339:1493-97.
4. Baker CJ, Melish ME, Hall RT, et al. Intravenous immune globulin for the prevention of nosocomial infection in low-birth-weight infants. N Engl J Med 1992;327(4):213-9.
5. Ballard RA, Ballard PL, Creasy RK, et al. Respiratory disease in very-low-birthweight infants after prenatal thyrotropin-releasing hormone and glucocorticoid. Lancet 1992;339:510-15.
6. Barrett BJ, Parfrey PS, Vavasour HM, O'Dea F, Kent G, Stone E. A comparison of nonionic low-osmolality radiocontrast agents with ionic, high-osmolality agents during cardiac catheterization. N Engl J Med 1992;326(7):431-6.
7. Beck R, Cleary P, Anderson M, et al. A randomized, controlled trial of corticosteroids in the treatment of acute optic neuritis. N Engl J Med 1992;326(9):581-8.
8. Belizan JM, Villar J, Gonzalez L, Campodonico L, Bergel E. Calcium supplementation to prevent hypertensive disorders of pregnancy. N Engl J Med 1991;325(20):1399-405.
9. Bloomfield P, Wheatley DJ, Prescott RJ, Miller HC. Twelve-year comparison of a Bjork-Shiley mechanical heart valve with porcine bioprosthesis. N Engl J Med 1991;324(9):573-9.
10. Bozzette SA, Sattler FR, Chiu J, et al. A controlled trial of early adjunctive treatment with corticosteroids for pneumocystis carinii pneumonia in the acquired immunodeficiency syndrome. N Engl J Med 1990;323(21):1451-7.
11. Brain Resuscitation Clinical Trial II Study Group. A randomized clinical study of a calcium-entry blocker (lidoflazine) in the treatment of comatose survivors of cardiac arrest. N Engl J Med 1991;324(18):1225-31.
12. Brown CG, Martin DR, Pepe PE, et al. A comparison of standard-dose and high-dose epinephrine in cardiac arrest outside the hospital. N Engl J Med 1992;327(15):1051-5.
13. Buchwald H, Varco RL, Matts JP, Long JM, Fitch LL, Campbell GS. Effect of partial ileal bypass surgery on mortality and morbidity from coronary heart disease in patients with hypercholesterolemia. N Engl J Med 1990;323(14):946-55.
14. Burgman LG, Christensen P, Christensen K, et al. Prevention of excess neonatal morbidity associated with broup B streptococci by vaginal chlorhexadine disinfection during labour. Lancet 1992;340:65-69.
15. Cabanes PA, Salmon RJ, Vilcoq JR, et al. Value of axillary dissection in addition to lumpectomy and radiotherapy in early breast cancer. Lancet 1992;339:1245-48.

16. Cady RK, Wendt JK, Kirschner JR, Sargent JD, Rothrock JF, Skaggs H, Jr. Treatment of acute migraine with subcutaneous sumatriptan. JAMA 1991;265(21):2831-2835.
17. Callaham M, Madsen CD, Barton CW, Saumders CE, Pointer J. A randomized clinical trial of high-dose epinephrine and norepinephrine vs standard-dose epinephrine in prehospital cardiac arrest. JAMA 1992;268(19):2667-2672.
18. Canellos GP, Anderson JR, Propert KJ, et al. Chemotherapy of advanced Hodgkin's disease with MOPP, ABVD, or MOPP alternating with ABVD. N Engl J Med 1992;327(21):1478-84.
19. Chapuy MC, Arlot ME, Duboeuf F, et al. Vitamin D3 and calcium to prevent hip fractures in elderly women. N Engl J Med 1992;327(23):1637-42.
20. Cocconi G, Bella M, Calabresi F, et al. Treatment of metastatic melanoma with dacarbazine plus tamoxifen. N Engl J Med 1992;327(8):516-23.
21. Cohn JN, Johnson G, Ziesche S, et al. A comparison of enalapril with hydralazine-isosorbide dinitrate in the treatment of chronic congestive heart failure. N Engl J Med 1991;325(5):303-10.
22. Condylomata International Collaborative Study Group. Recurrent condylomata acuminata treated with recombinant interferon alfa-2a: a multicenter double-blind placebo-controlled clinical trial. JAMA 1991;265(20):2684-2687.
23. Courtney MG, Nunes DP, Bergin CF, et al. Randomised comparison of olsalazine and mesalazine in prevention of relapses in ulcerative colitis. Lancet 1992;339:1279-81.
24. Crawford J, Ozer H, Stoller R, et al. Reduction by granulocyte colony-stimulating factor of fever and neutropenia induced by chemotherapy in patients with small-cell lung cancer. N Engl J Med 1991;325(3):164-70.
25. Czeizel AE, Dudas I. Prevention of the first occurrence of neural-tube defects by periconceptional vitamin supplementation. N Engl J Med 1992;327(26):1832-5.
26. Dahlof B, Lindholm LH, Hansson L, Schersten B, Ekbom T, Wester P-O. Morbidity and mortality in the Swedish trial in old patients with hypertension (STOP-hypertension). Lancet 1991;338:1281-85.
27. Dillman RO, Seagren SL, Propert KJ, et al. A randomized trial of induction chemotherapy plus high-dose radiation versus radiation alone in stage III non-small-cell lung cancer. N Engl J Med 1990;323(14):940-5.
28. Echt DS, Liebson PR, Mitchell B, et al. Mortality and morbidity in patients receving encainide, flecainide, or placebo. N Engl J Med 1991;324(12):781-8.
29. England MR, Gordon G, Salem M, Chernow B. Magnesium administration and dysrhythmias after cardiac surgery. JAMA 1992;268(17):2395-2402.
30. ETDRS Investigators. Aspirin effects on mortality and morbidity in patients with diabetes mellitus: early treatment diabetic retinopathy study report 14. JAMA 1992;268(10):1292-1300.
31. European Carotid Surgery Trialists' Collaborative Group. MRC European Carotid Surgery Trial: interim results for symptomatic patients with severe (70-99%) or with mild (0-29%) carotid stenosis. Lancet 1991;337:1235-43.
32. European CGRP in Subarachnoid Haemorrhage Study Group. Effect of calcitonin-gene-related peptide in patients with delayed postoperative cerebral ischaemia after aneurysmal subarachnoid haemorrhage. Lancet 1992;339:831-34.

33. Ezekowitz MD, Bridgers SL, James KE, et al. Warfarin in the prevention of stroke associated with nonrheumatic atrial fibrillation. N Engl J Med 1992;327(20):1406-12.
34. Fielding LP, Hittinger R, Grace RH, Fry JS. Randomised controlled trial of adjuvant chemotherapy by portal-vein perfusion after curative resection for colorectal adenocarcinoma. Lancet 1992;340:502-06.
35. Fiessinger JN, Schafer M. Trial of ilioprost versus Aspirin treatment for critical limb ischemia of thromboangiitis obliterans. Lancet 1990;335:555-57.
36. Francis CW, Pellegrini VD, Marder VJ, al e. Comparison of warfarin and external pneumatic compression in prevention of venous thrombosis after total hip replacement. JAMA 1992;267(21):2911-2915.
37. Gastinne H, Wolff M, Delatour F, Faurisson F, Chevret S. A controlled trial in intensive care untis of selective decontamination of the digestive tract with nonabsorbable antibiotics. N Engl J Med 1992;326(9):594-9.
38. Giannini EH, Brewer EJ, Kuzmina N, et al. Methotrexate in resistant juvenile rheumatoid arthritis. N Engl J Med 1992;326(16):1043-9.
39. Goodman JL, Winston DJ, Greenfield RA, et al. A controlled trial of fluconazole to prevent fungal infectionsin patients undergoing bone marrow transplantation. N Engl J Med 1992;326(13):845-51.
40. Gordon LI, Harrington D, Andersen J, et al. Comparison of a second-generation combination chemotherapeutic regimen (m-BACOD) with a standard regimen (CHOP) for advanced diffuse non-Hodgkin's lymphoma. N Engl J Med 1992;327(19):1342-9.
41. Greenman RL, Schein RMH, Martin MA, al e. A controlled clinical trial of E5 murine monoclonal IgM antibody to endotoxin in the treatment of gram-negative sepsis. JAMA 1991;266(8):1097-1102.
42. Grimm RH, Jr, Neaton JD, Elmer PJ, et al. The influence of oral potassium chloride on blood pressure in hypertensive men on a low-sodium die. N Engl J Med 1990;322(9):569-74.
43. Gruppo Italiano per lo Studio della Sopravvivenza nell'Infarcto Miocardio. GISSI-2: a factorial randomised trial of alteplase versus streptokinase and heparin versus no heparin among 12 490 patients with acute myocardial infarction. Lancet 1990;335:65-71.
44. Gutierrez G, Palizas F, Doglio G, et al. Gastric intramucosal pH as a therapeutic index of tissue oxygenation in critically ill patients. Lancet 1992;339:195-99.
45. Hall JC, Tarala R, Harris J, Tapper J, Chistiansen K. Incentive spirometry versus routine chest physiotherapy for prevention of pulmonary complications after abdominal surgery. Lancet 1991;337:953-56.
46. Hallman M, Bry K, Hoppu K, Lappi M, Pohjavuori M. Inositol sopplementation in premature infants with respiratory distress syndrome. N Engl J Med 1992;326(19):1233-9.
47. Hamilton JD, Hartigan PM, Simberkoff MS, et al. A controlled trial of early versus late treatment with zidovudine in symptomatic human immunodeficiency virus infection. N Engl J Med 1992;326(7):437-43.
48. Hammond JMJ, Potgieter PD, Saunders GL, Foster AA. Double-blind study of selective decontamination of the digestive tract in intensive care. Lancet 1992;340:5-9.
49. Hannah ME, Hannah WJ, Hellmann J, et al. Induction of labor as compared with serial antenatal monitoring in post-term pregnancy. N Engl J Med 1992;326(24):1587-92.

50. Hardy WD, Feinberg J, Finkelstein DM, et al. A controlled trial of trimethorim-sulfamethoxazole or aerosolized pentamidine for secondary prophylaxis of pneumocystis carinii pneumonia in patients with the acquired immunodeficiency syndrome. N Engl J Med 1992;327(26):1842-8.
51. Herrera MG, Nestel P, El Amin A, Fawzi WW, Mohamed KA, Weld L. Vitamin A supplementation and child survival. Lancet 1992;340:267-71.
52. Hersh EM, Brewton G, Abrams D, et al. Ditiocarb sodium (diethyldithiocarbamate) therapy in patients with symptomatic HIV infection and AIDS. JAMA 1991;265(12):1538-1544.
53. Herskovic A, Martz K, Al-Sarraf M, et al. Combined chemotherapy and radiotherapy compared with radiotherapy alone in patients with cancer of the esophagus. N Engl J Med 1992;326(24):1593-8.
54. Hirschel B, Lazzarin A, Chopard P, et al. A controlled study of inhaled pentamidine for primary prevention of pneumocystis carinii pneumonia. N Engl J Med 1991;324(16):1079-83.
55. Hogg KJ, Gemmill JD, Burns JMA, et al. Angiographic patency study of antistreplase versus streptokinase in acute myocardial infarction. Lancet 1990;335:254-58.
56. Hong WK, Lippman SM, Itri LM, et al. Prevention of second primary tumors with isotretinoin in squamous-cell carcinoma of the head and neck. N Engl J Med 1990;323(12):795-801.
57. Hsia J, Hamilton WP, Kleiman N, et al. A comparison between heparin and low-dose aspirin as adjunctive therpay with tissue plasminogen activator for acute myocardial infarction. N Engl J Med 1990;323(21):1433-7.
58. Hull RD, Raskob GE, Gent M, et al. Effectiveness of intermittent pneumatic leg compression for preventing deep vein thrombosis after total hip replacement. JAMA 1990;263(17):2313-2317.
59. Hussey GD, Klein M. A randomized, controlled trial of vitamin A in children with sever measles. N Engl J Med 1990;323(3):160-4.
60. Italian Group for Antiemetic Research. Ondansetron+dexamethasone vs metoclopramide+dezamethasone+diphenhydramine in prevention of cisplatin-induced emesis. Lancet 1992;340:96-99.
61. Johnson RE, Jaffe JH, Fudala PJ. A controlled trial of buprenorphine treatment for opioid dependence. JAMA 1992;267(20):2750-2755.
62. Jones AL, Hill AS, Soukop M, et al. Comparison of dexamethasone and ondansetron in the prophylaxis of emesis induced by moderately emetogenic chemotherapy. Lancet 1991;338:483-87.
63. Juul-Moller S, Edvardsson N, Jahnmatz B, Rosen A, Sorensen S, Omblus R. Double-blind trial of aspirin in primary prevention of myocardial infarction in patients with stable chronic angina pectoris. Lancet 1992;340:1421-25.
64. Kahn JO, Lagakos SW, Richman DD, et al. A controlled trial comparing continued zidovudine with didanosine in human immunodeficiency virus infection. N Engl J Med 1992;327(9):581-7.
65. Kamal GD, Pfaller MA, Rempe LE, Jebson PJR. Reduced intravascular catheter infection by antibiotic bonding. JAMA 1991;265(18):2364-2368.

66. Kaye SB, Lewis CR, Paul J, et al. Randomised study of two doses of cisplatin with cyclophosphamide in epithelial ovarian cancer. Lancet 1992;340:329-33.
67. Kendig JW, Notter RH, Cox C, et al. A comparison of surfactant as immediate prophylaxis and as rescue therapy in newborns of less than 30 weeks' gestation. N Engl J Med 1991;324(13):865-71.
68. Kennell J, Klaus M, McGrath S, Robertson S, Hinkley C. Continuous emotional support during labor in a US hospital. JAMA 1991;265(17):2197-2201.
69. Kreiss J, Ngugi E, Holmes K, al e. Efficacy of nonoxynol 9 contraceptive sponge use in preventing heterosexual acquisition of HIV in Nairobi prostitutes. JAMA 1992;268(4):477-482.
70. Krook JE, Moertel CG, Gunderson LL, et al. Effective surgical adjuvant therapy for high-risk rectal carcinoma. N Engl J Med 1991;324(11):709-15.
71. Lahtinen R, Laakso M, Palva I, Vrikkunen P, Elomaa I. Randomised, placebo-controlled multicentre trial of clodronate in multiple myeloma. Lancet 1992;340:1049-52.
72. Lammer J, Pilger E, Decrinis M, Quehenberger F, Klein GE, Stark G. Pulsed excimer laser versus continuous-wave Nd:YAG laser versus conventional angioplasty of peripheral arterial occlusions: prospective, controlled, randomised trial. Lancet 1992;340:1183-88.
73. Leoung GS, Feigal DW, Jr., Montgomery AB, et al. Aerosolized pentamidine for prophylaxis against pneumocystis carinii pneumonia. N Engl J Med 1990;323(12):769-75.
74. Light RW, O'Hara VS, Moritz TE, et al. Intrapleural tetracycline for the prevention of recurrent spontaneous pneumothorax. JAMA 1990;264(17):2224-2230.
75. Long W, Corbet A, Cotton R, et al. A controlled trial of synthetic surfactant in infants weighing 1250 g or more with respiratory distress syndrome. N Engl J Med 1991;325(24):1696-703.
76. Looareesuwam S, Viravan C, Vanijanonta S, et al. Randomised trial of artesunate and mefloquine alone and in sequence for acute uncomplicated falciparum malaria. Lancet 1992;339:821-24.
77. Lopez-Zeno JA, Peaceman AM, Adashek JA, Socol ML. A controlled trial of a program for the active management of labor. N Engl J Med 1992;326(7):450-4.
78. Maki DG, Ringer M, Alvarado CJ. Prospective randomised trial of povidone-iodine, alcohol, and chlorhexidine for prevention of infection associated with central venous and arterial catheters. Lancet 1991;338:339-43.
79. Mandelli F, Avvisati G, Amadori S, Boccadoro M, Gernone A, Lauta VM. Maintenance treatment with recombinant interferon alfa-2b in patients with multiple myeloma responding to conventional induction chemotherapy. N Engl J Med 1990;322(20):1430-4.
80. Martinez S, Marr JJ. Allopurinol in the treatment of American cutaneous leishmaniasis. N Engl J Med 1992;326(11):741-4.
81. Mayberg MR, Wilson SE, Yatsu F, et al. Carotid endarterectomy and prevention of cerebral ischemia in symptomatic carotid stenosis. JAMA 1991;266(23):3289-3294.
82. McParland P, Pearce JM, Chamberlain GVP. Doppler ultrasound and aspirin in recognition and prevention of pregnancy-induced hypertension. Lancet 1990;335:1552-55.

83. Merlan M, Vitale V, Rosso R, et al. Treatment of advanced squamous-cell carcinoma of the head and neck with alternating chemotherapy and radiotherapy. N Engl J Med 1992;327(16):1115-21.
84. Mishra M, Biswas UK, Jha DN, Khan AB. Amphotericin versus pentamidine in antimony-unresponsive kala-azar. Lancet 1992;340:1256-57.
85. Moertel CG, Lefkopoulo M, Lipsitz S, Hahn RG, Klaassen D. Streptozocin-doxorubicin, streptozocin-fluorouracil, or chlorozotocin in the treatment of advanced islet-cell carcinoma. N Engl J Med 1992;326(8):519-23.
86. Mofenson LM, Moye J, Jr., Bethel J, et al. Prophylactic intravenous immunoglobulin in HIV-infected children with CD4+ counts of 0.20x10^9/L or more. JAMA 1992;268(4):483-488.
87. MRC Vitamin Study Research Group. Prevention of neural tube defects: rsults of the medical research council vitamin study. Lancet 1991;338:131-37.
88. MRC Working Party on the Evaluation of Chorion Villus Sampling. Medical Research Council Eurpean Trial of chorion villus sampling. Lancet 1991;337:1491-99.
89. Niruthisard S, Roddy RE, Chutivongse S. Use of nonoxynol-9 and reduction in rate of gonococcal and chlamydial cervical infections. Lancet 1992;339:1371-75.
90. Nolan T, Debelle G, Oberklaid F, Coffey C. Randomised trial of laxatives in treatment of childhood encopresis. Lancet 1991;338:523-27.
91. North American Symptomatic Carotid Endarterectomy Trial Collaborators. Beneficial effect of carotid endarterectomy in symptomatic patients with high-grade carotid stenosis. N Engl J Med 1991;325(7):445-53.
92. Nyman I, Larsson H, Wallentin L, and the Research Group on Instability in Coronary Artery Disease in Southeast Sweden. Prevention of serious cardiac events by low-dose aspirin in patients with silent myocardial ishaemia. Lancet 1992;340:497-501.
93. Packer M, Carver JR, Rodeheffer RJ, et al. Effect of oral milrinone on mortality in severe chronic heart failure. N Engl J Med 1991;325(21):1468-75.
94. Pathy MSJ, Bayer A, Harding K, Dibble A. Randomised trial of case finding and surveillance of elderly people at home. Lancet 1992;340:890-93.
95. Pedersen C, Sandstrom E, Petersen CS, et al. The efficacy of inosine pranobex in preventing the acquired immunodeficiency syndrome in patients with human immunodeficiency virus infection. N Engl J Med 1990;322(25):1757-63.
96. Perrillo RB, Schiff ER, Davis GL, et al. A randomized, controlled trial of interferon alfa-2b alone and after prednisone withdrawal for the treatment of chronic hepatitis B. N Engl J Med 1990;323(5):295-301.
97. Pfeffer MA, Barunwald E, Moye LA, et al. Effect of captopril on mortality and morbidity in patients with left ventricular dysfunction after myocardial infarction. N Engl J Med 1992;327(10):669-77.
98. Platt R, Zaleznik DF, Hopkins CC, et al. Perioperative antibiotic prophylaxis for herniorrhaphy and breast surgery. N Engl J Med 1990;322(3):153-60.
99. Ploeg RJ, van Bockel JH, Langendijk PTH, et al. Effect of preservation solution on results of cadaveric kidney transplantation. Lancet 1992;340:129-37.
100. Poupon RE, Balkau B, Eschwege E, Poupon R, and the UDCA-PBC Study Group. A multicenter, controlled trial of ursodiol for the treatment of primary biliary cirrhosis. N Engl J Med 1991;324(22):1548-54.

101. Rahmathullah L, Underwood BA, Thulasiraj RD, et al. Reduced mortality among children in southern India receiving a small weekly dose of vitamin A. N Engl J Med 1990;323(14):929-35.
102. Research Committee of the British Thoracic Society. Optimum duration of anticoagulation for deep-vein thrombosis and pulmonary embolism. Lancet 1992;340:873-76.
103. Riggs BL, Hodgson SF, Fallon WMO, et al. Effect of fluoride treatment on the fracture rate in postmenopausal women with osteoporosis. N Engl J Med 1990;322(12):802-9.
104. Rock GA, Shumak KH, Buskard NA, et al. Comparison of plasma exchange with plasma infusion in hte treatment of thrombotic thrombocytopenic purpura. N Engl J Med 1991;325(6):393-7.
105. Roine RO, Kaste M, Kinnunen A, Nikki P, Sarna S, Kajaste S. Nimodipine after resuscitation from out-of-hospital ventricular fibrillation:a placebo-controlled, double-blind, randomized trial. JAMA 1990;264(24):3171-3177.
106. Saari-Kemppainen A, Karjalainen O, Ylostalo P, Heinonen OP. Ultrasound screening and perinatal mortality: controlled trial of systematic one-stage screening in pregnancy. Lancet 1990;336:387-91.
107. Sack JB, Kesselbrenner MB, Bregman D. Survival from in-hospital cardiac arrest with interposed abdominal counterpulsation during cardiopulmonary resuscitation. JAMA 1992;267(3):379-385.
108. Schaad UB, Suter S, Gianella-Borradori A, et al. A comparison of ceftriaxone and cefuroxime for the treament of bacterial meningitis in children. N Engl J Med 1990;322(3):141-7.
109. Schneider MME, Hoepelman AIM, Schattenkerk JKME, et al. A controlled trial of aerosolized pentamidine or trimethoprim-sulfamethoxazole as primary prophylaxis against pneumocystis carinii pneumonia in patients with human immunodeficiency virus infection. N Engl J Med 1992;327(26):1836-41.
110. Shapiro ET, Gerber MA, Holabird NB, et al. A controlled trial of antimicrobial prophylaxis for lyme disease after deer-tick bites. N Engl J Med 1992;327(25):1769-73.
111. SHEP Cooperative Research Group. Prevention of stroke by antihypertensive drug treatment in older persons with isolated systolic hypertension: final results of the systolic hypertension in the elderly program (SHEP). JAMA 1991;265(24):3255-3264.
112. Smidt-Jensen S, Permin M, Philip J, et al. Randomised comparison of amniocentesis and transabdominal and transcervical chorionic villus sampling. Lancet 1992;340:1237-44.
113. Smith P, Arnesen H, Holme I. The effect of warfarin on mortality and reinfarction after myocardial infarction. N Engl J Med 1990;323(3):147-52.
114. Spanish Ribavirin Trial Group. Comparison of ribavirin and placebo in CDC group III human immunodeficiency virus infection. Lancet 1991;338:6-9.
115. Stacpoole PW, Wright EC, Baumgartner TG, et al. A controlled clinical trial of dichloroacetate for treatment of lactic acidosis in adults. N Engl J Med 1992;327(22):1564-9.

116. Steinbeck G, Andresen D, Bach P, et al. A comparison of electrophysiologically guided antiarrhythmic drug therapy with beta-blocker therapy in patients with symptomatic, sustained ventricular tachyarrhythmias. N Engl J Med 1992;327(14):987-92.
117. Steinberg EP, Moore RD, Powe NR, et al. Safety and cost effectiveness of high-osmolality as compared with low-osmolality contrast material in patients undergoing cardiac angiography. N Engl J Med 1992;326(7):425-30.
118. Stiegmann GV, Goff JS, Michaletz-Onody PA, et al. Endoscopic sclerotherpay as compared with endoscopic ligation for bleeding esophageal varices. N Engl J Med 1992;326(23):1527-32.
119. Stiell IG, Hebert PC, Weitzman BN, et al. High-dose epinephrine in adult cardiac arrest. N Engl J Med 1992;327(15):1045-50.
120. Storck M, Hartl WH, Zimmerer E, Inthorn D. Comparison of pump-driven and spontaneous haemofiltration in postoperative acute renal failure. Lancet 1991;337:452-55.
121. Stroke prevention in atrial fibrillation study group investigators. Preliminary report of the stroke prevention in atrial fibrillation study. N Engl J Med 1990;322(12):863-8.
122. Studies of Ocular Complications of AIDS Research Group in Collaboration with the Aids Clinical Trials Group. Mortality in patients with the acquired immunodeficiency syndrome treated with either foscarnet or ganciclovir for cytomegalovirus retinitis. N Engl J Med 1992;326(4):213-20.
123. Sullivan KM, Kopecky KJ, Jocom J, et al. Immunomodulatory and antimicrobial efficacy of intravenous immunoglobulin in bone marrow transplantation. N Engl J Med 1990;323(11):705-12.
124. Sutherland G, Stapleton JA, Russell MAH, et al. Randomised controlled trial of nasal nicotine spray in smoking cessation. Lancet 1992;340:324-29.
125. Swedberg K, Held P, Kjekshus J, Rasmussen K, Ryden L, Wedel H. Effects of the early administration of enalapril on mortality in patients with acute myocardial infarction. N Engl J Med 1992;327(10):678-84.
126. Temkin NR, Dimken SS, Wilensky AJ, Keihm J, Chabal S, Winn HR. A randomized, double-blind study of phenytoin for the prevention of post-traumatic seizures. N Engl J Med 1990;323(8):497-502.
127. Tetteroo GWM, Wagenvoort JHT, Castelein A, Tilanus HW, Ince C, Bruining HA. Selective decontamination to reduce gram-negative colonisation and infections after oesophageal resection. Lancet 1990;335:704-07.
128. The Boston Area Anticoagulation Trial for Atrial Fibrillation Investigators. The effect of low-dose warfarin on the risk of stroke in patients with nonrheumatic atrial fibrillation. N Engl J Med 1990;323(22):1505-11.
129. The Canadian Preterm Labor Investigators Group. Treatment of preterm labor with the beta-adrenergic agonist ritodrine. N Engl J Med 1992;327(5):308-12.
130. The Candaian Cooperative Multiple Sclerosis Study Group. The Canadian cooperative trial of cyclophosphamide and plasma exchange in progressive multiple sclerosis. Lancet 1991;337:441-46.
131. The Cardiac Arrhythmia Suppression Trial II Investigators. Effect of the antiarrhythmic agent moricizine of survival after myocardial infarction. N Engl J Med 1992;327(4):227-33.

132. The International Chronic Granulomatous Disease Cooperative Study Group. A controlled trial of interferon gamma to prevent infection in chronic granulomatous disease. N Engl J Med 1991;324(8):509-16.
133. The International Study Group. In-hospital mortality and clinical course of 20 891 patients with suspected acute myocardial infarction randomised between alteplase and streptokinase with or without heparin. Lancet 1990;336:71-75.
134. The Intravenous Immunoglobulin Collaborative Study Group. Prophylactic intravenous administration of standard immune globulin as compared with core-lipopolysaccharide immune globulin in patients at high risk of postsurgical infection. N Engl J Med 1992;327(4):234-40.
135. The National Institute of Child Health and Human Development Intravenous Immunoglubulin Study Group. Intravenous immune globulin for the prevention of bacterial infections in children with symptomatic human immunodeficiency virus infection. N Engl J Med 1991;325(2):73-80.
136. The OSIRIS Collaborative Group. Early versus delayed neonatal administration of a synthetic surfactant - the judgment of OSIRIS. Lancet 1992;340:1363-69.
137. The SALT Collaborative Group. Swedish aspirin low-dose trial (SALT) of 75 mg aspirin as secondary prophylaxis after cerebrovascular ischemic events. Lancet 1991;338:1345-49.
138. The SOLVD Investigators. Effect of enalapril on survival in patients with reduced left ventricular ejection fractions and congestive heart failure. N Engl J Med 1991;325(5):293-302.
139. The SOLVD Investigators. Effect of enalapril on mortality and the development of heart failure in asymptomatic patients with reduced left ventricular ejection fractions. N Engl J Med 1992;327(10):685-91.
140. The Veterans Affairs Cooperative Variceal Sclerotherapy Group. Prophylactic sclerotherapy for esophageal varices in men with alcoholic liver disease. N Engl J Med 1991;324(25):1779-84.
141. The Veterans Affairs Total Parenteral Nutrition Cooperative Study Group. Perioperative total parenteral nutrition in surgical patients. N Engl J Med 1991;325(8):525-32.
142. Tilyard MW, Spears GFS, Thomson J, Dovey S. Treatment of postmenopausal osteoporosis with calcitrol or calcium. N Engl J Med 1992;326(6):357-62.
143. Tonnesen P, Norregaard J, Simonsen K, Sawe U. A double-blind trial of 16-hour transdermal nicotine patch in smoking cessation. N Engl J Med 1991;325(5):311-5.
144. Transdermal nicotine study group. Transdermal nicotine for smoking cessation. JAMA 1991;266(22):3133-3138.
145. Trust Study Group. Randomised, double-blind, placebo-controlled trial of nimodipine in acute stroke. Lancet 1990;336:1205-09.
146. UK Gabapentin Study Group. Gabapentin in partial epilepsy. Lancey 1990;335:1114-17.
147. van der Wall E, Verkooyen PP, Mintjes-de Groot J, et al. Prophylactic ciprofloxacin for catheter-associated urinary-tract infection. Lancet 1992;339:946-51.

148. Vijayaraghavan K, Radhaiah G, Prakasam BS, Sarma KVR, Redi V. Effect of massive dose vitamin A on morbidity and mortality in Indian children. Lancet 1990;336:1342-45.
149. Villar J, Farnot U, Barros F, et al. A randomized trial of psychosocial support during high-risk pregnancies. N Engl J Med 1992;327(18):1266-71.
150. Volberding PA, Lagakos SW, Koch MA, et al. Zidovudine in asymptomatic human immunodeficiency virus infection. N Engl J Med 1990;322(14):941-9.
151. Walt RP, Cottrell J, Mann SG, Freemantle NP, Langman MJS. Continuous intravenous famotidine for haemorrhage from peptic ulcer. Lancet 1992;340:1058-62.
152. West KP, Jr., Pokhrel RP, Katz J, et al. Efficacy of vitamin A in reducing preschool child mortality in Nepal. Lancet 1991;338:67-71.
153. Westgate J, Harris M, Curnow JSH, Greene KR. Randomised trial of cardiotocography alone or with ST waveform analysis for intrapartum monitoring. Lancet 1992;340:194-98.
154. Whitley R, Arvin A, Prober C, et al. A controlled trial comparing vidarabine with acyclovir in neonatal herpes simplex virus infection. N Engl J Med 1991;324(7):444-9.
155. Wilcox RG, von der Lippe G, Olsson CG, Jense G, Skene AM, Hampton JR. Effects of alteplase in acute myocardial infarction: 6-month results from the ASSET study. Lancet 1990;335:1175-78.
156. Woods KL, Fletcher S, Roffe C, Haider Y. Intravenous magnesium sulphate in suspected acute myocardial infarction: results of the second Leicester Intravenous Magnesium Intervention Trial (LIMIT-2). Lancet 1992;339:1553-58.
157. Young RC, Walton LA, Ellenberg SS, et al. Adjuvant therapy in stage I and stage II epithelial ovarian cancer. N Engl J Med 1990;322(15):1021-7.
158. Ziegler EJ, Fischer CJ, Jr., Sprung CL, et al. Treatment of gram-negative bacteremia and septic shock with HA-1A human monoclonal antibody against endotoxin. N Engl J Med 1991;324(7):429-36.

HARVEY CUSHING / JOHN HAY WHITNEY
MEDICAL LIBRARY

MANUSCRIPT THESES

Unpublished theses submitted for the Master's and Doctor's degrees and deposited in the Medical Library are to be used only with due regard to the rights of the authors. Bibliographical references may be noted, but passages must not be copied without permission of the authors, and without proper credit being given in subsequent written or published work.

This thesis by has been used by the following persons, whose signatures attest their acceptance of the above restrictions.

NAME AND ADDRESS DATE