Reanalysis of The National Institute of Mental Health Treatment of Depression Collaborative Research Program General Effectiveness Report

Donald F. Klein, M.D. and Donald C. Ross, Ph.D.

The National Institute of Mental Health (NIMH) Treatment of Depression Collaborative Research Program General Effectiveness Report statistical analyses are criticized. Their analyses, which fostered the belief that the active treatments were indistinguishable, were compromised by an inappropriately stringent level of significance with regard to both heterogeneity of slope and pairwise group differences. Once slope heterogeneity is detected, the Johnson-Neyman technique is more appropriate than arbitrary sample subdivision. All of these tactics lowered power substantially.

Our reanalysis indicates a reasonable ordering for the treatments with medication superior to the psychotherapies and the psychotherapies somewhat superior to placebo. These effects are particularly marked among the more symptomatic and impaired patients. The lack of dosage by severity analyses renders the severity findings ambiguous.

Scientific and public health implications are discussed. ***[Neuropsychopharmacology 8:241–251, 1993]***

KEY WORDS: *Depression; Clinical trials; Statistics; Analysis of variance; Psychotherapy; Drug therapy*

INTRODUCTION

The National Institute of Mental Health Treatment of Depression Collaborative Research Program was a 16-week, multisite, randomized study that compared two psychotherapies, cognitive behavior therapy (CBT) and interpersonal psychotherapy (IPT), with placebo case management (PCM) and imipramine case management (ICM) in the treatment of nonbipolar, nonpsychotic, depressed outpatients (Elkin et al. 1985, 1989).

Four outcome measures were reported: Hamilton Depression Rating Scale (HDRS), the Global Assessment Scale (GAS), the Beck Depression Inventory (BDI), and the Hopkins Symptom Check List-Total Score (HSCL-90). The data were analyzed by univariate 3 × 4 (sites × treatments) analyses of covariance (ANCOVA) for the scaled data and by chi square analyses for treatment × recovery status data. The ANCOVAs were followed by Bonferroni adjusted pairwise *t* tests and the chi square tests by Brunden adjusted pairwise comparisons. Tests for initial severity × treatment interactions were also made. Three sets of analyses on overlapping samples were performed: (1) completer analysis on only those patients who completed the course of treatment; (2) endpoint analysis on patients who completed 3.5 weeks of treatment; and (3) endpoint analysis on all patients who entered the study. Secondary analyses were conducted within less symptomatic and less impaired subgroups as well as more symptomatic and more impaired subgroups.

From the Department of Psychiatry, Columbia University; and the New York State Psychiatric Institute, New York, New York.

Address reprint requests to: Donald F. Klein, M.D., 722 West 168th Street, New York, New York 10032.

Received December 11, 1991; revised April 30, 1992 and August 3, 1992; accepted August 10, 1992.

© 1993 American College of Neuropsychopharmacology

Published by Elsevier Science Publishing Co., Inc.
655 Avenue of the Americas, New York, NY 10010

0893-133X/93/$6.00

The standard reference treatment was ICM. The conclusions were startling. "Thus, there is no evidence in the major analyses that either of the psychotherapies was inferior to the standard reference treatment at termination of treatment on measures of depressive symptoms or general functioning. These statistical analyses do not, of course, permit the inference that the psychotherapies and the standard reference treatment were "equal" in effectiveness. However, since we had satisfactory power in these analyses for detecting large effect size differences between pairs of treatments (in the total unstratified sample), it is unlikely that very large or important differences were missed." (Elkin et al. 1989). The initial paragraph of their conclusions reaffirms this.

Despite the authors' disclaimer that they were not asserting that the treatments were "equal," they felt that nothing important had been missed. Therefore, as a practical matter one could as well choose one treatment as the other. Their statement fostered the interpretation that medication and psychotherapy were equivalent for depression, as evidenced by a front page article in the *New York Times* of May 14, 1986, entitled "Psychotherapy is as Good as Drug in Curing Depression, Study Finds." As recently as November 13, 1989, *The Wall Street Journal* cited this study as follows, "Generally, the researchers cautiously conclude, there is no evidence that the psychotherapies are less effective than the antidepressant drug, but there was evidence the psychotherapies were better than the placebo treatment."

Elkin et al. (1989) reported only two significant treatment findings; imipramine is superior to placebo on two different measures in two overlapping patient groups. Three "trends" were reported. It was reported twice that ICM was superior to PCM and, once that IPT was superior to PCM. To sum up, in the undivided samples there was some evidence for ICM efficacy, minimal evidence for IPT efficacy, no evidence for CBT efficacy, but no evidence of treatment differences. One significant site-by-treatment interaction was discounted since "there was no treatment effect on this variable." (Elkin et al. 1989).

These surprising, meager results led to controversy. Klein (1990) objected to four aspects of the analysis. These were: (1) an inappropriately stringent level of significance used in the necessary preliminary analysis for heterogeneity of slope in the ANCOVA; (2) a lack of attention to initial severity once slope heterogeneity was detected; (3) the arbitrary, unnecessary, and stringent significance level adjustments in the multiple treatment comparisons; and (4) the arbitrary subdivision of the sample to explore severity effects, rather than using analyses that take advantage of slope heterogeneity and preserve the entire sample. Each of these tactics lowered power substantially, making it easier to miss real differences. We argue that real, clinically important, findings were obscured by a veil of "statistical insignificance" allowing the erroneous imputation of no important treatment differences.

Reanalysis Outline

We amplify their ANCOVA attending to the crucial issue of heterogeneity of slope, thereby showing that many between treatment contrasts were obscured. We present statistically meaningful treatment comparisons, unreported by Elkin et al., because of their unnecessary Bonferroni limitation, and both criticize and replace their "secondary analyses." We discuss multiple contrast control, attrition, dosage, comparison of IPT and CBT, and the procedural and scientific implications of this reanalysis.

Analytic Strategy of Elkin et al.

Elkin et al. conducted ANCOVAs on four measures; the HDRS, the GAS, the BDI, and the HSCL-90.

"Marital status, which was significantly related to outcome, was not distributed evenly across treatment groups . . . it was always used as a covariate in the ANCOVAs. Pretreatment scores on the dependent variable were also included as a covariate, except in those few instances on the HDRS and GAS in which there was significant ($p < .05$) heterogeneity of regression and the use of a pooled regression for the ANCOVA was not justified. In these instances, the ANCOVAs reported used only marital status, and not pretreatment score, as a covariate." The overall sample was divided into three overlapping subsamples: the completer sample ($n = 155$); the endpoint 204 sample, which included all patients who received 3.5 weeks of treatment; and the endpoint 239 sample, which consisted of all patients who entered treatment. The ANCOVA used a last observation carried forward technique within each subsample.

"To protect against inflation of the type I error rate associated with multiple comparisons, probability levels for comparisons between pairs of treatments were adjusted for the total number of comparisons, by means of the Bonferroni t test . . . an overall probability level of $p < .10$ was accepted for the F test between the 4 groups. However, for the crucial initial heterogeneity of slope analysis, the more stringent p value of .05 or less was required, this means that the probability level obtained must actually be $< .017$ to be considered significant at an adjusted alpha level $< .10$ . . . This approach left us with satisfactory statistical power (.81 to .95) to detect medium size effects in our primary ANCOVAs" (Elkin et al. 1989).

Reanalysis

The NIMH recently released a public access tape containing only the data that underlies published analyses. Despite statements that the published data were adequate to conduct alternative analyses (Hirschfeld 1990), only with this release could alternative ANCOVAs be done.

For simplicity, all *p* values presented are two-tailed, but the treatments versus placebo *p* values can be halved to yield one-tailed tests, since there are clear directional hypotheses of treatment superiority to placebo. Between treatment contrasts, not involving placebo, should be thought of as two-tailed tests.

Analysis of covariance assumes the same amount of benefit over the entire range of initial scores. Therefore, the slopes of final versus initial scores within each treatment group are parallel to each other. The parallel slopes are replaced by a common pooled slope that should fairly represent all the treatment slopes.

However, on placebo, those who start out badly often end up badly, whereas those who start out better end up better. In contrast, with a wonderfully active treatment, everyone would end up well regardless of the initial degree of illness. In the placebo case, the slope would be high, whereas for effective treatments, the slope would approximate 0.

Whether the amount of treatment benefit correlates with initial severity is made obvious by the first step of ANCOVA carried out to detect nonparallel slopes. This must be done prior to further analyses because the parallelism assumption is crucial. If the slopes are not parallel, there is no constant difference between groups, but rather varying differences, depending on initial severity.

We must estimate how far the slope differences exceed expectable sampling variation. If a stringent significance level is selected, then even widely different slopes might not invalidate the null hypothesis of homogeneous regressions and be considered parallel. Elkin et al. selected an α level of .05 to test all four slopes at once.

Cohen and Cohen (1983) say, "To reach a positive conclusion that the regression is homogeneous between groups, however, requires the logically impossible feat of proving the null hypothesis. We must therefore settle for results *consistent with* this null hypothesis, that is, we posit homogeneity in the absence of evidence to the contrary. A non-significant *F* ratio, particularly one well below the value at the conventional α = .05 criterion, ideally one that is close to the chance-expected value of one, constitutes such evidence."

Hays (1988) states, "It is probably a good idea to proceed with the ANCOVA when the *F* test for homogeneity of regression *fails* to reach significance even at the .25 level. However, if the homogeneity of regression test reached significance even for α of .10 or less, there may be good reason to doubt the validity of the *F* test in ANCOVA." Also, measurement error in the covariate (initial severity) lessens our ability to detect nonparallel slopes. Therefore, there is even more nonparallelism than our tests show (Huitema 1980).

Some strategy other than doing an omnibus *F* test for slope differences and then either proceeding with an ANCOVA if the null hypothesis is not rejected at the .05 level, or proceeding with a simple ANOVA if it is rejected is called for. We need a method sensitive to possible slope differences so that inappropriate ANCOVAs are not performed, and one that will use the information contained in slope differences if they are found.

We used the following strategy. First, an overall test for slope differences was performed. Alpha was set at .10. This is not as conservative as the authors cited suggest but is better than testing at the .05 level. The initial overall tests used the same model as Elkin et al.; the four treatment groups and the post- and pretreatment scores of the variable being tested were included. The slope estimates from these analyses are the ones reported in Table 1. If the null hypothesis was rejected, tests for pairwise slope differences were performed at

Table 1. Slope Descriptions Relevant to Detected Heterogeneity

Sample	Scale	Treatment	Slope	Slope SE
P239	GAS	CBT	.34	.22
		IPT	.47	.24
		ICM	−.21	.25
		PCM	.62	.21
	HDRS	CBT[a]	.68	.24
		IPT	.61	.21
		ICM	.17	.21
		PCM	1.06	.20
P204	GAS	CBT	.32	.24
		IPT	.40	.24
		ICM	−.40	.24
		PCM	.61	.23
	HDRS	CBT[a]	.55	.26
		IPT[a]	.46	.22
		ICM	.02	.20
		PCM	.85	.24
	BDI	CBT	.61	.20
		IPT	.01	.17
		ICM	.15	.13
		PCM[a]	.22	.21
Completer	HDRS	CBT	.55	.31
		IPT[a]	.16	.20
		ICM	−.07	.18
		PCM	.61	.23
	BDI	CBT	.58	.22
		IPT	.01	.12
		ICM	.17	.13
		PCM[a]	.27	.22

[a] Not involved in any statistically significant pairwise slope contrast.

the .05 level. Pairwise differences were tested in the context of all four treatment groups. Four individual regressions were fit, and then three regressions were fit with the two treatments hypothesized to have the same slope fit with the same regression. If the four-slope model fit significantly better than the three-slope model, the null hypothesis was rejected and the two groups in question were deemed to have different slopes. Treatment, site, site × treatment, and two marital status effects (never married versus all else, and in a stable relationship versus all else) were included in the model. Only slight changes were noted when these extra variables were omitted. In this data set, the same decisions would have been made to pursue pairwise contrasts if instead of $\alpha = .10$, we used the rule that the omnibus F had to be greater than one (see Table 2). That could be an alternative rule.

There is no way to simultaneously increase the power to detect real differences and to reduce the chances of falsely claiming to detect nonexisting differences, other than increasing the sample size. Given a fixed sample size, one can only try to sensibly balance these aims. We believe that Elkin et al. used a strategy that made it unduly difficult to detect slope differences, resulting in the ANCOVA being used inappropriately and in losing valuable information about the differential effect of treatments for different levels of initial pathology. The method used here allows us to detect treatment effects previously missed without incurring an unreasonable probability of type I error.

Reanalysis of Slope Heterogeneity

In only two analyses by Elkin et al., the HDRS in the endpoint 239 group and the GAS in the endpoint 204 group, did the four treatment slopes prove nonparallel at their preselected α value of .05. In 10 of these 12 tests, Elkin et al. did not report any indication of the degree of significance of the overall slope contrast.

Using the public access data tape, it became possible to reanalyze for slope heterogeneity. Table 1 contains the estimated slope and standard error of the slope for each contrast that showed a slope difference. The analyses are in Table 2. In 7 of 12 tests, both $F > 2.1$, and $p < .10$, thus invalidating the parallelism assumption.

Next, we tested pairwise slope contrasts within each overlapping group, given that the test for heterogeneity showed $p < .10$. It is unsatisfactory simply to state that these treatments differ in a fashion related to initial severity. One would like to be able to say, using the entire undivided sample, that above or below a certain point one treatment is superior to the other. Techniques for this have been proposed by Johnson and Neyman (1936) and Potthoff (1964). The Johnson-Neyman technique divides the independent variable axis into two regions; a region within which the α percent confidence limits for the mean difference between treatments do not include 0, which we will call the non-0 region, and a region within which they do, which we will call the region of possible intersection. Potthoff modified the Johnson-Neyman method so that one is α percent confident that the entire non-0 region, rather than each point within the region does not contain a 0 difference. The Johnson-Neyman non-0 regions are larger than the corresponding Potthoff regions. Johnson-Neyman regions are reported in this paper. Use of the Potthoff regions would not have substantially changed interpretations.

In solving for the endpoints of the Johnson-Neyman α percent non-0 region, an imaginary or complex answer may result. If α is reduced, a real answer may be obtained. Wherever possible, 95% regions are reported here. In some cases, it was necessary to reduce α to 90%.

Table 2. One-Way Four-Group ANCOVA with One Covariate *F* Tests for Slope Heterogeneity

	Reanalysis		NIMH
Endpoint 239 Analysis			
GAS	$F3,231 = 2.18$	$p = .091$	
HDRS	$F3,231 = 3.05$	$p = .029$	$p < .05$
BDI	$F3,231 = .64$	$p = .588$	
HSCL-90	$F3,230 = .65$	$p = .583$	
Endpoint 204 Analysis			
GAS	$F3,196 = 2.85$	$p = .039$	$p < .05$
HDRS	$F3,196 = 2.47$	$p = .063$	
BDI	$F3,196 = 2.26$	$p = .083$	
HSCL-90	$F3,195 = .43$	$p = .731$	
Completers Analysis			
GAS	$F3,147 = .96$	$p = .415$	
HDRS	$F3,147 = 2.15$	$p = .096$	
BDI	$F3,147 = 2.18$	$p = .093$	
HSCL-90	$F3,147 = .94$	$p = .425$	

Table 3. Significant Pairwise Slope Contrasts

Group	*p* for Pairwise Slope Contrast GAS	HDRS	BDI	HSCL-90
Endpoint 239				
ICM vs. PCM	.006	.011	X[a]	X[a]
IPT vs. PCM	NS	.102	X	X
CBT vs. PCM	NS	NS	X	X
ICM vs. IPT	.018	NS	X	X
ICM vs. CBT	.067	NS	X	X
CBT vs. IPT	NS	NS	X	X
Endpoint 204				
ICM vs. PCM	.001	.013	NS	X[a]
IPT vs. PCM	NS	NS	NS	X
CBT vs. PCM	NS	NS	NS	X
ICM vs. IPT	.008	NS	NS	X
ICM vs. CBT	.010	NS	.024	X
CBT vs. IPT	NS	NS	.007	X
Completers 155				
ICM vs. PCM (.08)[b]	X[a]	.050	NS	X
IPT vs. PCM	X	NS	NS	X
CBT vs. PCM	X	NS	NS	X
ICM vs. IPT	X	NS	NS	X
ICM vs. CBT	X	.102	.056	X
CBT vs. IPT	X	NS	.004	X

[a] Overall four-slope p value > .1.
[b] Pairwise contrast significant although overall $F < 1$.

The non-0 region may be of two types. It may consist of two subregions at each extreme of the axis, with the region of possible intersection between them, or it may consist of a single region bound at both ends by a subregion of possible intersection. In either case, the observed sample intersection will lie within the region of possible intersection.

Table 3 shows an unsuspected richness of meaningful contrasts between treatment groups. As we predicted, ICM versus PCM is distinguished by heterogeneity of slope on both the GAS and the HDRS in all groups, indicating ICM superiority for the more severe.

Previously there had been no indication that ICM was superior to the psychotherapies or that the psychotherapies may differ. The section below on Johnson-Neyman analyses gives details.

Since each contrast is meaningful it can be argued that one does not need a significant overall four-treatment test for slope heterogeneity to proceed with pairwise slope contrasts. However, only one more finding would be noted with this unconventional stand, ICM versus PCM on the GAS in the completer group.

Johnson-Neyman Analyses

What do these slope differences translate into viewed from the Johnson-Neyman perspective? For this data set, confidence limits are such that significant treatment contrasts are discerned only in the range of more severe pathology. Table 4 indicates the 95% and 90% two-tailed bounds, analogous to a conventional .05 finding and a ($p \leqslant .1$) trend. Table 4 contains the estimated intersection point, the boundaries of the Johnson-Neyman regions, the percentage of the sample that falls between the estimated intersection point, and each of the

Table 4. Johnson-Neyman Regions for Pairwise Slope Contrasts

						Proportions in Johnson-Neyman Regions			
			Johnson-Neyman Limits			Statistically Indeterminate		Region With	Confidence
Sample	Test	Comparison	Limits	Intersection	Limits	Regions		Benefits	Level (%)
P239	GAS	ICM vs. PCM	86.6	59.2	53.7	.21	.25	.54	.95
		IPT vs. PCM	66.9	53.1	44.3	.46	.41	.13	.90
		CBT vs. PCM	–	58.0	48.3	.46	.30	.24	.95
	HDRS	ICM vs. PCM	6.9	15.6	18.6	.18	.30	.52	.95
		IPT vs. PCM	–	11.6	17.8	.01	.38	.61	.95
P204	GAS	ICM vs. PCM	71.3	57.8	53.2	.26	.21	.53	.95
		IPT vs. PCM	73.6	54.1	46.4	.46	.32	.22	.95
		CBT vs. PCM	162.4	57.4	50.4	.26	.33	.41	.95
	HDRS	ICM vs. PCM	4.6	15.6	18.8	.20	.30	.50	.95
	BDI	CBT vs. ICM	–	19.6	28.3	.21	.40	.39	.95
		CBT vs. IPT	−2.2	23.2	31.0	.34	.40	.26	.95
Completers	HDRS	ICM vs. PCM	−2.0	16.6	20.3	.33	.32	.35	.95
		CBT vs. ICM	−3.5	17.9	23.1	.44	.38	.18	.90
	BDI	CBT vs. ICM	–	16.9	27.3	.12	.41	.47	.95
		CBT vs. IPT	−18.1	20.2	27.0	.23	.40	.37	.95
			← Less severe		More severe →	← Less severe		More severe →	

limits as well as the percentage that falls beyond the limits in the region with treatment benefit. The other possible regions were either empty or contained less than 1% of the sample and are not shown. The pathologic end of the scale is always presented to the right. In no case is as much as 1% of the sample beyond the confidence limit at the nonpathologic end of the scale nor in the farther out subregion of a two-part region of possible intersection.

Since the confidence limit at the less pathologic pole is always at or beyond the range of initial severity, there is never a real situation where one treatment is superior in the most pathologic range but treatment superiority is reversed in the least pathologic range.

At the conventional .05 level, ICM is superior to PCM on both the HDRS and the GAS, with similar cut points in the 204 and 239 groups. In the completer group, ICM is also superior to PCM on the HDRS at a point slightly higher along the severity dimension. In the 204 group, ICM is superior to both IPT and CBT on the GAS. Representative graphs of the Johnson-Neyman results appear in Figures 1 through 4. The axes are labeled to include only values actually observed. The intersection points and confidence limits, which occurred within the range of initial severity, are indicated.

In addition, CBT is superior to PCM on the GAS

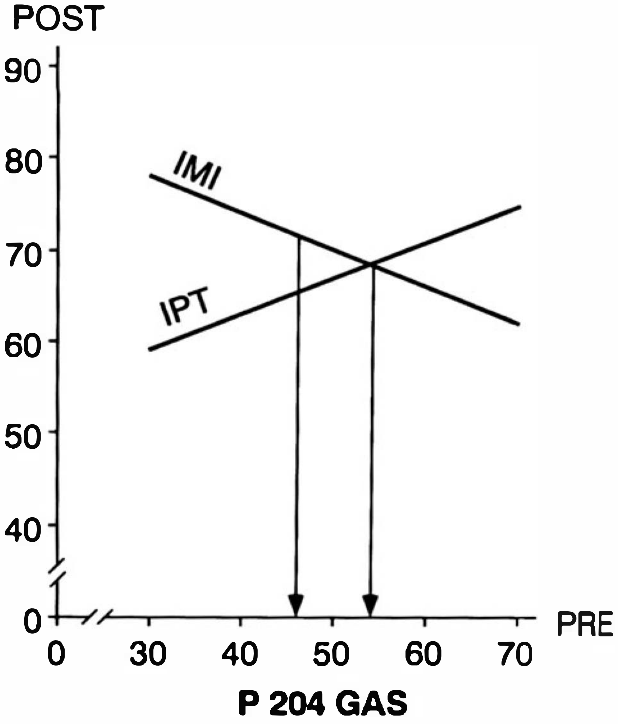

Figure 2. Johnson-Neyman analysis of P204 GAS ICM (IMI) versus IPT with intersection and more pathologic 95% limit.

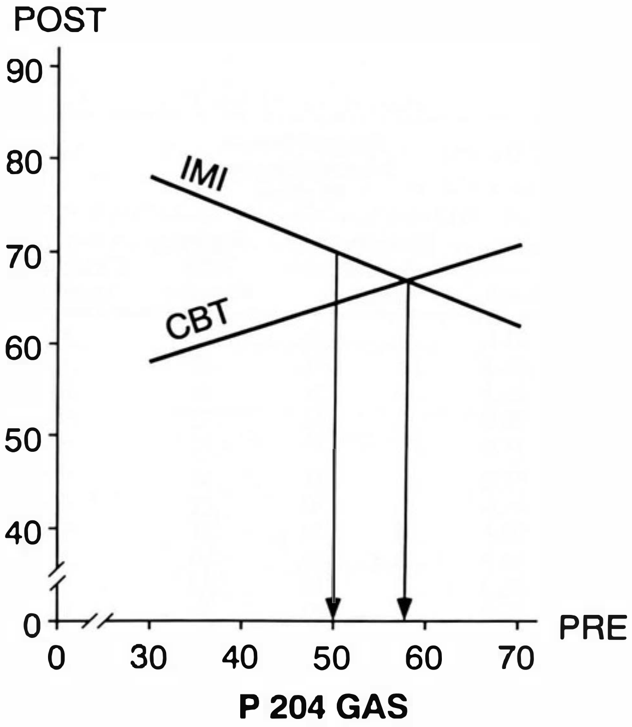

Figure 1. Johnson-Neyman analysis of P204 GAS ICM (IMI) versus CBT with intersection and more pathologic 95% confidence limit.

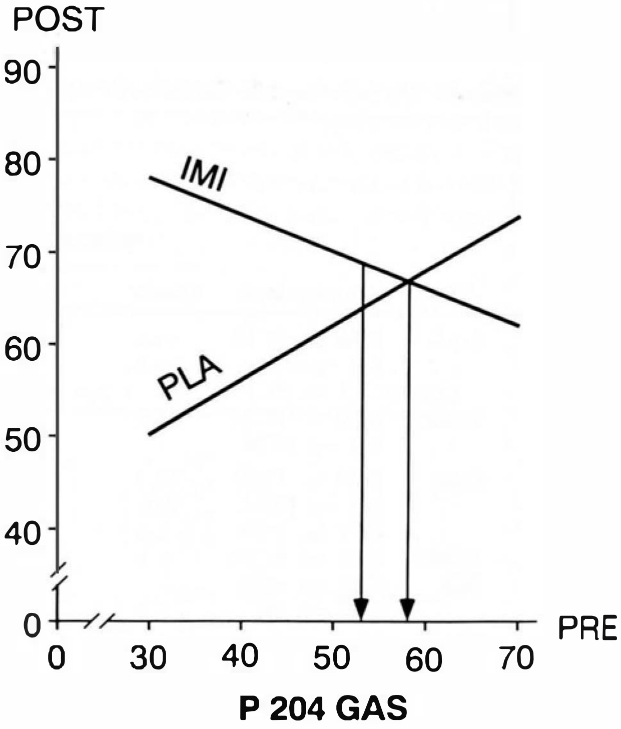

Figure 3. Johnson-Neyman analysis of P204 GAS ICM (IMI) versus PCM with intersection and more pathologic 95% limit.

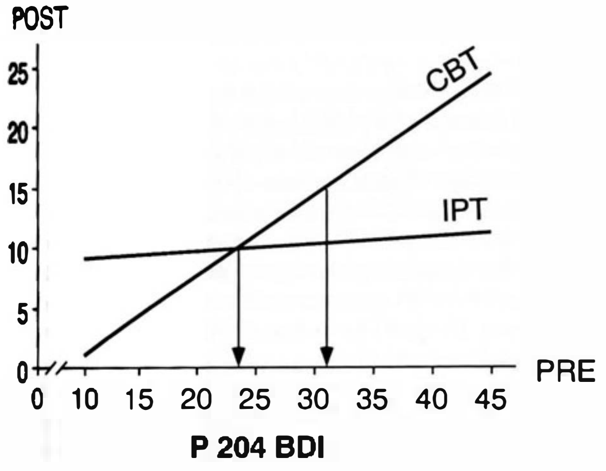

Figure 4. Johnson-Neyman analysis of P204 BDI CBT versus IPT with intersection and more pathologic 95% limit.

for both the 239 and 204 groups; IPT is superior to PCM on the HDRS in the 239 group and the GAS in the 204 group; ICM is superior to CBT on the BDI in the 204 and the completer groups; IPT is superior to CBT on the BDI for the 204 and the completer groups; and ICM is superior to PCM on the HDRS for the completer group; all using 95% limits.

Using 90% confidence limits, there is a trend in the completer group for ICM to be superior to CBT on the HDRS and in the 239 group for IPT to be superior to PCM on the GAS.

Elkin et al. found no significant difference between IPT and CBT in any major analysis or in their "secondary" split group analyses. Furthermore, they stated that the magnitude of the differences between these psychotherapies is small even when IPT was significantly different from PCM but CBT was not. Surprisingly, on the BDI scores among the completers, CBT is significantly inferior to IPT, which is repeated in the 204 group at a slightly greater level of severity. The BDI is the standard instrument for evaluation of CBT effects (see Figure 4).

The two psychotherapies lay on either side of the PCM slope so that their respective differences from PCM were not great. However, for slope, the p value for IPT versus PCM was .25 and for CBT versus PCM was .21. These two p values are at the outer limit of the values suggested by Hays as indicating possible heterogeneity of slope. However, since the psychotherapies lay on opposite sides of the PCM slope, they differed sharply from each other.

If confirmed in subsequent studies, this indicates that CBT is relatively inferior to IPT for patients with BDI scores greater than approximately 30, generally considered the boundary between moderate and severe depression.

If psychotherapy functions by antidemoralization, a psychotherapy that accentuated the positive and ignored eliminating the negative would be most effective. Perhaps for patients who score high on the BDI, the CBT focus upon rigid, dysfunctional attitudes produces a defeat experience. Interpersonal therapy, which focuses elsewhere, may have provided nonspecific support without direct confrontation.

Since efficacy differences between psychotherapies are rare, it is difficult not to be intrigued. However, the minimal differences from PCM make replication necessary. The requirement of a minimal credible comparison condition (e.g., PCM) in psychotherapy evaluation is supported by this trial. Studies that lack this comparison are irretrievably ambiguous (Klein and Rabkin 1984).

This analysis reveals meaningful differences between active treatments that were not detected in total group analyses by Elkin et al. Furthermore, this is not a secondary analysis, but derives directly from the same covariance analyses insensitively analyzed by the Elkin group.

Group Comparisons

For the analyses in which Elkin et al. found nonparallel slopes, only the final scores were contrasted, using two marital status variables as covariates. Since the initial scores were not used at all in these cases, power was again markedly attenuated.

Elkin et al. did report their contrasts in an unadjusted fashion, but given their framework of the Bonferroni correction, they did not report pairwise comparisons at $p < .025$. This ignores contrasts at the .05 level or the conventional .1 trend level. Does this matter?

Table 5 lists all contrasts significant at least at the .10 level that were not mentioned in the original paper, calculated from site-by-treatment ANCOVAs with either three (pretreatment score and two marital status variables, viz. never married versus all else and currently in a stable relationship versus all else) or two (marital status variables only) covariates, exactly consistent with the practice of Elkin et al.

Of these eight trend or better contrasts, six indicate some superiority of the psychotherapies to the PCM condition, one reaffirms the superiority of ICM to PCM and one shows a trend for superiority of ICM to CBT.

Responding to criticism, Elkin et al. performed "alternative" analyses that were still vitiated by low power analysis for heterogeneity of slope. They concluded, "We did not find, as Dr. Klein seems to assume we would, an overall superiority of medication to psychotherapy or both psychotherapies to placebo." (Elkin et al. 1990). However, our analyses, as reported in Tables 4 and 5, find just this.

Table 5. Adjusted Mean Treatment Contrasts Where $.1 > p > .025$

Group	Instrument	p Value
Endpoint 239		
IPT vs PCM	GAS	.04[a]
CBT vs. PCM	HDRS	.08
Endpoint 204		
IPT vs. PCM	HDRS	.10
Completer 155		
ICM vs. PCM	BDI	.03[a]
IPT vs. PCM	BDI	.09
IPT vs. PCM	HSCL-90	.05[a]
ICM vs. CBT	BDI	.06
CBT vs. PCM	HSCL-90	.07

[a] $p < .05$.

"Secondary" Analyses

Elkin et al. performed "secondary" analyses. "These analyses must be considered exploratory, since the design did not include stratification on this variable . . . patients' conditions were considered more severe if 20 or greater on their rescreening HDRS. . . (or) they were considered more severe if 50 or less on the GAS." (Elkin et al. 1989).

It is not correct that these analyses are secondary because the design did not include severity stratification. In fact, these data have already been analyzed for severity by including the initial score as a covariate. What makes their analyses secondary is not the lack of stratification (especially since ANCOVA is usually more powerful than stratification) but rather the post-hoc arbitrary subdivision. Since prior to the study they had not suggested that a split at these points would be fruitful, one cannot logically exclude the possibility of post-hoc data massage. The lack of power is due to stratifying the data into only two groups, which ignores meaningful outcome variance within strata, and the reduced n for estimating means and variances.

By referring to these analyses as secondary, the authors denigrate their importance, thus emphasizing the lack of difference between treatments. Their problem, however, stems directly from the fact that their low power analyses had not detected heterogeneity of slope in the first place. Once slope heterogeneity is detected, the correct Johnson-Neyman approach reveals their "secondary" analyses to be superfluous and misleading.

DISCUSSION

Multiple Contrast Control

Since the Bonferroni correction used in the analysis of Elkin et al. has been widely accepted, we wish to repeat (Klein 1990) our concerns for pedagogical purposes.

> "All six contrasts between the four treatments are of clinical interest and public health importance. We want to know the relative merits of all the treatments. The logic of multigroup data analysis often proves difficult for the non-statistically sophisticated reader, so I present a parable indicating why the Bonferroni correction, selected as the familywise type I error control strategy in this report, interferes with seeing real differences between the treatments. (A type I error results when chance variation is considered real.)
>
> Let us say you are hired to conduct a study comparing IPT versus CBT. You contrast the treatments on 40 variables and find that on 15 variables you have significant differences at the preset .05, two-tailed, type I error rate per comparison. Plainly this is fiction.
>
> You are about to happily write up your discoveries of real substantive differences, when your boss tells you that you are actually part of a larger study. Down in the basement, patients from the same pool were randomized to receive imipramine-CM [ICM]. Therefore, you are actually part of a three-group study. You should therefore Bonferroni adjust your p values to a critical value of .017 (.05/3) to preserve the .05 familywise α rate. Unfortunately, of your 40 comparisons, you only have 3 at the .017 level. But this still exceeds chance, so you write up a more conservative report that still affirms a few real differences between the two psychotherapies.
>
> However, your amnesic boss returns to say that he meant to tell you, up in the attic they were also conducting a PLA-CM [PCM] component of this trial. Since the patients were randomized to four different cells allowing six contrasts, you should only accept significant pairwise contrasts at the .008 (.05/6) level. Well, sad to say, you do not have any pairwise contrasts significant at the .008 level.
>
> Therefore, you tearfully burn your first manuscript, which presented your trailblazing discovery that the two psychotherapies were really different from each other, and now write a less interesting (and probably less valid) report indicating your inability to demonstrate differences. Mind you, this is not because your data have changed, but because there were other patients in the attic and basement. Further, it does not matter what actually happened to these other patients. Does that make sense?
>
> Since in the Elkin et al. study every pairwise comparison is meaningful, only a per-comparison type I error rate is important. The Bonferroni familywise "correction" simply loses power, with no compensating benefit.
>
> One might wonder when you should ever use a familywise error rate. Let us say you are manufacturing a motor that has 20 crucial components. If any of those components fail, the motor will not work. The economics are such that 1 in 20 defective motors is the maximum acceptable number. What should the acceptable failure rate per component be? Obviously it should not be 5%. The chance that any one component will not fail is .95. If the components fail independently, the chance that

all the components will not fail is $.95^{20}$ which is .36. Therefore, the chance that at least 1 component fails is 1-.36 or .64. If you accept a 5% component failure rate, you will end up with a 64% motor failure rate.

How can you get around this? Well, you can Bonferroni correct the .05 acceptable overall motor failure rate by dividing .05 by 20, yielding .0025. That means that the allowable chance that the individual component will not fail is very high, .9975, but $.9975^{20}$ is .951; therefore, only 4.9% of the motors will not work, so you are where you want to be."

Hochberg and Tamhane (1987) discuss research where one "has a finite number of inferences of interest specified prior to the study. If these inferences are unrelated in terms of their content or intended use (although they may be statistically dependent) then they should be treated separately and not jointly. If a decision (or conclusion) is to be based on these inferences and its accuracy depends on some joint measure of erroneous statements in the given set of inferences, then the collection of inferences should be considered jointly as a family."

If one overall decision will be made on the basis of numerous group comparisons (and if any group comparison was falsely positive, the entire decision would be erroneous), then you should have stringent rules for multiple comparisons. But that is not the case when evaluating several treatments since no overall decision is required, but rather a number of individually meaningful pairwise contrasts.

Recent articles (Rothman 1990; Saville 1990) emphasized that multiple comparison "corrections" are unnecessary and that even "*F* protection" results in inconsistent inferences. "Inconsistent" means that exactly the same contrast between two groups will sometimes be considered significant, and at other times insignificant, depending on the outcome of the other irrelevant groups in the trial.

Science progresses by constructive replication, which is particularly important in treatment evaluation. A well-attested treatment, such as ICM, may fail to show statistically significant differences from PCM (Klein and Davis 1969). Such failures are usually attributable to misselection of the appropriate patient population, but at times poor treatment conduct, inadequate dosage, poor power, bad luck, and/or misevaluation are the culprits.

Requiring severe significance levels is only appropriate for definitive experiments that require no replication. In our still-maturing field, plagued by samples of convenience, poor measures, attrition, and difficult to detect biases, this is never the case. Unadjusted *p* values for even slight trends should be presented, allowing the reader the freedom to be either stringent or to follow his nose towards replication.

Good design acknowledges an unpredictably high rate of spontaneous improvement for many anxious and depressed outpatients. Therefore, a standard treatment as the sole comparison is problematic. The major trouble is that you do not know if this particular sample is actually a sample where the standard treatment exerts specific effects. That can only be shown by internal calibration with a comparative placebo group. This was the major methodological advance incorporated in the study by Elkin et al., which sets a standard for all future studies in this area.

Treatment Technique and Dosage

Since the medication course has never been published or anyone given access to it, one cannot be sure that the flexible dosage pharmacotherapy was well done. It is quite possible that milder patients received ineffective, small doses, which could account for the lack of specific medication benefit for the less symptomatic or impaired patients.

Such dose-by-severity analyses should have been in the initial paper by Elkin et al. since they discerned a severity effect, and should have dealt with this possible confound. If the dose–severity relationship hypothesized here is true, all practical clinical implications about severity and medication benefit derived from this trial require revision.

It is also unfortunate that the data concerning the acceptability of the psychotherapeutic interventions are not available.

CONCLUSIONS

The paucity of findings in the original paper is largely due to unnecessarily stringent levels of significance for both slope heterogeneity and group contrasts.

Elkin et al. concluded that the value of imipramine has been shown but the active treatments could not be distinguished. We correct this by indicating superiority of ICM to the psychotherapies.

Our analyses also show some superiority of psychotherapy to PCM but the implications are not clear. Fawcett (1990), who supervised the psychopharmacology and case management approach stated that this was not free of strain and that some of the psychiatrists did not value their role and tended to cut short the sessions. These clinicians were initially hired to become trainers but were shifted into a direct treatment role.

Fawcett's statement raises the question of comparative treatment credibility. In evaluating a psychotherapy, it must be compared with a credible treatment. If measured, these data have not been presented. If the major motor of psychotherapy is relief of demoralization through credibility, as Frank has suggested, this is necessary (Klein and Rabkin 1984).

Implications

A leading scientific value recognizes that researchers should not simply impart conclusions. One wants to know enough about both methods and data that critical alternative analyses are possible.

The summary statistics provided in many papers do not allow for such analyses. Elkin et al. state "Since we do table the means and standard deviations [for the primary analyses], it is possible for a reader who is sufficiently determined, to compute his own *p* values if he is unwilling to accept our criteria." Similarly, Hirschfeld states that "Complete pretreatment and post treatment means, SDs and *N*s are provided for three samples on four outcome measures, as well as other further analyses. Few articles in literature, even those published in this Journal, can match this level of comprehensiveness.

"In light of this amount of data, I encourage Dr. Klein, or anyone else who so wishes, to reinterpret the published data."

However, these statements are incorrect since the regression slopes could not be calculated, despite our determination, until the release of the partial data tapes allowed alternative analyses. If it is difficult for an article to present the data necessary for alternative analyses, the data (or the more complex summary statistics) underlying the presented analyses should, ideally, be available on request. Means and sample sizes for each subgroup as well as the pooled within groups dispersion matrix for all variables in the analysis would most conveniently allow readers to reproduce the investigators' tests for adjusted mean differences. In addition, individual dispersion matrices for each group would be needed to check tests for slope differences and homogeneity of variance between groups and to perform Johnson-Neyman analyses.

There is a problem in intellectual property rights. Those who developed the ideas and did the work have a legitimate interest in receiving credit for the study, analyzing the data, and publishing the results. It is not unusual for there to be several years spent at data analysis and write up of a complex study. To release data prior to the completion of this task may result in others, rather than the original group, receiving the credit for making discoveries.

On the other hand, both the scientific and lay publics have an interest in rapid, thoughtful, data analysis so that public health implications as well as heuristic inferences can be properly considered. This should be a matter of widespread public interest, but has not been openly discussed with the scientific or lay community. Important public health issues often require widespread multisite collaborative efforts. The mechanism for achieving this goal is properly a subject for discussion by those affected by these procedures, including the general public. It would be natural for NIMH to sponsor a series of meetings to attempt to deal with these complex issues.

ACKNOWLEDGMENTS

This research is based on the NIMH Treatment of Depression Collaborative Research Program Public Use Data Tape and Documentation (First Release). The NIMH Treatment of Depression Collaborative Research Program is a multisite program initiated and sponsored by the Psychosocial Treatments Research Branch, Division of Extramural Research Programs (now part of the Mood, Anxiety, and Personality Disorders Research Branch, Division of Clinical Research), NIMH. The program was funded by Cooperative Agreements with six participating sites: George Washington University, MH 33762; University of Pittsburgh, MH 33753; University of Oklahoma, MH 33760; Yale University, MH 33827; Clarke Institute of Psychiatry, MH 38231; and Rush Presbyterian-St. Luke's Medical Center, MH 35017. The principal NIMH collaborators were Irene Elkin, Ph.D., Coordinator (now at the University of Chicago); M. Tracie Shea, Ph.D., Associate Coordinator (now at Brown University); John P. Docherty, M.D. (now at Nashua Brookside Hospital); and Morris B. Parloff, Ph.D. (now at Georgetown University). The Principal Investigators and Project Coordinators at the three participating research sites were: George Washington University, Stuart M. Sotsky, M.D. and David Glass, Ph.D.; University of Pittsburgh, Stanley D. Imber, Ph.D. and Paul A. Pilkonis, Ph.D.; and the University of Oklahoma, John T. Watkins, Ph.D. and William Leber, Ph.D. The Principal Investigators and Project Coordinators at the three sites responsible for training therapists were: Yale University, Myrna Weissman, Ph.D. (now at Columbia University), Eve Chevron, M.S. and Bruce J. Rounsaville, M.D.; Clarke Institute of Psychiatry, Brian F. Shaw, Ph.D. and T. Michael Vallis, Ph.D.; and Rush Presbyterian-St. Luke's Medical Center, Jan A. Fawcett, M.D. and Phillip Epstein, M.D. Collaborators in the data management and data analysis aspects of the program were C. James Klett, Ph.D., Joseph F. Collins, Sc.D., and Roderic Gillis of the Veterans Administration Cooperative Studies Program, Perry Point, Maryland.

REFERENCES

Cohen J, Cohen P (1983): Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, 2nd Ed. Hillsdale, NJ, Lawrence Erlbaum Associates

Elkin I, Parloff MB, Hadley SW, Autry JH (1985): NIMH Treatment of Depression Collaborative Research Program: Background and research plan. Arch Gen Psychiatry 42:305–316

Elkin I, Shea MT, Watkins JT, Imber SD, Sotsky SM, Collins JF, Glass DR, Pilkonis PA, Leber WR, Docherty JP, Fiester SJ, Parloff MB (1989): National Institute of Mental Health Treatment of Depression Collaborative Research Program; General Effectiveness of Treatments. Arch Gen Psychiatry 46:971–982

Elkin I, Shea MT, Collins JF, Klett CJ, Imber SD, Sotsky SM, Watkins JT, Parloff MB (1990): Comment, in reply. NIMH

Collaborative Research on Treatment of Depression. Arch Gen Psychiatry 47:684–685

Fawcett, J (1990): Remarks to Society for Clinical Psychosocial Research, May 1990

Hays, WL (1988): Statistics, 4th Ed. New York, Holt, Rinehart & Winston

Hirschfeld RMA (1990): Comment. NIMH Collaborative Research on Treatment of Depression. Arch Gen Psychiatry 47:685–686

Hochberg Y, Tamhane AC (1987): Multiple Comparison Procedures. New York, John Wiley & Sons

Huitema BE (1980): The Analysis of Covariance and Alternatives. New York, John Wiley & Sons

Johnson PO, Neyman J (1926): Tests of certain linear hypotheses and their application to some educational problems. Stat Res Memoirs 1:57–93

Klein DF (1990): NIMH Collaborative Research on Treatment of Depression. Arch Gen Psychiatry 47:682–684

Klein DF, Davis JM (1969): Diagnosis and Drug Treatment of Psychiatric Disorders. Baltimore, Williams & Wilkins

Klein DF, Rabkin JG (1984): Specificity and strategy in psychotherapy research and practice. In Williams J, Spitzer R (eds), Psychotherapy Research: Where are we and Where Should we Go? New York, Guilford Press

Potthoff RF (1964): On the Johnson-Neyman technique and some extensions thereof. Psychometrika 29:241–256

Rothman JK (1990): No adjustments are needed for multiple comparisons. Epidemiology 1:43–46

Saville DJ (1990): Multiple comparison procedures: The practical solution. Am Stat 44:174–180