The Quality of Clinical Trials Published in *The Journal of Family Practice*, 1974–1991

Jeffrey Sonis, MD, MPH, and Jerry Joines, MD
Chapel Hill, North Carolina

Background. Previous analyses of published clinical trials have identified major deficiencies in reporting, design, analysis, and overall quality. The purpose of this study was to determine the strengths and weaknesses of published clinical trials in family practice, and to identify predictors of quality in these trials.

Methods. Randomized controlled clinical trials published in *The Journal of Family Practice* from 1974 to 1991 were eligible for the study. Two raters independently evaluated the adequacy and appropriateness of reporting, design, and analysis for each clinical trial, using the Chalmers index for assessing clinical trial quality. Multiple linear regression was used to determine the predictors of quality.

Results. The 53 trials included in the study showed deficiencies in reporting, design, and analysis, although fundamental design issues, such as blinding, were a relative strength. On average, the trials scored 35% of the possible points on the scale. Three factors were positively associated with overall quality: year of publication, number of pages of the published report, and the type of intervention. Trials with pharmacologic and nonmedication therapy interventions, such as diet, had higher quality scores than did trials with psychosocial or educational interventions.

Conclusions. The overall quality of these clinical trials was less than optimal but comparable to previously analyzed groups of trials. The improvement in quality over time may be related to improvement in the quality of the trials themselves, or more exacting editorial standards, or a combination of the two.

Key words. Clinical trials; randomized controlled trials; meta-analysis; research design; quality of research.
(*J Fam Pract 1994; 39:225-235*)

Although many clinical trials are methodologically sound, even a casual review of published clinical trials reveals that many trials have not incorporated fundamental principles of clinical trial research. Analyses of clinical trials published in a wide variety of journals have identified large deficiencies in reporting, [1–4] design, [5–7] analysis,[8–11] and overall quality.[12–18]

These findings may not be generalizable to clinical trials in the family practice research literature for several reasons. First, to our knowledge, few if any of the previous analyses included trials from the family practice literature. Second, because family practice is a relatively new discipline, few clinical trials have been published in this specialty.[19–22] Third, given the importance of psychosocial factors in family practice theory, the content of clinical trials in family practice may be different from that of other disciplines.

Several previous analyses of clinical trials also have attempted to identify predictors of overall quality. In a study of breast cancer trials, Liberati and colleagues[17] showed that the quality could be predicted by the year the trial started and biostatistician involvement. In an analysis of clinical trials from a variety of disciplines, Emerson and colleagues[18] showed that quality could be predicted by year of publication and clinical content. However, there are several important factors that may be related to quality that were not considered in either study. The type of intervention (eg, medication vs patient education), research training of the authors (eg, authors with PhD or

Submitted, revised, February 28, 1994.

From the Cecil G. Sheps Center for Health Services Research and the Department of Family Practice (J.S.), and the Department of Internal Medicine (J.J.), University of North Carolina at Chapel Hill. Requests for reprints should be addressed to Jeffrey Sonis, MD, MPH, University of Michigan, Department of Family Practice, 1018 Fuller St, Ann Arbor, MI 48109-0708.

ISSN 0094-3509

MPH degrees), affiliation of the authors (university-based vs practice-based), and size of the research team (number of authors) all may be related to quality of the clinical trial. Since adequacy of reporting is one component of the quality of the published report, the number of pages also may be related to quality of the published report.

The purpose of this study was to answer the following three research questions: (1) What are the strengths and areas for improvement of published clinical trials in the discipline of family practice, as reflected by clinical trials published in *The Journal of Family Practice*? (2) How do trials published in *The Journal of Family Practice* compare with those of other disciplines? (3) What are the predictors of quality?

To answer these questions, we performed a cross-sectional analysis of clinical trials published in *The Journal of Family Practice,* using a standard instrument, the Chalmers index,[12] to assess trial quality. Although original research in family practice is published in a variety of medical journals, this study is limited to clinical trials in *The Journal of Family Practice* for two reasons. First, it is the primary journal for original research in the discipline. Faculty who seek or have been nominated for academic promotion[23,24] are far more likely to publish in *The Journal of Family Practice* than in any other single journal. As noted in a 1989 review of articles published in family practice, "*The Journal of Family Practice* remains the principal repository of original work in the field."[19] Second, if clinical trials from multiple family practice journals had been eligible for inclusion, it would have been impossible to disentangle the effect of year of publication from that of "start-up" difficulties for new journals.

Methods

Identification and Selection of Randomized Controlled Trials

A published study was eligible for the current study if it met all of the following inclusion criteria: (1) it was published in *The Journal of Family Practice* between 1974 (volume 1) and 1991 (volume 33), inclusive; (2) it was a prospective study comparing the effect of an intervention against a control or against another intervention in human subjects; (3) its intervention was allocated by randomization.

A study was excluded if it did not meet inclusion criteria or if it met any one of the following exclusion criteria: (1) the intervention was allocated to a group rather than individual subjects (community intervention trial); (2) the intervention was a diagnostic test and the purpose of the study was to assess the accuracy (sensitivity and specificity) of the test; (3) the study protocol, design, or methods were reported in a separate journal article.

Studies were identified by a MEDLINE search using the following key words: trial, control, controlled trial, randomized, randomization, placebo, blinded, crossover. The abstract and, if necessary, the methods sections of all articles identified by the search were reviewed to determine eligibility. Eligibility was determined by a single reviewer.

To determine the adequacy of this search strategy, an alternative strategy was used after the study had been completed. The abstracts of the clinical trials used in the study were reviewed to identify key words that were not included in the original search, but that might increase the sensitivity of the search. The following additional key words were identified: intervention, effectiveness, comparison, impact, improvement, therapy, treatment, assessment, evaluation, experimental, modification. These key words were then used in a repeat MEDLINE search and a manual search of the table of contents of volumes 1 through 33 of *The Journal of Family Practice.* If one of the key words was identified in the repeat MEDLINE search, or in the title, or the two-sentence description of the study in the manual search, the abstract was then reviewed to determine if the study was a randomized clinical trial. None of trials identified using this alternative strategy were included in the sample because evaluating the trials after the predictors of quality had been identified could have introduced substantial bias.

The Scoring System

The authors, who have training in epidemiology and biostatistics, independently read and scored each article using the quality index developed by Chalmers and colleagues.[12] In a deviation from the Chalmers protocol, we were not blinded to identifying information in the articles included in the sample because we also had to determine whether studies identified by the MEDLINE search met inclusion criteria for entry.

The purpose of the Chalmers instrument is to evaluate the quality of published clinical trials based on the adequacy and appropriateness of reporting, design, and analysis. The Chalmers instrument is designed to be flexible enough to evaluate clinical trials with any type of content or intervention. The items included in the current evaluation are shown in Table 1. Each item in the Chalmers instrument is weighted arbitrarily according to its putative relative contribution to overall quality of the

Table 1. Items from Chalmers Index Used to Assess the Quality of 53 Clinical Trials Published in *The Journal of Family Practice*, 1974–1991

Item	Definition or Description	Points*
1. Selection description	Detailed inclusion and exclusion criteria	3
2. Reject log	Number of subjects rejected and reasons	3
3. Withdrawals	Number of withdrawals <15% for long-term trials, or <10% for short-term trials	3
4. Therapeutic regimens definition	Completeness of description	3
5. Placebo appearance	Placebo similar to active treatment?	1.5
6. Placebo taste/sensation	Placebo similar to active treatment?	1.5
7. Randomization blinding	Randomization process itself blinded	10
8. Blinding of subjects	Subjects blinded to intervention	8
9. Blinding of observers	Observers blinded to intervention	8
10. Blinding of observers and subjects	Both blinded to ongoing results of trial	4
11. Sample size	Sample size calculated before subjects randomized	3
12. Testing randomization	Prognostic factors tested for comparability across intervention groups	3
13. Testing blinding	Subjects and observers tested for adequacy of blinding	3
14. Testing compliance	Subjects tested for adherence to intervention regimen	3
15. Biological equivalent	Physiologic surrogate (eg, blood levels of medication) measured, if appropriate	3
16. Endpoint duplicate variable	Subjective endpoints determined by multiple observers	3
17. Stopping criteria	Rules for terminating trial	3
18. Major endpoints statistics	Both test statistic value and *P* value reported	3
19. Posterior beta estimate of observed differences for negative trials	Type II error discussion if no difference between interventions	3
20. Statistical inference: confidence limits	Confidence intervals or standard errors reported for major findings	3
21. Statistical inference: life table	Survival analysis used for discrete endpoints	3
22. Timing of events (raw data)	Raw data presented for survival analysis	4
23. Appropriate statistical analysis	Is analysis appropriate?	3
24. Handling of withdrawals	Method of analysis of withdrawals	4
25. Side effects, statistical discussion	Side effects reported and analyzed	3
26. Blinding of statistician or analyst to results	Data submitted to statistician with intervention groups coded	2
27. Regression/correlation	Multivariate techniques used, where appropriate	2
28. Dates of starting and stopping	Beginning and end dates for trial presented	2
29. Results of prerandomization: data analysis	Differences in prognostic factors between intervention groups, if present, considered in interpretation of results	2

**Indicates maximum number of points per item.*

trial. For instance, blinding of subjects to the intervention (item 8) is assigned 8 points, but handling of withdrawals (item 24) is assigned only 4 points because incomplete blinding of subjects is thought to contribute more bias than inappropriate handling of withdrawals. After independent evaluation and scoring of articles, we met to resolve differences by consensus. The consensus opinion was considered the final result for each study.

The overall quality score for a trial equals the total number of points obtained by the trial divided by the total possible points for the trial. Items that were not applicable to a particular trial were not counted in the points scored (numerator) or the total possible points (denominator) and therefore do not affect the overall quality score for that trial. For instance, in a patient education trial, in which blinding of subjects to the intervention might be impossible, item 8 is scored as not applicable, and the 8 points assigned to that item are not included in either the numerator or denominator. If an item was applicable to a trial but the article provided either insufficient information or no information on that item, the trial received no points in the numerator for that item. The overall score for each trial represents the ratio of points obtained by the trial to the total possible points for that trial. Since the overall score is a proportion, the maximum is 1.00 and the minimum is 0.

Statistical Analysis

Three measures of interrater reliability are reported: percent agreement and kappa, which were calculated by the method of Fleiss,[25] and the intraclass correlation coefficient for the association between overall quality scores for each trial by each of the raters.[26] Descriptive statistics are reported for selected items on the index and follow standard formulas.[27]

A multiple linear regression analysis was performed to identify factors that could predict the quality of the studies. The overall quality score was used as the dependent variable. The following variables, representing factors that might predict the quality score, were used as independent variables in the regression analyses: (1) year

of publication, (2) biostatistician involvement, (3) affiliation (university vs other), (4) number of investigators, (5) sample size, (6) research training of authors, (PhD, PharmD, or MPH vs none), (7) number of pages, and (8) type of intervention. When the trials were initially reviewed, they were classified into one of the following six intervention categories: medication trial, nonmedication therapy (eg, diet), psychosocial (eg, hypnosis, stress-reduction program), patient education (eg, interventions to improve patient knowledge about diabetes or improve compliance with appointments), physician education (eg, intervention to improve physician compliance with preventive health guidelines), and other. For the regression analysis, four intervention categories (psychosocial, patient education, physician education, and other) were collapsed into a single category, and intervention type was then coded as two dummy variables: medication trial vs the collapsed category ("intervention 1"), and nonmedication therapy trial vs the collapsed category ("intervention 2").

The final model (best subset) was identified by backward elimination, with significant level for staying set at $\alpha=.10$.[28] Alpha was set at .10, rather than the traditional .05, to avoid underfitting the model. This was done because the sample size was small and fixed, given that nominal *P* values using backward elimination cannot be taken at face value because of multiple comparisons during the variable selection process.[29] For comparison purposes only, the model was compared with models obtained using forward selection with significant level for entering $\alpha=.10$[28]; stepwise with significant level for entering and significant level for staying $\alpha=.10$[28]; and all possible regressions, using Mallow's C_p as the criterion.[29] The final model was tested for collinearity and assessed for violation of the following assumptions: error normality, independence of errors, homogeneity of error variance, and linearity.[28-32] The data set was examined for outliers by graphical and statistical evaluation of jackknife residuals, leverage, and Cook's distance.[28] Cross validation (reliability) of the final model was determined by calculation of the $R^2_{\text{prediction}}$ from the PRESS statistic.[29] ($R^2_{\text{prediction}}$ from PRESS is a method of assessing the reliability [applicability of the model to new samples from the same population] when the sample size is too small to assess reliability through split-sample techniques.)

All statistical analyses were conducted with SAS PC software, version 6.04, except for χ^2 for linear trend and χ^2 for independence, for which Epi Info, version 5.0, was used.[33,34]

Results

Studies Identified

Sixty-one randomized clinical trials published in *The Journal of Family Practice* between 1974 and 1991, inclusive, were identified using the MEDLINE search strategy. Of these, 53 met inclusion criteria for entry into this study. Of the eight studies excluded, five were excluded because they were community intervention trials, two because their primary purpose was to determine the sensitivity and specificity of a diagnostic test, and one because the protocol was reported in an article published outside *The Journal of Family Practice*. (The list of trials identified, excluded, and selected is available on request.)

The sensitivity of the original search strategy was 87% (95% confidence interval [CI], 78% to 96%). Only eight randomized clinical trials that should have been included in the sample were missed from volumes 1 through 33.

Descriptive Features

Although *The Journal of Family Practice* was founded in 1974, the first randomized controlled trial was published in 1977. Both the number and proportion of total articles published as randomized trials increased over time. (χ^2 for linear trend = 22.181, $P<.001$). More than one half of the trials included in the sample were published after 1986. The content of the trials appeared to reflect family medicine's attention to nonbiomedical factors in health[35]; more than one half of the trials studied interventions other than medication. Descriptive features of the trials are listed in Table 2.

Interrater Reliability

There were 29 items on the scale. The overall percent agreement between the two raters was 81% (95% CI, 79% to 83%), resulting from 1239 agreements of 1537 items checked. Kappa for the entire index was .57. There were clear differences in kappa between different items on the scale. Items that required simple determination of the presence or absence of a feature, such as calculation of sample size, had very high interrater reliability, whereas items that required complex or subjective judgments, such as appropriateness of statistical analysis, had much lower interrater reliability. The overall interrater reliability in this study was comparable to that of similar studies and was within bounds that are widely considered fair to good.[2,3,15,17,18] Moreover, the intraclass correlation between the overall quality scores assigned by each of the raters was .82, indicating excellent agreement between the two raters regarding overall quality.

Table 2. Descriptive Features of 53 Randomized Trials Published in *The Journal of Family Practice*, 1974–1991

Feature	No. (%) Trials
Author affiliation	
University-based	36 (68)
Practice-based	3 (6)
Other	14 (27)
Author degree (PhD, MPH or PharmD)	
No	26 (49)
Yes	27 (51)
Intervention type	
Medication trial	26 (49)
Nonmedication therapy	9 (17)
Patient education	8 (15)
Physician education	5 (9)
Psychosocial intervention	3 (6)
Other	2 (4)
Number of published pages	
2-3	7 (13)
4-5	28 (53)
6-7	12 (23)
8-9	6 (11)
Sample size	
0-49	13 (25)
50-99	12 (23)
100-199	12 (23)
200-299	5 (9)
300-499	4 (8)
500-999	3 (6)
1000-1587	3 (6)

Note: Percentages may not add to 100 because of rounding.

Quality Scoring

The mean quality score (standard deviation [SD]) was .35. ± 17, (95% CI, .30 to .40), indicating that, on average, a clinical trial scored 35% of the possible points on the scale. The lowest score was .05 and the highest was .73. Most scores were between .1 and .6 (Figure 1).

There were large differences among individual items pertaining to the quality of reporting (Table 3). Almost all the trials reported selection criteria and the therapeutic regimen in sufficient detail to help readers assess the generalizability of the findings. Side effects were discussed by about one half of the trials. Although *P* values convey no information on the magnitude or precision of an effect, only 23% of the trials reported confidence intervals or standard errors (from which confidence intervals could be calculated) for effect measures on major trial endpoints.

There were also large differences among items pertaining to the quality of design and conduct of the trial (Table 4). The trials performed well on two of the three major design elements in a randomized trial: blinding of subjects and blinding of observers. Most (70%) of the trials blinded subjects appropriately when possible, and fewer, but still a majority (57%) of trials, blinded observers appropriately when possible. Only 19 (36%) of the 53

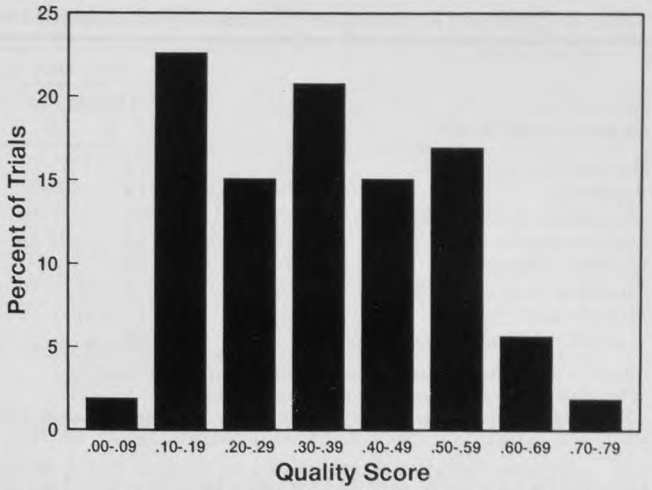

Figure 1. Frequency distribution of quality scores for 53 clinical trials published in *The Journal of Family Practice*, 1974–1991. Scores are based on the Chalmers index for assessing the quality of clinical trials.[12] Overall scores are a ratio of points scored to total possible points.

trials received full or partial credit for blinding of subjects. This item, however, did not apply to 26 (49%) of the trials because the subjects could not be blinded to the intervention; thus, 19 of the 27 trials for which the item was applicable, or 70%, received full or partial credit for blinding subjects. Although blinding of the randomization process is the fundamental guarantor of unbiased allocation, only 28% of trials received full or partial credit on this third major design element.

Other design elements appeared to be more problematic. Although 38 (72%) of the 53 trials received full or partial credit for comparing the distribution of demographic and prognostic factors across the intervention groups, it appeared that many of the 38 trials used this comparison as a test of confounding rather than as a test of the randomization process. Only 9% of the trials reported prior sample size calculations, which may partially explain why 68% showed no statistically significant difference between intervention groups on the major trial endpoint. None of the trials reported testing the effectiveness of any type of blinding. Twenty-seven percent of the trials had <15% withdrawals for trials of more than 3 months' or <10% withdrawals for trials of less than 3 months' duration.

A similar pattern of differences among items pertaining to analysis quality was observed (Table 5). Almost one half (49%) of the trials received full or partial credit for analyzing or discussing the impact of covariate imbalances on the observed results, or both. While 36 (68%) of the 53 trials reported negative findings on major endpoints, only 9 (25%) of the 36 calculated the power of the study to detect clinically meaningful differences. Although 74% of

Table 3. Quality of Reporting in 53 Clinical Trials Published in *The Journal of Family Practice*, 1974–1991

Chalmers Index Item*	Trials Receiving Full Credit, %	Partial Credit, %	No Credit, %	Not Applicable to Trials, %
Selection description	70	17	13	0
Reject log	15	13	72	0
Therapeutic regimens definition	94	6	0	0
Major endpoints statistics	23	62	15	0
Confidence limits	23	0	77	0
Timing of events	2 (7)‡	4 (14)	21 (79)	74†
Side effects	17 (18)	30 (31)	49 (51)	4
Dates of [trial] starting and stopping	38	0	62	0

**Based on the Chalmers index for assessing the quality of clinical trials.*[12]
†Percentages for an item may not equal 100 due to rounding.
‡Numbers in parentheses represent the percentage after excluding the trials for which an item was not applicable.

the trials had a large number of withdrawals, only 24% of the trials that had any withdrawals analyzed subjects in the original group to which they were randomized (ie, "intention to treat" analysis).

Predictors of Quality

Multiple linear regression analysis, using backward elimination, identified the following predictors of overall quality: year of publication, number of pages in the published report, medication trial intervention, and nonmedication therapy intervention. Each of these factors was positively associated with overall quality, as shown by the positive parameter estimates in Table 6. None of the following factors contributed to the prediction of overall quality score: biostatistician involvement, affiliation, number of investigators, sample size, or authors with PhD, PharmD or MPH degrees. The model predicted 41% of the variation in overall quality score.

The same four-variable model (year of publication, number of pages, medication trial intervention, and nonmedication therapy intervention) was also identified by forward selection, stepwise, and all possible regressions using Mallow's C_p,[29] supporting the validity of the model.

The multiple regression model shows that year of publication was by far the strongest predictor of quality score ($P=<.001$, $F>14.17$), confirming a similar finding from two previous studies.[17,18] The parameter estimate of .0177 for year of publication indicates that a trial published in 1991, for instance, would be expected to have a quality score .25 points higher than one published in 1977 (the year of publication of the first clinical trial in *The Journal of Family Practice,*) after adjusting for number of pages and type of intervention.

The type of intervention was also a strong predictor of quality score ($P=.005$, $F>8.54$ for the dummy variable coding for medication trials). The parameter estimate of

Table 4. Quality of Design and Conduct in 53 Clinical Trials Published in *The Journal of Family Practice*, 1974–1991

Chalmers Index Item*	Trials Receiving Full Credit, %	Partial Credit, %	No Credit, %	Not Applicable to Trials, %
Withdrawals	27	0	74	0
Placebo appearance	19 (48)†	0	21 (52)	60
Placebo sensation	6 (15)	0	32 (85)	62
Randomization blinding	26	2	72	0
Blinding of subjects	34 (66)	2 (4)	15 (30)	49
Blinding of observers	51 (55)‡	2 (2)	40 (43)	8
Blinding re: results	4	0	96	0
Prior estimate of sample size	9	0	91	0
Testing randomization	51	21	28	0
Testing blinding	0 (0)	0 (0)	62 (100)	38
Testing compliance	36 (37)	15 (15)	47 (48)	2
Biological equivalent	8 (80)	0	2 (20)	91
Endpoint duplicate variable	2 (5)	0 (0)	34 (95)	64
Stopping criteria	9	0	91	0
Blinding of statistician	4	0	96	0

**Based on the Chalmers index for assessing the quality of clinical trials.*[12]
†Numbers in parentheses represent the percentage after excluding the trials for which an item was not applicable.
‡Percentages for an item may not total 100 because of rounding.

Table 5. Quality of Analysis in 53 Clinical Trials Published in *The Journal of Family Practice*, 1974–1991

Chalmers Index Item*	Trials Receiving Full Credit, %	Partial Credit, %	No Credit, %	Not Applicable to Trials, %
Posterior beta for negative trial	17 (25)†	17 (25)	34 (50)	32
Life table analysis	0 (0)	0 (0)	25 (100)	76‡
Appropriate statistical analysis	13	64	23	0
Handling of withdrawals	11 (12)	11 (12)	71 (76)	6
Regression/Correlation	30	0	70	0
Results of prerandomization	40	9	51	0

**Based on the Chalmers index for assessing the quality of clinical trials.*[12]
†Numbers in parentheses represent the percentage after excluding the trials for which an item was not applicable.
‡Percentages for an item may not total 100 because of rounding.

.1227 indicates that a medication trial would be expected to have a quality score approximately .12 points higher than that of the group of trials with one of the following interventions: psychosocial, patient education, physician education, or other.

Model assumptions and absence of collinearity were confirmed. No observations were identified as being either outliers or extremely influential. The model had fair to good reliability ($R^2_{\text{prediction}}=.29$) explaining 29% of the variability in predicting new observations. Since the model predicted 41% of the variability in the original data ($R^2=.41$), there is mild to moderate degradation of the model, but this degradation is not severe enough to raise serious objections to the model. Moreover, when a trial regression was run deleting only one mildly influential observation, there was almost no degradation in the model.

Discussion

There were substantial deficiencies in reporting, design, and analysis of clinical trials published in *The Journal of Family Practice* between 1974 and 1991, although there were several areas of strength in important design elements, such as blinding of subjects and observers. The pattern of strengths and weaknesses exhibited by these trials is not unique to clinical trials in family practice. Two previous collections of clinical trials, analyzed using the Chalmers index, showed similar strengths and weaknesses.[16,17]

Table 6. Predictors of Quality in 53 Clinical Trials Published in *The Journal of Family Practice*, 1974–1991

Variable*	Parameter Estimate (β)	Standard Error	F Statistic	*P* Value
Intercept	−1.3730	0.4037	11.56	0.0014
Year	0.0177	0.0047	14.17	0.0005
Pages	0.0230	0.0115	4.00	0.0512
Intervention 1†	0.1227	0.0420	8.54	0.0053
Intervention 2†	0.1100	0.0570	3.72	0.0597

**Based on the Chalmers index for assessing the quality of clinical trials.*[12]
†Intervention 1 is dummy variable coding for type of intervention: medication trial vs patient education, psychosocial intervention, physician education, and other, combined. Intervention 2 is dummy variable coding for type of intervention: nonmedication therapy vs patient education, psychosocial intervention, physician education, and other, combined.

The overall quality score for this group of trials, .35 (95% CI, .30 to .40) appears to be within the general range of scores obtained by previously analyzed groups of trials (Figure 2). In the absence of information on the variability in quality scores within groups, it is impossible to determine whether the group means differ by more than chance. What is striking, though, is not the small differences in mean quality scores between the groups, including family practice, but the large discrepancy between the mean quality score for any of the groups and the ideal (a quality score of 1.0). The important finding is that clinical trials both in family practice and in longer-established biomedical disciplines have substantial room for improvement.

Year of publication was strongly and positively associated with quality score, confirming similar findings of two previous analyses of clinical trials.[17,18] This improvement in quality over time may be related to improvement in quality of the trials themselves, more exacting editorial standards over time, or some combination of the two. It is impossible to determine the relative contribution of each factor to the observed improvement in quality over time from the results of this study, but it is reasonable to speculate that both factors are important.

The strong association between the type of intervention and quality score has important implications for clinical trials in family practice. Less than one half of the trials in *The Journal of Family Practice* were medication trials, which is consistent with the expressed goals of family practice to look beyond the strictly biomedical context of health and illness.[35] However, since medication trials and trials of nonmedication therapies had significantly higher quality scores than did other types of trials, it appears that the methodologic and analytic strengths of these therapeutic trials have not been incorporated into trials of

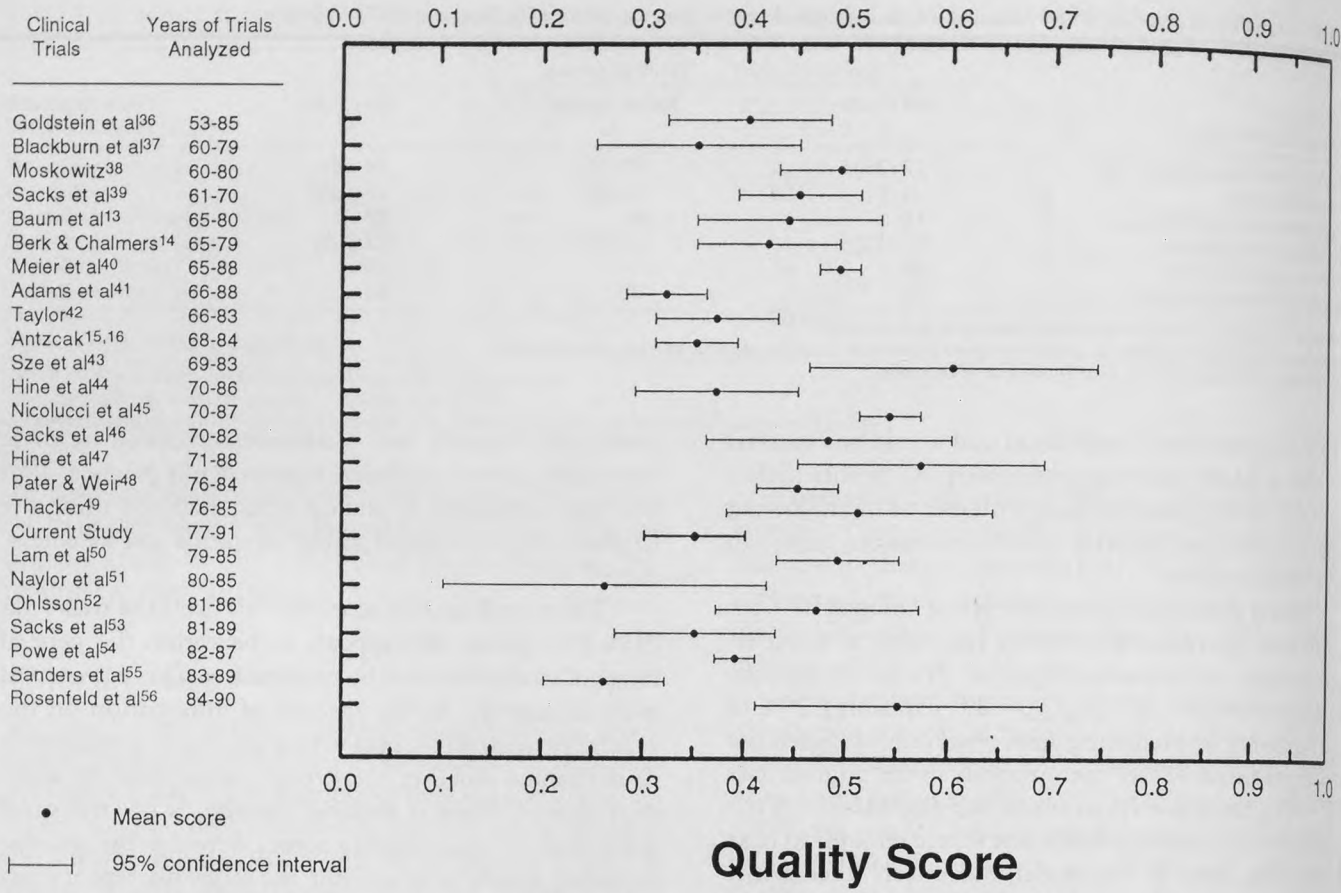

Figure 2. Quality scores of 25 studies that assessed the quality of clinical trials published in various journals, 1953–1991. Scores are based on the Chalmers index for assessing the quality of clinical trials.[12] Overall scores are a ratio of points scored to total possible points.

other types of interventions. If clinical trials in family practice are to continue to reflect the interests of this discipline and if these trials are to meet high scientific standards, much more attention must be devoted to strengthening the methodologic and analytic rigor of nontherapeutic clinical trials (eg, those involving patient and physician education interventions and psychosocial interventions).

There are several limitations to this study. First, the original search strategy did not capture all the clinical trials published in *The Journal of Family Practice* between 1974 and 1991. However, even marked selection bias would have had relatively little impact on the overall findings. Assuming that all eight of the trials missed by the original strategy would have received either the lowest (.05) or the highest (.73) quality scores, the mean overall quality score would be .31 and .40, respectively. Thus, even if all of the missed trials had received extreme scores, which is a highly unlikely scenario, the mean overall quality score would not have changed to a remarkable degree.

Second, since all the trials included in this study are from a single journal, the findings on study quality may not be generalizable to clinical trials that are conducted by family physicians and published in other journals. Further, the findings on predictors of quality may not be generalizable to published clinical trials in general. However, the similarity between trials in this study and other groups of analyzed clinical trials in both overall quality score and the pattern of strengths and weaknesses suggests that the findings regarding predictors of quality are probably generalizable to trials published in other journals. This assumption could be validated only by replicating this study in clinical trials from a wide variety of sources.

Third, the physician readers were not blinded to the identifying information in the articles. Observer bias cannot, therefore, be ruled out. It is possible, for instance, that knowledge of previous research indicating improvement in the quality of published clinical trials over time influenced the reviewers in subtle ways. However, as Detsky and colleagues[57] note, "the benefits of these elaborate precautions" (such as blinding reviewers to identifying information) are "purely speculative."

Fourth, the Chalmers index has not been definitively

demonstrated to be either reliable or valid. The index may have reasonable content validity, however, since it was developed by a group of experts with extensive experience in conducting and evaluating clinical trials. Criterion validity of the index is suggested by a study that showed that a series of trials that were "widely held to be truly high quality" had high quality scores on the Chalmers index, and that the rank ordering of 18 clinical trials did not change substantially when ranking by Chalmers score was compared with ranking based on scores on two other (unvalidated) instruments for assessing quality.[57] Both findings are suggestive of criterion validity but fall short of definitive validation because there is no clear "gold standard" for research "quality."

Fifth, since a trial was scored as zero for any item on the Chalmers index for which there was insufficient information in the article to determine whether the item had been carried out, the overall quality score reflects, at least to some degree, the completeness of the published report, rather than just the quality of the trial itself. This limitation was confirmed by the association in the current study between the number of pages and overall quality. However, a previous study of the quality of published clinical trials showed that the quality score was improved by an average of only 7% when the principal investigators of the trials were contacted by telephone to determine whether specific items that were unclear in the article had in fact been carried out.[17] Thus, while the quality score reflects to some degree the adequacy of reporting, it is probably a reasonably good indicator of the quality of design and analysis of the trial itself. The current study could have been improved had we contacted the authors of the trials rather than relying solely on the published report to determine whether specific design items had been carried out.

Sixth, excluding nonapplicable items (by not adding the points assigned to the nonapplicable item to either the points scored by a given trial or to the total points possible for that trial) probably introduced positive bias in the quality score, given that the average score for these trials was low. Trials with patient education, physician education, and psychosocial interventions have the largest number of items that do not apply, and therefore are the most likely to have artifactually high quality scores. As a result, the true difference in quality between both medication trials and nonmedication therapy trials, and those involving patient education, physician education, and psychosocial interventions, as a group, is probably larger than the difference reported in this study.

Finally, direction and magnitude of bias are not adequately incorporated into the overall score using the Chalmers scale. A clinical trial could receive a high quality score yet be completely invalid if the magnitude of bias in one area is large. For instance, the selection bias resulting from a large number of withdrawals in an otherwise well-designed and well-executed clinical trial could be large enough to invalidate the findings, regardless of the overall quality score.

Several interesting questions are raised by our findings and suggest fruitful areas for further research. First, can clinical trials using interventions other than medications or diet be improved simply by a more consistent application of existing methodology appropriate to medication trials, or will new intervention-appropriate methodologies need to be developed? This question is vital to future research in family practice. Second, what is the quality of observational (nonexperimental) research in family practice and what are the predictors of quality? Is the pattern of strengths and weaknesses similar to the pattern for clinical trials? Third, can an instrument be developed that assesses the quality of research articles and incorporates direction and magnitude of bias, instead of indicating simply the presence or absence of flaws as does the Chalmers index?

There are several ways that the quality of clinical trials in family practice could be improved. First, each of the items that received low scores in this study should be considered areas for special attention. For instance, family practice researchers should be especially careful to minimize withdrawals and analyze them appropriately, since this seems to be a problematic area. Second, the items on the Chalmers index (or a suitable alternative) could be used by researchers as a type of checklist during the planning stages of a clinical trial and by reviewers during the review of the manuscript for publication. Clinical trials are particularly suited to this type of planning aid, since many of the elements of sound design require attention to detail more than special expertise or large amounts of money. Since physician adherence to preventive service guidelines can be improved by simple checklists,[58,59] it is possible that researchers and reviewers can be influenced in a similar way. Although high-quality research cannot be guaranteed simply by adherence to guidelines, the items on the Chalmers index (or some subset of these items) could alert researchers to consider each of the areas before conducting a trial or submitting the results for publication, and alert reviewers to the same issues during the formal review of the manuscript. If an item does not apply to the specific trial, such as blinding of subjects to educational interventions, the researcher or reviewer could ignore the item. By using these guidelines, the researcher or reviewer would at least be guaranteed that important issues in reporting, design, and analysis have been considered. Authors of previous analyses of published clinical trials have offered similar recommendations.[9,10]

Unquestionably, there are disagreements among

clinical trial methodologists about the importance or validity of specific items on the Chalmers scale, and there are other important issues in reporting, design, and analysis that have not been incorporated into the Chalmers scale.[9] Nevertheless, we believe that the quality of clinical trials could be improved substantially if investigators conformed to a widely agreed upon set of principles. The Chalmers index is an important first step in that direction. A logical next step would be the development of a simpler checklist that would be easier to use while retaining important issues related to reporting, design, and analysis.

Acknowledgments

We would like to thank Herman A. Tyroler, MD, Donald Pathman, MD, MPH, Mary Sonis, RN, and the three anonymous reviewers for helpful comments on previous drafts of the manuscript.

References

1. Mosteller F, Gilbert JP, McPeek B. Reporting standards and research strategies for controlled clinical trials. Controlled Clin Trials 1980; 1:37–58.
2. DerSimonian R, Charette LJ, McPeek B, Mosteller F. Reporting on methods in clinical trials. N Engl J Med 1982; 306:1332–7.
3. Emerson JD, McPeek B, Mosteller F. Reporting clinical trials in general surgical journals. Surgery 1984; 95:572–9.
4. Kelen GD, Brown CG, Moser M, Ashton J, Rund DA. Reporting methodology protocols in three acute care journals. Ann Emerg Med 1985; 14:880–4.
5. Chalmers TC, Celano P, Sacks HS, Smith H. Bias in treatment assignment in controlled clinical trials. N Engl J Med 1983; 309: 1358–61.
6. Colditz GA, Miller JA, Mosteller F. How study design affects outcomes in comparisons of medical therapy. I: medical. Stat Med 1989; 8:441–54.
7. Miller JN, Colditz GA, Mosteller F. How study design affects outcomes in comparisons of therapy. II: surgical. Stat Med 1989; 8:455–66.
8. Lavori PW, Louis TA, Bailar JC III, Polansky M. Designs for experiments—parallel comparisons of treatment. N Engl J Med 1983; 309:1291–9.
9. Altman DG, Dore CJ. Randomization and baseline comparisons in clinical trials. Lancet 1990; 335:149–53.
10. Gardner MJ, Machin D, Campbell MJ. Use of check lists in assessing the statistical content of medical studies. BMJ 1986; 292:810–2.
11. Pocock SJ, Hughes MD, Lee RJ. Statistical problems in the reporting of clinical trials. N Engl J Med 1987; 317:426–32.
12. Chalmers TC, Smith H, Blackburn BA, Silverman B, Schroeder B, Reitman D, Ambroz A. A method for assessing the quality of a randomized control trial. Controlled Clin Trials 1981; 2:31–49.
13. Baum ML, Anish DS, Chalmers TC, Sacks HS, Smith H, Fagerstrom RM. A survey of clinical trials of antibiotic prophylaxis in colon surgery: evidence against further use of no-treatment controls. N Engl J Med 1981; 305:795–9.
14. Berk AA, Chalmers TC. Cost and efficacy of the substitution of ambulatory for inpatient care. N Engl J Med 1981; 304:393–7.
15. Antczak AA, Tang J, Chalmers TC. Quality assessment of randomized control trials in dental research I. Methods. J Periodont Res 1986; 21:305–14.
16. Antczak AA, Tang J, Chalmers TC. Quality assessment of randomized control trials in dental research II. Results: periodontal research. J Periodont Res 1986; 21:315–21.
17. Liberati A, Himel HN, Chalmers TC. A quality assessment of randomized control trials of primary treatment of breast cancer. J Clin Oncol 1986; 4:942–51.
18. Emerson JD, Burdick E, Hoaglin DC, Mosteller F, Chalmers TC. An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials. Controlled Clin Trials 1990; 11:339–52.
19. Geyman JP, Berg AO. The Journal of Family Practice 1974–1988: window to an evolving academic discipline. J Fam Pract 1989; 28:301–4.
20. Geyman JP, Berg AO. The Journal of Family Practice—1974–1983: analysis of an evolving literature base. J Fam Pract 1984; 18:47–51.
21. Fromm BS, Snyder VL. Research design and statistical procedures used in *The Journal Of Family Practice*. J Fam Pract 1986; 23: 564–6.
22. Franks P. Clinical trials. Fam Med 1988; 20:443–8.
23. Gjerde C, Clements W, Clements B. Publication characteristics of family practice faculty nominated for academic promotion. J Fam Pract 1982; 15:663–6.
24. Gjerde C. Where are articles by candidates for academic promotion published? J Fam Pract 1992; 34:449–53.
25. Fleiss JL. Statistical methods for rates and proportions. 2nd ed. New York, NY:Wiley, 1981.
26. Haas M. Statistical methodology for reliability studies. J Manipulative Physiol Ther 1991; 14:119–32.
27. Daniel WW. Biostatistics: a foundation for analysis in the health sciences. 4th ed. New York, NY: Wiley, 1987.
28. Kleinbaum DG, Kupper LL, Muller KE. Applied regression analysis and other multivariate methods. 2nd ed. Boston, Mass:PWS-Kent Publishing Co, 1988.
29. Draper NR, Smith H. Applied regression analysis. 2nd ed. New York, NY:Wiley, 1981.
30. Neter J, Wasserman W. Applied linear statistical models. Homewood, Ill:Richard D. Irwin, Inc, 1974.
31. Montgomery DC, Peck EA. Introduction to linear regression analysis. New York, NY:Wiley, 1982.
32. White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica 1980; 48:817–38.
33. SAS/STAT user's guide, release 6.03 edition. Cary, NC:SAS Institute, 1988.
34. Dean AD, Dean JA, Burton AH, Dicker RC. Epi Info, version 5. Stone Mountain, Ga:USD Incorporated, 1990.
35. Culpepper L. Family medicine research. In: Mayfield J, Grady ML. Primary care research: an agenda for the 90's. Rockville, Md:US Department of Health and Human Services, Agency for Health Care Policy and Research, report No. 90–17.
36. Goldstein P, Berrier J, Rosen S, Sacks HS, Chalmers TC. A meta-analysis of randomized control trials of progestational agents in pregnancy. Br J Obstet Gynecol 1989; 96:265–74.
37. Blackburn BA, Smith H, Chalmers TC. The inadequate evidence for short hospital stay after hernia or varicose vein stripping surgery. Mt Sinai J Med 1982; 49:383–90.
38. Moskowitz G, Chalmers TC, Sacks HS, Fagerstrom RM. Deficiencies of clinical trials of alcohol withdrawal. Alcohol Clin Exp Res 1983; 7:42–6.
39. Sacks HS, Ancona-Berk A, Berrier J, Nagalingam R, Chalmers TC. Dipyridamole in the treatment of angina pectoris: a meta-analysis. Clin Pharmacol Ther 1988; 43:610–5.
40. Meier WS, Schmitz PIM, Jeekel J. Meta-analysis of randomized controlled clinical trials of antibiotic prophylaxis in biliary tract surgery. Br J Surg 1990; 77:283–90.
41. Adams ME, McCall NT, Gray DT, Orza MJ, Chalmers TC. Economic analysis in randomized control trials. Med Care 1992; 30: 231–43.
42. Taylor SH. Secondary prevention after myocardial infarction: facts and fallacies. J Cardiovasc Pharmacol 1984; 6:5914–21.
43. Sze PC, Reitman D, Pincus MM, Sacks HS, Chalmers TC. Am

platelet agents in the secondary prevention of stroke: meta-analysis of randomized control trials. Stroke 1988; 19:436–42.
44. Hine LK, Laird N, Hewitt P, Chalmers TC. Meta-analytic evidence against prophylactic use of lidocaine in acute myocardial infarction. Arch Intern Med 1989; 149:2694–8.
45. Nicolucci A, Grilli R, Alexanian AA, Apolone G, Torri V, Liberati A. Quality, evolution, and clinical implications of randomized, controlled trials on the treatment of lung cancer: a lost opportunity for meta-analysis. JAMA 1989; 262:2101–7.
46. Sacks HS, Chalmers TC, Berk AA, Reitman D. Should mild hypertension be treated? An attempted meta-analysis of the clinical trials. Mt Sinai J Med 1985; 52:265–70.
47. Hine LK, Laird N, Hewitt P, Chalmers TC. Meta-analysis of empirical long-term antiarrhythmic therapy after myocardial infarction. JAMA 1989; 262:3037–40.
48. Pater JL, Weir L. Reporting the results of randomized trials of empiric antibiotics in febrile neutropenic patients—a critical survey. J Clin Oncol 1986; 4:346–52.
49. Thacker SB. The efficacy of intrapartum electronic fetal monitoring. Am J Obstet Gynecol 1987; 156:24–30.
50. Lam W, Sze PC, Sacks HS, Chalmers TC. Meta-analysis of randomized controlled trials of nicotine chewing gum. Lancet 1987; 2: 26–30.
51. Naylor CD, O'Rourke K, Detsky AS, Baker JP. Parenteral nutrition with branched-chain amino acids in hepatic encephalopathy. Gastroenterology 1989; 97:1033–42.
52. Ohlsson A. Treatments of preterm premature rupture of the membranes: a meta-analysis. Am J Obstet Gynecol 1989; 160:890–905.
53. Sacks HS, Chalmers TC, Blum Al, Berrier J, Pagano D. Endoscopic hemostasis: an effective therapy for bleeding peptic ulcers. JAMA 1990; 264:494–9.
54. Powe NR, Kinnison ML, Steinberg EP. Quality assessment of randomized controlled trials of constrast media. Radiology 1989; 170: 377–80.
55. Sanders JW, Powe NR, Moore RD. Ceftazidime monotherapy for empiric treatment of febrile neutropenic patients: a meta-analysis. J Infect Dis 1991; 164:907–16.
56. Rosenfeld RM, Mandel EM, Bluestone CD. Systemic steroids for otitis media with effusion in children. Arch Otolaryngol Head Neck Surg 1991; 117:984–9.
57. Detsky AS, Naylor CD, O'Rourke K, McGeer AJ, L'Abbe KA. Incorporating variations in the quality of individual randomized trials into meta-analysis. J Clin Epidemiol 1992; 45:255–65.
58. Becker MH, Janz NK. Practicing health promotion: the doctor's dilemma. Ann Intern Med 1990; 113:419–22.
59. Cheny C, Ramsdell JW. Effect of medical records' checklists on implementation of periodic health measures. Am J Med 1987; 83: 129–36.