Aid to the Evaluation of Therapeutic Studies

Joan S. Reisch, PhD, Jon E. Tyson, MD, and Susan G. Mize, BA

From Academic Computing Services, and the Department of Pediatrics, Southwestern Medical School, University of Texas Southwestern Medical Center at Dallas

ABSTRACT. Physicians are often faced with conflicting recommendations from therapeutic studies. An evaluation form is proposed to facilitate the evaluation of the quality of therapeutic studies and the resulting treatment or management recommendations in any area of medicine. Twelve major topics for evaluation include sample size determination, randomization, selection of control group(s), "blinding," and support for treatment recommendations. Emphasis is placed on study design and performance rather than data analysis. Thirty-four primary criteria based on accepted research standards are designated as most important, and examples from the literature are provided to illustrate their use. The form provides a comprehensive set of well-accepted standards of research in a format that encourages detailed, consistent, and thoughtful evaluation of therapeutic studies. The evaluation form is recommended as a tool for physicians who wish to develop and exercise skill in evaluating therapeutic studies. Pediatrics 1989;84:815–827; therapeutic study.

When deciding how to treat their patients, physicians often encounter conflicting treatment recommendations and study results in the medical literature. Critical assessment of therapeutic studies is often required to distinguish genuine therapeutic advances from ineffective or possibly hazardous treatment methods recommended without adequate study.1–3 Unfortunately, many physicians are inadequately prepared to evaluate medical research.4–7

A number of excellent articles devoted to specific topics in statistical analysis or experimental design have appeared in clinical journals but, in only a few, are methods to evaluate overall study quality or the justification for the author's treatment recommendations addressed.8–9 For this reason, we have developed a comprehensive evaluation form to facilitate the evaluation of the design and performance of therapeutic studies and the recommendations of investigators in any area of medicine (Appendix). The evaluation criteria are based on widely accepted standards for research. The form has been used in teaching critical reading skills to pediatric house officers and fellows and in conducting surveys of therapeutic studies in perinatal medicine.10–13 The original version of the form has been modified to make it simpler and more self-explanatory.

The purpose of this paper is to describe the form and to define and discuss the evaluation criteria in sufficient detail to make the form useful to the reader. Because we believe that studies with questionable results can usually be identified by physicians without a detailed understanding of statistics, emphasis is placed on study design and performance rather than data analysis. The rationale and use of these standards are described and illustrated with examples taken largely from our surveys of the perinatal literature. We provide an introduction to the evaluation of study design and performance with references for physicians interested in learning more.

FORM FORMAT AND USE

The form was designed for assessing studies in which different methods of treatment or management are compared or a recommendation for or against the use of a particular treatment or management method is based on the findings of the study. Each criterion is phrased in question form, requiring a specific response, usually "Yes," "No," "Unclear or Unknown," or "Not Applicable." Criteria should be considered "Not Applicable" only when they could not be applied if the study were properly designed to answer the question being investigated. For example, failure to "blind" care givers to treatment method should be considered not applicable to retrospective studies only if
“blinding” would not be possible in a prospective study of the same treatment method.

The criteria that we consider most important and most objective are designated as primary criteria. Some important criteria were considered important but too subjective to be used as primary criteria (eg, whether all treatment groups received the same level of supportive care [item 7C]). For each of the primary criteria, an asterisk appears on the form beside the most desirable response. A cross appears beside each “Not Applicable” response to primary criteria.

An evaluation index at the end of the form is the percentage of primary criteria fulfilled after excluding any considered not applicable. (As described at the end of the form, this index is calculated from the total number of responses marked by an asterisk and the total number marked by a cross.) Because even a single flaw in study design or performance may invalidate the study results, a high index does not preclude misleading findings. However, the index is of value to identify studies that have multiple deficiencies and considerable opportunity for bias or random error to compromise the validity of the results.

The evaluation criteria will be described and the number of the corresponding item on the form is noted in parentheses.

PURPOSE OF STUDY (1)

Common errors in therapeutic studies include a failure to define before initiating the study the specific question(s) to be answered (1B), the outcome or response variable(s) for assessing the treatments being investigated (1C), and the difference in outcome between treatment or management method groups that the investigation is designed to identify (eg, a 25% or greater difference in mortality) (1D).

Specifying the smallest outcome difference under investigation is important so that the clinician will know whether a “negative” study might have failed to discern an important difference between treatment methods. In a survey of 71 “negative” clinical trials, 50 studies (70%) had insufficient patients to identify a 50% difference in the effectiveness of the treatment or management methods. Few of the investigators noted a need to evaluate more patients.

Identification of the source of support (1E) for the study may provide a better understanding of the motivation for, and potential biases of, the study.

EXPERIMENTAL DESIGN (2)

When describing study design, investigators have used the terms “retrospective” and “prospective” in a variable fashion. For example, therapeutic studies have been designated as retrospective because of either the method of data collection (review of information, such as patient records, collected for purposes other than the study) or the method of subject selection (case-control studies in which patients are selected according to outcome and the investigator “follows” the subjects backward in time [from effect to cause] to review earlier care). Data from case-control studies usually, but not always, are collected by review of patient records.

The reader should evaluate both the collection of data and the selection of subjects. Bias or inaccuracies in measurement are less likely in studies in which data collection is planned before the use of the treatment or management method and conducted under specified conditions (2A1). Biases in the selection of patients are less likely when subjects are selected before treatment and followed forward in time to outcome. Case-control studies are less expensive and more easily performed but subject to a variety of potential sources of error. Such studies may be the only feasible way to study rare conditions and complications, and potential sources of bias may be minimized by careful planning and execution.

SAMPLE SIZE DETERMINATION (3)

To recognize studies with misleading results, the clinician needs to understand proper methods of determining the number of subjects (patients). Altman noted that “perhaps the most common design error is to have too small a sample size to get reliable and/or useful results.” Improper methods of determining sample size tend to reduce the number of patients studied and increase the likelihood of either false-negative results or false-positive results. The grounds for deciding that sufficient patients have been evaluated should be defined before beginning the study. Otherwise, as the study progresses, the investigators may be most likely to discontinue subject enrollment at times when the findings appear to confirm the hypothesis. This has been referred to as “stopping the race while your horse is ahead.”

False-positive results are especially likely if the study is discontinued on the basis of statistical tests performed repeatedly as the results accrued until a “statistically significant” P value (eg, P < .05) is obtained. When standard statistical methods are used, the cumulative probability of calculating such a P increases each time the analyses are repeated. For example, if tests of significance are performed five times without adjustment for repeated analyses, the likelihood of calculating a P < .05 is approximately 23%, not 5%. Even when statistical analyses are properly conducted, Chalmers and
Sinclair have shown “that statistically significant results in small trials tend to overestimate the differences.”

To avoid these problems and enroll sufficient subjects, the number of patients to be studied can be decided before study. A predetermined sample size is properly calculated from the magnitude of the treatment or management method effect being investigated (1D) and the largest chance of a false-positive result (type I error) or a false-negative result (type II error) acceptable to the investigators. Investigators virtually always describe the probability of a false-positive result (or significance level) but often fail to indicate either the probability of the false-negative result or the statistical power (1 minus the probability of a false-negative result). Der Simonian and colleagues conclude that, when neither the likelihood of a false-negative result nor statistical power is described, “the reader has a right to suspect the study was not large enough to detect important differences. . . . Most investigators seem not to realize their obligation to report on this item.”

A sequential design (3A1b) may be used in lieu of a predetermined sample size (3A1a). These designs, although somewhat complex and not always applicable, provide a legitimate method to consider results as they accrue, to calculate an appropriate P value despite multiple analyses, and to base sample size on study findings. Although used infrequently, sequential designs have been used in some well-done clinical trials. Expert monitoring committees uninvolved in the performance of the study may provide a legitimate method to decide when to discontinue a study, particularly when unanticipated hazards are identified (3A1c).

Other approaches to determining sample size are less satisfactory. A predetermined period of study (3A2) may not allow a suitable number of subjects but does prevent investigator bias from affecting sample size.

To allow the reader to assess the adequacy of the sample size (3C), the investigators should either describe the magnitude of treatment difference being investigated or calculate confidence intervals to indicate the range of values likely to contain the true difference in the outcome of the treatment groups. For example, a confidence interval of 5% to 35% would indicate that the true difference between treatment groups is likely to be no less than 5% and no more than 35%. (The confidence intervals when there is no significant difference between groups will include a value of 0.) Small studies are likely to have large confidence intervals. Such studies, however, can provide data useful in an overview (meta-analysis) combining the results of all studies of the same issue.

Unfortunately, often neither the treatment difference being investigated nor confidence intervals are provided. The reader who wishes to calculate confidence intervals can refer to standard statistical texts. However, Detrsky and Sackett have published tables to provide a simple method to assess whether enough subjects were evaluated in “negative” studies with no significant difference between the proportion of treatment failures in two treatment groups. The minimum number of patients in each treatment group to be confident that a true 25% difference has not been missed is given in the Table. P refers to the proportion of patients with an unfavorable outcome in the control group (group given no treatment, placebo treatment, or standard treatment); Pt refers to the proportion with an unfavorable outcome in the second or treatment group.

In studies discontinued because of unanticipated treatment hazards or toxicity, sample size may be considered adequate (regardless of that originally planned) providing the adverse effects appear to be sufficiently common or severe (3C). Greater benefit than expected may not justify early study termination unless appropriate adjustments were made for repeated analyses, the stopping rules were carefully considered before study, a convincing P was calculated (usually P < .01) for important outcome variable(s), and the findings are not attributable to factors other than treatment efficacy.

DESCRIPTION AND SUITABILITY OF SUBJECTS (4)

The criteria for subject inclusion and exclusion (eg, age, race, sex, disease/health status, complications, etc) should be identical for all treatment groups and should, to the extent possible, be clearly defined (4A) before study. Predefined exclusion criteria (4A6) are especially important in retrospective studies to avoid the possibility that the investigator’s awareness of patient outcome might unintentionally influence patient selection.

For a variety of reasons, many eligible patients may not be admitted to the study. These patients should be described (4B) to indicate whether they differ from those enrolled. Those enrolled should be adequately described with respect to all important demographic and medical factors (4C), in part to indicate whether they are typical of patients for whom the study was intended.

The subjects may be well described but unsuitable for the questions being investigated (4D). Studies of newborns cared for in regional referral centers may provide a misleading assessment of the effectiveness of neonatal intensive care due to a variety of selection biases that might influence referrals to


**TABLE.** Equivalence Testing: Sample Size Required (per Group) to Reject a True Risk Reduction of 25%*

<table>
<thead>
<tr>
<th>Pt</th>
<th>0.95</th>
<th>0.9</th>
<th>0.85</th>
<th>0.8</th>
<th>0.75</th>
<th>0.7</th>
<th>0.65</th>
<th>0.6</th>
<th>0.55</th>
<th>0.5</th>
<th>0.45</th>
<th>0.4</th>
<th>0.35</th>
<th>0.3</th>
<th>0.25</th>
<th>0.2</th>
<th>0.15</th>
<th>0.1</th>
<th>0.05</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>14</td>
<td>27</td>
<td>68</td>
<td>391</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.9</td>
<td>11</td>
<td>18</td>
<td>38</td>
<td>110</td>
<td>1057</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.85</td>
<td>14</td>
<td>25</td>
<td>54</td>
<td>185</td>
<td>4889</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>11</td>
<td>18</td>
<td>33</td>
<td>78</td>
<td>326</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.75</td>
<td>13</td>
<td>22</td>
<td>44</td>
<td>112</td>
<td>635</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.7</td>
<td>11</td>
<td>16</td>
<td>28</td>
<td>57</td>
<td>165</td>
<td>1524</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.65</td>
<td>13</td>
<td>20</td>
<td>35</td>
<td>75</td>
<td>250</td>
<td>6349</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td>10</td>
<td>15</td>
<td>24</td>
<td>43</td>
<td>99</td>
<td>402</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.55</td>
<td>12</td>
<td>17</td>
<td>28</td>
<td>53</td>
<td>132</td>
<td>722</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>13</td>
<td>20</td>
<td>33</td>
<td>65</td>
<td>180</td>
<td>1607</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.45</td>
<td>10 patients</td>
<td>15</td>
<td>22</td>
<td>38</td>
<td>79</td>
<td>254</td>
<td>6232</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td>per group</td>
<td>11</td>
<td>16</td>
<td>25</td>
<td>44</td>
<td>98</td>
<td>381</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.35</td>
<td>12</td>
<td>18</td>
<td>28</td>
<td>50</td>
<td>121</td>
<td>634</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.3</td>
<td>13</td>
<td>19</td>
<td>30</td>
<td>57</td>
<td>152</td>
<td>1296</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.25</td>
<td>10</td>
<td>13</td>
<td>20</td>
<td>33</td>
<td>64</td>
<td>196</td>
<td>4537</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>10</td>
<td>14</td>
<td>20</td>
<td>34</td>
<td>71</td>
<td>261</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.15</td>
<td>10</td>
<td>13</td>
<td>20</td>
<td>34</td>
<td>80</td>
<td>589</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>12</td>
<td>17</td>
<td>30</td>
<td>74</td>
<td>1245</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.05</td>
<td>12</td>
<td>17</td>
<td>30</td>
<td>74</td>
<td>1245</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* When a trial is finished and results are expressed in terms of the proportion of patients having some event in the control and experimental groups, this table shows how many patients were needed per group to be confident (α = 0.05, one-tailed) that a clinically important (25%) risk reduction in the experimental group was not missed. This table should be used only for interpreting finished trials, not for planning trials. Pt represents the observed event rate in the experimental group; Pc, the observed event rate in the control group. Table reprinted with permission from Arch Intern Med. 1985; 145:709–712. Copyright 1985, American Medical Association.

such units (eg, death of the sickest infants before transfer is feasible).

**RANDOMIZATION AND STRATIFICATION (5)**

Altman and colleagues contended that “any lack of randomization should be noted as a deficiency in design and the reasons given.” Sackett advised busy physicians to ignore nonrandomized studies when reading to “keep up” with therapeutic advances, noting that the efficacy of new treatment or management methods can be demonstrated by such studies “only when traditional therapy is invariably followed by death.” Despite negative results in randomized trials, the administration of diethylstilbestrol to approximately 3 million pregnant women in the United States was encouraged by a series of nonrandomized trials reporting reduced pregnancy losses. The serious long-term hazards to the offspring were not identified for many years.

Randomized studies are important not only to compare different treatment agents but also to define appropriate indications for treatment (eg, what blood levels of bilirubin, oxygen, or glucose require treatment). Often, an attempt is made to define such levels in observational studies. However, patient outcome (eg, the developmental delay associated with bilirubin levels of only 10 mg/dL in low birth weight infants) may be affected by a variety of confounding factors (eg, a higher incidence of intracranial hemorrhage in jaundiced than nonjaundiced infants). Experimental studies in which jaundiced infants are randomly assigned to relatively liberal or conservative indications for therapy are required to define the true risk of hyperbilirubinemia and the appropriate guidelines for therapy. The use of observational studies rather than randomized trials contributes to the persistence of many unresolved issues in neonatology, eg, the relation of retrolental fibroplasia to arterial PaO₂ values.

The reader should be alert to studies in which randomization may be improperly performed or inadequately documented (5B). Assigning treatments alternately or by day of the week or hospital identification number allows the investigator to anticipate treatment assignment. Subconscious bias may then influence which patients the investigator considers eligible for study and how diligently informed consent for study participation is sought. In a survey of treatment or management methods for myocardial infarction, significant differences in mortality were reported almost three times as often in randomized studies in which the investigators could anticipate treatment assignment than in randomized studies in which treatment assignment could not be foreseen. Appropriate randomization may involve calling a central office for the treatment assignment or having drugs or dietary items randomized and prepackaged for each patient in advance of study entry. The opening of opaque sealed envelopes containing the treat-
ment allocation may be satisfactory but is less “foolproof.”

Zelen and others have proposed alternatives to randomization including “play the winner,” “two-armed bandit,” and adaptive treatment assignment methods. For a variety of reasons, these methods have been used infrequently and their validity (as in a recent study of extracorporeal membrane oxygenation for seriously ill newborns) questioned. The usefulness and validity of other approaches (e.g., randomization before rather than after obtaining consent) also remain uncertain.

Randomized treatment assignment reduces, but does not eliminate, the possibility of important pretreatment differences between groups which make it difficult to interpret differences observed after treatment. Important pretreatment differences were observed in 14% of properly randomized studies but in 58% of nonrandomized studies in the survey cited before. Pretreatment differences may be avoided if investigators can stratify or group patients according to factors that influence outcome or response to therapy. In prospective studies, patients in each group or stratum can then be randomly assigned to the treatment methods, thereby increasing comparability of the patients in different treatment or management groups. For example, in a study relating fluid intake to the development of patent ductus arteriosus, prognostic stratification was used to group preterm infants according to three risk factors (birth weight, weight for gestational age, and respiratory disease). Within each group, patients were then randomly assigned to either a high or low fluid intake.

Stratification may be marked “Not Applicable” if no major risk factors are known or if the number of patients enrolled was so large that the likelihood of important pretreatment differences is remote (e.g., studies involving hundreds of patients). Stratification may be used retrospectively to analyze subgroups or to adjust for recognized imbalances between groups. However, important and unrecognized pretreatment differences are most likely when randomization is not or cannot be used. For example, the extent to which the differences between infants of alcoholic mothers and control mothers might result from factors other than alcohol is unclear.

Whether or not stratification is used, the reader should consider whether differences between treatment or management groups before treatment might limit the interpretability of the study (5E). The apparent benefits of steroid administration to infants with bronchopulmonary dysplasia in one randomized trial may reflect lower ventilatory and oxygen requirements in the steroid-treated group than in the controls. Too few patients were studied for a sizable difference to be statistically significant.

Despite the importance of randomized treatment assignment, the reader should be aware that randomized trials may provide misleading results not only because of inadequate sample size and pretreatment differences but also because of factors such as selection biases in exclusion of patients from analysis (9), biases in evaluation by “unblinded” examiners (8), or enrollment of a population that is not representative of the population treatment recommendations for which they are made (12).

**COMPARISON OR CONTROL GROUP USE (6)**

It has been said that “the best way to improve results is to omit controls.” The control or comparison group(s) in therapeutic studies may receive (1) no treatment; (2) another treatment or management method; or (3) a placebo, which should be identical with the active agent in color, size, volume, and any other distinguishing feature. An untreated control group as well as a placebo group may be used to assess the effects of treatment. Three methods of random assignment to the control therapies are listed: (1) subjects may be assigned at random to the treatment and control groups (6A1); (2) subjects may be used as their own control with the order of administering different treatment or management methods randomly assigned, with time between treatments allowed as necessary for treatment effects (e.g., drug effects) to subside before initiating another treatment or management method (6A2); (3) subjects may be matched according to risk factors before randomization (6A3) to treatment or management groups. In studies in which treatment is not randomly assigned, the comparison groups may be selected in any of a variety of ways (6C to 6I).

Some authorities advocate historical or other nonrandomized control groups in the belief that the number of persons exposed to an inferior treatment or management method will be minimized. However, uncontrolled studies have often been the basis for widespread use of ineffective or hazardous treatment or management methods. Despite difficult ethical problems of using controls in some kinds of studies, Shaw and Chalmers, as well as others, argued that randomization is a more ethical way of practicing medicine than is the routine use of possibly ineffective or hazardous treatment methods.

**PROCEDURES FOR TREATMENT/ MANAGEMENT (7)**

As a result of increasing professional, institutional, public, and governmental concern, specific
standards to protect human subjects have become “almost universal” in research in the United States. Among the standards is a provision for obtaining the consent of eligible subjects based on an informed consideration of relevant information and issues.

What constitutes informed consent (7A) is somewhat controversial. Moreover, the desirability or feasibility of obtaining truly informed consent in some circumstances is open to question. The requirement for informed consent may be considered not applicable for reviews of routinely collected clinical information and for some therapeutic studies, provided approval has been obtained from the institutional review board or ethics committee.

The following general comment from the Belmont Report may assist the reader in evaluating the adequacy of consent obtained in therapeutic studies:

Respect for persons requires that subjects, to the degree that they are capable, be given the opportunity to choose what shall or shall not happen to them. This opportunity is provided when adequate standards for informed consent are satisfied. While the importance of informed consent is unquestioned, controversy prevails over the nature and possibility of an informed consent. Nonetheless, there is widespread agreement that the consent process can be analyzed as containing three elements: information, comprehension and voluntariness.

The description of a therapeutic study should include all reasons to initiate, modify, or discontinue the treatment or management (7B8). All treatment or management groups should receive the same attention and care other than the treatment or management methods being investigated (7C). This criterion is important but difficult to evaluate when the care givers are aware of the treatment received by individual patients. Whether a treatment or management is adequately described (7D) should be assessed in terms of the indications, dose, and duration of treatment and all other factors important to efficacy and hazards (7E). An exemplary description of management methods appeared in one trial of electronic fetal heart rate monitoring. This report clearly described not only management protocols but also the percentage of patients for whom the prescribed monitoring was not feasible or the heart rate tracings were uninterpretable or misinterpreted.

BLINDING (MASKING) (8)

Whenever feasible, the investigators, care givers, and patients should be unaware of, or blinded (masked) to, which treatment or management is used. When the care givers and patients cannot be blinded, blinding is often feasible for the investigators who evaluate important outcome variables: those identified in the abstract or statement of purpose as measures of differences between groups resulting from their treatment or management. Blinding may be considered unnecessary for mortality data or for routine laboratory tests performed by hospital technicians unlikely to be aware of or interested in the investigation. However, without blinding, errors in recording and tabulating data may tend to favor the treatment method preferred by the investigators.

Blinding may be claimed despite treatment effects which may be obvious to patients, care givers, or investigators (8A) (eg, the effect of β-blockers on heart rate in a blinded trial of agents to stop preterm labor). Given the variety of ways that the treatment or management methods can be randomized, careful investigators may assess whether blinding was maintained by determining whether the subjects and staff could detect treatment assignments.

Failure or inability to use blinding should be considered likely to bias the results (8D) if knowledge of treatment method could be expected to influence patient care, evaluation, or outcome. Unblinded investigations should be evaluated carefully for subtle evidence of bias (eg, different time or attention devoted to patients in each treatment or management group), particularly in studies in which patient care is difficult or complex. This subtle bias may account for the initial report in an unblinded study that vitamin E therapy prevents bronchopulmonary dysplasia. No beneficial effect was found when the study was blinded and repeated in the same institution.

SUBJECT ATTRITION (9)

Differences in outcome of different treatment or management groups are more convincing if no subjects are excluded after randomization. All eligible subjects should be included in the analysis of primary response variables, if possible; this is referred to as “the intent to treat” principle. It is particularly important to adhere to this principle because exclusion of eligible subjects who were noncompliant with the protocol or who were lost to follow-up may dramatically change the treatment comparison results. The likelihood of bias in exclusion of subjects after study entry can be minimized if predefined criteria are used (9A).

Specific procedures should be used to minimize patient loss (9B). Loss of subjects or records is common in both prospective and retrospective studies. However, the opportunity for bias in loss or exclusion of patients is usually less in prospective studies than in retrospective studies. Whenever
possible, all subjects removed from the study and all lost records should be described (9C) in enough detail for the reader to determine whether their loss is likely to influence study outcome. The sickest or least compliant patients are often most likely to be lost or excluded from the study. Patients who fail to comply are likely to fare worse than other patients for a variety of reasons other than the efficacy of their treatment. In the Coronary Drug Project, the mortality of noncompliers in the placebo group was more than twice that of compliers (9.9 vs. 4.8%). Thus, noncompliers should not be excluded from one group unless they are excluded from all groups. Even then, it may be inappropriate to exclude noncompliers, in part because of differing reasons for noncompliance (eg, treatment or management side effects) in different groups. Attrition is likely to bias study results (9D) if the apparent outcome of any treatment group is likely to be affected more than the others. The importance of this issue is apparent in a report of a multicenter trial claiming that sulfinpyrazone (Anturan) reduced sudden death from myocardial infarction by 74%. The Food and Drug Administration refused to recognize this claim, noting that the investigators had failed to indicate that a disproportionate number of deaths excluded from the results were in the sulfinpyrazone-treated group.

EVALUATION OF SUBJECTS AND TREATMENT EVALUATION (10)

The investigators should include all clinical information important in evaluating the condition and response of patients in different treatment or management groups (10A).

Laboratory tests and other evaluation methods can be considered standardized and consistent (10B) if (1) for generally available tests and procedures, standard methods are used and reasonable values are obtained. However, the reader should appreciate that reliability may be surprisingly poor even for common procedures (eg, the laboratory determination of serum bilirubin or the radiologic diagnosis of necrotizing enterocolitis by pediatric radiologists); (2) when unusual procedures are used or unexpected results obtained, reliability is verified appropriately (eg, correlation coefficients for repeated analysis). References to previous reliability studies should not be considered relevant unless the patient population, clinical setting, and range of values are comparable to those being studied.

Whenever feasible, compliance should be evaluated for each treatment or management group (10C). Direct observation of the treatment or management, review of hospital records, measurement of drug levels in blood or urine, or assessment of the amount of medication remaining may be used. The reader will recognize that compliance is difficult to evaluate and that commonly used methods may overestimate compliance.

Evaluation methods can be considered appropriate (10E) if they appear to be both a reliable (reproducible) and a valid (correct) measure of outcomes appropriate to the question(s) posed by the investigators. For example, the interpretation of follow-up studies of high-risk newborns is often obscured by a failure to determine the reliability of the assessment techniques as well as a failure to blind the examiners to neonatal findings.

Potential treatment hazards (10F) may be evaluated in explanatory trials that are performed to determine treatment or management efficacy under restricted or ideal circumstances or to define the mechanism of treatment or management effects. Hazards should be evaluated in management trials that are performed to evaluate efficacy under usual clinical circumstances. Management trials represent the appropriate way to evaluate therapeutic recommendations for general use. Management trials should include a prospective evaluation of all likely important hazards (previously reported or theoretically likely hazards that might have a major effect on the risk to benefit ratio). These may include relatively minor, but common, side effects, or major, although uncommon, hazards.

The reader should consider whether sufficient patients were studied to evaluate important hazards. Failure to appreciate that insufficient patients had been studied when a 50% increase in mortality was not statistically significant contributed to the belief that the oxygen concentration administered to premature infants with severe respiratory distress could be safely restricted. As a result, the number of deaths resulting from well-meaning efforts to prevent retrolental fibroplasia appears to be many fold greater than the reduction in blind survivors.

Except in explanatory trials, cost-effectiveness should be discussed (10G) if the cost of care is likely to be substantially increased. In a careful evaluation, treatment or management methods may be compared in terms of cost per extra survivor, cost per year of survival gained, or cost per quality of adjusted year of survival gained. Economic analyses of newborn intensive care provides an excellent example of the use of these measures.

PRESENTATION AND ANALYSIS OF DATA (11)

The text should be clearly understandable (11A). Variation in number of subjects in different comparisons should be explained (11B). All descriptive
measures for important variables should be identified (11C). The reader may occasionally find contradictions or computation errors (11D).

A number of surveys report that approximately half of the articles in medical journals contain errors in statistical analysis. Thus, the clinician needs a basic understanding of at least the statistical measures and techniques most commonly used in medical journals. According to a recent survey in the New England Journal of Medicine, the statistical analyses in 73% of articles would be understandable to readers who understand descriptive statistics (eg, mean, standard deviation), the Student’s test, and contingency table (χ²) tests. Considerably more knowledge would be required to understand the analyses in the remainder of studies. There is no quick way to acquire this knowledge. However, Glantz indicates that errors in use of sophisticated statistical techniques are less common than are simple mistakes such as failure to include a control group, lack of random treatment assignment, or misuse of elementary hypothesis tests. Common errors that are readily recognized include failure to name or adequately describe methods of analysis (two-tailed vs one-tailed tests; degrees of freedom); failure to correct for repeated comparisons (thereby increasing the likelihood of “significant” results); and the use of post hoc statistical tests (‘data dredging’: testing for differences that the study was not designed to detect).

To be confident whether statistical tests are clearly identified, appropriately used, and appropriately interpreted (11F), the reader may need to consult a biostatistician or any of a number of excellent articles or texts, some written primarily for physicians. When the authors specify the statistical tests used and when they appear to have been appropriately used and interpreted, the “Yes” answer in item 11G may be selected.

RECOMMENDATIONS/CONCLUSIONS (12)

A statement that the treatment or management “has a role” or “should be considered” should be regarded as a recommendation encouraging the use of the treatment or management in at least some circumstances. Recommendations for the use of the treatment or management can be considered well supported if based on a controlled randomized study (if feasible); if convincing benefit is demonstrated; and, as described before, if all important likely hazards have been adequately assessed. The overall benefit of the treatment or management and support for recommending its use can be considered unclear if the treatment or management may be substantially more expensive than alternative methods and cost effectiveness is not demonstrated. Therapeutic recommendations should be applied only to subjects and conditions similar to those studied (12B1). Rapid infusion of hypertonic sodium bicarbonate, which was used widely for years before a relationship to intraventricular hemorrhage in premature infants was suspected, was recommended largely from observations in older patients and experimental animals.

A recommendation against the use of a treatment or management is justified by observations demonstrating that the treatment or management has unacceptable hazards or is less effective or more expensive than other methods. Before one concludes that a treatment method is less effective than the others studied, confidence intervals should be calculated to indicate the likely range of differences in effectiveness (12B2).

The investigator should make no treatment recommendations (12B3) if the criteria described for positive or negative treatment recommendations are not fulfilled.

OVERALL STUDY DESIGN AND PERFORMANCE (13)

The end of the evaluation form is devoted to calculation of an index which indicates the ratio of primary criteria fulfilled to the total number of primary criteria that are applicable. When interpreting the index, one should remember that studies with a high index may not address a clinically important issue and that even a single deficiency in study design or performance may result in invalid results. Nevertheless, a low index value is meaningful as an indicator of studies that have multiple deficiencies and considerable opportunity for bias or inaccuracies to influence study results. Therapeutic recommendations are often based on studies that satisfy fewer than 50% of primary evaluation criteria.

DISCUSSION

Methods to evaluate the quality of therapeutic or nontherapeutic studies have been described by others. In designing our form, we paid particular attention to making the criteria as specific and objective as possible to evaluate the quality of the study and the justification for the conclusions. Based on our experience in teaching residents and fellows, the form has been revised to make it more understandable and useful to physicians.

The form is not intended for case reports, review articles, or nontherapeutic studies. Given the time required for a thorough evaluation of therapeutic studies (approximately 30 minutes for form completion after the user gains experience), the form is
REFERENCES

27. Bell E, Warburton D, Stonestreet B, Oh W. Effect of fluid administration on the development of symptomatic patent
51. Fries E. Informed consent may be hazardous to your health. *Science.* 1979;204:11
57. *Coronary Drug Project Research Group, Clofibrate and niacin in coronary heart disease.* JAMA. 1975;231:369-381
72. Murphy E. *Biostatistics in Medicine.* Baltimore, MD: Johns Hopkins University Press, 1982
APPENDIX: CHECK LIST FOR ASSESSING THERAPEUTIC STUDIES

<table>
<thead>
<tr>
<th>JOURNAL</th>
<th>Volume</th>
<th>Number</th>
<th>Pages to</th>
<th>Year</th>
</tr>
</thead>
</table>

Y = Yes; N = No; U = Unclear or Unknown; NA = Not Applicable;  
T/M = Treatment or Management Method

A "*" is noted beside desirable responses to the criteria considered most important.  
A "+" appears beside "Not Applicable" responses to these criteria.

1. PURPOSE OF STUDY
   A. Title consistent with purpose of the study                  Y N U  
   B. Statement of purpose given                                  Y* N U  
   C. Outcome variables for therapeutic effects defined prior to study  Y* N U  
   D. Magnitude of difference in outcome of (T/M) groups under investigation specified prior to study  Y* N  
   E. Sources of support for study specified                        Y N U  

2. EXPERIMENTAL DESIGN
   A. Data Collection (Check only one)  
      1: Data collection planned prior to T/M of subjects; data collected prospectively under specified conditions  
   2. Data collection planned prior to T/M of subjects; data collected retrospectively by record review  
   3. Data collection not planned prior to T/M of subjects; data collected retrospectively  
   B. Selection of Subjects (Check only one)  
      1. Subjects selected prior to T/M and evaluated prospectively  
   2. Subjects followed from T/M to outcome but study planned after T/M  
   3. Subjects selected according to outcome and T/M evaluated retrospectively  
   4. Unclear time relation of subject selection to outcome of T/M  
   C. Carry-over or refractory effects avoided or considered in the design of the study  

3. SAMPLE SIZE DETERMINATION
   A. Method  
      1. Sample size determined by:  
   (Indicate which)  

   a. predetermined number of subjects  
   b. sequential experimental design  
   c. independent monitoring committee  
   2. Predetermined time period  
   3. Specified time period from  
   4. No method specified (Check if applicable)  
   5. Other (describe)  

4. TOTAL NUMBER OF SUBJECTS
   A. Entry criteria  
      1. Age of subjects given  
      2. Race of subjects given  
      3. Sex of subjects given  
      4. Socioeconomic status given  
      5. Disease/health status of subjects given  
      6. Contraindications for T/M (can include other diseases or treatments)  
   B. Eligible subjects who refuse to participate are adequately described  
   C. Subjects adequately described for all appropriate criteria including those listed in 4A  
   D. Subjects selected for this study suitable for question(s) posed by these researchers  

5. RANDOMIZATION AND STRATIFICATION
   A. It is possible to design a randomized study to evaluate the T/M under consideration  
   B. Randomization claimed and documented  
   C. Randomization not performed and bias is likely
D. Use of either prognostic stratification prior to study entry or retrospective stratification during data analyses Y* N U NA+
E. Group differences limit the interpretability of this study Y N U NA

6. COMPARISON GROUP(S) (CONTROL) USAGE
A. Random T/M assignment (indicate which below)
   1. Unmatched subjects with randomized T/M assignment Y* N U NA
   2. Subjects as own control with T/M order randomized
   3. Matched by subject with T/M assignment randomized
B. No assignment method described Y N U
C. Historical Y N U
D. Subjects matched/paired but assignment to T/M groups not randomized Y N U
E. Subjects as own control but T/M order not randomized Y N U
F. Subjects compared according to their response to the T/M procedure Y N U
G. Convenience (Subjects selected for availability) Y N U NA
H. Comparison (control) group not included Y N U
I. Other non-randomized (explain) Y N U NA

7. PROCEDURES FOR TREATMENT/MANAGEMENT
A. Informed consent obtained Y* N U NA+
B. Clear specification of:
   1. Dosage Y N U NA
   2. Time of day administered Y N U NA
   3. Frequency Y N U NA
   4. Time to complete T/M Y N U NA
   5. Route (IV, IM, PO, etc.) Y N U NA
   6. Presentation (Tablet, syrup, etc.) Y N U NA
   7. Source for drug or equipment in T/M under investigation Y N U NA
   8. Indications for a. Initiation of T/M Y* N U
   b. Modification of T/M Y* N U NA+
   c. Discontinuation of T/M Y* N U NA+
C. Subjects in different T/M groups appear to receive the same care other than that under investigation Y N U NA
D. T/M adequately described for above or other appropriate criteria Y* N U
E. T/M reasonable and appropriate to answer question(s) posed by these researchers Y* N U

8. BLINDING (MASKING)
A. Blinding claimed and appears realistic Y* N U NA
B. Blinding (masking) used where feasible for important variables* by the:
   1. investigators Y N Some U NA
   2. caregivers Y N Some U NA
   3. subjects (and family if appropriate) Y N Some U NA
C. Mark Y if 8B1, B2, B3 are marked Y or NA. Mark NA + if 8B1, B2, B3 are each marked NA Y* N NA+
D. Failure to use blinding likely to bias study results Y N U NA

---We consider a variable important only when it is clearly identified by the author(s) in the abstract or in the statement of purpose to describe differences between groups related to their treatment or management.

9. SUBJECT ATTITUATION
A. Predefined procedures for excluding subjects after entry Y N U NA
B. Specific procedures established to minimize loss of subjects from this study [Answer 'NA' to 9C and 9D if no subjects or records were lost or dropped] Y N U NA
C. Description of all subjects or their records which were lost or dropped Y* N U NA+
D. Any loss of subjects or their records likely to bias the results of this study Y N U NA

10. EVALUATION OF SUBJECTS AND TREATMENT/MANAGEMENT
A. All important clinical information reported Y* N U
B. Laboratory and other measurements appear standardized and consistent Y* N U
C. Treatment compliance assessed Y* N U
D. Evaluation methods adequately described Y* N U
E. Evaluation methods appropriate to answer question(s) posed by investigators Y* N U
F. Prospective evaluation of important hazards or toxicity Y* N U NA+
G. If use of T/M increases cost of care substantially, cost-effectiveness discussed Y* N U NA+

11. PRESENTATION AND ANALYSIS OF DATA
A. Text clearly understandable Y N U
B. All comparisons involve same number of subjects or any discrepancy is explained
   Y* N U NA+

C. Descriptive measures (mean, range, standard deviation, proportion, etc.) identified for all important variables
   All Some None

D. Computation errors or contradictions identified
   Y N* U

E. Statistical tests used for comparisons involving important variables
   All Some None

F. Reported statistical tests appear to be:
   1. clearly identified
      All Some None
   2. appropriately used
      All Some None
   3. appropriately interpreted
      All Some None

G. Responses to items 11E, F1, F2, F3 marked “ALL”
   Y* N

12. RECOMMENDATIONS/CONCLUSIONS
A. Recommendation(s) are:
   1. nonexistent
   2. unclear
   3. for further study
   4. for use of T/M method
   5. against use of T/M method

13. SUMMARY OF ITEMS REVIEWED

The summary of starred items can be used as an assessment of study quality by calculating the ratio of the starred items marked by the reviewer to the maximum total possible.

The maximum total possible is determined by subtracting the Total 'NA+' responses marked by the reviewer from 34. As many as 13 'NA+' responses may be recorded (Section 4, 5, 6, 7, 8, 9, 10, 11).

34 − _____________________________(Number of NA+ Responses) = _____________________________(Maximum total possible)
(Enter maximum total possible on line 14.)

SYNOPSIS OF ITEMS REVIEWED

<table>
<thead>
<tr>
<th>Item Number</th>
<th>PURPOSE OF STUDY</th>
<th>EXPERIMENTAL DESIGN</th>
<th>SAMPLE SIZE DETERMINATION</th>
<th>DESCRIPTION AND SUITABILITY OF SUBJECTS</th>
<th>RANDOMIZATION AND STRATIFICATION</th>
<th>COMPARISON GROUP (CONTROL) USAGE</th>
<th>PROCEDURES FOR TREATMENT/MANAGEMENT</th>
<th>BLINDING (MASKING)</th>
<th>SUBJECT ATTRITION</th>
<th>EVALUATION OF SUBJECTS AND T/M</th>
<th>PRESENTATION AND ANALYSIS OF DATA</th>
<th>RECOMMENDATIONS/CONCLUSIONS</th>
<th>Fulfilled</th>
<th>Starred Items Fulfilled</th>
<th>NA+ Possible</th>
<th>No. of Fulfilled</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

LINE
13 TOTAL
14 Maximum Total Possible
15 Ratio of total to maximum total possible

Downloaded from by guest on July 22, 2016
<table>
<thead>
<tr>
<th>Updated Information &amp; Services</th>
<th>including high resolution figures, can be found at: /content/84/5/815</th>
</tr>
</thead>
<tbody>
<tr>
<td>Citations</td>
<td>This article has been cited by 18 HighWire-hosted articles: /content/84/5/815#related-urls</td>
</tr>
<tr>
<td>Permissions &amp; Licensing</td>
<td>Information about reproducing this article in parts (figures, tables) or in its entirety can be found online at: /site/misc/Permissions.xhtml</td>
</tr>
<tr>
<td>Reprints</td>
<td>Information about ordering reprints can be found online: /site/misc/reprints.xhtml</td>
</tr>
</tbody>
</table>
Aid to the Evaluation of Therapeutic Studies
Joan S. Reisch, Jon E. Tyson and Susan G. Mize

Pediatrics 1989;84;815

The online version of this article, along with updated information and services, is located on the World Wide Web at:
/content/84/5/815