Statistical Science
1998, Vol. 13, No. 3, 209–235

Clinician Preferences and the Estimation of Causal Treatment Differences

Edward L. Korn and Sheldon Baumrind

Abstract. Clinician treatment preferences affect the ability to perform randomized clinical trials and the ability to analyze observational data for treatment effects. In clinical trials, clinician preferences that are based on a subjective analysis of the patient can make it difficult to define eligibility criteria for which clinicians would agree to randomize all patients who satisfy the criteria. In addition, since each clinician typically has some preference for the choice of treatment for a given patient, there are concerns about how strong that preference needs to be before it is inappropriate for him to randomize the choice of treatment. In observational studies, the fact that clinician preferences affect the choice of treatment is a major source of selection bias when estimating treatment effects. In this paper we review alternative designs that have been proposed in the literature for randomized clinical trials that utilize clinician preferences differently than the standard randomized trial design. We also examine the effects of clinician preferences on the ability to estimate causal treatment differences from observational data, and propose an alternative method of analysis for observational data that uses clinician preferences explicitly. We report on our experience to date in using our alternative randomized clinical trial design and our new method of observational analysis to compare two treatments at the orthodontic clinics at the University of California San Francisco and the University of the Pacific, San Francisco.

Key words and phrases: Bayesian methods, causal effects, ethics, instrumental variables, randomized clinical trials, observational studies, selection bias.

1. INTRODUCTION

The goal of many clinical studies is to compare treatments in order that patients treated in the future can receive the treatment that proves superior. Studies can be classified into two types, those that involve randomization (randomized clinical trials) and those that do not (observational studies, including nonrandomized clinical trials). The randomized clinical trial (RCT) is the "gold standard" for treatment comparisons. With sufficient sample sizes, one can assure with high probability that an observed treatment difference is due to the treatments in their observed implementation and not due to selection biases, that is, that it is a causal treatment difference. Of course, it may not always be practical to perform an RCT, for example, when the treatment and outcomes take very long to be observed. In such cases, one can attempt to estimate causal treatment differences from an observational study by retrospectively comparing outcomes for different treatments. As has long been known (Byar et al., 1976), estimation of causal effects from an observational study is much more difficult than estimation from an RCT.

Edward L. Korn is Head of Clinical Trials Section, Biometric Research Branch, Executive Plaza North 739, National Cancer Institute, Bethesda, Maryland 20892. Sheldon Baumrind is Professor of Orthodontics, University of the Pacific, San Francisco; Clinical Professor of Orthodontics, University of Medicine and Dentistry of New Jersey, Newark; and Professor Emeritus, University of California, San Francisco.

Outside of a clinical trial, a clinician will treat each patient with the treatment he prefers for that patient (provided that the treatment is available and the patient prefers it). This fact can lead to problems for the design of a standard RCT and for the analysis of an observational study: part of the ethical basis for conducting an RCT is that a clinician randomizing a patient must be unsure which treatment is better for that patient (Shaw and Chalmers, 1970). It may be difficult to define eligibility criteria so that participating clinicians would be willing to randomize all patients who satisfy the eligibility criteria. This problem can become acute when a clinician's belief about relative treatment efficacy is based on subjective evaluation of the patient. A second potential problem with the standard RCT design involves the question of just how unsure a clinician has to be (about the desirability of one treatment as compared to the other) for it to be appropriate for him to randomize a patient. On the one hand, Freedman (1987) proposes that as long as there is uncertainty in the expert medical community it is appropriate to randomize regardless of the beliefs of the clinician who is asking the patient to participate in the trial. As put by Clayton (1982, page 471): "the doctor has no *ethical* responsibility to treat a patient in the manner he believes to be the best, if that belief is unsupported by evidence or consensus." On the other hand, Hellman and Hellman (1991) suggest that any belief of the clinician, even if it is as likely to be wrong as right, needs to be acted on, which would preclude randomization. We think that it is useful to frame this issue in terms of a clinical setting; one would not want a brain surgeon to treat surgically against his treatment preference (because of the hands-on nature of the treatment), but might consider it appropriate to allow a clinician to treat with a drug therapy against his preference if a preponderance of experts thought it was appropriate. In section 3, we will describe some nonstandard RCT designs that attempt to alleviate the problems of defining eligibility criteria and of having clinicians treat against their preferences. These designs are useful when it is thought to be infeasible to conduct an RCT with a standard design because of clinician preferences.

In an observational study, clinician preferences are a potential major source of bias in estimating treatment difference, since different prognostic classes of patients may be given the different treatments. Standard observational analyses attempt to control for this bias by stratifying on patient prognostic covariates. Of course, identifying covariates that can accurately predict outcome may be difficult. In Section 4, we describe standard observational analyses and a new design for observational studies that makes use of explicit statements of clinician preferences to attempt to eliminate this clinician-preference bias. We also consider an instrumental variables approach to eliminating the bias, and show it will not work in our particular application.

The structure of this paper is as follows: Section 2 defines the notion of causal treatment difference that we will use throughout; it is an extension of the ideas of Rubin (1974) to encompass clinician effects. Section 3 gives a more detailed description of the potential problems that clinician preferences can lead to in a standard RCT and discusses some alternative RCT designs. Section 4 reviews the issues involved in estimating causal treatment differences from observational data in the presence of clinician preferences, and describes our new proposal. Throughout, we use as an example the estimation of a causal treatment difference for two orthodontic treatments that will be described in Section 3. In particular, we give a first report on our experience using a new RCT design for comparing the treatments in Section 3, and we describe our pilot study using a new observational study design in Section 4. We end in Section 5 with a further discussion of what these new trial and study designs are estimating, the generalizability of their results, and their feasibility.

2. CAUSAL TREATMENT DIFFERENCES AND CLINICIAN EFFECTS

Defining precisely what one means by a causal treatment difference can require some care; see Holland (1986) for a review from a statistical point of view. We informally define a causal treatment difference for treatment A versus treatment B for a given patient as the outcome if the patient had been treated with A minus the outcome if the patient had been treated with B. In general, this is a hypothetical construct, as patients can receive only one of the two treatments. There is a long history of using such informal definitions of causal effects (e.g, see Rubin, 1990a). The definition can be formalized with certain technical assumptions (Rubin, 1974, 1990b). These are implicitly assumed in the notation used below.

Since we are interested in clinician preferences, which clinician treats which patients is important. We assume that each patient is treated by one of J clinicians, and we allow for the possibility that

outcomes may depend upon which clinician has treated the patient. This generality is required for some treatments in which the relative skills of the clinicians are important (e.g., surgery, orthodontics, psychiatric therapy), but may not be necessary for other treatments (e.g., oral chemotherapy). We define a causal treatment difference for a given patient as the (hypothetical) difference between treatment outcomes if the treatment was given by the same clinician. This difference could depend on which clinician treats the patient.

Let $\mu_u^{(T,j)}$ be the outcome if patient u were treated with treatment T by clinician j, where $T = A$ or B, $j = 1, 2, \ldots, J$. We treat the $\mu_u^{(T,j)}$ as nonrandom quantities. The causal treatment difference for clinician j for this patient is defined as $\mu_u^{(A,j)} - \mu_u^{(B,j)}$. A strong null hypothesis is given by $\mu_u^{(A,j)} - \mu_u^{(B,j)} = 0$ for all u and j. We define a causal treatment difference as a weighted mean of the differences $\mu_u^{(A,j)} - \mu_u^{(B,j)}$ over a set of patients and a set of clinicians, with non-negative weights. For a given set of weights, a weak null hypothesis is given by the weighted mean equaling zero.

Patients have different pretreatment covariates, which we denote by the vector variable x, and their outcomes to treatment may be associated with x. In the present setting, we are also interested in a special set of pretreatment variables—the preferences for treatment of the clinicians who could have treated the patient. For a given patient, let $C = (C_1, \ldots, C_J)$ be the vector of these treatment preferences for the J clinicians, $C_j = A$ or B. By "preference" we mean which of the two treatments the clinician would have used if this patient had been treated as part of the clinician's regular clinical practice. Even though the clinician may have been unsure about which treatment was better, we assume that he would have used one of the treatments; this is defined as his preference. Although clinician preferences do not "cause" different outcomes, the preferences may be associated with different outcomes by their association with different patient subsets. When required to make explicit the dependence of x, C and C_j on the patient u, we use the notation $x(u)$, $C(u)$ and $C_j(u)$.

Before leaving this section, we note that a weighted mean of the $\mu_u^{(A,j)} - \mu_u^{(B,j)}$ is not always appropriate as the target parameter for estimating treatment differences. For example, suppose patients with a certain histology and stage of cancer are treated with either medical therapy (A) or radiation therapy (B). If medical therapy is being given by a different set of clinicians (medical oncologists) from those delivering the radiation therapy (radiation oncologists), a definition involving within-clinician treatment differences makes no sense. Instead, the treatments are "medical therapy performed by a medical oncologist" and "radiation as delivered by a radiation oncologist," with a reasonable measure of treatment difference being

$$(2.1)\quad \frac{1}{n}\sum_u \left(\frac{1}{\#J_A} \sum_{j \in J_A} \mu_u^{(A,j)} - \frac{1}{\#J_B} \sum_{j \in J_B} \mu_u^{(B,j)} \right),$$

with J_T representing the set of clinicians delivering treatment T, $\#J_T$ being the number of such clinicians and n being the total number of trial participants.

3. CLINICIAN PREFERENCES AND RANDOMIZED CLINICAL TRIALS (RCT)

3.1 Standard RCTs

One of the reasons RCTs are popular is that they can be used to estimate causal treatment differences. In the present setting, we assume that each eligible patient is seen by one of J clinicians, who randomizes the patient to either A or B. The simple difference between the mean outcome of those patients treated with A ($\bar{Y}^A$) and the mean outcome of those treated with B ($\bar{Y}^B$) estimates

$$(3.1)\quad E(\bar{Y}^A - \bar{Y}^B) = \frac{1}{n}\sum_j \sum_{u \in \Pi_j} \left[\mu_u^{(A,j)} - \mu_u^{(B,j)} \right],$$

where Π_j represents the set of the trial participants treated by clinician j. (We assume that the set of patients seen by each clinician is nonrandom.) One can control for covariates x in the analysis for variance reduction or utilize them in the randomization to lessen covariate imbalances between the treatment groups. Neither of these procedures is required for the mean treatment difference to estimate a causal treatment difference without bias.

Ethical issues involving clinician preferences. As mentioned in the Introduction, having to decide between following their treatment preferences and following a randomized protocol can induce a tension in physicians. This can be seen in some survey results. In a survey of 91 surgeons who were principal investigators in a breast cancer trial (Fisher et al., 1985), 73% of the 66 surgeons not entering all eligible patients gave as one of their reasons "concern with the doctor–patient relationship in a randomized clinical trial" (Taylor, Margolese and

Soskolne, 1984). In a survey of institutions participating in trials of the Eastern Cooperative Oncology Group (ECOG), approximately half of the medically eligible patients were not entered on studies, with approximately half the reasons given for nonentry being "Physicians' preference for specific treatment or alternate therapy" (Begg et al., 1983). Based on a patient log filled out by physicians participating in the Clinical Oncology Program, approximately one-third of the medically eligible patients were entered on studies, with approximately half of the reasons for nonentry being "physician decision" (Hunter et al., 1987). A report on a more recent survey of physician members of ECOG noted that "... 82% of the respondents were reluctant to relinquish individualized decision-making control in favor of randomization and adherence to a protocol; this was the key message of the interviews" (Taylor et al., 1994). A survey of physician members of the Illinois Cancer Center also identified various issues relating to the physician–patient relationship as negative aspects of clinical trials (Benson et al., 1991).

Another line of evidence concerning potential ethical problems with the standard RCT design comes from asking physicians if they would allow themselves to be randomized in specific RCTs. It has been suggested that a trial is not ethical if the clinicians would not allow a member of their own family to enter it (Atkins, 1966). Evidence from oncologists suggests that only about one-third of the time would they enter themselves in trials for which they were eligible (Mackillop, Ward and O'Sullivan, 1986; Moore, O'Sullivan and Tannock, 1988, 1990).

Feasibility and generalizability. A potential problem with the standard RCT design for some studies is that the clinicians may have strong treatment preferences for most of their patients. The sets Π_j in (3.1) may therefore be small or empty, making the trial infeasible. Controversial treatments can lead to strong preferences. For example, a discussion in this journal (Ware, 1989) concerning a randomized trial of extracorporeal membrane oxygenation for the treatment of persistent pulmonary hypertension of the newborn (O'Rourke et al., 1989) had some authors stating that the trial should never have taken place (Berry, 1989; Royall, 1989) and others stating that the trial was stopped too early (Begg, 1989); see also Royall (1991). Another example is given by a randomized trial of adenine arabinoside for the treatment of herpes simplex encephalitis (Whitley et al., 1977), where one author suggested the trial should never have taken place (McCartney, 1978) while another suggested it was stopped too early (Tager, 1977).

In a standard RCT, the eligibility criteria characterize the patients who the participating clinicians are willing to randomize. Provided all medically eligible patients are randomized, the eligibility criteria also define the population to which the results of the trial generalize. There can be controversy about this population when only a fraction of medically eligible patients is randomized. Such was the case with the Extracranial–Intracranial Arterial Bypass Study (EC/IC Bypass Study Group, 1985), in which more eligible patients underwent the trial surgery outside of the trial than within it (Relman, 1987; Sundt, 1987; Goldring, Zervas and Langfitt, 1987; Barnett et al., 1987). Another potential example is given by the recently begun trial in the United States for prostate cancer comparing radical prostatectomy versus watchful waiting (Moon, Brawer and Wilt, 1995). It is expected that only a small fraction of medically eligible patients will be randomized, partly because of the strengths of the clinicians's preferences for one or the other of the treatments (Oncology Bulletin, 1994; Kaplan, 1995). A third example is given by an ongoing randomized trial implantable cardioverter-defibrillators versus antiarrhythmic drugs to prevent deaths in patients with life-threatening ventricular arrhythmias (AVID Investigators, 1995), where some authors have questioned whether certain classes of patients should be eligible for this trial (Fogoros, 1994; Josephson and Nisam, 1996). One approach to the generalizability question is to record the outcomes of all eligible patients, whether they are randomized or not. A large effort of this sort was conducted by the German Breast Cancer Group between 1983 and 1989, but the investigators do not recommend this approach for routine use (Schmoor, Olschewski and Schumacher, 1996).

Besides questions of generalizability due to the fact that only a subset of patients is randomized, there can also be the question that the participating clinicians and hospitals are not representative of those available outside of the trial. For example, there was a controversy (Proudfit, 1978; Chalmers et al., 1978) over the quality of the surgeries in a randomized trial involving coronary bypass surgery (Murphy et al., 1977). A restricted set of clinicians and/or hospitals participating in a trial can lead to generalizability questions.

Design and monitoring. Clinician preferences typically enter into the sample size calculation for a planned RCT when determining the alternative hypothesis of interest. For example, a clinician might

feel that unless a new, more toxic and expensive drug improved survival by 20%, he would prefer the standard drug for his patients. During trial monitoring, if accruing information concerning the observed treatment difference were made available to the clinicians, their treatment preferences might become strong enough for them to be unwilling to continue to randomize patients. For this reason and others, accruing relative efficacy data are typically not made available to the participating clinicians, but instead are given confidentially to a data monitoring committee. This committee evaluates the accruing data using statistical guidelines to decide if the trial should be stopped.

Prerandomization. In a typical RCT, patients are asked to participate in the trial before their treatment assignment is chosen randomly. Zelen (1979, 1982) developed prerandomization designs in which patients are randomized to treatment before they are asked to participate; experience with these designs is reviewed in Zelen (1990). By requiring the clinician to discuss only a single treatment with the patient, these designs ease the informed consent process and in theory increase participation by clinicians and patients. However, the estimation of the treatment effect must take into account the fact that some patients will refuse their randomly assigned treatment (Ellenberg, 1984; Baker, 1997). We note that prerandomization does not address the issue of clinicians treating against their clinical preferences.

3.2 Eligibility Using Clinician Uncertainty

One possible solution to the problem of defining eligibility criteria that will be acceptable to all clinicians participating in a trial is to allow each clinician to randomize a patient whenever he is uncertain which treatment is better. Examples of this approach include the ISIS trials evaluating the effects of various treatments on survival after the onset of suspected acute myocardial infarction (ISIS-1 Collaborative Group, 1986; ISIS-2 Collaborative Group, 1988; ISIS-3 Collaborative Group, 1992; ISIS-4 Collaborative Group, 1995). These trials have informal eligibility criteria that include clinician uncertainty; for example, "The fundamental criterion for entry was that the responsible physician was *uncertain* whether, for a particular patient, treatment with streptokinase or with aspirin was indicated" (ISIS-2 Collaborative Group, 1988, page 350). The first four ISIS trials randomized over 132,563 patients at more than one thousand hospitals. In the cancer area, the ongoing AXIS trial is evaluating adjuvant chemotherapy and radiotherapy for colorectal cancer using clinician uncertainty for eligibility: "... eligibility is defined not by the protocol but by the clinician's own judgement..." (Gray, James, Mossman and Stenning, 1991, page 844). As of 1994, 2,450 patients have been randomized in this trial by over 200 clinicians, reaching half the targeted accrual of 4,000 patients (AXIS Steering Group, 1994). A final example is given by the Icon trials evaluating chemotherapy for ovarian cancer (Ghersi et al., 1992). This example is interesting in that if the clinician is uncertain whether or not to administer chemotherapy immediately, the patient is randomized between immediate versus delayed chemotherapy (Icon-1), otherwise the patient is randomized between two different chemotherapy regimens (Icon-2). As of 1996, Icon-2 closed to accrual with 1,526 patients randomized, while Icon-1 is still open to patient entry (Parmar, 1996).

This type of trial design avoids the problem of defining universally acceptable eligibility criteria. Because of this, and its simplicity, it should lead to greater clinician participation than a standard RCT design; it has also been suggested as a way to encourage patients and clinicians to participate in trials involving AIDS (Byar et al., 1990). There is the question of the patient population to which the trial results will generalize; we return to this point in the Discussion (Section 5).

3.3 Bayesian Methods for RCTs

Bayesian methods utilize a prior distribution on the true treatment difference to help with the design, monitoring and analysis of an RCT (Cornfield, 1966; Spiegelhalter, Freedman and Parmar, 1994; Berry, 1993). Types of prior distributions include those derived from previous trials, reference ("noninformative") priors, skeptical or enthusiastic priors, priors with a mass at the null hypothesis and clinician priors. We focus attention here on clinician priors, thinking of these as a more detailed specification of clinician preferences. Methods of elicitation of prior distributions from clinicians are discussed by Freedman and Spiegelhalter (1983), Chaloner, Church, Louis and Matts (1993) and Kadane and Wolfson (1996). Methods of combining prior distributions are reviewed by Genest and Zidek (1986).

At the design stage, clinician priors can be used for sample size calculations. The published examples of this technique have involved retrospective calculations done on trials already designed with classical frequentist methods (Spiegelhalter and

Freedman, 1986; Spiegelhalter, Freedman and Parmar, 1993, 1994; Parmar, Spiegelhalter, Freedman and CHART Steering Committee, 1994). Bayesian methods can also be used to adapt the proportion of patients randomized to each treatment ("play the winner rules"; Zelen, 1969), but reference priors (implying an equal chance of either treatment for the first patient) rather than clinician priors have been used or suggested for this application (Berry and Eick, 1995; Bartlett et al., 1985). For trial monitoring, the decision to stop a trial early can be based on the posterior distribution of the treatment difference that is continually updated by the accruing trial results. Published examples of trial monitoring using clinician priors have again been performed retrospectively (Spiegelhalter, Freedman and Parmar, 1993; Carlin et al., 1993; Berry, Wolff and Sack, 1994). Recent suggestions concerning this type of monitoring have not used clinician priors as the basis for the posterior distribution, but have instead used skeptical priors for stopping early for a large treatment difference, and enthusiastic priors for stopping early because of the lack of a treatment difference (Spiegelhalter, Freedman and Parmar, 1994). For Bayesian analysis and reporting of trial results, the use of priors other than clinician priors is most common (Pocock and Spiegelhalter, 1992; Hughes, 1993; Spiegelhalter, Freedman and Parmar, 1994). This is especially true for analyses more complex than simple treatment comparisons, for example, the analysis of crossover trials (Racine, Grieve, Fluhler and Smith, 1986) or subset analysis (Dixon and Simon, 1991).

In summary, the use of clinician priors in Bayesian methods for RCTs would appear to be somewhat limited. Whether or not clinician priors are used, the effects of clinician preferences on a Bayesian RCT are similar to those discussed previously for standard RCTs; that is, the same ethical and generalizability issues are raised. An exception to this statement is the Bayesian trial design of Kadane which will be discussed next.

3.4 A Trial Design Proposed by Kadane

Kadane and Sedransk (1980) and Kadane (1986) proposed a Bayesian trial design in which prior distributions as a function of the treatment and a set of prognostic covariates (defining prognostic groups) are elicited from each of a group of experts. These distributions are updated as the trial results accrue. As long as a treatment is best in terms of the posterior probability distribution of at least one of the experts for a given prognostic group, that treatment can be assigned to the next patient in that prognostic group. If more than one treatment is acceptable by these criteria, the patient can be randomized between the acceptable treatments, or some other optimization scheme can be used to assign the patient to one of the acceptable treatments.

The raison d'être of the Kadane design is to be able to modify who is eligible for the different treatments based on accruing trial data. The Bayesian monitoring substitutes for the deliberations of a data monitoring committee used in a standard RCT. The Bayesian monitoring has the advantage that the updating of the posterior distributions can be done after each patient's outcome data have become available, whereas a data monitoring committee would only be able to meet periodically. It is important to note that the prior probability distributions of the experts may not resemble the prior distribution of the clinician who is to treat the next patient. In theory, therefore, strong and opposing prior preferences from two of the experts could lead to a trial continuing to randomize patients past the time when a clinician (or data monitoring committee) might view the evidence in favor of one of the treatments as being overwhelming. The Kadane design also permits stopping the randomization for some of the patient prognostic groups as the trial proceeds, without stopping the whole trial. This raises the issue of the generalizability of the trial results; we return to this issue in Section 5.

Experience with the Kadane trial design. A trial was conducted comparing the effect of two drugs in lowering blood pressure after cardiopulmonary bypass surgery for those patients developing intraoperative hypertension; see Kadane (1996, Chapters 5–13) for full details. Five experts had their prior probability distributions for the outcome elicited for each drug for each of 16 prognostic groups. Forty-nine of 71 eligible patients consented to be in the study. Of the 49, 30 developed intraoperative hypertension and were treated with 1 of the 2 drugs. At the close of the study, randomization would not have been allowed in 5 of the 16 prognostic groups because the posterior distributions of the 5 experts favored the same drug.

3.5 A Trial Randomizing Patients to Clinicians with Different Specialties

It was noticed in the late 1970s that stage II breast cancer patients referred to the clinic at the Hamilton Regional Cancer Center who first saw a radiation oncologist at the clinic tended to be treated with radiation therapy, while those who first saw a

medical oncologist at the clinic tended to be treated with chemotherapy. At the time, both types of treatment were generally accepted by the oncological community. The specialty of the clinician first seeing the patient (radiation oncologist or medical oncologist) was determined by whichever clinician was assigned to the clinic that day. A trial was performed in which patients with stage II breast cancer over the age of 50 who were referred to the clinic were randomly assigned to be seen first by a radiation oncologist or by a medical oncologist. The idea was to use the results of this trial to perform an analysis of the effectiveness of radiation therapy versus chemotherapy for this patient population. Over 100 patients were randomized, but the results of the trial were not interpretable in large part because the radiation oncologists sent many of their patients (presumably those with poorer prognoses) to the medical oncologists to be treated with chemotherapy (Hryniuk, 1996).

Although this trial design does not estimate a causal treatment difference, it does estimate (2.1), which seems appropriate. The design neatly avoids the problem of clinicians treating against their preferences. Besides questions of feasibility (see Section 5), note that patients crossing over to the other treatment will need to be accounted for in the analysis (Baker, 1997). However, if one could define a subset of patients, based on pretreatment criteria, who do not cross over, then the analysis could be restricted to this subset, resulting in a more powerful analysis.

3.6 RCT with Clinician-Preferred Treatment

To avoid the potential problems with standard RCTs discussed earlier, we proposed the alternative design displayed in Figure 1 which we call the RCT with clinician-preferred treatment (Korn and Baumrind, 1991). Patients are first screened to see if they are eligible for the trial. The eligibility screening should be based on criteria using objective patient measurements and not on the feelings of the screener about appropriate therapy. After obtaining the informed consent of the patient to

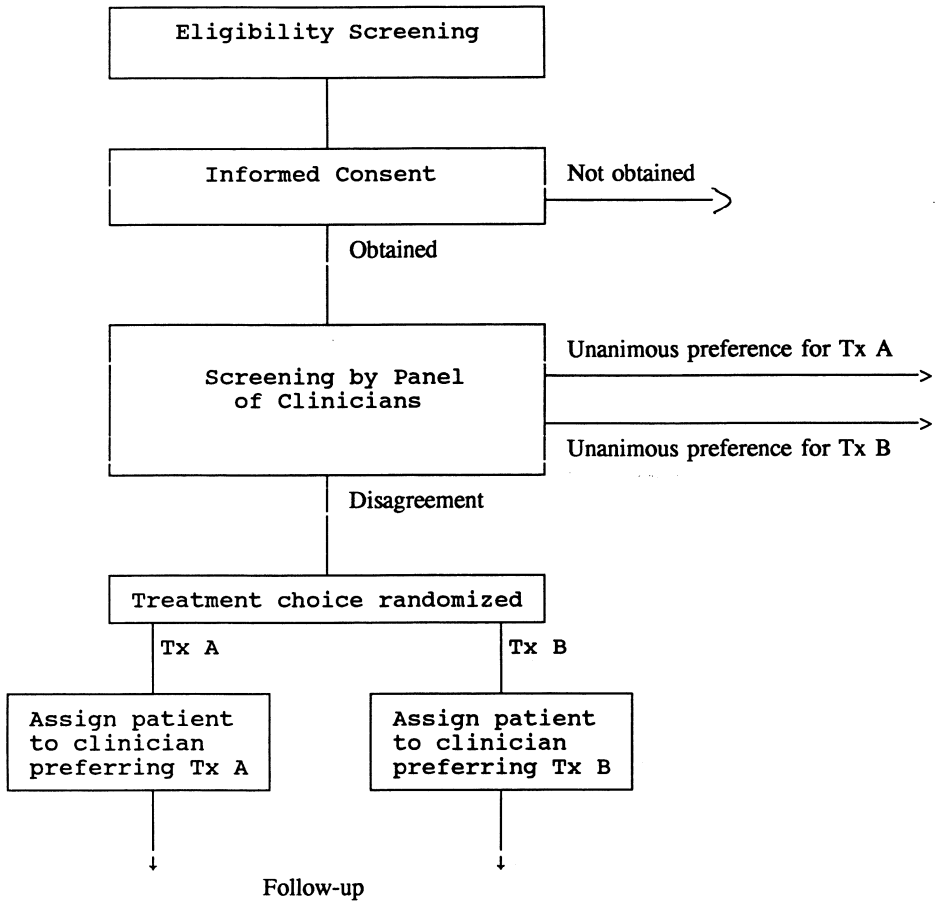

FIG. 1. *Design of randomized clinical trial with clinician-preferred treatment.*

participate in the trial, the patient and/or his records are reviewed independently by J clinicians who would be able to treat him. Each clinician independently states his treatment preference (A or B) for that patient. If the clinicians disagree concerning the preferred treatment, the patient is randomized to treatment A or B.

A patient randomized to treatment A (B) is assigned to be treated by one of the clinicians preferring A (B). Note that a clinician assigned a patient will be treating each patient with the treatment he preferred *for that patient*, so that the problem mentioned in Section 3.5 of patients being referred to other treatments does not arise. Typically, there will be more than one clinician who preferred the treatment to which the patient was randomized. In this case, the patient is assigned to one of these clinicians based on a designated probability distribution: for a patient with clinician-preference vector C, let $\lambda_j^T(C)$ be the designated assignment probability that clinician j will be assigned this patient if he is randomized to treatment T. For example, if $J = 5$ and $C = (ABBAA)$, then one might set $(\lambda_1^A(C), \ldots, \lambda_5^A(C)) = (1/3, 0, 0, 1/3, 1/3)$ and $(\lambda_1^B(C), \ldots, \lambda_5^B(C)) = (0, 1/2, 1/2, 0, 0)$. Other choices of the λ's will be useful in what follows, but always

$$\sum_j \lambda_j^A(C) = \sum_j \lambda_j^B(C) = 1 \quad \text{for each } C. \tag{3.2}$$

The mean treatment difference in an RCT with clinician-preferred treatment estimates

$$E\left(\bar{Y}_A - \bar{Y}_B\right) = \frac{1}{n_D} \sum_{u \in \Pi_D} \left\{ \sum_j \lambda_j^A(C(u)) \mu_u^{(A,j)} - \sum_j \lambda_j^B(C(u)) \mu_u^{(B,j)} \right\},$$

where Π_D is the (assumed nonrandom) subset of eligible patients for whom there was a disagreement about the preferred treatment, and n_D is the size of this subset. The expectation (E) is over the random assignment of the treatment, A versus B, and the random assignment of the clinician using the appropriate λ distribution. Defining

$$\mu_C^{(T,j)} = \frac{1}{\#\{u \mid C(u) = C\}} \sum_{\{u \mid C(u) = C\}} \mu_u^{(T,j)}, \qquad T = A \text{ or } B,$$

we have

$$E\left(\bar{Y}_A - \bar{Y}_B\right) = \sum_{C \in D} P(C \mid D) \cdot \left\{ \sum_j \lambda_j^A(C) \mu_C^{(A,j)} - \sum_j \lambda_j^B(C) \mu_C^{(B,j)} \right\}, \tag{3.3}$$

where D is the set of clinician-preference vectors with disagreement, and $P(C \mid D)$ is the conditional probability of C given $C \in D$. Although $\mu_C^{(A,j)} - \mu_C^{(B,j)}$ is a causal treatment difference, the right-hand side of (3.3) is not in general. To understand this better, we decompose $\mu_C^{(T,j)}$ in terms of a linear model:

$$\mu_C^{(T,j)} = \mu_C^{(T)} + \beta_j + (\mu\beta)_{Cj}^{(T)}. \tag{3.4}$$

If we assume the constraints that $\sum_j \beta_j = 0$ and $\sum_j (\mu\beta)_{Cj}^{(T)} = 0$, then

$$\mu_C^{(A)} - \mu_C^{(B)} = \frac{1}{J} \sum_j (\mu_C^{(A,j)} - \mu_C^{(B,j)})$$

is a causal treatment difference. The β_j in (3.4) can be thought of as the main effects for clinician skill and the $(\mu\beta)_{Cj}^{(T)}$ can be thought of as interactions of clinician skill with clinician preferences and treatment.

Using (3.4), (3.3) can be written as the sum of three terms:

$$\begin{aligned} E\left(\bar{Y}_A - \bar{Y}_B\right) &= \sum_{C \in D} P(C \mid D)\left[\mu_C^{(A)} - \mu_C^{(B)}\right] \\ &\quad + \sum_j \beta_j \sum_{C \in D} P(C \mid D)\left[\lambda_j^A(C) - \lambda_j^B(C)\right] \\ &\quad + \sum_{C \in D} P(C \mid D)\left[\sum_j \lambda_j^A(C)(\mu\beta)_{Cj}^{(A)} - \sum_j \lambda_j^B(C)(\mu\beta)_{Cj}^{(B)}\right]. \end{aligned} \tag{3.5}$$

The first term, derived using (3.2), is a causal treatment difference. It involves the main effects for treatment and clinician preference, and their interaction. The second term involves the β_j, and the third term involves the $(\mu\beta)_{Cj}^{(T)}$. In applications in which the skill of the clinicians is not expected to influence outcome, the β_j and the $(\mu\beta)_{Cj}^{(T)}$ would be zero so that the mean treatment difference would estimate a causal treatment difference.

In applications in which the skill of the clinicians is expected to be important, the second term in (3.5)

can sometimes be eliminated by a judicious choice of the λ's: suppose we can choose the λ's so that it is expected that each clinician will treat half his patients with A and half with B; that is, we can choose the λ's so that, for each j,

$$\sum_{C \in D} P(C \mid D)\left[\lambda_j^A(C) - \lambda_j^B(C)\right] = 0. \tag{3.6}$$

If (3.6) is satisfied, then we will say that the design is balanced. It is not possible always to balance the design, for example, if a clinician always prefers treatment A; we return to this issue below. When (3.6) is satisfied, however, the main effects for clinician skill are eliminated from (3.5). (An alternative strategy would involve assigning the patients to eligible clinicians at random, but using an analysis stratified on clinician. Although this stratified analysis eliminates the main clinician-skill effects, it is undesirable because clinician-preference effects would confound the stratified treatment difference.)

With a balanced design, we have

$$E\left(\bar{Y}_A - \bar{Y}_B\right) = \sum_{C \in D} P(C \mid D)\left[\mu_C^{(A)} - \mu_C^{(B)}\right] + \sum_{C \in D} P(C \mid D)\left[\sum_j \left\{\lambda_j^A(C)(\mu\beta)_{Cj}^{(A)} - \lambda_j^B(C)(\mu\beta)_{Cj}^{(B)}\right\}\right].$$

Further, if the clinician-skill by clinician-preference interaction is zero, we have

$$E\left(\bar{Y}_A - \bar{Y}_B\right) = \sum_{C \in D} P(C \mid D)\left[\mu_C^{(A)} - \mu_C^{(B)}\right],$$

a causal treatment difference. If the interaction is not zero, then even under the null hypothesis the expected mean treatment difference is not zero; we return to this point in Section 5.

The number of clinician judges. The larger J is, the more likely that there will be a disagreement about the preferred treatment, leading to a larger potential sample size available for randomization. However, patient factors induce a correlation in the treatment preferences of the clinicians, so that one would expect diminishing returns in terms of the number of disagreements as one increases J. This and practical considerations suggest J should be between 2 and 5. Although it is convenient for evaluation to have J be the same for all patients, the J clinicians will in general not be the same from patient to patient for logistical reasons.

A possible objection to the new design when J is greater than 2 is that a patient could be randomized to a treatment even though the majority of the J clinicians preferred the other treatment. To be specific, suppose $J = 5$ and four clinicians prefer treatment A and one prefers B for a given patient. There is a 50% probability that this patient would be randomized to B. If these clinicians were treating about equal numbers of patients in their practices, then in the absence of the trial there would only be a 20% chance that this patient would be treated with B. If one feels that this increase in probability from 20% to 50% is potentially harming the patient based on aggregate preferences that treatment A is "better," then one might object to the trial design. A possible modification of the design would be to allow only 2:3 or 3:2 disagreements to lead to randomization. Alternatively, one could use $J = 2$. The effect of either of these strategies would be to decrease the sample size of randomized patients.

Balancing the design. As mentioned previously, it is advantageous if one can define the assignment probabilities (λ's) in such a way that each clinician treats, on average, half his patients with A and half with B. The ability to make these definitions depends on the probabilities of observing the different clinician preference vectors C. For example, if $P(C) = 0$ whenever $C_1 = A$ so that clinician 1 never prefers treatment A, then the design cannot be balanced. If the $P(C \mid D)$ are known for all C, one can determine whether it is possible to balance the design, and define a set of balancing assignment probabilities. In practice, the $P(C \mid D)$ will not be known and may be hard to estimate. In this case, one can use adaptive allocation of the patients to clinicians to attempt to balance the design. In the present setting, the treatment choice is determined before the clinician assignment, so that the usual techniques (Efron, 1971; Pocock and Simon, 1975) need to be modified appropriately. There is, of course, no guarantee that adaptive allocation will lead to a balanced design, for example, if a clinician never prefers one of the treatments.

An objection that has been raised to the idea of balancing the design is as follows. Suppose clinician 1 prefers treatment A much more often than B, with the other clinicians preferring the treatments each about 50% of the time. For the infrequent patient for whom clinician 1 prefers treatment B, balancing the design will force that patient to be assigned to clinician 1 with high probability if he is randomized to treatment B. A consequence of the

balancing is that this patient would have a higher chance of being treated by clinician 1 with treatment B than if there was no trial being performed at all. If one presupposes that it is potentially harmful to have a patient treated by a clinician with a treatment that he infrequently prefers, then balancing the design would seem to be inappropriate. However, a principal investigator involved with the new trial design should believe that the clinicians participating in the trial are expert enough in both treatments so that he is comfortable with patients being treated by any of the clinicians using their preferred treatment. Thus, we believe it is not inappropriate to balance the design.

Eligibility screening and generalizability. Generalizability is restricted beyond the eligibility criteria by the requirement that the panel of clinicians disagree about the preferred treatment. At first glance, it would appear that a community clinician trying to decide whether the results of a completed trial are applicable to his patient must ensure that the panel from the trial would have disagreed on the preferred treatment for that patient. We attempt to avoid this problem by including as part of the eligibility criteria an objective screening to eliminate patients for whom there would likely be unanimous panel agreement about the preferred treatment. To the extent that the objective screening works, a community clinician need only check that his patient satisfies the eligibility criteria for the trial results to apply. The exact nature of this objective screening may depend on the composition of the panel. For example, fewer patients would need to be eliminated using a panel with diverse opinions concerning the preferred treatment than using a panel of similar-thinking clinicians. A disadvantage of using this type of objective screening is that it reduces the size of the population available for randomization.

The level of concern about the generalizability of trial results should depend on the likelihood that treatment differences are the same for those individuals in the trial as for those not randomized; we return to this point in Section 5. Along with the usual comparisons of prognostic variables, the RCT with clinician-preferred treatment offers another possible approach: one can compare the observed treatment differences for subsets of patients defined by different proportions of clinicians in the panel who preferred one of the treatments. For example, with five clinicians, suppose the observed treatment differences were the same when 1, 2, 3 or 4 clinicians preferred treatment A. This would provide some evidence (although not proof!) that the results of the trial generalize to the nonrandomized patients in which 0 or 5 of the clinicians preferred A.

Using covariates. The role of patient covariates in the RCT with clinician-preferred treatment is similar to their role in a standard RCT, the reduction of residual variation. Marginal control of one or two important covariates can be obtained with the same adaptive allocation that is used to balance the design for treating clinician. Alternatively, one can use analysis of covariance to estimate the treatment difference, rather than just the difference in means $\overline{Y}_A - \overline{Y}_B$.

Experience with an RCT with clinician-preferred treatment. There is much interest among orthodontists concerning the best way to treat patients with crowding and irregularities of their teeth and jaws. One can think of three groups of patients: one group for which experienced clinicians may reasonably disagree as to whether the preferred approach should involve tooth extraction or not; a second group for which almost all clinicians would use extraction; and a third group for which almost all clinicians would use nonextraction. To perform a standard RCT, it would be required to develop eligibility criteria that identified the set of patients for whom the preferred treatment was ambiguous. This may be difficult. Furthermore, even if we could identify such a set of patients using eligibility criteria, we believe that it would be inappropriate to have an orthodontist treat a patient with extraction when he preferred nonextraction for that patient, and vice versa. Another possible approach would be to use the clinician uncertainty as a criterion for eligibility as described in section 3.2, in which case an orthodontist would randomize a patient provided that he was uncertain as to the best treatment. However, we would expect that for fewer than 5% of patients would an orthodontist be sufficiently uncertain to be willing to randomize the treatment. This is true even though orthodontists would characterize a much larger percentage of their patients as "borderline" cases, that is, cases in which they would grant that other qualified clinicians might have different preferences.

Because of the difficulties in performing a standard RCT to address the extraction–nonextraction question, we conducted an RCT with clinician-preferred treatment at the University of California, San Francisco (UCSF). Five orthodontists selected out of a pool of 14 evaluated each patient. Our original plans were to enroll patients at the dental clinic at the University of the Pacific too, restricting

entry at both sites to adults. A sample size calculation suggested that 80 randomized adults would be sufficient to detect interesting differences in various orthodontic outcomes (e.g., a difference of 1 mm in a variable that has a standard deviation of 1.5 mm). We planned to randomize 90 adults (to allow for dropouts) over a period of up to 2.5 years. Because of reductions in grant funding, we decided to restrict the patient entry to the one site, but to open the trial to adolescents. Although we had initially planned to have objective screening as part of the eligibility criteria in place, we instead decided to rely on the less satisfactory solution of developing such criteria after the trial in order to help define the generalizable population.

The accrual to the trial began in October 1989 and ended two years later with only 41 patients randomized. As this was the first attempt at an RCT with clinician-preferred treatment, an examination of what went wrong is in order (Table 1). Although we had expected that more than half the patients treated in the clinic would be eligible for inclusion in the trial, only 19% (252 of 1,321) actually met the inclusion criteria. Expectations about numbers of eligible patients are apparently no substitute for hard data. A secondary problem involves the informed consent process. Of the 252 subjects who satisfied all the criteria for inclusion, 82 declined to participate, primarily because they rejected the idea of tooth extraction on any basis. Our original proposal for design of the RCT with clinician-preferred treatment had the informed consent take place after the randomization (Korn and Baumrind, 1991). Before implementing the trial, we decided to obtain consent prior to the panel screening and randomization (Figure 1). An alternative strategy would have been to obtain one consent before the screening to allow for the evaluation of the patient records, and another after the randomization. The second consent could be obtained by the clinician who would be treating the patient if the patient agreed to participate in the trial, offering advantages similar to the use of prerandomization in a standard trial.

For the 148 patients whose records were evaluated by 5 orthodontists, there was disagreement on approximately one-third (Table 1). With no objective screening to eliminate clear-cut cases, this amount of disagreement is consistent with our a priori expectations. If we had only randomized patients when there was a 3:2 or 2:3 disagreement, we would have randomized one-half of the disagreement cases (26/51); see Table 2. To estimate what would have happened if we had used only two orthodontists to evaluate each case, we tabulated the pairwise disagreement for pairs of orthodontists who evaluated at least 30 patients in common (Table 3). There were 16 such pairs (involving 8 of the participating orthodontists), with the proportions of disagreement ranging from 9.4% to 27.5%, mean = 16.7%, SD = 5.2%. Alternatively, one could estimate the pairwise disagreement rate directly from Table 2 as $17.3\% = [4(15 + 10) + 6(15 + 11)]/[10(59 + 38 + 15 + 10 + 15 + 11)]$. These numbers suggest that if one were concerned about randomizing patients with 4:1 or 1:4 disagreements, then the strategies of randomizing only 2:3 and 3:2 disagreements with five clinicians would yield about the same number of cases as just using two clinicians for evaluation in this orthodontic example. The latter strategy appears preferable since it involves less work.

The trial ended before the halfway accrual mark at which we planned to begin our attempts to balance the design using adaptive allocation. We note

TABLE 1
Information on orthodontic RCT with clinician-preferred treatment performed at UCSF (accrual period, October 1989 to October 1991)

Patients	Number of patients
Seen at clinic	1321
Eligible	252
Agreeing to participate	170
Whose records were evaluated by five orthodontists	148
For whom there was a disagreement about the preferred treatment	50
Randomized	41

TABLE 2
Clinician preferences for the 148 cases evaluated by five orthodontists as part of the RCT with clinician-preferred treatment performed at UCSF

Number of orthodontists favoring extraction:nonextraction	Adults	Adolescents	Total
5:0	23	36	59
4:1	5	10	15
3:2	3	12	15
2:3	6	5	11
1:4	1	9	10
0:5	10	28	38
Total	48	100	148

TABLE 3
Pairwise disagreement concerning treatment preference (XTR = extraction, Non-XTR = nonextraction) among orthodontists who evaluated at least 30 patients in common as part of the RCT with clinician-preferred treatment performed at UCSF

Clinician I I =	Clinician J J =	# I prefers Non-XTR & J prefers XTR	# I prefers XTR & J prefers Non-XTR	Total # disagreements	Total cases evaluated	Percentage disagreements
3	7	1	2	3	32	9.4%
6	12	4	1	5	32	15.6%
8	10	2	1	3	32	9.4%
3	12	5	3	8	36	22.2%
5	12	3	5	8	39	20.5%
3	8	4	2	6	40	15.0%
5	7	2	5	7	40	17.5%
7	8	6	5	11	40	27.5%
3	11	3	2	5	42	11.9%
7	11	7	3	10	42	23.8%
7	12	4	4	8	43	18.6%
11	12	0	5	5	48	10.4%
5	8	1	6	7	49	14.3%
8	12	4	5	9	50	18.0%
5	11	4	4	8	53	15.1%
8	11	7	3	10	58	17.2%

that for the 13 orthodontists who evaluated at least 10 cases, the proportions of cases for whom they preferred the extraction therapy ranged from 45% to 78% (Baumrind, Korn, Boyd and Maxwell, 1996). Therefore, we believe there would have been a reasonable chance that we would have been able to balance the design had the trial continued.

4. OBSERVATIONAL STUDIES

4.1 Standard Observational Analyses

In an observational study, clinicians are treating patients according to their preferences, so the ethical concerns regarding clinician preferences discussed previously are eliminated. However, as is well known, the simple difference between mean outcomes of patients treated with A and patients treated with B in an observational analysis does not in general estimate a causal treatment difference (Byar et al., 1976). Potential biases include those due to (a) differences in the patient populations treated with A and B because of differences (i) in diagnostic criteria used to determine inclusion in the analysis or (ii) in patient prognostic characteristics, (b) differences in supportive care available for the two treatments, (c) differences in the evaluation of the outcome for the two treatments and (d) the differing skills of the clinicians choosing to use the different treatments. If the patients were treated with A at a different calendar-time period than those treated with B, the potential for some of these biases could be larger. We focus here on the potential biases relating to the clinician, that is, biases because of clinician-preference effects and clinician-skill effects discussed in Section 3.6.

One approach to observational analyses is to perform trials with historical controls in which the outcomes of patients treated in one time period with one therapy are compared to the outcomes of patients treated in a different time period with another therapy. If a large proportion of eligible patients and clinicians participate in such trials, then the potential biases due to clinician-skill effects and clinician-preference effects can be minimized. This strategy has been used successfully in some analyses involving childhood cancers where it is estimated that 94% of children under the age of 15 years diagnosed with cancer in the United States are seen at an institution that is a member of one of the two large U.S. cooperative cancer groups (Ross, Severson, Pollock and Robison, 1996).

More generally, the approach to the problem of clinician-preference effects in observational studies is to ensure that the comparison groups are well-matched on pre-treatment patient covariates x that capture the patient prognosis (Gehan and Freireich, 1974), or to stratify the analysis by such covariates. One can additionally stratify on clinician to eliminate clinician-skill effects if they are thought to be important. A stratified analysis would be based on a linear combination of $(\overline{Y}_x^{(A,j)} - \overline{Y}_x^{(B,j)})$, where $\overline{Y}_x^{(T,j)}$ is the mean outcome for those patients with covariate value x who were treated

with T by clinician j. Using the notation of Section 2, the individual stratified difference $(\bar{Y}_x^{(A,j)} - \bar{Y}_x^{(B,j)})$ estimates

$$E\left(\bar{Y}_x^{(A,j)} - \bar{Y}_x^{(B,j)}\right) = \frac{\sum_{\{u \in \Pi_j | x(u)=x\}} I\left[C_j(u) = A\right] \mu_u^{(A,j)}}{\sum_{\{u \in \Pi_j | x(u)=x\}} I\left[C_j(u) = A\right]} - \frac{\sum_{\{u \in \Pi_j | x(u)=x\}} I\left[C_j(u) = B\right] \mu_u^{(B,j)}}{\sum_{\{u \in \Pi_j | x(u)=x\}} I\left[C_j(u) = B\right]}, \tag{4.1}$$

where Π_j represents the set of patients treated by clinician j.

A successful stratification strategy for capturing patient prognoses can be expressed as the treatment assignment (by each clinician) being strongly ignorable given the covariates x (Rosenbaum and Rubin, 1983). In the present notation, this assumption is that, for each clinician j, $(\mu_u^{(A,j)}, \mu_u^{(B,j)})$ is conditionally independent of $C_j(u)$ given $x(u)$, considered as random variables over Π_j with $P(u) = 1/\#\{\Pi_j\}$. With strong ignorability, (4.1) is equal to

$$E\left(\bar{Y}_x^{(A,j)} - \bar{Y}_x^{(B,j)}\right) = \frac{1}{\#\left\{u \in \Pi_j \,|\, x(u) = x\right\}} \cdot \sum_{\{u \in \Pi_j | x(u)=x\}} (\mu_u^{(A,j)} - \mu_u^{(B,j)}),$$

a causal treatment difference. However, finding a set of covariates x that will make the treatment assignment strongly ignorable may not be possible, which would leave standard observational analyses potentially biased due to clinician-preference effects.

4.2 Using Clinician Preferences in a Retrospective Analysis

We consider using clinician preferences in a retrospective analysis in order to lessen the impact of clinician-preference effects. The fundamental idea is to use the same clinicians who treated the patients (in the past) for evaluation (in the present) of the pretreatment records of each other's patients. The stated treatment preferences of these clinicians are then used in the analysis. Some assumptions are needed to proceed. The first concerns the comparability of the patient populations seen by the different clinicians.

ASSUMPTION 1. For each j, $C \in D$, and for each $T = A$ or B,

$$\frac{\sum_{u \in \Pi_s} I[C(u) = C]\, \mu_u^{(T,j)}}{\sum_{u \in \Pi_s} I[C(u) = C]}$$

does not depend on s, where Π_s represents the population of patients treated by clinician s, $s = 1, \ldots, J$. We denote this ratio by $\mu_C^{(T,j)}$ (this is consistent with the notation of Section 3.6).

Assumption 1 states that, given the treatment, treating clinician and clinician-preference vector, the mean treatment outcome does not depend on which patient population was treated. Assumption 1 would be automatically satisfied if the patients had been assigned essentially at random to the different clinicians, for example, assigned according to which clinician was attending at a university clinic the day of the patient's first visit. We consider below relaxing this assumption by using covariates in the analysis. Note that $\mu_C^{(A,j)} - \mu_C^{(B,j)}$ is a causal treatment difference.

With Assumption 1, we can consider the same linear decomposition of $\mu_C^{(T,j)}$ as given in (3.4). Just as in the RCT with clinician-preferred treatment, the retrospective analysis of causal treatment differences will be confounded by the interaction of clinician skill with clinician preference and treatment. We initially assume that these interactions are zero (but return to this issue in Section 5).

ASSUMPTION 2. For each j and for each $C \in D$,

$$\mu_C^{(T,j)} = \mu_C^{(T)} + \beta_j$$

for some $\mu_C^{(T)}$ and β_j, $\Sigma_j \beta_j = 0$.

Assumption 2 is satisfied with $\beta_j \equiv 0$ in the special case when there are no clinician-skill effects, for example, as might be expected with some oral medications. Without Assumption 1, the β_j's in Assumption 2 could not be interpreted simply as clinician-skill effects, but instead would implicitly incorporate differences in the patient populations seen by the different clinicians. Note that $\mu_C^{(A)} - \mu_C^{(B)}$ is a causal treatment difference, since $\mu_C^{(A)} - \mu_C^{(B)} \equiv \mu_C^{(A,j)} - \mu_C^{(B,j)}$ for all j.

With Assumptions 1 and 2, we can estimate a causal treatment difference as follows. Let Y_u be the outcome of the uth patient, and let $T(u)$, $j(u)$ and $C(u)$ be the treatment, treating clinician and clinician-preference vector for that patient. Consider the following analysis of variance model for the Y's:

$$Y_u = \mu + \alpha_{T(u)} + \gamma_{j(u)} + \delta_{C(u)} + \text{error}, \tag{4.2}$$

with constraints $\alpha_A + \alpha_B = 0$, $\Sigma\gamma_j = 0$ and $\Sigma\delta_C = 0$. Note that $T(u) = C_{j(u)}(u)$, so that not all possible combinations of T, j and C are possible. (The letters denoting the effects in (4.2) are different from those used previously to distinguish between the model being used to fit the data (4.2) and the models assumed to be generating the data.)

PROPOSITION 1. *Let $\hat{\alpha}_T$ be the usual least-squares estimator of α_T calculated using model* (4.2) *fitted to the restricted data set defined by $u \in \Pi_D$. Assume that Assumptions* 1 *and* 2 *are satisfied. Under a strong null hypothesis of no causal treatment difference, $E(2\hat{\alpha}_A) \equiv E(\hat{\alpha}_A - \hat{\alpha}_B) = 0$. For $J = 2$ clinicians, $E(2\hat{\alpha}_A)$ is a causal treatment difference. For $J > 2$, $E(2\hat{\alpha}_A)$ is a weighted mean of the differences $\mu_C^{(A)} - \mu_C^{(B)}$.*

CONJECTURE 1. *For $J > 2$, the weights are nonnegative so that $E(2\hat{\alpha}_A)$ is a causal treatment difference.*

The proof of the proposition is straightforward. For $J = 2$, the estimator is given by

$$2\hat{\alpha}_A = \frac{\left(\overline{Y}_{AB}^{(A,1)} - \overline{Y}_{AB}^{(B,2)}\right) + \left(\overline{Y}_{BA}^{(A,2)} - \overline{Y}_{BA}^{(B,1)}\right)}{2}, \tag{4.3}$$

where $\overline{Y}_C^{(T,j)}$ is the mean outcome of the individuals with clinician preference vector C ($= AB$ or BA) who were treated by clinician j with treatment T. (The notation is redundant since $T = C_j$, but helpful for descriptive purposes.) Estimator (4.3) estimates $[(\mu_{AB}^{(A)} - \mu_{AB}^{(B)}) + (\mu_{BA}^{(A)} - \mu_{BA}^{(B)})]/2$ under Assumptions 1 and 2. For $J > 2$, one can numerically verify whether the conjecture is true for any particular set of sample sizes.

The suggested analysis using model (4.2) involves only data from patients for whom $C \in D$. If one strengthens Assumptions 1 and 2 to hold for all C, then one can include all patients in the analysis. We do not recommend doing this; the efficiency gains for estimating the treatment effect α_A would be expected to be minor (since the $C \notin D$ patients only provide direct information on the γ_j), and the broadening of Assumption 2 to $C \notin D$ is potentially major. Moving in the other direction, Assumptions 1 and 2 can be weakened if one further restricts the patients to be analyzed. For example, in practical applications one may pair clinicians (and their patient populations) and estimate causal treatment differences separately for each pair. An overall estimate of the causal treatment difference can then be obtained by combining the pairwise estimates. Although a change in notation would be required since each patient's records would not be evaluated by every clinician, such a strategy would only require Assumptions 1 and 2 to hold "pairwise."

Use of covariates. Assumptions 1 and 2 can be made less constraining by including covariates in the analysis. For example, suppose Assumption 1 does not hold for two clinical practices because one practice treats a higher proportion of men. If one performs an analysis stratified on sex, then Assumption 1 will need only be satisfied separately for men and women. In general, including covariates in model (4.2) lessens the required assumptions for unbiased estimation of the causal treatment difference, or, stated differently, lessens the potential bias in estimating the difference. In addition, including covariates can reduce the residual variation and thereby improve the estimation. The disadvantage in including covariates is that the variability of the estimated treatment difference may be increased. We recommend including only very important covariates in the analysis, for example, those that are imbalanced across the clinical practices and are thought to interact strongly with the treatment effect.

Pilot study for orthodontic treatments. We are conducting a pilot study to see if we can approach the extraction–nonextraction issue discussed in Section 3.6 using the observational approach described above. The study is to evaluate the records of patients treated at the University of the Pacific Dental School for patients commencing treatment in 1986–1992. We limit the study to these years since there may have been secular changes in treatment strategy that would make treatment comparisons for patients treated before 1986 not comparable with those treated later. For patients beginning treatment after 1992, there would not be enough time for the outcomes which interest us to have occurred, although these patients could be included at a later date. As the first step of the pilot study, we examined the records of all patients commencing treatment in 1988–1991. We located 124 cases who had complete enough pretreatment and posttreatment records for our analyses and were treated by one of the six most active orthodontists (Table 4).

The second step of the pilot study will be to acquire the records for the appropriate cases who started treatment in 1986–1987 and 1992. When done, we expect to have approximately 220 useable

TABLE 4
Preliminary results from pilot study examining records of patients commencing treatment at the University of the Pacific dental school in 1988–1991 *treated by one of six orthodontists*

Supervising orthodontist	Number of cases with complete records: Patients treated with nonextraction	Patients treated with extraction	Total
A	12	14	26
B	14	12	26
C	10	12	22
D	11	9	20
E	6	12	18
F	8	4	12
Total	61	63	124

cases. Rather than have each clinician evaluate the pretreatment records from each of the other five clinicians, we plan to pair them, A with B, C with D and E with F, so that each clinician will be evaluating fewer cases. Each clinician will give his independent preference for how he would have treated the patient (extraction versus nonextraction). Included with the cases for evaluation will be some additional cases that the orthodontist treated himself. The inclusion of these self-treated cases allows a check on secular changes in preference patterns over calendar time. as well as an indication of whether there are factors other than clinician preference (e.g., patient preference) that determine the choice of treatment. Based on the analysis given in Section 3.6, we would expect disagreement in about 17% of the 220 cases. The six orthodontists have agreed to participate in this pilot study.

The aims of this pilot study are (1) to see if complete enough pretreatment records can be located for orthodontists to base preferences on, (2) to see whether the preferences of orthodontists reevaluating their own cases match with the way the patients were actually treated, (3) to count the number of disagreements among clinicians concerning treatment preference and (4) to perform an analysis of the outcomes on the disagreement cases as described earlier with an eye toward estimating variability for planning a future larger study. (The outcomes are all physical measurements, e.g., changes in linear distances.) This latter analysis will involve the evaluation of treatment outcome from the records of the patients, but only from the 35–40 disagreement cases. This evaluation will be done independently of the six orthodontists who treated the patients.

Comparison with RCT with clinician-preferred treatment. An observational analysis using clinician preferences can be performed retrospectively and use different patient populations seen at different single-clinician practices. An RCT with clinician-preferred treatment must be done prospectively, with at least two clinicians at each clinic. An observational study will require Assumption 1, which is not needed in an RCT because of the randomization between the treatments. However, this assumption might be satisfied by performing an observational study using data collected at a clinic in which multiple clinicians treat patients. If which patient sees which clinician is essentially random, Assumption 1 would be satisfied. If such an observational study were conducted prospectively, then the potential problem in a retrospective analysis of clinician preferences changing over time would also be eliminated.

With Assumption 2, both the observational and randomized studies estimate causal treatment differences, although potentially different causal treatment differences. One might expect that the observational analysis would be less efficient because of the necessity of including clinician-preference effects in the analysis. However, the inclusion of these effects might lessen the residual variation, making the relative merits less clear. In fact, one might wish to analyze data collected in an RCT with clinician-preferred treatment as if it were an observational study get similar reductions in residual variation.

In summary, the main use of randomization is to ensure Assumption 1 is satisfied. But if the same goal can be accomplished in a prospective or retrospective observational study, then logistical considerations may make an observational study the preferred approach.

4.3 An Instrumental-Variable Approach Using Clinician as the Instrumental Variable

Another approach to estimating causal effects for observational data, popular among economists, is to use instrumental variables. In this section we consider using "treating clinician" as an instrumental variable. We examine the required assumptions given by Angrist, Imbens and Rubin (1996) (denoted AIR) and compare them with those required by our approach above which uses clinician preferences explicitly. For simplicity, we consider the case of only two clinicians, clinician 1 and clinician 2. The instrumental variable (IV) estimator of the causal treatment difference is given by

$$\text{(4.4)}\quad \text{IV treatment effect estimator} = \frac{\overline{Y}^{(1)} - \overline{Y}^{(2)}}{\overline{A}^{(1)} - \overline{A}^{(2)}},$$

where $\overline{Y}^{(j)}$ is the mean outcome for patients treated by clinician j, and $\overline{A}^{(j)}$ is the proportion of patients treated with treatment A by clinician j, $j = 1, 2$.

The instrumental-variable approach requires that comparable patient populations be seen by the different clinicians, with random assignment of clinicians to patients one possibility (Assumption 2 in AIR). This is the same as our Assumption 1 (Section 4.2), and would apply in the same circumstances. The instrumental-variables approach requires that the average preference for treatment A (versus B) be different for the two clinicians (Assumption 4 in AIR). Given the denominator of (4.4), the larger the difference in average preference for treatment A, the more efficient one would expect this approach to be. This assumption may not be unreasonable in some clinical settings, although it is not required for our approach. The instrumental-variable approach requires no clinician-skill effects on outcome (Assumption 3 in AIR), whereas our approach controls for the main effects of clinician skill on outcome. To proceed, we assume a situation in which there are no clinician-skill effects.

Another assumption of the instrumental-variable approach is "monotonicity" (Assumption 5 in AIR), which states, in the present context, that if the clinicians disagree, they will disagree in only one direction. That is, if there are patients for whom clinician 1 prefers treatment A and clinician 2 prefers treatment B, then there can be no patients for whom clinician 1 prefers B and clinician 2 prefers A. One might take the monotonicity assumption as reasonable by hypothesizing a continuum of patient characteristics relating to the treatment decision, and that clinicians would divide the continuum differently in making their treatment decisions. Although this is usually an untestable assumption, we can check to see if it is satisfied in our orthodontic example because we have explicitly obtained clinician preferences in our RCT with clinician-preferred treatment. Recall that Table 3 displays the pairwise disagreements concerning treatment preference for the 16 clinician pairs who evaluated at least 30 patients in common. We see that, for all but one pair, the monotonicity assumption is not satisfied, so that the instrumental-variables approach cannot be used for this application.

If all the assumptions for the instrumental-variable estimator are satisfied, then (4.4) estimates (in our notation) $\mu_{AB}^{(A)} - \mu_{AB}^{(B)}$. This is the same estimand of the more efficient estimator that explicitly uses the treatment preferences, $\overline{Y}_{AB}^{(A,1)} - \overline{Y}_{AB}^{(B,2)}$. (With the monotonicity assumption satisfied, there would be no BA patients which would preclude the use of estimator (4.3).) The trade-off between the two approaches would be between this loss of efficiency and the practicality and costs of obtaining the clinician preferences.

5. DISCUSSION

With many of the alternative study designs discussed here, clinician preferences are involved in determining which patients are included in the analysis of the treatment difference. It may therefore be difficult to define objectively the analyzed population, raising the question of to what population the trial results will generalize. The same problem occurs in an RCT with a standard design when not all medically eligible patients are randomized or when the treatment effect for a subgroup of the patients analyzed may be different than for the whole group. The major concern is a qualitative interaction between the treatment and patient characteristics, in which treatment A is better than B for one patient subgroup, but worse than B for another (Byar 1985; Peto, 1995). It would seem best that this issue should be examined when considering applying one of the new designs, with the decision being made as to what population the study results will reasonably apply.

Unlike designs that involve randomizing patients, the RCT with clinician-preferred treatment (Section 3.6) and the observational analysis using clinician preferences (Section 4.2) do not estimate a causal treatment difference as defined in Section 2. This is because of the possibility of an interaction of clinician skill with clinician preference and treatment. We now examine this possibility in more detail. Table 5 contains a hypothetical example in which there are two clinicians and the mean outcomes only involve this kind of interaction: For patients for whom $C = (AB)$, the mean outcome is 11 if they are treated by clinician 1, and 9 if treated by clinician 2, regardless of which treatment (A or B) is used. For patients for whom $C = (BA)$, the

TABLE 5
Hypothetical example of an interaction of clinician skill with clinician preference and treatment under the null hypothesis of no causal treatment difference

Treatment preference of		Regardless of treatment (A or B), expected outcome if treated by	
Clinician 1	Clinician 2	Clinician 1	Clinician 2
A	B	11	9
B	A	9	11

mean outcome is 9 if they are treated by clinician 1, and 11 if treated by clinician 2, regardless of which treatment is used. Under these conditions, $E\overline{Y}_A = 11$ and $E\overline{Y}_B = 9$, although the causal treatment difference is zero. To see how such an interaction could arise, suppose the disease under study is a psychiatric disorder, and treatment A is psychoanalysis while treatment B is a drug therapy. Since psychoanalysis involves such a strong personal relationship of the clinician with the patient, if the clinician preferred it for a given patient, the outcome might be expected to be better than if he preferred to use drug therapy. This would lead to results similar to Table 5.

The possibility of such interactions does not discourage us from pursuing our designs that use clinician preferences for two reasons. The first is that, unlike main effects, it is hard to imagine an interaction of this type being large; therefore, the bias due to such interactions would be small. The second reason involves the appropriateness of the definition of a causal treatment difference discussed in Section 2, given the possibility of certain changes in clinician behavior after the completion of a study. Consider again the hypothetical example involving the comparison of psychoanalysis versus drug therapy. Suppose that a study using the RCT with clinician-preferred treatment was completed and showed that treatment A (psychoanalysis) was better. Clinicians who would have formerly preferred B for a patient might now prefer A. Additionally, these clinicians might modify their delivery of treatment A after discussing this type of patient with other clinicians who had preferred A all along. These two factors could result in the mean outcome being 11 for all future patients treated with A. Therefore, with this changed clinician behavior, treatment A is better than treatment B. One can therefore argue that, while the designs using clinician preferences are not estimating the causal treatment difference as defined in Section 2, they are sometimes estimating something more germane.

Another difference between the estimand of studies using the designs in Sections 3.6 and 4.2 and a causal treatment difference relates to the handling of placebo effects. Suppose that treatment A has a large placebo effect associated with it, whereas treatment B does not. Should the comparison of the treatments consider the placebo effect as part of treatment A or not? A double blind RCT with a standard design minimizes the placebo effect while the designs with clinician preferences include it.

We end by discussing the feasibility of the RCT trial designs that use clinician preferences. Trials with designs using clinician uncertainty as a major determinant of eligibility (Section 3.2) are probably easier to conduct than RCTs with a standard design, so feasibility is not an issue. The trial designs that involve randomizing patients to different clinicians (Sections 3.5 and 3.6) are applicable only in settings in which patients do not have a specific treating clinician before their entry into the trial. Such settings may be possible in university clinics and, with their larger numbers of patients and increasing popularity, in health maintenance organizations. Trial designs that involve elicitation of clinician preferences (Sections 3.4–3.6) can be a lot more work to implement than a standard RCT. Because of this, and the potential inferential difficulties previously discussed, only rarely would we recommend these designs and only when it was felt to be infeasible to conduct an RCT with a standard design. An alternative in these situations is to perform an observational study (Section 4.1), possibly using clinician preferences (Section 4.2).

ACKNOWLEDGMENTS

The authors thank the Editors and reviewers for their helpful comments.

REFERENCES

ANGRIST, J. D., IMBENS, G. W. and RUBIN, D. B. (1996). Identification of causal effects using instrumental variables (with discussion). *J. Amer. Statist. Assoc.* **91** 444–472.

ATKINS, H. (1966). Conduct of a controlled clinical trial. *British Medical Journal* **2** 377–379.

AVID INVESTIGATORS (1995). Antiarrhythmics versus implantable defibrillators (AVID)—rationale, design, and methods. *American Journal of Cardiology* **75** 470–475.

AXIS STEERING GROUP (1994). The AXIS colorectal cancer trial: randomization of over 2000 patients. *British Journal of Surgery* **81** 1672.

BAKER, S. G. (1997). Compliance, all-or-none. In *Encyclopedia of Statistical Sciences*, *Update*, **1**. Wiley, New York.

BARNETT, H. J. M., SACKETT, D., TAYLOR, D. W., HAYNES, B., PEERLESS, S. J., MEISSNER, I., HACHINSKI, V. and FOX, A. (1987). Are the results of the extracranial–intracranial arterial bypass study generalizable? *New England Journal of Medicine* **316** 820–824.

BARTLETT, R. H., ROLOFF, D. W., CORNELL, R. G., ANDREWS, A. F., DILLON, P. W. and ZWISCHENBERGER, J. B. (1985). Extracorporeal circulation in neonatal respiratory failure: a prospective randomized study. *Pediatrics* **76** 479–487.

BAUMRIND, S., KORN, E. L., BOYD, R. L. and MAXWELL, R. (1996). The decision to extract: part 1–inter-clinician agreement. *American Journal of Orthodontics and Dentofacial Orthopedics* **109** 297–309.

BEGG, C. B. (1989). Comment on "Investigating therapies of potentially great benefit: ECMO," by J. H. Ware. *Statist. Sci.* **4** 320–322.

BEGG, C. B., ZELEN, M., CARBONE, P. P., MCFADDEN, E. T., BRODOVSKY, H., ENGSTROM, P., HATFIELD, A., INGLE, J.,

SCHWARTZ, B. and STOLBACH, L. (1983). Cooperative groups and community hospitals: measurement of impact in the community hospitals. *Cancer* **52** 1760–1767.

BENSON, A. B., III, PREGLER, J. P., BEAN, J. A., RADEMAKER, A. W., ESHLER, B. and ANDERSON, K. (1991). Oncologists' reluctance to accrue patients onto clinical trials: an Illinois Cancer Center study. *Journal of Clinical Oncology* **9** 2067–2075.

BERRY, D. A. (1989). Comment: Ethics and ECMO, on "Investigating therapies of potentially great benefit: ECMO," by J. H. Ware. *Statist. Sci.* **4** 306–310.

BERRY, D. A. (1993). A case for Bayesianism in clinical trials. *Statistics in Medicine* **12** 1377–1393.

BERRY, D. A. and EICK, S. G. (1995). Adaptive assignment versus balanced randomization in clinical trials: a decision analysis. *Statistics in Medicine* **14** 231–246.

BERRY, D. A., WOLFF, M. C. and SACK, D. (1994). Decision making during a phase III randomized controlled trial. *Controlled Clinical Trials* **15** 360–378.

BYAR, D. P. (1985). Assessing apparent treatment–covariate interactions in randomized clinical trials. *Statistics in Medicine* **4** 255–263.

BYAR, D. P., SIMON, R. M., FRIEDEWALD, W. T., SCHLESSELMAN, J. J., DEMETS, D. L., ELLENBERG, J. H., GAIL, M. H. and WARE, J. H. (1976). Randomized clinical trials: perspectives on some recent ideas. *New England Journal of Medicine* **295** 74–80.

BYAR, D. P., SCHOENFELD, D. A. and GREEN, S. B., et al. (1990). Design considerations for AIDS trials. *New England Journal of Medicine* **323** 1343–1348.

CARLIN, B. P., CHALONER, K., CHURCH, T., LOUIS, T. A. and MATTS, J. P. (1993). Bayesian approaches for monitoring clinical trials with an application to toxoplasmic encephalitis prophylaxis. *The Statistician* **42** 355–367.

CHALMERS, T. C., SMITH, H., JR., AMBROZ, A., REITMAN, D. and SCHROEDER, B. J. (1978). In defense of the VA randomized control trial of coronary artery surgery. *Clinical Research* **26** 230–235.

CHALONER, K., CHURCH, T., LOUIS, T. A., and MATTS, J. P. (1993). Graphical elicitation of a prior distribution for a clinical trial. *The Statistician* **42** 341–353.

CLAYTON, D. G. (1982). Ethically optimised designs. *British Journal of Clinical Pharmacology* **13** 469–480.

CORNFIELD, J. (1966). A Bayesian test of some classical hypotheses—with applications to sequential clinical trials. *J. Amer. Statist. Assoc.* **61** 577–594.

DIXON, D. O. and SIMON, R. (1991). Bayesian subset analysis. *Biometrics* **47** 871–881.

EC / IC BYPASS STUDY GROUP (1985). Failure of extracranial–intracranial arterial bypass to reduce the risk of ischemic stroke. *New England Journal of Medicine* **313** 1191–1200.

EFRON, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika* **58** 403–417.

ELLENBERG, S. S. (1984). Randomization designs in comparative clinical trials. *New England Journal of Medicine* **310** 1404–1408.

FISHER, B., BAUER, M. and MARGOLESE, R. et al. (1985). Five-year results of a randomized clinical trial comparing total mastectomy and segmental mastectomy with and without radiation in the treatment of breast cancer. *New England Journal of Medicine* **312** 665–673.

FOGOROS, R. N. (1994). An AVID dissent. *PACE* **17** 1707–1711.

FREEDMAN, B. (1987). Equipoise and the ethics of clinical research. *New England Journal of Medicine* **317** 141–145.

FREEDMAN, L. S. and SPIEGELHALTER, D. J. (1983). The assessment of subjective opinion and its use in relation to stopping rules for clinical trials. *The Statistician* **32** 153–160.

GEHAN, E. A. and FREIREICH, E. J. (1974). Non-randomized controls in cancer clinical trials. *New England Journal of Medicine* **290** 198–203.

GENEST, C. and ZIDEK, J. V. (1986). Combining probability distributions: a critique and an annotated bibliography (with discussion). *Statist. Sci.* **1** 114–148.

GHERSI, D. PARMAR, M. K. B., STEWART, L. A., MARSONI, S. and WILLIAMS, C. J. (1992). Early ovarian cancer and the Icon Trials. *European Journal of Cancer* **28A** 1297.

GOLDRING, S., ZERVAS, N. and LANGFITT, T. (1987). The extracranial–intracranial arterial bypass study: a report of the committee appointed by the American Association of Neurological Surgeons to examine the study. *New England Journal of Medicine* **316** 817–820.

GRAY, R., JAMES, R., MOSSMAN, J. and STENNING, S. (1991). AXIS—a suitable case for treatment. *British Journal of Cancer* **63** 841–845.

HELLMAN, S. and HELLMAN, D. S. (1991). Of mice but not men, problems of the randomized clinical trial. *New England Journal of Medicine* **324** 1585–1589.

HOLLAND, P. W. (1986). Statistics and causal inference (with discussion). *J. Amer. Statist. Assoc.* **81** 945–970.

HRYNIUK, W. M. (1996). Personal communication.

HUGHES, M. D. (1993). Reporting Bayesian analyses of clinical trials. *Statistics in Medicine* **12** 1651–1663.

HUNTER, C. P., FRELICK, R. W., FELDMAN, A. R., BAVIER, A. R., DUNLAP, W. H., FORD, L., HENSON, D., MACFARLANE, D., SMART, C. R., YANCIK, R. and YATES, J. W. (1987). Selection factors in clinical trials: results for the Community Clinical Oncology Program Physician's Patient Log. *Cancer Treatment Reports* **71** 559–565.

ISIS-1 COLLABORATIVE GROUP (1986). Randomised trial of intravenous atenolol among 16027 cases of suspected acute myocardial infarction: ISIS-1. *The Lancet* **2**(8498) 57–66.

ISIS-2 COLLABORATIVE GROUP (1988). Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17187 cases of suspected acute myocardial infarction: ISIS-2. *The Lancet* **2**(8607) 349–360.

ISIS-3 COLLABORATIVE GROUP (1992). ISIS-3: a randomised comparison of streptokinase vs tissue plasminogen activator vs anistreplase and of aspirin plus heparin vs aspirin alone among 41299 cases of suspected acute myocardial infarction. *The Lancet* **339**(8796) 753–770.

ISIS-4 COLLABORATIVE GROUP (1995). ISIS-4: a randomised factorial trial assessing early oral captopril, oral mononitrate, and intravenous magnesium sulphate in 58050 patients with suspected acute myocardial infarction. *The Lancet* **345**(8951) 669–685.

JOSEPHSON, M. and NISAM, S. (1996). Prospective trials of implantable cardioverter defibrillators versus drugs: are they addressing the right question? *American Journal of Cardiology* **77** 859–863.

KADANE, J. B. (1986). Progress toward a more ethical method for clinical trials. *Journal of Medicine and Philosophy* **11** 385–404.

KADANE, J. B., ed. (1996). *Bayesian Methods and Ethics in a Clinical Trial Design*. Wiley, New York.

KADANE, J. B. and SEDRANSK, N. (1980). Toward a more ethical clinical trial. In *Bayesian Statistics: Proceedings of the First International Meeting Held in Valencia* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 329–338. Valencia Univ. Press.

KADANE, J. B. and WOLFSON, L. J. (1996). Priors for the design and analysis of clinical trials. In *Bayesian Biostatistics* (D. A. Berry and D. K. Stangl, eds.) 157–184. Dekker, New York.

KAPLAN, R. S. (1995). Personal communication.

KORN, E. L. and BAUMRIND, S. (1991). Randomised clinical trials with clinician-preferred treatment. *The Lancet* **337** 149–152.

MACKILLOP, W. J., WARD, G. K. and O'SULLIVAN, B. (1986). The use of expert surrogates to evaluate clinical trials in non-small cell lung cancer. *British Journal of Cancer* **54** 661–667.

MCCARTNEY, J. J. (1978). Randomized clinical trials in a fatal disease. *Hastings Center Report* **8** 5–7.

MOON, T. D., BRAWER, M. K. and WILT, T. J. (1995). Prostate Intervention Versus Observation Trial (PIVOT): a randomized trial comparing radical prostatectomy with palliative expectant management for treatment of clinically localized prostate cancer. *Journal of the National Cancer Institute Monographs* **19** 69–71.

MOORE, M. J., O'SULLIVAN, B. and TANNOCK, I. F. (1988). How expert physicians would wish to be treated if they had genitourinary cancer. *Journal of Clinical Oncology* **6** 1736–1745.

MOORE, M. J., O'SULLIVAN, B. and TANNOCK, I. F. (1990). Are treatment strategies of urologic oncologists influenced by the opinions of their colleagues? *British Journal of Cancer* **62** 988–991.

MURPHY, M. L., HULTGREN, H. N., DETRE, K., THOMSEN, J. and TAKARO, T. (1977). Treatment of chronic stable angina. *New England Journal of Medicine* **297** 621–627.

ONCOLOGY BULLETIN (1994). PIVOT: looking for answers to urgent questions. *Oncology Bulletin* December, 3–8.

O'ROURKE, P. P., CRONE, R. K., VACANTI, J. P., WARE, J. H., LILLIHEI, C. W., PARAD, R. B. and EPSTEIN, M. F. (1989). Extracorporeal membrane oxygenation and conventional medical therapy in neonates with persistent pulmonary hypertension of the newborn: a prospective randomized study. *Pediatrics* **84** 957–963.

PARMAR, M. K. B. (1996). Personal communication.

PARMAR, M. K. B., SPIEGELHALTER, D. J., FREEDMAN, L. S. and CHART STEERING COMMITTEE (1994). *Statistics in Medicine* **13** 1297–1312.

PETO, R. (1995). Clinical trials. In *Treatment of Cancer* (P. Price and K. Sikora, eds.) 1039–1043. Chapman & Hall, London.

POCOCK, S. J. and SIMON, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics* **31** 103–115.

POCOCK, S. J. and SPIEGELHALTER, D. J. (1992). Grampian region early anistreplase trial (letter). *British Medical Journal* **305** 1015.

PROUDFIT, W. L. (1978). Criticisms of the VA randomized study of coronary bypass surgery. *Clinical Research* **26** 236–240.

RACINE, A., GRIEVE, A. P., FLUHLER, H. and SMITH, A. F. M. (1986). Bayesian methods in practice: experiences in the pharmaceutical industry (with discussion). *J. Roy. Statist. Soc. Ser. C* **35** 93–150.

RELMAN, A. S. (1987). The extracranial–intracranial arterial bypass study: what have we learned? *New England Journal of Medicine* **316** 809–810.

ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55.

ROSS, J. A., SEVERSON, R. K., POLLOCK, B. H. and ROBISON, L. L. (1996). Childhood cancer in the United States. *Cancer* **77** 201–207.

ROYALL, R. (1989). Comment on "Investigating therapies of potentially great benefit: ECMO," by J. H. Ware. *Statist. Sci.* **4** 318–319.

ROYALL, R. (1991). Ethics and statistics in randomized clinical trials (with discussion). *Statist. Sci.* **6** 52–88.

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66** 688–701.

RUBIN, D. B. (1990a). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statist. Sci.* **5** 472–480.

RUBIN, D. B. (1990b). Formal modes of statistical inference for causal effects. *J. Statist. Plann. Inference* **35** 279–292.

SCHMOOR, C., OLSCHEWSKI, M. and SCHUMACHER, M. (1996). Randomized and non-randomized patients in clinical trials: experiences with comprehensive cohort studies. *Statistics in Medicine* **15** 263–271.

SHAW, L. W. and CHALMERS, T. C. (1970). Ethics in cooperative clinical trials. *Ann. New York Acad. Sci.* **169** 487–495.

SPIEGELHALTER, D. J. and FREEDMAN, L. S. (1986). A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Statistics in Medicine* **5** 1–13.

SPIEGELHALTER, D. J., FREEDMAN, L. S. and PARMAR, M. K. B. (1993). Applying Bayesian ideas in drug development and clinical trials. *Statistics in Medicine* **12** 1501–1511.

SPIEGELHALTER, D. J., FREEDMAN, L. S. and PARMAR, M. K. B. (1994). Bayesian approaches to randomized trials. *J. Roy. Statist. Soc. Ser. A* **157** 357–416.

SUNDT, T. M., JR. (1987). Was the international randomized trial of extracranial–intracranial arterial bypass representative of the population at risk? *New England Journal of Medicine* **316** 814–816.

TAGER, I. B. (1977). ARA-A for herpes encephalitis (letter). *New England Journal of Medicine* **297** 1289.

TAYLOR, K. M., MARGOLESE, R. G. and SOSKOLNE, C. L. (1984). Physicians' reasons for not entering eligible patients in a randomized clinical trial of surgery for breast cancer. *New England Journal of Medicine* **310** 1363–1367.

TAYLOR, K. M., FELDSTEIN, M. L., SKEEL, R. T., PANDYA, K. J., NG, P. and CARBONE, P. P. (1994). Fundamental dilemmas of the randomized clinical trial process: results of a survey of the 1,737 Eastern Cooperative Oncology Group investigators. *Journal of Clinical Oncology* **12** 1796–1805.

WARE, J. H. (1989). Investigating therapies of potentially great benefit: ECMO (with discussion). *Statist. Sci.* **4** 298–340.

WHITLEY, R. J., SOONG, S.-J., DOLIN, R., GALASSO, G. J., CHIEN, L. T. and ALFORD, C. A. (1977). Adenine arabinoside therapy of biopsy-proved herpes simplex encephalitis. *New England Journal of Medicine* **297** 289–294.

ZELEN, M. (1969). Play the winner rule and the controlled clinical trial. *J. Amer. Statist. Assoc.* **64** 131–146.

ZELEN, M. (1979). A new design for randomized clinical trials. *New England Journal of Medicine* **300** 1242–1245.

ZELEN, M. (1982). Strategy and alternate randomized designs in cancer clinical trials. *Cancer Treatment Reports* **66** 1095–1100.

ZELEN, M. (1990). Randomized consent designs for clinical trials: an update. *Statistics in Medicine* **9** 645–656.

Comment

Marvin Zelen

Korn and Baumrind discuss the issue of physician preference in the evaluation of the causal difference between therapies which are being studied in a clinical trial. Little attention has been given to this issue and their paper illustrates the difficulty of the problem. The randomized clinical trial (RCT) is widely regarded as the ideal way to evaluate therapies. However, these trials have inherent limitations which are not often discussed. These limitations affect the Korn and Baumrind discussion.

The group of patients participating in an RCT can be divided into two major groups. They are either a random sample of patients from a population conforming to eligibility requirements and positive patient consent or are considered as a "collection" of patients conforming to the same eligibility criteria and positive patient consent. A "collection" of patients is defined as the complement of a random sample. If the patients in the trial are a random sample, then the conclusions of the trial apply to the population from which they have been sampled. This is termed a global inference. Alternatively, the inference based on a collection of patients can only relate to conclusions for the actual patients entering the study. This inference is termed a local inference. In other words, the local inference can determine the best treatment for the finite population of patients who entered the study. The global inference is targeted at determining the best treatment for treating people with disease. Unfortunately nearly all RCT's are based on collections of patients. However, the scientific community interprets the reported outcome of an RCT as a global inference. The leap from a local to a global inference when the clinical trial is based on a patient collection can only be justified subjectively.

As a result, any modifications of the RCT which further restrict the randomization process serve to narrow the local inference. Furthermore the jump from a local inference to a global inference may be made more difficult. Is the panel of physicians evaluating the patients a random sample of physicians or a collection of physicians? Clearly the physicians evaluating the patients are likely to be a collection. Different physicians may have different evaluations resulting in a different collection of patients entering the RCT.

On a more technical level, if there do exist interactions between the treatment and the physician, then the main effect (causal effect) among treatments is nonestimable. The only estimable functions are the interaction effects. Such a situation would indicate that carrying out a clinical trial in the presence of such interactions would be premature. This could arise (say) in evaluating a new surgical procedure in which some of the participating surgeons have not yet become completely skilled in the new procedure.

One of the basic tenets of scientific experimentation is that the conclusions of an experiment can be duplicated by others. This may be difficult in a clinical trial setting due to ethical considerations. However, the multicenter clinical trial can be considered as carrying out the same clinical trial in different institutions, the execution being done in parallel. If the outcome of a therapy is physician dependent, then there are serious problems in extending the benefits of a therapy to other institutions. It is possible that there would be institution–treatment interactions. The presence of such interactions implies that the transfer of benefit does not apply to all institutions. In such instances, initiating a clinical trial may be premature until the dependence of outcome on the physician is eliminated.

Adopting new drug therapies is one example where the dependence of the physician on therapy outcome is minimal. There may be disagreements about the choice of therapy, but the same drug therapy adopted by different physicians will have the same outcome. It is of some interest that the bulk of RCT's are drug trials. There are very few RCT's comparing different surgical procedures, especially when one of the surgical procedures is newly developed. One reason for this is that, as the surgeon obtains more experience with a new procedure, the procedure is likely to undergo change. As a result, the actual procedure may be evolving as

Marvin Zelen is Professor, Harvard School of Public Health and Dana-Farber Cancer Institute, Boston, Massachusetts (*e-mail: zelen@jimmy.harvard.edu*).

the surgeons become more experienced. In such instances an RCT does not seem appropriate until the skill levels of participating surgeons are comparable. Acquiring skill levels for large numbers of surgeons may require so many patients that the benefit of the procedure is apparent and may preclude the initiation of a clinical trial.

In this discussant's opinion, the value of the Korn–Baumrind experimental design is to make it possible for physicians having different judgments on choice of therapy to participate in a joint trial. Some physicians may feel that ethically, if they favor a treatment, the patient cannot be registered in an RCT in which the favored treatment may not be available. However, there should be concern when colleagues may disagree with the choice of the favored treatment for the particular patient. When differences exist about the most appropriate therapy for a patient, only an RCT will be able to supply information which can settle the issue. However, it can only be carried out in the absence of a physician–treatment interaction.

Comment

David Freedman

1. PRELIMINARY REMARKS

The main topic in Korn and Baumrind is "clinician preference"; I will discuss this idea after making some preliminary remarks. I agree with the authors that randomized controlled experiments constitute the "gold standard" for clinical comparisons. Such experiments are often difficult to carry out. In some cases, experiments are impossible, for practical or ethical reasons. There is, therefore, a large role for epidemiology—even though causal inferences from observational studies are generally much less certain than causal inferences from experiments.

The clinical trials literature is uneven. There are many successes, and some failures. A few of the latter are cited by Korn and Baumrind, including the Extracranial–Intracranial Bypass Study, where the principal investigators apparently failed to ensure compliance with protocol, or to maintain adequate records, or to disclose the problems in their report (EC/IC Bypass Study Group, 1985). Disasters can be instructive, especially if they teach us what not to do. On the other hand, the EC/IC study is exceptional. Thousands of clinical trials have been published and reported in MEDLINE since 1985, and few of them have experienced difficulties of a similar order. In my view—Korn and Baumrind must agree—clinical trials are generally manageable, although some do go off the rails.

David Freedman is Professor, Department of Statistics, University of California, Berkeley, California 94720 (e-mail: freedman@stat.berkeley.edu).

The authors' skepticism about instrumental-variable approaches seems justified (Section 4.3). There is also some skepticism about Bayesian designs (Section 3.3). Here, Korn and Baumrind could have gone farther, because randomization does not fit at all coherently into the Bayesian framework. If subjects are exchangeable, randomization is expensive and irrelevant; if subjects are not exchangeable, randomization is counterproductive, because it is unlikely to yield the samples that are—in your prior opinion—the most informative.

Korn and Baumrind cite Rubin (1974, 1990b) for a theory of causal inference. That could be misleading. The "Rubin model" was developed by Neyman, used by Fisher, and discussed in the textbook literature, long before Rubin's 1974 paper. See, for instance, Dabrowska and Speed (1990) or Hodges and Lehmann (1964, Section 9.4). Rubin's (1990a) response to these facts may be of interest to some readers.

2. CLINICIAN PREFERENCE

One problem in analyzing observational studies or conducting experiments is "clinician preference." For example, if the surgeon decides who is to get surgical treatment and who is to get medical care, the two groups are unlikely to be comparable. On the other hand, in experimental settings, the physician may be unwilling to randomize some kinds of patients. A standard solution to this problem for a clinical trial is to specify in the protocol the target group of patients—for whom there is considerable uncertainty about the benefits of the different treatment plans.

Korn and Baumrind propose an extension of this idea, in effect, blocking patients on the basis of clinician preference and then randomizing within blocks. That idea may be helpful in some contexts, although it would significantly complicate the task of managing the trial. Indeed, the authors attempted to run a trial using their idea, and failed (Section 3.6). This is a strong signal for the rest of us.

The authors also propose to measure clinician preferences and use the results as covariates when analyzing observational data. A "pilot study" is being conducted as a demonstration (Section 4.2). Detailed comments should await publication of the data from the full study. There may be some exceptional circumstances in which the technique would be useful, but caution is in order. Real preferences may not be easy to elicit, as demonstrated by the Hamilton breast cancer study. When nobody is counting, clinicians recommend the treatments in which they specialize; when data collection starts, clinicians refer the weaker patients to other specialties (Section 3.5).

Even if the right covariates can be identified and measured, blocking may require enormous samples; recourse to modeling then suggests itself. But the adequacy of textbook models is in many cases open to serious doubt. That is, after all, why experiments are the gold standard, rather than observational studies.

Randomized controlled experiments are among the chief accomplishments of statistical science. In the clinical context, such experiments are hard to do—and even harder to explain. The medical community needs the help of statisticians in these respects among others. Korn and Baumrind are to be congratulated for focusing our attention on this important topic.

Comment

Deborah Ashby and Jayne E. Harrison

The randomized controlled trial (RCT) is increasingly being recognized as the gold standard for evaluating therapies. In some clinical areas, RCT's are widely used, but in others areas they are very slow to gain ground. The reasons for this are complex. For pharmaceutical products, regulatory procedures in the United States, Europe and elsewhere mean that new treatments can only rarely gain a license without conducting trials. It is also relatively straightforward to define what is meant by a course of treatment, the effects of which may be regarded as independent of the prescribing physician. By contrast, for example, surgical procedures are not regulated and may vary more subtly, not least, with the skill of the clinician. A particularly difficult area is that of orthodontic treatment. A recent review (Harrison, Ashby and Lennon, 1996) of two journals in the field, the *British Journal of Orthodontics* and the *European Journal of Orthodontics*, between 1989 and 1993 found that RCT's accounted for only 2.8% of clinical research papers. The rest used either nonrandomized controls or were uncontrolled. Against this background, innovative approaches to clinical research should be first welcomed, then carefully evaluated.

Our main thrust in this commentary will be the authors' work in orthodontics, but their reviews of other methods warrant a couple of remarks. Kadane's design presupposes that information is accruing quickly relative to patients. In orthodontics, we are looking at long-term outcomes, with the first meaningful data available perhaps two years after entry to the study, so the design is not feasible in this context. When reviewing Bayesian methods, Korn and Baumrind claim that published reports of trial monitoring are "retrospective." In fact Fayers,

Deborah Ashby is Professor of Medical Statistics, Department of Environmental and Preventive Medicine, Wolfson Institute of Preventive Medicine, St. Bartholomew's and the Royal London School of Medicine and Dentistry, Queen Mary and Westfield College, London, EC1M 6BQ United Kingdom (e-mail: d.ashby@mds.qmw.ac.uk). Jayne E. Harrison is Research Senior Registrar in Orthodontics, Liverpool University Dental Hospital, Pembroke Place, Liverpool L3 5PS United Kingdom. She is undertaking a Ph.D. in which she is investigating the application of RCT methods to clinical orthodontic research and is a member of the editorial team of the Cochrane Collaboration Oral Health Group.

Ashby and Parmar (1997) report on examples in the cancer field where Bayesian methods have formed the basis of the monitoring rules.

A challenge in any clinical research, but especially that requiring surgical techniques, is clinical preferences for patient treatment. This paper explores these issues both for randomized controlled trials and for observational studies. Two themes are entwined in the paper: the clinical challenge of designing and carrying out practically feasible research studies; and the statistical modeling of causal treatment differences using the data from such studies.

Korn and Baumrind define carefully what is meant by a causal treatment difference when the clinician carrying out the procedure is expected to have an influence on outcome, and then use this as a framework for trials where randomization involves allocation to a particular clinician using a particular procedure. They then consider observational studies, using their causal framework to give two assumptions under which it is possible to estimate a causal treatment difference. They motivate their model with reference to surgical trials, but, in passing, we note that even with pharmaceutical treatments clinician effects are plausible: through manner and monitoring there may be effects on compliance and hence on outcome.

Although the paper is more generally applicable, consideration of both the randomized controlled trial and the observational study designs include substantive examples from orthodontics, so we shall consider the degree to which they are likely to make genuine contributions to treatment decisions in this field.

1. RCTs WITH CLINICIAN-PREFERRED TREATMENT

The design of an RCT with clinician preferred treatments as presented has the advantage that patients are only entered if there is demonstrable clinical uncertainty about the best treatment. From the patient's point of view this design would appear "ethical," and an interesting by-product of the design is that there is systematic documentation of the characteristics of patients for whom there is and is not "clinical certainty." However, unless that certainty is based on good scientific reasoning, or hard evidence, there is then the further problem of to whom the trial results are pertinent. In a traditional trial the apparent answer to this question is straightforward: to all patients meeting the entry criteria. However, insofar as clinicians (and patients) have been exercising further judgment over whom to enter, the realities may be essentially similar, but undocumented. Korn and Baumrind give several examples of this.

A further advantage of the trial is that it fits with the realities of clinical practice. Most specialties have rather different clinical arrangements, so this trial design is unlikely to have wide application, but the broad principle of using this imaginative design within specific clinical constraints is a good one.

A complex design issue is the need for "balancing the design." As a mathematical exercise, it simply reduces to "making the right choice of λ's." In practice this may be an impossible task: in this trial the original plans to apply "objective" entry criteria soon fell by the wayside. A prime requirement for a clinical trial is that it is straightforward to implement alongside routine clinical practice. Unless there is a sophisticated trials infrastructure in place, this usually means a simple design is preferable.

There is potential for bias in implementation. Entry is conditional upon disagreement between a panel of 5 clinicians selected from a pool of 14. Somebody has to make the choice of which five for which patient. If that choice rests in the hands of more junior staff, their selection may influence the panel decision, because in a field like orthodontics, trainees soon tune in to individual consultants' preferences for different types of patients and may select particular clinicians to make up the panel in specific situations. Also, if the junior staff are aware of the ongoing decisions of successive panel members they have the potential to "balance" or "skew" the panel decision by selecting the appropriate clinician to be the next member of the panel.

Perhaps the question which needs to be faced is just "on what evidence is the 'clinical certainty,' of clinicians who feel unable to randomize patients, based?" Sometimes constraints will genuinely rule out the use of some treatments. In practice clinicians may have "personal" cutoff points as to who they think should have chemotherapy or, in the orthodontic case, at what level of crowding it becomes justifiable or necessary to extract teeth. Also different clinicians will randomize different types of patients and the "typing" may not be related to disease severity but to social and other factors. For example, due to peer pressure children at boarding school may not feel able to wear headgear (the appliance needed to push their teeth back and make extra space) in the dormitory at night so may be more likely to have extraction where the decision as whether to extract or not is borderline. From Table 1, it is clear that in Korn and Baumrind's study patients and, in this case, their parents also have strong opinions about treatment. More work on

their perspective, both on treatments and on attitudes to randomization in clinical trials, is likely to be useful.

Challenging a culture in which clinicians are thought to be willing to randomize only 5% of their patients is not easy. In a specialty where randomized trials are rare, this design represents a bold step forward in accepting the existence of uncertainty, and then randomizing to obtain unbiased evidence to reduce that uncertainty.

2. OBSERVATIONAL STUDIES WITH CLINICIAN PREFERENCES

From a "clinical" perspective, the observational approach has much to recommend it. As it is retrospective, it is a relatively quick way of obtaining answers, at least about therapies already in use, and it saves setting up special studies, requesting consent for randomization and so on. However, the arguments against the use of databases for this purpose are well rehearsed as Korn and Baumrind document. The authors use their causal model to identify conditions under which such an analysis might be acceptable. The idea of documenting "clinician preferences" helps to augment more traditionally measured covariates, which, it may be argued, do not fully capture patient prognosis. However, what their general approach boils down to is to assuming that, conditional on covariates, patient populations on different treatments are comparable. The authors recognize that finding a set of covariates that make treatment assignment strongly ignorable may be difficult in practice. This approach actually raises a rather wider issue. The justification for the observational analysis is essentially a post hoc recognition of uncertainty on the parts of the clinicians. If we are asked to accept, retrospectively, that, for patients with the same covariates, treatment was as good as given at random, why could they not have been formally randomized to start with?

3. TOWARD EVIDENCE-BASED ORTHODONTICS

The overall aim of the research described in this paper is to improve treatment for patients, and in particular, orthodontic patients. We wish to use rigorous statistical practice as an aid to this end. In practice we will always need a blend of the observational approach and the experimental: perhaps the main question is with what balance or emphasis? There is no doubt that studying or understanding clinician preferences and areas of disagreement is a valuable activity in its own right. There is scope for more formal study of outcomes and the utilities that they attach to them as well as a patient's own feelings. Similarly, the intelligent analysis of observational data is useful, not least in obtaining preliminary estimates for the planning of trials, and studying whether treatments are used, and how they fare in practice after more formal evaluation. However, clever analyses do not, in our opinion, overcome the need for randomized controlled trials. Before further orthodontic RCTs are planned, systematic reviews of existing evidence are required. To this end the Oral Health Group of the Cochrane Collaboration is preparing and maintaining such reviews. The authors are preparing a systematic review on the treatment for posterior crossbites and have designed an RCT based on questions raised by this review (Harrison and Ashby, 1997). Data obtained from the systematic review were used when designing the RCT and for the sample size and power calculations needed for this RCT. The trial is currently accruing patients.

For the question of extraction versus nonextraction for patients with crowding or irregularities, what contribution do these two studies make? Both the observational study and the RCT highlight where there is currently agreement or disagreement. But, however careful the analysis, the observational data is unlikely to demonstrably fulfill the assumptions for useful information on treatment comparisons. RCTs are needed, with better power than the one reported here. We do not underestimate the difficulties involved, but we take heart from the debate on cleft palate treatment: Berkowitz (1995) argued that doing trials in the area was unethical and impossible largely due to the problems of asking surgeons to perform procedures which they do not believe in or are as familiar with. However, Shaw (1995) reports that an international RCT is now in progress which is comparing two surgical procedures to treat nasopharyngeal incompetency in patients born with a cleft lip and palate. Prior to this RCT being established each procedure was used exclusively by the surgeons practicing in each of the first two centers participating in the trial, but following discussion and demonstration the surgeons now use, and are randomizing patients to, both techniques. The potential biasing effect of the surgeons' learning curve will be built into the analysis so that a significant payoff of the study will be to find out whether the learning curve of a particular effective technique is

easier. The management of bias in surgical skill in such a way would not be possible in a retrospective study.

As Korn and Baumrind say in their Introduction, randomized trials *are* the gold standard for treatment comparisons. If orthodontics is to become more evidence based and not at the mercy of the market place (Johnston, 1990) or a branch of the cosmetics industry (Vig, 1986), we believe RCT's could and should be used more widely with profit. Insofar as this paper furthers that end, both directly and indirectly, it is to be welcomed.

Rejoinder

Edward L. Korn and Sheldon Baumrind

We thank the discussants for their knowledgeable and thoughtful comments. We stress that the standard RCT remains our first choice for answering clinical questions when clinician preferences are not so strong as to interfere with its implementation. This would include studies in orthodontics of less ideologically charged issues: for example, bonding versus banding; kinds of cementing materials; alternative bracket and archwire designs and materials; and rapid palatal expansion versus slow palatal expansion. We have been pursuing alternative study designs for situations in which standard RCT's cannot be performed because of strong clinician preferences. We take this opportunity to clarify some issues raised by the discussants and to present some results of our pilot observational study using clinician preferences.

1. GENERALIZABILITY

Even in the context of a standard RCT, Zelen indicates rightly that there are questions about global inference when a nonrandom sample of patients is participating (as is practically always the case). In situations where clinicians may be reluctant to randomize patients, for example, because of their preferences and the hands-on nature of the treatments, an even smaller proportion of eligible patients may be randomized, compounding the problem. However, we agree with Ashby and Harrison that in these situations the problem is of a similar nature whether the data were collected in a standard RCT, an RCT using clinician uncertainty as the major eligibility criterion (Section 3.2) or one of the proposed randomized or nonrandomized study designs. An advantage of the proposed designs is that they may make it easier to keep track of the numbers of patients whose outcomes are not analyzed.

In all study designs, the ability to generalize to larger populations requires the assumption that the relative benefits of the treatments are similar for patients participating and not participating in the study. Although we do not minimize the importance of this assumption, the argument can sometimes be reasonably made that qualitative interactions between treatment efficacy and patient subsets are unlikely, so that the assumption is approximately satisfied.

2. RCT WITH CLINICIAN-PREFERRED TREATMENT

Unlike Freedman, we do not view the fact that our first attempt at conducting an RCT with clinician-preferred treatment did not succeed as a "strong signal." Although one should not underestimate the resources needed to conduct such a trial, the main reason for its failure was the overestimation of the number of patients who would be eligible for inclusion in the trial (Table 1). Many standard RCT's in well-established clinical settings also fail for the same reason. Given that, and given that our trial was the first time a randomized trial of any design had been attempted at the Department of Orthodontics, UCSF, we believe that the RCT with clinician-preferred treatment could profitably be tried again.

Ashby and Harrison question how the 5 clinicians were chosen from the panel of 14 to evaluate each patient, and whether a directed choice could lead to bias. There was not a directed choice; each patient's records were evaluated by the first five clinicians who were available and willing to participate in the study. However, even if the five clinicians had been selected based on patient characteristics, this should not have lead to bias since

the randomization was done after their selection and independent evaluation of the records.

3. ANALYSIS OF OBSERVATIONAL DATA USING CLINICIAN PREFERENCES

Ashby and Harrison suggest that our approach for using clinician preferences relies on the assumption that, conditional on covariates, patient populations on different treatments are comparable. Although this is the assumption that is used for standard observational analyses, the proposed designs require only that the patient populations seen by the different clinicians are comparable. It can be very difficult to find covariates that make the former assumption reasonable, whereas the latter assumption is sometimes reasonable a priori, for example, in a university clinic. The reason we can avoid the former assumption is that treatment preferences of treating clinicians are very special variables, since they constitute a surrogate for the treatment assignment mechanism.

Freedman warns that preferences may not be easy to elicit. This has not been our experience; clinicians have had no problem stating how they would treat a patient, which is our definition of treatment preference. We do believe that one cannot assume that a certain type of clinician (e.g., a radiation oncologist) will always treat patients the same way (e.g., with radiation therapy). One must record how the clinicians treat patients or have the clinicians state their preferences on a patient-by-patient basis. This is the message we perceive from the Hamilton breast cancer study.

4. RESULTS OF OBSERVATIONAL STUDY USING CLINICIAN PREFERENCES

The results on pairwise disagreements and self-disagreements for the three clinician pairs are given in Table A. The clinicians disagreed with each other 24% (= 38/156) of the time, slightly above our expected rate of 17%. However, they also disagreed with themselves 17% (= 10/58) of the time; we had hoped that this would be a rare event. Clearly, it is untenable to assume that the current treatment preference given by a clinician evaluating pretreatment records is the same as his pretreatment preference would have been. Note, however, that the proportion of self-disagreements that are XTR(then) → Non-XTR(now) is the same as the proportion that are Non-XTR(then) → XTR(now). This suggests that it might not be unreasonable to model, for each clinician, three types of patients: those for whom the clinician would prefer extraction at either evaluation; those for whom the clinician would prefer nonextraction at either evaluation; and those for whom the preference of the clinician at either of the two evaluations can be modeled as an independent Bernoulli random variable.

TABLE A

Agreements–disagreements and self-agreements–self-disagreements pooled across three pairs of clinicians (treating clinician for the patient and one nontreating clinician) concerning the appropriate orthodontic treatment for 156 *patients (XTR = extraction, Non-XTR = nonextraction)*

Patient treated with	Current treatment preference			
	Nontreating clinician		Treating clinician	
	XTR	Non-XTR	XTR	Non-XTR
XTR	59	18	19	4
Non-XTR	20	59	6	29

Using this model we estimated the treatment effect (extraction versus nonextraction) on three dependent variables: change in the vertical length of the face (face height); change in the angle the lower jaw makes with the base of the cranium (mandibular plane angle); and the change in the distance between the front cheek-side cusp of the upper right first molar and the middle of the edge of the upper right central incisor (arch length). An open research question concerns the effect of extraction on these first two variables. The null hypothesis is that there is no treatment difference. Clinically interesting alternative hypotheses are (1) the treatment effect for change in face height is less than or equal to −2 mm and (2) the treatment effect for change in mandibular plane angle is less than or equal to −2 degrees, both for extraction versus nonextraction. Change in arch length was included as a "positive control," as most orthodontists would expect a treatment effect of about −5 mm, about two-thirds the width of the extracted tooth.

The results of the analysis are given in Table B. The estimated treatment effect on change in arch length is consistent with what was expected. The confidence interval for the treatment effect on change in face height is so wide as to contain the null and alternative hypothesis; a larger sample size would be required to make a clinically meaningful statement. The estimated treatment effect for change in mandibular plane angle was 0.55 degrees, with a lower confidence interval of −1.19 degrees; we can eliminate with high probability the possibility of a clinically meaningful treatment effect of −2 degrees. We are pleased that we could

TABLE B
Linear regression analysis for three dependent variables on treatment, clinician-preference vector and clinician; estimated regression coefficient and 90% confidence interval for the treatment effect (extraction versus nonextraction)

Dependent variable	Treatment effect	90% confidence interval
Change in arch length (millimeters)	−2.79	(−5.43, −0.15)
Change in face height (millimeters)	−2.79	(−6.08, 0.50)
Change in mandibular plane angle (degrees)	0.55	(−1.19, 2.29)

make such a statement even with these limited pilot data.

Further results and description of the methods of analysis of this pilot study are given elsewhere (Korn et al.).

ADDITIONAL REFERENCES

BERKOWITZ, S. (1995). Ethical issues in the case of surgical repair of cleft palates (with discussion). *Cleft Palate–Cranofacial Journal* **32** 271–280.

DABROWSKA, D. and SPEED, T. (1990). Translation of J. Neyman, "Sur les applications de la théorie des probabilités aux experiences agricoles: essai des principes" (with discussion). *Statist. Sci.* **5** 463–480. [Originally published in *Roczniki Nauk Rolniczki* **10** (1923) 1–51, in Polish.]

FAYERS, P. M., ASHBY, D. and PARMAR, M. K. B. (1997). Tutorial in biostatistics: Bayesian data monitoring in clinical trials. *Statistics in Medicine* **16** 1413–1430.

HARRISON, J. E. and ASHBY, D. (1997). Orthodontic treatment for posterior crossbites. In *Oral Health Module of the Cochrane Database of Systematic Reviews* (W. C. Shaw, H. Worthington, A. Antczuk-Boukoms, J. F. C. Tulloch, S. Reisine and J. E. Harrison, eds.). [(Updated 2 July 1997.) Available in the Cochrane Library (database on disk and CD-ROM) The Cochrane Collaboration, Issue 3. Update Software, Oxford 1997. Updated quarterly. Available from BMJ Publishing Group, London.]

HARRISON, J. E., ASHBY, D. and LENNON, M. A. (1996). An analysis of papers published in the British and European Journals of Orthodontics. *British Journal of Orthodontics* **23** 203–209.

HODGES, J. L., JR. and LEHMANN, E. (1964). *Basic Concepts of Probability and Statistics*. Holden-Day, San Francisco.

JOHNSTON, L. E., JR. (1990). Fear and loathing in orthodontics: notes on the death of theory. *British Journal of Orthodontics* **17** 333–341.

KORN, E. L., TEETER, D. M. and BAUMRIND, S. (1998). Using explicit clinician preferences in nonrandomized study designs. *Journal of Statistical Planning and Inference*. To appear.

SHAW, W. C. (1995). Comment on "Ethical issues in the case of surgical repair of cleft palates," by S. Berkowitz. *Cleft Palate–Cranofacial Journal* **32** 277–280.

VIG, P. S. (1986). Reflections on the rationality of orthodontics: towards a new paradigm. Science and clinical judgement in orthodontics. In *Science and Clinical Judgement in Orthodontics. Craniofacial Growth Series* (K. D. Vig and P. S. Vig, eds.) **19** 31. Center for Human Growth and Development, Univ. Michigan, Ann Arbor.