Pergamon

0895-4356(93)E0003-T

J Clin Epidemiol Vol. 47, No. 3, pp. 261–270, 1994
Copyright © 1994 Elsevier Science Ltd
Printed in Great Britain. All rights reserved
0895-4356/94 $6.00 + 0.00

# AN ILLUSTRATIVE STATISTICAL ANALYSIS OF CUTOFF-BASED RANDOMIZED CLINICAL TRIALS

JOSEPH C. CAPPELLERI[1]* and WILLIAM M. K. TROCHIM[2]

[1]New England Medical Center, Boston, MA 02111, U.S.A. and [2]Cornell University, Ithaca, NY 14853, U.S.A.

*(Received 4 August 1993)*

**Abstract**—Cutoff-based randomized clinical trials (RCTs) are designed to balance ethical and scientific concerns. Patients scoring below a cutoff score on a baseline measure (i.e. the least severely ill) are assigned to the control-treated group, those scoring above a second cutoff score (i.e. the most severely ill) are assigned to the test-treated group, and those scoring within the interval (i.e. the moderately ill) are randomly assigned. This paper provides a formal illustration on the statistical analysis of cutoff-based RCTs using data from the Xanax Cross-National Collaborative Study. To overcome problems specific to cutoff-based designs, we generally recommend a backward elimination approach that tests interactions before main effects.

Cutoff-based RCTs  Clinical trials  Randomization  Experimental designs
Quasi-experimental designs  Methodology

## INTRODUCTION

In principle the conventional or traditional randomized clinical trial (RCT) provides the most powerful and scientifically rigorous method for comparing the efficacy of treatments. In recent years, however, ethical concerns about the RCT have been raised when strong *a priori* information exists that the test treatment is more effective than the control treatment. Critics of the RCT claim that randomization is unethical because research subjects are assigned arbitrarily to test-treated and control-treated groups irrespective of their needs or willingness to incur risk at the time of assignment. Under the assumption that accumulating data suggest that the test treatment may be better than a control treatment, and that the disease under investigation may be potentially very serious, critics of the RCT argue that random assignment denies the test treatment to some study patients who could benefit most by receiving the test treatment or who could be more willing to undertake the risks (i.e. side effects) that may come with the test treatment. They also claim that random assignment assigns some patients to the control-treated group who are not as much in need of the test treatment (and can therefore afford to forego the test treatment for now) or who are not as willing to chance any adverse effects that may be caused by the test treatment.

Conventional RCTs can raise ethical difficulties, even when the undertaking is important enough, when there is strong evidence that a test treatment may offer a greater benefit than a control treatment. Such evidence can come from case series, observational studies, and non-randomized trials with concurrent controls, historical controls, or no controls. For instance, recent controversies regarding the ethics of implementing RCTs in such life-threat-

*All correspondence should be addressed to: Joseph C. Cappelleri, PhD., M.P.H., New England Medical Center, Center for Health Services Research & Study Design, 750 Washington Street—Box 63; Boston, MA 02111, U.S.A.

ening diseases as extracorporeal membrane oxygenation in neonatal intensive care, AIDS, and cancer have stirred national debates about the feasibility of RCTs [1–3].

Of course, no complete certainty exists favoring a test treatment over a control treatment; otherwise there would be no need to conduct an experiment. While existing data and judgments are not sufficient, there may be a reasonable amount of existing and accumulating evidence that favors a new therapy over a control therapy, enough to incorporate into the study design.

Recent ethical and logistical criticism of the traditional RCT have become so heated that statisticians and methodologists have proposed new variations of the traditional RCT when the test treatment is believed, either *a priori* or from study data, to be more beneficial than a control treatment [4–6]. In an effort to balance ethical and scientific concerns, Trochim and Cappelleri [6] offered a new hybrid design that combines random assignment with assignment by one or more cutoff values on a baseline variable (e.g. severity of illness). In such a "cutoff-based" RCT, persons scoring below a cutoff score on a baseline measure (i.e. the least severely ill) are automatically assigned to the control-treated groups, those scoring above a second, higher cutoff (i.e. the most ill) are automatically assigned to the test-treated group, and those scoring in the interval between the cutoff scores (i.e. the moderately ill) are randomly assigned to either group. Depending on the baseline score, the patient is assigned either to treatment randomly or by the need-based, clinically-related baseline score. Their article also considered a single cutoff-point design with no randomization, known as the regression-discontinuity (RD) design, whereby all subjects scoring above a cutoff value are automatically placed in one group, while all subjects scoring below the same cutoff value are automatically placed in the other group.

In selecting the cutoff point(s) and deciding where "low risk" patients lie along the baseline continuum, the analyst should consider the following factors: the treatments involved, the width of the randomization interval, the baseline indicator itself, the nature of the disease, substantive grounds, statistical power, the desired interval and overall proportions assigned to the treatments, ethical considerations, program resources, and a sufficient number of baseline values to the left and right of the cutoff point(s) to enable adequate estimation of the true outcome-baseline functional form [7]. These factors were considered, for instance, in deciding the two cutoff points and the "low risk" patients of a cutoff-based RCT in the Cocaine Treatment Study at the University of California at San Francisco [8]. The baseline measure was a composite of four subscales, each with a different set of ordinal items. These four subscale items were weighted and added based on clinically sensibly criteria from staff members.

Trochim and Cappelleri [6] provided a review of cutoff-based assignment, conducted Monte Carlo simulations on six cutoff-based RCT design variations, and compared them to the traditional RCT design and the single cutoff (RD) design. Moreover, they performed a secondary analysis of data from the Cross-National Collaborative Study [9–13] to illustrate the cutoff-based configurations, with the Sheehan Clinician Rated Anxiety Scale [14] used as the baseline and outcome indicator. The secondary analysis confirmed the simulations and illustrated how cutoff-based designs might look with real data.

Intended to be only preliminary, the illustrative secondary analysis was limited in that an analysis of covariance model was directly applied for each design variation instead of developing an appropriate model. This paper adds to the earlier analysis by detailing the modeling building process, with emphasis on the statistical analysis of the cutoff-based designs, and comparing treatment estimates of cutoff-based RCTs with those of the original RCT. It should be kept in mind that cutoff-based RCTs are *not post-hoc* procedures applied to the traditional RCT in which all patients are randomized, but rather are alternatives to the traditional RCT when randomization cannot be done for at least a portion of the enrolled patients.

## THE REANALYSIS OF THE XANAX STUDY

Only those aspects relevant to the methodological purposes at hand are mentioned. The study was a double-blind, cross-national RCT undertaken to evaluate the comparative efficacy of Alprazolam (Xanax) to placebo primarily in the treatment of panic disorder and associated agoraphobia. The original sample consisted of 542 subjects who were randomly assigned, with equal probability of being in either the Xanax group ($n = 270$) or the placebo group ($n = 272$).

The reanalysis investigates the immediate (i.e. 1-week) effect of Xanax relative to placebo. The Hamilton Anxiety Rating Scale [15], which consists of a clinician's rating of 14 items (e.g. "anxious mood", "tension", "insomnia") on a 0 (none) to 4 (severe, grossly disabling) ordinal scale, was the sole measurement instrument that we examined. Unweighted average scores from the Hamilton Anxiety Scale, averaged across its 14 items, were taken from a clinician's assessment of a given patient and were used as both baseline and outcome scores. Patients with higher average scores suffered (presumably) higher levels of overall anxiety and, by implication, were more in need of the drug intended to reduce anxiety.

There are three reasons for selecting the Hamilton anxiety rating scale. First, in a real study employing cutoff values, clinician ratings (as well as objective measures of symptomology) would probably be a primary candidate in constructing a baseline assignment measure. Second, because the effectiveness of Xanax on anxiety is well-established, the Hamilton scale, being valid and reliable to diagnose anxiety levels, is likely to show detectable efficacy of Xanax treatment relative to placebo treatment. Third, average scaled scores of the scale should allow sufficient variability in the baseline measurement, a desirable characteristic when implementing cutoff-based designs.

## CONSTRUCTION OF THE CUTOFF-BASED DESIGNS

Cutoff-based designs were constructed by selectively discarding cases from the original, fully randomized data set in order to simulate cutoff-based assignment. Four cutoff-based experimental designs—three cutoff-based RCTs and the single-cutoff RD design—were considered along with two fully randomized (i.e. traditional) experimental designs, one containing the maximum number of baseline-outcome observations and the other containing about half as many. Based on the original set of 539 patients (out of 542) with recorded baseline measurements on the Hamilton Anxiety Scale, the baseline variable was approximately normally distributed, with very slight right-skewness, as evidenced by its mean of 1.55 being close to its median of 1.50.

We measured possible treatment effects of all six models at the baseline value of 1.55 on the baseline Hamilton Anxiety Rating Scale. Choosing the baseline value of 1.55, which was near the center of the approximately normal baseline distribution, also served to partition about 50% of the patients to each group in the cutoff-based designs, allowing them to be evenly compared among themselves and with the equally-balanced randomized designs with respect to efficiency. Therefore, each cutoff-based RCT had about 50% of its randomized patients in each group and each of the six models had about half of all its observations in each group.

Specifications for constructing the two traditional RCTs (and sample sizes) particular to the Xanax study were as follows:

- *The full-sampled RCT* (*sample size, $n = 516$*)
  The full-sampled RCT model was based on a fully randomized clinical trial of those patients who had both pre- and post-measurements on the Hamilton Anxiety Rating Scale.
- *The half-sampled RCT* (n = *243*)
  We randomly discarded 47% of the original cases so that it had about the same number of cases as the cutoff-based designs.

In each of the cutoff-based RCT, all Xanax-treated cases that resided below the interval of randomization and all placebo-treated cases that resided above it were discarded. All patients that fell within the interval were randomly assigned with equal probability to each treatment condition. Specifications for constructing the four cutoff-based models in the Xanax study were:

- *The single cutoff RD design* (n = *246*)
  All placebo-treated cases that fell above the single cutoff value of 1.55 and all Xanax-treated cases that fell below it were discarded.
- *The small cutoff-interval RCT* (n = *272*)
  A small-sized cutoff interval was arbitrarily defined here as an interval containing 17% of all cases considered. This was accomplished by using the baseline values of 1.5 and 1.6 (inclusive) to bracket the interval of randomization.
- *The medium cutoff-interval RCT* (n = *296*)
  A medium-sized cutoff interval was (arbitrarily) defined here as an interval capturing 31% of all cases considered. This was accomplished by using the baseline values of 1.4 and 1.7 to form an interval of randomization.
- *The large cutoff-interval RCT* (n = *315*)
  A large-sized cutoff interval was defined here as an interval capturing 46% of all cases considered. To accomplish this, baseline values of 1.35 and 1.75 (inclusive) were used as

the two cutoff points bracketing the cutoff interval

## THE STATISTICAL ANALYSIS OF CUTOFF-BASED DESIGNS

### *Specifying the baseline-outcome distribution*

Perhaps the most challenging and critical step in the statistical analysis of cutoff-based experimental designs is to specify the correct baseline-outcome functional form [16]. Cutoff-based designs, with or without some randomization, require extrapolation of a linear, polynomial, or other functional baseline-outcome relationship (e.g. log) to a range of the baseline values not covered by the data for that treatment. No such extrapolation is needed, of course, in fully randomized designs as both control-treated and test-treated group regression lines spread over the entire range of the baseline distribution. Extrapolation of regression lines in cutoff-based designs can result in a biased treatment estimate when the true regression line of either treatment group in the area where that group does not receive the other treatment is not a mere extension of the observed (fitted) regression line of that group [7, 16, 17].

In this illustrative analysis the "gold standard"—the conventional RCT—is known; however, in practice the "gold standard" will not be known. As such, in practical applications there is less than full certainty on whether the cutoff-based model has captured the truth in the region where data are not observed. A few complementary guidelines are offered to increase the validity and robustness of the cutoff-based design [18].

One approach, which is especially promising in the typical situation when the control treatment is the standard one, resembles a short pilot study and aims at first determining the expected regression line of the control-treated group in the area that will be eventually assigned only to test treatment. Here *all* subjects are measured on baseline, then given the control treatment, and measured on outcome after a period of time on the control treatment. The bivariate relationship between the baseline and outcome measures would probably lead to the correct functional form needed for the range of baseline values not covered by the control treatment in the actual cutoff-based study that follows. After a reasonable wash-out period, a further enhancement has all subjects temporarily on test treatment in order to determine the expected regression line of the test-treated group in the area that will be eventually assigned only to control treatment. This approach assumes that this functional form will not change in the actual study, a fairly benign assumption in most cases.

A second strategy is to smooth the data. This can be accomplished, for example, by using a moving average procedure such as one that plots average values of the outcome variable for narrowly defined columns of the baseline variable. A third option is based on strong *a priori* information regarding the functional form (whenever such is available) as when, for instance, historical data are available. A fourth guidelines adopts the recently proposed method by Robbins-Zhang [19–22], which makes no assumption about the nature of the baseline-outcome regression, that relies on empirical Bayes analysis to arrive at an unbiased treatment estimator under certain conditions. A fifth strategy, which is adopted in this article, uses a standard backward elimination approach.

Undoubtedly, the best alternative is to include as much randomization as possible. No benchmark can be given on when a cutoff-based design is "close enough" to a conventional RCT, because the degree of similarity between the two design types is essentially a sample phenomenon; what may be "close enough" for one data set may not be so for another data set.

### *Specifying the model*

Any appropriate functional form between outcome and baseline can be used. We follow a polynomial regression approach to specify the baseline-outcome functional form and to analyze the cutoff-based models, because the outcome-baseline scatterplots (see Fig. 1 for instance) show that the outcome measure can be describable as a polynomial in the baseline assignment variable, which is an assumption of the approach. Two types of standard backward elimination strategies have been proposed; both of them will be employed here. One type statistically evaluates the highest main effect term in tandem with its corresponding interaction term [7, 16]. The second type is a hierarchical approach in which interaction terms are considered before main effects [23]. No reason has been previously given to prefer one approach over the other.

The general polynomial regression model setup can be expressed as follows:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 \tilde{x}_i + \hat{\beta}_2 z_i + \hat{\beta}_3 \tilde{x}_i z_i + \hat{\beta}_4 \tilde{x}_i^2 + \hat{\beta}_5 \tilde{x}_i^2 z_i + \hat{\beta}_6 \tilde{x}_i^3 + \hat{\beta}_6 \tilde{x}_i^3 z_i + \ldots + e_i$$

where

- $\tilde{x}_i$ = baseline measure for individual $i$ minus the cutoff value
- $y_i$ = outcome measure for individual $i$
- $z_i$ = treatment group variable (1 if test-treated participant; 0 if control-treated participant)
- $\hat{\beta}_0$ = intercept estimator
- $\hat{\beta}_1$ = linear slope estimator
- $\hat{\beta}_2$ = treatment effect estimator
- $\hat{\beta}_3$ = linear interaction estimator
- $e_i$ = sample regression disturbance term

The other regression estimators are the coefficients for powers of $\tilde{x}_i$ higher than one and for higher order interaction. The major null hypothesis of interest

$$H_0\colon \beta_2 = 0 \text{ (i.e. the treatment effect parameter is zero)}$$

is tested against the alternative hypothesis

$$H_1\colon \beta_2 \neq 0.$$

Trochim [7, 16] gave a rule of thumb that starts with an initial model that goes two orders of polynomial higher than that indicated by the number of times the bivariate baseline-outcome distribution "bends" or "flexes". If a polynomial relationship is not warranted, an appropriate transformation (e.g. log, square root) on either baseline or outcome or both should be considered.

We define PREHAM as the (average) baseline Hamilton score minus the cutoff value of 1.55, and POSTHAM as the (average) original outcome Hamilton score. Figure 2 clearly shows a positive linear relationship between POSTHAM and PREHAM for all observations

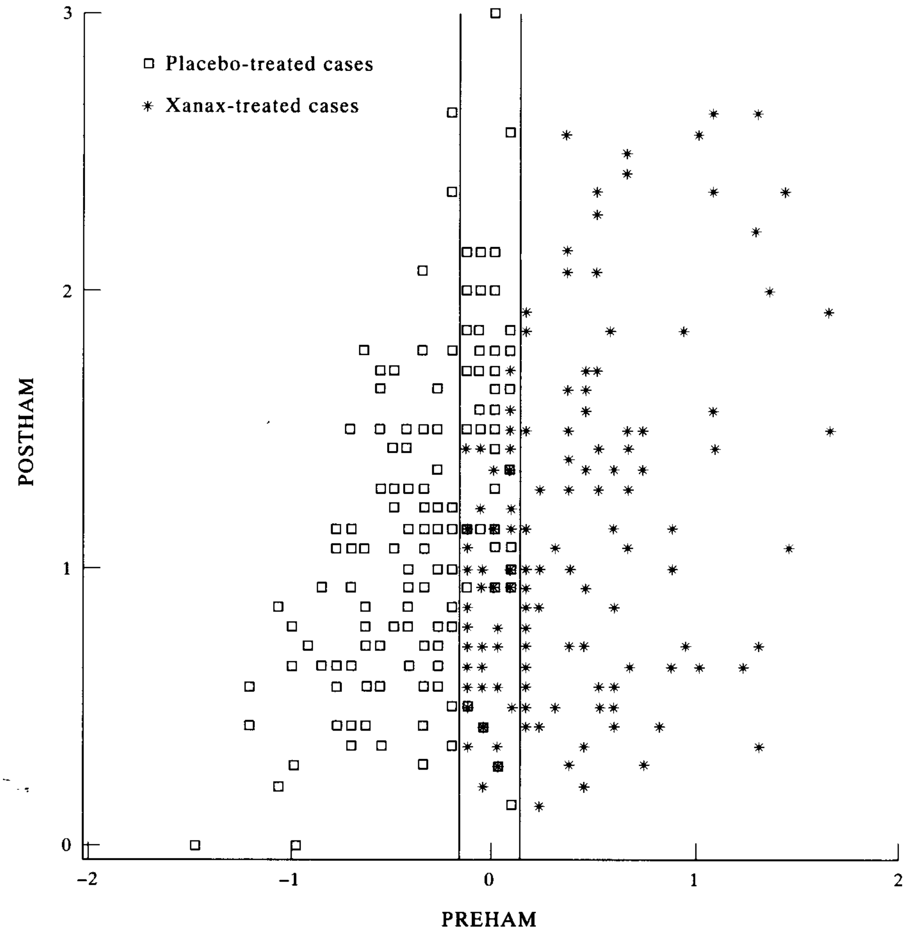

Fig. 1. Scatterplot for medium cutoff-interval RCT.

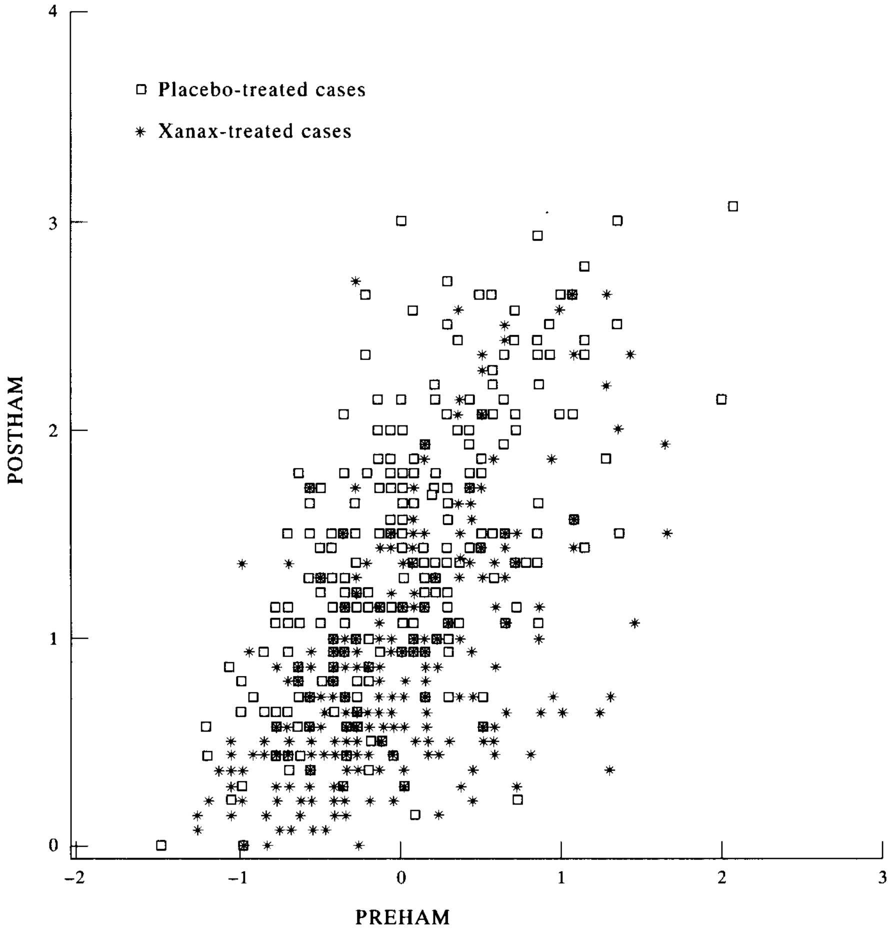

Fig. 2. Scatterplot for full-sampled RCT.

from the full-sampled RCT; the scatterplot for the half-sampled RCT (similar to Fig. 2) concurred, as expected.

The full and half-sampled RCT models shared the same set of initial variables as the cutoff-based models so that their coefficients and standard errors of their predictor variables can be directly compared. The full-sampled RCT was taken as the "gold standard" model, upon which the truth was based for these data, by which the other five models were compared and evaluated. A two-tailed hypothesis test was implemented at the 0.05 level of significance to test whether or not to reject a given null hypothesis.

Figure 1 shows that the scatterplot from the medium cutoff-interval RCT had a linear baseline-outcome relationship; scatterplots for the other cutoff-based designs (which are similar to Fig. 1) also showed a prominent positive linear relationship. So, with no visually apparent bends in the baseline-outcome scatterplot, the "flexion point" rule of thumb suggests that the initial model for each cutoff-based model should regress POSTHAM on PREHAM, the binary treatment group variable (TRT, coded 0 for placebo-treated, 1 for Xanax-treated), the linear interaction term (INTER = PREHAM*TRT), the quadratic term (PREHAM$^2$), and the quadratic interaction term (QINTER = PREHAM$^2$*TRT). (An asterisk (*) indicates multiplication.)

### *Model building with main effects and interactions tested together*

Evaluating main effects and their interactions first considers the highest order term and its corresponding interaction. The results of all six models showed that the coefficients for the quadratic term and the quadratic interaction term were not statistically significant. The coefficients of QINTER had $p$-values that ranged from as high as 0.958 for the RD model to as low as 0.439 for the large cutoff-interval RCT; the coefficients of PREHAM$^2$ had $p$-values that ranged from as high as 0.925 for the small cut-off interval RCT to as low as 0.167 for the large cutoff-interval RCT. These terms were therefore dropped from all equations.

The results based on the regression of POSTHAM on PREHAM, TRT, and INTER showed that no strong evidence of linear interaction for the full-sampled RCT ($p$-value = 0.119), half-sampled RCT ($p$-value = 0.357), and large cutoff-interval RCT ($p$-value = 0.126), but did show evidence of linear interaction for the RD design ($p$-value = 0.05), small cutoff-interval RCT ($p$-value = 0.025), and medium cutoff-interval RCT ($p$-value = 0.041). Therefore, the final models for the two RCTs and large cutoff-interval RCT were analysis of covariance models, whereas the final models for the other designs included an interaction term in addition to the treatment and baseline variables.

Table 1 contains the results of the best-fitting model for each of the six design strategies. Each of four cutoff-based experimental designs (the RD design and the three cutoff-based RCTs) and the half-sampled RCT gave a comparable significant treatment effect estimate that fell within about one standard error of the corresponding value of −0.464 from the full-sampled RCT. Results from the RD design, small cutoff-interval RCT, and medium cutoff-interval RCT indicated that Xanax proved more effective than placebo in lowering anxiety on average, but its relative effectiveness was even more for patients who suffered higher baseline levels of anxiety. Results from the two RCTs and large cutoff-interval RCT, on the other hand, indicated that the beneficial effect of Xanax over placebo did not depend on a patient's baseline anxiety level.

Table 1. Final results from the backward elimination approach based on first testing the highest main effect term and its corresponding interaction term

| Model and variables | Estimate | Standard error | $p$-Value |
|---|---|---|---|
| Full-sampled RCT ($R^2 = 0.439$) | | | |
| Intercept | 1.399 | 0.031 | 0.000 |
| PREHAM | 0.620 | 0.037 | 0.000 |
| TRT | −0.464 | 0.044 | 0.000 |
| Half-sampled RCT ($R^2 = 0.460$) | | | |
| Intercept | 1.338 | 0.047 | 0.000 |
| PREHAM | 0.686 | 0.054 | 0.000 |
| TRT | −0.400 | 0.066 | 0.000 |
| Basic RD ($R^2 = 0.183$) | | | |
| Intercept | 1.487 | 0.087 | 0.000 |
| PREHAM | 0.903 | 0.158 | 0.000 |
| TRT | −0.551 | 0.113 | 0.000 |
| INTER | −0.388 | 0.197 | 0.050 |
| Small cutoff-interval RCT ($R^2 = 0.199$) | | | |
| Intercept | 1.499 | 0.072 | 0.000 |
| PREHAM | 0.922 | 0.139 | 0.000 |
| TRT | −0.571 | 0.097 | 0.000 |
| INTER | −0.397 | 0.177 | 0.025 |
| Medium cutoff-interval RCT ($R^2 = 0.197$) | | | |
| Intercept | 1.462 | 0.064 | 0.000 |
| PREHAM | 0.861 | 0.128 | 0.000 |
| TRT | −0.535 | 0.086 | 0.000 |
| INTER | −0.334 | 0.163 | 0.041 |
| Large cut-off interval RCT ($R^2 = 0.194$) | | | |
| Intercept | 1.361 | 0.048 | 0.000 |
| PREHAM | 0.631 | 0.074 | 0.000 |
| TRT | −0.472 | 0.077 | 0.000 |

### *Hierarchical model building*

The hierarchical backward elimination approach first considers the highest interaction term in the model and, after all interaction terms are tested, then considers the highest main effect term. The coefficient of the quadratic interaction term for each of the six models was clearly not significant, with its $p$-values ranging from 0.439 for the large cutoff-interval RCT to 0.958 for the RD design. After we deleted the quadratic interaction term from each model, the coefficient of the linear interaction term was not significant, with its $p$-values ranging from 0.108 for the full-sampled RCT to 0.933 for the RD design. After we deleted the linear interaction term from each model, the coefficient of the PREHAM$^2$ term was significant for the four cutoff-based experiments but not for the half-sampled RCT ($p$-value = 0.411) and full-sampled RCT ($p$-value = 0.316).

Table 2 presents the results of the six models based on the hierarchical regression approach. Again each of the four cutoff-based experimental designs gave a comparable significant treatment effect estimate that fell within about one standard error of the corresponding value of −0.464 from the full-sampled RCT. The two RCTs and cutoff-based experiments are now in full agreement about a strong treatment effect but no interaction effect.

Again, standard errors of the treatment effect exhibited a discernible hierarchy [5]: More randomization implies more efficiency. The major difference is that the regression models for the two RCTs did not contain the PREHAM$^2$ term, but the four cutoff-based experiments did.

## AN EXPLANATION FOR THE DISCREPANCY ABOUT TREATMENT

In the hierarchical regression approach, there was no treatment-related discrepancy across models. However, in the backward elimination

Table 2. Final results from the backward elimination approach in which interaction terms are considered before main effects

| Model and variables | Estimate | Standard error | $p$-Value |
|---|---|---|---|
| Full-sampled RCT | | | |
| Intercept | 1.399 | 0.031 | 0.000 |
| PREHAM | 0.620 | 0.037 | 0.000 |
| TRT | 0.464 | 0.044 | 0.000 |
| Half-sampled RCT | | | |
| Intercept | 1.338 | 0.047 | 0.000 |
| PREHAM | 0.686 | 0.054 | 0.000 |
| TRT | −0.400 | 0.066 | 0.000 |
| Basic RD | | | |
| Intercept | 1.449 | 0.075 | 0.000 |
| PREHAM | 0.723 | 0.100 | 0.000 |
| TRT | −0.554 | 0.113 | 0.000 |
| PREHAM$^2$ | −0.150 | 0.073 | 0.040 |
| Small cutoff-interval RCT | | | |
| Intercept | 1.468 | 0.064 | 0.000 |
| PREHAM | 0.740 | 0.090 | 0.000 |
| TRT | −0.575 | 0.097 | 0.000 |
| PREHAM$^2$ | −0.161 | 0.070 | 0.023 |
| Medium cutoff-interval RCT | | | |
| Intercept | 1.440 | 0.058 | 0.001 |
| PREHAM | 0.709 | 0.083 | 0.001 |
| TRT | −0.536 | 0.086 | 0.001 |
| PREHAM$^2$ | −0.150 | 0.068 | 0.029 |
| Large cutoff-interval RCT | | | |
| Intercept | 1.406 | 0.053 | 0.000 |
| PREHAM | 0.672 | 0.077 | 0.000 |
| TRT | −0.493 | 0.077 | 0.000 |
| PREHAM$^2$ | −0.131 | 0.067 | 0.050 |

approach where main effects are evaluated jointly with their corresponding interactions, a discrepancy arose between the two traditional RCTs and three of the four cutoff-based designs with respect to the baseline-treatment interaction term.

One exploratory technique to understand the source of the discrepancy is to model the original RCT data for both Xanax-treated and placebo-treated cases separately for observations below and above the Hamilton anxiety cutoff score of 1.55. A regression analysis for baseline values below 1.55 with POSTHAM regressed on PREHAM, TRT, and INTER gave a significant treatment estimate of −0.588 with standard error of 0.095 ($p$-value = 0.00) and a significant linear interaction estimate of 0.34 with a standard error of 0.17 ($p$-value = 0.05). The corresponding regression analysis for baseline values at or above 1.55 gave a significant treatment estimate of −0.446 with a standard error of 0.110 ($p$-value = 0.00) but the linear interaction estimate of −0.114 with standard error of 0.177 was not statistically significant ($p$-value = 0.517). In both of these analyses there was no risk of curvilinearity masquerading, spuriously, as interaction; the coefficient of PREHAM$^2$ was clearly not significant ($p$-value of about 0.95 in both regressions).

Additional evidence shows that for the Xanax-treated group, but not for the placebo-treated group, each of the 14 subscale items in the outcome measurement scale of average Hamilton anxiety scores had substantially more cases with a value of zero than the corresponding subscale item on the baseline measure. Even though only a few *average* outcome scores were exactly zero, the sufficient number of subscale items that were zero for low risk Xanax-treated patients leveled their average outcome scores.

These two pieces of evidence suggest that there may have been a floor effect on the POSTHAM outcome measure for low risk patients given Xanax. Trochim [16] elaborated on measurement-related artifacts in the RD design. Because Xanax seemed to minimize anxiety scores for those patients who had lower baseline anxiety scores, and because the measurement scale reached its lowest bound of zero for "no symptoms present", low risk Xanax-treated patients cannot have done better than "no symptoms present". This may have caused low risk Xanax-treated regression line in three of the four cutoff-based designs to be more level or flatter than the regression line for low risk placebo-treated patients as evidenced by a significant interaction effect when main effects and their associated interactions were jointly evaluated. The hierarchical regression approach appeared to successfully adjust for this by using a significant quadratic baseline term instead of an interaction term. The two RCTs overcame the floor effect contamination on outcome by fitting the lines over all baseline values, across both low and high risk patients.

It should be noted that each model generally abided by the set of assumptions used for multiple regression models. If an assumption were seriously violated, standard approaches would be used to remedy the violation.

## GENERAL IMPLICATIONS

While based on only one case study, this paper has general utility. The more the regression lines in a cutoff-based RCTs cover a wider range of the baseline continuum, the less susceptible they are to extraneous factors (such as measurement-related limitations, treatment efficacy, or both) that can influence the fit in that limited range. Note that this issue in cutoff-based methodology is related to but separate from the more obvious extrapolation issue. A regression line fit over a wider range of the baseline continuum requires less extrapolation and hence gives a more reliable fit. Yet the correct form of the extrapolated regression line (e.g. a linear one) may be still obtained over a narrower range, but the fitted regression line (upon which the extrapolation is based) may have a different slope if it were based instead on a wider range of baseline values. If the form of the extrapolated regression lines is incorrect, both main and interaction estimates are likely to be biased. When the form of the extrapolations are correct, however, flooring and ceiling effects are more likely to render a biased interaction estimate than a biased treatment estimate. Even a RCT with its fitted regression line based on only a portion of the baseline range may render a different conclusion about treatment efficacy than a RCT with its regression lines spread over the entire baseline range.

Other than incorporating more randomization, one way to counter a spurious interaction effect due to the influence of measurement-related artifacts and other factors in cutoff-based designs is to use a hierarchical backward elimination approach of polynomials in which interaction terms are considered before main effects. Hierarchical backward elimination is the generally recommended strategy for polynomial modelling. Covariates should be included in the hierarchical regression analysis if they explain additional variation in outcome.

An artifact masquerading as an interaction effect in cutoff-based designs is not limited to this data set. It is a general problem that extends to all types of baseline-outcome relationships, whether or not baseline and outcome are pre- and post-measures, and to all types of diseases. Regardless of the magnitude of the treatment efficacy, a baseline or outcome measure that does not adequately differentiate among subjects who are either low scorers or high scorers can give rise to a floor or ceiling effect on that measure in cutoff-based designs [16].

The motivation for the cutoff-based design could stem from uncontrolled and potentially biased studies that treatment is better than control. It is not uncommon in clinical trials to observe the opposite of what was hypothesized. In fact, this occurred when the regression–discontinuity design showed a slightly negative program effect for Title I compensatory education programs [24]. But observing a contrary finding has no impact on the analysis of the cutoff-based design, just as it has no impact on the analysis of the fully randomized design.

Cutoff-based designs can accommodate accumulating evidence during a cutoff-based study that the control treatment is better by increasing the proportion of patients assigned to control treatment. One way to implement this is by changing the randomization interval or assignment proportions within the interval of randomization [6]. Even if control treatment turns out to be better than test treatment, the use of the cutoff-based design is justified given the particular ethical situation, where patients who are more sick before the study have a greater need to experiment with the test treatment and are more willing to assume potential risks.

This empirical case study gives practical insight that is not apparent from theoretical or simulation work. It is the first and only study that we know that formally creates and models cutoff-based assignment from a real medical data set based on complete random assignment in order to find what would have happened, relative to a conventional RCT, if an actual cutoff-based RCT were undertaken instead. More empirical studies like it are encouraged.

*Acknowledgements*—The authors gratefully acknowledge the comments of Louis M. Ryan and Steven J. Schwager as well as an anonymous reviewer. This work was supported in part by grant number HS-07782-01 awarded by the Agency for Health Care Policy and Research (Dr. Joseph Lau, principal investigator.

## REFERENCES

1. Mikè V. Suspended judgment: Ethics, evidence, and uncertainty. **Controlled Clin Trials** 1990; 11: 153–156.
2. Marshall E. Quick release of AIDS drugs. **Science** 1989; 245: 346–347.
3. Marx JL. Drug availability is an issue for cancer patients, too. **Science** 1989; 245: 345–346.
4. Ware JH. Investigating therapies of potentially great benefit: ECMO (with discussion). **Stat Sci** 1989; 4: 298–340.
5. Cappelleri JC. Cutoff-based designs in comparison and combination with randomized clinical trials. Unpublished Doctoral Dissertation. Ithaca, New York: Cornell University; May 1991.
6. Trochim WMK, Cappelleri, JC. Cutoff assignment strategies for enhancing randomized clinical trials. **Controlled Clin Trials** 1992; 13: 190–212.
7. Trochim WMK. The regression-discontinuity design. In: Sechrest L, Perrin P, Bunker J, Eds. **Research Methodology: Strengthening Causal Interpretations of Nonexperimental Data**. Washington, DC: Agency for Health Care Policy and Research, U.S. Public Health Service; 1990: 119–139.

8. Havassey B. **Efficacy of Cocaine Treatments: A Collaborative Study**. San Francisco, CA: NIDA Grant Number DA05582 awarded to UCSF; 1988.
9. Klerman GL, Coleman JH, Purpura MD. The design and conduct of the Upjohn cross-national collaborative panic study. **Psychopharmacol Bull** 1986; 22: 59–64.
10. Klerman GL. Overview of the cross-national collaborative panic study. **Arch Gen Psychiatry** 1988; 45: 407–412.
11. Ballenger JC, Burrows GD, DuPont RL, Lesser IM, Noyes R, Pecknold JC, Rifkin A, Swinson RP. Alprazolam in panic disorder and agoraphobia: Results from a multi-center trial, I. Efficacy of short-term treatment. **Arch Gen Psychiatry** 1988; 45: 413–422.
12. Noyes R, DuPont RL, Pecknold JC, Rifkin A, Rubin RT, Swinson RP, Ballenger JC, Burrows GD. Alprazolam in panic disorder and agoraphobia: Results from a multi-center trial, II. Patient acceptance, side effects and safety. **Arch Gen Psychiatry** 1988; 45: 423–428.
13. Pecknold JC, Swinson RP, Kuch K, Lewis CP. Alprazolam in panic disorder and agorophobia: Results from a multi-center trial, III. Discontinuation effects. **Arch Gen Psychiatry** 1988; 45: 429–436.
14. Sheehan DV, Ballenger JC, Jacobsen G. Treatment of endogenous anxiety with phobic hysterical and hypochondriachal symptoms. **Arch Gen Psychiatry** 1980; 37: 51–59.
15. Hamilton M. Diagnosis and rating of anxiety. **Br J Psychiatry** 1969; 69: 76–79.
16. Trochim WMK. **Research Design for Program Evaluation: The Regression–Discontinuity Approach**. Beverly Hills: Sage; 1984.
17. Cook RD, Campbell DT. **Quasi-experimentation: Design and Analysis Issues for Field Settings**. Boston: Houghton-Mifflin; 1979.
18. Trochim WMK, Cappelleri JC, Reichardt CS. Random measurement error does not bias the treatment effect estimate in the regression–discontinuity design: II. When an interaction effect is present. **Eval Rev** 1991; 15: 571–604.
19. Robbins H, Zhang CH. Estimating a treatment effect under bias sampling. **Proc Nat Acad Sci USA** 1988; 85: 3670–3672.
20. Robbins H. Zhang CH. Estimating the superiority of a drug to a placebo when all and only those patients at risk are treated with a drug. **Proc Nat Acad Sci USA** 1989; 86: 3003–3005.
21. Robbins H, Zhang CH. Estimating a treatment effect under biased allocation. New Brunswick, NJ: Working paper, Rutgers University, Institute of Biostatistics and Department of Statistics; 1990.
22. Ross DC. Monte carlo test of the Robbins-Zhang method of assigning those at risk to active treatment in clinical trials. New York, NY: Unpublished manuscript, New York State Psychiatric Institute, 1991.
23. Judd CM, Kenny DA. **Estimating the Effects of Social Interventions**. Cambridge, MA: Cambridge University Press; 1981.
24. Trochim WMK. Methodologically based discrepancies in compensatory education evaluations. **Eval Rev** 1982; 6: 443–480.