Power and Bias in Adaptively Randomized Clinical Trials
Technical Report UTMDABTR-002-06

J. Kyle Wathen
John D. Cook
Department of Biostatistics, Box 447
The University of Texas, M. D. Anderson Cancer Center
1515 Holcombe Blvd., Houston, Texas 77030, USA
jkwathen@mdanderson.org
cook@mdanderson.org

March 7, 2006
revised July 12, 2006

Abstract

This report examines the operating characteristics of adaptively randomized trials relative to equally randomized trials in regard to power and bias. We also examine the number of patients in the trial assigned to the superior treatment. The effects of prior selection, sample size, and patient prognostic factors are investigated for both binary and time-to-event outcomes.

Contents

Chapter 1

Introduction

In clinical trials, patients often are randomized between two or possibly more treatments in order to obtain statistical estimates of the treatment differences. Using standard randomization schemes, patients have an equal chance of receiving each treatment; the information obtained during the trial is not used to alter the randomization probabilities. By contrast, outcome-adaptive randomization uses interim data to unbalance the randomization probabilities in favor of the treatment or treatments having comparatively superior outcomes. This gives patients a higher chance of receiving the treatment that, on average, appears to be superior. Bayesian models are especially well-suited to such adaptive decision-making in clinical trials.

The goal of outcome-adaptive randomization is to obtain statistical estimates of efficacy while assigning treatments in a more ethical manner. While this type of randomization is clearly ethically attractive, it introduces various complications.

In this report we examine how the operating characteristics of adaptively randomized trials compare to those of equally randomized trials. We consider the effects of prior specification, sample size, and patient prognostic factors on operating characteristics. These issues are examined in the context of trials in which the primary end point is either binary or time-to-event.

1.1 Binary data

The first model is rather simple and straight forward, but has many practical applications. The primary outcome of interest is binary. Data of this type arise in situations where the outcome that is measured has two values such as response/no response to therapy, toxicity/no toxicity to treatment, and white blood count that has/has not reached a specified level.

This model assumes that the outcome of patient i on treatment j, denoted $X_{i,j}$,

has a Binomial(π_j) distribution and that *a priori* π_j is distributed as beta(α_j, β_j).

1.2 Time-to-event data

The primary outcome for this model is the time from treatment until a predefined outcome event occurs. Often the event is death or relapse of disease. However, this model can also be applied in other situations, including ones in which the event is desirable. For this design the goal is to select the treatment that has the longest (or shortest if the outcome is desirable) time-to-event (TTE).

This model assumes that the outcome $X_{i,j}$ has an exponential distribution with median parameter η_j. It is assumed that η_j will follow an inverse gamma distribution with shape parameter α_j and scale parameter β_j.

Chapter 2

Adaptive randomization models

The process of designing a clinical trial using novel statistical methods may require an iterative process of elicitation, simulation, and examination. The elicitation phase may involve analyzing historical data or questioning an expert about his or her beliefs concerning the trial. One must also gather other pertinent information about the trial, such as maximum trial duration, number of treatments, and maximum sample size. The simulation and examination stages are the focus of this report and will be discussed at length for both models. During the examination phase one examines the operating characteristics (OCs) obtained during the simulation phase, observing how the model would perform under various hypothetical scenarios. If the design does not perform well, the process of elicitation, simulation, and examination may be repeated as necessary, possibly changing the approach.

2.1 Common design parameters

There are design parameters common to both of the proposed models. These parameters control how to adapt the randomization and when to stop a trial. The trial-specific parameters include rate of accrual, total patient accrual or length of accrual, and length of additional follow-up. Typically, these parameters are determined by the nature of the trial. For example, a rare disease may have a slower rate of accrual, which would influence the decision of maximum trial duration as well as maximum accrual. Parameters that are important to the statistical aspect of running the trial include: a tuning parameter, initiation of adapting, dropping rule, stopping rule, and minimum criterion for selecting.

2.1.1 Accrual

Adaptive randomization is more sensitive to accrual rate than equal randomization because the former depends on accruing information. If patients are accruing much faster than information is accruing, the process of learning is compromised. To see this, consider the extreme case of a line of patients waiting to be treated the day the trial opens. All patients are treated before any patient observations have been made, and the so-called adaptive trial had no opportunity to adapt to data.

Erratic accrual may also be a problem. Suppose a few patients have been treated at a leisurely pace and then the long line of patients appears. The outcomes of a small number of patients would effect the randomization probabilities for the remainder of the trial with no chance for additional data to have any effect.

Of course such infinite accrual rates do not occur in practice. However, rapid accrual can exhibit the same problems, albeit to a lesser extent.

2.1.2 Length of additional follow-up

Once patient accrual has terminated, if the outcome of interest is not immediately observed a substantial amount of information may be acquired by continuing to follow patient status. Following patient status beyond the end of the accrual phase of the trial allows the last patients enrolled to contribute information to the knowledge obtained from the trial. This concept is particularly important in time-to-event trials where the time measured is large.

2.1.3 Adaptation tuning parameter

In the proposed method of AR the randomization probabilities are based on the posterior probability of each arm being the "best". It is often of interest to use a function of these posterior probabilities rather than the actual probabilities. This is often accomplished by using a tuning parameter, λ, which controls the extent to which the randomization probabilities respond to the data. For each j, $1 \leq j \leq J$, let p_j represent the posterior probability that arm j is the best. That is, if θ_j is the random variable representing the probability of response on arm j, p_j is the probability that if one were to draw a sample from all the θ's, the sample from θ_j would be the largest. For more detail, see [7] and [9] for more detail on these probabilities and their numerical calculation.

For $\lambda \geq 0$, set probability of assigning arm j to

$$\rho_j = \frac{p_j^\lambda}{p_1^\lambda + \cdots + p_J^\lambda}.$$

Setting $\lambda = 0$ results no adapting, *i.e.*, equal randomization. When $0 < \lambda < 1$ the randomization is shrunk toward ER. If $\lambda = 1$, then no tuning is done. If $\lambda > 1$ the randomization probabilities are pushed away from ER. As $\lambda \to \infty$, adaptive randomization approaches a deterministic play-the-winner rule. The values of λ that have been used to date in trials conducted at M. D. Anderson Cancer Center are 0.5, 1.0, and 2.0.

Values of λ less than 1 result in designs that are less likely to favor one arm earlier in the trial but that also assign fewer patients to the superior treatment. In contrast, larger values of λ lead to designs that are more likely favor an arm earlier in the trial, even when the treatments are equal, but that also assign more patients to the superior treatment. The simulations in this report use $\lambda = 1$.

2.1.4 Initiation of adapting

If an adaptive randomization trial beings with an uninformative prior on the response probabilities and unbalances the randomization probabilities immediately, early outcomes may unduely influence the course of the trial. There are two solutions to this problem: begin with a moderately informative prior, or require a period of equal randomization before unbalancing.

2.1.5 Dropping rule

A dropping rule is a condition that must be met in order to suspend, or possibly terminate, randomization to a particular arm. This type of rule is used to prevent a patient from being randomized to a treatment that currently has a high probability of being an inferior treatment. Formally, if

$$\Pr(\text{arm } j \text{ is the best given data}) < P_L$$

then do not allow the current patient to be randomized to treatment j. In some circumstances one may wish to terminate randomization to this arm permanently, effectively removing the arm from the trial.

2.1.6 Stopping rule

An early stopping rule is a set of one or more conditions that is used to terminate a trial before all of the patients have been enrolled. The stopping rules for this application will be described in terms of posterior probabilities of each treatment being the best treatment in the trial. Formally, if

$$\Pr(\text{arm } j \text{ is the best given data}) > P_U$$

then the trial will be terminated early and arm j selected as the best treatment in the trial. That is, if it is very likely that one of the treatments is better than all of the other treatments then the trial will be stopped.

2.1.7 Minimum criterion for selecting

Once patient accrual has terminated, one must decide whether adequate information has been acquired to conclude that one of the treatments is significantly better than all other treatments in the trial. That is, select a treatment at the end of the trial if

$$\Pr(\text{arm } j \text{ is the best given data}) > P_{\text{U final}}.$$

This criterion helps avoid selecting a treatment which is only marginally superior.

2.2 Model-specific design parameters

The technical details of each model are presented in the Appendix. This section will describe the parameters specific to each model and their interpretation.

2.2.1 Binary model

Here we discuss the model used when the primary outcome of interest has only two possible states. For example, we may be interested in whether a patient has had response to therapy, whether experienced toxicity, or whether their white blood count has reached a specified level. Typically, the time between treatment and examination is significant and is typically not 0. Because AR uses the outcome data to unbalance the randomization probabilities, time between treatment and evaluation should be short enough that few patients are accrued during the observation interval.

Let J denote the number of arms in the trial, and for each $j = 1, 2, \ldots, J$ let n_j denote the number of patients on arm j. Let $X_{i,j}$ be the indicator of response for patient i on treatment j. This model assumes that $X_{i,j}$ has a binomial distribution with probability parameter π_j. It is assumed that *a priori* the probability parameters $\pi_1, \pi_2, \ldots, \pi_J$ follow independent beta distributions with parameters α_j, β_j.

The beta distribution is a conjugate prior for the binomial likelihood. When specifying the prior distribution one may refer to the formulas for mean and variance listed in equation 5.1 of the Appendix. One possibility for specifying the prior is to set the mean of the prior equal to the historical response rate and choose a

large variance if the estimate is based on little information. See [8] for software to calculate distribution parameters for the beta given a mean and variance. Alternatively, one may specify the prior sample size, n_H, and historical rate, π_H. This corresponds to parameters of $\alpha = \pi_H n_h$ and $\beta = n_H(1 - \pi_H)$.

One may interpret α as the prior number of patients who had a response, and β as the prior number of patients who did not have a response. Thus, $\alpha + \beta$ is equivalent to the prior sample size. Figure 5.1 A is a plot of the beta distribution for various parameters.

As an example, consider a case in which one has significant prior information. Suppose the historical data includes 100 patients of which 20 responded. Using the above, $\alpha = \pi_H n_h = 20$ and $\beta = n_H(1 - \pi_H) = 80$, which corresponds to a fairly informative prior. In a small trial, say less than 100 patients, the posterior will be dominated by the prior. A common solution to this problem is to down-weight the historical data, for example using a beta(2, 8) prior, corresponding to using 10% of the data. For a plot of the distribution of this prior see figure 5.1 A. Using this prior, suppose further that after enrolling 10 patients in the trial there were 2 responses and 8 failures. The posterior would be beta(4, 16), a graph for this as well as the posterior at 30 and 60 patients, all assuming the historical rate of response, are provided in figure 5.1 B. Notice how the posterior becomes more peaked, informative, as the data accrue. Figure 5.1 C provides the same example but having specified the prior as beta(8, 2), this figure also assumes that 20% of the patients enrolled respond. Figure 5.1 D is a plot of the posteriors corresponding to the two priors.

2.2.2 Time-to-event model

This model is used when the primary outcome of interest is the time from treatment to occurrence of a pre-specified outcome. Typically, the outcome is death or relapse of disease. We assumed that the time-to-event, TTE, denoted $X_{i,j}$, of patient i on treatment j, has an exponential distribution with median η_j. We assume that *a priori* the median parameters $\eta_1, \eta_2, \ldots, \eta_J$ follow independent inverse gamma (ING) distributions with shape parameter α_j and scale parameter β_j. The data for each patient consist of a pair of the form (t, I). Here t is the elapsed time between initial treatment and either the event of interest or the patient's last follow up. The I is this patient's indicator of the event. In other words, if the patient has experienced the event, $I = 1$ and t represents the time from treatment to that event. If the event has not occurred, $I = 0$ and t represents the elapsed time from treatment until last observation. For more detail please refer to the Appendix.

The inverse gamma distribution has support $(0, \infty)$ and is a conjugate prior for the exponential likelihood. When specifying the parameters for the prior one can use the formulas for mean and variance given in equation 5.2 in the Appendix.

Usually, the mean of the prior is set equal to the expected median survival. A large variance is used if median estimate is based on little information. The α parameter may be interpreted as the prior number of patients who had experienced the event. The β parameter may be interpreted as $\log 2$ times the prior total-time-on-test. The factor of $\log 2$ comes from the fact for an exponential distribution, the mean is $\log 2$ times the median.

It is useful to note that given the survival percentage, p, at time t the median TTE is $-t \log(2 - p)$.

For example, suppose that one has historical data for the standard treatment where 110 deaths/relapses were observed with a median time-to-event of 7 months. One could set $\alpha = 111$ and $\beta = 7 \cdot 110 = 770$. However, this prior would tend to overwhelm the data in most trials and so it would be appropriate use a less informative prior, especially in a smaller trial. Two possibilities would be to use 20% or 10% of the data in calibrating the prior. This would correspond to an inverse gamma(23, 154) or an inverse gamma(12, 77) respectively. Another alternative would be to set the mean of the prior equal to the observed median, 7, and select a large variance, say 100. (See [8] for software to calculate the distribution parameters corresponding a specified mean and variance.) This leads to an inverse gamma (2.49, 10.4) distribution. Each of the priors are plotted in figure 5.10. Suppose that the true median TTE were 10.5 and one wanted to know what a typical posterior would look like after enrolling 60 patients in the trail. For two of the above mentioned priors the posteriors have been plotted in figure 5.11. For the most informative prior, inverse gamma(111, 770), *a priori* the 95% credible interval is (5.8, 8.4) after enrolling 60 patients one could expect the 95% posterior credible interval to be similar to (7.1, 9.6), which does not contain true median, 10.5. However, for the less informative inverse gamma(12, 77) prior, *a priori* the 95% credible interval is (3.9, 12.4). After enrolling 60 patients one could expect the 95% posterior credible interval to be similar to (7.9, 12.5), which does contain the true median survival.

For more details concerning the TTE model, see [19].

Chapter 3

Simulation study

For both models the OCs can only be obtained via simulation. Before discussing the particulars of the simulations carried out for each model, it is useful to give an overview of the concept of simulating a clinical trial. The use of computer simulation in trial design is important because it informs person designing the trial of what is likely to happen under various circumstances. It provides a way of comparing the properties of different designs and eliminating poor designs before ever enrolling a patient.

A simulation must adequately represent the entity being simulated. If the simulation of a trial does not reflect how the trial will actually be conducted then the results may be misleading. As a basic example, consider the case of a binary outcome. If there is a long time lag between treatment and evaluation, a simulation which assumes outcomes are known immediately could be misleading.

It is vitally important to understand the difference between the statistical model that is used to design and make decisions and the model that is used to generate the data. The binary and TTE models described above are used to calculate the randomization probabilities and to evaluate dropping and stopping rules. But the model used to simulate the data may (and often *should*) be different than the statical design model. The distinction is easily understood with a simple example.

Example

Suppose one wanted to conduct a trial where the primary outcome is binary and there are two treatments under investigation. Assume that the probability of response for a patient that receives treatment 1 is π_1 and for one receiving treatment 2 is π_2. Suppose that based on historical data or expert opinion, *a priori* $\pi_1 \sim$ beta(2, 8) and assume that π_2 follows the same distribution, independent of π_1. That is the statistical model. In order to simulate the trial one must now decide how data should be simulated. This is playing a "what if" game. For example, what if the patients treated on arm 1 have a 30% response and those on arm 2

have a 60% response rate? Also, one could generate patient outcomes based on a prognostic covariate, even though the statistical design model did not include such a covariate.

The simulation study presented below is intended to address the following points.

> Q1) When comparing the power for AR to a standard method using ER, how do the OCs compare and how much are they affected by the amount of information used in specifying the prior for the standard treatment?
>
> Q2) Does AR assign more patients to the better treatment?
>
> Q3) Are the estimates obtained biased by the use of AR, and if so, what is the nature of that bias?

Sensitivity Analysis

After choosing the design and carefully selected the parameters to obtain good OCs, it is often of interest to see how much the OCs will be changed if the underlying assumptions of the model are not met. The three questions that most commonly arise are as follows.

> S1) What happens if the standard treatment has different results than it did in the historical data?
>
> S2) How much does the presence of a prognostic variable that is not accounted for in statistical model effect the OCs?
>
> S3) Can AR be used in small-sample trials?

3.1 Simulation study for binary model

3.1.1 Design details

Assume a trial has two treatment arms: standard, S, and experimental, E. Patients arrive according to a Poisson process with an average of 5 patients per month. There is one month between treatment and outcome evaluation. The first 25% of patients were randomized with equal chance of receiving either treatment. For convenience,

assume exactly half of these early patients would be assigned to each treatment. Alternatively, assume a balanced randomization was used rather than purely equal randomization so that the initial patients were evenly distributed.

Since the standard of care is included in the trial, and for ease of understanding, no early stopping rules were enforced. Assume historical data were available on 100 patients and that the historical response rate was 30%.

For item Q1) the standard design that will be used for power comparisons is the two-sample test of proportions which can easily be obtained using S-PLUS or other statistical software. We use varying percentages of the data in order to calibrate the prior for the standard treatment: 3% corresponding to a vague prior; 10%, a slightly more informative prior; and 20%, the most informative under consideration, though still not unreasonably informative. Thus, *a priori* $\pi_S \sim$ beta(1.2, 2.8), beta(3, 7), or beta(6, 14). Assuming there is little information on the experimental treatment, we take $\pi_E \sim$ beta(1.2, 2.8) *a priori*. This prior corresponds to a prior belief that, on average, the two treatments have the same percent response. This is important because ethics dictate that one must have equipoise in order to run a randomized trial. We use the posterior mean as an estimate for the probability of response in order to study bias.

Item S1) is easily addressed by allowing the response rate for the standard treatment to be different than its historical mean. For this particular we let $\pi_S = 0.15$ and $\pi_S = 0.45$ in addition to the historical $\pi_S = 0.3$.

For item S2), we simulated the data under the assumption that a prognostic covariate, Z, was present that was not accounted for in the statistical model. This was done by fixing the covariate effect and treating the standard arm as baseline. The covariate effect is taken to be an indicator variable. For example, if a patient's performance status (PS) is good ($Z = 0$) then they have one probability of response, and a lower probability of response if their PS is poor ($Z = 1$). We assume that $Pr(Z = 1) = 0.3$. Specifically, for generating the data, for patient i who receives treatment E, $T_i = 1$, the probability of response is $\text{logit}^{-1}(\beta_0 + \beta_1 Z_i + \beta_2 T_i)$. Patients who receive S, $T_i = 0$, with a good PS have a response rate equal to the historical rate of 30%. Therefore, we set $\beta_0 = -0.8472979$ so that patients with a good PS have a 30% response rate. To better understand the effect of a covariate, we examine two covariate effect sizes. For the first, we set $\beta_1 = -0.8873032$ so that the patients who have a poor PS have a 15% response rate. For the second covariate effect size, we set $\beta_1 = -0.5389965$ so that the patients who have a poor PS have a 20% response rate. The parameter β_2 is easily calculated to give the desired response rate for a patient with Z = 0 receiving treatment E.

In order to address S3) sample sizes of 120 and 60 are considered.

Each scenario was simulated with 5000 trials.

3.1.2 Results

Figure 5.2 A is a plot of the power curves for each of the three priors as well as that of ER for a trial with 120 patients. As one can see the power curves are similar for each of the three priors and ER when $\pi_E \leq \pi_S$ but, are different when $\pi_S > \pi_E$. Figure 5.2 B suggests one may expect a substantial proportion of the patients to receive the superior treatment when using AR. Also, as the difference in the treatments becomes smaller, the number of additional patients assigned to the better treatment decreases. But, for an increase in the response rate from 0.3 to 0.45, an additional 27 patients on average would receive the superior treatment. Figure 5.2 C shows the number of additional responses for adaptive randomization compared to equal randomization as a function of the response probability. For the same increase, from 0.3 to 0.45, four addition patients respond on average.

Bias

Figure 5.2 D is a plot of the bias for both the experimental treatment and the standard treatment. For this calculation the posterior mean was used as the estimate for probability of response. There are two interesting aspects of this graph. First, when $\pi_E = \pi_S = 0.3$ the bias is not 0 for either treatment. Second, the bias for the standard treatment becomes negative when π_E is close to π_S.

Why is adaptive randomization negatively biased when π_E and π_S are equal? It seems as though everything is symmetric, especially if the standard and experimental arms have the same prior, and thus unbiased. But there is an important asymmetry at the heart of adaptive randomization: better performing arms get more patients. A given arm may perform poorly for the first few patients in one trial but perform better in another trial. These random fluctuations do not entirely cancel out each other.

When an arm performs better than it "should" early on, it receives more patients and thus has more opportunity to bring the posterior mean closer to the mean of the simulated data. But when an arm performs worse than it "should" early on, it gets fewer patients and thus has less chance to change its posterior value. Thus early over-estimates tend to regress to the mean more efficiently than early under-estimates.

For unequal values of π_E and π_S, the bias is best explained by noting that the posterior mean is a compromise between the observed data and the prior. Thus the effectiveness of poorly performing treatments is over-estimated and the effectiveness of well performing treatments is under-estimated.

However, the extent of the bias is not large and may be an acceptable price to pay for the benefit of treating more patients on the more effective arms. Fortunately,

the bias is smallest where the power is smallest, and largest where the power is largest.

Response rates different from historical

To further understand how AR works another "what if" question comes to mind. What if π_S is not the same as the historical rate? To explore this question, set $\pi_S = 0.15$ and then to 0.45, leaving all other parameters unchanged. Figures 5.3 A - D and figure 5.4 A - D are graphs of the results. One interesting point is that when $\pi_S = 0.15$, ER has more power, as measured by a 2-sample test of proportions. This is explained by recalling that π_S is a compromise between the data and prior. This suggests that the standard treatment has a higher response rate than what is being seen in the trial. In fact, the beta(6, 14) prior has the lowest power due to the amount of shrinkage. However, figures 5.3 B-C shows that more patients still receive the better treatment and so you can expect to have more patients respond using AR. The graphs of bias look just as one might expect. If $\pi_S = 0.15$ the bias is positive for $\pi_E < \pi_S$, increases as the two values become closer, and then levels off. The bias does not return to near as zero as in the case when $\pi_S = 0.3$ because the estimate is a compromise between the data and prior.

In the case of $\pi_S = 0.45$, larger than the historical rate, AR has more power, since we have incorporated prior belief that would indicate that π_S is actually less than what is currently being observed in the trial.

Heterogeneity in patient population

As described in Section 3.1.1, heterogeneity in a patient population was simulated by introducing a covariate in the simulation model but not accounting for it in the statistical model. In order to obtain the power for the standard two-sample test, a short script was written in S-PLUS to generate the patient population. Then using the same two-sample test, but this time with a population that is heterogeneous, the power was obtained by simulation. For this simulation only the less informative prior, beta(1.2, 2.8) was used. The heterogeneity effects the power of both AR and ER. See figures 5.5 A and B for plots. Introducing a covariate into the patient population that is not accounted for in the statistical model compromises the power; a larger sample size would be needed to maintain the same power. Figure 5.5 C is a plot of the power curve for $\beta_1 = 0, -0.539$ and -0.887, where $\beta_1 = 0$ corresponds to no covariate effect. These graphs show that for a covariate which occurs in 30% of the patients and reduces the response rate from 30% to 15%, there is only a small loss in terms of power. In addition, the same is also true for the number of patients assigned to the superior arm. This could be explained by the larger sample size.

Small sample size

The above simulation was useful in understanding how AR works and exploring its use in a large trial. However, often the available resources may limit the sample size. The simulation study above was repeated with only 60 patients to illustrate the effects of smaller sample sizes. Figure 5.6 A is a plot of the power curve for AR and ER using the same test as before. In contrast to the large sample, for $\pi_E > \pi_S$ the power for AR is better using any of the three priors. In addition, even for this relatively small trial, one may expect to have approximately 5 more patient responses if π_S is 0.6. The estimates for π_S and π_E are slightly more biased due to the smaller sample.

For the case of $\pi_S = 0.15$ see figure 5.7. Very little power is lost even if π_S is different for the historical rate. However, in the case where $\pi_S = 0.45$ the power is much greater using AR rather than the standard ER design. In either case, the estimates of π_S and π_R are slightly more biased than in the case with 120 patients.

3.2 Simulation study for time-to-event model

3.2.1 Design details

As in the binary model example, we consider a trial having two treatment arms: standard, S, and experimental, E. Also as before, patients accrue according to a Poisson process with mean of 5 patients per month, no early stopping rules were enforced, and the first 25% of the patients were equally randomized.

The outcome in this section is survival time. We assumed that there were historical data available on 110 patients and that the historical median TTE was 7 months. We use a variety of priors. Using 10% of the data, the corresponding prior is $\eta_S \sim$ inverse gamma(12, 77). Retaining the historical mean and setting the prior variance to 100 results in the vague prior $\eta_S \sim$ inverse gamma(2.49, 10.4). Graphs of these prior can be seen in figure 5.10. Assuming there is little information available on the experimental arm, we use the vague prior for the experimental arm. Bias was calculated using the posterior mean as an estimate for the median TTE for each treatment.

To address Q1) for comparison purposes to ER, a log-rank test for survival was used.

Item S1) is investigated by allowing the median TTE to vary while using the same prior. For simulation, median TTE was set to either 3.5 or 14 while using the same priors and trial parameters as above. This simulation also allows investigation of sensitivity to the prior specification.

There are many ways to include a covariate in the simulation model to investigate S2. Here we assume each patient has a single covariate, Z_i, and that $Pr(Z_i = 1) = 0.30$. The patient's TTE is generated from an exponential distribution with median $= (1 + \beta Z_i)\eta_j$ for $\beta > -1$ where η_j is the median TTE for treatment j when $Z_i = 0$. When $\beta < 0$ a patient with $Z = 1$ is expected to have a shorter TTE than a patient with $Z = 0$. For $\beta = 0$ there is no patient heterogeneity. When $\beta > 0$ a patient with $Z = 1$ is expected to have a longer TTE than a patient with $Z = 0$. To better understand the effect of the covariate, we simulate using two covariate effect sizes. That is, we set β equal to -0.5 and -0.75. These values correspond to a patient with $Z = 1$ having a median TTE equal to 50% or 25%, respectively, of the median TTE for a patient with $Z = 0$.

Each of the above simulations was carried out with a total accrual of 120 and 60 in order to test sensitivity to sample size.

3.2.2 Results

Figure 5.12 A is a plot of the power curves using the log-rank test for AR using each of the two priors and for ER. When $\eta_E < \eta_S$ all three power curves are nearly identical. However, when $\eta_E > \eta_S$ the design using the prior incorporating 10% of the data has greater power, while the design using the vague prior still have a power nearly identical to that obtained using ER. However, graph B illustrates that if patients receiving the experimental treatment have less than a 50% increase in median survival, then you could expect to assign an additional 25 patients to the superior treatment using AR.

Figure 5.12 C is a plot of the bias for both the experimental arm and standard arm. As was seen in section 3.1.2, the bias becomes smaller in magnitude when the treatments are equal. The bias for the standard treatment is never worse than -3%. The discussion of bias from the binary model is also relevant here.

Standard median TTE different from historical

In order to investigate the effect of a different median TTE from the standard treatment, η_S was set to either 3.5 or 14 and the simulations re-run with the same design and prior specifications as above. The results for $\eta_S = 3.5$ are in figures 5.13 A - C and $\eta_S = 14$ are in figures 5.14 A - C.

Similar to the binary case, when $\eta_S = 3.5$, less than the historical value of η, the prior incorporating more prior information has lower power. When $\eta_E \in (3.5, 4.5)$ more patients are assigned to the inferior arm. When η_S is slightly smaller than η_E about 67% of the patients are assigned to the standard treatment and 33% to the experimental. When $\eta_E = 4.5$ approximately 50% of the patients are randomized

to each treatment. As η_E increases, the bias for η_S also increases. This is due largely to that fact that when η_E is large the standard arm will only receive about 20 patients, which leads to a larger bias. The inverse gamma(12, 77) prior causes more shrinkage than the less informative prior and thus leads to a more biased estimate of the median TTE for the standard treatment.

When $\eta_S = 14$, larger than the historical value, the power for the more informative prior is larger than both the less informative prior and ER, as can be seen in figure 5.14 A. Figure 5.14 B is a plot of the additional patients receiving the superior treatment. For the inverse gamma(22, 77) prior, when $\eta_E \in (11, 14)$ the number of additional patients assigned to the superior treatment, standard in this case, is negative. But for the inverse gamma(2.49, 10.4) prior, the number of additional patents receiving the superior treatment is never negative.

Heterogeneity in patient population

As described in Section 3.2.1, we simulated patient survival times from a heterogeneous population. We examined two covariate effect sizes. This report examines power comparison for AR when the patient population is or is not heterogeneous.

The first covariate effect, $\beta = -0.5$, reduces the median survival by 50% for patients with the covariate, see figures 5.15 A - C and the second covariate effect, $\beta = -.25$,, reduces the median survival by 75% for patients with the covariate, see figures 5.16 A - C. Plot A shows that the prior has little effect on the power in either case. But of more interest is the amount of power lost by having a heterogeneous population. Graph B is a plot of the loss of power, power for a population with no covariate minus power for a population with a covariate. For the inverse gamma(12,77) prior, the loss of power is always less than 0.15 for the first covariate effect, $\beta = -0.5$, and 0.26 when using the second covariate effect, $\beta = -0.25$. For the inverse gamma(2.49, 10) prior the loss of power is less than 0.06, 0.13 for covariate effects 1 and 2 respectively. In terms of the additional patients receiving the better treatment there is little difference for covariate effect 1, and only a moderate difference for covariate effect 2.

Small sample size

The resources needed to run a large 120 patient trial are not always available and it is therefore sometimes necessary to run a trial with fewer patients. We repeat our simulations for trial with 60 patients and examined the same ideas as above. Figures 5.17 A - C are plots of the power, additional patients receiving the superior treatment, and the bias for a smaller trial. In order to calibrate the design parameters, the null case, $\eta_S = \eta_E = 7.0$, was simulated until a false-positive rate of 5% was achieved. As one would expect, using a more informative prior gives

higher power, figure 5.17 A. In addition, a more informative prior also leads to more patients being assigned to the superior treatment. The bias is smaller for the more informative prior.

However, figures 5.18 A - C for $\eta_S = 3.5$ and figures 5.19 A - C for, $\eta_S = 14$, show that the choice of a more informative prior may not be the best way to go. Using the inverse gamma(12, 77) prior leads to inferior power when η_S is less than the historical value of η and superior power when η_S is greater than the historical value. In either case, the use of the more informative prior has the possibility of randomizing slightly more patients to the inferior treatment.

In a smaller trial where the patient population is heterogeneous there is potential for significant loss in power when compared to the same design used in a homogenous population. Figures 5.20 A - C and figures 5.21 A - C are graphs for the simulations for two different covariate effects sizes. If a covariate reduces the median TTE by 50% ($\beta = -0.5$) there is potential for a 0.2 reduction in power. For the case when the covariate is more extreme and reduces the median TTE by 75% there is a possibility for a 0.3 reduction in power. In either case, using the less informative prior one can still expect to assign more patients to the superior treatment. However, with the more informative prior there is a small possibility to assign a few (less than 5) more patients to the inferior arm.

Chapter 4

Discussion

The main purpose of adaptive randomization is to provide a more ethical way of randomizing patients in a clinical trial. The simulations above show that when the assumptions of the model are met, AR can be expected to assign more patients to the superior treatment. AR appears to have similar power to ER when using a standard test for difference at the end of the trial. However, one concern with using AR is the bias that is introduced when using the posterior mean as the estimate for the parameter of interest. Even if the assumptions of the model are met and the treatments are equal at the end of the trial the posterior mean tends to underestimate the probability of response, median time-to-outcome, and over estimate the probability of treatment failure.

When the patient outcomes on the standard treatment are different than what was observed in the historical data, incorporation of more data into the prior can have adverse effects. When the standard treatment is less effective than it was in the historical data the power is less than one could expect using ER. However, one could still expect to assign more patients to the superior treatment. When the standard treatment is more effective than it was historically, the power is generally higher using AR than it is using ER. However, if the experimental arm is just slightly worse than the standard and the standard treatment performs more poorly than it had historically, the use of a more informative prior leads to the possibility of assigning more patients to the inferior arm.

When a covariate is present that is not accounted for in the statistical model, the OCs are mildly affected. The power decreases slightly in a larger trial ($N = 120$ patients) but more in a smaller trial ($N = 60$ patients). Even with the loss of power AR is still expected to assign more patients to the superior treatment. While the loss of power is not always substantial, it could be reported in cases where there is a possibility of a covariate that is not being accounted for.

Chapter 5

Appendix

For a trial with J treatment arms, let $X_{i,j}$ denote the outcome of patient i on treatment j.

5.1 Binary data model

Set $X_{i,j} = 1$ if the i^{th} patient's outcome is a "success" and $X_{i,j} = 0$ if the i^{th} patient's outcome is a "failure".

Let π_j be the probability of "success" on treatment j = 1, 2, . . . , J.

$$X_{1,j}, X_{2,j}, \ldots, X_{n_j,j} \mid \pi_j \sim \text{ i.i.d binomial}(\pi_j)$$

$$\pi_j \mid \alpha_j, \beta_j \sim \text{beta}(\alpha_j, \beta_j)$$

Suppressing the j index:

$$p(X_i = x_i \mid \pi) \;=\; \pi^{x_i}\,(1-\pi)^{1-x_i}$$

Let $S^+ = \sum_{i=1}^{n} X_i$ which is the number of "successes" and $F^+ \;=\; \sum_{i=1}^{n}(1-X_i)$, which is the number of "failures" on treatment j.

Likelihood:

$$\begin{aligned}\mathcal{L}(\boldsymbol{X} \mid \pi) &= \prod_{i=1}^{n} p(X_i = x_i \mid \pi) \\ &= \prod_{i=1}^{n} \pi^{x_i} \ (1-\pi)^{1-x_i} \\ &= \pi^{S^+} \ (1-\pi)^{F^+}\end{aligned}$$

Prior:

$$p(\pi \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \ \pi^{\alpha-1} \ (1-\pi)^{\beta-1}; \quad \alpha > 0, \beta > 0$$

$$\mathrm{E}(\pi) = \frac{\alpha}{\alpha+\beta}; \quad \mathrm{var}(\pi) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{5.1}$$

Posterior:

$$\begin{aligned}p(\pi \mid \boldsymbol{X}) &\propto \pi^{S^+} \ (1-\pi)^{F^+} \ \pi^{\alpha-1} \ (1-\pi)^{\beta-1} \\ &\propto \pi^{\alpha+S^+-1} \ (1-\pi)^{\beta+F^+-1}\end{aligned}$$

Therefore

$$\pi \mid \boldsymbol{X} \sim \mathrm{beta}(\alpha + S^+, \beta + F^+).$$

5.2 Time-to-event data model

Let η_j be the median TTE of a patient given treatment j.

$$X_{1,j}, X_{2,j}, \ldots, X_{n_j,j} \mid \eta_j \ \sim \ \text{i.i.d. exponential with median} = \eta_j$$

$$\eta_j \mid \alpha_j, \beta_j \ \sim \text{inverse gamma}(\alpha_j, \beta_j)$$

Therefore, suppressing the j index

$$p(X_i = x_i \mid \eta) \;=\; \frac{\log(2)}{\eta} \exp\left\{ \frac{-x_i \log(2)}{\eta} \right\}$$

Note: For this parameterization

$$\begin{aligned} S(X_i \mid \eta) \;&=\; p(X_i > x_i \mid \eta) \\ &= \int_{x_i}^{\infty} \frac{\log(2)}{\eta} \exp\left\{ \frac{-x_i \log(2)}{\eta} \right\} \\ &= \exp\left\{ \frac{-x_i \log(2)}{\eta} \right\} \end{aligned}$$

Let $\delta_i = 1$ if patient i has had the outcome and 0 otherwise, $T^+ = \sum_{i=1}^n x_i, \quad E^+ = \sum_{i=1}^n \delta_i$

Likelihood:

$$\begin{aligned} \mathcal{L}(\boldsymbol{X} \mid \eta) \;&=\; \prod_{i=1}^{n} \left\{ p(X_i = x_i \mid \eta) \right\}^{\delta_i} \;\; \left\{ S(X_i \mid \eta) \right\}^{1-\delta_i} \\ &=\; \left\{ \frac{\log(2)}{\eta} \right\}^{E^+} \exp\left\{ -\frac{\log(2)\; T^+}{\eta} \right\} \end{aligned}$$

Prior:

$$p(\eta \mid \alpha, \beta) \;=\; \frac{\beta^\alpha}{\Gamma(\alpha)} \eta^{-(\alpha+1)} \exp(\beta/\eta)$$

$$\mathrm{E}(\eta) = \frac{\beta}{\alpha - 1}; \quad \mathrm{var}(\eta) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)} \tag{5.2}$$

Posterior:

$$\begin{aligned} p(\eta \mid \boldsymbol{X}) \;&\propto\; \eta^{-E^+} \exp\left\{ -\frac{\log(2)\; T^+}{\eta} \right\} * \eta^{-(\alpha+1)} \exp(\beta/\eta) \\ &\propto \eta^{-(\alpha+1+E^+)} \exp\left\{ -\frac{\beta + \log(2)\; T^+}{\eta} \right\} \end{aligned}$$

Therefore

$$\eta \mid \boldsymbol{X} \sim \text{inverse gamma}(\alpha + E^+, \beta + \log(2)\; T^+)$$

Bibliography

[1] Donald A. Berry and Bert Fristedt (1979). Bernoulli One-Armed Bandits-Arbitrary Discount Sequences. *The Annals of Statistics* **7**, 1086-1105.

[2] Donald A. Berry (1989a). Investigating therapies of potentially great benefit: ECMO. *Statistical Science* **4**, 306-310.

[3] Donald A. Berry (1989b). Monitoring accumulating data in a clinical trial. *Biometrics* **46**. 1197-1211.

[4] Donald A. Berry and Stephen G. Eick (1995). Adaptive assignment versus balanced randomization in clinical trials: a decision analysis. *Statistics in Medicine*, **14**, 231-246.

[5] Donald A. Berry (2004). Bayesian statistics and the ethics of clinical trials. *Statistical Science* **19**, 175-187.

[6] Ying Kuen Cheung, Lourdes Y. T. Inoue, J. Kyle Wathen, and Peter F. Thall (2005). Continuous Bayesian adaptive randomization based on event times with covariates. *Statistics in Medicine*, in press.

[7] John D. Cook (2003). Numerical Computation of Stochastic Inequality Probabilities. M. D. Anderson technical report UTMDABTR-008-03.

http://www.mdanderson.org/pdf/biostats_utmdabtr00803.pdf.

[8] John D. Cook, Hoang Nguyen, and J. Kyle Wathen (2005). Parameter Solver [Computer software].

[9] John D. Cook and Saralees Nadarajah (2006). Stochastic Inequality Probabilities for Adaptively Randomized Clinical Trials Biometrical Journal. In press.

[10] Janis Hardwick, Robert Oehmke, and Quentin F. Stout (1999). A program for sequential allocation of three Bernoulli populations. *Computational Statistics & Data Analysis* **31**, 397-416.

[11] S. J. Pocock (1977). Adaptive treatment assignment methods and clinical trials. *Biometrics* **33**, 743-749.

[12] S. J. Pocock (1979). Allocation of patients to treatments in clinical trials. *Biometrics* **35**, 183-197.

[13] S. J. Pocock (1991). A decade of progress in statistical methodology for clinical trials. *Statistics In Medicine* **10**, 1789-1817.

[14] W. F. Rosenberger (1996). New directions in adaptive designs. *Statistical Science* **11**, 137-149.

[15] Richard M. Royall (1991). Ethics and Statistics in Randomized Clinical Trials *Statistical Science* **6**, 52-62.

[16] W. F. Rosenberger and J. M. Lachin (1993). The use of response-adaptive designs in clinical trials. *Controlled Clinical Trials* **14,** 471-484.

[17] David Spiegelhalter, Laurence Freedma, and Mahesh Parmar (1994). Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society Ser. A* **157**, 357-416.

[18] Peter F. Thall and J. Kyle Wathen (2005). Covariate-adjusted adaptive randomization in a sarcoma trial with multi-stage treatments. *Statistics in Medicine* **27**, 1947-1964.

[19] Peter F. Thall, Leiko H. Wooten, and Nizar M. Tannir (2005). Monitoring Event Times in Early Phase Clinical Trials: Some Practical Issues, *Clinical Trials* **2**, 467-478.

[20] William R. Thompson (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* **25**, 285-294.

[21] Wei, L. J. (1978). The adaptive biased coin design for sequential experiments. *The Annals of Statistics* **6**, 92–100.

[22] J. Kyle Wathen, Hoang Nguyen, and John D. Cook (2005). Inequality Calculator [Computer software].

[23] J. Kyle Wathen and Odis Wooten (2005) Adaptive Randomization 3.1.1 [Computer software].

Software applications listed above are freely available at

http://biostatistics.mdanderson.org/SoftwareDownload.

MDACC biostatistics technical reports are available at http://tinyurl.com/743xv.

Figure 5.1: Example for Beta Prior

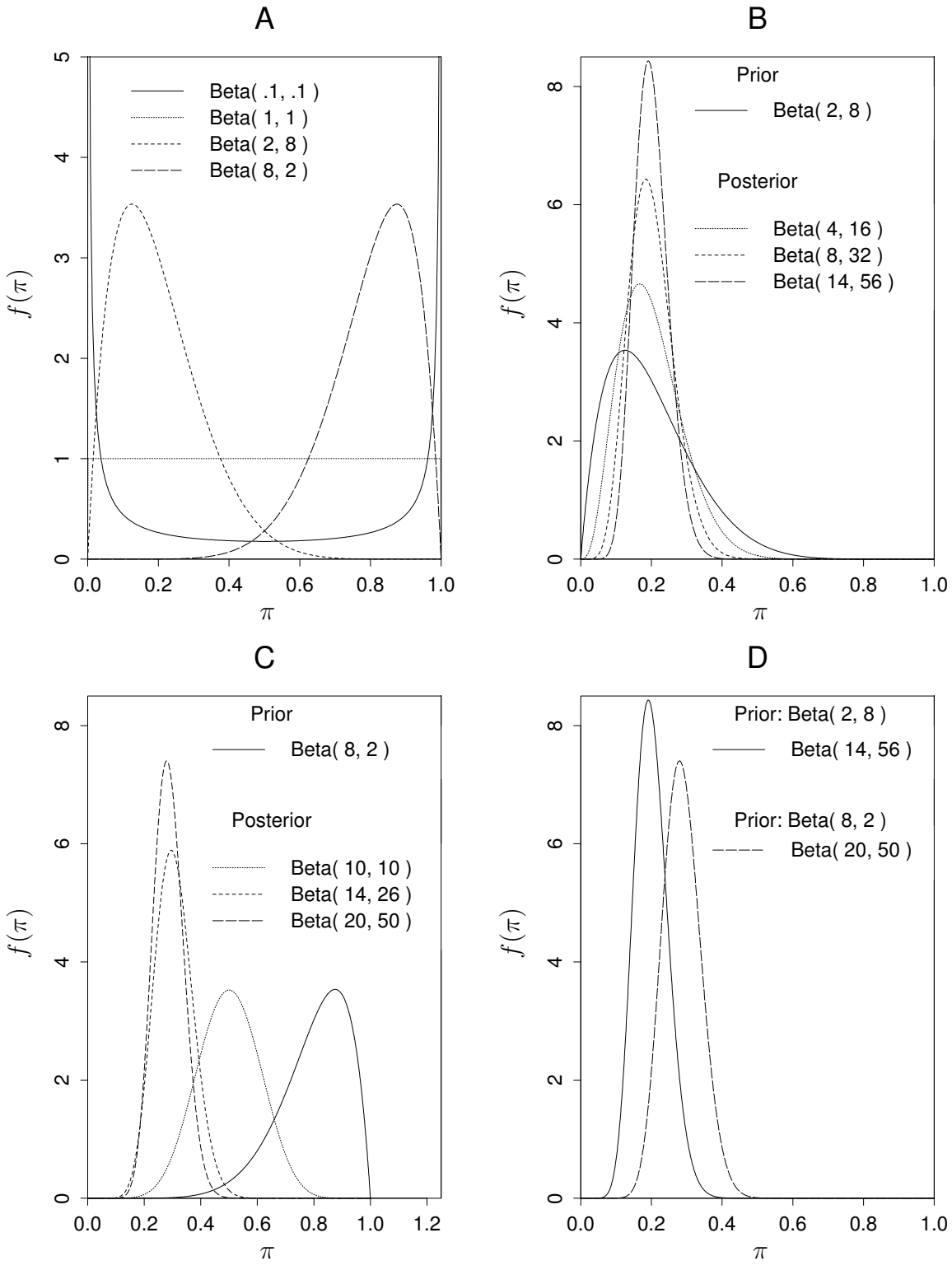

Figure 5.2: Simulation Results for Binary Model $N = 120$, $\pi_S = 0.30$

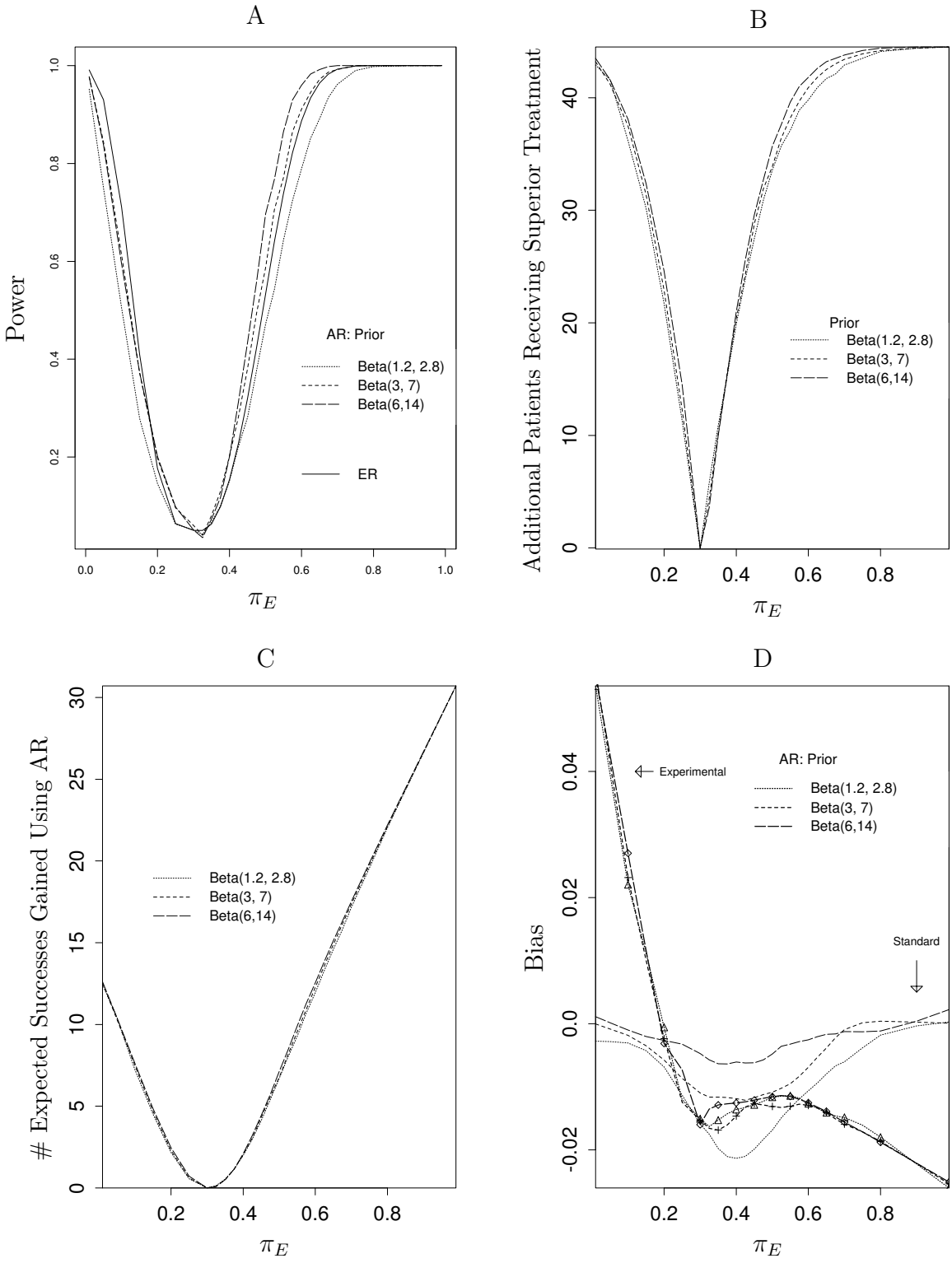

Figure 5.3: Simulation Results for Binary Model $N = 120$, $\pi_S = 0.15$

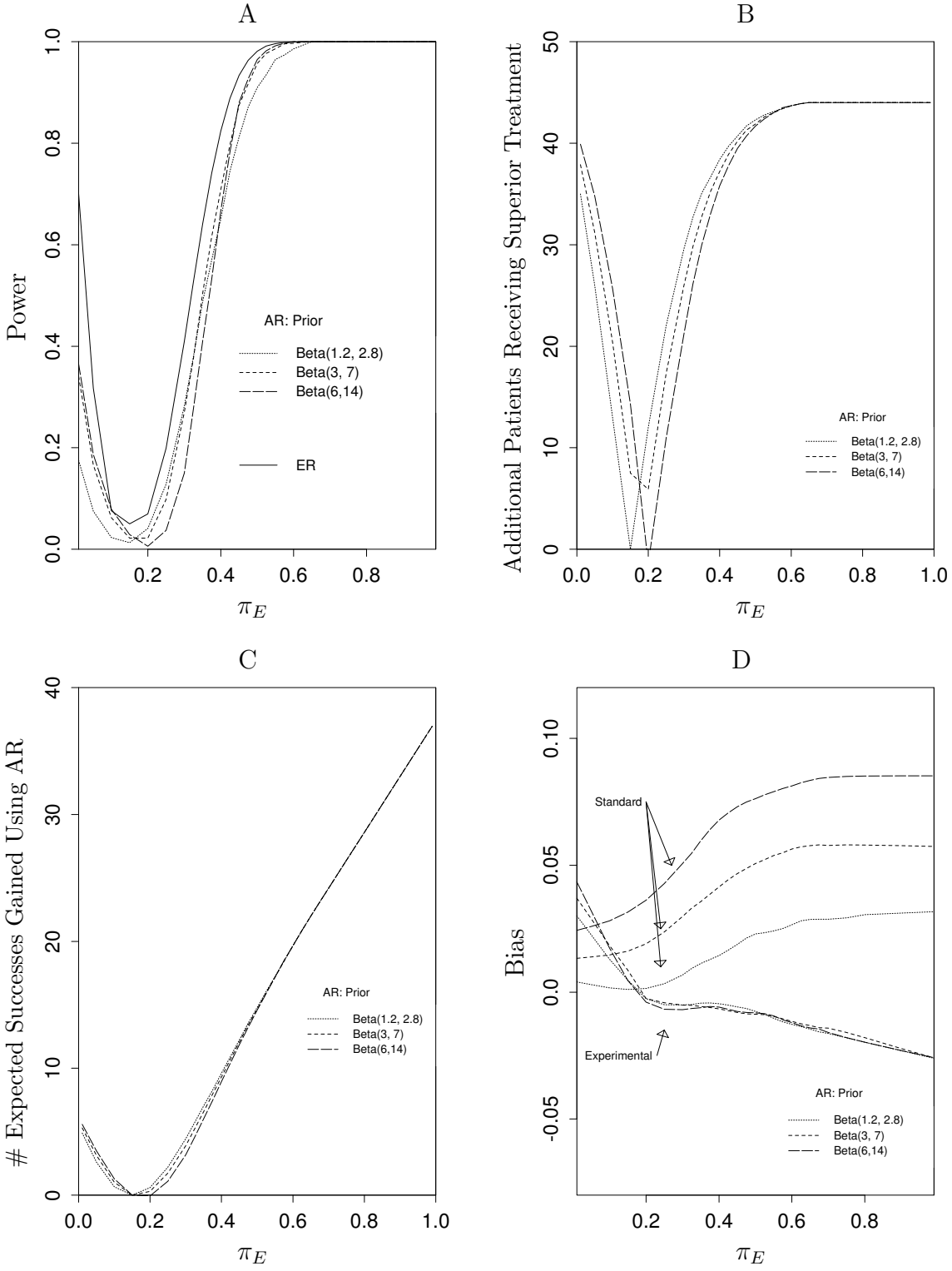

Figure 5.4: Simulation Results for Model 1 $N = 120$, $\pi_S = 0.45$

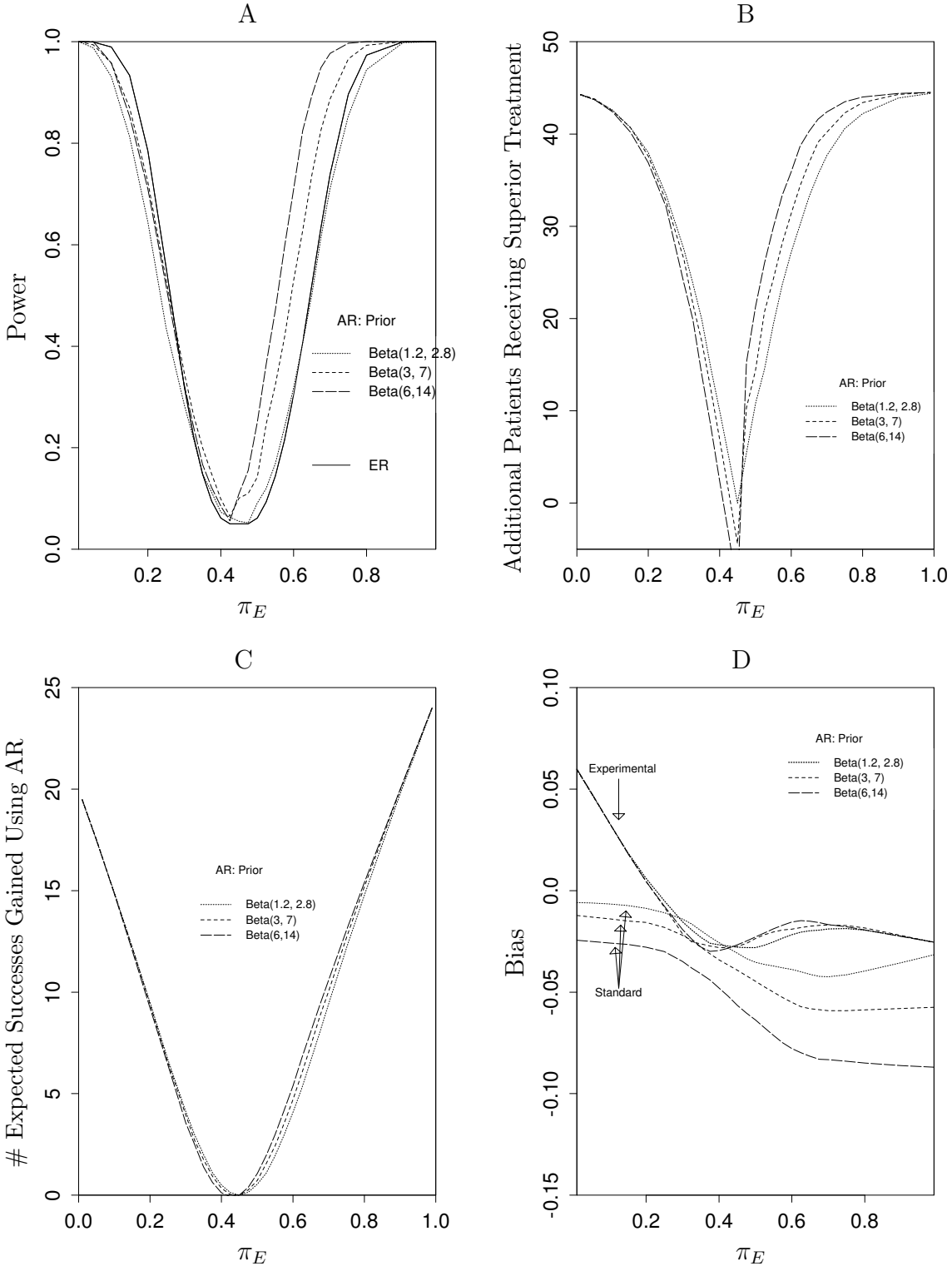

Figure 5.5: Effects of a covariate that is not accounted for in the statistical model, $N = 120$. The data was simulated where a patient had a Pr(Resp.) $= \text{logit}^{-1}(-0.847 + \beta_1 Z + \beta_2 Trt)$, where Pr($Z = 1$) $= 30\%$. This corresponds to a patient who receives S having a 30%, 15% response for $Z = 0, Z = 1$, respectively if $\beta_1 = -0.887$. and 30% and 20% if $\beta_1 = -0.539$.

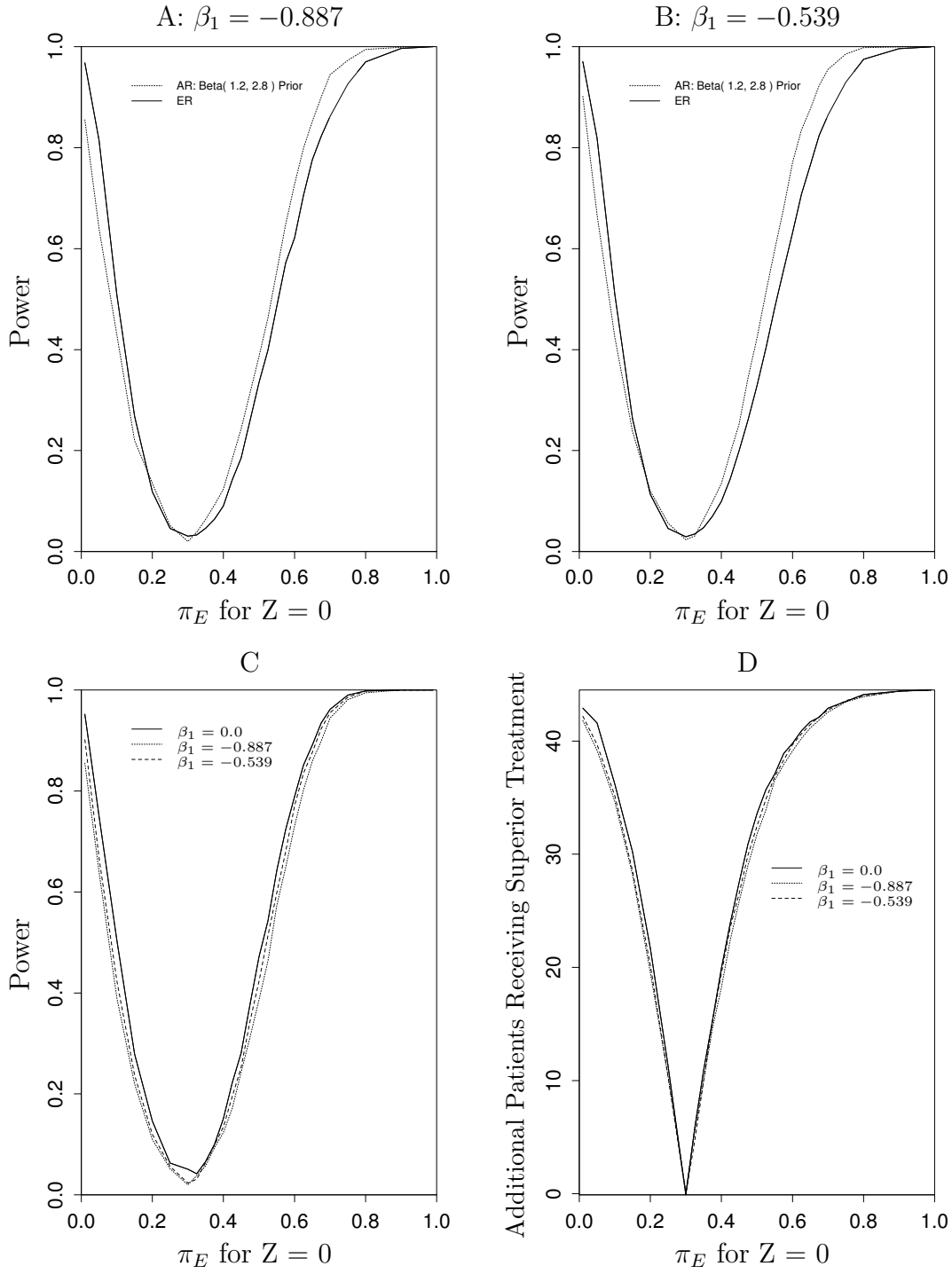

Figure 5.6: Simulation Results for Binary Model $N = 60$, $\pi_S = 0.30$

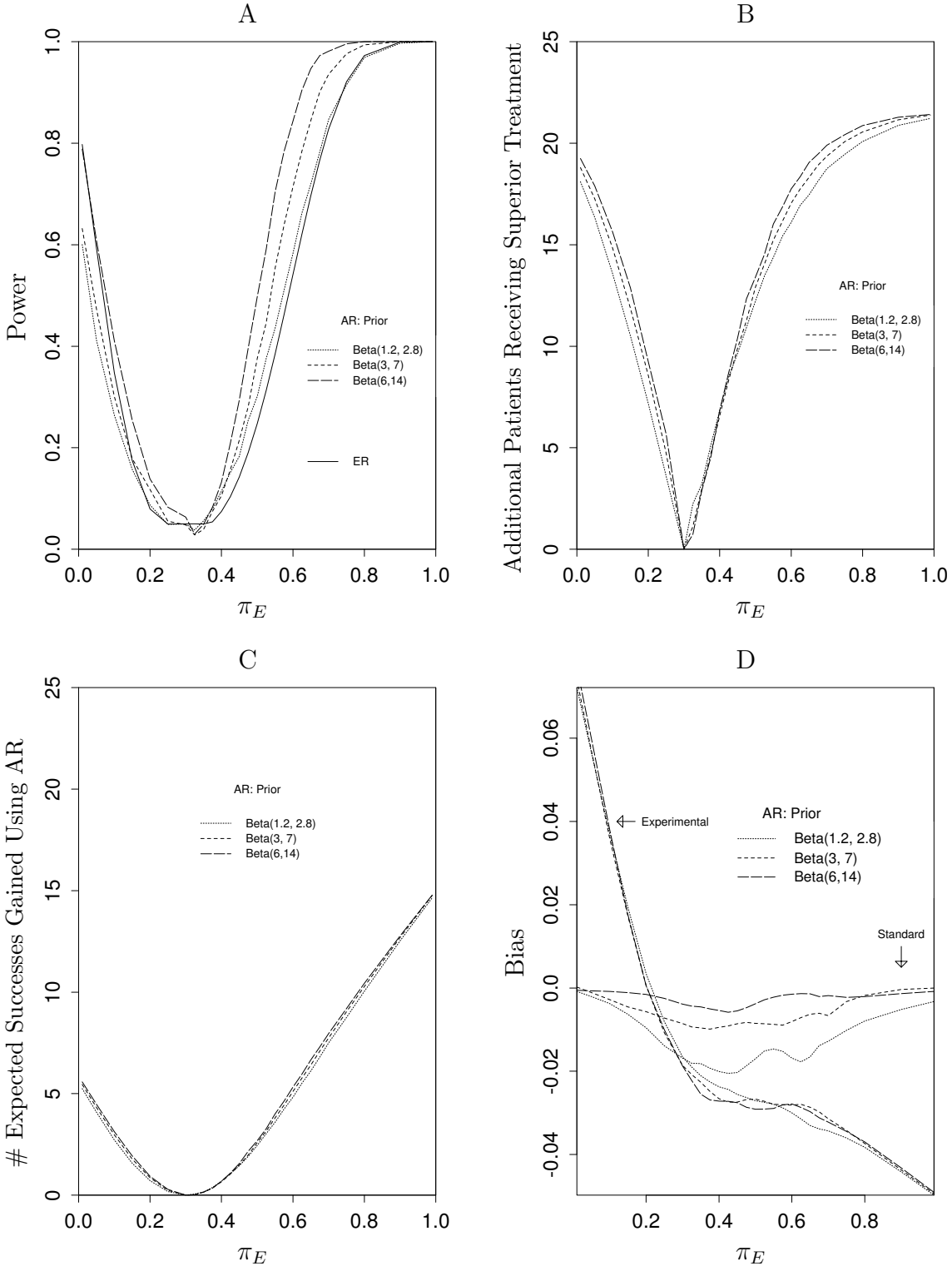

Figure 5.7: Simulation Results for Binary Model $N = 60$, $\pi_S = 0.15$

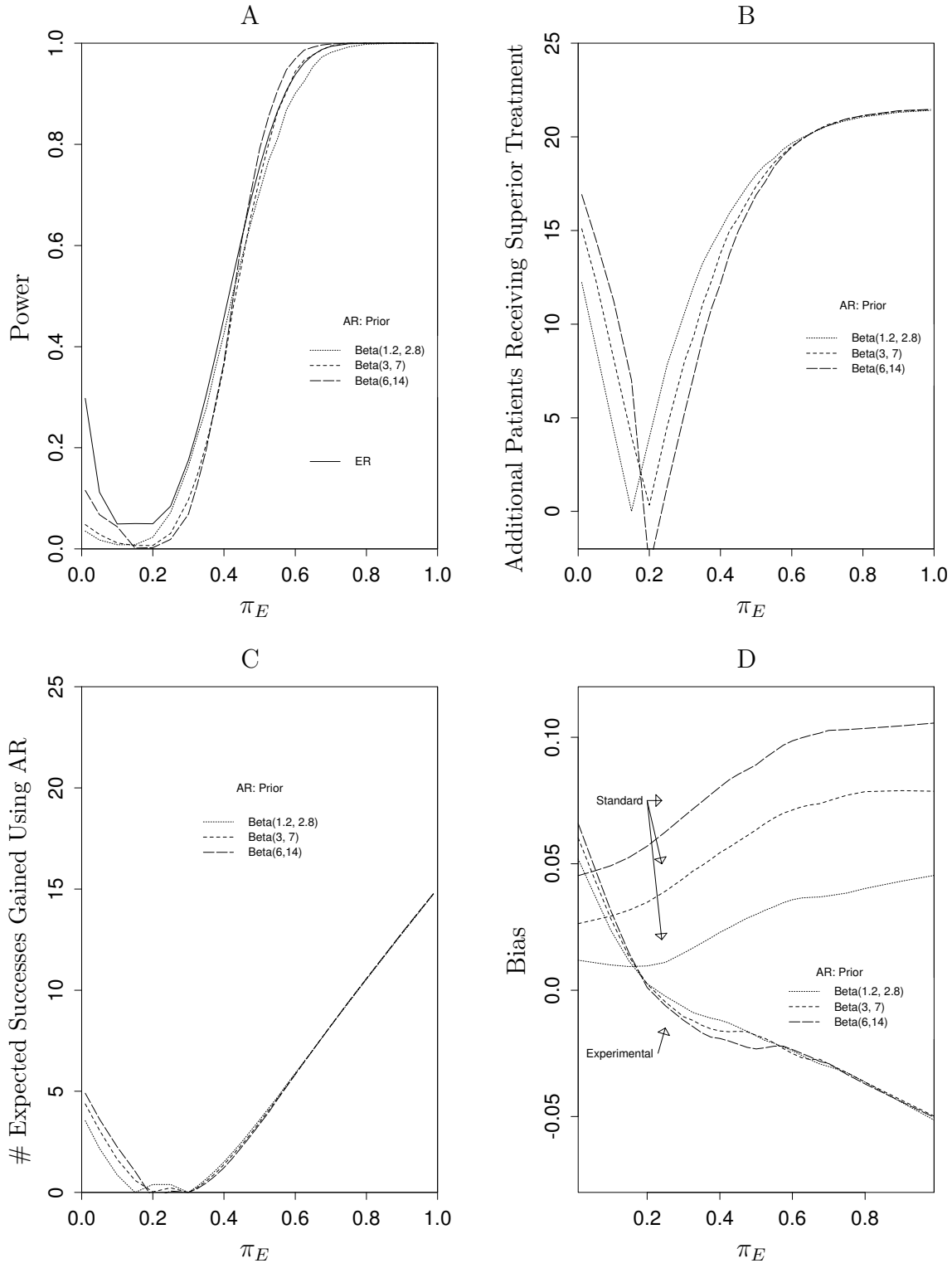

Figure 5.8: Simulation Results for Binary Model $N = 60$, $\pi_S = 0.45$

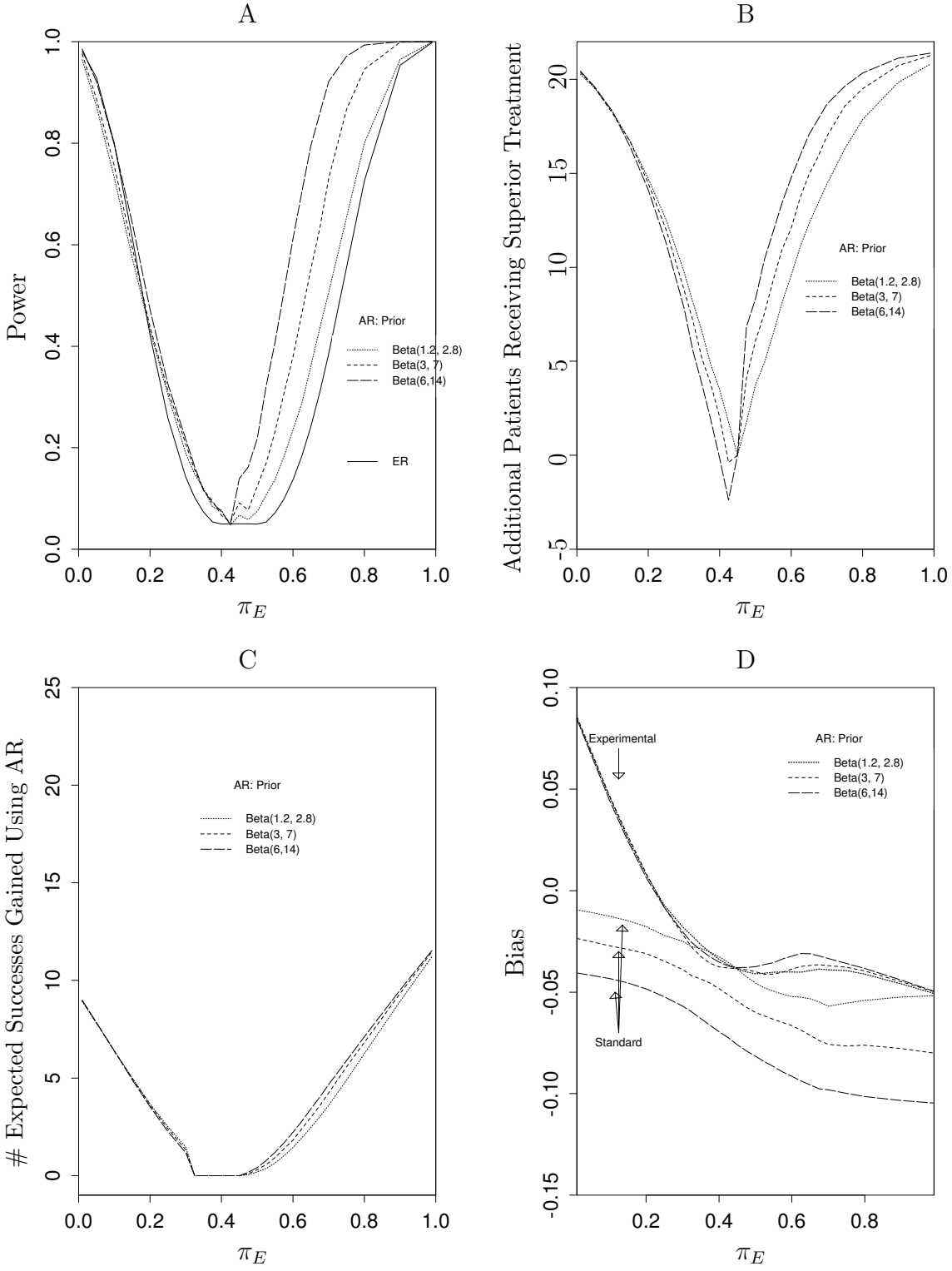

Figure 5.9: Effects of a covariate that is not accounted for in the statistical model, $N = 60$. The data was simulated where a patient had a Pr(Resp.) $= \text{logit}^{-1}(-0.847 + \beta_1 Z + \beta_2 Trt)$, where Pr($Z = 1$) $= 30\%$. This corresponds to a patient who receives S having a 30%, 15% response for $Z = 0, Z = 1$, respectively if $\beta_1 = -0.887$. and 30% and 20% if $\beta_1 = -0.539$.

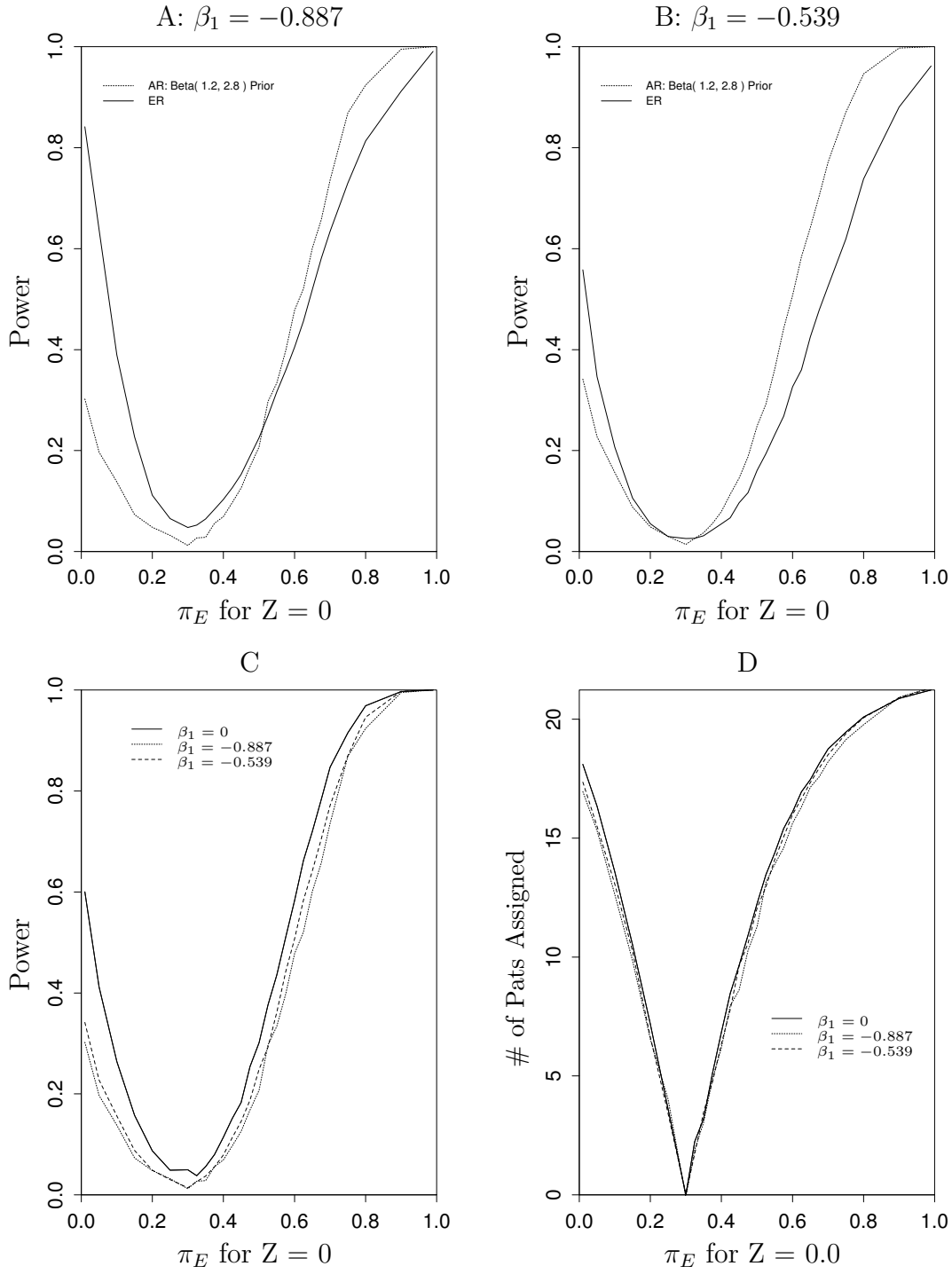

Figure 5.10: Example Priors TTE Model

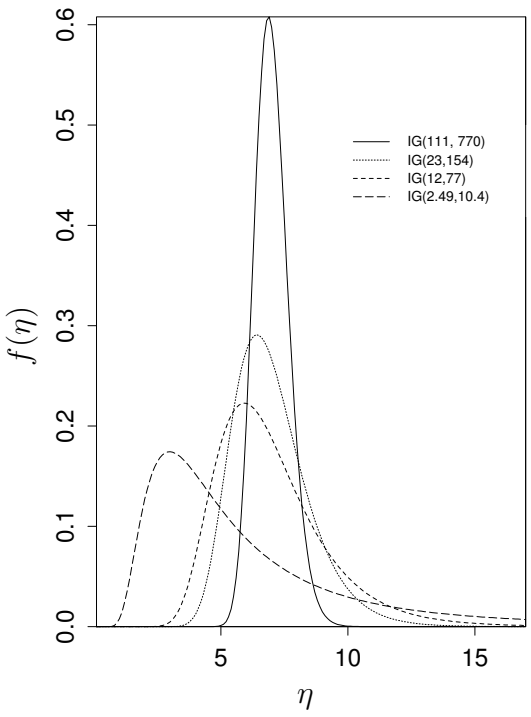

Figure 5.11: Example posteriors for various priors where the true median TTE is 10.5.

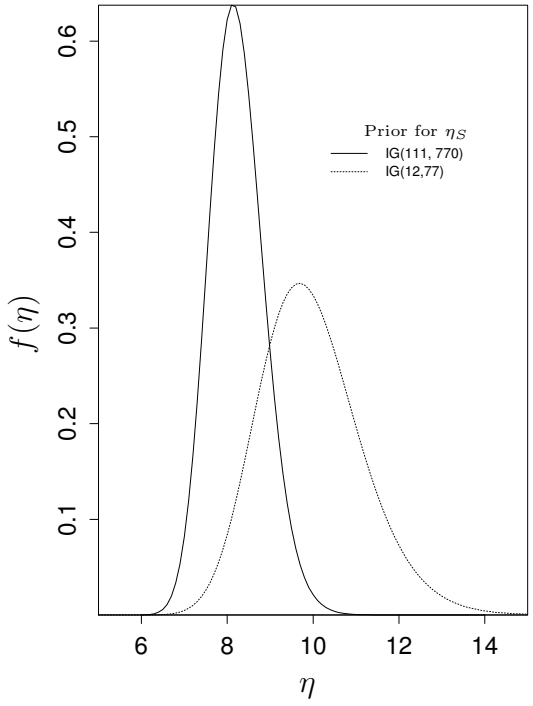

Figure 5.12: Simulation Results for TTE Model: $N = 120$, $\eta_S = 7.0$

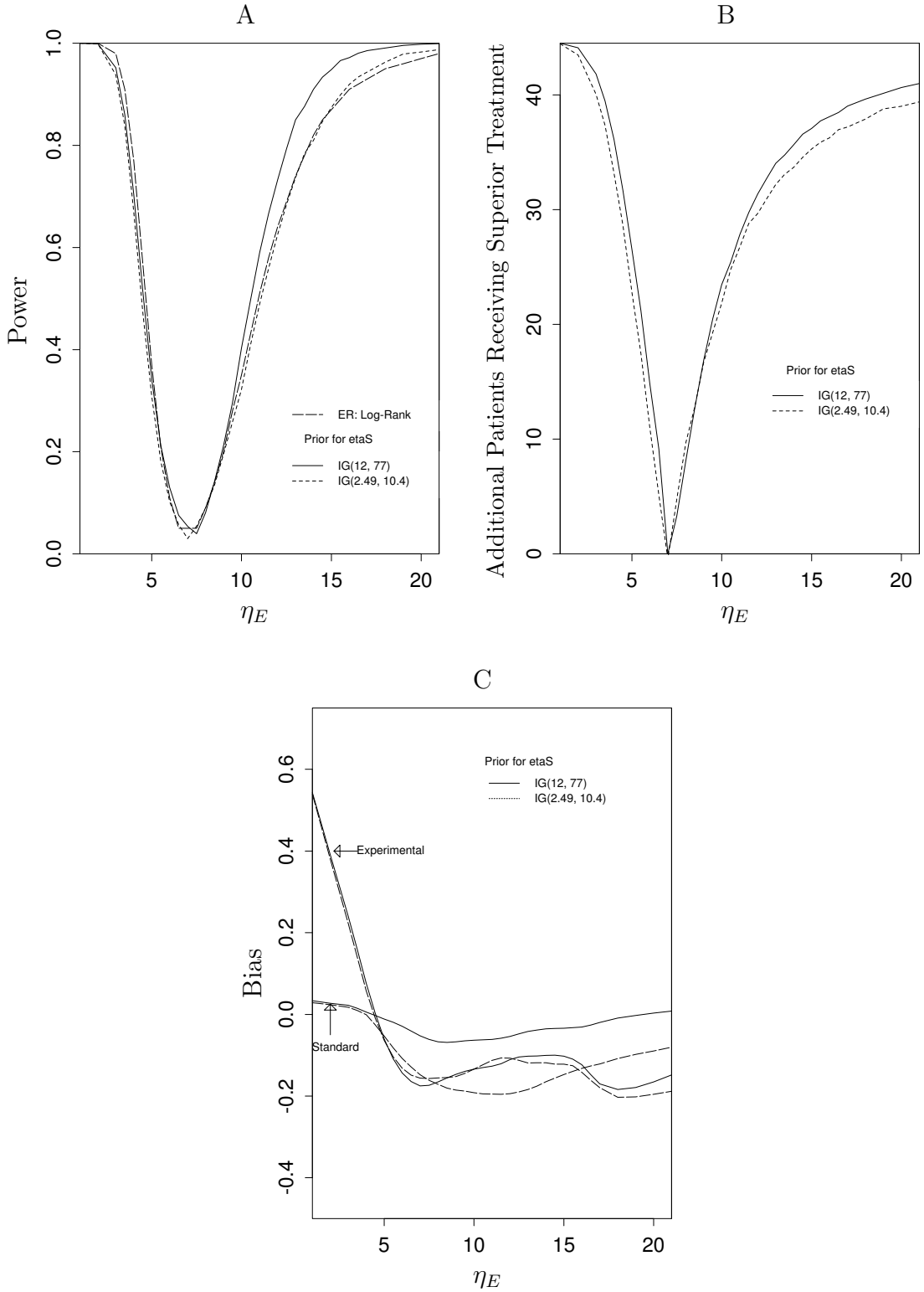

Figure 5.13: Simulation Results for TTE Model: $N = 120$, $\eta_S = 3.5$

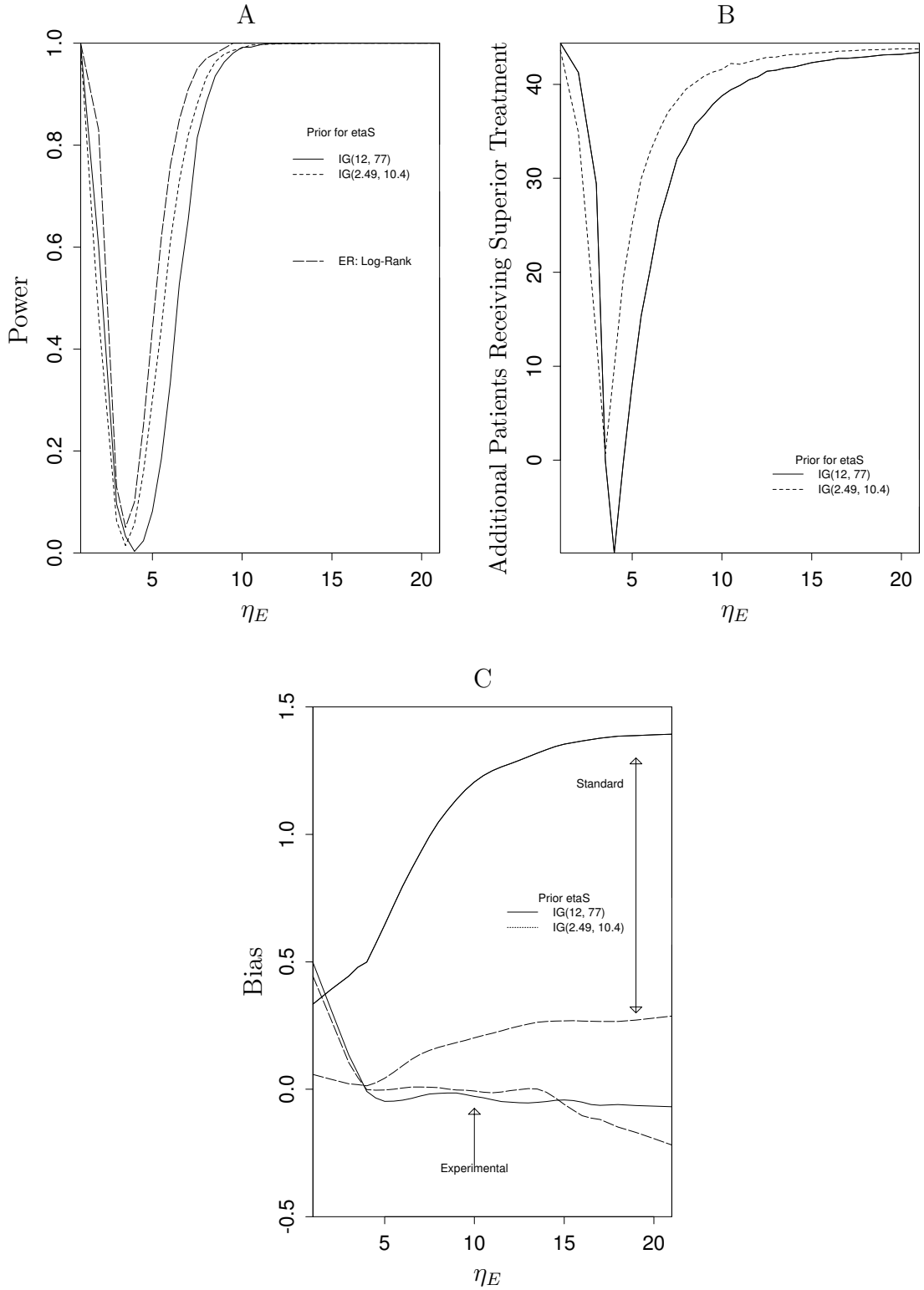

Figure 5.14: Simulation Results for TTE Model: $N = 120$, $\eta_S = 14$

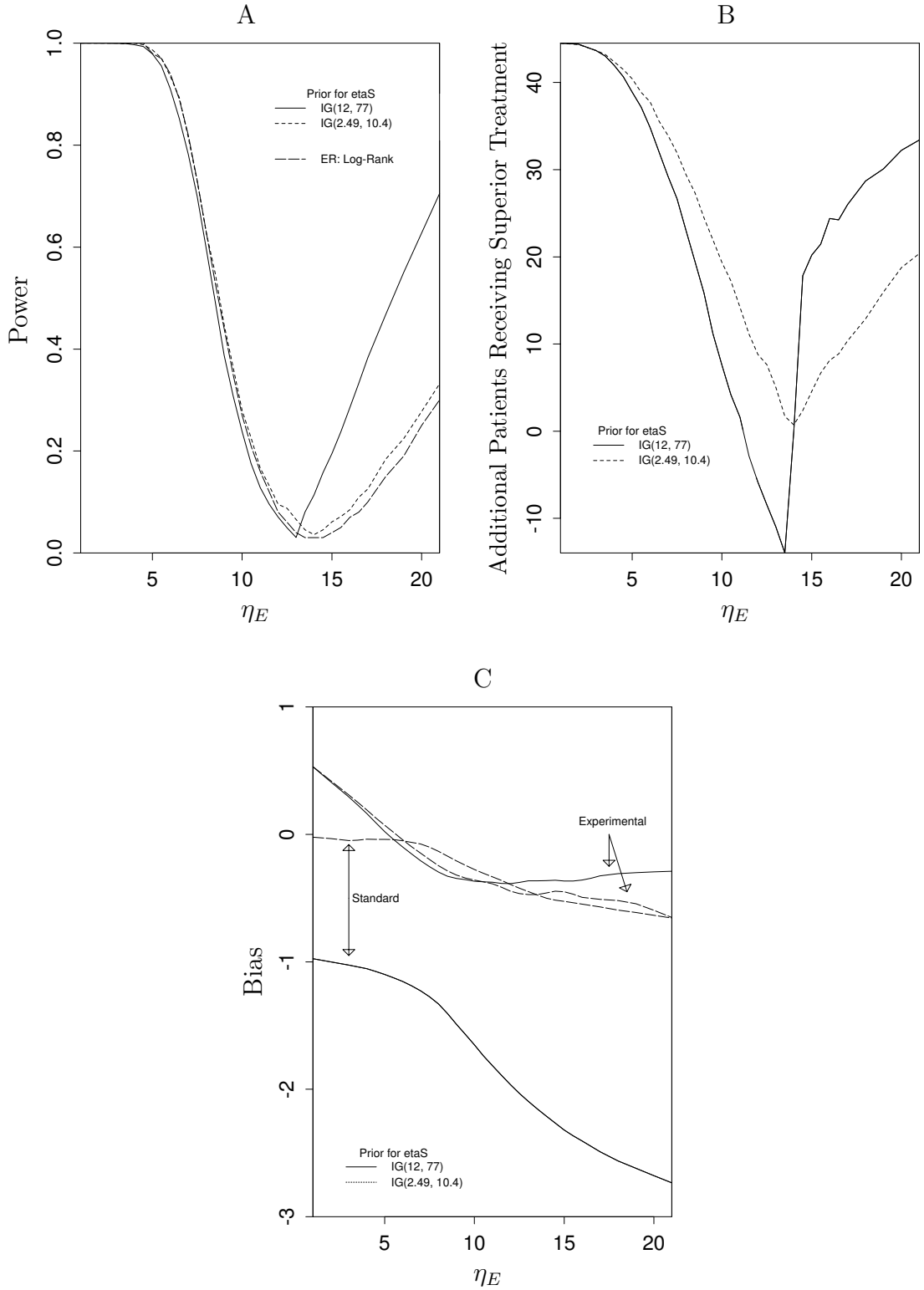

Figure 5.15: TTE Model: Effect of a covariate that was not accounted for in the statistical model, $N = 120$. A patient's TTE was simulated from an exponential$(\eta_j(1 + \beta Z))$ where $\beta = -0.5$ and $Pr(Z = 1) = 0.30$ and $\eta_S = 7$.

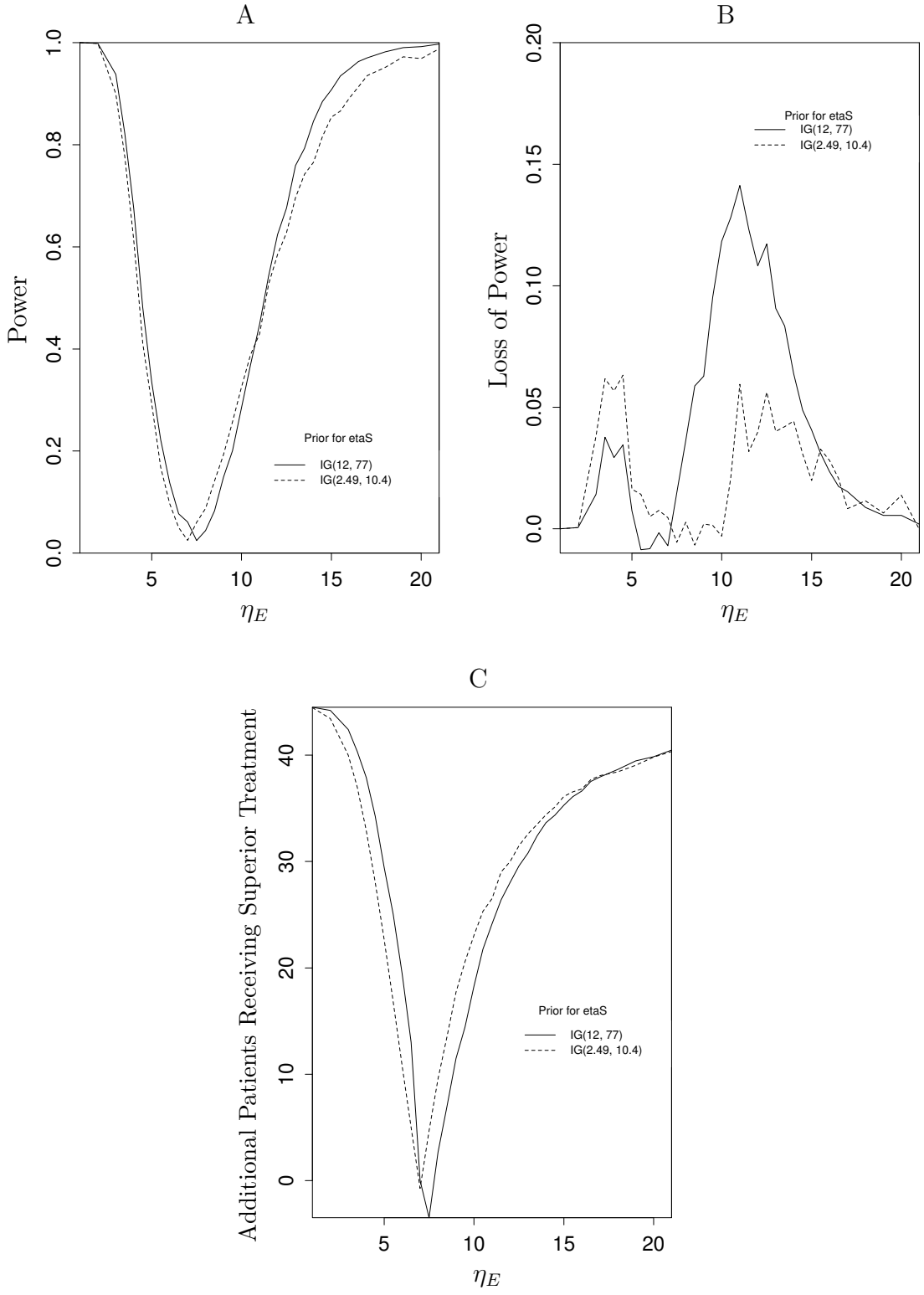

Figure 5.16: TTE Model: Effect of a covariate that was not accounted for in the statistical model, $N = 120$. A patient's TTE was simulated from an Exponential$(\eta_j(1 + \beta Z))$ where $\beta = -0.75$ and $Pr(Z = 1) = 0.30$ and $\eta_S = 7$

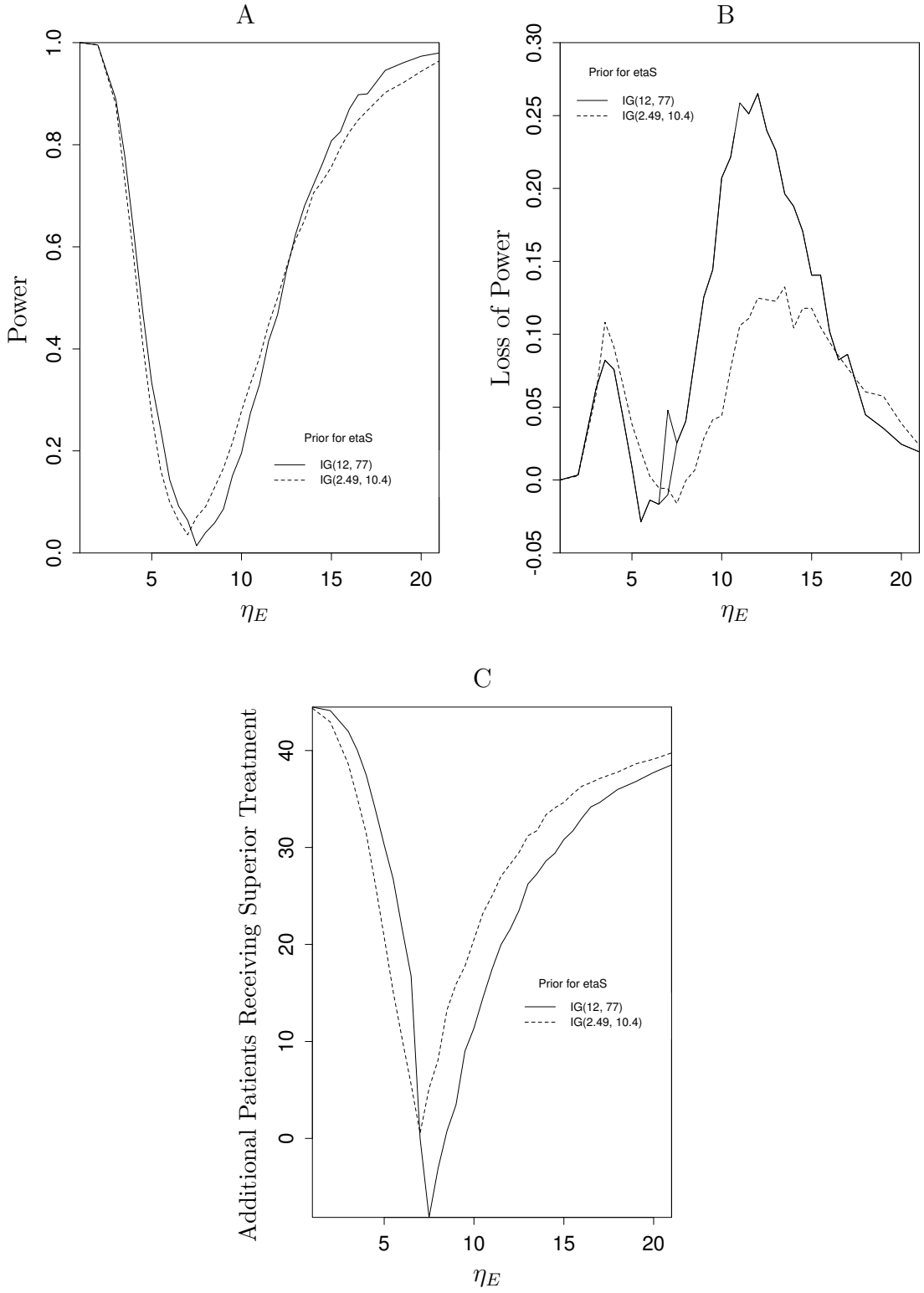

Figure 5.17: Simulation Results for TTE Model: $N = 60$, $\eta_S = 7.0$

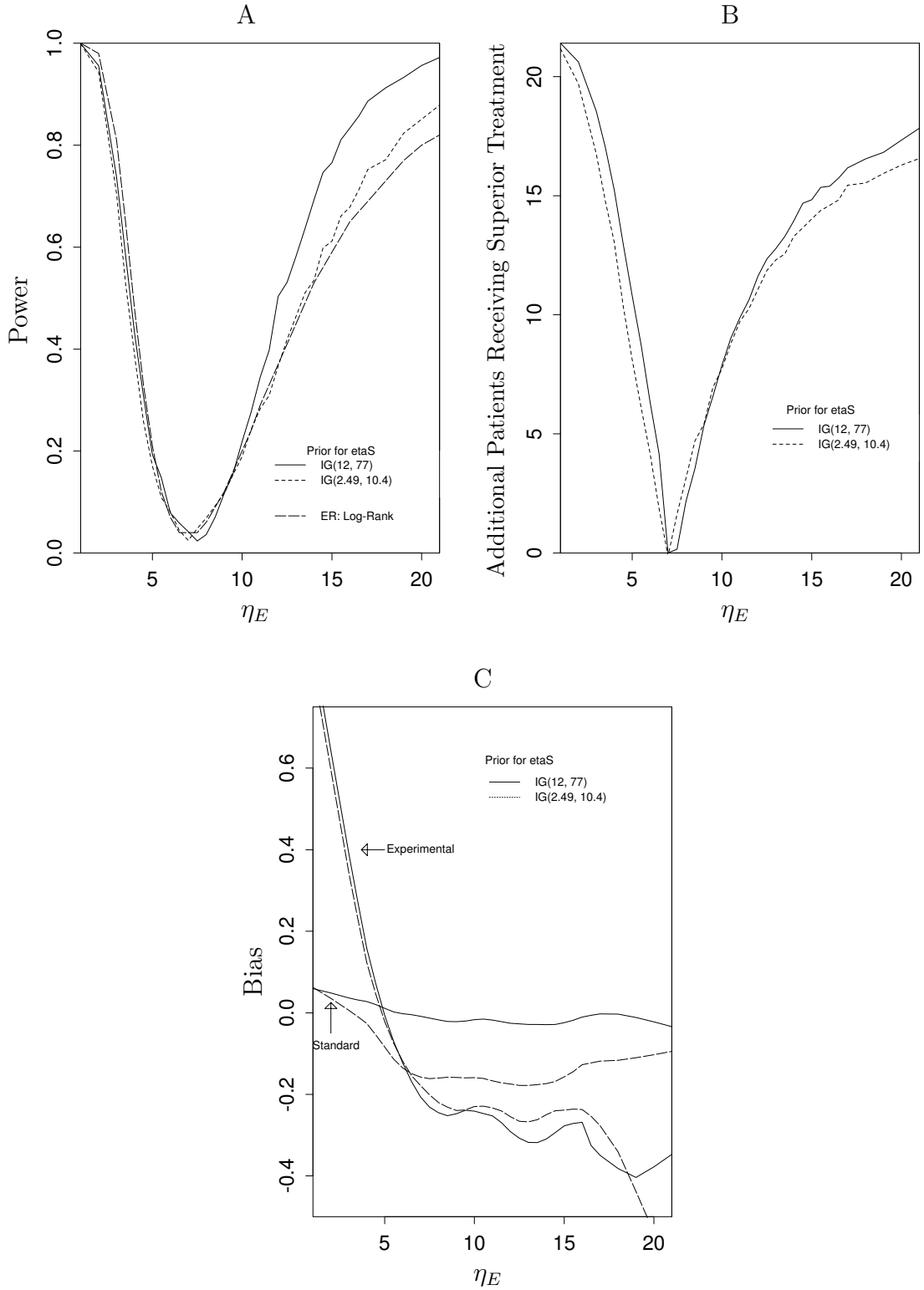

Figure 5.18: Simulation Results for TTE Model $N = 60$, $\eta_S = 3.5$

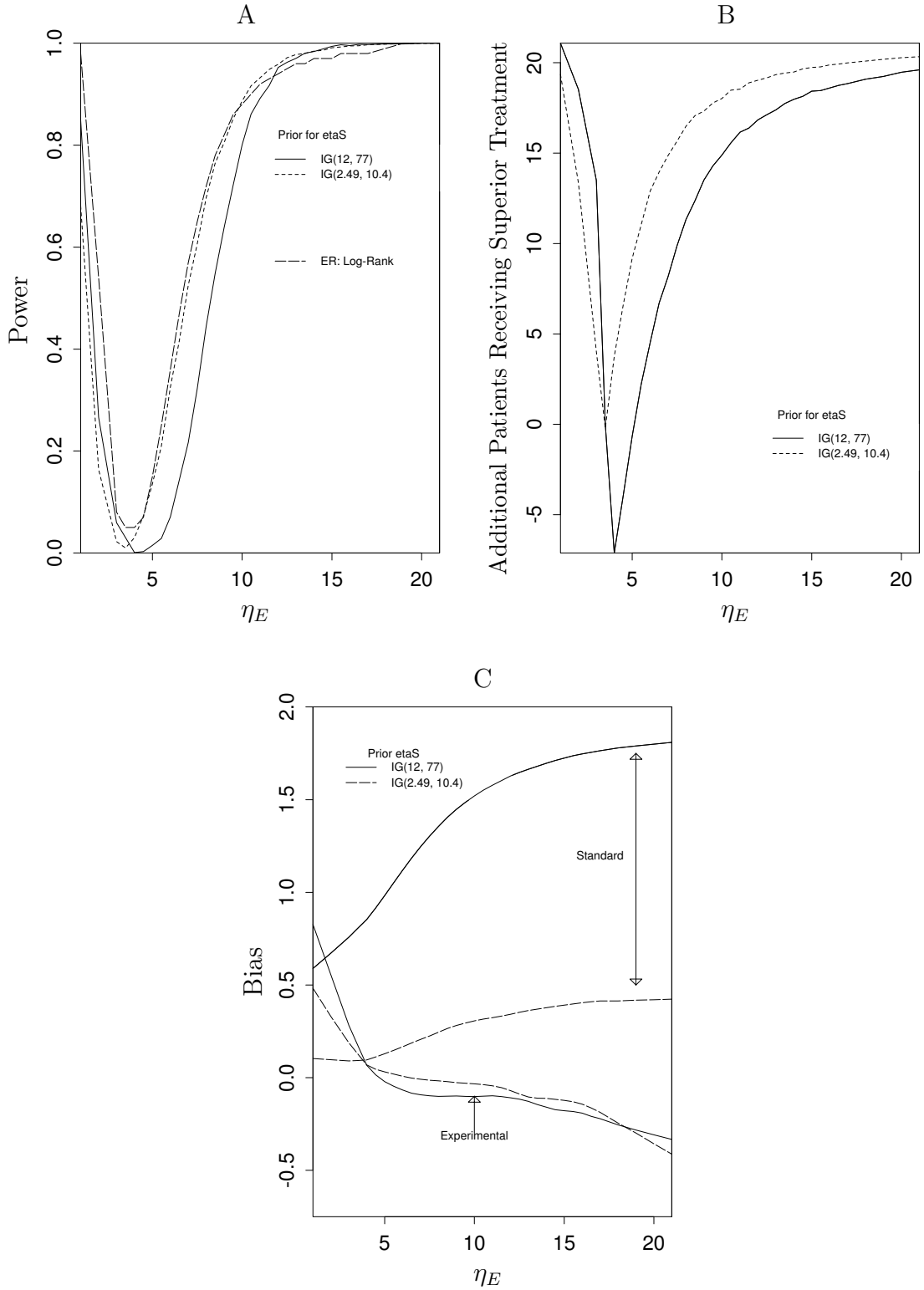

Figure 5.19: Simulation Results for TTE Model: $N = 60$, $\eta_S = 14$

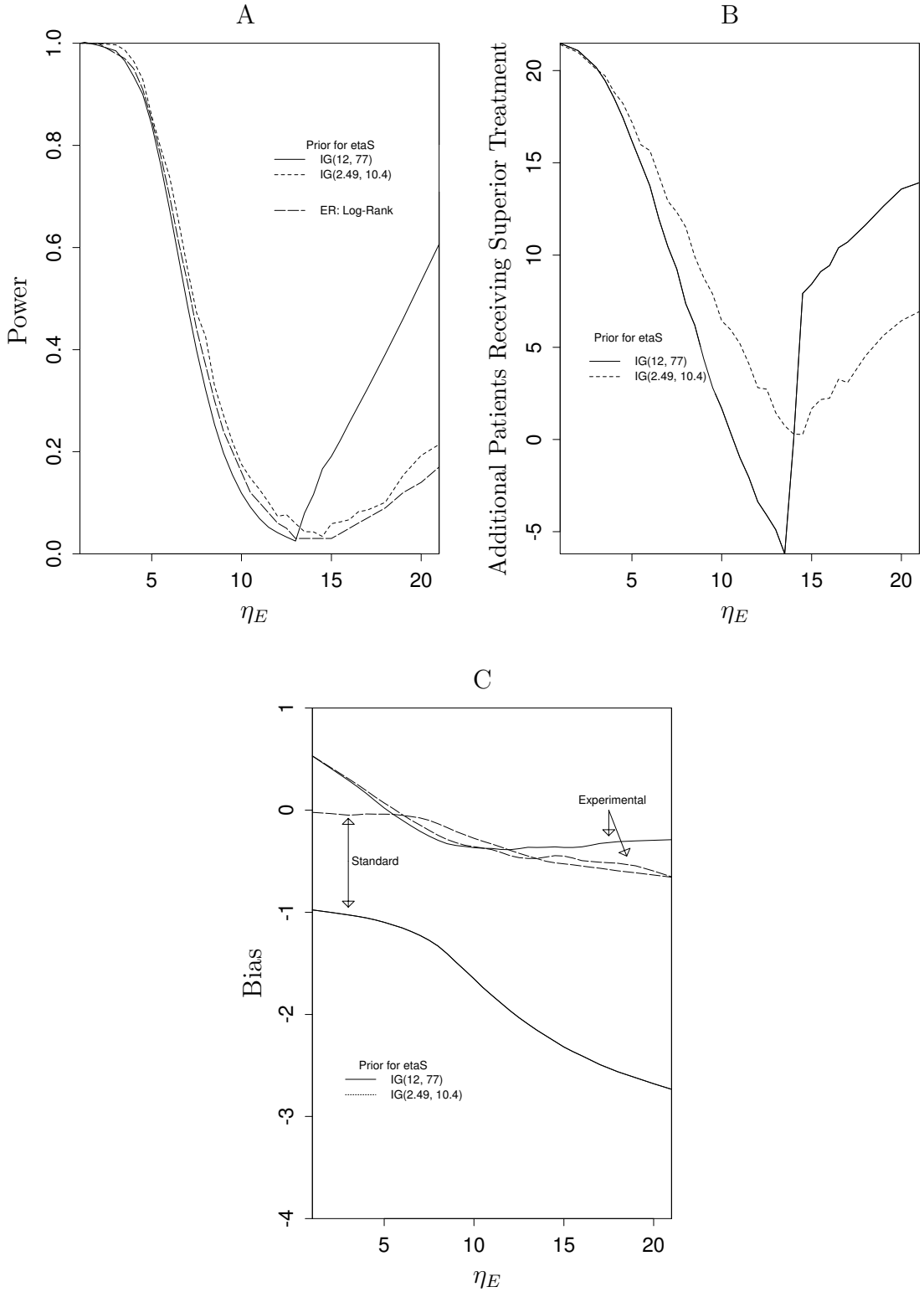

Figure 5.20: TTE Model: Effect of a covariate that was not accounted for in the statistical model, $N = 60$. A patient's TTE was simulated from an exponential$(\eta_j(1 + \beta Z))$ where $\beta = -0.5$ and $Pr(Z = 1) = 0.30$ and $\eta_S = 7$.

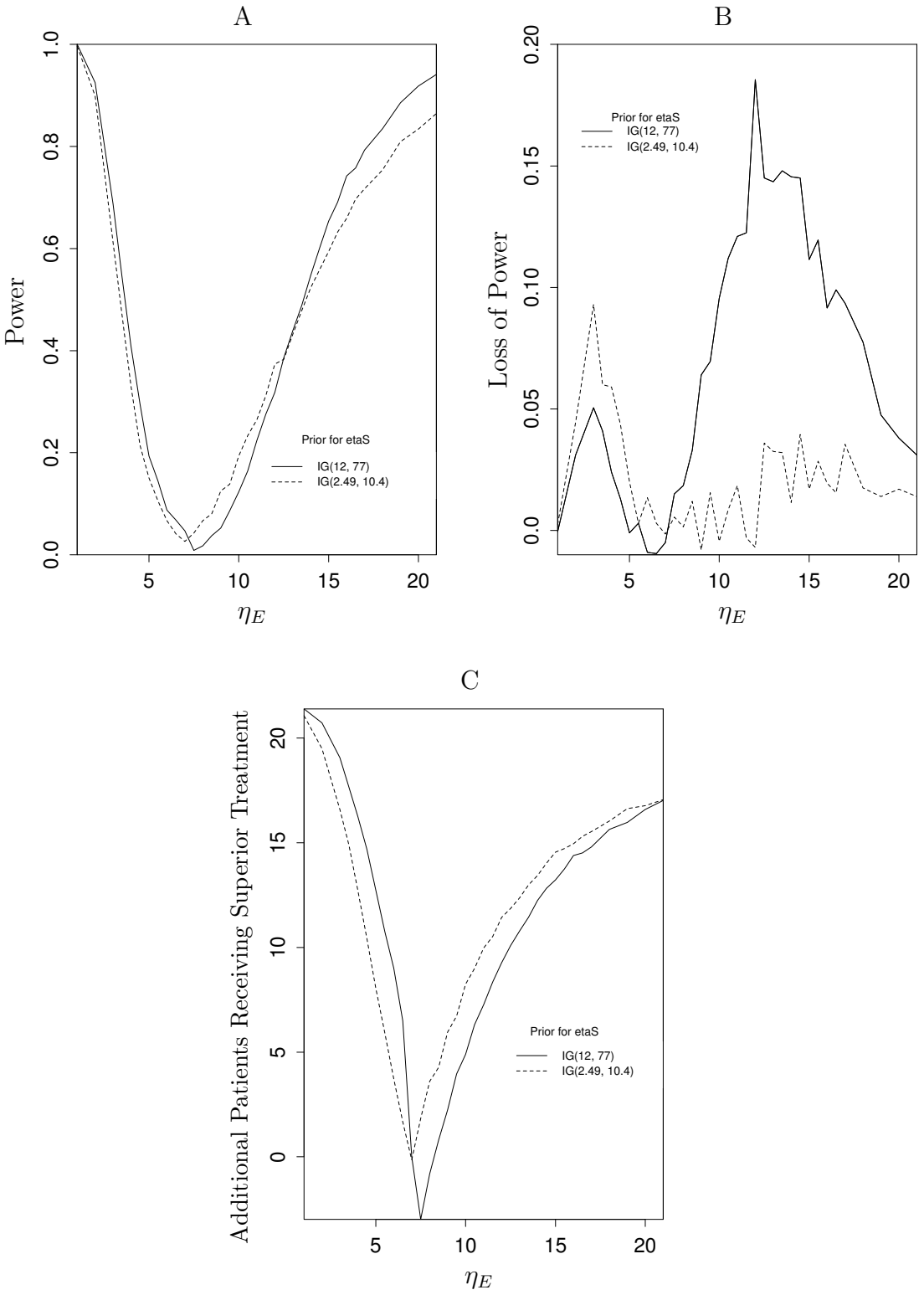

Figure 5.21: TTE Model: Effect of a covariate that was not accounted for in the statistical model, $N = 60$. A patient's TTE was simulated from an Exponential$(\eta_j(1 + \beta Z))$ where $\beta = -0.75$ and $Pr(Z = 1) = 0.30$ and $\eta_S = 7$