Multi-Course Treatment Strategies for Clinical Trials of Rapidly Fatal Diseases

Peter F. Thall,[1] Hsi–Guang Sung[2] and Elihu H. Estey[3]

[1,2] *Department of Biostatistics, Box 447*

[3] *Department of Leukemia, Box 428*

M.D. Anderson Cancer Center

1515 Holcombe Boulevard

Houston, Texas 77030, U.S.A.

[1]*E–mail rex@mdanderson.org*

May 4, 2001

SUMMARY

Therapy of rapidly fatal diseases often requires multiple courses of treatment. In each course, the treatment may achieve the desired clinical goal, “response,” the patient may survive without response, “failure,” or the patient may die. When treatment fails in a given course, it is common medical practice to switch to a different treatment for the next course. Most statistical approaches to such settings simply ignore the multi-course structure. They characterize patient outcome as a single binary variable, combine death and failure, and identify only one treatment for each patient. Such approaches waste important information. We provide a statistical framework, based on a family of generalized logistic regression models, that incorporates historical data while accommodating multiple treatment courses, a trinary outcome in each course, and patient prognostic covariates. The framework serves as a basis for both data analysis and outcome–adaptive clinical trial conduct. Rather than focusing on individual treatments, we evaluate multi–course treatment strategies that specify which treatment to give in each course within each prognostic subgroup. We describe a general approach for constructing clinical trial designs that may be tailored to different multi–course settings. For each prognostic subgroup, based on a real-valued function of the covariate-adjusted probabilities of response and death, the design drops inferior treatment strategies during the trial and selects the best strategy at the end. The methodology is illustrated in the context of a randomized two–course, three–treatment acute leukemia trial with two prognostic covariates. The model is first fit to an historical data set to obtain a reasonably informative prior on non-treatment related parameters for use in trial design and conduct. We describe a simulation study of the design under several clinical scenarios. The simulations show that the method can reliably identify treatment–subgroup interactions based on moderate sample sizes. Extensions of the leukemia trial design to more complex multi-course settings are discussed.

1. Introduction

Therapy of rapidly fatal diseases often requires multiple courses of treatment. The clinical goal is to achieve a "response," such as remission of leukemia, 50% shrinkage of a solid tumor, or resolution of infection. Such responses are presumed to predict longer survival. The other therapeutic outcomes are death during treatment and "failure," in which case the patient survives therapy but does not respond. Death during therapy is an unavoidable risk in oncology trials involving acute or advanced disease where only very aggressive, life–threatening treatments have any substantive anti–disease effect. Thus, in general, each treatment course results in one of three possible outcomes: response, death, or failure. When treatment fails after a given course, it is common medical practice to switch to a different treatment for the next course. We consider settings where it is reasonable to define outcome as a discrete variable observed within a time period sufficiently short that interim monitoring is feasible. Most statistical approaches to this or similar settings characterize patient outcome as a single binary variable by collapsing the multi–course structure and combining death and failure, and moreover they typically evaluate only one treatment for each patient. Such approaches waste important information, since each patient may receive several different treatments over successive courses, these treatments may have interactive effects, and the distinction between death and treatment failure is very important clinically.

In this paper, we provide a statistical framework for clinical trial design and conduct in multi-course settings. In contrast with the usual approach of evaluating individual treatments, we propose an outcome–adaptive, multi–course treatment design that specifies which treatment to give in each course in each of several prognostic subgroups. We take a Bayesian approach, which provides a natural basis for incorporating historical data and making inferences sequentially during the trial and upon its completion. The methodology is presented in the context of the following chemotherapy trial, which motivated this research.

The trial involves acute myelogenous leukemia (AML) patients who previously achieved a complete remission (CR), the criteria for which are normal blood counts and a normal-appearing bone marrow, but who subsequently relapsed in less than 24 months. For these patients, the principal covariates predicting success are the duration of the first CR, denoted D_{CR}, and age at diagnosis. Patients with initial D_{CR} less than one year are unlikely to achieve a second CR with standard therapy and thus are candidates for investigational regimens. In contrast, controversy exists among physicians regarding management of patients with longer D_{CR}. While some physicians believe that their prognoses are sufficiently poor that only investigational therapies should be tried, others believe that these same prognoses justify initial use of standard therapy. This controversy motivates this trial.

Each patient receives either one or two courses of chemotherapy. The three possible outcomes for each course, determined within one month from initiation of that course's treatment, are CR, death, or failure, the event that the patient is alive but has not achieved a CR. The occurrence of either death or CR, or the completion of two courses that are both failures, marks the end of the patient's therapy. Patients who fail two courses are given palliative care subsequently. This definition of therapeutic outcome is motivated by the necessity of CR for long-term survival in AML and the very low probability of a subsequent CR once failure has occurred in each of two courses. The trial includes the standard chemotherapy combination idarubicin + high–dose cytosine arabinoside (IDA), and two experimental treatments, IDA + mylotarg (M) and IDA + topotecan (T). For the first course, all patients are randomized fairly among the three treatments. A patient for whom IDA fails in the first course is randomized between IDA+M and IDA+T for the second course. A patient for whom either IDA+M or IDA+T fails in the first course must receive IDA in the second course, however, because it was considered unacceptable to give a patient experimental treatments in both courses. Figure 1 illustrates this treatment assignment algorithm.

[Figure 1 about here]

The primary scientific goal of the trial is to select the best two–course treatment strategy within each prognostic subgroup based on the probabilities of CR and death. Trial conduct is outcome-adaptive in that, if interim data show a particular treatment strategy to be substantially inferior to at least one other strategy within a subgroup, then the inferior strategy is dropped within that subgroup. The design consists of an algorithm for assigning a treatment to each patient in each course, the above interim safety monitoring rules and, at the end of the trial, treatment strategy selection within prognostic subgroups. The method requires a real–valued objective function of the probabilities of CR and death that quantifies the clinically acceptable trade–off between these two outcomes. This function is used as a basis for interim decision–making and inferences at the conclusion of the trial. In practice, the trade–off function is elicited from the physicians planning the trial, and we illustrate how this may be done using contour plots as a graphical aid.

We employ a generalized logistic regression model to characterize the probabilities of CR and death in each course as functions of the patient's treatments and prognostic covariates. The model also allows pairwise interactions between treatment strategy, course, and covariates. Because the probabilities of CR and death vary greatly with patient prognosis, different prognostic subgroups may have different optimal treatment strategies. In addition to its application in the context of trial design and conduct, the regression model is also a useful analytic tool for evaluating covariate and treatment strategy effects on the probabilities of response and death based on existing data. While the overall trial sample size is 96 patients, the numbers of patients in the subgroups determined by the various combinations of treatment strategies and patient covariates may be quite small. Consequently, the ability of the model to borrow strength across these subgroups is essential, and we will show that a non-model based approach with this sample size is simply not feasible.

As the first step in developing a design, we fit the model to historical data arising from 714 AML patients treated at M.D. Anderson Cancer Center between 1990 and 1999. This analysis served to validate the model, obtain informative distributions for model parameters unrelated to treatment, and also obtain reasonable numerical values of parameters for use in a simulation study of the design. Like the patients in the trial being planned, each historical patient previously achieved CR but later relapsed and then received one or two courses of salvage therapy in an attempt to re–induce remission. The salvage treatments were an allogeneic bone marrow transplant, combination chemotherapy containing high dose cytosine arabinoside ("ara–C"), or chemotherapy not including ara–C. The data for each patient consisted of prognostic covariates and the treatment and outcome in each of one or two courses. A summary of the empirical outcome probabilities in each course, ignoring prognostic covariates, is given in Table 1.

[Table 1 about here]

The remainder of the paper is organized as follows. In Section 2, we explain current medical practice for diagnosis and treatment of patients with AML, as well as the type of clinical trial designs currently used in most medical centers. Section 3 describes a general Bayesian strategy for using historical data in constructing a clinical trial design, and how we will apply this strategy in designing the AML trial. In Sections 4 and 5 we present the probability model and numerical methods that will serve as the basis for treatment evaluation and trial design. Section 6 describes the objective function that we used to combine the probabilities of CR and death. Section 7 summarizes our analysis of the historical data. The AML trial design is described in Section 8. In Section 9 we summarize a simulation study of the design's operating characteristics (OCs), and we present some graphical methods that may be used to evaluate and compare the effects of various combinations of treatment, course, and covariates on the outcome probabilities. We close in Section 10 with descriptions of extensions that deal with other multi-course settings.

2. Developmental Therapeutics in Acute Leukemia

AML is a disorder of blood cell formation, "hematopoiesis." This process occurs in the bone marrow, with the cells subsequently released into the blood. The cells of interest are red cells, white cells, and platelets. Red cells carry oxygen, white cells prevent or ameliorate infection, and platelets prevent or minimize bleeding. Normally, the bone marrow's rate of production of each type of blood cell equals its rate of loss from the blood. Each type of cell arises from an immature cell, a "stem cell" or "blast," with these immature cells maturing in the marrow prior to entry into the blood. AML is characterized by the presence of abnormal immature hematopoietic cells. A hallmark of this abnormality is the inability to mature. A second feature is the capacity to inhibit normal hematopoiesis. The result is a decrease in the numbers of normal blood cells. Such “bone marrow failure” leads to death just as would failure of other vital organs such as the heart or lungs. The diagnosis of AML is straightforward. Examination of the bone marrow, generally performed because of symptoms such as fatigue together with the finding of low blood counts, shows an increased proportion of immature cells. Reflecting the failure of maturation, this increase is *prima facie* evidence of the abnormal nature of the blasts. In a variable percentage of cases, the blasts demonstrate chromosomal abnormalities that provide further support for the diagnosis.

Prior to the mid 1960s, treatment for AML essentially consisted of transfusions of red cells and platelets and the use of antibiotics for infections. The median survival was approximately six months, with survival beyond one year very unlikely. Since about 1970, the great majority of patients have been given drugs, “chemotherapy,” intended to kill AML cells but not normal blasts. During the past 30 years, there have been well-documented improvements in both transfusion practices and in antibiotics. Thus, it is difficult to determine whether the prognosis of patients not given chemotherapy has changed since the 1960s. Clearly, however, there are AML patients who can live as long as two years with untreated disease. Such patients are discovered when a bone marrow sample not considered to show AML when initially examined is subsequently re-examined several years later and found to indeed be diagnostic of AML. Given the risk of early death inherent in the administration of anti-AML chemotherapy, it would be highly desirable to identify patients with such prognoses in the absence of specific treatment. The number of untreated patients is insufficient for this purpose, and of course it not ethically possible to randomize patients between chemotherapy and no-treatment in order to obtain an unbiased estimator of the effect of chemotherapy. It is clear, however, that patients who present with high white blood cell counts are very unlikely to have extended survival without treatment.

There is great prognostic heterogeneity in patients given current anti-AML chemotherapy. This usually consists of two drugs: idarubicin (or equivalently daunorubicin) and ara-C. We will refer to such therapy as “standard.” The first objective of therapy is to produce and maintain a CR. While the probability of CR varies substantially as a function of prognostic covariates, patients who achieve a CR live longer than patients who do not, with the survival difference between the two groups almost entirely due to the time spent in remission. Furthermore, only patients in whom CR occurs are potentially cured, with this term applicable to patients whose disease remains in CR for approximately three years from CR date, after which time the risk of relapse declines sharply.

Response to standard chemotherapy for newly-diagnosed AML is so variable that to speak of an average outcome is potentially quite misleading. Rather, we should speak of a series of outcomes whose likelihoods are dictated by a set of well-described prognostic variables. Failure of chemotherapy can result either because the therapy leads to death or because the therapy is ineffective. Therapy-induced death is a direct consequence of the lack of selectivity of current chemotherapy. Simply put, normal cells, in particular normal bone marrow blasts, are only slightly less vulnerable to chemotherapy than are AML blasts. The life-threatening toxicity of standard chemotherapy is bone marrow failure, with death resulting from infection and hemorrhage as in untreated AML. Rates of therapy-induced mortality increase with increasing age, abnormal organ function and, particularly, poor performance status (PS). Table 2 summarizes the induction chemotherapy mortality rate for AML patients treated at M.D. Anderson during the period 1991 - 1999. For example, an ambulatory ($PS \leq 2$) adult aged under 50 would be expected to have a 28-day mortality rate of about 5%, while this rate jumps to about 39% if the patient has the same age but is bedridden ($PS \geq 3$). Table 2 also shows the important effect of age, as both the 28-day and the 56-day death rates increase monotonically with age within each performance status group. Comparison of the 28-day and 56-day death rates shows the speed with which AML patients die during the first eight weeks of therapy.

[Table 2 about here]

Except in bedridden patients over age 50, the primary cause of failure with standard treatment has been resistant AML. This can be manifested as failure to achieve initial CR despite a survival time of 4-8 weeks, the usual time required to observe CR. More usually, resistance presents as relapse after an initial CR, which may last anywhere from one month to two years, with a median of approximately one year. Once resistance is demonstrated subsequent success is unlikely; the probability of a second CR increases with D_{CR}, and patients with D_{CR} less than one year have very poor prognosis.

In previously untreated patients, the principal predictor of resistance to standard therapy is the particular abnormal chromosomal composition of the AML cells, or "cytogenetics." Three cytogenetic groups can be distinguished. A good prognosis group, characterized by an inversion of the 16^{th} chromosome or a translocation of chromosomes 8 and 21, constitutes about 10 % of patients. Typically, such patients are aged under 60. A poor prognosis group, characterized by the loss of portions of chromosomes 5 or 7, comprises about 30 to 40 % of all patients. These patients are on average older and are more likely to have either a history of abnormal blood counts prior to the diagnosis of AML, known as an "antecedent hematologic disorder" (AHD), or have received previous chemotherapy for another condition, frequently lymphoma, breast, or ovarian cancer. The remaining 50 to 60 % of patients comprise a final, intermediate prognosis cytogenetic group, although their prognosis is closer to that of the poor than to the good group. Less than 10% of patients in the best prognostic group fail to achieve a CR with standard therapy, with median CR duration close to 2 years, and about 50% of such patients are cured, as defined above. In stark contrast, resistance to initial therapy occurs in 30 to 40% of patients in the poor prognosis group, with median remission duration is four months and a cure expected in less than 5 to 10% of these cases, depending on other covariates. Potential cure rates in the intermediate group range from less than 10 % to 30 % with a median remission duration of 6 - 24 months, again depending on other prognostic variables. In any case, once resistance has been established D_{CR} prognostically supersedes the patient's initial cytogenetic abnormality.

While cytogenetic information is highly prognostic of patient outcome, within each cytogenetic group this is still highly variable, particularly in the intermediate cytogenetic group described above. Inclusion of other covariates reduces this variability. Chief among these is an AHD, often preceding the diagnosis of AML by months or occasionally years. An AHD is seen in about one-third of patients. The longer the AHD the less the likelihood of cure, with the cure rate only about 10% in patients with an AHD and a normal karyotype.

Physicians treating patients with AML must make management decisions based on analysis of the prognostic factors described above. Broadly speaking, there are three options: palliative care without chemotherapy, standard chemotherapy, and investigational chemotherapy preferably within the context of a clinical trial. While by definition little is known about the efficacy or toxicity of a particular investigational treatment, the data described above indicate that the option of standard therapy would be satisfactory only for a minority of patients. Candidates for standard therapy certainly include patients in the best prognosis cytogenetic group. Another group in whom standard therapy could be considered appropriate are ambulatory patients in the intermediate cytogenetic group under age 60-70 and without an AHD. In the remaining patients, expectations with standard therapy are so low that its use can be readily questioned. Such patients have a substantial probability of either (a) death during remission induction, especially if bed-ridden and aged > 50, or aged over 80 regardless of performance status, or (b) resistance to therapy, especially if in the worse prognosis cytogenetic group, or in the intermediate prognosis cytogenetic group but with an AHD. Given these expectations, such patients, who constitute approximately 60% to 75% of all AML patients, optimally should be referred to academic centers for investigational approaches. However, such referral seldom occurs in the United States, where less than 10% of patients with AML are registered on clinical trials testing new therapies.

Numerous new therapies are now becoming available for clinical trials in patients with AML. These include not only new chemotherapeutic agents but also therapies believed to target specific abnormalities in AML blasts. For example, the AML blasts of some patients contain tumor suppressor genes that are ”hypermethylated” relative to similar genes in normal blasts. Hypermethylation prevents normal functioning of the gene. Drugs that induce hypomethylation have been developed, with the hypothesis that use of these drugs will permit normal functioning of the tumor suppressor genes with resultant beneficial effects. Another example is a gene for a protein called RAS that is thought to stop AML blasts from dying in the same manner as normal blasts. This motivates the use of treatments aimed at deactivating this gene and thus inhibiting RAS, in order to increase the death rate in the AML blasts and thus improve patient outcome.

Given the heterogeneity of AML described above, it is unlikely that all patients will have hypermethylated tumor suppressor genes, abnormally active RAS, or any other specific abnormality. It follows that any given "targeted" therapy may be appropriate for some patients but not for others. It is already known that increases in the dose of ara-C benefits patients in the better, but not in the worse, cytogenetic group. Nonetheless, nearly all currently used statistical designs do not account for prognostic covariates but rather proceed on the unlikely assumption that the effect of a given therapy is homogeneous across all AML patient subgroups.

Adding to this complexity is the likelihood that the sequence in which therapies are delivered may be important. Treatment of elderly patients in the intermediate cytogenetic group provides a simple example. It is highly probable that targeted therapies such as hypomethylating agents or RAS inhibitors are less toxic than usual chemotherapy. However, at least in their initial stages of development, they also may be less effective. This raises questions such as whether an elderly patient would live longer if a RAS inhibitor were given first, in order to lower the chance of therapy-induced death, with chemotherapy given subsequently only if the first treatment fails. It also may be argued that it is preferable to treat with chemotherapy in the first course, and then use the RAS inhibitor only if the chemotherapy fails. The rationale for treating more aggressively in the first course of therapy is the high death rate if this first course fails to achieve a CR. On a more fundamental level, it is possible that a given new therapy may affect various biologic parameters so as to influence the likelihood of success with a subsequent agent. This reflects the ability of the prior drug to sensitize leukemia cells to the actions of the subsequently administered drug. An obvious example is the use of lymphocyte infusions from normal donors. Such infusions have some efficacy in AML only if patients have previously received chemotherapy which induces sufficient suppression of the immune system to allow the donated lymphocytes to survive and exert a graft-vs-leukemia effect. Statistical designs in current use ignore the issue of how to sequence treatments, instead regarding each therapy as a distinct entity.

response to one regimen depends on which treatment was previously given motivated the design described herein. As noted above, standard phase II designs pay little attention to these issues. In particular, these designs focus on an average result and ignore treatment history. In contrast, our design's explicit purpose is to determine which sequence of therapies is best for each patient subgroup (young, long first remission vs young, short first remission vs old, long first remission vs old, short first remission).

3. A Bayesian Strategy for Constructing Designs

The first step in designing a clinical trial, regardless of the particular statistical methodology used, is to determine its essential elements. These are (1) the disease and specific patient group to be studied, that is, the trial's "entry criteria," (2) the treatments to be studied, (3) the therapeutic paradigm, including specification of multiple courses of therapy, (4) patient outcomes to be recorded after each course, as well as any long-term outcomes such as survival time, (5) the therapeutic and scientific objectives, (6) logistical constraints, including the anticipated accrual rate and upper limits on financial resources, trial duration, or drug availability, (7) patient covariates or subgroups, (8) the institutions that will participate, and (9) historical data from previous trials in the same or similar patient groups.

Once this basic information is in hand, one may apply the following statistical strategy to construct a trial design. The main idea is to utilize available historical data to first estimate non-treatment related parameters, such as patient covariate effects and baseline outcome rates. Assuming that these parameters will follow the same distribution in the planned trial as they did historically, the historical data may be combined with the trial data to improve the reliability of the treatment effect estimates. Essentially, this is just Bayesian covariate adjustment that is informed by historical data.

We begin the formal modeling process by specifying the patient outcomes and covariates of interest and formulating a Bayesian model describing both the historical data, $\mathcal{X}_H$, and the data, $\mathcal{X}$, that is anticipated from the trial being planned. The key step is to partition the model parameter vector $\boldsymbol{\theta}$ into two subvectors: the baseline parameters $\boldsymbol{\theta}_B$ that do not pertain to treatments and the parameters involving treatment effects. We denote the historical treatment-related parameters by $\boldsymbol{\theta}_{T(H)}$ and those for the upcoming trial by $\boldsymbol{\theta}_T$, since typically these are not the same. Our main focus is evaluating $\boldsymbol{\theta}_T$, and the values of $\boldsymbol{\theta}_{T(H)}$ are not relevant to inferences about either $\boldsymbol{\theta}_T$ or the treatment strategies in the trial. Specifying reasonably uninformative priors on $\boldsymbol{\theta}_B$, $\boldsymbol{\theta}_{T(H)}$, and $\boldsymbol{\theta}_T$, the historical data are fit to obtain the marginal posterior of the baseline parameters,

$$f(\boldsymbol{\theta}_B \mid \mathcal{X}_H) \;=\; \int_{\boldsymbol{\theta}_{T(H)}} f(\boldsymbol{\theta}_B, \boldsymbol{\theta}_{T(H)} \mid \mathcal{X}_H)\, d\boldsymbol{\theta}_{T(H)}.$$

At the start of the trial, we assume that $\boldsymbol{\theta} = (\boldsymbol{\theta}_B,\, \boldsymbol{\theta}_T)$ follows the prior $f(\boldsymbol{\theta}_B \mid \mathcal{X}_H)\, f(\boldsymbol{\theta}_T)$. The posterior used at each interim analysis during the trial or in a final analysis is then

$$f(\boldsymbol{\theta}_T \mid \mathcal{X},\, \mathcal{X}_H) \;=\; \int_{\boldsymbol{\theta}_B} f(\boldsymbol{\theta}_T,\, \boldsymbol{\theta}_B \mid \mathcal{X}, \mathcal{X}_H)\, d\boldsymbol{\theta}_B,$$

where $\mathcal{X}$ now denotes the most recent trial data at the time of the analysis. The historical data thus provide initial information about baseline parameters while the data $\mathcal{X}$ accumulating during the trial provide new information about the parameter of primary interest, $\boldsymbol{\theta}_T$, as well as additional information about $\boldsymbol{\theta}_B$.

If one were to assume that $\boldsymbol{\theta}_T = \boldsymbol{\theta}_{T(H)}$ then the appropriate prior at the start of the trial would be $f(\boldsymbol{\theta}_T, \boldsymbol{\theta}_B \mid \mathcal{X}_H)$. The resulting inferences based on $\mathcal{X}$ and $\mathcal{X}_H$ would then constitute a meta-analysis, and the issue would arise of accounting for trial effects. Our aim here is not to pool treatment effect information from different trials, however, since $\boldsymbol{\theta}_T \neq \boldsymbol{\theta}_{T(H)}$. Rather, we wish to do a reasonable job of accounting for covariate effects while evaluating $\boldsymbol{\theta}_T$.

The next step is to specify the trial design, which of course may be done in numerous ways depending on the particular setting at hand. In general, we determine numerical values of design parameters by first specifying several clinical scenarios in terms of fixed values of the parameters $(\boldsymbol{\theta}_B,\, \boldsymbol{\theta}_T)$ and simulating the trial under each scenario. Frequentist properties of the design obtained in this way may be used to calibrate design parameters and explain the design's properties to the physicians involved. This is analogous to the widespread practice of determining an experiment's sample size to achieve a given power under a conventional test of hypothesis. In our experience, however, frequentist tests often greatly oversimply both the data structure and the actual decisions made during a clinical trial.

Denoting a given J–course treatment strategy by $\boldsymbol{\tau} = (t_1, \ldots, t_J)$ and the vector of probabilities of the possible outcomes with $\boldsymbol{\tau}$ over J courses by $\boldsymbol{\xi}(\boldsymbol{\tau})$, we base comparison of different multi–course treatment strategies on a real–valued objective function $\phi(\boldsymbol{\xi}(\boldsymbol{\tau}))$ elicited from the physician(s) planning the trial. Interim decisions to drop comparatively inferior strategies and selection of a best strategy at the end of the trial may be based on posterior probabilities such as $\Pr[\phi(\boldsymbol{\xi}(\boldsymbol{\tau}_1)) < \phi(\boldsymbol{\xi}(\boldsymbol{\tau}_2)) \mid \mathcal{X}]$, posterior means, or predictive probabilities. Inferences may be made for prognostic subgroups so that, for example, strategy $\boldsymbol{\tau}_1$ may be best for one subgroup while strategy $\boldsymbol{\tau}_2$ is best for another.

4. Probability Models

4.1 A Two–Course Model for the AML Trial

We now develop a probability model and design for the AML trial. Denote by (s, t) the two–course treatment strategy wherein the patient receives treatment T_s in the first course and, if the first course results in failure, receives T_t in the second course. Denoting IDA, IDA+M, and IDA+T by T_0, T_1, and T_2 for brevity, the AML trial design allows the four two-course strategies $\mathcal{S} = \{(1,\ 0), (2,\ 0), (0,\ 1), (0,\ 2)\}$. While strategies $(1, 2)$ and $(2,\ 1)$ are not permitted in the AML trial, in general the methodology allows $\mathcal{S}$ to contain any two–course combination, including strategies of the form (s, s) that give the same treatment in both courses. Each two–course strategy (s, t) has five possible outcomes. Therapy may end in the first course with either response or death with T_s, or T_s may fail in the first course, followed by response, death, or failure with T_t in the second course.

For each patient baseline prognostic covariate vector $\mathbf{Z} = (Z_1, \ldots, Z_q)$, the goal is to select the best two–course treatment strategy (s, t) from $\mathcal{S}$ based on the probabilities $\xi_R(s, t, \mathbf{Z})$ of achieving CR and $\xi_D(s, t, \mathbf{Z})$ of death. For course c = 1 or 2, let τ_c denote the treatment index 0, 1, or 2, and let Y_{Rc} and Y_{Dc} denote the indicators of response and death, so that Y_{Fc} $= 1 - Y_{Rc} - Y_{Dc}$ indicates failure. Since there is no second course if $Y_{F1} = 0$, for completeness we define $Y_{R2} = Y_{D2} = 0$ and $\tau_2 = 0$ in this case. Denote the probability of outcome $k = R, D$, or F with T_s in course 1 by

$$\pi_{k1}(s, \mathbf{Z}) = Pr[Y_{k1} = 1 \,|\, \mathbf{Z},\, \tau_1 = s], \tag{1}$$

and the probability of outcome k with T_t in course 2 after a failure with T_s in course 1 by

$$\pi_{k2}(s, t, \mathbf{Z}) = Pr[Y_{k2} = 1 \,|\, \mathbf{Z},\, \tau_1 = s, Y_{F1} = 1, \tau_2 = t]. \tag{2}$$

Aside from covariates, π_{k1} is a function of τ_1 alone while π_{k2} is a function of both τ_1 and τ_2. Since one of R, D, or F must occur in each course and the occurrence of either R or D, or two treatment failures, marks the end of the patient's therapy, for any strategy (s, t)

$$\pi_{R1}(s, \mathbf{Z}) \;+\; \pi_{D1}(s, \mathbf{Z}) \;+\; \pi_{F1}(s, \mathbf{Z}) \sum_{k=R,D,F} \pi_{k2}(s, t, \mathbf{Z}) \;=\; 1. \tag{3}$$

The likelihood function of the i^{th} patient thus takes the form

$$\mathcal{L}_i = \prod_{k=R,D,F} \Big\{\pi_{k1}(\tau_{i1}, \mathbf{Z}_i)\Big\}^{Y_{i,k1}} \Bigg\{ \prod_{r=R,D,F} \Big[\pi_{r2}(\tau_{i1}, \tau_{i2}, \mathbf{Z}_i)\Big]^{Y_{i,r2}} \Bigg\}^{Y_{i,F1}}, \tag{4}$$

with $\mathcal{L} = \prod_{i=1}^{n} \mathcal{L}_i$ the likelihood of a sample of n patients.

4.2 A Generalized Logistic Model

The following generalized logistic model (cf. Agresti, 1990, Chapter 9.2) accounts for trinary outcomes, the two–course treatment structure, and prognostic covariates. The formulation also accommodates an arbitrary number of treatments and any collection of two–course treatment sequences formed from them. In addition to its use as a basis for clinical trial design and conduct, this regression model is also very useful *per se* when the primary goal is to analyze existing data consisting of trinary outcomes with covariates.

For outcome $k = R$ or D, treatment strategy (s, t), and covariates $\mathbf{Z}$, denote the linear components corresponding to courses 1 and 2 by

$$\eta_{k1}(s, \mathbf{Z}) \;=\; \mu_k + \alpha_k(s) + \sum_{j=1}^{q} \Big\{\gamma_{kj} + \zeta_{kj}(s)\Big\} Z_j \;, \tag{5}$$

and

$$\eta_{k2}(s, t, \mathbf{Z}) \;=\; \mu_k + \alpha_k(t) + \beta_k(s, t) + \sum_{j=1}^{q} \Big\{\gamma_{kj} + \zeta_{kj}(t) + \delta_{kj}\Big\} Z_j \;, \tag{6}$$

respectively, subject to the $2(q+1)$ constraints

$$\sum_s \alpha_R(s) = \sum_s \alpha_D(s) = 0 \ \text{ and } \ \sum_s \zeta_{Rj}(s) = \sum_s \zeta_{Dj}(s) = 0, \quad j = 1, \ldots, q. \tag{7}$$

Our application includes $q = 2$ covariates, hence 6 constraints, so we set $\alpha_k(0) = 0$ and $\zeta_{kj}(0) = 0$ for $k = R, D$ and $j = 1, 2$. That is, we use $s = 0$ as the baseline treatment group.

We characterize the regression of the outcomes $\mathbf{Y}_1 = (Y_{R1}, Y_{D1})$ and $\mathbf{Y}_2 = (Y_{R2}, Y_{D2})$ on treatment strategy (s, t) and covariates $\mathbf{Z}$ by the probability functions

$$\pi_{k1}(s, \mathbf{Z}) \ = \ \frac{\exp\{\eta_{k1}(s, \mathbf{Z})\}}{1 + \{\eta_{R1}(s, \mathbf{Z})\} + \exp\{\eta_{D1}(s, \mathbf{Z})\}} \tag{8}$$

and

$$\pi_{k2}(s, t, \mathbf{Z}) \ = \ \frac{\exp\{\eta_{k2}(s, t, \mathbf{Z})\}}{1 + \exp\{\eta_{R2}(s, t, \mathbf{Z})\} + \exp\{\eta_{D2}(s, t, \mathbf{Z})\}} \ , \tag{9}$$

where $k = R$ or D, with each $\pi_{Fc} = 1 - (\pi_{Rc} + \pi_{Dc}) = 1/[1 + \exp(\eta_{Rc}) + exp(\eta_{Dc})]$. Under this generalized logistic model, for each $k = R$ or D, the intercept of η_{kc} is decomposed into the baseline mean μ_k, the main effect $\alpha_k(s)$ of treatment s and, for course 2, the additional effect $\beta_k(s, t)$ of t as a salvage treatment following failure with s. Similarly, the coefficient of Z_j for outcome Y_k is decomposed into the baseline parameter γ_{kj}, the treatment effect $\zeta_{kj}(s)$, and the course 2 effect δ_{kj}. Viewing (π_{k1}, π_{k2}) as a function of treatment, course, and covariates, if we write $\beta_k(s, t) = \beta_k + \beta^*_k(s, t)$ with $\sum_{(s,t)} \beta^*_k(s, t) = 0$ for each k, then $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ are the treatment, course, and covariate main effects, while $\boldsymbol{\beta}^*$, $\boldsymbol{\zeta}$, and $\boldsymbol{\delta}$ are the [treatment × course], [treatment × covariate], and [covariate × course] interactions. Thus, the treatment-related parameters are $\boldsymbol{\theta}_T = (\boldsymbol{\alpha}, \boldsymbol{\beta}^*, \boldsymbol{\zeta})$ and the baseline parameters are $\boldsymbol{\theta}_B = (\boldsymbol{\mu}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\beta})$.

If there are m treatments and r_m two–course strategies then, subject to the constraints, $\dim(\theta_T) = 2\{(m-1)(q+1) + r_m - 1\}$, $\dim(\theta_B) = 4(q+1)$ and the overall model dimension is $p = 2\,(q + m + q\,m + r_m)$. Alternatively, it is also useful to decompose θ_T in terms of the vector $(\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\beta}^*)$ of intercept parameters, which has dimension $2(m + r_m)$, and the vector $(\boldsymbol{\gamma}, \boldsymbol{\zeta}, \boldsymbol{\delta})$ of covariate effect parameters, which has dimension $2q(m+1)$. For the AML trial $p = 30$, since $q = 2$, $m = 3$, and $r_m = 4$.

5. Computational Methods

We employ the following computational approximation to facilitate simulation study of the trial during the design stage. Assume *a priori* that $\boldsymbol{\theta}$ is multivariate normal, denoted $\boldsymbol{\theta} \sim N(\boldsymbol{\mu}_\theta,\, \Omega)$. Under the usual frequentist large sample theory, the maximum likelihood estimate (MLE) $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is approximately multivariate normal, denoted $\widehat{\boldsymbol{\theta}} \mid \boldsymbol{\theta} \mathrel{\dot\sim} N(\boldsymbol{\theta},\, \Sigma)$. It follows from Bayes' Theorem that, *a posteriori*, $\boldsymbol{\theta} \mid \widehat{\boldsymbol{\theta}} \mathrel{\dot\sim} N(B\,\mathbf{b},\, B)$, where $B = (\Sigma^{-1} + \Omega^{-1})^{-1}$ and $\mathbf{b} = \Sigma^{-1}\widehat{\boldsymbol{\theta}} + \Omega^{-1}\boldsymbol{\mu}_\theta$ (Lindley and Smith, 1972). This approach has been used by many authors, including Dixon and Simon (1991) in the context of Bayesian subset selection, by Faraggi and Simon (1997) in proportional hazards regression and by Thall, Simon and Shen (2000) in evaluating multidimensional treatment effects. The method is straightforward, since it relies on multivariate normal distributions. The necessary computations include deriving the MLE, computing an estimator $\widehat{\Sigma}$ of the covariance matrix, and generating multivariate normal posterior samples using a Cholesky decomposition. It may be implemented with standard statistical software and provides a practical alternative to more computationally intensive Markov chain Monte Carlo (MCMC) methods.

6. An Objective Function

The overall probability of outcome $k = R$ or D in either one or two courses with the treatment strategy (s,t) for a patient with covariates $\mathbf{Z}$ is

$$\xi_k(s,t,\mathbf{Z}) \;=\; \pi_{k1}(s,\mathbf{Z}) \;+\; \pi_{F1}(s,\mathbf{Z})\,\pi_{k2}(s,t,\mathbf{Z}). \tag{10}$$

We will use the overall probabilities $\boldsymbol{\xi}(s,t,\mathbf{Z}) = (\xi_R(s,t,\mathbf{Z}), \xi_D(s,t,\mathbf{Z}))$ of response and death as the basis for both interim safety monitoring and treatment strategy selection, since these are what matter clinically. Because $\boldsymbol{\xi}(s,t,\mathbf{Z})$ is two–dimensional, the use of this criterion to compare treatment strategies is problematic. We thus define an objective function to reduce $\boldsymbol{\xi}(s,t,\mathbf{Z})$ to a single real number by quantifying the trade–off between the likelihood of response and the risk of death. We we will use it as a basis for both interim monitoring and treatment selection. Temporarily suppress the argument $(s,t,\mathbf{Z})$. The function ϕ is constructed so that all pairs (ξ_R, ξ_D) for which $\phi(\xi_R, \xi_D)$ equals a given constant are equally desirable. The process of eliciting ϕ from the physicians planning the trial may be facilitated by interactively modifying $\phi(\xi_R, \xi_D)$ while viewing its contours on a computer screen.

For the AML trial, we began with the family of linear objective functions $\phi = a\,\xi_R + b\,\xi_D$ in the triangular two–dimensional domain of (ξ_R, ξ_D) over a range of (a, b) values with $a > 0 > b$. We determined ϕ by specifying two equations and solving for a and b. The null value (ξ_R, ξ_D) = (.40,.40) corresponding to all patients in the historical data was assigned $\phi = 0$, and the desirable goal $(\xi_R, \xi_D) = (.50,.15)$ was assigned $\phi = 1$. The values 0 and 1 for ϕ in these two cases were chosen purely for numerical convenience. After examining plots of the resulting linear contours, it was decided that ϕ should increase more rapidly in ξ_R for smaller values of ξ_D, especially for ξ_D near 0. We thus considered functions of the more general form

$$\phi(\xi_R,\ \xi_D) = a\,\xi_R + b\,\xi_D{}^c \tag{11}$$

with $a > 0 > b$ and $c > 0$. Given the above two constraints, a third equation to determine c was given by the value of ξ_R that would be required to still have $\phi = 1$ if there were no fatalities, that is, the value of ξ_R such that $\phi(\xi_R, 0) = \phi(.50, .15) = 1$. After examining contour plots corresponding to several different values of ξ_R using the three–parameter version of ϕ, this was specified to be $\xi_R = .30$. A contour plot of the resulting ϕ, which is characterized by $a = 3.333$, $b = –2.548$ and $c = .707$, is given in Figure 2.

[Figure 2 about here]

Other functional forms for ϕ could be used, provided that ϕ increases in ξ_R and decreases in ξ_D. The shape of its contours should provide a reasonably flexible graphical representation of the trade–off between ξ_R and ξ_D that reflects the physicians' goals and opinions. The particular shape of our trade–off function contours is one of several geometries in the two–dimensional parameter plane that have been proposed to characterize the trade–off between safety and efficacy. To define hypotheses for tests based on bivariate outcomes, Willan and Pater (1985) use two parallel lines that partition the plane into three hypotheses, Jennison and Turnbull (1993) and Bryant and Day (1995) utilize various rectangular regions, while Thall and Cheng (1999) propose polygonal regions.

The probability model, probabilities of response and death in each course, overall probabilities of response and death, and objective function comprise a parametric hierarchy. The mapping $\theta \longrightarrow \boldsymbol{\pi}(s,t,\mathbf{Z}) = (\pi_{R1}(s,\mathbf{Z}), \pi_{D1}(s,\mathbf{Z}), \pi_{R2}(s,t,\mathbf{Z}), \pi_{D2}(s,t,\mathbf{Z}))$ reduces the parameter vector to the probabilities of response and death in each course using the strategy (s,t) in prognostic group $\mathbf{Z}$. Next mapping $\boldsymbol{\pi}(s,t,\mathbf{Z}) \longrightarrow \boldsymbol{\xi}(s,t,\mathbf{Z})$ into the two–dimensional triangular region illustrated in Figure 2 limits attention to the overall two–course probabilities of response and death. The final real–valued mapping $\boldsymbol{\xi}(s,t,\mathbf{Z}) \longrightarrow \phi(\boldsymbol{\xi}(s,t,\mathbf{Z}))$ induces an ordering among the strategies, thus providing a basis for comparison and selection.

7. Analysis of the Historical Data

For all model fits reported here, both in analysis of the historical data and in fitting models to simulated data sets, *a priori* all parameters in each model were assumed to be iid normal random variables with mean 0 and variance 10.

The prognostic covariates used in the model–based analysis of the historical data were the binary indicators of whether the patient's age was < 50 years and whether the patient's initial remission duration prior to entering the trial was at least one year. Thus, q=2 and there were four prognostic subgroups. For example, the group having worst prognosis consisted of the older patients with short initial CR duration, while the best prognostic group had the younger patients with long initial CR duration. There were m=3 treatment groups, and the treatment effects in the model correspond to allogeneic bone marrow transplant (s=1) and chemotherapy not including ara–C ($s = 2$) relative to the baseline treatment group consisting of chemotherapy containing high dose ara–C ($s = 0$). Because there were $r_m = 9$ different two–course treatment combinations, the full model has a total of $p = 40$ parameters.

Starting with the full model and including $\boldsymbol{\theta}_B$ throughout, we obtained a more parsimonious model by successively eliminating entries from the parameter vector $\boldsymbol{\theta}_{T(H)}$ pertaining to treatments in the historical data. We considered only hierarchical models. Because the two elements of each pair $\boldsymbol{\zeta}_j(s) = (\zeta_{R,j}(s), \zeta_{D,j}(s))$ act together, we either included both entries or deleted the pair. As criteria for model comparison, we used the maximized log likelihood, variability of posterior parameter means, and the *Bayes information criterion*

$$\mathrm{BIC}(\mathcal{M}) \;=\; \log \mathcal{L}(\widehat{\boldsymbol{\theta}}_{\mathcal{M}}) \;-\; \frac{1}{2}\, p_{\mathcal{M}} \log(n), \tag{12}$$

where $p_{\mathcal{M}}$ is the number of parameters in model $\mathcal{M}$. In particular, the BIC penalizes the log likelihood for larger $p_{\mathcal{M}}$. A discussion of the BIC is given by Kass and Raftery (1995).

The fitted models that we considered are summarized in Table 3. We first eliminated the vector $\boldsymbol{\beta}^*$ of treatment-course interactions because this increased the BIC greatly (Model 3), much more than the increase obtained by eliminating $\boldsymbol{\zeta}$ (Model 2). Moreover, under the full model the absolute value of the posterior mean of each entry of $\boldsymbol{\beta}^*$ was small compared to its standard deviation. Next focusing on the four pairs of treatment-covariate interactions, $\{\boldsymbol{\zeta}_j(s),\ j$ = AGE, DUR, s = 1,2$\}$, we successively eliminated pairs in a stepdown manner. We stopped with Model 6, our final model because $\Pr\{\zeta_{D,DUR}(1) > 0 \mid data\} = .95$, hence it was appropriate to retain the pair $\boldsymbol{\zeta}_{DUR}(1)$.

[Table 3 about here]

To check the approximate Bayesian method, we also computed the posteriors under several models in Table 3 using MCMC (Gilks, et al. 1996). Each MCMC computation was based on 100,000 runs with a burn-in sample of 10,000. The two methods gave similar posteriors, with a few large differences for parameters with a posterior mean very small relative to its standard deviation, that is, with marginal posterior centered around 0 and very disperse. Under Model 6, the posterior approximate mean(std) of $\zeta_{R,DUR}(1)$ was $-0.277(.713)$ compared to $-0.162(.719)$ using MCMC; the approximate mean(std) of $\gamma_{D,DUR}$ was $-0.007(.312)$ compared to $-0.018(.286)$ using MCMC; the approximate mean(std) of $\delta_{R,AGE}$ was $0.081(.636)$ compared to $0.095(.481)$ using MCMC. The posterior means of the remaining 15 parameters differed by $< 8\%$, with each difference well within the posterior standard deviation.

Table 4 gives posterior means of the parameters in the final model computed using the approximate method and MCMC and corresponding MLEs. The signs of the estimates $\{\hat{\alpha}_j(s), j = 1,2,\ s = 1,2\}$ of the main treatment effects show that, relative to high dose ara-C, transplant had higher rates of both CR and death while non-ara-C chemo had lower rates of both events. The well-known fact that the CR rate decreases and the death rate increases in a second course of treatment following failure in a previous course is borne out by the relationship $\hat{\beta}_R < 0 < \hat{\beta}_D$.

[TABLE 4 ABOUT HERE]

The signs of the remaining parameter estimates in Table 4 should be interpreted in the context of the generalized logistic model's algebraic structure, which differs from that of the usual logistic model. For example, although the fact that $\widehat{\gamma}_{D,DUR} > 0$ considered *per se* might seem to imply the model predicts a higher probability of death for patients with a longer initial CR duration, this is not the case. The effect of a given covariate on π_{Dc} is determined by all of that covariate's coefficients, including those indexed by both R and D. The numerical values of $\widehat{\gamma}_{R,DUR}$, $\widehat{\gamma}_{D,DUR}$, $\widehat{\zeta}_{R,DUR}(1)$ and $\widehat{\zeta}_{D,DUR}(1)$ act together so that $\widehat{\pi}_{D1}$ decreases and $\widehat{\pi}_{R1}$ increases with longer initial CR duration, as should be the case on medical grounds. This illustrates the fact that these four parameters act together algebraically for each treatment to determine the course 1 probabilities. Similarly, these parameters and $(\delta_{R,2}, \delta_{D,2})$ act together to determine the effect of Z_2 on the course 2 probabilities. Table 5, which gives the predicted and empirical overall CR and death probabilities within each prognostic group for patients who received only high dose ara-C in either course, shows that the fitted model gives predictions that make sense for the four prognostic subgroups. In particular, Table 5 illustrates the importance of accounting for prognostic group, since $\widehat{\xi}_R$ increases and $\widehat{\xi}_D$ decreases with increasing CR duration and with younger age, and these changes are quite large. Moreover, the good agreement between the model–based estimates and the corresponding empirical values provide a further validation of the model.

[TABLE 5 ABOUT HERE]

8. Trial Conduct

Aside from technical details related to accounting for two courses, the trinary outcome, and four prognostic subgroups, in principle conduct of the AML trial is straightforward.

Patients are randomized fairly among the acceptable treatments at each of two stages, using the Pocock–Simon dynamic allocation procedure to balance on the two covariates. Halfway through the trial, a safety monitoring rule is applied within each prognostic subgroup to drop any treatment strategies that are comparatively inferior. The stage 2 randomization thus accounts for any strategies that are dropped in any prognostic subgroups after stage 1. At the end of the trial, the best strategy for each prognostic subgroup is selected. Formally, the AML trial is conducted as follows:

Stage 1. Randomize $n/2$ patients fairly among the three treatments for their first course of therapy, using the Pocock–Simon algorithm to balance on Z_1 and Z_2. Patients who fail with T_0 in course 1 are randomized between T_1 and T_2 for their second course. All patients who fail with either T_1 or T_2 in course 1 are treated with T_0 in course 2. If

$$\Pr[\ \phi(s,t,\mathbf{Z})\ >\ \phi(u,v,\mathbf{Z})\ \mid\ \mathit{data}\ ]\ >\ .95 \tag{13}$$

for distinct strategies (s,t) and (u,v), then drop strategy (u,v) in patient subgroup $\mathbf{Z}$.

Stage 2. Randomize $n/2$ additional patients among the treatments in each course as in Stage 1, subject to the constraints imposed by dropping any treatment strategies. Once n patients have been treated and evaluated, for each $\mathbf{Z}$ select the two–course strategy, among those not dropped in that subgroup, for which the posterior mean of $\phi(s,t,\mathbf{Z})$ is largest.

9. Simulation Study

The simulations were designed to provide a reasonable reflection of actual trial conduct. As explained earlier, the scientific purpose of the trial is to learn about the treatment–related parameters $\boldsymbol{\theta}_T = (\boldsymbol{\alpha}, \boldsymbol{\beta}^*, \boldsymbol{\zeta})$, and we use the historical data to obtain preliminary knowledge about the non–treatment–related parameters $\boldsymbol{\theta}_B = (\boldsymbol{\mu}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\beta})$. Thus, in fitting each simulated data set, we applied the approximate Bayesian method using the posterior $f(\boldsymbol{\theta}_B \mid \mathcal{X}_H)$ from the fitted historical data, under the model summarized in Table 4, as the the prior of the non–treatment–related parameters $\boldsymbol{\theta}_B$, and iid $N(0, 10)$ priors for the treatment effect parameters in $\boldsymbol{\theta}_T$.

9.1 Clinical Scenarios

Because the two–course, trinary outcome setting considered here is more complex than a single–course selection trial based on a univariate outcome, necessarily our design and criteria for evaluating its performance also are more complex. To provide a conceptual framework for what follows, we first briefly review the analogous single–course setting with a univariate outcome where the goal is to select the best among k treatments based on estimates of their means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)$. Without loss of generality assume that $\mu_1 \leq \ldots \leq \mu_k$. For a randomized trial to select a single best treatment, let μ_0 be a null value and $\mu_0 + \delta$ a desirable target, where δ is a clinically significant improvement over μ_0. The *null configuration* $\boldsymbol{\mu}^0$ is the k–vector having all $\mu_j = \mu_0$, while the *least favorable configuration* (LFC) $\boldsymbol{\mu}^*$ has $\mu_1 = \ldots = \mu_{k-1} = \mu_0$ and $\mu_k = \mu_0 + \delta$ (cf. Gibbons, Olkin and Sobel, 1977, 1.3). It can easily be shown that, among the set of $\boldsymbol{\mu}$ having no entries between μ_0 and $\mu_0 + \delta$ and at least one entry $\geq \mu_0 + \delta$, the LFC minimizes the probability of correct selection (PCS) of treatment k. Since the PCS under $\boldsymbol{\mu}^*$ increases with sample size, n, given $\boldsymbol{\mu}_0$ and δ one may determine n to achieve a given PCS. In the present setting, one may regard the two contours on which $\phi(\xi_R, \xi_D) = 0$ and 1, respectively, as two–dimensional generalizations of the points μ_0 and $\mu_0 + \delta$ in the one–dimensional case.

The clinical scenarios given in Table 6 may be regarded as multi–dimensional generalizations of the null vector and LFC in the one–dimensional case. We will use these scenarios, and one more complex scenario that is not tabled, as a basis for evaluating the selection design, and for determining sample size. Because we account for trinary outcomes, two treatment courses and four patient prognostic groups, our parametric characterizations of clinical settings are necessarily more complex than those in the univariate single–course case. Consequently, there are more qualitatively different cases than the two, noted above, that typically are considered in the one–dimensional case. The five scenarios under which we evaluate the design's OCs were chosen to cover a reasonable range of cases that may actually obtain in practice, and should illustrate the design's essential properties.

[Table 6 about here]

We determined each scenario in Table 6 by first specifying values for the 14 probabilities $\{\pi_{k1}(s,\mathbf{0}), \pi_{k2}(s,t,\mathbf{0})\}$ corresponding to $\mathbf{Z} = \mathbf{0}$ and then using these values to determine the 14 parameters $(\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ via a one–to–one transformation. These parameters in turn determine the linear components $\eta_{k1}(s,\mathbf{0})$ and $\eta_{k2}(s,t,\mathbf{0})$. The probabilities $\{\pi_{k1}(s,\mathbf{Z}), \pi_{k2}(s,t,\mathbf{Z})\}$ for $\mathbf{Z} \neq \mathbf{0}$ were obtained by adding the covariate adjustment terms $(\boldsymbol{\gamma} + \boldsymbol{\delta}\, I[c = 2])'\mathbf{Z}$ to the $\eta_{k1}(s,\mathbf{0})$'s and $\eta_{k2}(s,t,\mathbf{0})$'s using the posterior means of $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ from the historical data. To obtain covariate-adjusted probabilities in the simulations, the value of $\mathbf{Z}$ for each simulated patient was chosen randomly using the historical frequencies of the four prognostic groups, which were .42 for (CR duration, Age) = (Short, Old), .35 for (Short, Young), .11 for (Long, Old), and .12 for (Long, Young).

Because the probabilities for each scenario vary with $\mathbf{Z}$, to conserve space we present numerical values corresponding to the prognostic group having short CR duration and younger age, $\mathbf{Z}$ = (0,1), since this is a reasonably representative subgroup. The scenarios are given in Table 6 and the simulation results in Table 7. For each scenario in Table 6, the corresponding probabilities and values of ϕ for the other three prognostic subgroups vary in a manner analogous to the estimates in Table 5. Suppressing the argument $\mathbf{Z}$ = (0,1) in π_{kc} for brevity, the null scenario corresponds to (π_{R1}, π_{D1}) = (.16,.22) and (π_{R2}, π_{D2}) = (.09,.17) regardless of treatments, using the historical probabilities in this prognostic group. Scenario A is obtained by changing the course 2 probabilities $(\pi_{R2}(0,1), \pi_{D2}(0,1))$, corresponding to salvage with T_1 following a course 1 failure with T_0, from the null values (.09,.17) to (.47,.06). The result is that, in terms of the objective function ϕ, strategy (0,1) is greatly superior to the other three strategies. Scenario A is analogous to the LFC in the one–dimensional setting, although cases with $\phi(0,1) > \phi(0,2) = \phi(1,0) = \phi(2,0)$ may be obtained in a variety of different ways. Scenario B is obtained from the null scenario by increasing $\pi_{R1}(1)$ from .16 to .51 and $\pi_{R2}(0,1)$ from .09 to .26. That is, T_1 improves the response rates in both courses without affecting the death rates. In this case, $\phi(1,0) > \phi(0,1) > \phi(0,2) = \phi(2,0)$. Scenario C is obtained from the null scenario by changing $(\pi_{R1}(1), \pi_{D1}(1))$ from (.16,.22) to (.52,.27) and $(\pi_{R2}(0,1), \pi_{D2}(0,1))$ from (.09,.17) to (.30,.46). That is, under scenario C, T_1 increases the probabilities of both response and death in both courses, which is a phenomenon commonly encountered in testing experimental treatments for AML. In this case, $\phi(1,0) > \phi(0,2) = \phi(2,0) > \phi(0,1)$, so that strategy (0,1) is worst and (1,0) is best. Scenario D includes a treatment–covariate interaction in which T_1 is a superior salvage treatment overall, but also increases the death rate among older patients. The probabilities characterizing this scenario were obtained by parameterizing the model using the indicator Z^*_{AGE} of older age, so that larger values of $\zeta^*_{D,AGE}(1)$ correspond to higher death rates among older patients treated with T_1 in either course. We obtained the probabilities for this scenario by starting with Scenario A and increasing $\zeta^*_{D,AGE}(1)$ from 0 to 3. For example, among older patients with short CR duration treated with strategy (0,1), this has the effect of changing the two–course response and death rates from $\xi_R(0,1) = .35$ and $\xi_D(0,1) = .37$ under scenario A to $\xi_R(0,1) = .17$ and $\xi_D(0,1) = .70$ under scenario D.

[Table 7 about here]

9.2 Simulation Results

The trial was simulated 4000 times under each clinical scenario. The values in Tables 7 and 8 and reported in the text are the means over these repetitions. Each simulated data set was fit via maximum likelihood using the full 30–parameter model specified by equations (5) – (9) for s = 0, 1 or 2 and the four strategies (s,t) = (0,1), (0,2), (1,0), (2,0). The Bayesian decision criteria used in each simulated trial were computed using the approximate method described in Section 2.3. The sample size of 96 patients used throughout was chosen to obtain a correct selection probability $\geq .75$ in the (Short CR duration, Younger age) prognostic subgroup under Scenario A.

The OCs in Table 7 indicate that, under each of a reasonable set of possible clinical scenarios, the design has a good probability of correctly selecting the best two–course treatment strategy. The tabled correct selection probabilities are substantial improvements over the probability .25 of guessing the best strategy in the absence of empirical evidence. Unfortunately, this practice is quite common in clinical settings where several treatment strategies are available and one must be selected. The numerical results should be interpreted in terms of the numerical values of the probabilities that characterize each scenario and the fact that, of the 96 patients in the trial, on average the sample sizes in the subgroups are only 39.2 in (Short, Old), 34.4 in (Short, Young), 11.2 in (Long, Old), and 11.2 in (Long, Old).

The variation in the selection probabilities of the three inferior strategies under scenario A, from .04 to .11, is due to the facts that the course 2 sample sizes are not fixed. Rather, they depend on the number of failures in each course 1 treatment group and the imbalance in the course 2 randomization. In the (Short, Young) subgroup, on average $(1/3)\times 34.4 = 11.5$ patients are randomized to each of the three treatments in course 1. Since all three treatments have the same course 1 failure rate $\pi_{F1}(s,(0,1)) = .6185$ in the (Short, Young) subgroup under Scenario A, this yields about $\pi_{F1}(0,(0,1))\,11.5 = 7.1$ patients who fail in course 1 with T_0 and are randomized equally between T_1 and T_2 in course 2, hence about 3.6 patients receive each of strategies (0,1) and (0,2). In contrast, on average 7.1 patients receive strategy (1,0) and 7.1 receive strategy (2,0). This also explains why on average fewer patients receive the best strategy (0,1) than receive either (1,0) or (2,0) under scenario A, which otherwise may seem counterintuitive.

Table 8 summarizes the results under scenario D, which illustrate the design's ability to select the best strategy within each prognostic subgroup. Since treatment 1 has a higher death rate among older patients under this scenario, it is desirable for the design to have a relatively low probability of selecting either strategy (0,1) or (1,0) in either of the two prognostic groups having older patients. Equivalently, it is desirable to select either (2,0) or (0,2) for older patients. The probabilities of correctly selecting either strategy (0,2) or (2,0) are .87 in the (Short, Old) subgroup and .76 in the (Long, Old) subgroup, and on average only 39.3 and 11.0 patients, respectively, are treated in these two subgroups. These subgroup–specific selection probabilities should be compared to the value .50 that would be obtained by guessing. The much smaller correct selection probability .42 for the optimal strategy (0,1) in the (Long, Young) subgroup is due to its much smaller sample size of 11.3, although this probability is still much larger than the value .25 obtained by guessing. The fact that the design performs well under scenario D may be attributed to the adaptive nature of the two-course treatment strategy and borrowing of strength by the parametric model across the various treatment strategy and prognostic subgroup combinations.

[Table 8 about here]

The interim decision rule that drops inferior treatments has a very small effect on the selection probabilities, but yields a design that on average treats more patients with the superior strategies. For example, under scenario A, if the interim rule is not used then the total number of patients in all prognostic groups treated with the best strategy (0,1) drops from 21.1 to 15.4, so that about six more patients among the 96 receive the best treatment strategy due to interim monitoring. The effect of interim monitoring under scenario D is greater, with on average 41.4 - 32.7 = 8.7 more patients among the 96 receiving one of the best strategies in their prognostic group due to the interim monitoring rule. Since dropping the interim monitoring rule (13) corresponds to using an upper probability cutoff of 1, the question arises of whether this cutoff may be calibrated to improve the design's OCs. We thus repeated the simulations summarized in Tables 5 and 6 using cutoffs .90 and .99. As the cutoff is increased over this range the design's overall safety drops, but there is no clear pattern in its effect on the selection probabilities. It appears that the design's safety and selection probabilities may depend in a complex way on both the cutoff and the parameterization of each scenario.

The underlying probability model includes parameters characterizing not only treatments, courses, and covariate effects, but also all pairwise interactions between these three factors. A much simpler version of the model containing only main effects is given by $\eta_{k1}(s, \mathbf{Z}) = \mu_k + \alpha_k(s) + \sum_{j=1}^{q} \gamma_{kj} Z_j$ and $\eta_{k2}(s, t, \mathbf{Z}) = \eta_{k1}(t, \mathbf{Z}) + \beta_k$. While this model's comparative simplicity may seem appealing, its use results in greatly degraded OCs. For example, the probability of correctly selecting the optimal strategy (0,1) for (Short, Young) patients under scenario A decreases from .76 under the full model to .24 under the simpler model, while the respective probabilities of dropping the three inferior strategies decrease from .32, .34 and .29 (Table 7) to .12, .05 and .18. A similar question is what may result from basing the design on the empirical probabilities of response and death, rather than using model–based estimates.

This empirical approach reduces the correct selection probability by about .10 under each of the four scenarios. This is as expected, since the regression model borrows strength across prognostic subgroups and courses while the purely empirical approach does not.

Recall that, in developing a trial design as described in Section 3, the historical data are used only to provide the marginal posterior $f(\boldsymbol{\theta}_B \mid \mathcal{X}_H)$ of the baseline, non-treatment-related parameters. Repeating the simulations under models other than Model 6 in Table 3 showed that the operating characteristics of the AML trial design were relatively insensitive to which model was chosen, apparently because $f(\boldsymbol{\theta}_B \mid \boldsymbol{\chi}_H)$ changed very little between these models. For example, under Scenario A, using either Model 1 or Model 3 yielded selection probabilities all within .019 and early dropping probabilities all within .024 of the corresponding values for Model 6 given in Table 7, with most of the probabilities identical to two decimal places and no systematic variation. Similarly, the numbers of patients treated in each course were all within 0.16 of the corresponding values for Model 6. These differences appear to be due mainly to simulation variability.

9.3 Some Graphical Methods

Because the design accounts for multiple factors, graphical methods are especially useful for evaluating their effects. Figure 3 compares the posterior distributions of the overall response and death rates, ξ_R and ξ_D, and of the utility function, ϕ, corresponding to treatment strategies (0,1) and (0,2) in the patient subgroup with short initial CR duration ($<$ one year) and younger age (< 50). The posteriors were obtained from a single representative simulated data set under Scenario A. The left column gives the posteriors after a total of 48 patients, when the interim decision rules are applied, and the right column shows what these posteriors become once the entire trial is completed with 96 patients. The first two rows illustrate how the posterior learns that, with high probability, $\xi_R(0,1) > \xi_R(0,2)$ and $\xi_D(0,1) < \xi_D(0,2)$. It also shows that the posterior of $\xi_D(0,2)$ is still comparatively disperse even at the end of the trial. The third row illustrates the manner in which ϕ combines ξ_R and ξ_D, and that this is a useful device for combining these two probability criteria to make overall treatment comparisons.

[Figure 3 about here]

Figure 4 provides a similar illustration, but this time using only the utility function and based on a representative simulated data set under Scenario D. This figure compares the four treatment strategies within each of the four patient prognostic subgroups. Referring to the numerical values of ϕ in Table 8, the upper left graph illustrates the comparative inferiority of strategy (1,0) in the (short, old) subgroup; the upper right graph illustrates the comparative superiority of strategy (0,1) in the (short, young) subgroup; the lower left graph illustrates the comparative inferiority of strategy (1, 0) in the (long, old) subgroup; and the lower right graph illustrates the comparative superiority of strategy (0, 1) in the (long, old) subgroup. The bottom row shows that there is more variability between the four strategies in patients with a longer initial CR duration. Overall, the figure shows the great advantages of younger age and of longer initial CR duration.

[Figure 4 about here]

Figure 5 has the same structure as that of Figure 4, but with the roles of the four treatment strategies and the four patient subgroups reversed. This figure shows the great variability of the outcome criterion function ϕ across the the four prognostic groups within each treatment strategy.

[Figure 5 about here]

10. Generalizations

Numerous extensions and modifications of the design described here are possible. For example, Lavori and Dawson (2000) propose a biased-coin within–subject adaptive randomization method to compare multi-course treatment strategies. A simple generalization of the AML trial design is to allow more than two courses. This would arise, for example, in a trial of multiple treatments for a life–threatening infection, with the trinary outcome {alive and not infected, alive and infected, dead} in each course. In such settings, a patient may be treated until either the infection is resolved, the patient dies, or death is nearly certain regardless of additional treatment. This motivates the following generalization of the two–course model given by (1) – (4) to accommodate an arbitrary number of courses. Denote $\mathbf{Y}_j = (Y_{Rj}, Y_{Dj}, Y_{Fj})$ and τ_j as before, but now for $j = 1, \ldots, J$, where J is the maximum number of treatment courses. Since the patient's therapy continues beyond the j^{th} course only if it results in a failure, for notational consistency if $Y_{Fj} = 0$ we define $\mathbf{Y}_r = (0{,}0{,}0)$ for all $r > j$. Define $\pi_{k1}(s, \mathbf{Z})$ as before. For each $j > 1$, denoting $\mathbf{t}_j = (t_1, ..., t_j)$ and $\boldsymbol{\tau}_j = (\tau_1, ..., \tau_j)$, we define

$$\begin{aligned}\pi_{kj}(\mathbf{t}_j, \mathbf{Z}) &= \Pr[Y_{kj} = 1 \mid \mathbf{Z},\ Y_{F1} = ... = Y_{F,j-1} = 1 \text{ and } \tau_r = t_r,\ r = 1, ..., j] \\ &= \frac{\exp\{\eta_{kj}(\mathbf{t}_j, \mathbf{Z})\}}{1 + \ \exp\{\eta_{Rj}(\mathbf{t}_j, \mathbf{Z}) + \exp\{\eta_{Dj}(\mathbf{t}_j, \mathbf{Z})\}}, \end{aligned} \tag{14}$$

for $k = R$, D or F, where

$$\eta_{kj}(\mathbf{t}_j, \mathbf{Z}) \ = \ \mu(\mathbf{t}_j, \mathbf{Z}) + \boldsymbol{\gamma}_k \mathbf{Z} + \beta_k(\mathbf{t}_j) \tag{15}$$

with all $\eta_F(\mathbf{t}_j) = 0$. The probability of overall outcome $k = R$ or D under J–course treatment strategy $\mathbf{t}_J$ is

$$\xi_k(\mathbf{t}_J, \mathbf{Z}) \ = \ \sum_{j=1}^{J} \left\{ \prod_{r=0}^{j} \pi_{F,r}(\mathbf{t}_r, \mathbf{Z}) \right\} \pi_{k,j}(\mathbf{t}_j, \mathbf{Z}), \tag{16}$$

where $\pi_{F,0} = 1$. The likelihood (4) may now be extended to the general form

$$\mathcal{L}_i = \prod_{j=1}^{J} \left\{ \prod_{r_j = R,D,F} \left[\pi_{i,r_j,j}(\boldsymbol{\tau}_{i,j}, \mathbf{Z}_i) \right]^{Y_{i,r_j,j}} \right\}^{Y_{i,F,j-1}}, \tag{17}$$

denoting $Y_{i,F,0} \equiv 1$. Assuming a reasonably small number of possible covariate vectors $\mathbf{Z}$, the likelihood for n patients is the product multinomial

$$\mathcal{L} \ = \ \prod_{\mathbf{Z}} \prod_{j=1}^{J} \prod_{\{t_j : \mathbf{t}_j \in \mathcal{S}_j(\mathbf{Z})\}} \prod_{k_j = R,D,F} \left\{ \pi_{k_j,j}(\mathbf{t}_j, \mathbf{Z}) \right\}^{X_{k_j,j}(\mathbf{t}_j, \mathbf{Z})}, \tag{18}$$

where $X_{k,j}(\mathbf{t}_j, \mathbf{Z})$ is the number of patients in subgroup $\mathbf{Z}$ who have outcome k with treatment t_j following $j - 1$ consecutive failures with $\mathbf{t}_{j-1}$, and $\mathcal{S}_j(\mathbf{Z})$ is the set of admissible treatment sequences in that subgroup through j courses. Since

$$\sum_{\{t_j : \mathbf{t}_j \in \mathcal{S}_j(\mathbf{Z})\}} \ \sum_{k} X_{k,j}\,(\mathbf{t}_j, \mathbf{Z}) \ = \ X_{F,j-1}(\mathbf{t}_{j-1}, \mathbf{Z}),$$

in practice even for three or four courses a reasonably parsimonious parameterization of the $\pi_{k,j}(\mathbf{t}_j, \mathbf{Z})$'s will be needed.

A more complicated extension in the context of AML therapy would be to give patients who achieve CR with the treatment T_s a second course of T_s, so-called "consolidation" therapy. This is similar to the definition of patient success used in the prostate cancer trial described by Thall et al. (2000) where death during therapy is very unlikely. In this case, patient success is defined as two consecutive successful courses with the same treatment. The set of possible outcomes would be more complex since each initial CR is now partitioned into three sub-events. Given the realistic constraint that a patient may receive at most three treatment courses, each patient's therapy ends with two failures, death, or patient success as defined. The treatment assignment algorithm is illustrated by Figure 6. As before, each patient may receive either one or two different treatments. Denote the probability of outcome k with T_s in course 1 by $\pi_k(s, \mathbf{Z})$; of outcomes k_1 with T_s in course 1 and k_2 with T_t in course 2 by $\pi_{k_1,k_2}(s, t, \mathbf{Z})$; and of outcomes k_1 with T_s in the first course followed by k_2 and k_3 with T_t in courses 2 and 3 by $\pi_{k_1,k_2,k_3}(s, t, t, \mathbf{Z})$. Because the course 1 treatment is changed in course 2 if $Y_{F1} = 1$ and repeated in course 2 if $Y_{R1} = 1$, all of the course 2 probabilities are of the form $\pi_{F,k}(s, t, \mathbf{Z})$ or $\pi_{R,k}(s, s, \mathbf{Z})$, for $k = R, D, F$. Similarly, a patient receives a third course only if the first two courses have outcomes F and R with two different treatments. Thus, all three-course probabilities are of the form $\pi_{F,R,k}(s, t, t, \mathbf{Z})$.

[Figure 6 about here]

To extend the two-course model to accommodate the above three-course structure, we first note that the course 1 probabilities are unchanged. Now, recall that the course 2 outcome probability given by (2) is defined conditionally. Thus, we express the above course 2 and course 3 joint outcome probabilities as the following products of conditional probabilities:

$$\pi_{k_1,k_2}(s, t, \mathbf{Z}) \;=\; \pi_{k_2 \mid k_1}(s, t, \mathbf{Z})\; \pi_{k_1}(s, \mathbf{Z})$$

and

$$\pi_{k_1,k_2,k_3}(s, t, t, \mathbf{Z}) \;=\; \pi_{k_3 \mid k_1,k_2}(s, t, t, \mathbf{Z})\; \pi_{k_2 \mid k_1}(s, t, \mathbf{Z})\; \pi_{k_1}(s, \mathbf{Z}).$$

The linear components corresponding to each of the above conditional probabilities may be parameterized as before, but now generalizing $\beta_k(s,t)$ and the coefficient δ_{kj} of Z_j in (6). These terms become $\beta_{k|F}(s,t)$ and $\delta_{k|F,j}$ in $\eta_{k|F}(s,t,\mathbf{Z})$; and $\beta_{k|R}(s,t)$ and $\delta_{k|R,j}$ in $\eta_{k|R}(s,t,\mathbf{Z})$, for $k = R, D, F$. Similarly, for the linear terms characterizing $\pi_{k_3 \ | \ k_1,k_2}(s,t,t,\mathbf{Z})$, these parameters are $\beta_{k|F,R}(s,t,t)$ and $\delta_{k|F,R,j}$. In practice, sample size limitations likely will necessitate much more parsimonious versions of this parameterization.

A very different type of extension would utilize the times to the events, rather than discretizing them. This would require a multivariate event time model in place of the generalized logistic model, treating the times to response and failure as non-fatal competing risks, with the distribution of subsequent survival time depending on whether response or failure has occurred. The multivariate event time models proposed by Shen and Thall (1998) or Chang and Wang (1999) might be useful for this type of design. Utilizing event times could potentially provide a more informed evaluation of treatment strategies, especially since the time to achieve response has a profound effect on subsequent survival time in AML. Such a design also could account for relapse after response and the salvage therapy administered at relapse. Practical implementation would require addressing the issues of model complexity, the logistics of continuously monitoring multiple event times, and sample size.

An important question is whether Bayesian decision theory may yield a design with better properties. Such an approach could be based on the use of ϕ as a utility function, or possibly a more complex utility that also accounts for costs, as in Stallard et al. (1999). Because such an approach is very different from that taken here, it is a topic for future research.

References

Agresti, A. (1990) *Categorical Data Analysis*, New York: Wiley Interscience.

Bryant, J. and Day, R. (1995) Incorporating toxicity considerations into the design of two-stage phase II clinical trials. *Biometrics* **51**, 1372-1383.

Chang, S-H. and Wang, M-C. (1999) Conditional regression analysis of recurrence time data. *J. American Statistical Association* **94**, 1221-1230.

Dixon, D.O. and Simon, R. (1991) Bayesian subset analysis. *Biometrics* **47**, 871-881.

Faraggi, D. and Simon, R. (1998) Bayesian variable selection method for censored survival data. *Biometrics* **54**, 1475-1485.

Gibbons, J.D., Olkin, I. and Sobel, M. (1977) *Selecting and Ordering Populations: A New Statistical Methodology*, New York: John Wiley and Sons.

Gilks, W., Richardson, S. and Spiegelhalter, D. (1996) *Markov Chain Monte Carlo in Practice*, London: Chapman and Hall.

Jennison, C. and Turnbull, B.W. (1993) Group sequential tests for bivariate response: Interim analyses of clinical trials with both efficacy and safety endpoints. *Biometrics* **49**, 741-752.

Kass, R.E. and Raftery, A.E. (1995) Bayes Factors. *J. American Statistical Association* **90**, 773-795.

Lavori, P.W. and Dawson, R. (2000) A design for testing clinical strategies: biased adaptive within-subject randomization. *J. Royal Statistical Society, A* **163**, Part 1, 29-38.

Lindley, D.V. and Smith, A.F.M. (1972) Bayesian estimates for the linear model (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 1-41.

Shen, Y. and Thall, P.F. (1998) Parametric likelihoods for multiple non-fatal competing risks and death. *Statistics in Medicine*, **17**, 999-1016.

Stallard, N., Thall, P.F. and Whitehead, J. (1999) Decision theoretic designs for phase II clinical trials with multiple outcomes. *Biometrics*, **55**, 971–977.

Thall, P.F. and Cheng, S.-C. (1999) Treatment comparisons based on two-dimensional safety and efficacy alternatives in oncology trials. *Biometrics*, **55**, 746-753.

Thall, P.F., Millikan, R.E. and Sung, H.G. (2000) Evaluating multiple treatment courses in clinical trials. *Statistics in Medicine*, **19**, 1011-1028.

Thall, P.F., Simon, R. and Shen, Y. (2000) Approximate Bayesian evaluation of multiple treatment effects. *Biometrics*, **56**, 213-219.

Willan, A.R. and Pater, J.L (1985) Hypothesis testing and sample size for bivariate binomial response in the comparison of two groups. *J. Chronic Diseases* **38**, 603-608.

Table 1. Outcome counts for each course and treatment combination in the historical AML data. Row probabilities are given in parentheses. 0 = high dose ara–C, 1 = allogeneic bone marrow transplant, and 2 = chemotherapy without ara–C

Treatment t_1	Course 1 Outcome CR	 Death	 Failure	Total
0	84 (.27)	66 (.21)	166 (.52)	316
1	50 (.56)	18 (.20)	21 (.24)	89
2	13 (.04)	41 (.13)	255 (.82)	309

Treatments t_1	 t_2	Course 2 Outcome CR	 Death	 Failure	Total
0	0	14 (.17)	24 (.29)	44 (.54)	82
0	1	5 (.36)	5 (.36)	4 (.29)	14
0	2	0 (.00)	5 (.22)	18 (.78)	23
1	0	1 (.14)	5 (.71)	1 (.14)	7
1	1	1 (.50)	0 (.00)	1 (.50)	2
1	2	0 (.00)	0 (.00)	3 (1.00)	3
2	0	4 (.11)	12 (.34)	19 (.54)	35
2	1	3 (.33)	3 (.33)	3 (.33)	9
2	2	4 (.02)	26 (.16)	129 (.81)	159

Table 2. Induction mortality rates in AML at M.D. Anderson, 1991-1999

			Number of Deaths (%)	
Performance Status	Age	No. Patients	by Day 28	by Day 56
0, 1, or 2	< 50	372	19 (5)	30 (8)
	50 - 59	235	16 (7)	26 (11)
	60 - 69	313	31 (10)	53 (17)
	70 - 79	260	34 (13)	55 (21)
	> 79	43	13 (30)	17 (40)
3 or 4	< 50	41	16 (39)	18 (44)
	50 - 69	76	35 (46)	42 (55)
	> 69	54	31 (57)	36 (67)

Table 3. Summary of models fit to the historical data

Model	Description	p	$\log \mathcal{L}(\hat{\boldsymbol{\theta}})$	BIC
1	Full	40	–795.96	–927.38
2	$\boldsymbol{\zeta} = \mathbf{0}$	32	–800.98	–906.12
3	$\boldsymbol{\beta}^{*} = \mathbf{0}$	24	–801.19	–880.94
4	Model 3 – $\boldsymbol{\zeta}_{AGE}(1)$	22	–801.48	–873.76
5	Model 4 – $\boldsymbol{\zeta}_{AGE}(2)$	20	–801.91	–867.62
6	Model 5 – $\boldsymbol{\zeta}_{DUR}(2)$	18	–803.22	–862.36
7	$\boldsymbol{\zeta} = \mathbf{0}, \boldsymbol{\beta}^{*} = \mathbf{0}$	16	–806.55	-859.12

Table 4. Maximum likelihood estimates and posterior means under generalized logistic model 6 for the historical data. Standard deviations are given in parentheses.

		Bayesian Estimates	
Parameter	ML Estimate	Approximate	MCMC
μ_R	–1.350 (.249)	–1.338 (.244)	–1.352 (.207)
$\alpha_R(1)$	1.740 (.304)	1.723 (.301)	1.744 (.303)
$\alpha_R(2)$	–2.143 (.281)	–2.132 (.280)	–2.166 (.282)
μ_D	–0.685 (.197)	–0.682 (.195)	–.685 (.165)
$\alpha_D(1)$	0.563 (.355)	0.562 (.349)	0.560 (.338)
$\alpha_D(2)$	–1.061 (.189)	–1.058 (.188)	–1.067 (.178)
β_R	–0.458 (.601)	–0.474 (.576)	–0.507 (.428)
β_D	0.467 (.280)	0.456 (.276)	0.458 (.240)
$\gamma_{R,DUR}$	1.570 (.270)	1.545 (.266)	1.562 (.253)
$\gamma_{R,AGE}$	0.223 (.279)	0.222 (.273)	0.225 (.231)
$\gamma_{D,DUR}$	0.004 (.317)	–0.007 (.312)	–0.018 (.286)
$\gamma_{D,AGE}$	–0.440 (.253)	–0.439 (.250)	–0.447 (.220)
$\zeta_{R,DUR}(1)$	–0.263 (.746)	–0.277 (.713)	–0.162 (.719)
$\zeta_{D,DUR}(1)$	1.365 (.850)	1.307 (.809)	1.421 (.794)
$\delta_{R,DUR}$	–0.639 (.564)	–0.599 (.549)	–0.618 (.502)
$\delta_{R,AGE}$	0.078 (.661)	0.081 (.636)	0.095 (.481)
$\delta_{D,DUR}$	–0.989 (.543)	–0.949 (.532)	–0.990 (.508)
$\delta_{D,AGE}$	0.140 (.400)	0.141 (.394)	0.144 (.346)

Table 5. Estimated two–course probabilities of response and death, and objective function values, by prognostic group, for historical patients treated with high–dose ara–C in both courses. Standard deviations are given in parentheses.

	Model Based			Empirical			
(CR Duration, Age)	$\widehat{\xi}_R$	$\widehat{\xi}_D$	$\widehat{\phi}$	$\widehat{\xi}_R$	$\widehat{\xi}_D$	$\widehat{\phi}$	n
(Short, Old)	.19 (.05)	.52 (.07)	–.95 (.31)	.16 (.04)	.44 (.05)	–.90 (.17)	29
(Short, Young)	.27 (.10)	.40 (.10)	–.42 (.54)	.31 (.05)	.40 (.06)	–.31 (.22)	30
(Long, Old)	.54 (.12)	.25 (.09)	.85 (.60)	.63 (.07)	.27 (.06)	1.10 (.26)	11
(Long, Young)	.65 (.13)	.16 (.08)	1.48 (.65)	.62 (.06)	.16 (.05)	1.40 (.25)	12
Average	.33 (.04)	.38 (.04)	–.16 (.24)	.35 (.03)	.36 (.03)	–.08 (.11)	82

Table 6. The first four clinical scenarios. Tabled values correspond to the patient prognostic group with short CR duration and younger age

Null Scenario

Strategy	ξ_R	ξ_D	ϕ
(0,1)	.22	.32	–.39
(0,2)	.22	.32	–.39
(1,0)	.22	.32	–.39
(2,0)	.22	.32	–.39

Scenario A

Strategy	ξ_R	ξ_D	ϕ
(0,1)	.46	.25	.55
(0,2)	.22	.32	–.39
(1,0)	.22	.32	–.39
(2,0)	.22	.32	–.39

Scenario B

Strategy	ξ_R	ξ_D	ϕ
(0,1)	.33	.32	–.04
(0,2)	.22	.32	–.39
(1,0)	.54	.25	.83
(2,0)	.22	.32	–.39

Scenario C

Strategy	ξ_R	ξ_D	ϕ
(0,1)	.35	.50	–.40
(0,2)	.22	.32	–.39
(1,0)	.54	.30	.72
(2,0)	.22	.32	–.39

Table 7. Operating characteristics of the design under scenarios A, B and C for the prognostic subgroup with short CR duration and younger age. Correct decision probabilities are enclosed in boxes.

	Treatment		Decision Probabilities		# Patients Treated	
Scenario	Strategy	ϕ	Selected	Dropped Early	Course 1 Only	Two Courses
A	(0,1)	.55	$\boxed{.76}$	.03	3.0	4.7
	(0,2)	–.39	.04	$\boxed{.32}$	2.0	3.3
	(1,0)	–.39	.08	$\boxed{.34}$	3.9	6.5
	(2,0)	–.39	.11	$\boxed{.29}$	4.2	6.7
B	(0,1)	–.04	.12	.27	2.1	3.4
	(0,2)	–.39	.04	$\boxed{.37}$	1.9	3.0
	(1,0)	.83	$\boxed{.78}$	.04	9.8	4.0
	(2,0)	–.39	.06	$\boxed{.34}$	3.8	6.3
C	(0,1)	–.40	.09	$\boxed{.36}$	2.0	3.2
	(0,2)	–.39	.06	$\boxed{.31}$	2.0	3.2
	(1,0)	.72	$\boxed{.77}$	.04	10.8	2.8
	(2,0)	–.39	.08	$\boxed{.30}$	4.0	6.4

Table 8. Operating characteristics of the design under scenario D, where strategy $(0, 1)$ is superior overall but but T_1 in either course greatly increases the death rate in older patients. Correct decision probabilities are enclosed in boxes.

Prognostic Group	Treatment		Decision Probabilities		# Patients Treated	
(CR Dur, Age)	Strategy	ϕ	Selected	Dropped Early	Course 1 Only	Two Courses
(Short, Old)	(0,1)	–1.33	.13	$\boxed{.38}$	2.8	3.8
	(0,2)	–.91	$\boxed{.44}$	.07	4.0	5.1
	(1,0)	–2.32	.00	$\boxed{.89}$	7.0	0.6
	(2,0)	–.91	$\boxed{.43}$	.16	6.9	9.2
(Short, Young)	(0,1)	.55	$\boxed{.73}$	.06	2.8	4.6
	(0,2)	–.39	.06	$\boxed{.30}$	2.0	3.3
	(1,0)	–.39	.10	$\boxed{.28}$	4.2	6.8
	(2,0)	–.39	.12	$\boxed{.30}$	4.0	6.6
(Long, Old)	(0,1)	–.17	.24	$\boxed{.30}$	1.4	0.8
	(0,2)	–.03	$\boxed{.35}$	.17	1.6	0.9
	(1,0)	–2.21	.00	$\boxed{.86}$	2.1	0.1
	(2,0)	–.03	$\boxed{.42}$	.27	2.7	1.5
(Long, Young)	(0,1)	1.48	$\boxed{.42}$	.19	1.3	0.7
	(0,2)	.73	.10	$\boxed{.36}$	1.1	0.6
	(1,0)	.73	.24	$\boxed{.28}$	2.4	1.4
	(2,0)	.73	.24	$\boxed{.27}$	2.4	1.3

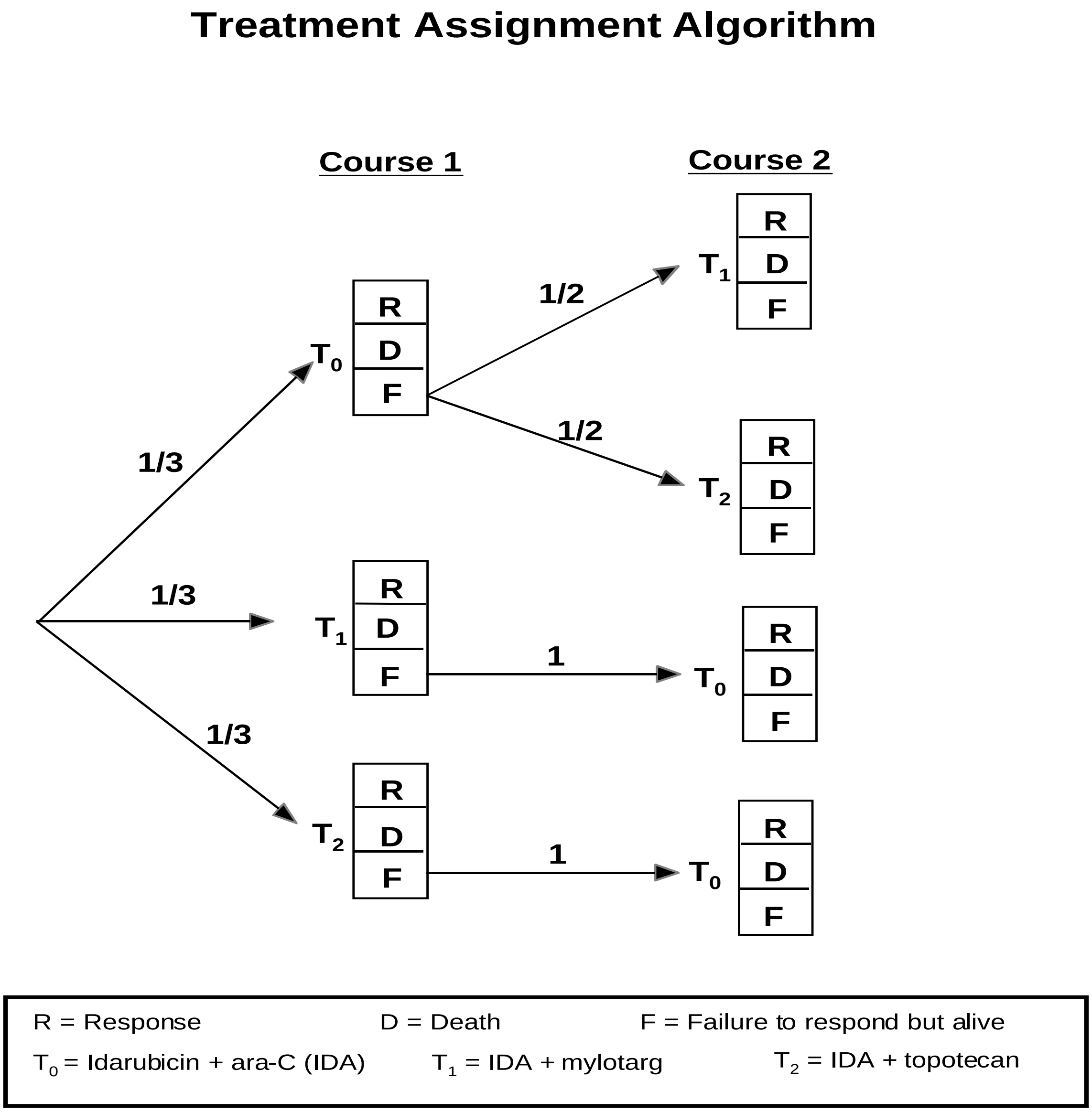

Figure 1: Treatment assignment algorithm for the AML trial

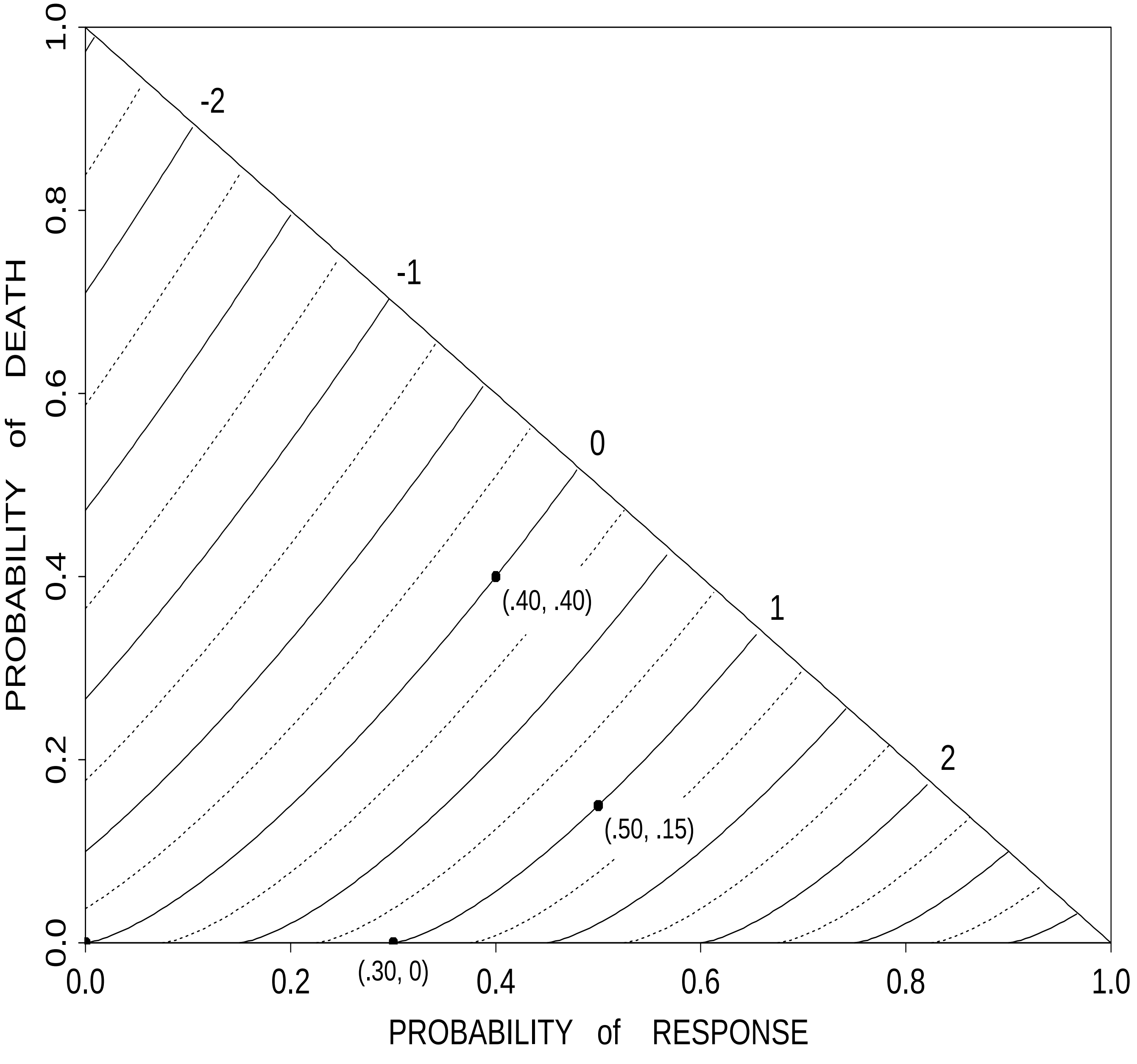

Figure 2: Contours of the objective function $\phi = 3.333\ \xi_R - 2.548\ \xi_D^{.707}$

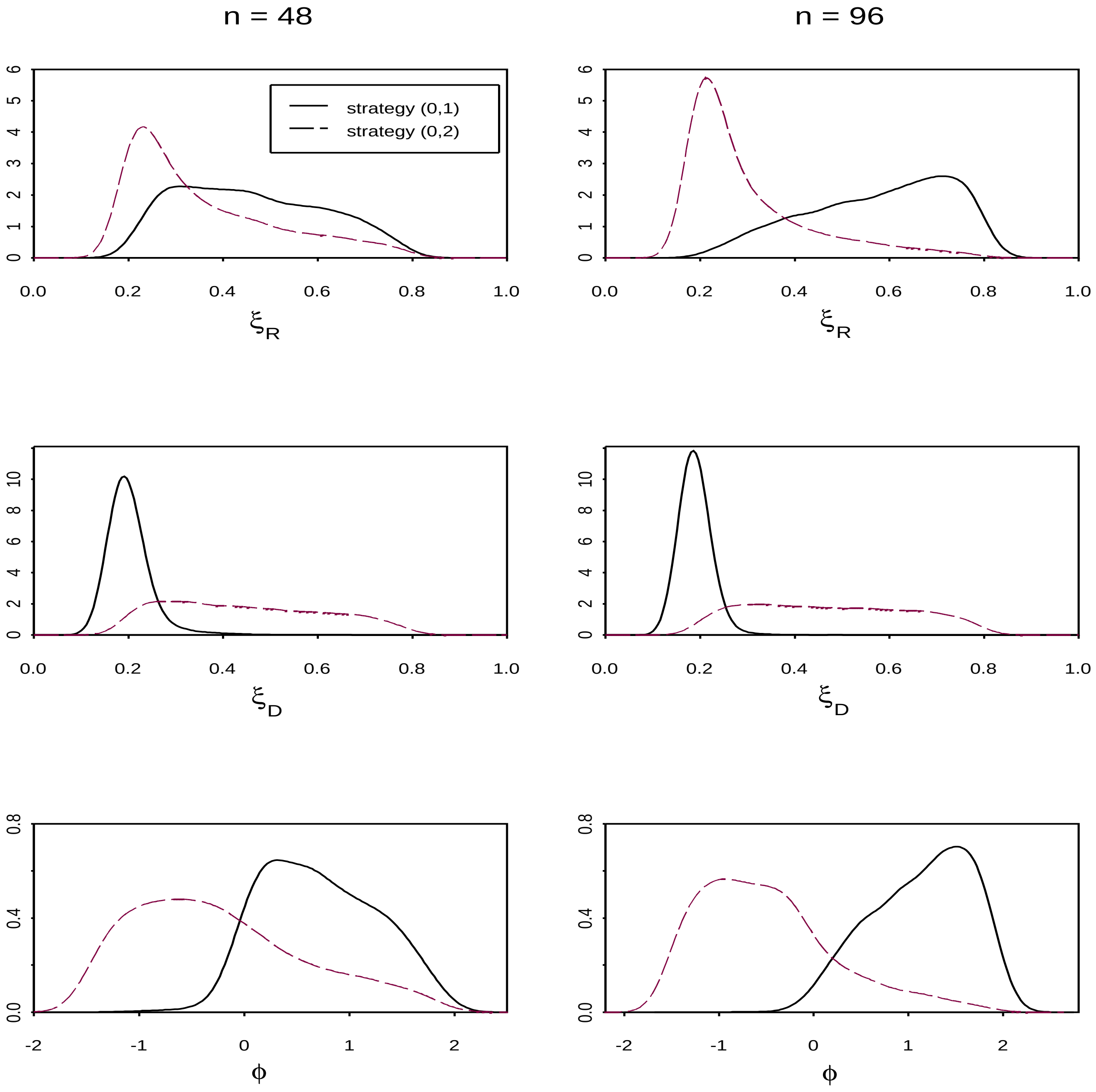

Figure 3: Comparisons of treatment strategies (0,1) and (0,2) in terms of ξ_R, ξ_D, and ϕ in the patient subgroup with short initial CR duration and younger age under scenario A

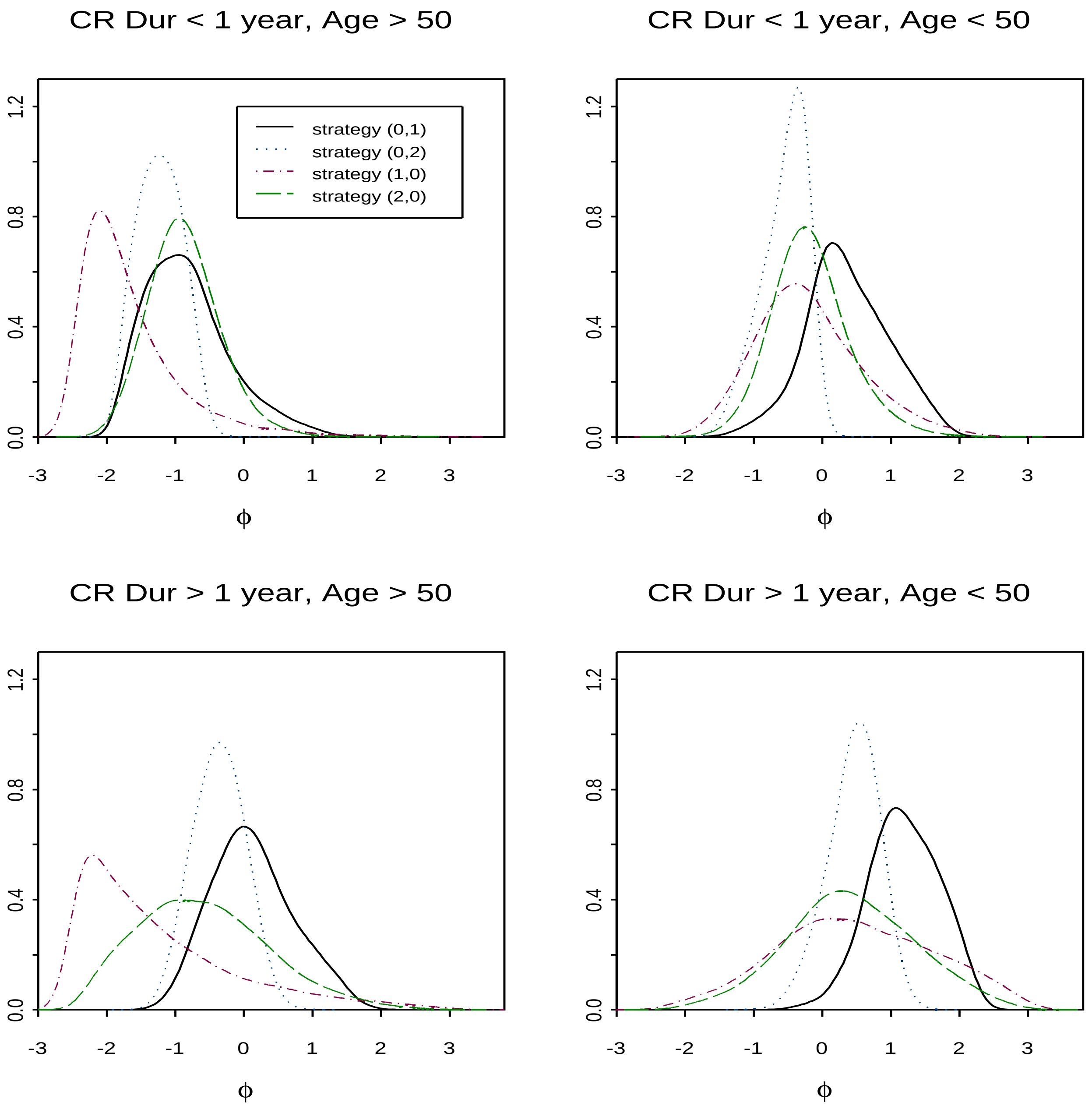

Figure 4: Comparisons of the four treatment strategies in terms of ϕ in each patient subgroup under scenario D

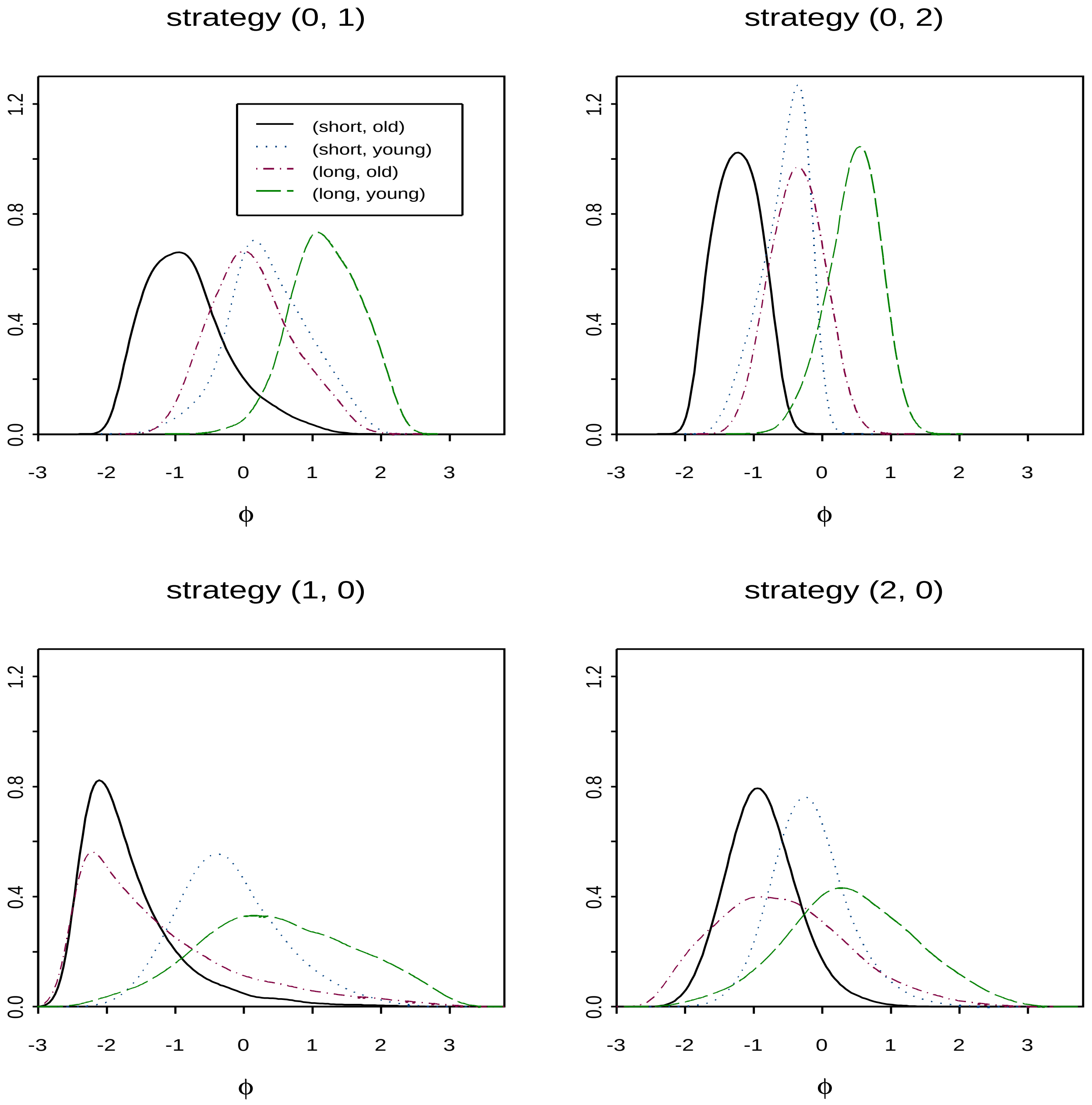

Figure 5: Comparisons of the four patient subgroups under each treatment strategy in terms of ϕ under scenario D

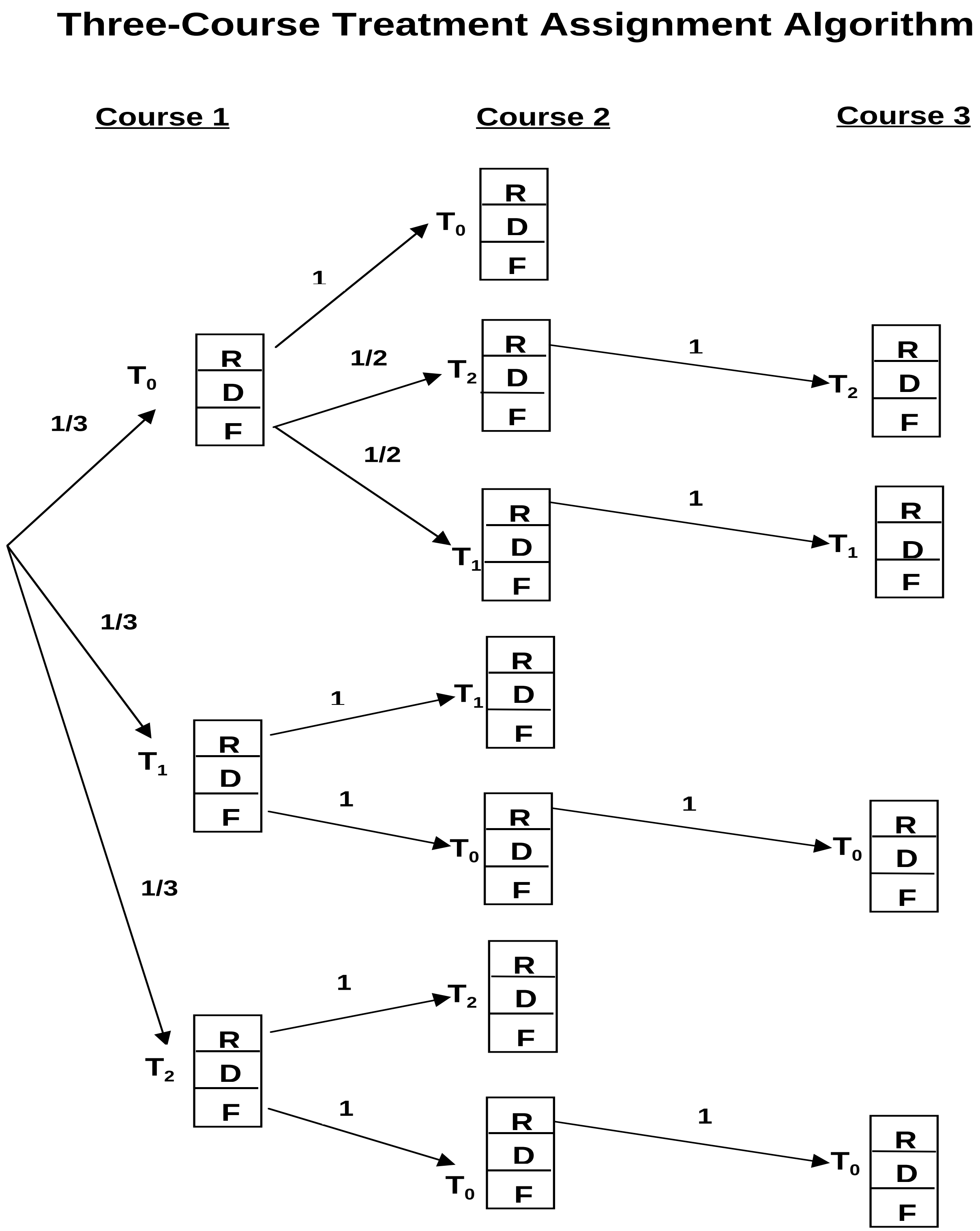

Figure 6: Treatment assignment algorithm for an extended design defining patient success as two consecutive CRs