

*University of Texas, MD Anderson Cancer
Center*

UT MD Anderson Cancer Center Department of Biostatistics
Working Paper Series

Year 2005

Paper 17

Some Ethical Issues in Phase II Trials in Acute
Leukemia

Peter F. Thall*

Elihu H. Estey[†]

*U.T.M.D. Anderson Cancer Center, rex@mdanderson.org

[†]U.T.M.D. Anderson Cancer Center, ehestey@mdanderson.org

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

<http://biostats.bepress.com/mdandersonbiostat/paper17>

Copyright ©2005 by the authors.

Some Ethical Issues in Phase II Trials in Acute Leukemia

Peter F. Thall and Elihu H. Estey

Departments of Biostatistics and Applied Mathematics (PFT) and Leukemia (EHE),
The University of Texas M.D. Anderson Cancer Center
1515 Holcombe Boulevard, Houston, Texas 77030

August 1, 2005

To appear in *Clinical Advances in Hematology and Oncology*



Introduction

Three medical approaches are possible for any given illness: no treatment, treatment with standard therapy, and treatment with investigational therapy. Because few untreated patients with acute myeloid leukemia (AML) live more than one year, and because, in most cases, standard treatment does not improve prognosis, guidelines promulgated by academic cancer centers typically recommend investigational therapy for many such patients. These guidelines also advise that investigational treatments be administered within the context of a formal clinical trial in order to gain knowledge about their actual effects on patients. Because in general patients rely on the expertise of their physicians when making treatment decisions, and a wide variety of treatments for AML are available, nearly all AML patients simply accept their physician's recommendation. Thus, when a physician advises an AML patient to enter a trial of an investigational agent, the patient undoubtedly assumes that this will serve his or her best interests. It has become apparent to us, however, that for many patients the recommended trial, as designed, is not their best option. This gap between assumption and reality results from two aspects of clinical trial design. First, the ethical underpinnings of such designs can be questioned, especially with regard to the direct consequences to a trial participant. Second, the scientific structure of these designs is frequently inadequate, limiting what may be learned about new treatments. In particular, we will argue that many designs waste patient data. This is at odds with good statistical and scientific practice, as well as patients' often cited desire to benefit future patients as motivation for participation in a clinical trial.

Because we believe that the first issue, the dubious ethical bases for some clinical trials, is the more serious, this will be our primary focus. We will discuss the following specific points: The common practice of targeting inappropriately low response rates, the fact that many designs are insufficiently adaptive, patient heterogeneity, and the practice of focusing on the activity of a single regimen. We will provide numerical illustrations of these points based on comparisons of alternative clinical trial designs. Although we will focus on trials in acute leukemia, the methods that we will discuss have a much wider range of application.

Low Target Response Rates

As a basis for illustration, we will use a multi-center phase 2 trial of the new drug VNP. The trial began in 2004 and enrolled patients if they had relapsed or refractory AML and a first remission lasting < 1 year; or if they had untreated AML and were age 59 or above ("older"). The trial used an optimal 2-stage Simon design (1989). The protocol made no distinction between relapsed/refractory and untreated older patients. The targeted response (CR) rate was set at 20% ($p_1 = 0.20$), with a null rate of 5% ($p_0 = 0.05$), with type I and type II error probabilities both 0.10. That is, the design had a 90% probability

of rejecting the drug for future study if its true response rate was “of no interest”, corresponding to the null hypothesis that $p = \text{prob}(\text{CR}) = p_0 = 0.05$, and it had a 90% probability (power) of accepting the drug for future study if its true response rate was “of interest”, here the alternative hypothesis that $p = p_1 = 0.20$. The Simon optimal 2-stage design with these operating characteristics treats 12 patients in stage 1, stops accrual with acceptance of the null hypothesis if 0 responses are observed in these 12 patients, enters 25 more patients in a second stage if at least 1 of the first 12 respond, and accepts the alternative hypothesis if 4 or more responses are observed in the 37 patients accrued in both stages combined. All of this was based on the assumption that the probability of CR with VPN would be the same in relapsed/refractory and untreated older patients.

Since the null CR rate of $p_0 = 0.05$ used in applying the Simon design presumably reflected historical data, it is instructive to examine actual CR rates following the use of standard treatment (ara-C-containing regimens, and specifically ara-C + idarubicin) as given at M.D. Anderson Cancer Center (MDACC) from 1990-2004. The CR rate for patients with relapsed or refractory AML was 0.19 (69/356, 95% confidence interval ci, 0.16 – 0.24), while that for untreated patients age 59 and above was .45 (46/103, 95% ci 0.35 – 0.54). Thus, the targeted rate of interest specified in the protocol, ($p_0 = 0.20$) represented no improvement over the CR rate available with standard therapy for relapsed/refractory patients, while this target well below the CR rate that could almost certainly be obtained with standard therapy in untreated older patients. This illustrates two obvious flaws with the design: It failed to discriminate between two patient subgroups having very different prognoses, and the targeted rate was inappropriate each of the subgroups, being too low in one and far too low in the other.

We now evaluate the consequences of using the 2-stage Simon design with a null rate of $p_0 = 0.05$ and a desired improvement of $p_1 - p_0 = 0.15$ to achieve the targeted $p_1 = 0.20$, i.e. the design used in the VNP trial. For simplicity, we will focus on the subgroup of untreated older patients and assume a null CR rate of 0.40, which is slightly below the historical rate in these patients. Once we have explored this simple case, we will return to the issue of patient heterogeneity. We thus focus on a design with $p_0 = 0.40$, and use the same type I and type II error rates of 0.10 and the same desired 0.15 improvement, so that the targeted alternative is $p_1 = 0.55$. Because this design has a much larger null rate of $p_0 = 0.40$ rather than $p_0 = 0.05$, it requires a much larger number of patients to detect the same improvement of 0.15 with the same reliability. Specifically, for the optimal 2-stage Simon design in this case, 38 patients are enrolled in stage 1 and accrual stops with acceptance of the null if 16 or fewer responses are seen in these first 38 patients. If at least 17 responses are observed in the first stage, then another 50 patients are entered in the second stage for a maximum total sample size of $n_{\text{max}} = 88$. The alternative is accepted, and the drug is considered promising, if at least 41 of the 88 patients respond.

With both designs, we will also assume that if the trial is terminated after the first stage, then the remaining patients (i.e. those who would have been accrued in the second stage) are given the standard therapy, recalling that under the null hypothesis, as based on historical data, this therapy produces (at least) a 40% response rate in untreated patients. We will then compute (table 1) the expected number of responses lost compared to the simple plan of treating all n_{max} patients (37 with the design that was used, 88 with the alternative design) with standard therapy. For example when the true response rate with the experimental drug (here VNP) is $p = 0.40$, the expected number

of responses lost is zero, since this is also the response rate with standard therapy. However, the employed design is very unlikely to stop the trial after the first stage even when $p = 0.10$ (0.30 below the standard CR rate), with a probability of early termination, hereafter PET, of 0.28. In contrast, the PET values are much higher with the alternative design for small values of p , so this design has a much smaller percentage of lost responses compared to treating all patients with standard therapy. Thus with the design that was used, characterized by specifying $p_0 = 0.05$, if the true response rate is $p = 0.20$, the targeted value, then 7.7 responses would be expected vs. 14.8 responses if all 37 patients had received standard therapy. On average, the design that was used achieves only 52% of the number of responses that would be expected by simply treating all patients with standard therapy, equivalently, 48% of the expected responses are lost. Hereafter, this quantity is called ERL. In contrast, under the same scenario, the design that uses the more appropriate $p_0 = 0.40$, based on historical data, results in an ERL of only 22%.

The VNP design is not exceptional; many phase II designs specify inappropriately low values for p_0 . We suspect that this may be done due to confusion between a response rate that may be of interest based on evidence that a drug is “active,” and a response rate that is of interest based on evidence that the drug is better than standard therapy. The above illustration is motivated by the belief that the latter is of greater fundamental interest to patients, and thus should be the basis of an ethical design. Regardless of the explanation, however, we hope that the above example, and the more extensive analysis in table 1, make clear the clinical consequences of targeting an inappropriately low response rate.

A More Adaptive Design

Although the second design in table 1 is preferable to the first, it still suffers from the fact that, because it is a 2-stage design, in any case it cannot stop until all patients in stage 1, here 38, have received the experimental therapy. Thus, even if the true CR rate is a very small value substantively inferior to standard therapy, with $p = .20$ or less, it still must treat at least 38 patients with the experimental regimen. For example, suppose none of the first 18 patients achieved a CR. Then assuming, from a Bayesian viewpoint, that p followed a $\text{beta}(.20, .80)$ prior, which has expected value 0.20 and effective prior sample size equal to one patient, the posterior probability based on a run of 18 treatment failures that $p > 0.20$ would be 0.001, very strong evidence that even the 0.20 CR rate is unlikely. This is just a statistical way of quantifying how unpromising a treatment that yields 0/18 responses actually is, and it leads to the natural question of why a physician would want to treat an additional 20 patients before applying a stopping rule, or even enroll a 19th patient in the trial. Clearly, it would be very desirable to use a design that allows one to stop earlier in such cases. That is, a more adaptive design with more and earlier decisions based on the incoming data is needed. While there are numerous “frequentist” designs based on tests of hypotheses that do this (cf. Fleming, 1989; Chen, 1997), for simplicity we will describe a Bayesian design that allows several interim looks. This is an example of a family of Bayesian phase II designs that are extremely flexible (Thall, Simon and Estey, 1995, 1996), and whose use has been validated in numerous phase II trials over the past decade.

To provide more adaptive monitoring, we constructed a Bayesian design with the same $n_{\max} = 88$ patients and based on the same average null and alternative values $p_0 = 0.40$ and $p_1 = .55$, but with interim stopping rules applied after every 15 patients have been

treated and evaluated. Thus, there are up to five interim analyses, at 15, 30, 45, 60 and 75 patients. The Bayesian design assumes that p_0 and p are random, with p_0 following $\text{beta}(400,600)$ prior, and p following a $\text{beta}(.80,1.20)$ prior. Both priors have mean 0.40, but the historical standard prior is very informative while that of the experimental agent's CR probability is very uninformative, equivalent to having information on 2 patients. The trial is stopped early if the posterior probability $\Pr(p_0 + .15 < p \mid \text{data}) < 0.04$ at any interim look, which translates to stopping if the $[\# \text{ CRs}]/[\# \text{ patients evaluated}] \leq 4/15, 11/30, 18/45, 26/60$ or $33/75$. Details of how to construct this sort of design are given in Thall and Sung (1998), and a computer program for carrying out the computations is freely available at <http://biostatistics.mdanderson.org>. Table 2 provides a comparison of the operating characteristics of this design to the Simon two-stage design with these values of p_0 and p_1 . Because the decision that the new treatment is not superior and that the trial should be stopped early can be made much more quickly with multiple interim looks at the data, the multi-stage design is much more efficient. While the two designs have nearly identical values of PET for the targeted $p = 0.55$, the multi-stage design has much larger PET values for smaller values of p . The ethical consequence of this is seen in the expected numbers of responses achieved for $p = 0.40$ or smaller, which are much larger for the multi-stage design, and so the corresponding number responses lost is much smaller. For desirably large values of $p = 0.50$ or 0.55 , the two designs have virtually identical operating characteristics. This illustrates the general fact that a properly calibrated design with more interim looks at the data is safer without sacrificing the ability to identify a true treatment advance.

It is important to emphasize that we are not arguing against the use of 2-stage designs, but rather against their misuse. Indeed, when Simon first introduced the 2-stage designs in 1989 along with good quality computer code for implementation, they provided a substantive improvement over single-arm trials conducted without any interim stopping rule at all. Our first point has been that early stopping rules do not function well when the design has not been parameterized properly to reflect comparison to the actual historical rate. Such an improper parameterization is likely to render any statistical design dysfunctional. The second point is that an early stopping procedure should be constructed so that it stops the trial as soon as the interim data show that it is likely that an experimental treatment is not promising. In a trial where the best available treatment has $p_0 = .05$ or smaller, the goal is essentially to detect whether the new agent has any anti-disease activity, and in such cases a target $p_1 = 0.20$ is scientifically and ethically appropriate. In such activity trials, stopping early does not really protect patients, but rather it clears the way for study of newer experimental agents. When there is a standard therapy with a substantively large value of p_0 , however, early stopping rules have very important ethical consequences (Thall, 2002).

Patient Heterogeneity

The previous example has focused on a trial of the subgroup of untreated older patients for the sake of illustration. Recall, however, that the VNP trial also included relapsed/refractory patients, and that the historical CR rates of 0.19 and 0.45 in the two groups were very different. Thus, for example, a new treatment achieving an actual CR rate of 0.40 in relapsed/refractory patients would be a desirable treatment advance, while this rate would not provide an improvement over the standard in untreated older patients. The point is that, since these two groups have very different CR rates with standard therapy, what constitutes an improvement over standard therapy also is different in the two groups. Consequently, it does not make sense to conduct a trial

including both groups that has one overall targeted CR rate. There are two sensible alternative approaches to deal with this problem. The first, which is very simple, is to conduct separate trials in the two groups. For example, the trial in the relapsed/refractory group might be based on $p_0 = 0.19$ and $p_1 = .19 + .15 = .34$, and the trial in the untreated older patients might use $p_0 = 0.45$ and $p_1 = .45 + .15 = .60$, or possibly other improvements than .15 could be targeted. This solves the problem of patient heterogeneity, but it has the undesirable property that it does not allow one to borrow strength between the two subgroups. That is, if an improvement over the historical rate is seen in relapsed/refractory patients, this should provide evidence that an improvement in untreated older patients also is likely. To account for this, one may run a single trial including both subgroups, while using a regression model to account for prognosis. This may be done in numerous ways. To illustrate this sort of statistical methodology, one simple approach is to define the covariate $Z = 1$ if the patient is untreated older and $Z = 0$ if the patient is relapsed/refractory, let p_Z denote the CR rate in subgroup Z , and assume the logistic regression model $\text{logit}(p_Z) = \log\{p_Z/(1-p_Z)\} = \alpha + \beta Z$. Then $\text{logit}(p_Z) = \alpha + \beta$ for untreated older patients and $\text{logit}(p_Z) = \alpha$ for relapsed refractory patients, with the parameter β accounting for prognostic subgroup. Early stopping criteria could then utilize data from both subgroups. Many versions of such covariate-adjusted phase II designs are possible (cf. Thall, Sung and Estey, 2002; Thall, et al., 2003; Thall and Wathen, 2005). The point is that it is important to account for patient heterogeneity when it is large enough so that a design that ignores heterogeneity does not make sense.

The Fallacy of Single Arm Phase 2 Trials

We have argued that it is ethically important to set standards in a phase II trial that reflect comparison to what can be achieved with standard therapy, and moreover that it may be very desirable to monitor the accruing data more intensively than taking only one interim look. However, the cases that we have examined greatly simplify actual clinical settings, and numerous complicating issues remain. These include multiple outcomes including adverse treatment effects, the fact that early patient outcomes such as CR may be inadequate surrogates for survival time, and settings where multiple courses of therapy are given over an extended period of time. Although we cannot deal with all of these issues here, we will briefly address the issue of treatment-trial confounding.

The observed difference in outcome between two treatments as administered in two separate single arm trials is the sum of (a) the actual difference between the treatments, typically called the “treatment effect” (b) differences due to observable patient prognostic covariates, and (c) differences due to unobserved variables in the patients and the therapeutic environments of the two trials. The variables causing this third class of effects are sometimes referred to as “latent variables” and their combined effect as “trial effect” (Thall and Wang, 2005). Examples include differences in supportive care practices, the types of patients that are enrolled, and the skill of the physicians and nurses caring for the patients. Although differences in quantifiable covariates (e.g. age, cytogenetics) can be accounted for, the same is not true of trial effects. Thus, when comparing treatments that have been evaluated in separate trials, which in particular includes comparison to historical data, the treatment and trial effects are completely confounded. Indeed, it has been demonstrated that differences in outcome beyond those attributable to the play of chance arise when the same treatment is given in two separate trials, and that these differences persist even after accounting for the effects of known prognostic covariates (Estey and Thall, 2002). This observation of course

motivates the accepted use of the randomized phase 3 trial as the arbiter of the superiority of one treatment over another. Given this, it seems peculiar that the decision to proceed to a phase 3 trial of a new drug is commonly based on the performance of the drug in a single arm phase 2 trial, despite the confounding between treatment and trial effects inherent in the evaluation of data from such trials. This practice can also be criticized from a patient's perspective. The most common question posed by patients to physicians involved in trials of new drugs is "which of the drugs that you have available is best". This question indicates that patients view phase 2 trials as inherently comparative. It follows that patients' interests are poorly served by single arm phase 2 trials.

A simple alternative is to conduct a randomized phase 2 trial that employs a "selection design" that randomizes patients among several treatments, including one or more experimental regimens and possibly the standard. The objective is to select one experimental treatment that is best (Thall, Simon and Ellenberg, 1988), although designs that allow more than one experimental treatment to be selected certainly may be used (Schaid, Wieand and Therneau, 1990). This does away with trial effects, and thus ensures unbiased comparisons of the experimental therapies to each other and to the standard. Designs that do not include the standard, but rather randomize patients among two or more experimental treatments, are also useful because they provide unbiased comparisons among the treatments studied, hence the selection is much more reliable than if a sequence of single-arm trials were conducted (Simon, Wittes and Ellenberg, 1985). In any case, the selected therapy or therapies may be studied further subsequently, e.g. in comparison to the standard treatment in terms of survival or disease-free survival (DFS) time.

A selection design that aims to select the best among several experimental therapies regardless of the difference in success rates among the therapies requires fewer patients (e.g. a maximum of 15-20 per treatment) than are required by designs (e.g. the Simon 2-stage) whose goal is to test the hypothesis that a given therapy is better than another by at least a given amount. Simulation studies typically indicate that the probability of selecting a truly superior therapy is 60-70%, corresponding to a power of 60-70%. Because physicians are accustomed to 80-90% power, the selection design is sometimes criticized as an "underpowered phase 3 trial". However such criticism ignores the fact that selection designs have the less demanding goal of choosing a best treatment, rather than demonstrating a given degree of improvement, as well as the fact that any experimental therapy selected in this way still must reliably show an improvement in survival or DFS in a subsequent trial. That is, the randomized selection design is not a substitute for a confirmatory phase III trial, but simply a much more efficient way to screen new therapies. Randomized selection designs are especially useful in settings where there are several new agents that might be tested. Experience suggests that, at least in AML, pre-clinical rationale cannot substitute for clinical data in deciding which new agent is best. Yet experience also suggests that the decision as to which new agent to test is made informally, i.e. in the absence of such data. For example, suppose that three new agents are available for evaluation in patients with relapsed/refractory AML, one must be selected for a subsequent comparison with standard therapy, and the pre-clinical rationale for using each seems equally compelling. It follows that the probability of selecting the best agent is 0.33. It is this figure, not 80-90%, that should be compared with the 60-70% correct selection probabilities that typify selection designs. Such designs thus should be viewed as an attempt, using relatively

few patients, to substitute empiricism for informality in selecting which new drugs are worthy of further investigation.

By extension, we view the phase 2-3 dichotomy as artificial. In addition to the non-randomized nature of the single arm phase 2 trial, the distinction between phase 2 and phase 3 prevents use of phase 2 data in final evaluation of a drug because the data did not arise from a randomized trial. This is particularly unfortunate because, with respect to important endpoints such as survival, the phase 2 data are the most mature. Inoue, Berry, and Thall (2002) and Liu and Pledger (2005) have proposed phase 2-3 designs that randomize patients throughout. In the Inoue et al. design, at each of several interim analyses a decision is made to stop the trial if one treatment is superior, to stop if it is implausible that any treatment will be superior (“futility”), or to expand the trial to include other centers if the accumulating data indicate that many more patients will be needed to reach conclusions. At this point the phase 3 aspect of the phase 2-3 design is said to start. Such designs are intuitively appealing, and have been demonstrated to result in substantial savings of both time and sample size.

Conclusions

There currently is a cultural gap between physician and statistician, with the former often viewing the latter as divorced from clinical reality. While, unfortunately, this viewpoint is accurate in many cases, the segment of the biostatistical community that works closely with physicians has provided a wide array of practical clinical trial designs that have ethically desirable properties. We hope that this paper will convince the reader that medical statistics has important ethical dimensions, and that science and ethics cannot be separated when designing a clinical trial. In particular, we have argued that setting target response rates too low or looking too infrequently at available clinical trial data may result in lost responses, and that the focus on single arm phase 2 clinical trials may be antithetical to patients’ best interests. We have highlighted statistical methods that we firmly believe are more relevant than conventional methodologies to the scientific and ethical imperatives of clinical research.



Table 1 Comparison Between the Design Employed in VNP Trial with an Alternative Design Based On a Null Rate Equal to the Actual Historical Rate

	Employed Design			Alternative Design		
	$p_0=.05, p_1=.20$			$p_0=.40, p_1.55$		
	$n_{max}=37$			$n_{max}=88$		
Interim Test	At n=12			At n=38		
p_E	PET	ER	ERL (%)	PET	ER	ERL (%)
.05	.54	6.6	8.2 (56)	1.00	21.9	13.3 (38)
.10	.28	5.8	9.0 (61)	1.00	23.8	11.4 (32)
.20	.07	7.7	7.1 (48)	1.00	27.6	7.6 (22)
.30	.01	11.1	3.7 (25)	.96	31.2	4.0 (11)
.40	.002	14.8	0	.67	35.2	0
.50	<.001	18.5	-3.7 (-25)	.21	43.0	-7.8 (-22)
.55	<.001	20.3	-5.6 (-37)	.08	47.8	-12.6 (-36)

PET = probability of early termination

ER = expected number of responses

ERL= expected number of responses lost (see text)

Table 2 Contrasts Between Simon Optimal 2-Stage and Bayesian Multistage Designs, both based on $p_0 = 0.40$ and $p_1 = 0.55$ with maximum sample size 88 patients.

Interim Tests	Simon Optimal 2-stage			Bayesian multi-stage		
	One at n=38			Up to five at n=15, 30, 45, 60, 75		
True p	PET	ER	ERL (%)	PET	ER	ERL (%)
.05	1.00	21.9	13.3 (38)	1.00	31.1	4.1 (12)
.10	1.00	23.8	11.4 (32)	1.00	31.5	3.7 (10)
.20	1.00	27.6	7.6 (22)	1.00	32.1	3.1 (9)
.30	.96	31.2	4.0 (11)	1.00	32.9	2.3 (6)
.40	.67	35.2	0	.85	35.2	0
.50	.21	43.0	-7.8 (-22)	.29	42.7	-7.5 (-21)
.55	.08	47.8	-12.6 (-36)	.09	47.7	12.5 (-36)

PET = probability of early termination

ER = expected number of responses

ERL= expected number of responses lost (see text)

References

- Chen, TT. Optimal three-stage designs for phase II cancer clinical trials. *Statistics in Medicine*, 16:2701-2711, 1997.
- Estey EH, Thall PF. New designs for phase 2 clinical trials. *Blood*, 102: 442-448, 2003.
- Fleming, TR. One-sample multiple testing procedure for phase II clinical trials *Biometrics*, 38:143-151, 1982.
- Inoue LYT, Thall PF, Berry, DA. Seamlessly expanding a randomized phase II trial to phase III. *Biometrics*, 58:823-831, 2002.
- Liu, Q and Pledger, GW. Phase 2 and 3 combination designs to accelerate drug development *Journal of the American Statistical Association*, 100, 493-502, 2005.
- Schaid, DJ, Wieand, S and Therneau, TM. Optimal two-stage screening designs for survival comparisons *Biometrika*, 77, 507-513, 1990.
- Simon, R. Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials*, 10:1-10, 1989.
- Simon, R, Wittes, RE and Ellenberg, SS. Randomized phase II clinical trials. *Cancer Treatment Reports*. 69:1375-1381, 1985.
- Thall PF, Sung H-G. Some extensions and applications of a Bayesian strategy for monitoring multiple outcomes in clinical trials. *Statistics in Medicine*, 17:1563-1580, 1998.
- Thall PF. Ethical issues in oncology biostatistics. *Statistical Methods in Medical Research*, 11:429-448, 2002.
- Thall PF, Simon R, Ellenberg SS. Two-stage selection and testing designs for comparative clinical trials. *Biometrika* 75: 303-310, 1988.
- Thall PF, Simon R, Estey EH. Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. *Statistics in Medicine* 14:357-379, 1995.
- Thall PF, Simon RM, Estey EH. New statistical strategy for monitoring safety and efficacy in single-arm clinical trials. *J Clinical Oncology* 14:296-303, 1996.
- Thall PF, Sung H-G, Estey EH. Selecting therapeutic strategies based on efficacy and death in multi-course clinical trials. *J American Statistical Assoc*, 97:29-39, 2002.
- Thall PF, Wang X. Bayesian sensitivity analyses of confounded treatment effects. In: J. Crowley and D. Pauler (eds.), *Handbook of Statistics in Clinical Oncology: Second Edition, Revised and Expanded*, New York: Marcel-Dekker, 2005. In press.

Thall PF, Wathen JK, Bekele BN, Champlin RE, Baker LO, Benjamin RS. Hierarchical Bayesian approaches to phase II trials in diseases with multiple subtypes. *Statistics in Medicine*, 22: 763-780, 2003.

Thall PF, Wathen JK. Covariate-adjusted adaptive randomization in a sarcoma trial with multi-stage treatments. *Statistics in Medicine*, 27:1947-1964, 2005.

